UNIVERSITY OF CALIFORNIA SAN DIEGO

Super-resolution under Extreme Sampling Constraints: Theory and Algorithms

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Electrical Engineering
(Signal and Image Processing)

by

Pulak Sarangi

Committee in charge:

Professor Piya Pal, Chair
Professor Yeshaiahu (Shaya) Fainman
Professor Takaki Komiyama
Professor Truong Q. Nguyen
Professor Bhaskar D. Rao

2023

The Dissertation of Pulak Sarangi is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First of all, I am extremely grateful to my advisor Professor Piya Pal for accepting me as a graduate student in her group that made a tremendous impact on my career trajectory. I was inspired by her lectures and her teaching style that truly made me appreciate how building a strong foundation can lead to the demystification of seemingly complicated concepts. Throughout this journey, she has always been accessible, extremely generous with her time, and genuinely invested in my growth and success. She has always been extremely patient with me, providing constructive feedback at every stage starting from the presentation of technical ideas to writing them down in the form of a research paper. Another lifelong lesson that I learnt from her is the practice of scientific integrity, where we always try to reinspect our results several times from different perspectives.

Next, I would like to thank Professor Shaya Fainman, Professor Takaki Komiyama, Professor Truong Nguyen, and Professor Bhaskar Rao for serving on my committee and providing their valuable time and feedback.

I would like to acknowledge the stimulating conversations with Dr. Ryoma Hattori and Professor Takaki Komiyama regarding calcium imaging that inspired several of the results relating to binary priors in this thesis. I also had a fruitful collaboration with Dr. Phuong Nguyen and Dr. Shimon Rubin and I got to learn about Surface-enhanced Raman spectroscopy (SERS).

I would like to thank all my labmates Dr. Heng Qiao, Dr. Ali Koochakzadeh, Dr. Robin Rajamäki, Mehmet Can Hücümenoglu, Jiawen Chen, Sina Shahsavari, Parthasarathi Khirwadkar and Yinyan Bu for intellectual discussions during the group meetings, collaborations and acting as a second family away from home.

I would also like to thank my friends in San Diego Mainak Biswas, Nishant Bhaskar, Anant Dhayal, Suganth Krishna, Thyagarajan Venkatanarayanan for the fun memories that will last for a lifetime.

Finally, I would like to thank my parents Mr. Himansu Sarangi and Mrs. Sangeeta Sarangi for their patience and support.

Chapter 2, in part, is a reprint of the material as it appears in the following papers:

- P. Sarangi, Ryoma Hattori, Takaki Komiyama and P. Pal, “Super-resolution with Binary Priors: Theory and Algorithms,” IEEE Transactions on Signal Processing, 2023.

- P. Sarangi and P. Pal, “No Relaxation: Guaranteed Recovery of Finite-Valued Signals from Undersampled Measurements,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5440-5444.

The dissertation author was the primary investigator and author of these papers.

Chapter 3, in part, is a reprint of the material as it appears in P. Sarangi and P. Pal, “Measurement Matrix Design for Sample-Efficient Binary Compressed Sensing,” in IEEE Signal Processing Letters, vol. 29, pp. 1307-1311, 2022.

The dissertation author was the primary investigator and author of these papers.

Chapter 4, in part, is a reprint of the material as it appears in the following papers:

- P. Sarangi, M. C. Hücümenoglu and P. Pal, “Beyond Coarray MUSIC: Harnessing the Difference Sets of Nested Arrays With Limited Snapshots,” in IEEE Signal Processing Letters, vol. 28, pp. 2172-2176, 2021.

- P. Sarangi, M. C. Hücümenoglu and P. Pal, “Single-Snapshot Nested Virtual Array Completion: Necessary and Sufficient Conditions,” IEEE Signal Processing Letters, vol. 29, pp. 2113-2117, 2022.

The dissertation author was one of the primary investigator and author of these papers.

Chapter 5, in part, is a reprint of the material as it appears in the following papers:

- P. Sarangi, M. C. Hücümenoglu and P. Pal, “Effect of Undersampling on Non-Negative Blind Deconvolution with Autoregressive Filters,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 5725-5729.

- P. Sarangi, M. C. Hücümenoglu and P. Pal, ”Understanding Sample Complexities for Structured Signal Recovery from Non-Linear Measurements,” 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Le gosier, Guadeloupe, 2019, pp. 81-85.

- P. Sarangi, H. Qiao and P. Pal, “On the role of sampling and sparsity in phase retrieval for optical coherence tomography,” 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Curacao, 2017, pp. 1-5.

The dissertation author was one of the primary investigator and author of these papers.

VITA

2015	Bachelor of Technology, Indian Institute of Technology, Kharagpur
2018	Master of Science, University of California San Diego
2023	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

P. Sarangi, R. Hattori, T. Komiyama and P. Pal, “Super-resolution with Binary Priors: Theory and Algorithms,” IEEE Transactions on Signal Processing, 2023.

P. Sarangi, M. C. Hücümenoglu, Robin Rajamäki and P. Pal, “Super-resolution with Sparse Arrays: A Non-Asymptotic Analysis of Spatio-temporal Trade-offs,” Submitted to IEEE Transactions on Signal Processing, 2023.

P. Sarangi, M. C. Hücümenoglu and P. Pal, “Single-Snapshot Nested Virtual Array Completion: Necessary and Sufficient Conditions,” IEEE Signal Processing Letters, vol. 29, pp. 2113-2117, 2022.

P. Sarangi and P. Pal, “Measurement Matrix Design for Sample-Efficient Binary Compressed Sensing,” IEEE Signal Processing Letters, vol. 29, pp. 1307-1311, 2022.

P. H. L. Nguyen, S. Rubin, P. Sarangi, P. Pal, and Y. Fainman, “SERS-based ssDNA composition analysis with inhomogeneous peak broadening and reservoir computing,” Appl. Phys. Lett., vol. 120, no. 2, p. 023701, 2022.

S. Shahsavari, P. Sarangi, and P. Pal, “Beamspace esprit for mmwave channel sensing: Performance analysis and beamformer design,” Frontiers in Signal Processing 1, p. 20., 2022. [2022 Outstanding Article award (Signal Processing for Communications section), Invited Paper]

P. Sarangi, M. C. Hücümenoglu and P. Pal, “Beyond Coarray MUSIC: Harnessing the Difference Sets of Nested Arrays With Limited Snapshots,” IEEE Signal Processing Letters, vol. 28, pp. 2172-2176, 2021.

M. C. Hücümenoglu, P. Sarangi, Robin Rajamäki and P. Pal, “To Regularize or Not to Regularize: The Role of Positivity in Sparse Array Interpolation with a Single Snapshot,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.

S. Shahsavari, P. Sarangi, M. C. Hücümenoglu and P. Pal, “Ada-JSR: Sample Efficient Adaptive Joint Support Recovery From Extremely Compressed Measurement Vectors,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 9077-9081.

J. Millhiser, P. Sarangi and P. Pal, “Initialization-Free Implicit-Focusing (IF2) for Wideband Direction-of-Arrival Estimation,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 4973-4977.

M. C. Hücümenoglu, P. Sarangi and P. Pal, “Exploring Fundamental Limits of Spatiotemporal Sensing for Non-Linear Inverse problems,” 55th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2021, pp. 1409-1413.

S. Shahsavari, P. Sarangi and P. Pal, “KR-LISTA: Re-Thinking Unrolling for Covariance-Driven Sparse Inverse Problems,” 55th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2021, pp. 1403-1408.

P. Sarangi and P. Pal, “No Relaxation: Guaranteed Recovery of Finite-Valued Signals from Undersampled Measurements,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5440-5444.

P. Sarangi, S. Shahsavari and P. Pal, “Robust DOA and Subspace Estimation for Hybrid Channel Sensing,” 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2020, pp. 236-240.

H. Qiao, P. Sarangi, Y. Alnumay and P. Pal, “Sample complexity trade-offs for synthetic aperture based high-resolution estimation and detection,” IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM), Hangzhou, China, 2020, pp. 1-5. **[Finalist, Best Student Paper Award]**

P. Sarangi, M. C. Hücümenoglu and P. Pal, “Effect of Undersampling on Non-Negative Blind Deconvolution with Autoregressive Filters,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 5725-5729.

P. Sarangi, M. C. Hücümenoglu and P. Pal, ”Understanding Sample Complexities for Structured Signal Recovery from Non-Linear Measurements,” IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Le gosier, Guadeloupe, 2019, pp. 81-85. **[Best Student Paper Award (First Position)]**

P. Sarangi and P. Pal, “Robust Sparse Phase Retrieval from Differential Measurements Using Reweighted L1 Minimization,” IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM), Sheffield, UK, 2018, pp. 223-227.

P. Sarangi and P. Pal. “Superresolution via bilinear fusion of multimodal imaging data,” In Big Data: Learning, Analytics, and Applications, volume 10989, pages 128 – 134. International Society for Optics and Photonics, SPIE, 2019. [Invited Paper]

P. Sarangi, H. Qiao and P. Pal, “On the role of sampling and sparsity in phase retrieval for optical coherence tomography,” IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Curacao, 2017, pp. 1-5.

ABSTRACT OF THE DISSERTATION

Super-resolution under Extreme Sampling Constraints: Theory and Algorithms

by

Pulak Sarangi

Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)

University of California San Diego, 2023

Professor Piya Pal, Chair

High dimensional inverse problems are at the heart of numerous modern signal processing and machine learning applications, where the goal is to sense the physical environment and infer parameters of interest residing in a high-dimensional ambient space, from low-dimensional (non)-linear measurements. Despite the rapid growth in the volume of data that is being generated in the form of images, videos, and sensors, there are restrictions imposed by the physical constraints of the sensing system. For instance, in a scanning microscopy system, the frame rate and hence the temporal resolution is limited by the speed of the scanning mirrors and the size of the field of view (FOV). Similarly, in the modern hybrid mmWave systems, although a massive number

of antennas are deployed, the number of Radio Frequency (RF) chains is scarce due to their high cost and power consumption. Therefore, it is crucial to design sensing paradigms that can reliably recover the information of interest under such stringent sampling budgets. This requires leveraging the underlying "low-dimensional" geometry of the signal (known as priors) to enable reconstruction with relatively few measurements. Over the last two decades, several theoretical and algorithmic techniques have been developed for tackling these under-determined systems, the most well-known among them being sparse and low-rank signal/image reconstruction.

In this thesis, we primarily focus on the "ill-posed" inverse problem of super-resolution with "extreme spatial/temporal sampling constraints" arising from real-world applications in Neural Spike Deconvolution and Sensor Array Signal Processing. We especially focus on unconventional regimes where existing approaches based on sparsity, correlation and/or low rank priors may fail. Broadly, super-resolution is concerned with the reconstruction of temporally/spatially localized events (or spikes) from samples of their convolution with a low-pass filter. The problem becomes ill-posed due to systematic attenuation of the high-frequency content of the underlying spikes. Unlike classical compressed sensing, the under-sampling operation in super-resolution does not correspond to observing random linear projections of the unknown signal of interest. This prevents direct application of existing theoretical guarantees developed in the compressed sensing literature. In contrast to prior works in super-resolution which exploit the role of sparsity and/or non-negativity priors to solve the resulting ill-posed problem, this thesis explores the problem of *Binary Super-resolution*, i.e., when the spike amplitudes are known apriori to be binary-valued. We demonstrate that enforcing binary priors in under-determined linear inverse problems can allow exact recovery (in absence of noise) in the non-standard regime where the sparsity level of the spikes far exceeds the number of acquired measurements — which henceforth is referred to as "extreme compression" regime. Past works have shown that it is possible to operate in such a regime by leveraging multiple snapshots and statistical priors. On the contrary, this thesis shows that multiple measurements may not be necessary to operate in extreme compression in the face of binary constraints. We also show that standard convex-

relaxation techniques for the binary constraint, such as box-constraints (that are widely used in Binary compressed sensing), are inadequate for operating in extreme compression regimes and they can introduce a certain bias in the support of the recovered spikes. We overcome the computational challenges of enforcing binary constraints by exploiting the special structure of the measurements that allow us to reformulate the problem as a binary search.

Despite the ability to recover finite-valued signals from uniformly downsampled measurements for most filters, there might exist certain "adversarial filters" that result in ambiguity upon uniform downsampling . We exhibit that the additional flexibility of designing the measurement matrix (beyond uniform-sampling) can mitigate the effects of adversarial filters. We propose a novel algorithm-measurement co-design framework where the measurement matrix is designed as a function of the filter. In absence of noise, this framework can provably achieve the optimal complexity of $\Omega(1)$, which is independent of dimension as well as sparsity of the binary signal. Moreover, the recovery can be performed using a greedy sequential decoding algorithm with low computational complexity.

The second half of this dissertation studies another class of problems with stringent requirements on available spatial and temporal measurements. This problem arises in passive sensing scenarios with sparse sensor arrays, where the goal is to perform source localization with very few spatial sensors. Sparse array geometries such as nested arrays and co-prime arrays have gained popularity due to their ability to identify more sources than sensors and offer very high resolution compared to a uniform array with the same number of physical sensors. The benefit is attributed to the ability of sparse arrays to exploit certain correlation structures in the source signals, without increasing the spatial sensing budget. However, it is commonly believed that sparse arrays inherently require a large number of temporal snapshots to obtain reliable correlation estimation, and therefore, they may not be preferred in the so-called "sample-starved" regimes, where the temporal snapshots are scarce. In applications like automotive radar and mmWave communication systems, the sources/multipaths may be coherent and the environment is dynamic due to the high mobility of the sources. Motivated by these challenging scenarios, it is desirable to identify the sources or multipath components with superior resolution while using very few temporal snapshots (only a single snapshot in the extreme case). The techniques developed in these sample-starved scenarios are largely based on heuristics. This thesis debunks some of the myths associated with sparse arrays in the limited snapshot regime by providing provable ways to leverage benefits of deterministic sparse arrays (nested arrays) in the following largely unexplored regimes: (i) with few snapshots (of the order of number of sources) and (ii) extreme-sample starved scenarios with only a single snapshot.

Chapter 1

Introduction

1.1 Background and Motivation

An integral part of future wireless systems is the inclusion of sensing capabilities in the communication network. This introduces many more devices beyond mobile phones into the wireless network such as smart sensors for automation, as well as intelligent monitoring systems for manufacturing and healthcare, to name a few. A naive acquisition strategy for sensing the environment from such a massive number of devices could result in large volumes of data. Most of these measurements are often used for important downstream tasks such as obstacle detection, surveillance, or smart monitoring systems in factories. This naturally involves either communicating or processing these massive amounts of data to make critical decisions with low-power hardware, and under strict memory and time constraints. The signals that we expect to encounter in these applications (and beyond) often have a latent low-dimensional representation. The underlying channel has a low-rank structure due to the sparse scattering which can be leveraged during channel sensing stage. During the communication stage, it is common to transmit messages that are typically chosen from a finite constellation, which can be leveraged for efficient decoding. Therefore, it is possible to deploy clever sensing strategies (such as sparse arrays) for signal acquisition and extract the information of interest in a fast and efficient manner by solving an inverse problem, which captures the underlying low-dimensional structure, described variously in terms of sparsity, low rank, geometric channel model and finite-value constraints.

Over the last decade, compressed sensing has become a well-known framework for reconstruction of high dimensional signals from very few measurements by leveraging suitable priors. Frequently studied priors include sparsity, non-negativity, and low-rank structure. In specific applications, the acquisition system itself imposes restrictions on the type of measurement that can be acquired. This can implicitly constrain both the sensing matrix and the sampling budget. For example, in the context of hybrid mmWave communication system, it is impractical to have a dedicated RF chain for each antenna and perform a fully digital processing due to the large number of antennas and high power consumption of the RF chains. Therefore, an analog front-end linearly combines the signal received at the antenna array to obtain low-dimensional measurements that can significantly reduce the hardware constraint associated with the ADCs. The number of measurements that are available to decode the communication message will be determined by the number of RF chains. In two-photon calcium imaging (used to detect neural spiking activity), the available measurements are samples obtained from temporal blurring of the spiking signal (blurring filter is controlled by the choice of the calcium indicator). In addition, the temporal sampling rate is determined by the FOV of imaging since the microscope performs raster scanning. Therefore, the temporal resolution achievable by such an acquisition strategy is severely limited. However, as we will demonstrate in this thesis, it is possible to overcome the conventional resolution bottleneck by leveraging binary priors and algorithmic reconstruction.

One of the central themes of this thesis is to *provably* operate in regimes with *stringent constraints on the sampling budget* (spatial or temporal) where existing approaches based on inappropriately chosen priors may fail. This will be primarily illustrated through the lens of real-world problems arising in two distinct applications, namely Neural Spike Deconvolution and Sensor Array Signal Processing with applications in autonomous sensing. An important consideration in these applications will be the ability to perform either temporal or spatial super-resolution, i.e., the ability to resolve two closely located spikes (temporal) or targets (spatial) with limited data. We are interested in both the theory behind these problems and the development of efficient algorithms. Depending on the setting, we design computationally efficient techniques operating with optimal sample complexity by exploiting the measurement structure and/or co-designing the sensing operator and the algorithm.

1.2 Outline and Summary of Key Contributions in the Dissertation

We briefly summarize our key contributions in three classes of inverse problems within the overarching theme of this thesis, and distinguish them from prior work.

The first part of the thesis focuses on underdetermined signal models involving finite-valued constraints (Binary compressed sensing). It is devoted to development of new lower bounds on sample complexities, and relaxation-free algorithmic schemes to attain those bounds. The contributions are summarized below:

Role of Binary Priors in Linear Inverse Problems

- **Structured Linear Inverse Problems with Binary Constraints:** In Chapter 2, we provide the first identifiability results for a compressive convolutional model with binary constraints. We specifically consider *structured* linear operators which take the form $\mathbf{A} = \Phi\mathbf{H}$. The matrix $\mathbf{H}$ models the convolution operation with a known filter, whereas Φ captures the effect of under-sampling, which can be either a dense or uniform sub-sampling operator. The existing results developed for arbitrary linear operators cannot be trivially applied to obtain identifiability results for the structured measurement model. One of the key contributions of our results is to bring out the nuances of the interaction between the filter and the measurement matrix in determining the overall sample-complexity.

- **Sensing of Finite-Valued Signals with Uniform Undersampling: A super-resolution framework**: Sparse signal recovery typically utilizes dense compression/sketching operator, since in the worst case (in terms of the support of the sparse signal) adopting a uniform sub-sampling operator fails to capture contribution from certain non-zero elements altogether. Unlike standard results in compressed sensing, our results are first to demonstrate that it is possible to exactly recover finite-valued signals even from *uniformly downsampled* measurements (Section 2.2 and 2.5) without exploiting any additional structure such as sparsity. A direct consequence of this is the ability to achieve "super-resolution" where we can sense/infer signals with features on a much finer scale from measurements that are sampled at a significantly lower sampling rate. Our analysis leads to new insights into the interplay between binary/finite-valued priors and the undersampling limit, which is essentially determined by the length of the filter. In this thesis, neural spike deconvolution in calcium imaging serves as a prototypical example to illustrate how the theoretical insights can be applied to achieve super-resolution capability. However, these ideas can be translated to many other practical problems such as massive MIMO communication, and detection using DNA microarrays, where the signal of interest is binary.

- **Extreme Compression with Binary Priors**: We show that binary constraints allow us to operate in the so-called "extreme compression" regime, where the number of measurements can be significantly smaller than the sparsity level of the signal. The ability to operate in extreme compression regime has been showcased in the context of the Multiple Measurement Vector (MMV) model, where it is possible to identify supports of size which are much larger than the dimension of each measurement vector [1–4]. This feat is made possible by exploiting multiple measurements along with a statistical priors on the sparse signals. In contrast to the prior works in MMV setting, we show that it is fundamentally possible to operate in the extreme compression regime even in the Single Measurement Vector (SMV) model due to the binary constraint.

- **Relaxation-free Algorithms for enforcing Binary Constraints**: The strong identifiability guarantees for recovering a binary vector often do not necessarily translate to computationally efficient algorithms. If the problem is identifiable, one can always perform exhaustive search to recover the desired binary vector, however, such a solution is impractical. Therefore, a major focus of binary compressed sensing has been on developing computationally efficient algorithms, often by designing relaxations such as box constraints. An inevitable consequence of this relaxation is an increase in the required sample complexity. One of the key questions addressed in this thesis is to understand if the benefits of finite-valued constraints can be maximally leveraged in a computationally efficient manner, avoiding potentially suboptimal relaxations. Our algorithmic solutions are inspired by the notion of "β-expansion", which is concerned with finding the generalized radix representation of real numbers [5–7].

- **Measurement-algorithm co-design**: Despite the fact that for a broad class of filters, any binary vector can be recovered from uniformly downsampled convolutional measurements, the characteristics of the underlying filter controls recoverability. As we will show, there could exist "adversarial filters" for which either uniform sub-sampling may introduce ambiguities or it becomes challenging to guarantee recovery using a low-complexity decoding algorithm, when uniform sub-sampling is employed. To mitigate these challenges posed by adversarial filters, Chapter 3 of this thesis investigates the question "Can the flexibility to design a filter-dependent sampler overcome the challenges posed by an adversarial filter?" We provide an affirmative answer to this question by proposing an "*algorithm-measurement co-design framework*" that can attain optimum sample complexity with a computationally efficient sequential decoding algorithm.

Harnessing Benefits of Sparse Arrays in Sample-Starved Regime

Structured sparse array geometries are being actively studied due to their superior-resolution capabilities and ability to identify more sources than sensors. Typically, these benefits rely on estimating a "virtual coarray covariance matrix" depicting the correlation between the received signals at different sensors. The number of unknowns in this "virtual coarray covariance matrix" can grow up to quadratically with the number of physical sensors. Therefore, it is often believed that reliably estimating the coarray covariance matrix requires large number of snapshots. This leads to the impression that the benefits of sparse arrays come at the expense of additional temporal measurements. A contribution of this thesis is to answer the question: “Is it possible to harness the benefits offered by sparse arrays with a limited number of snapshots (in the sample-starved scenarios) if the goal is to identify fewer sources than the number of sensors with high-resolution?” Our investigation is motivated by contemporary applications such as autonomous sensing and mmWave channel estimation where identifying more sources than sensors is not necessary, rather it is crucial to operate under a severe restriction on the number of snapshots either due to coherent multipaths or a rapidly changing environment. The contributions specific to this problem are given below:

- **Proxy covariance matrix**: With only a few snapshots (of the order of the number of sources), popularly used algorithms for sparse arrays, such as Coarray MUSIC, incurs large estimation error, which saturates away from zero even as the signal-to-noise ratio (SNR) tends to infinity. Such a behaviour is undesirable especially when there are fewer sources than sensors where a Uniform Linear Array (ULA) can succeed. In Chapter 4 of this thesis, we propose moving away from estimating the coarray covariance matrix when snapshots are limited, and instead obtaining a biased estimate of the covariance matrix called a “proxy covariance matrix” (Prox-Cov). (Prox-Cov) aims to identify the coarray subspace (not the source powers) instead of the entire covariance matrix. We prove that in the noiseless setting, when the number of sources is of the order of the number of sensors, it is possible to exactly identify the desired subspace by formulating a convex optimization problem and thereby, overcoming a key limitation of coarray MUSIC-type algorithms.

- **Extreme-Sample Starved Regime**: In the extreme scenario, when only a single-snapshot of measurement is available for DOA estimation, application of subspace based techniques such as MUSIC or ESPRIT is no longer straightforward. For a ULA, techniques such as spatial-smoothing can be adopted. However, it is not straightforward to translate these ideas for sparse arrays. Recently, interpolation techniques have been developed that attempt to synthesize a virtual ULA by estimating the missing measurements. The virtual measurements can be arranged in the form of a low-rank Hankel/Toeplitz matrix, and the measurements acquired by the sparse array only reveal certain entries of this matrix, leading to a structured low-rank matrix completion problem. However, existing guarantees from matrix completion cannot be directly applied for deterministic sparse arrays such as nested array. Chapter 4 of this thesis provides matching necessary and sufficient conditions to perfectly interpolate the virtual array of a nested array via rank-minimization, provided the number of sources is not too large.

Non-Linear Inverse Problems We conclude the thesis by exploring the role of binary constraints for blind-deconvolution and sparsity in phase retrieval. These are bilinear and quadratic inverse problems, respectively, and are often considered as prototypical examples of ill-posed inverse problems beyond linear measurement models. The contributions are listed below:

- **Parametric Blind-Deconvolution**: In the binary super-resolution work discussed earlier, we assumed the underlying filter to be known apriori. However, in practice, only a parametric representation for the kernel maybe known leading to a "blind" super-resolution problem. In Chapter 5, we show that if the spikes are generated according to a Bernoulli model, it is possible to uniquely identify both the signal and the kernel with high probability from uniformly downsampled measurements.

- **Redundancy of Priors in Sparse Phase Retrieval:** In Chapter 5, we consider the problem of recovering a sparse signal from its quadratic measurements also known as *sparse phase retrieval*. It can be shown that after applying a well-known linearization technique called lifting, the sparse phase retrieval problem can be cast as searching for a simultaneously *sparse, low-rank and positive semi-definite matrix* that is consistent with the measurements. However, adopting convex relations that incorporate penalties for both low-rankness and sparsity result in a sub-optimal sample complexity (scaling quadratically in sparsity). One of the contributions of this thesis is a modified formulation in the lifted space that can exactly recover the signal with an optimal sample complexity (scaling linearly in sparsity). Our formulation demonstrates that it is possible to only impose positive semi-definite (PSD) and sparsity constraints on the lifted variable and completely eliminate the need for trace minimization without sacrificing sample-complexity.

Chapter 2

Basics of Linear Inverse Problems with Finite-Valued Priors

In this chapter, we provide a brief overview of the widely studied underdetermined linear inverse problems, and the different types of priors that are commonly used to make them well-posed. Consider an underdetermined system of linear equations given by:

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $M < N$. If we assume $\text{rank}(\mathbf{A}) = M$, this system of equations has infinitely many solutions. In engineering applications, we are interested in recovering a ground-truth signal with certain desirable properties, which is known as the prior on $\mathbf{x}$. In the standard compressed sensing problem, the goal is to recover the sparsest solution, i.e., the vector with the smallest number of non-zero entries. An important property called the Kruskal rank of the measurement matrix $\mathbf{A}$ determines the uniqueness and recoverability of the sparsest solution. A matrix $\mathbf{A}$ is said to have a Kruskal rank of r if any subset of r columns of $\mathbf{A}$ are linearly independent and there exists at least one subset of $r+1$ columns that are linearly dependent. It is well-known that the condition $\text{krank}(\mathbf{A}) \geq 2s$ is necessary and sufficient to recover all sparse signals with sparsity at most s [8]. Therefore, the number of measurements M required to identify s-sparse vector scales only linearly with the sparsity level $M = \Omega(s)$ and is independent of the ambient dimension N.

In this thesis, instead of sparsity, we leverage the prior knowledge that the non-zero entries of the ground truth $\mathbf{x}_0$ take values from a finite-set ($\mathscr{A}$) of cardinality q:

$$[\mathbf{x}_0]_i \in \mathscr{A} \subset \mathbb{R} \quad i = 1, 2, \cdots, N \text{ where } |\mathscr{A}| = q.$$

For instance, the underlying signal can be binary-valued, i.e., $q = 2$. Binary signals are encountered in a diverse set of applications such as neural spike decoding [9–11], communication systems [12], DNA compressed sensing microarrays [13] and discrete tomography [14, 15].

In neural spike decoding, the neural spiking activity at a fine temporal scale is modeled as a binary-valued signal (spike or no spike) [9,11]. The neural spiking activity is typically measured indirectly, and only a limited number of samples are available to decode the underlying activity. This constraint on the number of samples is imposed by the acquisition hardware. In massive Multiple-Input and Multiple-Output (MIMO) communication systems, the downlink system can be “Overloaded MIMO” system, where the number of antennas at the mobile user can be significantly smaller than the number of antennas at the base station due to size/power constraints [16, 17]. The decoding problem in such scenarios involves recovering the transmit signals belonging to a finite alphabet/constellation. Another contemporary application is motivated by the growing demand for massive machine to machine type communication in Internet of Things (IoT) applications. In such scenarios, a large number of users transmit data to the same receiver. Recently, in the “unsourced random access” model multiple users communicate using a shared codebook [18]. A binary signal can be used to indicate which message from the codebook was transmitted by one of the active users. The task of the decoder is to identify the binary vector, which in turn can be mapped to a list of transmitted messages. In single antenna systems, it is desirable to support a large number of users with very few channel uses. All these applications give rise to the core mathematical problem of solving an underdetermined linear inverse problem where the signal of interest comes from a finite set.

One of our central focus is to characterize the identifiability conditions for linear inverse problems with such finite-valued priors analogous to the sparse recovery problem. These identifiability results are complemented by developing computationally efficient algorithms which can achieve the optimal sample complexity. In this chapter, a special emphasis will be placed on the problem of "super-resolution" where the measurement matrix is structured (instead of being a generic linear operator), consisting of the composition of a convolutional operator followed by uniform down-sampling.

2.1 Prior Works

Early works on super-resolution date back to algebraic/subspace-based techniques such as Prony's method, MUSIC [19, 20], ESPRIT [21, 22] and matrix pencil [23, 24]. Following the seminal work in [25], substantial progress has been made in understanding the role of sparsity as a prior for super-resolution [26–28]. In recent times, convex optimization-based techniques have been developed that employ Total Variational (TV) norm and atomic norm regularizers, in order to promote sparsity [26–30] and/or non-negativity [31–33]. These techniques primarily employ sampling in the Fourier/frequency domain by assuming the kernel $h(t)$ to be (approximately) bandlimited. However, selecting the appropriate cut-off frequency is crucial for super-resolution and needs careful consideration [27, 34]. Unlike subspace-based methods, theoretical guarantees for these convex algorithms rely on a minimum separation between the spikes, which is also shown to be necessary even in absence of noise [35]. The finite rate of innovation (FRI) framework [36–40] also considers the recovery of spikes from measurements acquired using an exponentially decaying kernel, which includes the AR(1) filter considered in this section. In the absence of noise, FRI enables the exact recovery of K spikes with *arbitrary amplitudes* from $M = \Omega(K)$ measurements, without any separation condition [38]. It is to be noted that all of the above methods require $M > K$ measurements for resolving K spikes. In contrast, we will show that it is possible to recover K spikes from $M \ll K$ measurements by exploiting the *binary nature of the spiking signal*. The above algorithms are designed to handle *arbitrary real-valued* *amplitudes* and as such, they are oblivious to binary priors. Therefore, they cannot successfully recover spikes in the regime $M < K$, which is henceforth referred to as the *extreme compression regime*.

The problem of recovering binary signals from underdetermined linear measurements (with more unknowns than equations/measurements) has been recently studied under the parlance of Binary Compressed Sensing (BCS) [41–50]. In BCS, the undersampling operation employs random (and typically dense) sampling matrices, whereas we consider a deterministic and structured measurement matrix derived from a filter, followed by uniform downsampling. Moreover, existing theoretical guarantees for BCS crucially rely on sparsity assumptions that will be shown to be inadequate for our problem (discussed in Section II-C). Most importantly, in order to achieve computational tractability, BCS relaxes the binary constraints and solves continuous-valued optimization problems. Consequently, their theoretical guarantees do not apply in the extreme compression regime $M < K$.

One of the motivations for our study is the problem of neural spike deconvolution arising in calcium imaging [10, 11, 38, 51–54]. A majority of the existing spike deconvolution techniques [11, 51, 53] infer the spiking activity at *the same (low) rate at which the fluorescence signal is sampled*, and a single estimate such as spike counts or rates are obtained over a temporal bin equal to the resolution of the imaging rate. Although sequential Monte-Carlo based techniques have been proposed that generate spikes at a rate higher than the calcium frame rate [10], no theoretical guarantees are available that prove that these methods can indeed uniquely identify the high-rate spiking activity. Algorithms that rely on sparsity and non-negativity [51, 53] alone are ineffective for inferring the neural spiking activity that occurs at a much higher rate than the calcium sampling rate. On the other hand, at the high-rate, the spiking activity is often assumed to be binary since the probability of two or more spikes occurring within two time instants on the fine temporal grid is negligible [9, 55]. Therefore, we propose to exploit the inherent binary nature of the neural spikes and provide the first theoretical guarantees that it is indeed possible to resolve the high-rate binary neural spikes from calcium fluorescence signal acquired at a much lower rate.

2.2 Convolutional Compressive Model

Consider the problem of recovering a finite-valued signal $\mathbf{x} \in \mathscr{A}^N$ from compressed measurements of its convolution with a known filter:

$$\mathbf{z} = \Phi\mathbf{H}\mathbf{x} \tag{2.2}$$

Here, $\mathbf{H}$ is the linear operator associated with the filtering operation whereas Φ is an under-sampling operator. In this chapter, we analyze the case when Φ denotes a uniform under-sampling operator.

2.2.1 Identifiability of Finite-Valued Signal: Finite-Impulse Response Filter

We consider the problem of recovering an unknown unipolar finite-valued signal $\mathbf{x} = [x_0, x_1, \cdots, x_{N-1}]^T$ whose entries x_i's take values from the set of integers $\mathscr{A} = \{0, 1, \cdots, q-1\}$ and $q > 0$. Our goal is to recover $\mathbf{x}$ from undersampled measurements of its convolution with a known finite impulse response filter $\mathbf{h} = [h_0, h_1, \cdots, h_{L-1}]^T \in \mathbb{R}^L$ $(L < N)$. The measurement model is

$$\mathbf{z} = \mathbf{D}_\Omega \underbrace{\mathbf{H}\mathbf{x}}_{\mathbf{y}} \tag{2.3}$$

Here $\mathbf{H} \in \mathbb{R}^{P \times N}$ is a Toeplitz matrix with $P = N + L - 1$ that represents the discrete convolution operation:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & 0 & \cdots & 0 & 0 \\ h_1 & h_0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & h_{L-1} & h_{L-2} \\ 0 & 0 & 0 & \cdots & 0 & h_{L-1} \end{bmatrix}$$

and $\mathbf{D}_\Omega \in \mathbb{R}^{M \times P}$ (where $|\Omega| = M$) denotes a uniform downsampling operator given by:

$$[\mathbf{D}_\Omega]_{i,j} = \begin{cases} 1, & \text{if } j = \Omega_i \\ 0, & \text{otherwise} \end{cases}$$

A special case of this model involves binary valued signals (i.e. $q = 2$) and such binary valued signals, shapes or images have been considered in [42,44,48]. However, they relax the binary constraint to recover $\mathbf{x}$ using convex optimization, and theoretical guarantees are limited to random sampling. The recovery of finite-valued signal $\mathbf{x}$ from undersampled measurements is potentially an ill posed problem when $M \ll N$. Most existing approaches utilize sparsity constraints and tools from compressed sensing to solve this problem [41,42,44,45,48]. However, these approaches end up relaxing the finite-valued (binary) constraint to develop convex relaxations of the original non-convex problem. In this section, we will take a different approach and directly enforce the finite-value constraint. As shown later, this will enable us to recover $\mathbf{x}$ from undersampled measurements without having to explicitly enforce sparsity promoting penalties. We begin by developing conditions on the filter $\mathbf{h}, M$ and N under which the following linear map $\Phi : \mathscr{A}^N \to \mathbb{R}^M$ is injective:

$$\Phi(\mathbf{x}) := \mathbf{D}_\Omega \mathbf{H}\mathbf{x} \tag{2.4}$$

Theorem 1. *Let $\mathbf{D}_\Omega$ be a uniform subsampling operator where the index set is given by:*

$$\Omega = \{0, m, 2m, \cdots, (M-1)m\} \tag{2.5}$$

and $M = \lfloor N/m \rfloor$. Suppose the filter $\mathbf{h}$ is a random vector drawn from a distribution which is absolutely continuous with respect to the Lebesgue measure over $\mathbb{R}^L$. Then with probability 1, $\mathbf{x} \in \mathscr{A}^N$ is the unique solution of (2.3) if

$$M > \frac{N}{L}.$$

Conversely, if $M \leq \frac{N}{L}$ then the map $\Phi : \mathscr{A}^N \to \mathbb{R}^M$ is non-injective for every $\mathbf{h}$.

Proof. First assume $M > N/L$. Notice,

$$z_k = \mathbf{h}^\top \tilde{\mathbf{x}}_k,\ k \geq 1$$

where $\tilde{\mathbf{x}}_k \in \mathscr{A}^{L+1}$ is given by

$$[\tilde{\mathbf{x}}_k]_j = \begin{cases} x_{mk-j}, & j \leq mk \\ 0, & \text{else} \end{cases} ,0 \leq j \leq L \tag{2.6}$$

It is easy to see that $\Phi : \mathscr{A}^N \to \mathbb{R}^M$ is injective if the map $f_\mathbf{h} : \mathscr{A}^{L+1} \to \mathbb{R}$, $f_\mathbf{h}(\mathbf{x}) = \mathbf{h}^\top \mathbf{x}$ is injective. Define:

$$\mathscr{B} = \{\mathbf{h} \in \mathbb{R}^L, \exists\, \mathbf{x}, \mathbf{y} \in \mathscr{A}^{L+1}, \mathbf{x} \neq \mathbf{y} \text{ s. t. } \mathbf{h}^T(\mathbf{x} - \mathbf{y}) = \mathbf{0}\}$$

Therefore, $f_{\mathbf{h}}(\mathbf{x})$ is injective if $\mathbf{h} \notin \mathscr{B}$. Note that if $\mathbf{x}, \mathbf{y} \in \mathscr{A}^L$ then $\mathbf{x} - \mathbf{y} \in \bar{\mathbb{S}}^L$ where $\bar{\mathbb{S}} = \{-(q-1), -(q-2), \cdots, 0, \cdots, (q-2), (q-1)\}$ and $|\bar{\mathbb{S}}^L| = (2q-1)^L$. For every $\mathbf{v} \in \bar{\mathbb{S}}^L$, define

$$\mathscr{B}_{\mathbf{v}} = \{\mathbf{h} \in \mathbb{R}^L | \mathbf{h}^T \mathbf{v} = 0\}$$

Then, it is easy to see that $\mathscr{B} = \bigcup\limits_{v \in \bar{\mathbb{S}}^L \setminus \mathbf{0}} \mathscr{B}_{\mathbf{v}}$. Notice that $\mathscr{B}$ is a union of $(2q-1)^L$ sets. If $\mathbf{h}$ is a random vector generated from a continuous distribution over $\mathbb{R}^L$, for any fixed $\mathbf{v} \in \mathbb{R}^L$ we have

$$\mathbb{P}(\mathbf{h} \in \mathscr{B}_{\mathbf{v}}) = \mathbb{P}(\mathbf{h}^T \mathbf{v} = 0) = 0.$$

Therefore,

$$\mathbb{P}(\mathbf{h} \in \mathscr{B}) = \mathbb{P}(\mathbf{h} \in \bigcup_{v \in \bar{\mathbb{S}}^L \setminus \mathbf{0}} \mathscr{B}_{\mathbf{v}}) \leq (2q-1)^L \mathbb{P}(\mathbf{h} \in \mathscr{B}_{\mathbf{v}}) = 0$$

Hence, $\mathbf{h} \notin \mathscr{B}$ with probability 1, implying that Φ is injective.

Now, suppose $M \leq N/L$. We will show that there exists $\mathbf{x}' \in \mathscr{A}^L (\neq \mathbf{x})$ that satisfies $\mathbf{z} = \Phi(\mathbf{x}) = \Phi(x')$. Let $m = L+1$ then we have

$$z_0 = h_0 x_0, z_k := y_{k(L+1)} = \sum_{j=0}^{L-1} h_j x_{k(L+1)-j}, \; 1 \leq k \leq M$$

It is clear that none of the measurements $\{z_k\}_{k=0}^{M-1}$ is a function of x_1. Therefore, we can construct a signal $\mathbf{x}' \in \mathscr{A}^N$ as follows which satisfy $\mathbf{z} = \Phi(\mathbf{x}) = \Phi(\mathbf{x}')$:

$$x_i' = x_i \quad \forall \; i \neq 1 \text{ and } x_1' \neq x_1$$

□

Remark 1. *Theorem 1 shows that it is possible to downsample the output of the filter* $\mathbf{h}$ *by a factor of L and yet uniquely identify a finite-valued input to the filter. In contrast to existing* *approaches [42,44,48], this show that the finite valued constraint alone ensures injectivity of the overall map and it is not necessary to impose any additional sparsity constraint.*

2.2.2 Computationally Efficient Decoding Algorithm and Theoretical Guarantees

Since the map $\Phi : \mathscr{A}^N \to \mathbb{R}^M$ is injective with probability 1, it is always possible to design an exhaustive search decoder that searches over all $(D+1)^N$ possible signals in $\mathscr{A}^N$ to uniquely recover **x**. However, this is obviously computationally intractable. In this section, we present an efficient decoding algorithm whose complexity is $O(ML+L^2)$ that can provably recover **x** under a mild decay condition on the filter. This decoding is inspired by the notion of β-expansion introduced in [5] and further studied in [6,56]. However, our approach significantly departs from β-expansion as the filter coefficients maybe arbitrary and may not necessarily be exponents of a single positive real β.

Algorithm 1. Sequential Decoding with Filter Sorting

```
Input: Measurement $\mathbf{z}$, Filter $\mathbf{h}$, Subsampling operator $\Omega$,
Output: Estimate $\hat{\mathbf{x}} \in \mathscr{A}^N$
SORTING STEP
    $[\mathbf{h}', \Pi] \leftarrow$ SORT($\mathbf{h}$,'ascend') //Sorted filt. and perm. map
SEQUENTIAL BLOCK-WISE DECODING
$i \leftarrow 0$
Repeat
    $r \leftarrow z_i$
    $j \leftarrow L-1$
    Repeat
        $\hat{x}^b_{ij} \leftarrow \lfloor r/h'_j \rfloor$    //Divide and Round
        $r \leftarrow r - \hat{x}^b_{ij} h'_j$   //Update the residual
        $j \leftarrow j-1$
    until $j >= 0$
PERMUTATION STEP
$\hat{\mathbf{x}}^b_i \leftarrow PERM(\hat{\mathbf{x}}^b_i, \Pi)$ //Block perm. using sorted index
$i \leftarrow i+1$
until $i <= M$
$\hat{\mathbf{x}} \leftarrow$ MERGE($\{\hat{\mathbf{x}}^b_i\}_{i=0}^{M}$) //Merge decoded blocks
```

2.2.3 Sequential Decoding with Sorted Filters

We present our decoding algorithm in the form of Algorithm 1. Our algorithm includes a non-trivial sorting step where the filter coefficients are first sorted in an ascending order. Using these sorted coefficients, the decoding step sequentially recovers blocks of length L using a single measurement per block. Finally, the algorithm requires unscrambling to recover the correct representation.

Lemma 1. *The decoding complexity of Algorithm 1 is $O(ML+L^2)$.*

Proof. The first step involving sorting incurs a complexity of $O(L^2)$ snice the length of the filter $\mathbf{h}$ is L. Subsequent decoding of each block requires L operations of division, rounding, subtraction. This is repeated for each of the M blocks leading to an overall complexity of $O(ML+L^2)$. □

We next provide sufficient conditions on the kernel $\mathbf{h}$ for our algorithm to succeed. To this end, we introduce the notion of kernel decay.

Definition 1. *(**Kernel Decay**) Suppose $\mathbf{h} \in \mathbb{R}_+^L$ is a filter with non-negative coefficients where $\{h_i\}_{i=0}^{L-1}$ are distinct. Let $\mathbf{h}' \in \mathbb{R}^{L+1}$ be the sorted coefficients of the filter $\mathbf{h} \in \mathbb{R}_+^{L+1}$ such that:*

$$0 < h'_0 < h'_1 < \cdots < h'_L$$

The kernel decay parameter is defined as the maximum ratio

$$\rho(\mathbf{h}) = \max_{1 \le i \le L} \frac{h'_{i-1}}{h'_i} \tag{2.7}$$

Note that the kernel decay satisfies $0 < \rho(\mathbf{h}) < 1$.

Theorem 2. *Consider the measurement model (2.3) where Ω is given by (2.5). The output of the proposed Algorithm 1 coincides with the ground truth $\mathbf{x}$ if*

$$\rho(\mathbf{h}) \le \frac{1}{2D} \quad \text{and} \quad M \ge \frac{N}{L}$$

Proof. Recall that z_k can be represented as $z_k = \mathbf{h}^T \tilde{\mathbf{x}}_k$, where $\tilde{\mathbf{x}}_k \in \mathscr{A}^L$ is defined in (2.6). We will now establish that for each k, Algorithm 1 can exactly recover the block $\tilde{\mathbf{x}}_k$ from a single measurement z_k. Let $\mathbf{h}' = [h_{i_0}, h_{i_1}, \cdots, h_{i_{L-1}}]^T$ be the sorted filter coefficients (in ascending order). We define a permutation $\Pi : [L] \to [L]$ as

$$\Pi(i_j) = j, \quad 0 \le j \le L-1$$

. Using this permutation, for $0 \le j \le L-1$ we have

$$h'_j = h_{\Pi(i_j)}, \quad p^k_j = [\tilde{x}_k]_{\Pi(i_j)}$$

where $\mathbf{p}^k$ is a permuted block of $\tilde{\mathbf{x}}_k$. The measurements can be written as:

$$z_k = \sum_{j=0}^{L-1} h'_j p^k_j = \sum_{j=0}^{L-1} h_{\Pi(i_j)} [\tilde{x}_k]_{\Pi(i_j)}$$

Our proof proceeds by induction. For step $P = 1$, we observe that

$$\frac{z_k}{h'_{L-1}} = \sum_{j=0}^{L-2} p^k_j \frac{h'_j}{h'_{L-1}} + p^k_{L-1}$$

As a result of the decay assumption $\frac{h_{L-1}}{h_{L-1-i}} \le \frac{1}{(2D)^i}$ we have

$$\sum_{j=0}^{L-2} p^k_j \frac{h'_j}{h'_{L-1}} < 1$$

Therefore,

$$p^k_{L-1} \le \frac{z_k}{h'_{L-1}} < p^k_{L-1} + 1 \Rightarrow \left\lfloor \frac{z_k}{h'_{L-1}} \right\rfloor = p^k_{L-1}$$

and p^k_{L-1} is correctly recovered in the first step. We now assume that after $P < L$ iterations we have correctly identified $p^k_{L-1}, p^k_{L-2}, \cdots, p^k_{L-P}$. The residual is given by:

$$r = \sum_{j=0}^{L-P-2} p^k_j h'_j + h'_{L-P-1} p^k_{L-P-1}$$

Since $\sum_{j=0}^{L-P-2} p^k_j \frac{h'_j}{h'_{L-P-1}} < 1$, we can show that

$$\lfloor r/h'_{L-P-1} \rfloor = p^k_{L-P-1}$$

Therefore, we can correctly identify p^k_{L-P-1}. The proof therefore follows by induction. □

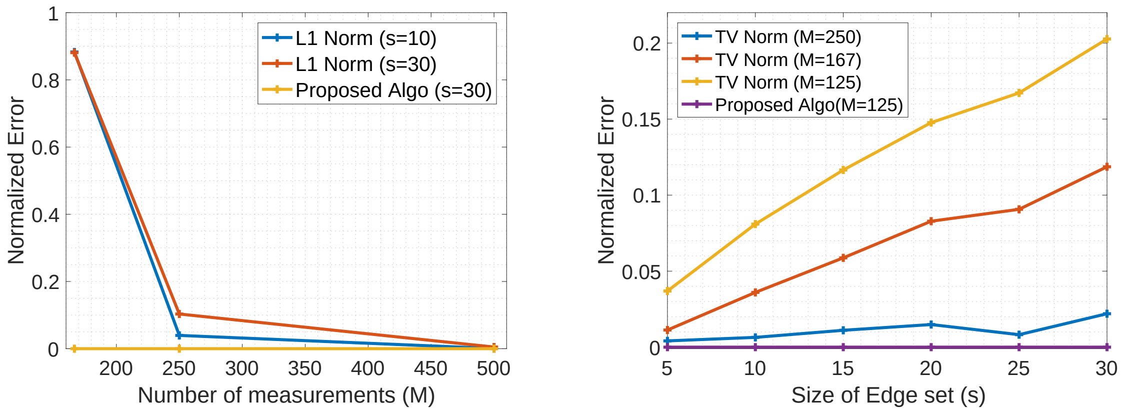
Remark 2. ***Sample Complexity:*** *Theorem 2 shows that if the decay condition is satisfied, then Algorithm 1 only requires around N/L measurements to succeed, irrespective of the sparsity level of* $\mathbf{x}$ *(even if $\|\mathbf{x}\|_0 \geq N/L$). Therefore, Algorithm 1 operates at the optimal subsampling regime determined by our injectivity result in Theorem 1.*

Remark 3. ***Decay condition:*** *For binary signals ($D = 1$), it can be shown that the decay condition $\rho(\mathbf{h}) \leq \frac{1}{2}$ can be satisfied by commonly used kernels such as Gaussian and exponential kernels by suitably choosing the sampling step and variance of the kernel.*

Uniform vs Non-uniform Sampling: *In contrast to [48], we develop recovery guarantees for uniform downsampling and our sample complexity does not depend on kernel incoherence parameter [48] or sparsity of the signal.*

2.3 Simulations

In the first experiment, we consider recovering a sparse binary signal x with $N = 500$ and the filter h is assumed to be a (truncated) $1-$D Gaussian filter of length 5 and variance 1. We evaluate the performance using the normalized error $\frac{\|\hat{x}-x\|_2}{\|x\|_2}$. Figure 2.1 (a) shows the average normalized error (over 100 Monte Carlo runs) for M between 135 to 500. We compare



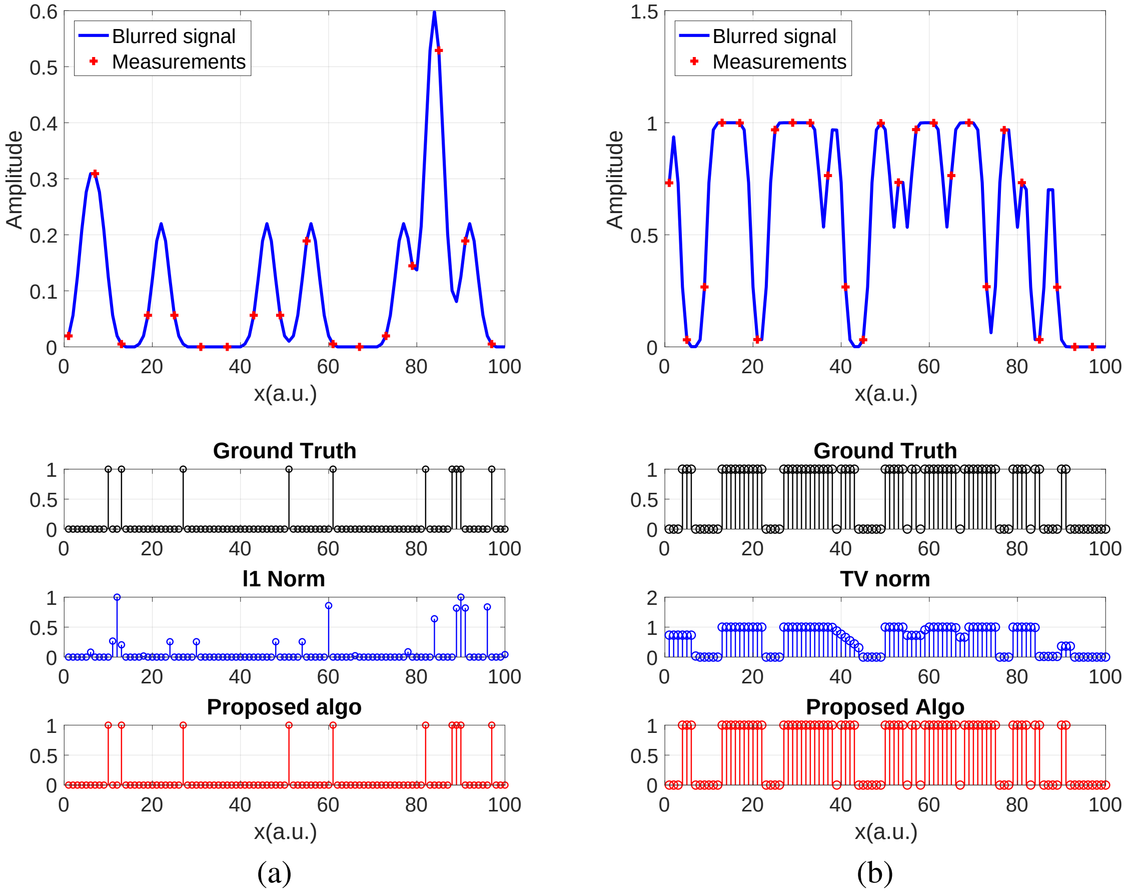
(a) (b)

Figure 2.1. (a) Normalized Error vs. M (Number of Measurements) for binary signal recovery using relaxed l_1 minimization and the proposed algorithm (b) Normalized Error vs. Size of edge set (s) for piece-wise constant signal recovery using relaxed TV norm and proposed algorithm

the proposed algorithm against a relaxed l_1 minimization approach as used in [41, 42].

$$\min_{u} \|u\|_1 \quad \text{s.t} \quad \Phi u = z, \quad \mathbf{0} \leq u \leq \mathbf{1}$$

As predicted by our theorem, the proposed algorithm can exactly identify the signal with maximum sparsity with only $M = 125$ measurements. However, the relaxation fails to operate in this regime as there are no guarantees for it to succeed with uniform subsampling. In Figure 2.2 (a), we display the recovered binary signals of length $N = 100$ with sparsity $s = 10$ recovered by both algorithms. It is clear that l_1 norm often predicts two spikes of smaller amplitude instead of a single spike with unit amplitude as a result of the relaxation. In the next experiment, we consider the case when x is a bi-level discrete piece-wise constant signal and the convolutional kernel is a $1-$D Gaussian filter with length 7 and variance 1. We compare the recovery performance of a relaxed discrete Total variational (TV) norm approach [57, 58] where the binary constraint is relaxed as $0 \leq x \leq 1$. The observation in Figure 2.1 (b) is consistent with our theoretical results as the proposed algorithm successfully recovers (with zero error) the piece-wise signal (binary) with only $M = 125$ measurements irrespective of the size of the edge set.



(a) (b)

Figure 2.2. (a)(Top) Blurred binary signal using Gaussian kernel (Bottom) Recovered binary signal from l_1 norm and proposed algorithm (b)(Top) Blurred piece-wise signal using Gaussian kernel (Bottom) Recovered signal using TV norm and proposed algorithm

2.4 Binary Super-resolution: Infinite Impulse Response

The problem of super-resolution is concerned with the reconstruction of temporally or spatially localized events (or spikes) from samples of their convolution with a low-pass filter. Distinct from prior works which exploit sparsity in appropriate domains in order to solve the resulting ill-posed problem, this chapter explores the role of binary priors in super-resolution, where the spike (or source) amplitudes are assumed to be binary-valued. Our study is inspired by the problem of neural spike deconvolution, but also applies to other applications such as symbol detection in hybrid millimeter wave communication systems. This chapter makes several theoretical and algorithmic contributions to enable binary super-resolution with very few measurements. Our results show that binary constraints offer much stronger identifiability guarantees than sparsity, allowing us to operate in “extreme compression” regimes, where the number of measurements can be significantly smaller than the sparsity level of the spikes. To ensure exact recovery in this ”extreme compression” regime, it becomes necessary to design algorithms that exactly enforce binary constraints without relaxation. In order to overcome the ensuing computational challenges, we consider a first order auto-regressive filter (which appears in neural spike deconvolution), and exploit its special structure. This results in a novel formulation of the super-resolution binary spike recovery in terms of binary search in one dimension. We perform numerical experiments that validate our theory and also show the benefits of binary constraints in neural spike deconvolution from real calcium imaging datasets.

2.4.1 Fundamental Sample Complexity Of Binary Super-Resolution

Let $y_{\mathrm{hi}}[n]$ be the output of a stable first-order Autoregressive AR(1) filter with parameter α, $0 < \alpha < 1$, driven by an unknown binary-valued input signal $x_{\mathrm{hi}}[n] \in \{0, A\}$, $A > 0$:

$$y_{\mathrm{hi}}[n] = \alpha y_{\mathrm{hi}}[n-1] + x_{\mathrm{hi}}[n] \tag{2.8}$$

In this section, we consider a super-resolution setting where we do not directly observe $y_{\mathrm{hi}}[n]$, and instead acquire M measurements $\{y_{\mathrm{lo}}[n]\}_{n=0}^{M-1}$ at a lower-rate by uniformly subsampling $y_{\mathrm{hi}}[n]$ by a factor of D:

$$y_{\mathrm{lo}}[n] = y_{\mathrm{hi}}[\mathrm{D}n], \quad n = 0, 1, \cdots, M-1, \tag{2.9}$$

The signal $y_{\mathrm{lo}}[n]$ corresponds to a filtered and downsampled version of the signal $x_{\mathrm{hi}}[n]$ where the filter is an infinite impulse response (IIR) filter with a single pole at α. Let $\mathbf{y}_{\mathrm{lo}} \in \mathbb{R}^M$ be a vector obtained by stacking the low-rate measurements $\{y_{\text{lo}}[n]\}_{n=0}^{M-1}$:

$$\mathbf{y}_{\text{lo}} = [y_{\text{lo}}[0], y_{\text{lo}}[1], \cdots, y_{\text{lo}}[M-1]]^\top$$

Since (2.8) represents a causal filtering operation, the low rate signal $\mathbf{y}_{\text{lo}}$ only depends on the present and past high-rate binary signal. Denote $L := (M-1)\mathrm{D}+1$. The M low-rate measurements in $\mathbf{y}_{\text{lo}}$ are a function of L samples of the high rate binary input signal $\{x_{\text{hi}}[n]\}_{n=0}^{L-1}$. These L samples are given by the following vector $\mathbf{x}_{\text{hi}} \in \{0, A\}^L$:

$$\mathbf{x}_{\text{hi}} := [x_{\text{hi}}[0], x_{\text{hi}}[1], \cdots, x_{\text{hi}}[L-1]]^\top.$$

Assuming the system to be initially at rest, i.e., $y_{\text{hi}}[n] = 0, n < 0$, we can represent the M samples from (2.9) in a compact matrix-vector form as:

$$\mathbf{y}_{\text{lo}} := \mathbf{S}_{\mathrm{D}}\mathbf{y}_{\text{hi}} = \mathbf{S}_{\mathrm{D}}\mathbf{G}_\alpha\mathbf{x}_{\text{hi}} \tag{2.10}$$

where $\mathbf{G}_\alpha \in \mathbb{R}^{L\times L}$ is a Toeplitz matrix given by:

$$\mathbf{G}_\alpha = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{L-1} & \alpha^{L-2} & \cdots & 1 \end{bmatrix} \tag{2.11}$$

and $\mathbf{S}_{\mathrm{D}} \in \mathbb{R}^{M\times L}$ is defined as:

$$[\mathbf{S}_{\mathrm{D}}]_{i,j} = \begin{cases} 1, & j = (i-1)\mathrm{D}+1 \\ 0, & \text{else} \end{cases}.$$

The matrix $\mathbf{S}_{\mathrm{D}}$ represents the $\mathrm{D}-$fold downsampling operation. Our goal is to infer the unknown high-rate binary input signal $x_{\text{hi}}[n]$ from the low-rate measurements $y_{\text{lo}}[n]$. This is essentially a "super-resolution" problem because the AR(1) filter first attenuates the high-frequency components of $x_{\text{hi}}[n]$, and the uniform downsampling operation systematically discards measurements. As a result, it may seem that the spiking activity $\{x_{\text{hi}}[(n-1)\text{D}+k]\}_{k=1}^{\text{D}}$ occurring "in-between" two low-rate measurements $y_{\text{lo}}[n-1]$ and $y_{\text{lo}}[n]$ is apparently lost. One can potentially interpolate arbitrarily, making the problem hopeless. In the next section, we will show that surprisingly, $\mathbf{x}_{\text{hi}}$ still remains identifiable from $\mathbf{y}_{\text{lo}}$ in the absence of noise, due to the binary nature of $\mathbf{x}_{\text{hi}}$ and "infinite memory" of the AR(1) filter.

Identifiability Conditions for Binary super-resolution

Consider the following partition of $\mathbf{x}_{\text{hi}}$ into M disjoint blocks, where the first block is a scalar and the remaining $M-1$ blocks are of length D, $\mathbf{x}_{\text{hi}} = [x_{\text{hi}}{}^{(0)}, \mathbf{x}_{\text{hi}}{}^{(1)\top}, \dots, \mathbf{x}_{\text{hi}}{}^{(M-1)\top}]^\top$. Here, $x_{\text{hi}}{}^{(0)} = x_{\text{hi}}[0]$ and $\mathbf{x}_{\text{hi}}{}^{(n)} \in \{0, A\}^{\text{D}}$ is given by:

$$[\mathbf{x}_{\text{hi}}{}^{(n)}]_k = x_{\text{hi}}[(n-1)\text{D}+k], \quad 1 \le n \le M-1 \tag{2.12}$$

The sub-vectors $\mathbf{x}_{\text{hi}}{}^{(n)}$, and $\mathbf{x}_{\text{hi}}{}^{(n-1)}$ ($n \geq 1$) represent consecutive and disjoint blocks (of length D) of the high-rate binary spike signal. In order to study the identifiability of $\mathbf{x}_{\text{hi}}$ from $\mathbf{y}_{\text{lo}}$, we first introduce an alternative (but equivalent) representation for (2.10), by constructing a sequence $c[n]$ as follows $c[0] = y_{\text{lo}}[0]$,

$$c[n] = y_{\text{lo}}[n] - \alpha^{\text{D}} y_{\text{lo}}[n-1], \; 1 \le n \le M-1 \tag{2.13}$$

Given the high rate AR(1) model defined in (2.8), it is possible to recursively represent $y_{\text{hi}}[\text{D}n]$ in terms of $y_{\text{hi}}[\text{D}n-1]$, which in turn, can be represented in terms of $y_{\text{hi}}[\text{D}n-2]$, and so on. By this recursive relation, we can represent $y_{\text{hi}}[\text{D}n-1]$ in terms of $y_{\text{hi}}[\text{D}n-\text{D}]$ and $\{x_{\text{hi}}[\text{D}n-i]\}_{i=0}^{\text{D}-1}$

and re-write $y_{\mathrm{lo}}[n]$ as

$$\begin{aligned} y_{\mathrm{lo}}[n] &= y_{\mathrm{hi}}[\mathrm{D}n] = \alpha y_{\mathrm{hi}}[\mathrm{D}n-1] + x_{\mathrm{hi}}[\mathrm{D}n] \\ &= \alpha^{\mathrm{D}} y_{\mathrm{hi}}[\mathrm{D}n-\mathrm{D}] + \alpha^{\mathrm{D}-1} x_{\mathrm{hi}}[\mathrm{D}(n-1)+1] + \cdots + \alpha x_{\mathrm{hi}}[\mathrm{D}n-1] + x_{\mathrm{hi}}[\mathrm{D}n], \\ y_{\mathrm{lo}}[n] - \alpha^{\mathrm{D}} y_{\mathrm{lo}}[n-1] &= \alpha^{\mathrm{D}-1} x_{\mathrm{hi}}[\mathrm{D}(n-1)+1] + \cdots + \alpha x_{\mathrm{hi}}[\mathrm{D}n-1] + x_{\mathrm{hi}}[\mathrm{D}n] \end{aligned} \tag{2.14}$$

The last equality holds due to the fact that $y_{\mathrm{lo}}[n-1] = y_{\mathrm{hi}}[\mathrm{D}n-\mathrm{D}]$. Combining (2.13) and (2.14), the sequence $c[n]$ can be re-written as $c[0] = y_{\mathrm{lo}}[0] = x_{\mathrm{hi}}{}^{(0)}$, and for $1 \le n \le M-1$

$$c[n] = \sum_{i=1}^{\mathrm{D}} \alpha^{\mathrm{D}-i} x_{\mathrm{hi}}[(n-1)\mathrm{D}+i] = \mathbf{h}_{\alpha}^{T} \mathbf{x}_{\mathrm{hi}}{}^{(n)} \tag{2.15}$$

where $\mathbf{h}_{\alpha} = [\alpha^{\mathrm{D}-1}, \alpha^{\mathrm{D}-2}, \ldots, \alpha, 1]^{T} \in \mathbb{R}^{\mathrm{D}}$. This implies that $c[n]$ depends only on the block $\mathbf{x}_{\mathrm{hi}}{}^{(n)}$. Denote $\mathbf{c} := [c[0], c[1], \ldots, c[M-1]]^{\top} \in \mathbb{R}^{M}$. For any D, (2.15) can be compactly represented as:

$$\mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha) \mathbf{x}_{\mathrm{hi}} \tag{2.16}$$

where $\mathbf{H}_{\mathrm{D}}(\alpha) \in \mathbb{R}^{M \times L}$ is given by:

$$\mathbf{H}_{\mathrm{D}}(\alpha) = \begin{bmatrix} 1 & \mathbf{0}^{\top} & \mathbf{0}^{\top} & \cdots & \mathbf{0}^{\top} \\ 0 & \mathbf{h}_{\alpha}^{\top} & \mathbf{0}^{\top} & \cdots & \mathbf{0}^{\top} \\ 0 & \mathbf{0}^{\top} & \mathbf{h}_{\alpha}^{\top} & \cdots & \mathbf{0}^{\top} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mathbf{0}^{\top} & \mathbf{0}^{\top} & \cdots & \mathbf{h}_{\alpha}^{\top} \end{bmatrix}$$

The following Lemma establishes the equivalence between (2.10) and (2.16).

Lemma 2. *Given* $\mathbf{y}_{\mathrm{lo}}$, *construct* $\mathbf{c}$ *following* (2.13)*. Then, there is a unique binary* $\mathbf{x}_{\mathrm{hi}} \in \{0, A\}^{L}$ *satisfying* (2.10) *if and only if* $\mathbf{x}_{\mathrm{hi}}$ *is a unique binary vector satisfying* (2.16)*.*

Proof. First suppose that there is a unique binary $\mathbf{x}_{\mathrm{hi}} \in \{0,A\}^L$ satisfying (2.10) but (2.16) has a non-unique binary solution, i.e., there exists ${\mathbf{x}_{\mathrm{hi}}}' \in \{0,A\}^L$, ${\mathbf{x}_{\mathrm{hi}}}' \neq \mathbf{x}_{\mathrm{hi}}$, such that

$$\mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{x}_{\mathrm{hi}} = \mathbf{H}_{\mathrm{D}}(\alpha){\mathbf{x}_{\mathrm{hi}}}' \tag{2.17}$$

Define ${\mathbf{y}_{\mathrm{hi}}}' := \mathbf{G}_{\alpha}{\mathbf{x}_{\mathrm{hi}}}'$ whose entries are given by:

$${y_{\mathrm{hi}}}'[n] = \sum_{k=0}^{n} \alpha^{n-k}{x_{\mathrm{hi}}}'[k], \quad 0 \leq n \leq L-1 \tag{2.18}$$

Notice that (2.13) can be re-written as

$$\begin{aligned}
&y_{\mathrm{lo}}[0] = c[0] = x_{\mathrm{hi}}[0], y_{\mathrm{lo}}[1] = c[1] + \alpha^{\mathrm{D}}y_{\mathrm{lo}}[0] = c[1] + \alpha^{\mathrm{D}}c[0] \\
&y_{\mathrm{lo}}[2] = c[2] + \alpha^{\mathrm{D}}y_{\mathrm{lo}}[1] = c[2] + \alpha^{\mathrm{D}}c[1] + \alpha^{2\mathrm{D}}c[0] \\
&\vdots
\end{aligned}$$

Following this recursive relation, and using (2.15) and (2.17), we can further re-write $y_{\mathrm{lo}}[n]$ as:

$$\begin{aligned}
y_{\mathrm{lo}}[n] &= \sum_{i=0}^{n} \alpha^{(n-i)\mathrm{D}}c[i] = \alpha^{n\mathrm{D}}{x'}_{\mathrm{hi}}^{(0)} + \sum_{i=1}^{n} \alpha^{(n-i)\mathrm{D}}\mathbf{h}_{\alpha}^{\top}{\mathbf{x}_{\mathrm{hi}}}'^{(i)} \\
&= \alpha^{n\mathrm{D}}{x'}_{\mathrm{hi}}^{(0)} + \sum_{i=1}^{n}\sum_{j=1}^{\mathrm{D}} \alpha^{n\mathrm{D}-(i-1)\mathrm{D}-j}{x'}_{\mathrm{hi}}[(i-1)\mathrm{D}+j] \\
&\overset{(a)}{=} \sum_{k=0}^{n\mathrm{D}} \alpha^{n\mathrm{D}-k}{x'}_{\mathrm{hi}}[k] \overset{(b)}{=} {y'}_{\mathrm{hi}}[n\mathrm{D}]
\end{aligned} \tag{2.19}$$

The equality (a) follows by a re-indexing of the summation into a single sum, and (b) follows from (2.18). By arranging (2.19) in a matrix form we obtain the following relation:

$$\mathbf{y}_{\mathrm{lo}} = \mathbf{S}_{\mathrm{D}}\mathbf{G}_{\alpha}{\mathbf{x}_{\mathrm{hi}}}'$$

However from (2.10), we have $\mathbf{y}_{\mathrm{lo}} = \mathbf{S}_{\mathrm{D}}\mathbf{G}_{\alpha}\mathbf{x}_{\mathrm{hi}}$. This contradicts the supposition that (2.10) has a unique binary solution.

Next, suppose that (2.16) has a unique binary solution but the binary solution to (2.10) is non-unique, i.e., there exists $\mathbf{x}_{\text{hi}}{}' \in \{0, A\}^L$, $\mathbf{x}_{\text{hi}}{}' \neq \mathbf{x}_{\text{hi}}$ such that

$$\mathbf{y}_{\text{lo}} = \mathbf{S}_{\text{D}} \mathbf{G}_{\alpha} \mathbf{x}_{\text{hi}}{}' = \mathbf{S}_{\text{D}} \mathbf{G}_{\alpha} \mathbf{x}_{\text{hi}}$$

By following (2.13) and (2.16), we also have $\mathbf{c} = \mathbf{H}_{\text{D}}(\alpha)\mathbf{x}_{\text{hi}}{}' = \mathbf{H}_{\text{D}}(\alpha)\mathbf{x}_{\text{hi}}$ which contradicts the assumption that solution of (2.16) is unique. □

Lemma 2 assures that a binary $\mathbf{x}_{\text{hi}}$ is uniquely identifiable from measurements $\mathbf{y}_{\text{lo}}$ if and only if there is a unique binary solution $\mathbf{x}_{\text{hi}} \in \{0, A\}^L$ to (2.16). From (2.15), it can be seen that $c[n]$ and $c[n-1]$ have contributions from only disjoint blocks of high rate spikes $\mathbf{x}_{\text{hi}}{}^{(n)}$, and $\mathbf{x}_{\text{hi}}{}^{(n-1)}$. Hence effectively, we only have a *single scalar measurement* $c[n]$ to decode an entire block $\mathbf{x}_{\text{hi}}{}^{(n)}$ of length D, regardless of how sparse it is. The task of decoding $\mathbf{x}_{\text{hi}}{}^{(n)}$ from a single measurement seems like a hopelessly "ill-posed" problem, caused by the uniform downsampling operation. But this is precisely where the binary nature of $\mathbf{x}_{\text{hi}}$ can be used as a powerful prior to make the problem well-posed. Theorem 3 specifies conditions under which it is possible to do so.

Theorem 3. *(Identifiability) For any $\alpha \in (0,1)$, with the possible exception of α belonging to a set of Lebesgue measure zero, there is a unique $\mathbf{x}_{\text{hi}} \in \{0, A\}^L$ that satisfies (2.16) for every* $\text{D} \geq 1$.

Proof. In Appendix A. □

Using Lemma 2 and Theorem 3, we can conclude that $\mathbf{x}_{\text{hi}}$ is uniquely identifiable from $\mathbf{y}_{\text{lo}}$ for almost all $\alpha \in (0,1)$. It can be verified that for $\alpha = 1$ the mapping is non-injective. Theorem 3 establishes that it is fundamentally possible to decode each block $\mathbf{x}_{\text{hi}}{}^{(n)}$ of length D, from effectively a single measurement $c[n]$. Since $\mathbf{x}_{\text{hi}}{}^{(n)}$ can take 2^{D} possible values, in principle, one can always perform an exhaustive search over these 2^{D} possible binary sequences and by Theorem 3, only one of them will satisfy $c[n] = \mathbf{h}_{\alpha}^{\top}\mathbf{x}_{\text{hi}}{}^{(n)}$. Since exhaustive search is computationally prohibitive, this leads to the natural question regarding alternative solutions. In Section 2.4.2, we will develop an alternative algorithm that leverages the trade-off between memory and computation to achieve a significantly lower run-time decoding complexity.

Comparison with Finite Rate of Innovation Approach

In a related line of work [36–38,40], the FRI framework has been developed to reconstruct spikes from the measurement model considered here. However, in the general FRI framework, there is no assumption on the amplitude of the spikes, and there are a total of 2D real valued unknowns corresponding to the locations and amplitudes of D spikes. In [38], it was shown that by leveraging the property of exponentially reproducing kernels, it is possible to recover arbitrary amplitudes and spike locations using Prony-type algorithms, provided at least $2\mathrm{D}+1(>\mathrm{D})$ low-rate measurements are available. However, since we exploit the binary nature of spiking activity, we can operate at a much smaller sample complexity than FRI. In fact, Theorem 3 shows that when we exploit the fact that the spikes occur on a high-resolution grid with binary amplitudes, $M = \Omega(1)$ measurements suffice to identify D spikes regardless of how large D is. A direct application of the FRI approach cannot succeed in this regime, since the number of spikes is larger than the number of measurements. That being said, with enough measurements, FRI techniques are powerful, and they can also identify off-grid spikes. In future, it would be interesting to combine the two approaches by incorporating binary priors to FRI based techniques and remove the grid assumptions.

Curse of Uniform Downsampling: Inadequacy of sparsity and non-negativity

By virtue of being a binary signal, $\mathbf{x}_{\text{hi}}$ is naturally sparse and non-negative. Therefore, one may ask if sparsity and/or non-negativity are sufficient to uniquely identify $\mathbf{x}_{\text{hi}}$ from $\mathbf{c}$, without the need for imposing any binary constraints. In particular, we would like to understand if the solution to the following problem that seeks the sparsest non-negative vector in $\mathbb{R}^L$ satisfying

(2.16) indeed coincides with the true $\mathbf{x}_{\mathrm{hi}} \in \{0, A\}^L$

$$\min_{\mathbf{x} \in \mathbb{R}^L} \quad \|\mathbf{x}\|_0 \quad \text{subject to } \mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{x}, \quad \mathbf{x} \geq \mathbf{0} \tag{P0}$$

Lemma 3. *For every $\mathbf{x}_{\mathrm{hi}} \in \{0, A\}^L$ (except $\mathbf{x}_{\mathrm{hi}} = A\mathbf{e}_1$), and $\mathbf{c} \in \mathbb{R}^M$ satisfying* (2.16)*, the following are true*

(i) There exists a solution $\mathbf{x}^{\star} \neq \mathbf{x}_{\mathrm{hi}}$ to (P0) *satisfying*

$$\|\mathbf{x}^{\star}\|_0 \leq \|\mathbf{x}_{\mathrm{hi}}\|_0 \tag{2.20}$$

(ii) The inequality in (2.20) *is strict as long as there exists an integer $n_0 \geq 1$ such that the block $\mathbf{x}_{\mathrm{hi}}^{(n_0)}$ of $\mathbf{x}_{\mathrm{hi}}$ (defined in* (2.12)*) satisfies $\|\mathbf{x}_{\mathrm{hi}}^{(n_0)}\|_0 \geq 2$.*

Proof. The proof is in Appendix B. □

Lemma 3 shows there exist other non-binary solution(s) to (2.16) (different from $\mathbf{x}_{\mathrm{hi}}$) that have the same or smaller sparsity as the binary signal $\mathbf{x}_{\mathrm{hi}} \in \{0, A\}^L$. Furthermore, there exist problem instances where the sparsest solution to (P0) is strictly sparser than $\mathbf{x}_{\mathrm{hi}}$. Hence, sparsity and/or non-negativity are inadequate to identify the ground truth $\mathbf{x}_{\mathrm{hi}}$ uniquely.

Implicit Bias of Relaxation: The optimization problem (P0) is non-convex and the binary constraints are not enforced. However, for computational tractability, it is common to instead solve the following relaxed optimization problem that seeks a non-negative vector in $\mathbb{R}^L$ with the smallest l_1 norm (instead of l_0 norm) satisfying $\mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{x}_{\mathrm{hi}}$ indeed deviates from the true $\mathbf{x}_{\mathrm{hi}} \in \{0, A\}^L$ and has a special property as characterized by lemma 4.

$$\min_{\mathbf{x} \in \mathbb{R}^L} \quad \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{x}, \quad \mathbf{x} \geq \mathbf{0} \tag{PL1}$$

Lemma 4. *For every* $\mathbf{x}_{\text{hi}} \in \{0, A\}^L$, *and* $\mathbf{c} \in \mathbb{R}^M$ *satisfying* $\mathbf{c} = \mathbf{H}_{\text{D}}(\alpha)\mathbf{x}_{\text{hi}}$, *the solution* $\mathbf{x}^\star$ *to* (PL1) *satisfies the following:*

$$\|\mathbf{x}^\star\|_1 \leq \|\mathbf{x}_{\text{hi}}\|_1 \tag{2.21}$$

and the support of the optimal solution $\mathscr{S}^*$ *obeys:*

$$\mathscr{S}^* \subseteq \{1, \text{D}+1, 2\text{D}+1, \cdots, (M-1)\text{D}+1\}. \tag{2.22}$$

Proof. We will construct a vector $\mathbf{v} \in \mathbb{R}^L$ with support of the form (2.22), that is feasible for (PL1) and prove that it has the smallest l_1 norm. Consider the vector $\mathbf{v} = [v^{(0)}, \mathbf{v}^{(1)\top}, \cdots, \mathbf{v}^{(M-1)\top}]^\top$, with the blocks $\mathbf{v}^{(n)}$ constructed as follows:

$$\mathbf{v}^{(0)} = c[0], \quad [\mathbf{v}^{(n)}]_k = \begin{cases} c[n], & \text{if } k = \text{D} \\ 0, & \text{else} \end{cases}$$

It is easy to verify that $c[n] = \mathbf{h}_\alpha^\top \mathbf{v}^{(n)}$ for all $n \geq 1$ and $\|\mathbf{v}\|_1 = \sum_{k=0}^{M-1} c[k]$. Let $\mathbf{v}_f \in \mathbb{R}^L$ be any feasible point of (PL1) which must be of the form:

$$\mathbf{v}_f^{(0)} = c[0], \quad \mathbf{v}_f^{(n)} = \mathbf{v}^{(n)} + \mathbf{r}^{(n)}$$

where $\mathbf{r}^{(n)} \in \mathscr{N}(\mathbf{h}_\alpha^\top)$ is a vector in the null-space of $\mathbf{h}_\alpha^\top$. It can be verified that for $1 \leq t \leq \text{D}-1$ the vectors $\mathbf{w}_t \in \mathbb{R}^{\text{D}}$ form a basis for $\mathscr{N}(\mathbf{h}_\alpha^\top)$ where:

$$[\mathbf{w}_t]_k = \begin{cases} 1, & k = t \\ -\alpha, & k = t+1 \\ 0, & \text{else} \end{cases}$$

Therefore, there exists $\{\beta_i^{(n)}\}_{i=1}^{\mathrm{D}-1}$ such that $\mathbf{r}^{(n)} = \sum_{j=1}^{\mathrm{D}-1} \beta_i^{(n)} \mathbf{w}_t$. The vector $\mathbf{v}_f^{(n)}$ has the following structure:

$$[\mathbf{v}_f^{(n)}]_k = \begin{cases} \beta_k^{(n)}, & k = 1 \\ -\alpha\beta_{k-1}^{(n)} + \beta_k^{(n)}, & 2 \leq k \leq \mathrm{D} - 1 \\ c[n] - \alpha\beta_{k-1}^{(n)}. & k = \mathrm{D} \end{cases}$$

To ensure $\mathbf{v}_f^{(n)}$ is a non-negative vector (feasible point of (PL1)), the following must hold:

$$\beta_1^{(n)} \geq 0, \quad \beta_k^{(n)} \geq \alpha\beta_{k-1}^{(n)} \text{ for } 2 \leq k \leq \mathrm{D} - 1$$

which implies $\beta_k^{(n)} \geq 0$ for all k as $\beta_1^{(n)} \geq 0$. Since $\mathbf{v}_f^{(n)}$ is a non-negative vector:

$$\|\mathbf{v}_f^{(n)}\|_1 = \sum_{k=1}^{\mathrm{D}} [\mathbf{v}_f^{(n)}]_k = c[n] + \sum_{k=1}^{\mathrm{D}-1} (1-\alpha)\beta_k^{(n)} = \|\mathbf{v}^{(n)}\|_1 + \underbrace{\sum_{k=1}^{\mathrm{D}-1} (1-\alpha)\beta_k^{(n)}}_{\geq 0}$$

We used the fact that $\sum_{k=1}^{D} \sum_{t=1}^{D-1} \beta_t^{(n)} [\mathbf{w}_t]_k = \sum_{t=1}^{\mathrm{D}-1} (1-\alpha)\beta_t^{(n)}$. Therefore, for any $\mathbf{v}_f^{(n)} \neq \mathbf{v}^{(n)}$, i.e., at least some $\beta_k^{(n)} \neq 0$ and we can conclude that

$$\|\mathbf{v}_f^{(n)}\|_1 > \|\mathbf{v}^{(n)}\|_1.$$

Therefore, the vector $\mathbf{v}$ constructed with the support (2.22) has the minimum l_1 norm among all possible feasible points of (PL1). □

This result provides a mathematical justification for why l_1 minimization (with non-negativity) is not suitable for the super-resolution spike reconstruction problem at hand. It reveals an interesting bias introduced of the minimum l_1 norm solution is always biased to contain spikes that are restricted to the low-resolution grid.

The aforementioned relaxation does not account for the binary constraints, we now take a look at a different class of relaxation designed for binary constraints. In binary compressed sensing [41, 42], it is common to relax the binary constraints using box-constraint and l_0 norm is relaxed to l_1 norm in the following manner:

$$\min_{\mathbf{x}\in\mathbb{R}^L} \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{x},\ \mathbf{0} \leq \mathbf{x} \leq A\mathbf{1} \tag{P1-B}$$

In the following Lemma, we show that there is an implicit bias introduced to the solution of (P1-B).

Lemma 5. *For every* $\mathbf{x}_{\mathrm{hi}} \in \{0, A\}^L$*, and* $\mathbf{c} \in \mathbb{R}^M$ *satisfying* (2.16)*. There exists a solution* $\mathbf{x}^\star$ *to* (P1-B) *satisfying*

$$\|\mathbf{x}^\star\|_1 \leq \|\mathbf{x}_{\mathrm{hi}}\|_1. \tag{2.23}$$

Moreover, for all $n \geq 1$*, the blocks* $\mathbf{x}^{(n)\star} \in \mathbb{R}^{\mathrm{D}}$ *of* $\mathbf{x}^\star$ *satisfy:*

$$supp(\mathbf{x}^{(n)\star}) = \{\mathrm{D}, \mathrm{D}-1, \cdots, \mathrm{D}-j_n\},\ \textit{if}\ c[n] \neq 0 \tag{2.24}$$

for some $0 \leq j_n \leq \mathrm{D}-1$ *and* $\mathbf{x}^{(n)\star} = \mathbf{0}$ *if* $c[n] = 0$*, irrespective of the support of* $\mathbf{x}_{\mathrm{hi}}$.

Proof. The proof is in Appendix B. □

Lemma 5 shows that even in the noiseless setting, introducing the box-constraint as a means of relaxing the binary constraint introduces a bias in the support of the recovered spikes. The optimal solution always results in spikes with support clustered towards the end of each block of length D, irrespective of the ground truth spiking pattern $\mathbf{x}_{\mathrm{hi}}$ that generated the measurements. This bias is a consequence of the nature of relaxation, as well as the specific structure of the measurement matrix $\mathbf{H}_{\mathrm{D}}(\alpha)$ arising in the problem.

Role of Memory in Super-resolution: IIR vs. FIR filters

The ability to identify the high-rate binary signal $\mathbf{x}_{\text{hi}} \in \{0, A\}^L$ from D$-$fold undersampled measurements $\mathbf{y}_{\text{lo}}$ (for arbitrarily large D) in the absence of noise, is in parts also due to the "infinite memory" or infinite impulse response of the AR(1) filter. Indeed, for an Finite Impulse Response (FIR) filter, there is a limit to downsampling without losing identifiability. This was recently studied in our earlier work [46] where we showed that the undersampling limit is determined by the length of the FIR filter. To see this, consider the convolution of a binary valued signal $\mathbf{x}_{\text{hi}}$ with a FIR filter $\mathbf{u} = [u[0], u[1], \cdots, u[r-1]]^T \in \mathbb{R}^r$ of length r: $z_f[n] = \sum_{i=0}^{r-1} u[r-1-i]x_{\text{hi}}[n+i]$. These samples are represented in the vector form as $\mathbf{z}_f := \mathbf{u} \star \mathbf{x}_{\text{hi}} \in \mathbb{R}^L$ (by suitable zero padding). Suppose, as before, we only observe a D$-$fold downsampling of the output $z_{\text{D}}[n] = z_f[\text{D}n]$. Two consecutive samples $z_{\text{D}}[p], z_{\text{D}}[p+1]$ of the low-rate observation are given by:

$$\begin{aligned} z_{\text{D}}[p] &= \sum_{i=0}^{r-1} u[r-1-i]x_{\text{hi}}[\text{D}p+i], \\ z_{\text{D}}[p+1] &= \sum_{i=0}^{r-1} u[r-1-i]x_{\text{hi}}[\text{D}(p+1)+i] \end{aligned}$$

If $\text{D} > r$, notice that none of the measurements is a function of the samples $x_{\text{hi}}[\text{D}p+r], x_{\text{hi}}[\text{D}p+r+1], \cdots, x_{\text{hi}}[\text{D}(p+1)-1]$. Hence, it is possible to assign them arbitrary binary values and yet be consistent with the low-rate measurements $z_{\text{D}}[n]$. This makes it impossible to exactly recover $\mathbf{x}_{\text{hi}}$ (even if it is known to be binary valued) if the decimation is larger than the filter length ($\text{D} > r$). The following lemma summarizes this result.

Lemma 6. *For every FIR filter $\mathbf{u} \in \mathbb{R}^r$, if the undersampling factor exceeds the filter length, i.e. $\text{D} > r$, there exist $\mathbf{x}_0, \mathbf{x}_1 \in \{0, A\}^L$, $\mathbf{x}_0 \neq \mathbf{x}_1$ such that $\mathbf{S}_{\text{D}}(\mathbf{u} \star \mathbf{x}_0) = \mathbf{S}_{\text{D}}(\mathbf{u} \star \mathbf{x}_1)$.*

This shows that the identifiability result presented in Theorem 1 is not merely a consequence of binary priors but the infinite memory of the autoregressive process is also critical in allowing arbitrary undersampling $\text{D} > 1$ in absence of noise. For such IIR filters, the memory of all past (binary) spiking activity is encoded (with suitable weighting) into every measurement captured after the spike, which would not be the case for a finite impulse response filter.

2.4.2 Efficient Binary Super-Resolution Using Binary Search with Structured Measurements

By Theorem 3, we already know that it is possible to uniquely identify $\mathbf{x}_{\mathrm{hi}}$ from $\mathbf{c}$ (or equivalently, each block ${\mathbf{x}_{\mathrm{hi}}}^{(n)}$ from a single measurement $c[n]$) by exhaustive search. We now demonstrate how this exhaustive search can be avoided by formulating the decoding problem in terms of "binary search" over an appropriate set, and thereby attaining computational efficiency. We begin by introducing some notations and definitions. Given a non-negative integer $k, 0 \leq k \leq 2^{\mathrm{D}} - 1$, let $(b_1(k), b_2(k), \cdots, b_{\mathrm{D}}(k))$ be the unique D-bit binary representation of k: $k = \sum_{d=1}^{\mathrm{D}} 2^{\mathrm{D}-d} b_d(k), \quad b_d(k) \in \{0,1\}\ \forall\ 1 \leq d \leq \mathrm{D}$. Here $b_1(k)$ is the most significant bit and $b_{\mathrm{D}}(k)$ is the least significant bit. Using this notation, we define the following set:

$$\mathscr{S}_{\mathrm{all}} := \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{2^{\mathrm{D}}-1}\}, \tag{2.25}$$

where each $\mathbf{v}_k \in \{0, A\}^{\mathrm{D}}$ is a binary vector given by

$$[\mathbf{v}_k]_d = A b_d(k). \quad 1 \leq d \leq \mathrm{D} \tag{2.26}$$

In other words, the binary vector $\frac{1}{A}\mathbf{v}_k$ is the D-bit binary representation of its index k. Using this convention, $\mathbf{v}_0 = \mathbf{0}$ (i.e., a binary sequence of all $0'$s) and $\mathbf{v}_{2^{\mathrm{D}}-1} = A\mathbf{1}$ (i.e., a binary sequence of all A's). Recall the partition of $\mathbf{x}_{\mathrm{hi}}$ defined in (2.12), where each block ${\mathbf{x}_{\mathrm{hi}}}^{(n)}$ $(n \geq 1)$ is a binary vector of length D and ${x_{\mathrm{hi}}}^{(0)} \in \{0, A\}$ is a scalar. It is easy to see that (2.25) comprises of all possible values that each block ${\mathbf{x}_{\mathrm{hi}}}^{(n)}$ can assume. According to (2.15) each scalar measurement $c[n]$ can be written as: $c[0] = x^{(0)}, \quad c[n] = {\mathbf{h}_\alpha}^{\top} {\mathbf{x}_{\mathrm{hi}}}^{(n)},\ 1 \leq n \leq M-1$. For every α, we define the following set:

$$\Theta_\alpha := \{\theta_0, \theta_1, \cdots, \theta_{2^{\mathrm{D}}-1}\}, \text{ where } \theta_k := \mathbf{h}_\alpha^\top \mathbf{v}_k \tag{2.27}$$

Observe that every measurement $c[n] = \sum_{i=1}^{\mathrm{D}} \alpha^{\mathrm{D}-i} x_{\mathrm{hi}}[(n-1)\mathrm{D}+i]$ takes values from this set Θ_α, depending on the value taken by the underlying block of spiking pattern from $\mathscr{S}_{\mathrm{all}}$. Our goal is to recover the spikes $\{x_{\mathrm{hi}}[(n-1)\mathrm{D}+i]\}_{i=1}^{\mathrm{D}}$ from $c[n]$.

In the following, we show that this problem is equivalent to finding the representation of a real number over an arbitrary radix, which is known as "β-expansion" [5]. Given a real (potentially non-integer) number $\beta > 1$, the representation of another real number $p \geq 0$ of the form:

$$p = \sum_{n=1}^{\infty} a_n \beta^{-n}, \text{ where } 0 \leq a_n < \lfloor \beta \rfloor \tag{2.28}$$

is referred to as a β-expansion of p. The coefficients $0 \leq a_n < \lfloor \beta \rfloor$ are integers. This is a generalization of the representation of numbers beyond integer-radix to a system where the radix can be chosen as an arbitrary real number. This notion of representation over arbitrary radix was first introduced by Renyi in [5], and since then has been extensively studied [6,7,56]. There is a direct connection between β-expansion and the binary super-resolution problem considered here. In the problem at hand, any element $\theta_k \in \Theta_\alpha$ can be written as:

$$\theta_k = \mathbf{h}_\alpha^\top \mathbf{v}_k = \sum_{i=1}^{\mathrm{D}} \alpha^{\mathrm{D}-i} [\mathbf{v}_k]_i$$

When $1/2 < \alpha < 1$, by letting $\beta = 1/\alpha$, we see that the coefficients in (2.28) must satisfy $0 \leq a_n < \lfloor 1/\alpha \rfloor < 2$, i.e., they are restricted to be binary valued $a_n \in \{0, 1\}$. *Therefore, decoding the spikes $\mathbf{v}_k$ from the observation θ_k is equivalent to finding a* D$-$*bit representation for the number θ_k/A over the non-integer radix $\beta = 1/\alpha$.* Questions regarding the existence of β-expansion, and finding the coefficients of a finite $\beta-$expansion (whenever it exists) has been an active topic of research [6,7,56,59]. When $\beta \geq 2$ (equivalently, $0 < \alpha \leq 1/2$), it is possible to find the coefficients using a greedy algorithm which proceeds in a fashion similar to finding the D-bit binary representation of an integer [7,59]. However, the regime $\beta \in (1,2)$ (equivalently $1/2 < \alpha < 1$), is significantly more complicated and is of continued research interest [6,7,56]. To the best of our knowledge, there are no known computationally efficient ways to find the finite β-expansion when $1/2 < \alpha < 1$ (if it exists) [N. Sidorov, personal communication, May 24, 2022]. In practice, we encounter filter values α $(= 1/\beta)$ that are much closer to 1, and hence, we need an alternative approach to find this finite β-radix representation for θ_k. In the next section, we show that by performing a suitable preprocessing, finite β-radix representation can be formulated as a binary search problem which is guaranteed to succeed for all values of β that permit unique finite $\beta-$expansions.

Formulation as a Binary Search Problem

Before describing the algorithm, we first introduce the notion of a *collision-free* set.

Definition 2 (Collision Free set)**.** *Given an undersampling factor* D*, define a class of "collision free" AR(1) filters as:*

$$\mathscr{G}_{\mathrm{D}} = \{\alpha \in (0,1) \text{ s.t. } \mathbf{h}_{\alpha}^{\top}\mathbf{v}_i \neq \mathbf{h}_{\alpha}^{\top}\mathbf{v}_j \ \forall\, i \neq j, \mathbf{v}_i, \mathbf{v}_j \in \mathscr{S}_{\mathrm{all}}\}$$

The set $\mathscr{G}_{\mathrm{D}}$ denotes permissible values of the AR(1) filter parameter α such that each of the 2^{D} binary sequences in $\mathscr{S}_{\mathrm{all}}$ maps to a unique element in the set Θ_{α}. In other words, every $\theta_k \in \Theta_{\alpha}$ has a unique D$-$bit expansion for all $\alpha \in \mathscr{G}_{\mathrm{D}}$. This naturally raises the question "How large is the set $\mathscr{G}_{\mathrm{D}}$?". Theorem 3 already provided the answer to this question, where the identifiability result implies that for every D, almost all $\alpha \in (0,1)$ belong to this set $\mathscr{G}_{\mathrm{D}}$ (with the possible exception of a measure zero set). Hence, Theorem 3 ensures that there are infinite choices for collision-free filter parameters.

Lemma 7. *For every $\alpha \in \mathscr{G}_{\mathrm{D}}$, the mapping $\Phi_\alpha(.) : \mathscr{S}_{\mathrm{all}} \rightarrow \Theta_\alpha$, $\Phi_\alpha(\mathbf{v}) = \mathbf{h}_\alpha^\top \mathbf{v}$ forms a bijection between $\mathscr{S}_{\mathrm{all}}$ and Θ_α.*

Proof. Since $\alpha \in \mathscr{G}_{\mathrm{D}}$, from the definition of the set $\mathscr{G}_{\mathrm{D}}$, it is clear that for any $\mathbf{v}_i, \mathbf{v}_j \in \mathscr{S}_{\mathrm{all}}$, $\mathbf{v}_i \neq \mathbf{v}_j$ we have $\mathbf{h}_\alpha{}^\top \mathbf{v}_i \neq \mathbf{h}_\alpha{}^\top \mathbf{v}_j$. Therefore, the mapping is injective. Furthermore, from (2.27) we also have $|\Theta_\alpha| \leq |\mathscr{S}_{\mathrm{all}}| = 2^{\mathrm{D}}$. Since $\Phi_\alpha(\cdot)$ is injective, we must also have $|\Theta_\alpha| = 2^{\mathrm{D}}$ and hence the mapping $\Phi_\alpha(.)$ forms a bijection between $\mathscr{S}_{\mathrm{all}}$ and Θ_α. □

When $\alpha \in \mathscr{G}_{\mathrm{D}}$, Lemma 7 states that the finite beta expansion for every $\theta_k \in \Theta_\alpha$ is unique. Lemma 7 provides a way to avoid exhaustive search over $\mathscr{S}_{\mathrm{all}}$, and yet identify $\mathbf{x}_{\mathrm{hi}}{}^{(n)}$ from $c[n]$ in a computationally efficient way. From Lemma 7, we know that each of the 2^{D} spiking patterns in $\mathscr{S}_{\mathrm{all}}$ maps to a unique element in Θ_α, and each element in Θ_α has a corresponding spiking pattern. Hence instead of searching $\mathscr{S}_{\mathrm{all}}$, we can equivalently search the set Θ_α in order to determine the unknown spiking pattern. Since Θ_α permits "ordering", searching Θ_α has a distinct computational advantage over searching $\mathscr{S}_{\mathrm{all}}$. This ordering enables us to employ binary search over (an ordered) Θ_α and find the desired element in a computationally efficient manner. To do this, we first sort the set Θ_α (in ascending order) and *arrange the corresponding elements of $\mathscr{S}_{all}$ in the same order*. Given Θ_α as an input, the function SORT$(\cdot)$ returns a sorted list $\Theta_\alpha^{\mathrm{sort}}$, and an index set $\mathscr{I} = \{i_0, i_1, \cdots, i_{2^{\mathrm{D}}-1}\}$ containing the indices of the sorted elements in the list Θ_α.

$$\Theta_\alpha^{\mathrm{sort}}, \mathscr{I} \leftarrow \mathrm{SORT}(\Theta_\alpha)$$

Let us denote the elements of the sorted lists as $\Theta_\alpha^{\mathrm{sort}} = \{\tilde{\theta}_0, \cdots, \tilde{\theta}_{2^{\mathrm{D}}-1}\}$, and $\mathscr{S}_{\mathrm{all}}^{\mathrm{sort}} = \{\tilde{\mathbf{v}}_0, \cdots, \tilde{\mathbf{v}}_{2^{\mathrm{D}}-1}\}$ where:

$$\tilde{\theta}_0 < \tilde{\theta}_1 < \cdots < \tilde{\theta}_{2^{\mathrm{D}}-1} \quad \text{and } \tilde{\theta}_j = \theta_{i_j}, \quad \tilde{\mathbf{v}}_j = \mathbf{v}_{i_j} \quad \forall j.$$

It is important to note that this sorting step does not depend on the measurements $\mathbf{c}$, and can therefore be part of a pre-processing pipeline that can be performed offline. However, it does require memory to store the sorted lists. In the noiseless setting, we know that every scalar

Algorithm 2. Noiseless Spike Recovery

```
1: **Input:** Measurement $c[n]$, Sorted list $\Theta_{\alpha}^{\text{sort}}$ and the corresponding (ordered) spike patterns $\mathscr{S}_{\text{all}}^{\text{sort}}$
2: **Output:** Decoded spike block $\hat{\mathbf{x}}_{\text{hi}}{}^{(n)}$
3: $i^{\star} \leftarrow \text{BINSEARCH}(\Theta_{\alpha}^{\text{sort}}, c[n])$
4: Return $\hat{\mathbf{x}}_{\text{hi}}{}^{(n)} \leftarrow \tilde{\mathbf{v}}_{i^{\star}}$
```

measurement $c[n] = \mathbf{h}_{\alpha}^{\top}\mathbf{x}_{\text{hi}}{}^{(n)}$ belongs to the set $\Theta_{\alpha}^{\text{sort}}$. Therefore, if we identify its index, say $i^{\star}$, then we can successfully recover $\mathbf{x}_{\text{hi}}{}^{(n)}$ by returning the corresponding binary vector $\tilde{\mathbf{v}}_{i^{\star}}$ from $\mathscr{S}_{\text{all}}^{\text{sort}}$. Therefore, we can formulate the decoding problem as searching for the input $c[n]$ in the sorted list $\Theta_{\alpha}^{\text{sort}}$. This can be efficiently done by using "Binary Search". The noiseless spike decoding procedure is summarized as Algorithm 2. Since the complexity of performing a binary search over an ordered list of N elements is $O(\log N)$, the complexity of Algorithm 2 is logarithmic in the cardinality of $\Theta_{\alpha}^{\text{sort}}$, which results in a complexity of $O(\log(2^{\text{D}})) = O(\text{D})$. We summarize this result in the following Lemma.

Lemma 8. *Assume $\alpha \in \mathscr{G}_{\text{D}}$. Given the ordered set Θ_{α}^{sort} , and an input $c[n] = \mathbf{h}_{\alpha}^{\top}\mathbf{x}_{\text{hi}}{}^{(n)}$, Algorithm 2 terminates in $O(\text{D})$ steps and its output $\hat{\mathbf{x}}_{\text{hi}}{}^{(n)}$ satisfies $\hat{\mathbf{x}}_{\text{hi}}{}^{(n)} = \mathbf{x}_{\text{hi}}{}^{(n)}$.*

Noisy Measurements and 1- D Nearest Neighbor Search

We demonstrate how binary search can still be useful in presence of noise by formulating noisy spike detection as a one dimensional *nearest neighbor search* problem. Suppose $\{z_{\text{lo}}[n]\}_{n=0}^{M-1}$ denote noisy D-fold decimated filter output

$$z_{\text{lo}}[n] = y_{\text{lo}}[n] + w[n], \quad 0 \leq n \leq M-1 \tag{2.29}$$

Here $w[n]$ represents the additive noise term that corrupts the (noiseless) low-rate measurements $y_{\text{lo}}[n]$. Similar to (2.13), we compute $c_e[n]$ from $z_{\text{lo}}[n]$ as follows:

$$c_e[n] = z_{\text{lo}}[n] - \alpha^{\text{D}} z_{\text{lo}}[n-1] \tag{2.30}$$

$$= \sum_{i=1}^{\text{D}} \alpha^{\text{D}-i} x_{\text{hi}}[(n-1)\text{D}+i] + e[n] = c[n] + e[n] \tag{2.31}$$

where $c[n] = \mathbf{h}_{\alpha}^{\top} \mathbf{x}_{\text{hi}}{}^{(n)} \in \Theta_{\alpha}^{\text{sort}}$, and $e[n] = w[n] - \alpha^{\text{D}} w[n-1]$. We can interpret $c_e[n]$ as a noisy/perturbed version of an element $c[n] \in \Theta_{\alpha}^{\text{sort}}$, with $e[n]$ representing the noise. This perturbed signal may no longer belong to $\Theta_{\alpha}^{\text{sort}}$ (i.e. $c_e[n] \notin \Theta_{\alpha}^{\text{sort}}$) and hence, we cannot find an exact match in the set $\Theta_{\alpha}^{\text{sort}}$. Instead, we aim to find the closest element in $\Theta_{\alpha}^{\text{sort}}$ (the nearest neighbor of $c_e[n]$) by solving the following problem:

$$\hat{\mathbf{x}}_{\text{hi}}{}^{(n)} = \arg \min_{\mathbf{v} \in \mathscr{S}_{\text{all}}^{\text{sort}}} |c_e[n] - \mathbf{h}_{\alpha}^{\top} \mathbf{v}| \tag{2.32}$$

Solving (2.32) is equivalent to finding the spike sequence $\tilde{\mathbf{v}} \in \mathscr{S}_{\text{all}}^{\text{sort}}$ that maps to the nearest neighbor of $c_e[n]$ in the set $\Theta_{\alpha}^{\text{sort}}$. By leveraging the sorted list $\Theta_{\alpha}^{\text{sort}}$, it is no longer necessary to parse the list sequentially (which would incur $O(2^{\text{D}})$ complexity), instead we can perform a modified binary search as summarized in Algorithm 3, that keeps track of additional indices compared to the vanilla binary search. Finally, we return the unique spiking pattern from $\mathscr{S}_{\alpha}^{\text{sort}}$ that gets mapped to the nearest neighbor of the noisy measurement $c_e[n]$. It is well-known that the nearest neighbor for any query could be found in $O(\log(2^{\text{D}})) = O(\text{D})$ steps, instead of the linear complexity of $O(2^{\text{D}})$. This guarantees a computationally efficient decoding of spikes by solving (2.32).

Next, we characterize the error events that lead to erroneous detection of a block of spikes. Recall that the set $\Theta_{\alpha}^{\text{sort}}$ is sorted, and its elements satisfy the ordering:

$$0 = \tilde{\theta}_0 < \tilde{\theta}_1 < \cdots < \tilde{\theta}_{l_{\text{D}}} = 1 + \alpha + \cdots + \alpha^{\text{D}-1}$$

where $l_\mathrm{D} := 2^\mathrm{D} - 1$. We also have $\tilde{\theta}_k = \mathbf{h}_\alpha^\top \tilde{\mathbf{v}}_k$, where $\tilde{\mathbf{v}}_k \in \mathscr{S}_\mathrm{all}^\mathrm{sort}$ is a binary spiking sequence of length D.

For each $\tilde{\mathbf{v}}_k$ and each n, we will determine the error event $\hat{\mathbf{x}}_\mathrm{hi}{}^{(n)} \neq \mathbf{x}_\mathrm{hi}{}^{(n)}$, when $\mathbf{x}_\mathrm{hi}{}^{(n)} = \tilde{\mathbf{v}}_k$. First, consider the scenario when $\mathbf{x}_\mathrm{hi}{}^{(n)} = \tilde{\mathbf{v}}_k$ for some $0 < k < l_\mathrm{D}$ (excluding $\tilde{\mathbf{v}}_0, \tilde{\mathbf{v}}_{l_D}$). The corresponding noiseless measurement is $c[n] = \tilde{\theta}_k = \mathbf{h}_\alpha^\top \tilde{\mathbf{v}}_k$ which satisfies $\tilde{\theta}_{k-1} < c[n] = \tilde{\theta}_k < \tilde{\theta}_{k+1}$. Since $\Theta_\alpha^\mathrm{sort}$ is sorted, it can be easily verified that the nearest neighbor of $c_e[n]$ will be $\tilde{\theta}_k$, if and only if $c_e[n]$ satisfies the following condition:

$$(\tilde{\theta}_{k-1} + \tilde{\theta}_k)/2 \leq c_e[n] \leq (\tilde{\theta}_{k+1} + \tilde{\theta}_k)/2 \tag{2.33}$$

Since $\tilde{\theta}_k = \mathbf{h}_\alpha^\top \tilde{\mathbf{v}}_k$, the solution to (2.32) is attained at $\tilde{\mathbf{v}}_k \in \mathscr{S}_\mathrm{all}^\mathrm{sort}$, and the decoding is successful. Therefore Algorithm 3 produces an erroneous estimate of $\tilde{\mathbf{v}}_k$ if and only if $c_e[n]$ violates (2.33). The event $c_e[n] \notin [\frac{\tilde{\theta}_{k-1}+\tilde{\theta}_k}{2}, \frac{\tilde{\theta}_{k+1}+\tilde{\theta}_k}{2}]$ is equivalent to $e[n] \in \mathscr{E}_k$ ($e[n]$ is defined earlier in (2.31)), where

$$\mathscr{E}_k = \{e[n] < -\frac{\tilde{\theta}_k - \tilde{\theta}_{k-1}}{2}, \text{ or } e[n] > \frac{\tilde{\theta}_{k+1} - \tilde{\theta}_k}{2}\} \tag{2.34}$$

Finally, we characterize the error events for $k = 0, l_\mathrm{D}$. The error events for $c[n] = \theta_0 = 0$ or $c[n] = \theta_{l_\mathrm{D}}$ are given by:

$$\mathscr{E}_0 = \{e[n] \geq \tilde{\theta}_1/2\},\ \mathscr{E}_{l_\mathrm{D}} = \{e[n] \leq -(\tilde{\theta}_{l_\mathrm{D}} - \tilde{\theta}_{l_\mathrm{D}-1})/2\} \tag{2.35}$$

Define the "minimum distance" between points in $\Theta_\alpha^\mathrm{sort}$:

$$\Delta\theta_\mathrm{min}(\alpha, \mathrm{D}) = \min_{1 \leq k \leq l_\mathrm{D}} |\tilde{\theta}_k - \tilde{\theta}_{k-1}|.$$

This minimum distance depends on A, α and D. From (2.34), (2.35) it can be verified that if $2|w[n]| < \Delta\theta_\mathrm{min}(\alpha, \mathrm{D})/2$ (which would imply $|e[n]| < \Delta\theta_\mathrm{min}(\alpha, \mathrm{D})/2$) for all n, then $\hat{\mathbf{x}}_\mathrm{hi}{}^{(n)} =$

${\mathbf{x}_{\text{hi}}}^{(n)}$. As summarized in Theorem 4, Algorithm 3 can exactly recover the ground truth spikes from measurements corrupted by bounded adversarial noise, the extent of the robustness is determined by the parameters A, α, D.

Algorithm 3. Noisy Spike Recovery

```
Input: Measurement $c_e[n]$, Sorted list $\Theta_\alpha^{\text{sort}}$ and the corresponding (ordered) spike patterns $\mathscr{S}_{\text{all}}^{\text{sort}}$
Output: Decoded spike block ${\hat{\mathbf{x}}_{\text{hi}}}^{(n)}$
Set $l \leftarrow 0, u \leftarrow 2^{\mathrm{D}} - 1$
while $u - l > 1$
    Set $m \leftarrow l + \lfloor (u-l)/2 \rfloor$
    if $\tilde{\theta}_m > c_e[n]$ then
        $u \leftarrow m$
    else
        $l \leftarrow m$
    end if
end while
Find the nearest neighbor $i^\star = \arg\min_{i \in \{l,u\}} (c_e[n] - \tilde{\theta}_i)^2$
Return ${\hat{\mathbf{x}}_{\text{hi}}}^{(n)} \leftarrow \tilde{\mathbf{v}}_{i^\star}$
```

Theorem 4. *Assume $\alpha \in \mathscr{G}_{\mathrm{D}}$. Given the ordered set Θ_α^{sort}, the output of Algorithm 3 with input $c_e[n]$ exactly coincides with the solution of the optimization problem* (2.32) *in at most $O(\mathrm{D})$ steps. Furthermore, if for all n, $|w[n]| < \Delta\theta_{\min}(\alpha, \mathrm{D})/4$, then the output of Algorithm 3 satisfies ${\hat{\mathbf{x}}_{\text{hi}}}^{(n)} = {\mathbf{x}_{\text{hi}}}^{(n)}$.*

From Theorem 4, it is evident that $\Delta\theta_{\min}(\alpha, \mathrm{D})$ plays an important role in characterizing the upper bound on noise. We attempt to gain insight into how $\Delta\theta_{\min}(\alpha, \mathrm{D})$ varies as a function of α when D is held fixed.

Lemma 9. *Given* D, $\Delta\theta_{\min}(\alpha, \mathrm{D}) = A\alpha^{\mathrm{D}-1}$ *for* $\alpha \in (0, 0.5]$.

Proof. The proof for $A = 1$ is in Appendix C and it can be scaled to obtain the desired bound. □

When $\alpha \in (0, 0.5]$, $\Delta\theta_{\min}(\alpha, \mathrm{D})$ is monotonically increasing with α. However, for $\alpha > 0.5$ the trend fluctuates with α differently for different D, and becomes quite challenging to predict. This is also confirmed by the empirical plot in Figure 2.3. A refined analysis of $\Delta\theta_{\min}(\alpha, \mathrm{D})$ to gain insight into desirable filter parameters α is an interesting direction for future work.

Trade-off between memory and computational complexity

A crucial aspect of Algorithms 2 and 3 is that they achieve efficient run-time complexity by leveraging the off-line construction of the sorted list $\Theta_\alpha^{\text{sort}}$ and $\mathscr{S}_{\text{all}}^{\text{sort}}$. These lists, each with 2^{D} elements, need to be stored in memory and made available during run-time. Since there is no free lunch, the resulting computational efficiency of $O(\mathrm{D})$ at run-time is attained at the expense of the additional memory that is required to store the sorted lists $\Theta_\alpha^{\text{sort}}, \mathscr{S}_{\text{all}}^{\text{sort}}$.

Parallelizable Implementation

Algorithm 3 (also Algorithm 2) only takes $c_e[n](c[n])$ as input and returns $\hat{\mathbf{x}}_{\text{hi}}{}^{(n)}$, and is completely de-coupled from any other $\hat{\mathbf{x}}_{\text{hi}}{}^{(n')}$, $n' \neq n$. Recall that in reality, we are provided with measurements $z_{\text{lo}}[n](y_{\text{lo}}[n])$, and $c_e[n]$(respectively $c[n]$) needs to be computed. Due to this de-coupling, we can compute $c_e[n]'s$ in parallel using two consecutive low-rate samples $z_{\text{lo}}[n], z_{\text{lo}}[n-1]$ and perform a nearest neighbor search without waiting for any previously decoded spikes. Therefore, the total decoding complexity can be further improved depending on the available parallel computing resources.

2.4.3 Error Analysis for Gaussian Noise

Algorithm 3 solves (2.32) without requiring any knowledge of the noise statistics. However, in order to analyze its performance, we will make the following (standard) assumptions on the statistics of the high-rate spiking signal $\mathbf{x}_{\text{hi}}$ and the measurement noise $w[n]$ as follows:

- (**A1**) The entries of the binary vector $\mathbf{x}_{\text{hi}} \in \{0, A\}^L$ are i.i.d random variables distributed as $x_{\text{hi}}[n] \sim A\text{Bern}(p)$.

- (**A2**) The additive noise $w[n], 0 \leq n \leq M-1$ is independent of $x_{\text{hi}}[n]$, and distributed as $w[n] \sim \mathscr{N}(0, \sigma^2)$

Probability of Erroneous Decoding

Under assumption (**A2**), the ML estimate of $\mathbf{x}_{\text{hi}}$ is given by the solution to the following problem:

$$\hat{\mathbf{x}}_{\text{ML}} = \arg \min_{\mathbf{v} \in \{0,A\}^L} \|\mathbf{z}_{\text{lo}} - \mathbf{S}_{\text{D}} \mathbf{G}_{\alpha} \mathbf{v}\|_2 \quad (P_{\text{NN}})$$

The proposed Algorithm 3 does not attempt to solve (P_{NN}), which is computationally intractable. Instead, it solves a set of $M-1$ one dimensional nearest neighbor search problems, by finding the nearest neighbor of $c_e[n]$ for each $n = 1, 2, \cdots, M-1$. This scalar nearest neighbor search is implemented in a computationally efficient manner by using parallel binary search on a pre-sorted list. Notice that by the operation (2.30), the variance of the equivalent noise term $e[n]$ gets amplified by a factor of at most $(1 + \alpha^{2\text{D}}) < 2$. This can be thought of as a price paid to achieve computational efficiency and parallelizability. The following theorem characterizes the dependence of certain key quantities of interest, such as the signal-to-noise ratio (SNR), undersampling factor D, and filter's frequency response (controlled by α) on the performance of Algorithm 3.

Theorem 5. *Suppose $\alpha \in \mathscr{G}_{\text{D}}$ and assumptions* **(A1-A2)** *hold. Given $\delta > 0$, if the following condition is satisfied:*

$$\Delta\theta^2_{\min}(\alpha, \text{D})/\sigma^2 \geq 4 \ln(2M/\delta) \tag{2.36}$$

then Algorithm 3 can exactly recover the binary signal $\mathbf{x}_{\text{hi}}$ with probability at least $1 - \delta$.

Proof. The proof follows standard arguments for computing the probability of error for symbol detection in Gaussian noise, followed by certain simplifications and is included in Appendix *D* for completeness. □

In Figure 2.3, we plot $\Delta\theta_{\min}(\alpha, \text{D})$ as a function of D for different values of α. As expected, $\Delta\theta_{\min}(\alpha, \text{D})$ decays as the D increases. Understandably, for a fixed α, as D increases,

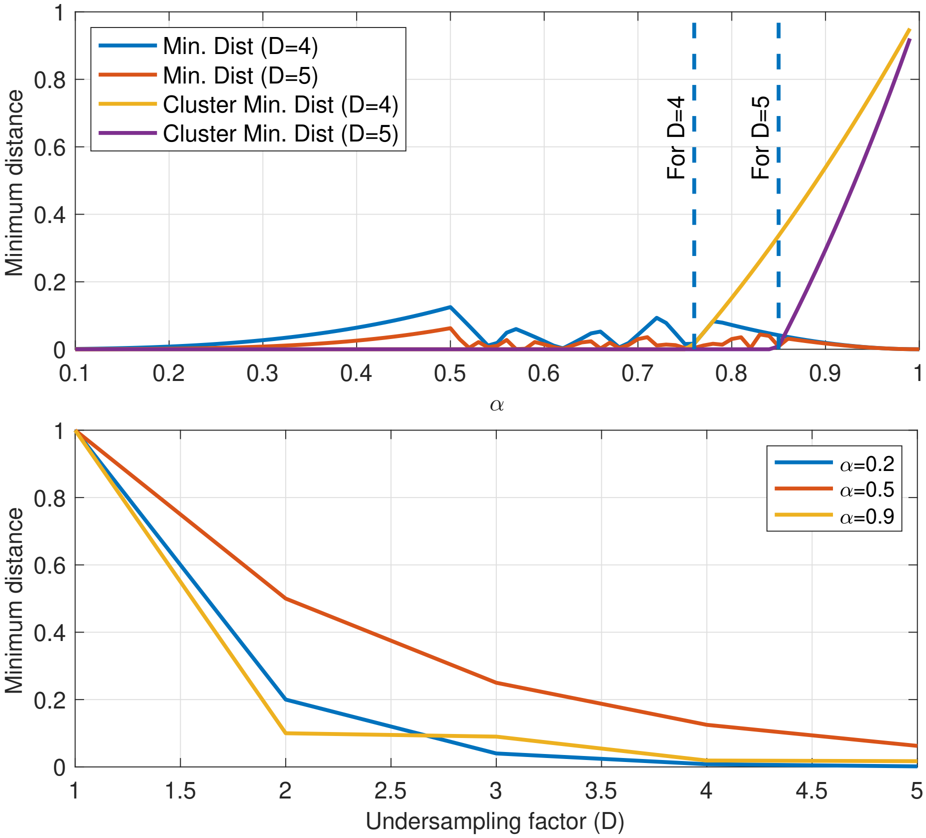


Figure 2.3. Variation of $\Delta\theta_{\min}(\alpha, \mathrm{D})$ as a function of undersampling factor D and α. The cluster-distance $\Delta^{c}_{\min}(\alpha, \mathrm{D})$ vs. α is also overlaid. Each dotted line denotes the start of the interval $\mathscr{F}_{\mathrm{D}}$.

it becomes harder to recover the spikes exactly, and higher SNR is needed to compensate for the lower sampling rate. This can be interpreted as the price paid for super-resolution in presence of noise. This phenomenon is also reminiscent of the noise amplification effect in super-resolution, where the ability to super-resolve point sources becomes more severely hindered by noise as the target resolution grid becomes finer [25]. In Figure 2.3, we plot $\Delta\theta_{\min}(\alpha, \mathrm{D})$ as a function of α and as predicted by Lemma 9, it monotonically increases upto 0.5, but for $\alpha > 0.5$, the behavior becomes much more erratic and a precise characterization becomes challenging. It is to be noted that in Theorem 5, we aim to *exactly recover* $\mathbf{x}_{\mathrm{hi}}$. The SNR requirement can be relaxed if our goal is to recover only spike counts instead of the true spikes as discussed in the next subsection. One can define other notions of approximate recovery, the analysis of which will be a topic of future research.

2.4.4 Relaxed Spike reconstruction: Count Estimation

As shown in Theorem 4, exact recovery of spikes is possible under somewhat restrictive condition on the noise in terms of $\Delta\theta_{\min}(\alpha, \mathrm{D})$, which becomes quite small as D increases. This naturally calls for other relaxed notions of recovery which can handle larger noise levels. In neuroscience, it is believed that information is encoded as either the spike timing (temporal code) or the firing rates (rate coding) of individual neurons in the brain. Therefore, the spike counts over an interval can be informative to understand neural functions, even when it is impossible to temporally localize the neural spikes. For example, neurons in the visual cortex encode stimulus orientations as their firing rates [60]. We will therefore focus on spike count as an approximate recovery metric, which concerns estimating the number of spikes occurring between two consecutive low-rate measurements instead of resolving the individual spiking activity at a higher resolution.

Let $\gamma[n]$ denote the total number of spikes occurring between two consecutive low-rate samples $z_{\mathrm{lo}}[n]$ and $z_{\mathrm{lo}}[n-1]$. Since $\mathbf{x}_{\mathrm{hi}}$ and its estimate $\widehat{\mathbf{x}}_{\mathrm{hi}}$ are both binary valued (amplitude A), the true spike count ($\gamma[n]$) and estimated count ($\hat{\gamma}[n]$) are given by: $\gamma[n] = \|{\mathbf{x}_{\mathrm{hi}}}^{(n)}\|_0, \quad \hat{\gamma}[n] =$

$\|\hat{\mathbf{x}}_{\mathrm{hi}}^{(n)}\|_0$, $n = 1, \cdots, M-1$, $\gamma[0] = x_{\mathrm{hi}}[0]/A$ and $\hat{\gamma}[0] = \hat{x}_{\mathrm{hi}}[0]/A$ since the first block is of size 1 as described in (2.12). Define a set $\mathscr{C}_k^{\mathrm{D}}$ as:

$$\mathscr{C}_k^{\mathrm{D}} := \{\mathbf{v} \in \{0, A\}^{\mathrm{D}}, \|\mathbf{v}\|_0 = k\}, \quad 0 \leq k \leq \mathrm{D}$$

It is a collection of all binary vectors (of length D) with spike count k. The ground truth spike block belongs to $\mathscr{C}_{\gamma[n]}^{\mathrm{D}}$. Any element from $\mathscr{C}_{\gamma[n]}^{\mathrm{D}}$ will give the true spike count. Hence, exact recovery of count can be possible even when spikes cannot be recovered.

For a fixed D, we define a set of α denoted by $\mathscr{F}_{\mathrm{D}}$:

$$\mathscr{F}_{\mathrm{D}} := \{\alpha \in (0,1) | \alpha^{\mathrm{D}} - \alpha^{\mathrm{D}-k_0-1} - \alpha^{k_0} + 1 < 0\} \tag{2.37}$$

where $k_0 = \lfloor \mathrm{D}/2 \rfloor$. We will obtain a sufficient condition for robust spike count estimation when $\alpha \in \mathscr{F}_{\mathrm{D}}$. It can be shown that for any D, $\mathscr{F}_{\mathrm{D}}$ will always be non-empty. Define

$$\theta_{\min}^k := \min_{\mathbf{u} \in \mathscr{C}_k^{\mathrm{D}}} \mathbf{h}_\alpha^\top \mathbf{u} \quad \theta_{\max}^k := \max_{\mathbf{u} \in \mathscr{C}_k^{\mathrm{D}}} \mathbf{h}_\alpha^\top \mathbf{u} \tag{2.38}$$

Observe that if

$$\theta_{\min}^{k+1} > \theta_{\max}^k, k = 0, 1, \cdots, \mathrm{D}-1 \tag{2.39}$$

then all spike patterns $\mathbf{u}_i \in \mathscr{C}_k^{\mathrm{D}}$ (with the same spike count k) are clustered together when mapped on to the real line by the transformation $\mathbf{h}_\alpha^\top \mathbf{u}$ as shown in Figure 2.4. When (2.39) holds, we can define a "cluster-restricted minimum distance" as:

$$\Delta_{\min}^c(\alpha, \mathrm{D}) := \min_{0 \leq k \leq \mathrm{D}-1} \theta_{\min}^{k+1} - \theta_{\max}^k \tag{2.40}$$

Given a noisy observation $c_e[n] = \mathbf{h}_\alpha^\top \mathbf{x}_{\mathrm{hi}}{}^{(n)} + e[n]$, the solution to the nearest neighbor problem (2.32) may return an incorrect neighbor $\theta_j \neq \mathbf{h}_\alpha^\top \mathbf{x}_{\mathrm{hi}}{}^{(n)}$. However, when (2.39) holds and if the noisy observation satisfies the following conditions:

$$(\theta_{\min}^{\gamma[n]} + \theta_{\max}^{\gamma[n]-1})/2 < c_e[n] < (\theta_{\min}^{\gamma[n]+1} + \theta_{\max}^{\gamma[n]})/2 \tag{2.41}$$

then the nearest-neighbor decision rule in Algorithm 3 will still ensure that $\theta_j \in \mathscr{C}_{\gamma[n]}^{\mathrm{D}}$. This has also been visualized in Figure 2.4 where each colored band represents the “safe-zone” for each count and the black dotted-line denotes the boundary. This will result in correct identification of the spike count but will incur error in terms of spiking pattern. We formally summarize this in the following Theorem that provides robustness guarantee for exact count recovery from measurements corrupted by adversarial noise (similar to Theorem 4 for spike recovery).

Theorem 6. *Assume $\alpha \in \mathscr{F}_{\mathrm{D}}$. Given the ordered set Θ_{α}^{sort}, let $\hat{\gamma}[n]$ be the estimated spike count obtained from Algorithm 3 with input $c_e[n]$. If for all n, $|w[n]| < \Delta_{\min}^{c}(\alpha, \mathrm{D})/4$, then the count can be exactly recovered, i.e., $\hat{\gamma}[n] = \gamma[n]$.*

Proof. Proof is in Appendix E. □

It is clear that when (2.39) holds, $\Delta_{\min}^{\mathrm{c}}(\alpha, \mathrm{D})$ is no smaller than $\Delta\theta_{\min}(\alpha, \mathrm{D})$, since the former is computed over neighboring elements of the cluster whereas $\Delta\theta_{\min}(\mathrm{D}, \alpha)$ computes the minimum distance over all consecutive elements (both inter-cluster as well as intra-cluster) in $\Theta_{\alpha}^{\mathrm{sort}}$. This essentially suggests that estimation of counts (for this range of α and D) can be more robust compared to inferring the individual spiking patterns. We also illustrate this numerically in Figure 2.3 (top), where we plot both $\Delta_{\min}^{c}$ and $\Delta\theta_{\min}$ as a function of α and the start of the interval $\mathscr{F}_{\mathrm{D}}$ (computed numerically) is denoted using dotted lines. For both values of D, we can see that $\Delta_{\min}^{c} > \Delta\theta_{\min}$ and the gap grows as α increases.

2.5 Numerical Experiments

We conduct numerical experiments to evaluate the performance of the proposed super-resolution spike decoding algorithm on both synthetic and real calcium imaging datasets.

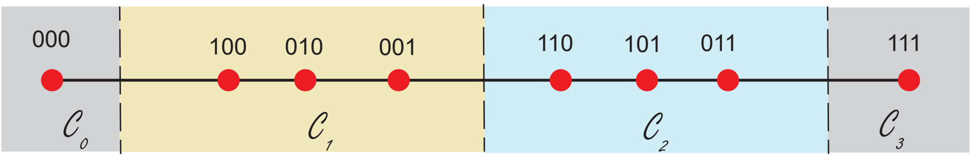


Figure 2.4. Visualization of the sets $\mathscr{C}_k^{\mathrm{D}}$ for $\mathrm{D}=3$. In this scenario, the spiking patterns corresponding to the same count are clustered together and hence, are favorable for robust count estimation.

2.5.1 Synthetic Data Generation and Evaluation Metrics

We create a synthetic dataset by generating high-rate binary spike sequence $\mathbf{x}_{\mathrm{hi}} \in \{0,1\}^L$ ($A=1$ and $L=1000$) that satisfies assumption (**A1**). The spiking probability p controls the average sparsity level given by $s := \mathbb{E}[\|\mathbf{x}_{\mathrm{hi}}\|_0] = Lp$. We aim to reconstruct $\mathbf{x}_{\mathrm{hi}}$ from $M \approx L/\mathrm{D}$ low-rate measurements $z_{\mathrm{lo}}[n]$ defined in (2.29). Notice that we operate in a regime where the expected sparsity is greater than the total number of low-rate measurements, i.e., $s > M$. We employ the widely-used *F-score* metric to evaluate the accuracy of spike detection [11,32]. The F-score is computed by first matching the estimated and ground truth spikes. An estimated spike is considered a "match" to a ground truth spike if it is within a distance of t_0 of the ground truth (many-to-one matching is not allowed) [11,32]. Let K and K' be the total number of ground truth and estimated spikes, respectively. The number of spikes declared as true positives is denoted by T_p. After the matching procedure, we compute the recall ($R = \frac{T_p}{K}$) which is defined as the ratio of true positives (T_p) and the total number of ground truth spikes (K). Precision ($P = \frac{T_p}{K'}$) measures the fraction of the total detected spikes which were correct. Finally, the F-score is given by the harmonic mean of recall and precision F-score $= 2PR/(P+R)$.

Noiseless Recovery: Role of Binary priors and memory

We first consider the noiseless setting ($w[n]=0$ in (2.29)). We compare the performance of Algorithm 3 against box-constrained l_1 minimization method [41,42], where we solve:

$$\min_{\mathbf{x}\in\mathbb{R}^L} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y}_{\mathrm{lo}} - \mathbf{S}_{\mathrm{D}}\mathbf{G}_\alpha\mathbf{x}\|_2 \le \varepsilon, \mathbf{0} \le \mathbf{x} \le A\mathbf{1} \qquad \text{(P1)}$$

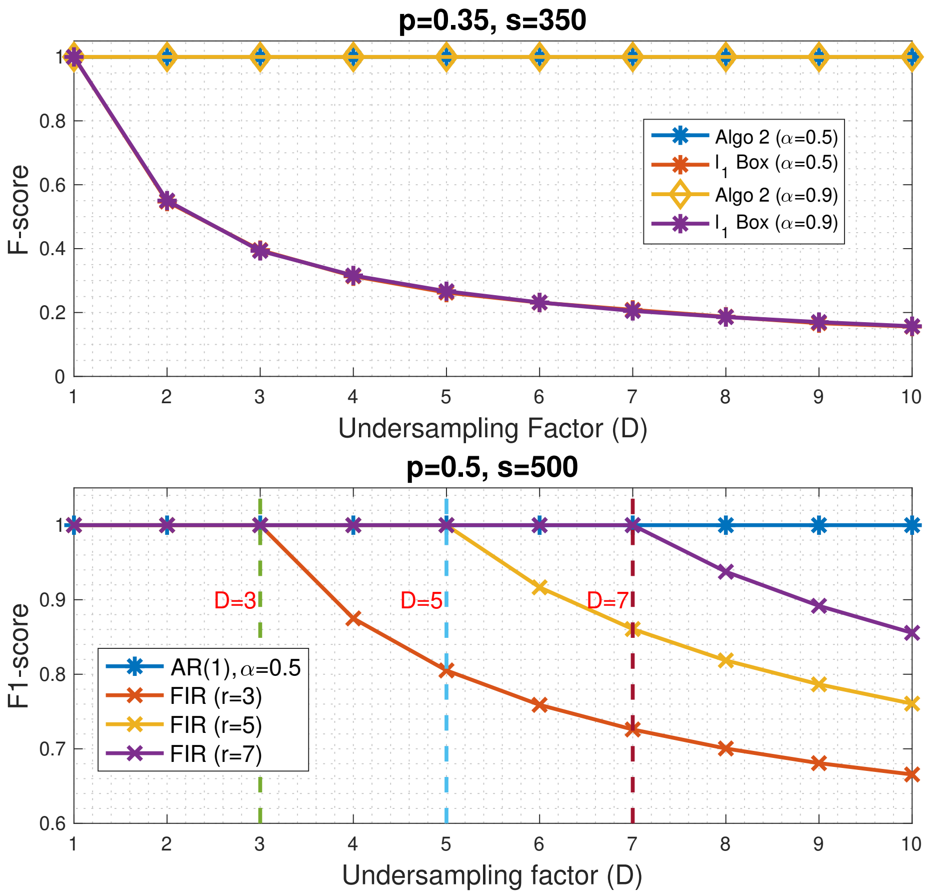


Figure 2.5. (Top) Quantitative comparison of Algorithm 3 against box-constrained l_1 minimization method with noiseless measurements (with tolerance $t_0 = 0$). (Bottom) (Role of Filter Memory): Average F-score vs. D for FIR and IIR (AR(1)) filters. Each dotted line indicates the corresponding theoretical transition point (D $= r$).

For synthetic data, ε is chosen using the norm of the noise term $\|\mathbf{w}\|_2$. This *oracle* choice ensures most favorable parameter tuning for the (P1), although a more realistic choice would be to set $\varepsilon = \sqrt{M}\sigma$ according to the noise power (σ). In the noiseless setting, we choose $\varepsilon = 0$. The problem (P1) is a standard convex relaxation of (P0) which promotes sparsity as well as tries to impose the binary constraint via the box-relaxation (introduced in Section II-C). In Figure 2.5 (Top), we plot the F-score ($t_0 = 0$) as a function of D. As can be observed, Algorithm 3 consistently achieves an F-score of 1, whereas the F-score of l_1 minimization shows a decay as D increases. This confirms Lemma 5 that for D > 1, using box-constraints with l_1 norm minimization is not enough to enable exact recovery from low rate measurements. In absence

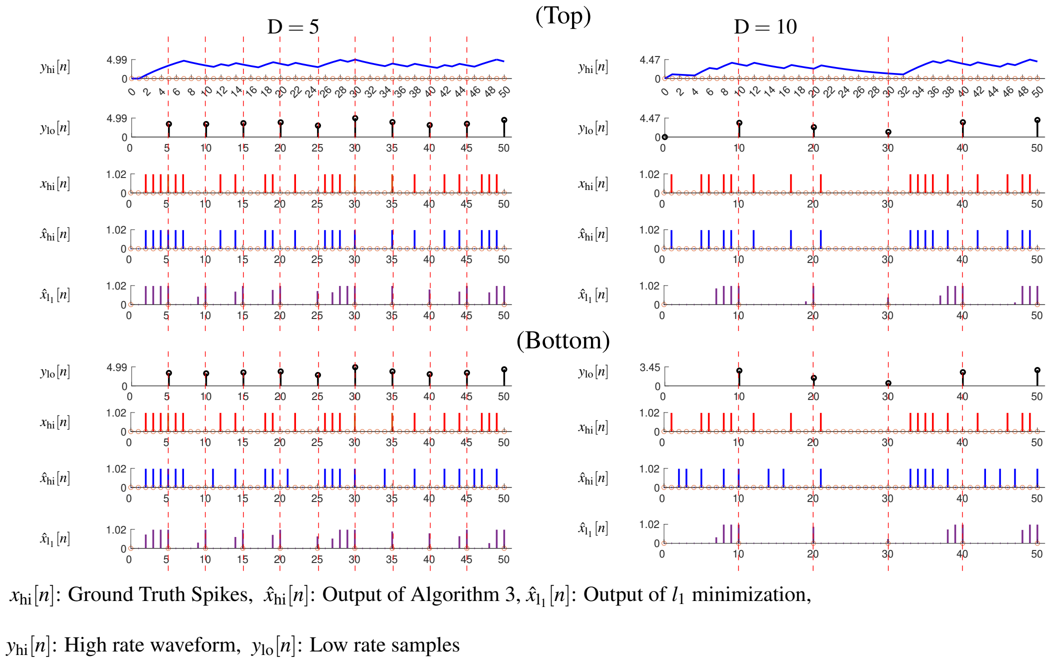


Figure 2.6. Qualitative comparison of Algorithm 3 and box-constrained l_1 minimization on simulated data. For each simulation noisy measurements are generated with $\alpha = 0.9$ such that the noise realization (Top) obeys the bound $|w[n]| \leq \Delta\theta_{\text{min}}$ (from Theorem 4) and (Bottom) violates the bound. For larger noise (Bottom), the spike recovery is imperfect but the spike count can still be exactly recovered using Algorithm 3.

of noise, the performance of Algorithm 3 is not affected by the filter parameter α as shown in Figure 2.5 (Top).

Next, we compare the reconstruction from the decimated output of (i) an AR(1) filter and (ii) an FIR filter of length r driven by the same input $\mathbf{x}_{\text{hi}} \in \{0,1\}^{1000}$. We choose the FIR filter $\mathbf{h} = [1, \alpha, \cdots, \alpha^{r-1}]^\top$ (truncation of the IIR filter) with $\alpha = 0.5$. Algorithm 3 is applied to the low-rate AR(1) measurements, whereas the algorithm proposed in [46] is used for the FIR case. The algorithm applied for the FIR case can provably operate with the optimal number of measurements when $\alpha = 0.5$ and hence, we chose this specific value for the filter parameter. In Figure 2.5 (Bottom), we again compare the average F-score as a function of D, averaged over 10000 Monte Carlo runs, for $p = 0.5$. As predicted by Lemma 6, despite utilizing binary priors, the error for the FIR filter shows a phase transition when D $> r$. This demonstrates the critical role played by the infinite memory of the AR(1) filter in achieving exact recovery with arbitrary

D.

Performance of noisy spike decoding

We generate noisy measurements of the form (2.29), where $w[n]$ and $x_{\text{hi}}[n]$ satisfy assumptions **(A1-A2)**. We illustrate some representative examples of recovered spikes on synthetic data. In Figure 2.6, we display the recovered super-resolution estimates on synthetically generated measurements for two undersampling factors $\mathrm{D} = 5$ (left), 10 (right). For each D, the top plots show the spikes recovered using Algorithm 3 and l_1 minimization with box-constraint where the noise realization obeys the bound in Theorem 4, while the bottom plots show the same for noise realization violating the bound. The output of l_1 minimization with box-constraint is inaccurate, and the spikes are clustered towards the end of each block of length D. This bias is consistent with the prediction made by our theoretical results in Lemma 5. When the noise is small enough (top), Algorithm 3 exactly decodes the spikes, including the ones occurring between two consecutive low-rate samples as predicted by Theorem 4. In presence of larger noise (violating the bound), the spikes estimated using l_1 minimization continue to be biased to be clustered towards the end of the block. Although the spikes recovered using Algorithm 3 are not exact, most of the detected spikes are within a tolerance window of ground truth spikes. In fact, the spike count estimation is perfect as predicted by Theorem 6. We next quantitatively evaluate the performance in presence of noise, where the metrics are computed with $t_0 = 2$. In Figure 2.7 (Top), we plot the F-score as a function of D for different values of α. For a fixed α, the F-score of both methods decays with increasing D, but Algorithm 3 consistently attains a higher F-score compared to l_1 minimization. We observe that $\alpha = 0.5$ leads to a higher F-score potentially due to having a larger $\Delta\theta_{\text{min}}(\alpha, \mathrm{D})$ compared to $\alpha = 0.9$. Next, in Figure 2.9, we study the behavior of spike detection as a function of the spiking probability p, while keeping D fixed at $\mathrm{D} = 5$. When σ is fixed, the performance trend is not significantly affected by the spiking probability. At first, this may seem surprising as the expected sparsity is growing while the number of measurements is unchanged. However, since our algorithm exploits the binary nature

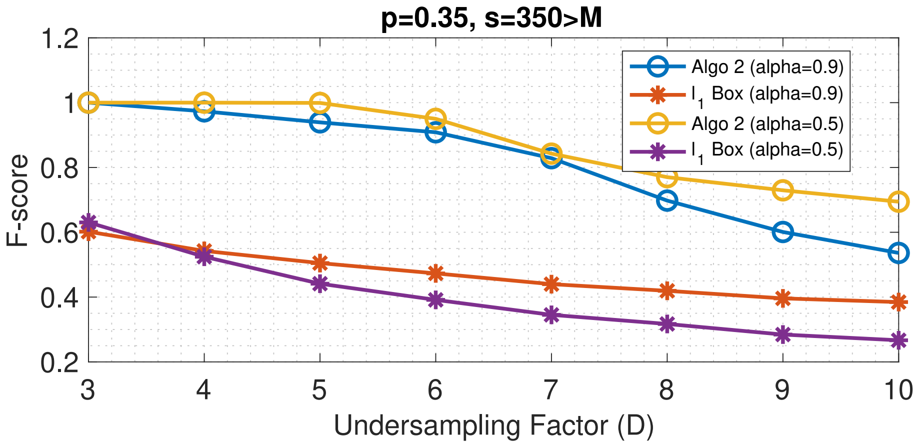


Figure 2.7. Spike detection performance with noisy measurements. (Top) F-score vs. D for different filter parameters α ($\sigma = 0.01$). Here, $L = 1000$ and expected sparsity $s = 350$ where we operate in the regime $s > M$. The F-score is computed with a tolerance of $t_0 = 2$.

of the spikes (and not just sparsity), it can handle larger sparsity levels. The spikes reconstructed using l_1 minimization achieve a much lower F-score than Algorithm 3 since the former fails to succeed when the sparsity is large. As expected, smaller σ leads to higher F-scores.

In Figure 2.10, we study the probability of erroneous spike detection as a function of D and validate the upper bound derived in Theorem 5. Recall that the decoding is considered successful if "every" spike is detected correctly. Therefore, it becomes more challenging to "exactly super-resolve" all the spikes in presence of noise as the desired resolution becomes finer. We calculate the empirical probability of error and overlay the corresponding theoretical bound. As shown in Figure 2.10, the empirical probability of error is indeed upper bounded by the bound computed by our analysis. The empirical probability of error increases as a function of undersampling factor D.

Finally, we evaluate the noise tolerance of the proposed methodology by comparing the average F-score as a function of the noise level σ, while keeping the spiking rate and undersampling factor fixed at $p = 0.35$ and $\text{D} = 5$, respectively. As seen in Figure 2.8 (Top), the performance of both algorithms degrades with increasing noise level and this is also consistent with the intuition that it becomes harder to super-resolve spikes with more noise. However, for both filter parameters considered in this experiment Algorithm 3 has a higher F-score compared

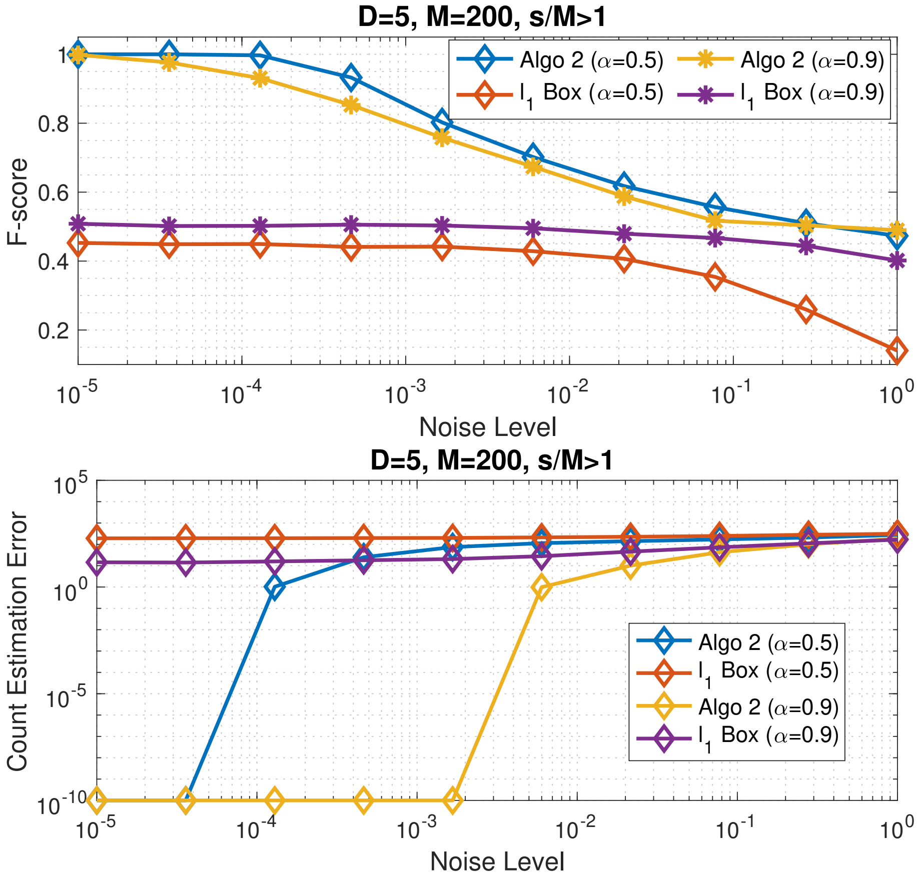


Figure 2.8. Spike detection performance with noisy measurements for different filter parameters α. (Top) F-score vs. noise level (σ) (Bottom) Count estimation error vs. noise level. Here, $L = 1000$ and expected sparsity is fixed at $s = 350$ where we operate in the regime $s > M$. The F-score is computed with a tolerance of $t_0 = 2$.

to box-constrained l_1 minimization. For large noise levels (comparable to spike amplitude $A = 1$), the performance gap decreases for $\alpha = 0.9$ but Algorithm 3 achieves a much higher F-score for $\alpha = 0.5$ at all noise levels.

As discussed in Section 2.4.4, we next study a relaxed notion of spike recovery which focuses on the spike counts occurring between two consecutive low-rate samples. Let $\Gamma = [\gamma[0], \cdots, \gamma[M-1]]^\top$ be the vector of counts and $\widehat{\Gamma}$ be its estimate. In Figure 2.8 (Bottom) we plot the average l_1 distance $\|\Gamma - \widehat{\Gamma}\|_1$ as a function of the noise level. We observe that for $\alpha = 0.9$ (it can be verified from Figure 2.3 (Top) that $0.9 \in \mathscr{F}_5$), it is possible to exactly recover the spike counts at higher noise even though the F-score (for timing recovery) has dropped below 1.

However, this is not the case for $\alpha = 0.5$, since $0.5 \notin \mathscr{F}_5$. This is consistent with the conclusion of Theorem 6 which states that when $\alpha \in \mathscr{F}_{\mathrm{D}}$, the noise tolerance for exact count recovery can be much larger than exact spike recovery since $\Delta^c_{\min}(\alpha, \mathrm{D}) > \Delta\theta_{\min}(\alpha, \mathrm{D})$.

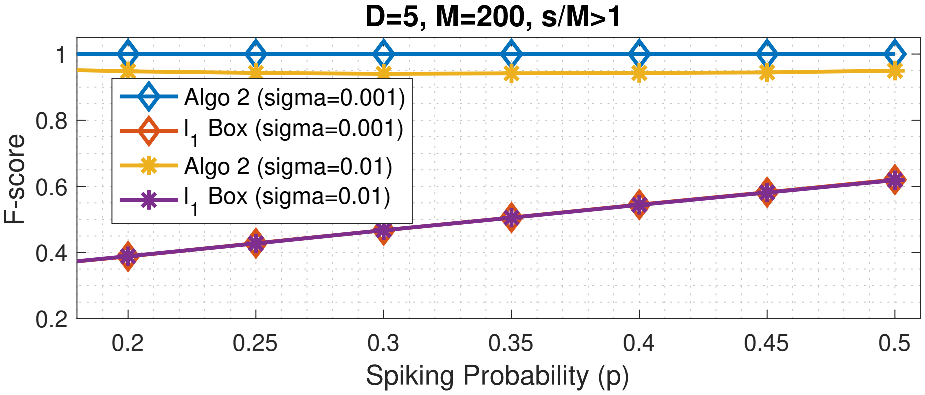


Figure 2.9. Spike detection performance with noisy measurements. F-score vs. spiking probability (p) for different noise levels σ (fix $\alpha = 0.9$, $\mathrm{D} = 5, L = 1000$) in the extreme compression regime $s > M$.

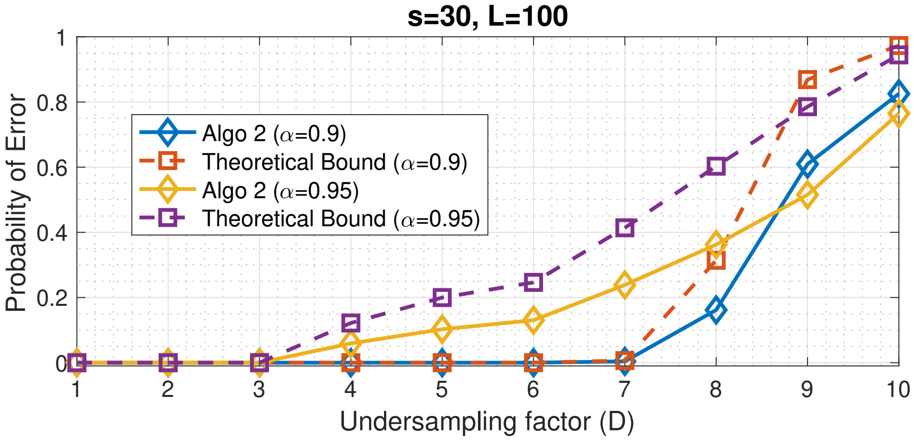


Figure 2.10. Probability of erroneous detection of high-rate spikes $\mathbf{x}_{\mathrm{hi}} \in \{0, 1\}^L$ as a function of the undersampling factor D. Theoretical upper bounds are overlaid using dotted lines. Here, $L = 100$.

Spike Deconvolution from Real Calcium Imaging Datasets

We now discuss how the mathematical framework developed in this chapter can be used for super-resolution spike deconvolution in calcium imaging. Two-photon calcium imaging is a widely used imaging technique for large scale recording of neural activity with high spatial but poor temporal resolution. In calcium imaging, the signal $\mathbf{x}_{\text{hi}}$ corresponds to the underlying neural spikes which is modeled to be binary valued on a finer temporal scale [9, 55]. Each neural spike results in a sharp rise in Ca^{2+} concentration followed by a slow exponential decay, leading to superposition of the responses from nearby spiking events [9–11]. This calcium transient can be modeled by the first order autoregressive model introduced in Section 2.4.1. The decay time constant depends on the calcium indicator and essentially determines the filter parameter α. The signal $y_{\text{hi}}[n]$ is an unobserved signal corresponding to sampling the calcium fluorescence at a high sampling rate (at the same rate as the underlying spikes). The observed calcium signal $y_{\text{lo}}[n]$ corresponds to downsampling $y_{\text{hi}}[n]$ at an interval determined by the frame rate of the microscope. The frame rate of a typical scanning microscopy system (that captures the changes in the calcium fluorescence) is determined by the amount of time required to spatially scan the desired field of view, which makes it significantly slower compared to the temporal scale of the neural spiking activity. We model this discrepancy by the downsampling operation (by a factor D). Therefore, the mathematical framework developed in this chapter can be directly applied to reconstruct the underlying spiking activity at a temporal scale finer than the sampling rate of the calcium signal. Using real calcium imaging data, we demonstrate a way to fuse our algorithm with a popular spike deconvolution algorithm called OASIS [51]. OASIS solves an l_1 minimization problem similar to (P1) with only the non-negativity constraint, in order to exploit the sparse nature of the spiking activity. Unlike our approach where we wish to obtain spikes representation on a finer temporal scale, OASIS returns the spike estimates on the low-resolution grid. This is typically used to infer the spiking rate over a temporal bin equal to the sampling interval. We demonstrate that our proposed framework can be integrated with OASIS and improve its performance. As we saw in the synthetic experiments, the noise level is an important consideration. By augmenting Algorithm 3 with OASIS, referred as "B-OASIS", the denoising power of l_1 minimization can be leveraged.Let $\hat{\mathbf{x}}_{\text{l1}} \in \mathbb{R}^M$ be the estimate obtained on a low-resolution grid by solving the l_1 minimization problem such as the one implemented in OASIS. We can obtain an estimate of the denoised calcium signal as $\hat{y}_{\text{lo}}[n] = \alpha^{\text{D}}\hat{y}_{\text{lo}}[n] + \hat{x}_{\text{l1}}[n], n \geq 1$ and $\hat{y}_{\text{lo}}[0] = \hat{x}_{\text{l1}}[0]$. We can now utilize the denoised calcium signal $\hat{y}_{\text{lo}}[n]$ generated by OASIS to obtain the estimate $c_e[n]$ indirectly. Due to the non-linear processing done by OASIS, it is difficult to obtain the resulting noise statistics. An important advantage of Algorithm 3 is that it does not rely on the knowledge of the noise statistics. Hence, we can directly apply Algorithm 3 on $\hat{c}_e[n] = \hat{y}_{\text{lo}}[n] - \alpha^{\text{D}}\hat{y}_{\text{lo}}[n-1]$ (instead of $c_e[n]$) to obtain a binary "fused super-resolution spike estimate".

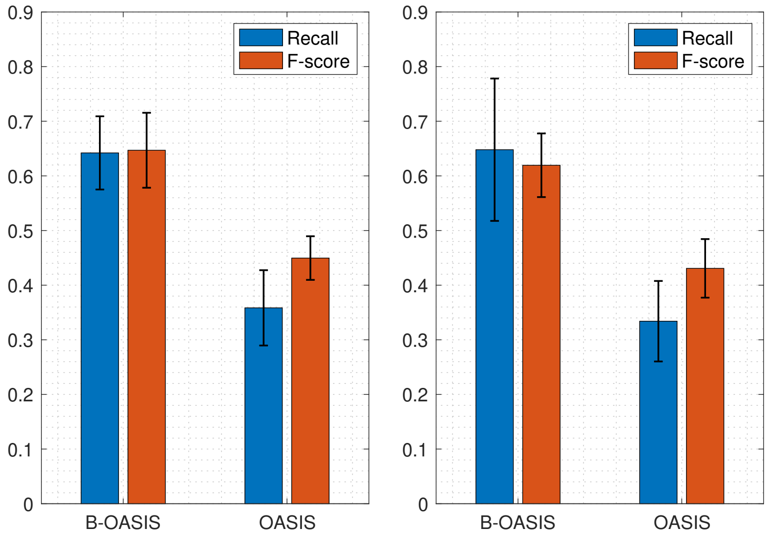


Figure 2.11. Spike detection performance of OASIS and B-OASIS on GCaMP6f dataset sampled at (Left) 60 Hz and (Right) 30 Hz. We compare the average F-score of data points where the F-score of OASIS is < 0.5. Standard deviation is depicted using the error bars.

Results on Real Data

We evaluate the algorithms on the publicly available GENIE dataset [61, 62] which consists of simultaneous calcium imaging and *in vivo* cell-attached recording from the mouse visual cortex using genetically encoded GCaMP6f calcium indicator GCaMP6f [61, 62]. The calcium images were acquired at a frame rate of 60 Hz and the ground truth electrophysiology signal was digitized at 10 KHz and synchronized with the calcium frames. In addition to using the original data, we also synthetically downsample it to emulate the effect of a lower frame rate of 30 Hz, and evaluate how the performance changes by this downsampling operation.

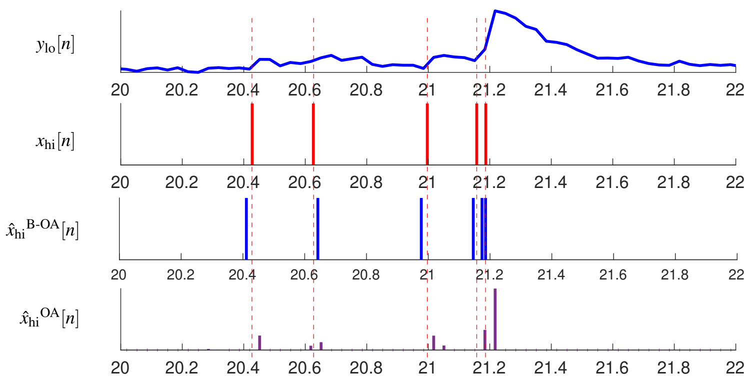


Figure 2.12. Example of spike reconstruction on GENIE dataset (GCaMP6f indicator) using OASIS and B-OASIS (binary augmented) with calcium signal sampled at 30Hz.

In Figure 2.12, we extract an interval of ~ 2 sec (from the neuron 1 of the GCaMP6f indicator dataset) and qualitatively compare the detected spikes with the ground truth. We downsample the data by a factor of 2 to emulate frame rate of 30 Hz, the low-rate grid becomes coarser. As a result of which, we observe an offset between ground truth spikes and estimate produced by OASIS. However, with the help of binary priors (B-OASIS), we can output spikes that are not restricted to be on the coarser scale, and this mitigates the offset observed in the raw estimates obtained by OASIS.

We quantify the improvement in the performance by comparing the F-scores of OASIS and B-OASIS at both sampling rates (60 and 30 Hz). Since the output of OASIS is non-binary, the estimated spikes are binarized by thresholding. To ensure a fair comparison, we select the threshold by a $80-20$ cross-validation scheme that maximizes the average F-score on a held-out validation set (averaged over 3-random selections of the validation set). The tolerance for the F-score was set at 100 ms. The dataset consisted of 34 traces of length ~ 234 s. The OASIS algorithm has an automated routine to estimate the parameter α, which we utilize for our experiments. The amplitude A is estimated using the procedure described in Appendix F. We use $\mathrm{D} = 12$ to obtain the spike representation for B-OASIS. In order to quantify the performance boost achieved by augmentation, we isolate the traces where the $F-$score of OASIS drops below 0.5 and compare the average F-score and recall for these data points. As shown in Figure 2.11, at both sampling rates, we see a significant improvement in the average F-score of B-OASIS over OASIS, attributed to an increase in recall while keeping the precision unchanged. Additionally, despite downsampling, the spike detection performance is not significantly degraded with binary priors, although the detection criteria were unchanged.

2.6 Conclusion

In the first part of this chapter, we addressed the problem of identifying a finite-valued input from uniformly downsampled measurements at the output of a known finite length filter. We established that the overall linear map remains injective over the set of finite-valued signals provided the number of measurements exceed N/L. Under a certain decay condition on the filter, we show that it is possible to design a computationally efficient sequential decoding algorithm that leverages the finite-valued constraint instead of relaxing these conditions. The proposed algorithm can recover signals with sparsity larger than N/L with only $O(N/L)$ measurements which is also supported by our numerical simulations. Our results establish that it is indeed possible to develop computationally efficient approaches without relaxation and thereby avoid the performance degradation incurred due to relaxation.

In the second part of this chapter, we theoretically established the benefits of binary priors in super-resolution, and showed that it is possible to achieve significant reduction in sample complexity over sparsity-based techniques. Using an AR(1) model, we developed and analyzed an efficient algorithm that can operate in the extreme compression regime ($M \ll K$) by exploiting the special structure of measurements and trading memory for computational efficiency at run-time. We also demonstrated that binary priors can be used to boost the performance of existing neural spike deconvolution algorithms. In the future, we will develop algorithmic frameworks for incorporating binary priors into different neural spike deconvolution pipelines and evaluate the performance gain on diverse datasets. The extension of this binary framework for higher-order AR filters is another exciting future direction.

Chapter 2, in part, is a reprint of the material as it appears in the following papers:

- P. Sarangi, Ryoma Hattori, Takaki Komiyama and P. Pal, “Super-resolution with Binary Priors: Theory and Algorithms,” IEEE Transactions on Signal Processing, 2023.
- P. Sarangi and P. Pal, “No Relaxation: Guaranteed Recovery of Finite-Valued Signals from Undersampled Measurements,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5440-5444.

The dissertation author was the primary investigator and author of these papers.

2.7 Appendices

2.7.1 Appendix A: Proof of Theorem 3

Proof. We show that for any α in $0 < \alpha < 1$, except possibly for a set consisting of only a finite number of points, (2.16) always has a unique binary solution. Consider all possible $\mathrm{D}-$dimensional ternary vectors with their entries chosen from $\{-1,0,1\}$, and denote them as $\mathbf{v}^{(i)} = [v_1^{(i)}, v_2^{(i)}, \cdots, v_{\mathrm{D}}^{(i)}]^T \in \{-1,0,1\}^{\mathrm{D}}, 0 \le i \le 3^{\mathrm{D}} - 1$. We use the convention that $\mathbf{v}^{(0)} = 0$. For every $i > 0$, we define a set $\mathscr{Z}_{\mathbf{v}^{(i)}}$ determined by $\mathbf{v}^{(i)}$ as $\mathscr{Z}_{\mathbf{v}^{(i)}} := \left\{x \in (0,1) \middle| \sum_{k=1}^{\mathrm{D}} v_k^{(i)} x^{\mathrm{D}-k} = 0\right\}$. Notice that $p_i(x) := \sum_{k=1}^{\mathrm{D}} v_k^{(i)} x^{\mathrm{D}-k}$ denotes a polynomial (in x) of degree at most $\mathrm{D}-1$, whose coefficients are given by the ternary vector $\mathbf{v}^{(i)}$. The set $\mathscr{Z}_{\mathbf{v}^{(i)}}$ denotes the set of zeros of $p_i(x)$ that are contained in $(0,1)$. Since the degree of $p_i(x)$ is at most $\mathrm{D}-1$, $\mathscr{Z}_{\mathbf{v}^{(i)}}$ is a finite set with cardinality at most $\mathrm{D}-1$.

Now suppose that the binary solution of (2.16) is non-unique, i.e., there exist $\mathbf{u}, \mathbf{w} \in \{0, A\}^L$, $\mathbf{u} \neq \mathbf{w}$, such that

$$\mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{u} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{w} \Rightarrow \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{u} - \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{w} = \mathbf{0} \tag{2.42}$$

By partitioning $\mathbf{u}, \mathbf{w}$ into blocks $\mathbf{u}^{(n)}, \mathbf{w}^{(n)}$ in the same way as in (2.12), we can re-write (2.42) as

$u^{(0)} = w^{(0)}$ and

$$\sum_{i=1}^{\mathrm{D}} \frac{1}{A}([\mathbf{u}^{(j)}]_i - [\mathbf{w}^{(j)}]_i)\alpha^{\mathrm{D}-i} = 0, \quad 1 \leq j \leq M-1 \tag{2.43}$$

Since $\mathbf{u} \neq \mathbf{w}$, they differ at least at one block, i.e., there exists some $j_0, 1 \leq j_0 \leq M-1$ such that $\mathbf{u}^{(j_0)} \neq \mathbf{w}^{(j_0)}$. Define $\mathbf{b} := \frac{1}{A}(\mathbf{u}^{(j_0)} - \mathbf{w}^{(j_0)})$. Then, $\mathbf{b}$ is a non-zero ternary vector, i.e., $\mathbf{b} \in \{-1,0,1\}^{\mathrm{D}}$. Now from (2.43), we have

$$\sum_{i=1}^{\mathrm{D}} [\mathbf{b}]_i \alpha^{\mathrm{D}-i} = 0, \tag{2.44}$$

which implies that $\alpha \in \mathscr{Z}_{\mathbf{b}}$. Since $\mathbf{b}$ can be any one of the $3^{\mathrm{D}} - 1$ ternary vectors $\{\mathbf{v}^{(i)}\}_{i=1}^{3^{\mathrm{D}}-1}$, (2.44) holds if and only if $\alpha \in \mathbb{S} := \bigcup_{i=1}^{3^{\mathrm{D}}-1} \mathscr{Z}_{\mathbf{v}^{(i)}}$, i.e., α is a root of at least one of the polynomials $p_i(x)$ defined by the vectors $\mathbf{v}^{(i)}$ as their coefficients. For each $\mathbf{v}^{(i)}$, since the cardinality of $\mathscr{Z}_{\mathbf{v}^{(i)}}$ is at most $\mathrm{D}-1$, $\mathbb{S}$ is a finite set (of cardinality at most $(\mathrm{D}-1)(3^{\mathrm{D}}-1)$), and therefore its Lebesgue measure is 0. This implies that (2.16) has a non-unique binary solution only if α belongs to the measure zero set $\mathbb{S}$, thereby proving the theorem. □

2.7.2 Appendix B: Proof of Lemma 3 and Lemma 5

Proof. (i) Let s_n denote the sparsity (number of non-zero elements) of the n^{th} block $\mathbf{x_{\text{hi}}}^{(n)}$ of $\mathbf{x}_{\text{hi}}$. Then, the total sparsity is $\|\mathbf{x}_{\text{hi}}\|_0 = \sum_{n=0}^{M-1} s_n$. We will construct a vector $\mathbf{v} \in \mathbb{R}^L$, $\mathbf{v} \neq \mathbf{x}_{\text{hi}}$ that satisfies $\mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{v}$ and $\|\mathbf{x}_{\text{hi}}\|_0 \geq \|\mathbf{v}\|_0$. Following (2.12), consider the partition of $\mathbf{v}$ $\mathbf{v} = [v^{(0)}, \mathbf{v}^{(1)\top}, \cdots, \mathbf{v}^{(M-1)\top}]^\top$. Firstly, we assign $v^{(0)} = c[0] = {x_{\text{hi}}}^{(0)}$. We construct $\mathbf{v}^{(n)}$ as follows. For each $n \geq 1$, there are three cases:

Case I: $s_n = 0$. In this case, $\mathbf{x_{\text{hi}}}^{(n)} = \mathbf{0}$ and hence $c[n] = 0$. Therefore, we assign $\mathbf{v}^{(n)} = \mathbf{x_{\text{hi}}}^{(n)} = \mathbf{0}$.

Case II: $s_n = 1$. First suppose that $[\mathbf{x_{\text{hi}}}^{(n)}]_{\mathrm{D}} = 0$. We construct $\mathbf{v}^{(n)}$ as follows:

$$[\mathbf{v}^{(n)}]_k = \begin{cases} c[n], & \text{if } k = \mathrm{D} \\ 0, & \text{else} \end{cases}. \tag{2.45}$$

Next suppose that $[{\mathbf{x}_{\mathrm{hi}}}^{(n)}]_{\mathrm{D}} \neq 0$. Since $s_n = 1$, this implies that $[{\mathbf{x}_{\mathrm{hi}}}^{(n)}]_k = 0, k = 1, \cdots, \mathrm{D}-1$. In this case, we construct $\mathbf{v}^{(n)}$ as follows:

$$[\mathbf{v}^{(n)}]_k = \begin{cases} c[n]/\alpha, & \text{if } k = \mathrm{D}-1 \\ 0, & \text{else} \end{cases}. \tag{2.46}$$

Notice that both (2.45) and (2.46) ensure that $\mathbf{v}^{(n)} \neq {\mathbf{x}_{\mathrm{hi}}}^{(n)}$ and $c[n] = \mathbf{h}_\alpha^T \mathbf{v}^{(n)}$. Moreover, $\|\mathbf{v}^{(n)}\|_0 = s_n$.

Case III: $s_n \geq 2$. In this case, we follow the same construction as (2.45). As before $\mathbf{v}^{(n)}$ satisfies $c[n] = \mathbf{h}_\alpha^\top \mathbf{v}^{(n)}$. Since $\|{\mathbf{x}_{\mathrm{hi}}}^{(n)}\|_0 \geq 2$ and $\|\mathbf{v}^{(n)}\|_0 = 1$, we automatically have $\mathbf{v}^{(n)} \neq {\mathbf{x}_{\mathrm{hi}}}^{(n)}$, and $\|\mathbf{v}^{(n)}\|_0 < s_n$. Therefore, combining the three cases, we can construct the desired vector $\mathbf{v}$ that satisfies $\mathbf{v} \neq \mathbf{x}_{\mathrm{hi}}$, $\mathbf{c} = \mathbf{H}_{\mathrm{D}}(\alpha)\mathbf{v}$, and $\|\mathbf{v}\|_0 \leq \sum_{n=0}^{M-1} s_n = \|{\mathbf{x}_{\mathrm{hi}}}^{(n)}\|_0$. Therefore, the solution $\mathbf{x}^\star$ to (P0) satisfies $\|\mathbf{x}^\star\|_0 \leq \|\mathbf{v}\|_0 \leq \|{\mathbf{x}_{\mathrm{hi}}}^{(n)}\|_0$.

(ii) In this case, we construct $\mathbf{v}^{(n_0)}$ according to Case III. Since $\|\mathbf{v}^{(n_0)}\|_0 < s_{n_0}$, and $\|\mathbf{v}^{(n)}\|_0 \leq s_n, n \neq n_0$, we have $\|\mathbf{v}\|_0 < \|\mathbf{x}_{\mathrm{hi}}\|_0$, implying $\|\mathbf{x}^\star\|_0 \leq \|\mathbf{v}\|_0 < \|\mathbf{x}_{\mathrm{hi}}\|_0$. □

2.7.3 Proof of Lemma 5

Proof. We will construct a vector $\mathbf{v} \in \mathbb{R}^L$ whose support is of the form (2.24), that is feasible for (P1-B), and we will prove that it has the smallest l_1 norm. Using the block structure given by (2.12), we choose $\mathbf{v}^{(0)} = c[0]$. For each $n \geq 1$, we construct $\mathbf{v}^{(n)}$ based on the following two cases:

Case I: $c[n] \geq A$. Let k_n be the largest integer such that the following holds: $\mu[n] := A(1 + \alpha + \cdots + \alpha^{k_n - 1}) \leq c[n]$, where $1 \leq k_n \leq \mathrm{D}$. Note that $k_n = 1$ always produces a valid lower bound. However, we are interested in the largest lower bound on $c[n]$ of the above form. We choose

$$[\mathbf{v}^{(n)}]_k = \begin{cases} A, & \text{if } \mathrm{D} - k_n + 1 \leq k \leq \mathrm{D} \\ (c[n] - \mu[n])/\alpha^{k_n}, \text{ if } k = \mathrm{D} - k_n \\ 0, \text{ else} \end{cases}$$

It is easy to verify that $\mathbf{h}_\alpha^\top \mathbf{v}^{(n)} = c[n]$. From the definition of k_n, it follows that $\mu[n] \leq c[n] < \mu[n] + A\alpha^{k_n}$ and hence, $0 \leq (c[n] - \mu[n])/\alpha^{k_n} < A$, which ensures that $\mathbf{v}$ obeys the box-constraints in (P1-B). Now, let $\mathbf{v}_f \in \mathbb{R}^L$ be any feasible point of (P1-B) which must be of the form $\mathbf{v}_f^{(0)} = c[0], \mathbf{v}_f^{(n)} = \mathbf{v}^{(n)} + \mathbf{r}^{(n)}$, where $\mathbf{r}^{(n)} \in \mathscr{N}(\mathbf{h}_\alpha^\top)$ is a vector in the null-space of $\mathbf{h}_\alpha^\top$. It can be verified that the following vectors $\{\mathbf{w}_t\}_{t=1}^{D-1}$ form a basis for $\mathscr{N}(\mathbf{h}_\alpha^\top)$:

$$[\mathbf{w}_t]_k = \begin{cases} 1, & k = t \\ -\alpha, & k = t+1 \\ 0, & \text{else} \end{cases},$$

Therefore, $\exists \, \{\beta_t^{(n)}\}_{t=1}^{\mathrm{D}-1}$ such that $\mathbf{r}^{(n)} = \sum_{t=1}^{\mathrm{D}-1} \beta_t^{(n)} \mathbf{w}_t$.

We further consider two scenarios: (i) $1 \leq k_n \leq \mathrm{D} - 2$. In this case $[\mathbf{v}^{(n)}]_1 = 0$, and for $k = 1, 2, \cdots D$, $[\mathbf{v}_f^{(n)}]_k$ satisfies [1]

$$[\mathbf{v}_f^{(n)}]_k = \begin{cases} \beta_k^{(n)}, \text{ if } k = 1 \\ \beta_k^{(n)} - \alpha\beta_{k-1}^{(n)}, \text{ if } 2 \leq k \leq \mathrm{D} - k_n - 1 \\ [\mathbf{v}^{(n)}]_k + \beta_k^{(n)} - \alpha\beta_{k-1}^{(n)}, \text{ if } k = \mathrm{D} - k_n \\ A + \beta_k^{(n)} - \alpha\beta_{k-1}^{(n)}, \text{ if } \mathrm{D} - k_n + 1 \leq k \leq \mathrm{D} - 1 \\ A - \alpha\beta_{k-1}^{(n)}, \text{ if } k = \mathrm{D} \end{cases}$$

To ensure $\mathbf{v}_f^{(n)}$ is a feasible point for (P1-B), the following must hold: $0 \leq \beta_{\mathrm{D}-1}^{(n)} \leq A/\alpha$ and $0 \leq \beta_1^{(n)} \leq A$. For $2 \leq k \leq \mathrm{D} - k_n - 1$, the constraint $[\mathbf{v}_f^{(n)}]_k \geq 0$ implies $\beta_k^{(n)} \geq \alpha\beta_{k-1}^{(n)}$. Since $\beta_1^{(n)} \geq 0$, it follows that $\beta_k^{(n)} \geq 0$ for all $2 \leq k \leq \mathrm{D} - k_n - 1$. For $\mathrm{D} - k_n + 1 \leq k \leq \mathrm{D} - 1$, the constraint $[\mathbf{v}_f^{(n)}]_k \leq A$ implies $\beta_{k-1}^{(n)} \geq \beta_k^{(n)}/\alpha$. Since $\beta_{\mathrm{D}-1}^{(n)} \geq 0$, it follows that $\beta_k^{(n)} \geq 0$ for all

[1] In the definition of $\mathbf{v}_f^{(n)}$, an assignment will be ignored if the specified interval for k is empty.

$\mathrm{D}-k_n \leq k \leq \mathrm{D}-1$. (ii) $k_n \in \{\mathrm{D}-1, \mathrm{D}\}$. In this case, for $k = 1, 2, \cdots, D$, $[\mathbf{v}_f^{(n)}]_k$ satisfies

$$[\mathbf{v}_f^{(n)}]_k = \begin{cases} [\mathbf{v}^{(n)}]_1 + \beta_1^{(n)}, \text{ if } k = 1 \\ A + \beta_k^{(n)} - \alpha\beta_{k-1}^{(n)}, \text{ if } 2 \leq k \leq \mathrm{D}-1 \\ A - \alpha\beta_{k-1}^{(n)}, \text{ if } k = \mathrm{D} \end{cases}$$

For $2 \leq k \leq \mathrm{D}-1$, the box-constraint $[\mathbf{v}_f^{(n)}]_k \leq A$ implies $\beta_{k-1}^{(n)} \geq \beta_k^{(n)}/\alpha$. Since $\beta_{\mathrm{D}-1}^{(n)} \geq 0$, it follows that $\beta_k^{(n)} \geq 0$ for all $1 \leq k \leq \mathrm{D}-1$. Summarizing, we have established that $\beta_i^{(n)} \geq 0, \forall i$.

Case II: $c[n] < A$. In this case, $\mathbf{v}^{(n)}$ is constructed following (2.45), and hence $\mathbf{v}_f^{(n)}$ has the following structure:

$$[\mathbf{v}_f^{(n)}]_k = \begin{cases} \beta_k^{(n)}, \text{if } k = 1 \\ -\alpha\beta_{k-1}^{(n)} + \beta_k^{(n)}, \text{ if } 2 \leq k \leq \mathrm{D}-1 \\ c[n] - \alpha\beta_{k-1}^{(n)}, \text{ if } k = \mathrm{D} \end{cases}$$

To ensure $\mathbf{v}_f^{(n)}$ is a feasible point, it must hold that $\beta_1^{(n)} \geq 0, \beta_k^{(n)} \geq \alpha\beta_{k-1}^{(n)} \geq 0$ for $2 \leq k \leq \mathrm{D}-1$. Hence, in both Cases I and II, we established that $\beta_k^{(n)} \geq 0$. For each case, since $\mathbf{v}_f^{(n)}$ is a non-negative vector $\forall n$, it can be verified that

$$\begin{aligned} \|\mathbf{v}_f\|_1 &= \sum_{n=0}^{M-1} \|\mathbf{v}_f^{(n)}\|_1 = v_f^{(0)} + \sum_{n=1}^{M-1} \sum_{k=1}^{\mathrm{D}} [\mathbf{v}_f^{(n)}]_k \\ &= \underbrace{c[0] + \sum_{n=1}^{M-1} \sum_{k=1}^{\mathrm{D}} [\mathbf{v}^{(n)}]_k}_{\|\mathbf{v}\|_1} + \sum_{n=1}^{M-1} \sum_{k=1}^{\mathrm{D}-1} (1-\alpha)\beta_k^{(n)} \end{aligned}$$

We used the fact that $\sum_{k=1}^{D} \sum_{t=1}^{D-1} \beta_t^{(n)} [\mathbf{w}_t]_k = \sum_{t=1}^{\mathrm{D}-1} (1-\alpha)\beta_t^{(n)}$. If $\mathbf{v}_f \neq \mathbf{v}$, we must have $\beta_k^{(n)} \neq 0$ for some k and $n > 0$. This implies that $\|\mathbf{v}_f\|_1 > \|\mathbf{v}\|_1$. It is easy to see that the support of the constructed vector is of the form (2.24). Moreover, based on the above argument, $\mathbf{v}$ is the only vector that has the minimum l_1 norm among all possible feasible points of (P1-B). □

2.7.4 Appendix C: Proof of Lemma 9

Proof. For any $0 < \alpha \leq 0.5$, we begin by showing that for an integer $p \geq 1$ the following inequality holds:

$$\sum_{k=1}^{p} \alpha^{\mathrm{D}-k} = \alpha^{\mathrm{D}-p-1}\left(\frac{1-\alpha^{p}}{1/\alpha-1}\right) < \alpha^{\mathrm{D}-p-1} \tag{2.47}$$

since $1/\alpha - 1 \geq 1$ and $1-\alpha^{p} < 1$ in the regime $0 < \alpha \leq 0.5$. Let $\mathscr{S}_1 = \{0, \alpha^{\mathrm{D}-1}, \alpha^{\mathrm{D}-2}, \alpha^{\mathrm{D}-1} + \alpha^{\mathrm{D}-2}\}$. Notice that the elements of $\mathscr{S}_1$ are sorted in ascending order for any α and D. Now, we recursively define the sets $\mathscr{S}_i$ as follows:

$$\mathscr{S}_i := \{\mathscr{S}_{i-1}, \mathscr{S}_{i-1} + \alpha^{\mathrm{D}-1-i}\},\ 2 \leq i \leq \mathrm{D}-1 \tag{2.48}$$

Our hypothesis is that for every $2 \leq i \leq \mathrm{D}-1$ $\alpha \in (0, 0.5]$ and D, the set $\mathscr{S}_i$ as defined in (2.48), is automatically sorted in ascending order. We prove this via induction. For $i = 2$, the sets $\mathscr{S}_1$ and $\mathscr{S}_1 + \alpha^{\mathrm{D}-3}$ are individually sorted. Moreover from (2.47), we can show that: $\max_{a \in \mathscr{S}_1} a = \alpha^{\mathrm{D}-1} + \alpha^{\mathrm{D}-2} < \alpha^{\mathrm{D}-3} = \min_{b \in \mathscr{S}_1 + \alpha^{\mathrm{D}-3}} b$. This shows that $\mathscr{S}_2$ is ordered, establishing the the base case of our induction. Now, assume $\mathscr{S}_i$ is ordered for some $2 \leq i \leq \mathrm{D}-2$. We need to show that $\mathscr{S}_{i+1}$ is also ordered. As a result of the induction hypothesis, both $\mathscr{S}_i$ and $\mathscr{S}_i + \alpha^{\mathrm{D}-2-i}$ are ordered. Using the ordering of $\mathscr{S}_i$, we have: $\max_{a \in \mathscr{S}_i} a = \sum_{j=1}^{i+1} \alpha^{\mathrm{D}-j}, \min_{b \in \mathscr{S}_i + \alpha^{\mathrm{D}-2-i}} b = \alpha^{\mathrm{D}-(i+1)-1}$. From (2.47), we can conclude that $\max_{a \in \mathscr{S}_i} a < \min_{b \in \mathscr{S}_i + \alpha^{\mathrm{D}-2-i}} b$ and hence, $\mathscr{S}_{i+1}$ is also ordered. This completes the induction proof. Also, note that for $\alpha \in (0, 0.5]$, we have $\Theta_{\alpha}^{\mathrm{sort}} = \mathscr{S}_{\mathrm{D}-1}$.

Let $\Delta_{\min}(\mathscr{S}_i)$ be the min. distance between the elements of the set $\mathscr{S}_i$. It is easy to see that

$\Delta_{\min}(\mathscr{S}_i) = \Delta_{\min}(\mathscr{S}_i + \alpha^{\mathrm{D}-2-i})$. Since $\mathscr{S}_i$ is sorted for $\alpha \in (0, 0.5]$, $\Delta_{\min}(\mathscr{S}_i)$ is given by:

$$\begin{aligned}\Delta_{\min}(\mathscr{S}_i) &= \min(\Delta_{\min}(\mathscr{S}_{i-1}), \min_{x \in \mathscr{S}_{i-1} + \alpha^{\mathrm{D}-1-i}} x - \max_{y \in \mathscr{S}_{i-1}} y) \\ &= \min\{\Delta_{\min}(\mathscr{S}_{i-1}), \alpha^{\mathrm{D}-i-1} - \sum_{j=1}^{i} \alpha^{\mathrm{D}-j}\}. \end{aligned} \tag{2.49}$$

Now, we use induction to establish the following conjecture:

$$\Delta_{\min}(\mathscr{S}_i) = \alpha^{\mathrm{D}-1},\ 1 \le i \le \mathrm{D}-1 \tag{2.50}$$

For the base case $i = 1$, $\Delta_{\min}(\mathscr{S}_1) = \min(\alpha^{\mathrm{D}-1}, \alpha^{\mathrm{D}-2} - \alpha^{\mathrm{D}-1}) = \alpha^{\mathrm{D}-1}$, where the last equality holds since $\alpha \in (0, 0.5] \Rightarrow \alpha^{\mathrm{D}-1}(1/\alpha - 1) \ge \alpha^{\mathrm{D}-1}$. Suppose (2.50) holds for some $1 \le i \le \mathrm{D}-2$. From the definition of $\Delta_{\min}(\mathscr{S}_{i+1})$ and the induction hypothesis that $\Delta_{\min}(\mathscr{S}_i) = \alpha^{D-1}$, it follows that $\Delta_{\min}(\mathscr{S}_{i+1}) = \min\{\alpha^{\mathrm{D}-1}, \alpha^{\mathrm{D}-(i+1)-1} - \sum_{j=1}^{i+1} \alpha^{\mathrm{D}-j}\}$. Again, from the definition of $\Delta_{\min}(\mathscr{S}_i)$ in (2.49), and the induction hypothesis we also have $\alpha^{\mathrm{D}-i-1} - \sum_{j=1}^{i} \alpha^{\mathrm{D}-j} \ge \Delta_{\min}(\mathscr{S}_i) = \alpha^{\mathrm{D}-1}$. Using this and the fact that $\alpha \le 0.5$, we can show:

$$\alpha^{\mathrm{D}-i-2} - \alpha^{\mathrm{D}-i-1} - \textstyle\sum_{j=1}^{i} \alpha^{\mathrm{D}-j} \ge \alpha^{\mathrm{D}-i-2} - 2\alpha^{\mathrm{D}-i-1} + \alpha^{\mathrm{D}-1} \ge \alpha^{\mathrm{D}-1} + \alpha^{\mathrm{D}-i-1}(1/\alpha - 2) \ge \alpha^{\mathrm{D}-1}$$

Therefore $\Delta_{\min}(\mathscr{S}_{i+1}) = \min\{\alpha^{\mathrm{D}-1}, \alpha^{\mathrm{D}-i-2} - \sum_{j=1}^{i+1} \alpha^{\mathrm{D}-j}\} = \alpha^{D-1}$. Thus, we can conclude that $\Delta_{\min}(\alpha, \mathrm{D}) = \Delta_{\min}(\mathscr{S}_{\mathrm{D}-1}) = \alpha^{\mathrm{D}-1}$. □

2.7.5 Appendix D: Proof of Theorem 5

Proof. The probability of incorrectly identifying ${\mathbf{x}_{\mathrm{hi}}}^{(n)}$ from a single measurement $c_e[n]$ is given by

$$p_e := \mathbb{P}({\hat{\mathbf{x}}_{\mathrm{hi}}}^{(n)} \ne {\mathbf{x}_{\mathrm{hi}}}^{(n)}) = \sum_{k=0}^{l_{\mathrm{D}}} \mathbb{P}({\hat{\mathbf{x}}_{\mathrm{hi}}}^{(n)} \ne {\mathbf{x}_{\mathrm{hi}}}^{(n)} | {\mathbf{x}_{\mathrm{hi}}}^{(n)} = \tilde{\mathbf{v}}_k)\mathbb{P}({\mathbf{x}_{\mathrm{hi}}}^{(n)} = \tilde{\mathbf{v}}_k)$$

Given a binary vector $\mathbf{z} \in \{0,1\}^{\text{D}}$, define the function $\psi(\mathbf{z}) := \sum_{k=1}^{\text{D}} z_k$, which denotes the count of ones in $\mathbf{z}$. Since the noisy observations are given by $c_e[n] = c[n] + e[n]$, where $e[n] = w[n] - \alpha^D w[n-1]$, it follows from assumption (A2) that $e[n] \sim \mathscr{N}(0, \sigma_1^2)$ where $\sigma_1^2 = (1+\alpha^{2\text{D}})\sigma^2$. From (2.35), we obtain $\mathbb{P}(\hat{\mathbf{x}}_{\text{hi}}{}^{(n)} \neq \mathbf{x}_{\text{hi}}{}^{(n)} | \mathbf{x}_{\text{hi}}{}^{(n)} = \tilde{\mathbf{v}}_0) = \mathbb{P}(e[n] \in \mathscr{E}_0) = Q(\alpha^{\text{D}-1}/(2\sigma_1))$. Similarly, $\mathbb{P}(\hat{\mathbf{x}}_{\text{hi}}{}^{(n)} \neq \mathbf{x}_{\text{hi}}{}^{(n)} | \mathbf{x}_{\text{hi}}{}^{(n)} = \tilde{\mathbf{v}}_{l_{\text{D}}}) = \mathbb{P}(e[n] \in \mathscr{E}_{l_D}) = Q((\tilde{\theta}_{l_{\text{D}}} - \tilde{\theta}_{l_{\text{D}}-1})/(2\sigma_1))$
$= Q(\alpha^{\text{D}-1}/(2\sigma_1))$. The last equality follows from the fact that $\tilde{\theta}_{l_{\text{D}}} - \tilde{\theta}_{l_{\text{D}}-1} = \alpha^{\text{D}-1}$. Finally, when conditioned on $\mathbf{x}_{\text{hi}}{}^{(n)} = \tilde{\mathbf{v}}_k$ for $0 < k < l_{\text{D}}$, from (2.34), we obtain $\mathbb{P}(\hat{\mathbf{x}}^{(n)} \neq \mathbf{x}_{\text{hi}}{}^{(n)} | \mathbf{x}_{\text{hi}}{}^{(n)} = \tilde{\mathbf{v}}_k) = \mathbb{P}(e[n] \in \mathscr{E}_k) = Q(\frac{\tilde{\theta}_k - \tilde{\theta}_{k-1}}{2\sigma_1}) + Q(\frac{\tilde{\theta}_{k+1} - \tilde{\theta}_k}{2\sigma_1})$. Due to Assumption (**A1**) on $\mathbf{x}_{\text{hi}}$, we have $\mathbb{P}(\mathbf{x}_{\text{hi}}{}^{(n)} = \tilde{\mathbf{v}}_k) = p^{\psi(\tilde{\mathbf{v}}_k)}(1-p)^{\text{D}-\psi(\tilde{\mathbf{v}}_k)}$. Therefore, p_e is given by

$$\begin{aligned} &p_e = Q(\alpha^{\text{D}-1}/(2\sigma_1))(1-p)^{\text{D}} + Q(\alpha^{\text{D}-1}/(2\sigma_1))p^{\text{D}} + \\ &\sum_{k=1}^{l_{\text{D}}-1} \left(Q(\frac{\tilde{\theta}_k - \tilde{\theta}_{k-1}}{2\sigma_1}) + Q(\frac{\tilde{\theta}_{k+1} - \tilde{\theta}_k}{2\sigma_1}) \right) p^{\psi(\mathbf{v}_k)}(1-p)^{\text{D}-\psi(\mathbf{v}_k)} \end{aligned} \tag{2.51}$$

The spike train $\mathbf{x}_{\text{hi}}$ is incorrectly decoded if at least one of the blocks are decoded incorrectly, hence, the total probability of error is given by:

$$\begin{aligned} &\mathbb{P}(\bigcup_{n=0}^{M-1} \hat{\mathbf{x}}^{(n)} \neq \mathbf{x}_{\text{hi}}{}^{(n)}) \leq \sum_{n=0}^{M-1} \mathbb{P}(\hat{\mathbf{x}}^{(n)} \neq \mathbf{x}_{\text{hi}}{}^{(n)}) = M p_e \\ &\overset{(a)}{\leq} 2MQ(\Delta\theta_{\min}(\alpha, \text{D})/(2\sigma_1)) \sum_{j=0}^{\text{D}} p^j (1-p)^{\text{D}-j} \binom{\text{D}}{j} \\ &\overset{(b)}{\leq} 2M \exp(-\Delta\theta_{\min}^2(\alpha, \text{D})/(4\sigma_1^2)) \end{aligned} \tag{2.52}$$

where the first inequality follows from union bound and second equality is a consequence of (2.51). The inequality (a) follows from the monotonically decreasing property of $Q(.)$ function and the sum can be re-written by grouping all terms with the same count, i.e., $\psi(\mathbf{v}_k) = j$. The inequality (b) follows from the inequality $Q(x) \leq \exp(-x^2/2)$ for $x > 0$. If the SNR condition (2.36) holds then from (2.52) the total probability of error is bounded by δ. □

2.7.6 Appendix E: Proof of Theorem 6

Proof. We first begin by showing that $\alpha \in \mathscr{F}_{\mathrm{D}}$ implies that (2.39) holds and hence the mapping of spikes with the same counts are clustered. Notice that for $k = 0$, $\theta^k_{\max} = \theta^k_{\min} = 0$. For $k \geq 1$, it is easy to verify that $\theta^k_{\max}$ and $\theta^k_{\min}$ are attained by the spiking patterns 00...1111 (with k consecutive spikes at the indices $\mathrm{D}-k+1$ to D) and 111...000 (with consecutive spikes at the indices 1 to k), which allows us to simplify (2.39) as $\alpha^{\mathrm{D}-1} > 0$ for $k = 0$ and $\sum_{i=1}^{k+1} \alpha^{\mathrm{D}-i} > \sum_{j=0}^{k-1} \alpha^j$, $k = 1, \cdots, \mathrm{D}-1$. The values of α that satisfy each of these relations can be described by the following sets:

$$\mathscr{G}_0 = \{\alpha \in (0,1) | \alpha^{\mathrm{D}-1} > 0\}, \mathscr{G}_k = \{\alpha \in (0,1) | r_k(\alpha) < 0\},$$

where $r_k(\alpha) = \alpha^{\mathrm{D}} - \alpha^{\mathrm{D}-k-1} - \alpha^k + 1$ for $1 \leq k \leq \mathrm{D}-1$. It is easy to see that $\mathscr{F}_{\mathrm{D}} = \mathscr{G}_{k_0}$. Observe that the relations are symmetric, i.e., $\mathscr{G}_k = \mathscr{G}_{\mathrm{D}-k-1}$. Furthermore, for $1 \leq k \leq \mathrm{D}/2$, we show that $\mathscr{G}_k \subseteq \mathscr{G}_{k-1}$ as follows. Trivially, $\mathscr{G}_1 \subset \mathscr{G}_0$. For $2 \leq k \leq D/2$, observe that $r_k(\alpha) - r_{k-1}(\alpha) = \alpha^{\mathrm{D}-k}(1-1/\alpha) - \alpha^k(1-1/\alpha) = (1/\alpha - 1)(\alpha^k - \alpha^{\mathrm{D}-k}) \geq 0$. Therefore, $\alpha \in \mathscr{G}_k \Rightarrow \alpha \in \mathscr{G}_{k-1}$, $k = 1, 2 \cdots, k_0$. Moreover, since $\mathscr{G}_k = \mathscr{G}_{\mathrm{D}-k-1}$, it follows that $\mathscr{F}_{\mathrm{D}} = \mathscr{G}_{k_0} = \cap_{k=0}^{\mathrm{D}-1} \mathscr{G}_k$. Hence, $\alpha \in \mathscr{F}_{\mathrm{D}} \Rightarrow \alpha \in \mathscr{G}_i$ for all $0 \leq i \leq \mathrm{D}-1$, which implies that (2.39) holds. If the noise perturbation satisfies $|w[n]| < \Delta^{\mathrm{c}}_{\min}(\alpha, \mathrm{D})/4$, it implies $|e[n]| < \Delta^{\mathrm{c}}_{\min}(\alpha, \mathrm{D})/2$. For any block $\mathbf{x}_{\mathrm{hi}}{}^{(n)} \in \mathscr{C}^{\mathrm{D}}_k$, $\theta^k_{\min} \leq \mathbf{h}_\alpha^\top \mathbf{x}_{\mathrm{hi}}{}^{(n)} \leq \theta^k_{\max}$. If $|e[n]| < \Delta^{\mathrm{c}}_{\min}(\alpha, \mathrm{D})/2$, we have

$$\begin{aligned}
\mathbf{h}_\alpha^\top \mathbf{x}_{\mathrm{hi}}{}^{(n)} + e[n] &< \theta^k_{\max} + \frac{\Delta^c_{\min}(\alpha, \mathrm{D})}{2} < \theta^k_{\max} + \frac{\theta^{k+1}_{\min} - \theta^k_{\max}}{2} \\
\mathbf{h}_\alpha^\top \mathbf{x}_{\mathrm{hi}}{}^{(n)} + e[n] &> \theta^k_{\min} - \frac{\Delta^c_{\min}(\alpha, \mathrm{D})}{2} > \theta^k_{\min} - \frac{\theta^k_{\min} - \theta^{k-1}_{\max}}{2}
\end{aligned}$$

This shows that whenever $\alpha \in \mathscr{F}_{\mathrm{D}}$, the condition $|e[n]| < \Delta^{\mathrm{c}}_{\min}(\alpha, \mathrm{D})/2$ is sufficient for (2.41) to hold $\forall$ $\gamma[n]$ and hence the spike count can be exactly recovered. □

2.7.7 Appendix F: Amplitude Estimation

We suggest a procedure to estimate the binary amplitude A, if it is unknown. We first evaluate the signal $c[n]$ from different time instants $n = 1, 2, \cdots, M-1$. For some $1 \leq n_0 \leq M-1$, we estimate a set $\mathscr{A} = \{A_k\}$ of candidate amplitudes: $A_k := c[n_0]/\mathbf{h}_\alpha^T \mathbf{v}_k$ where $\mathbf{v}_k \in \mathscr{S}_{\text{all}}$. Only a certain amplitudes can generate $c[n_0]$ from a valid binary spiking pattern $\mathbf{v}_k \in \mathscr{S}_{\text{all}}$. Our goal is to prune $\mathscr{A}$ by sequentially eliminating certain candidate amplitudes from the set based on a consistency test across the remaining measurements $c[n]$. At the t^{th} stage ($t = 2, 3, \cdots$), for every remaining candidate amplitude $A_k \in \mathscr{A}$, we perform the following consistency test with $c[n]$, to identify if a candidate amplitude can potentially generate the corresponding measurement $c[n]$. Suppose there exists a spiking pattern $\mathbf{v}_l \in \mathscr{S}_{\text{all}}$ such that

$$c[n] = A_k \mathbf{h}_\alpha^T \mathbf{v}_l \tag{2.53}$$

then A_k remains a valid candidate. If we cannot find a corresponding $\mathbf{v}_l \in \mathscr{S}_{\text{all}}$ for an amplitude A_k, we remove it, $\mathscr{A} = \mathscr{A} \setminus A_k$. In presence of noise, (2.53) can be modified to allow a tolerance γ as we may not find an exact match. The tolerance γ is chosen to be 0.5 in the experiments on the GENIE dataset. This procedure prunes out possible values for the amplitude by leveraging the shared amplitude across multiple measurements $c[n]$.

Chapter 3

Measurement-Algorithm Co-Design: Exact Recovery of Binary Signals from finite-memory filtered measurements

In this chapter, we continue to investigate the problem of recovering a binary-valued signal from compressed measurements of its convolution with a known finite impulse response filter. However, the compression matrix is no longer restricted to be a uniform sub-sampling operator. We show that it is possible to attain optimum sample complexity for exact recovery (in absence of noise) with a computationally efficient algorithm. We achieve this by adopting an algorithm-measurement co-design strategy where the measurement matrix is designed as a function of the filter, such that the recovery of binary signals with arbitrary sparsity is possible using sequential decoding. Such a filter-dependent sampler design overcomes the computational challenges associated with enforcing binary constraints, and enables us to operate in "extreme compression" regimes, where the number of measurements can be much smaller than the sparsity level ($M < s$).

3.1 Prior Works

In this chapter, we re-visit the special class of the binary compressed sensing problem, where we observe compressive measurements of the convolution of a binary signal with a known finite impulse response (FIR) filter. Recovering binary signals from such compressive convolutional measurements is of interest to several applications such as neural spike detection from fluorescence measurements [9], medical imaging [63], binary shape recovery from blurred images [48, 64], image segmentation [65], and discrete tomography [14, 15]. A concrete application is in millimeter-wave communication, where the goal is to decode binary (or finite alphabet) symbols from low-dimensional measurements obtained by a compressive spatial filtering/beamforming.

As discussed in [49, 50], there exist measurement matrices such that the linear mapping between the unknown binary vectors and the real valued compressive measurement is injective even with a single (scalar) measurement. In this case, the desired binary vector can be recovered via exhaustive search, however, it is computationally prohibitive to do so. Therefore, a major focus of binary compressed sensing has been on algorithmic developments (often via relaxations) that are computationally efficient, and at the same time, exploit additional structure such as sparsity of the binary signal [41, 42, 44, 45, 66, 67]. A common approach is to relax the (non-convex) binary constraints with box-constraints, and formulate various continuous valued optimization problems for recovering the binary vectors. These include l_1/l_∞-norm minimization [41, 42, 68], and semidefinite relaxation [44]. Theoretical guarantees for l_1/l_∞ minimization have been established in [41, 42, 68], but the results are mostly applicable to random measurement matrices (drawn from suitable centered distributions). Recently, the benefit of using a biased measurement matrix was established in [43], which allows recovery of binary signals simply by solving a least squares problem with box-constraints, and thereby eliminating the need for l_1/l_∞ minimization. Alternative lines of work that modify classical sparse recovery algorithms to exploit finite-valued structure include greedy orthogonal matching pursuit (OMP) algorithm [66, 69], Bayesian formulations [45, 70], graph-based decoding techniques [67], and iterative reweighting techniques [71]. A common feature of all the aforementioned approaches is that their theoretical guarantees (whenever they exist) are applicable when the number of measurements (M) is larger than the sparsity (s), similar to standard results in compressed sensing. To the best of our knowledge, exact recovery guarantees for these techniques are unavailable when $M < s < N/2$ (where N is the signal dimension).

In a recent work [46], we moved away from relaxation-based techniques and showed that it is possible to exactly recover binary signals from *uniformly downsampled* measurements of the filter output, without imposing any sparsity constraints. Specifically, we developed a new computationally efficient decoding algorithm that was inspired by successive cancellation (SC) or decision feedback decoding [70, 72] used in multiuser detection, and decoding of polar codes [73, 74]. We showed that $M > N/L$ (L being the filter length) measurements are necessary for exact recovery of any binary vector from *uniformly downsampled* convolutional measurements, and the algorithm was able to attain this under a certain decay condition on the filter.

Summary of contributions: We establish that by appropriately designing the measurement matrix (beyond uniform downsamplers), it is possible to achieve a sample complexity of $M \geq 1$ for the exact recovery of binary signals.[1] We achieve this by (i) developing a modified version of the sequential decoding algorithm from [46], and (ii) proposing compressive measurement design techniques that are *dependent on the filter*. This algorithm-measurement co-design strategy achieves the optimal sample complexity of $M \geq 1$ (for any sparsity level), without requiring any strong decay assumptions on the filter. The measurement matrix itself can be designed in a computationally efficient manner, by solving a linear program. **Notations:** For a matrix $\mathbf{A}$, $\mathscr{N}(\mathbf{A})$ denotes its null-space, and $\mathbf{I}_N$ is the $N \times N$ identity matrix. For an integer n, define $[n] := \{1, 2, \cdots, n\}$.

[1] In [44], it was noted that $M = 1$ may be achievable provided the SDP returned a rank 1 solution. However, conditions under which the SDP solution is guaranteed to be rank one with $M = 1$ measurement, are currently unavailable.

3.2 Motivation for Measurement-Design

Recall the convolutional compressive measurement model (introduced in Chapter 2):

$$\mathbf{z} = \Phi \mathbf{H} \mathbf{x}_0 \tag{3.1}$$

where $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_N] \in \mathbb{R}^{P \times N}$ is a Toeplitz matrix with $P = N + L - 1$, L is the filter length, $\Phi \in \mathbb{R}^{M \times P}, M < P$ is a compressive measurement matrix and $\mathbf{H}\mathbf{x}_0$ is the output of the filter. In Chapter 2, we constrained $\Phi \in \{0,1\}^{M \times P}$ to be a structured binary matrix, which models the uniform downsampling of the filter output. In Theorem 1, we showed that for almost all filters h, it is possible to exactly recover any binary signal from the uniformly downsampled measurements of the filter output, without imposing any sparsity constraints, provided $M \geq N/L$. This bring us to the question "What are the benefits of designing the measurement matrix Φ (without being restricted to uniform downsampling)?"

We motivate the need for measurement design through an illustrative example. Consider the FIR filter to be a moving average filter with $L = 5$, i.e., $h = [1,1,1,1,1]^T$ and $N = 15$. In Figure 3.1, we consider two different binary ground truth signals (top row) and plot the output of the convolution with the moving average filter (bottom row). The final observation after downsampling by a factor of 5 is shown in Figure 3.1 (middle row). From Figure 3.1, it is evident that two distinct binary signals can result in identical under-sampled measurements. Hence, this moving average filter h is an example of an adversarial filter which does not permit uniformly downsampling the output of the convolution. Note that this example does not contradict our previous result Theorem 1, since the event of sampling such a FIR filter $h = [1,1,1,1,1]^T$ has zero probability under the probabilistic model considered in Theorem 1 . By moving away from uniform downsampling and instead adopting a dense compression matrix Φ, we can overcome the challenge posed by such an adversarial filter. This simple example motivates exploring the notion of a filter-dependent measurement design.

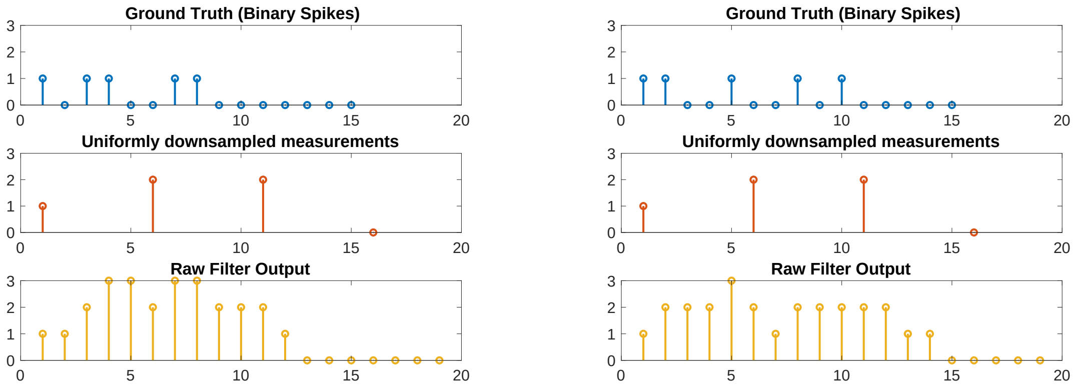


Figure 3.1. Example of an adversarial filter that results in ambiguous binary solutions with uniform sub-sampling operation. (Top) Ground Truth binary signal (Middle) Measurements after under-sampling by a factor of 5 (Bottom) Output of convolution of the ground truth binary signal with a moving average filter of length $L = 5$.

3.3 Measurement-Algorithm Co-Design

It has been shown that a linear map $\mathbf{A} : \{0,1\}^N \rightarrow \mathbb{R}^M$ can be injective (over $\{0,1\}^N$) even when $M = 1$ [49,50]. [2] The linear map of interest to us has a specific structure $\mathbf{A} = \Phi\mathbf{H}$. We begin by showing that for any filter $\mathbf{h}$, there exist infinite choices of real-valued sensing matrices $\Phi \in \mathbb{R}^{M\times P}$ such that the map $\mathbf{A}$ is injective for every $M \geq 1$.

Theorem 7. *Assume rank*$(\mathbf{H}) = N$. *Let* $\Phi \in \mathbb{R}^{M\times P}$ *be a random matrix whose rows* $\{\phi_m\}_{m=1}^M$ *are drawn independently from a distribution which is absolutely continuous with respect to the Lebesgue measure over* $\mathbb{R}^P$. *With probability* 1, $\mathbf{x}_0$ *is the unique binary vector that satisfies* $\mathbf{z} = \Phi\mathbf{H}\mathbf{x}$ *for every* $M \geq 1$, *where* $\mathbf{z}$ *is given by* (2.3).

Proof. Suppose there exist $\mathbf{x},\mathbf{y} \in \{0,1\}^N (\mathbf{x} \neq \mathbf{y})$ such that $\Phi\mathbf{H}\mathbf{x} = \Phi\mathbf{H}\mathbf{y} \Rightarrow \Phi\mathbf{H}(\mathbf{x}-\mathbf{y}) = \mathbf{0}$. This means that there is a non-zero ternary vector $\mathbf{x}-\mathbf{y} \in \mathscr{S}^N$, $\mathbb{S} := \{-1,0,1\}$, that belongs to the null space of $\Phi\mathbf{H}$. We will show that this will happen with zero probability. Notice that the cardinality of $\mathscr{S}^N$ is 3^N. We denote each vector in $\mathscr{S}^N$ as $\{\mathbf{v}_k\}_{k=0}^{3^N-1}$, with the convention $\mathbf{v}_0 = \mathbf{0}$

[2] In [50], it is shown that $\mathbf{A}$ can be linearly dependent over $\mathbb{R}$, but linearly independent over $\{0,1\}$, and [49] shows the existence of such a rational $\mathbf{A}$.

for notational ease. Let $\mathscr{E} = \left\{\Phi \mid \mathscr{N}(\Phi\mathbf{H})\bigcap\mathscr{A}^N \neq \{\mathbf{0}\}\right\}$. Then,

$$\begin{aligned}\mathbb{P}(\Phi \in \mathscr{E}) &= \mathbb{P}\big(\exists\, \mathbf{v} \in \mathscr{A}^N \backslash \{\mathbf{0}\},\ \text{s.t. } \Phi\mathbf{H}\mathbf{v} = \mathbf{0}\big)\\ &= \mathbb{P}\Big(\bigcup_{k=1}^{3^N-1} \{\Phi\mathbf{H}\mathbf{v}_k = \mathbf{0}\}\Big) \underset{(a)}{=} \mathbb{P}\Big(\bigcup_{k=1}^{3^N-1}\bigcap_{i=1}^{M}\{\phi_i \in \mathscr{N}(\mathbf{v}_k^T\mathbf{H}^T)\}\Big)\\ &\underset{(b)}{\leq} \sum_{k=1}^{3^N-1}\prod_{i=1}^{M}\mathbb{P}\left(\phi_i \in \mathscr{N}(\mathbf{v}_k^T\mathbf{H}^T)\right) \qquad (3.2)\end{aligned}$$

where (a) follows from the fact that $\Phi\mathbf{H}\mathbf{v}_k = \mathbf{0}$ if and only if $\phi_i \in \mathscr{N}(\mathbf{v}_k^T\mathbf{H}^T)$ for all $i \in [M]$. The inequality (b) follows from union bound, and the independence assumption on the rows of Φ. Rank$(\mathbf{H}) = N$ implies that $\mathbf{v}_k^T\mathbf{H}^T \neq \mathbf{0}$ for non-zero $\mathbf{v}_k$, and therefore $\mathscr{N}(\mathbf{v}_k^T\mathbf{H}^T)$ is a $P-1$ dimensional subspace of $\mathbb{R}^P$ whose Lebesgue measure is zero. Since ϕ_i is generated from a distribution that is absolutely continuous with respect to the Lebesgue measure over $\mathbb{R}^P$, we have $\mathbb{P}(\phi_i \in \mathscr{N}(\mathbf{v}_k^T\mathbf{H}^T)) = 0$. Using (3.2), we can conclude that $\mathbb{P}(\Phi \in \mathscr{E}) = 0$. □

Relaxation-based techniques succeed in a regime where M is larger than the sparsity of $\mathbf{x}_0$, and exact recovery may not be possible with $M = \Omega(1)$ measurements. We now present a simple and computationally efficient algorithm that sequentially decodes the binary entries of $\mathbf{x}_0$. We further show that by using the idea of *filter-dependent* sampler design, it is possible to achieve $M = \Omega(1)$ with this algorithm.

3.3.1 Sequential Block-Wise Decoding and Performance Guarantees

The proposed Sequential Block-wise Decoding Algorithm is summarized as Algorithm 4. The main idea is to partition the entries of $\mathbf{x}_0$ into $b = \lceil N/M \rceil$ disjoint blocks, one corresponding to each scalar measurement z_m, and decode the entries of a block sequentially. For the m^{th} block, suppose that the first $k < b$ indices within the block, denoted by the set $\mathscr{J}_{m,k} = \left\{(m-1)b+i\right\}_{i=1}^{k}$ have already been decoded. The sequential decoding algorithm computes a residual $r = z_m - \sum_{i\in\mathscr{J}_{m,k}} A_{m,i}\hat{x}_i$, and compares it against a suitable threshold determined by $\varepsilon(\geq 0)$, in order to estimate the $(k+1)^{\text{th}}$ element. The estimate is given as $\hat{x}_{(m-1)b+k+1} = \mathbb{1}_{\{r/A_{m,(m-1)b+k+1} \geq (1-\varepsilon)\}}$

Algorithm 4. Sequential Block-wise Decoding Algorithm

```
Input: Measurement $\mathbf{z}$, Sensing matrix $\Phi \in \mathbb{R}^{M\times P}$, Filter $\mathbf{H}$, Tolerance $\varepsilon \geq 0$.
Output: $\hat{\mathbf{x}} \in \{0,1\}^N$ //Estimate of $\mathbf{x}_0$
$\mathbf{A} \leftarrow \Phi\mathbf{H}, b \leftarrow \lceil N/M \rceil, m \leftarrow 1$, $\hat{\mathbf{x}} \leftarrow \mathbf{0}$ //Initialization
Repeat
    $l \leftarrow 1$, $r \leftarrow z_m$//Reset Residual
    Repeat
        if $r/A_{m,(m-1)b+l} \geq 1-\varepsilon$ //Detection threshold
            $\hat{x}_{(m-1)b+l} \leftarrow 1$
        end
        $r \leftarrow r - \hat{x}_{(m-1)b+l}A_{m,(m-1)b+l}$, $l \leftarrow l+1$ //Update residual
    until $l \leq b$ or $(m-1)b+l \leq N$
    $m \leftarrow m+1$
until $m \leq M$
```

where $\mathbb{1}_{\{x\geq\gamma\}} : \mathbb{R} \to \{0,1\}$ denotes an indicator function defined as

$$\mathbb{1}_{\{x\geq\gamma\}} = \begin{cases} 1, & \text{if } x \geq \gamma \\ 0, & \text{else} \end{cases} .$$

It is important to note that the residual computation subtracts only the elements that have been decoded within the *current block*. The previous blocks that have already been decoded, are not subtracted out. This algorithmic choice has been made to avoid error propagation between blocks. Such disjoint decoding can be especially beneficial in presence of noise. Decoding each block requires $O(b)$ operations, resulting in a total computational complexity of $O(Mb) = O(N)$ for decoding all M blocks.

Given integers $m \in [M]$, and $l \in [b]$ we define $\eta(m,l) := (m-1)b+l$. For ease of exposition, we assume that N/M is an integer. The following theorem specifies sufficient conditions on $\mathbf{A} = \Phi\mathbf{H}$ under which Algorithm 4 exactly recovers $\mathbf{x}_0$ from noiseless measurements (2.3).

Theorem 8. *Let* $\mathbf{A} := \Phi\mathbf{H}$ *and* $\varepsilon = 0$. *For any* $M \geq 1$, *Algorithm 4 recovers* $\mathbf{x}_0$, *if for each* $m \in [M]$, *the following holds*

$$A_{m,l} > 0, \forall\, l \in [N],\ A_{m,\eta(m,l)} > \sum_{\substack{k=1 \\ k\notin \mathscr{J}_{m,l}}}^{N} A_{m,k}, \forall\, l \in [b-1] \tag{3.3}$$

Proof. Condition (3.3) implies that for any binary $\mathbf{x} \in \{0,1\}^N$ and $l \in [b-1]$,

$$x_{\eta(m,l)} \leq \sum_{k \notin \mathscr{I}_{m,l}} x_k \frac{A_{m,k}}{A_{m,\eta(m,l)}} + x_{\eta(m,l)} < 1 + x_{\eta(m,l)} \tag{3.4}$$

We now show that for every m, we can decode the indices of $\mathbf{x}_0$ given by $\{\eta(m,l)\}_{l=1}^{b}$. Fix m. Our proof proceeds via induction on l. For $l = 1$, we have $r = z_m = \sum_{k \notin \mathscr{I}_{m,1}} x_k A_{m,k} + x_{\eta(m,1)} A_{m,\eta(m,1)}$ (Line 5 of Algorithm 4). Hence from (3.4) we have $x_{\eta(m,1)} \leq r/A_{m,\eta(m,1)} < x_{\eta(m,1)} + 1$. Since $\hat{x}_{\eta(m,1)} = \mathbb{1}_{\{r/A_{m,\eta(m,1)} \geq 1\}}$, it follows that $\hat{x}_{\eta(m,1)} = 1$ if $x_{\eta(m,1)} = 1$, and 0 otherwise, implying $\hat{x}_{\eta(m,1)} = x_{\eta(m,1)}$. Next assume that after $l-1 < b$ iterations we have correctly decoded $\{x_{\eta(m,k)}\}_{k=1}^{l-1}$. The residual satisfies:

$$\begin{aligned} r/A_{m,\eta(m,l)} &= \big(z_m - \sum_{k \in \mathscr{I}_{m,l-1}} \hat{x}_k A_{m,k}\big)/A_{m,\eta(m,l)} \\ &\overset{(a)}{=} \sum_{k \notin \mathscr{I}_{m,l}} x_k \frac{A_{m,k}}{A_{m,\eta(m,l)}} + x_{\eta(m,l)} \end{aligned} \tag{3.5}$$

where (a) holds due to the induction hypothesis. Using a similar argument as $l = 1$, from (3.4), (3.5) we can again show that $\hat{x}_{\eta(m,l)} = \mathbb{1}_{\{r/A_{m,\eta(m,l)} \geq 1\}} = x_{\eta(m,l)}$, which concludes the proof. □

The success of Algorithm 4 therefore depends on condition (3.3), which reveals the dependence of the sampler Φ on the filter $\mathbf{h}$, and implicitly governs the sample complexity. If the entries of Φ are drawn randomly, agnostic to $\mathbf{h}$, the condition (3.3) may not be satisfied with high probability. Therefore, it becomes essential to explicitly tune the design of sampler Φ to the structure of the filter $\mathbf{h}$.

3.3.2 Filter-Dependent Sampler Design via Linear Program

It can be verified that condition in (3.3) is satisfied if and only if for every $m \in [M]$, ϕ_m belongs to the following set

$$\begin{aligned}\mathscr{F}_{\mathbf{h}}^{(m)} =& \{\phi \in \mathbb{R}^P \mid \phi^T \mathbf{h}_i > 0, \quad i = 1,2,\cdots,N,\\ & \phi^T (\mathbf{h}_{(m-1)b+j} - \sum_{\substack{k=1\\ k \notin \mathscr{J}_{m,j}}}^{N} \mathbf{h}_k) > 0, 1 \leq j \leq b-1\}\end{aligned}$$

Notice that $\mathscr{F}_{\mathbf{h}}^{(m)}$ is a polyhedral set, whose geometry depends on the choice of the filter $\mathbf{h}$. Hence the sampling operators ϕ_m can be designed to satisfy (3.3), by solving the following linear program for every m:

$$\text{find} \quad \phi_m \quad \text{subject to} \quad \phi_m \in \mathscr{F}_{\mathbf{h}}^{(m)} \tag{LPH}$$

For any $\phi \in \mathscr{F}_{\mathbf{h}}^{(m)}$, the scaled vector $\alpha\phi$ for any $\alpha > 0$ is also a valid solution, i.e., $\alpha\phi \in \mathscr{F}_{\mathbf{h}}^{(m)}$. Therefore, the sensing matrix can always be scaled to avoid solutions close to $\mathbf{0}$, as well as meet any desired power constraint. The following lemma, whose proof is in the Appendix, ensures that $\mathscr{F}_{\mathbf{h}}^{(m)}$ is non-empty under mild conditions on $\mathbf{h}$.

Lemma 10. *For any $\mathbf{h} \in \mathbb{R}^L$ satisfying rank$(\mathbf{H}) = N$ and $M \geq 1$, the set $\mathscr{F}_{\mathbf{h}}^{(m)}$ is non-empty for every $m \in [M]$.*

We obtain the following exact recovery guarantee for Algorithm 4 by combining Theorem 8 and Lemma 10.

Theorem 9. *Let $\Phi^\star$ be a sensing matrix whose mth row is a solution to* (LPH)*, $m \in [M]$. Consider noiseless measurements $\mathbf{z} \in \mathbb{R}^M$ acquired using $\Phi^\star$ as $\mathbf{z} = \Phi^* \mathbf{H} \mathbf{x}_0$, where $\mathbf{x}_0 \in \{0,1\}^N$. For every $M \geq 1$, Algorithm 4 recovers $\mathbf{x}_0$, regardless of its sparsity.*

3.3.3 Remarks on Noise Resilience and Sampler Design

The main objective of this chapter was to achieve optimum sample complexity for exact recovery of binary signals in absence of noise with a computationally efficient algorithm. In presence of noise, the threshold ε in Algorithm 4 should be optimized based on the noise level. Increasing M increases the number of blocks which is also important to promote noise resilience, since Algorithm 4 prevents error propagation from one block to another. Another important, but perhaps less obvious consideration is the effect of block length on the dynamic range of the sampler. By decreasing b (increasing M), measurement matrices with smaller dynamic range can be designed, which leads to better numerical stability and more reliable decoding. Determining the optimal choices of b and ε based on the noise level and dynamic range considerations will be of future interest. Other directions will be to design Φ by using a suitable optimization criterion over the set $\mathscr{F}_{\mathbf{h}}^{(m)}$ (instead of a simple feasibility search), and explore adaptive filter-dependent sensing strategies.

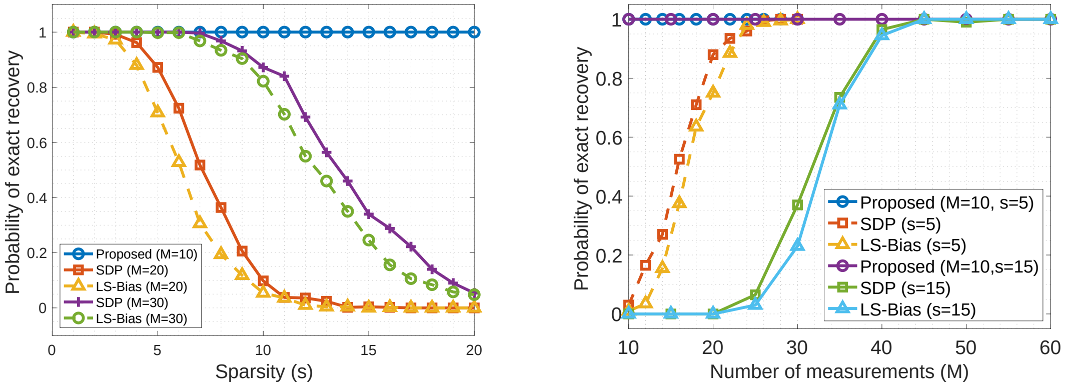


Figure 3.2. Noiseless recovery performance of different methods (Left) versus s for different M, and (Right) versus M for $s = 5, 15$.

3.4 Simulations

We consider binary signals of dimension $N = 100$, and FIR filters of length $L = 5$. The filter coefficients are generated independently as product of two independent random variables $h_i = s_i d_i$ where $d_i \sim \mathscr{U}[1,2]$, and s_i is a Rademacher random variable, and these coefficients are kept fixed throughout the experiments. We compare the performance of Algorithm 4 and the filter-dependent sampler design strategy (LPH) against two recent binary compressed sensing algorithms (i) SDP relaxation [44] and (ii) box-constrained least squares with biased measurement (LS-Bias) [43]. We generate noiseless measurements of the form $z = \mathbf{A}x_0$. For our approach, $\mathbf{A} = \Phi\mathbf{H}$ where Φ is obtained by solving (LPH). For SDP relaxation, the entries of $\mathbf{A}$ are generated i.i.d as $A_{i,j} \sim \mathscr{N}(0,1)$. For LS-Bias, following [43], the entries of $\mathbf{A}$ are generated i.i.d as $A_{i,j} \sim \mathscr{N}(1,1)$, where the non-zero mean of 1 acts as the bias. In the first experiment, we study the noiseless performance of each technique as a function of sparsity s. The probability of exact recovery, i.e., number of times $\hat{x} = x_0$ over 100 Monte Carlo runs, is used as the performance metric. Figure 3.2 shows that the proposed strategy exactly recovers x_0 regardless of the sparsity level (even when $M < s$), whereas SDP relaxation and LS-Bias, the probability of exact recovery falls below 0.5 when s exceed $M/2$.

In the next experiment, we study the performance of SDP relaxation and LS-Bias as a function of M, keeping the sparsity fixed at $s = 5$ and 15. For the proposed strategy, we keep M fixed at $M = 10$ ($b = 10$). Figure 3.2 (b) shows SDP and LS-Bias require $M = 25$ (for $s = 5$) and $M = 45$ (for $s = 15$) to exactly recover x_0 whereas the proposed strategy is able to do so with only $M = 10$ filter-dependent measurements.

Next, we evaluate the performance of the proposed strategy in presence of noise. We generate noisy measurements of the form $\mathbf{z} = \mathbf{A}x_0 + \mathbf{n}$, where the additive noise $\mathbf{n}$ is distributed as $\mathbf{n} \sim \mathscr{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_M)$. The signal-to-noise ratio (SNR) for each sensing strategy is defined as $10\log_{10}\left(\frac{\|\mathbf{A}x_0\|_2^2}{M\sigma_n^2}\right)$. To ensure a consistent SNR across different approaches, we normalize the measurement matrices (in our case, we normalize Φ) such that $\|\mathbf{A}x_0\|_2$ is the same for each

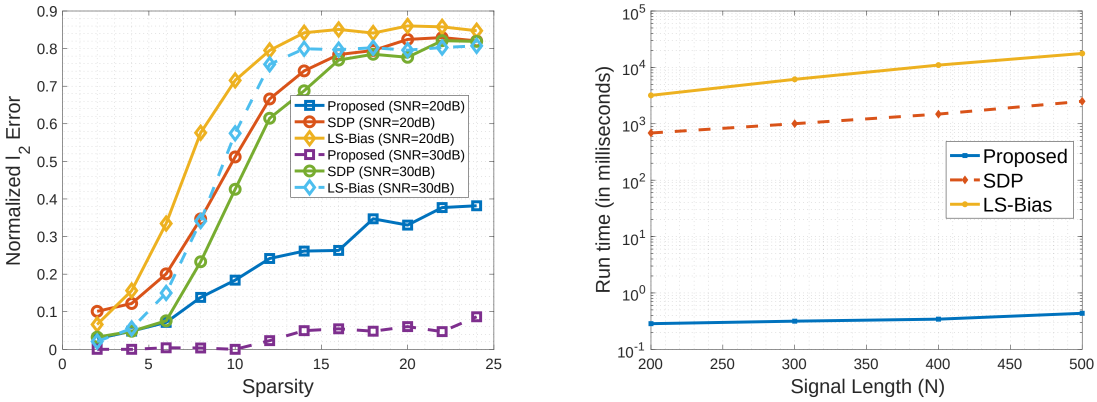


Figure 3.3. (Left) Noisy Reconstruction: Normalized l_2 error vs. sparsity s, for different SNR. Here $M = 25$ and threshold is fixed at $\varepsilon = 0.1$. (Right) Comparison of runtime versus N ($M = \lceil 0.2N \rceil$, $s = 5$)

method, and σ_n is chosen according to the desired SNR. In Figure 3.3 (a), we plot the l_2 error averaged over 100 Monte Carlo runs for each sparsity level. The proposed algorithm achieves a significantly smaller error especially when the sparsity increases. We operate in the regime $s < M$ to ensure sufficient measurements for all algorithms.

Finally, in Figure 3.3 (b) we compare the average run-time (averaged over 10 runs) of all three algorithms as a function of N. We choose $s = 5$, and $M = \lceil 0.2N \rceil$. The run-time of Algorithm 4 is significantly smaller than the others, which were implemented using off-the-shelf convex solver (CVX) [75].

3.5 Conclusion

We proposed a measurement matrix design framework for recovery of binary-valued signals from compressed convolutional measurements. The filter-dependent sensing matrix design guarantees exact recovery in absence of noise, using a computationally efficient sequential block-wise decoding algorithm. The overall strategy achieves an optimal sample complexity of $M \geq 1$. The proposed framework also paves way for several interesting future directions such as optimizing the algorithm and measurement design parameters using the knowledge of noise level, and extending the strategy to more general alphabets and different classes of filters.

Chapter 3, in part, is a reprint of the material as it appears in P. Sarangi and P. Pal, "Measurement Matrix Design for Sample-Efficient Binary Compressed Sensing," in IEEE Signal Processing Letters, vol. 29, pp. 1307-1311, 2022.

The dissertation author was the primary investigator and author of these papers.

3.6 Appendix

3.6.1 Proof of Lemma 10

Proof. We will first establish that if $\mathscr{F}_h^{(1)}$ is non-empty then $\mathscr{F}_h^{(m)}$ is also non-empty for every $m \in [M]$. For each m, we define a permutation matrix $\Pi_m \in \mathbb{R}^{N \times N}$ as follows:

$$[\Pi_m \mathbf{x}]_j = \begin{cases} x_{(m-1)b+j}, & 1 \leq j \leq b \\ x_{j-(m-1)b}, & (m-1)b+1 \leq j \leq mb \\ x_j, & \text{otherwise} \end{cases}$$

This permutation swaps the first and m^{th} block (of size b) of the vector $\mathbf{x}$. The set $\mathscr{F}_h^{(m)}$ is described by $N+b-1$ inequalities, which can be compactly represented as $\mathbf{B}\Pi_m\mathbf{H}^T\phi \succ \mathbf{0}$. Here, $\succ$ denotes element-wise inequality constraints, and $\mathbf{B} = [\mathbf{I}_N, \tilde{\mathbf{B}}^T]^T$ is a $(N+b-1) \times N$ matrix, with $\tilde{\mathbf{B}}_{i,j} = 1$ if $i = j$, $\tilde{\mathbf{B}}_{i,j} = -1$ if $i < j$, and 0 otherwise. If $\mathscr{F}_h^{(1)}$ is non-empty, then $\exists\ \phi_1$ such that $\mathbf{B}\mathbf{H}^T\phi_1 \succ \mathbf{0}$ (since $\Pi_1 = \mathbf{I}_N$). Since $\text{rank}(\mathbf{H}) = N$, we can always find $\tilde{\phi} \in \mathbb{R}^P$ satisfying $\mathbf{H}^T\tilde{\phi} = \Pi_m^T\mathbf{H}^T\phi_1$. Such a vector $\tilde{\phi}$ also satisfies $\mathbf{B}\Pi_m\mathbf{H}^T\tilde{\phi} = \mathbf{B}\Pi_m\Pi_m^T\mathbf{H}^T\phi_1 \overset{(a)}{=} \mathbf{B}\mathbf{H}^T\phi_1 \succ \mathbf{0}$, since Π_m is a permutation matrix with $\Pi_m\Pi_m^T = \mathbf{I}_N$. Therefore, $\tilde{\phi} \in \mathscr{F}_h^{(m)}$, whenever $\mathscr{F}_h^{(1)}$ is non-empty.

We now establish that $\mathscr{F}_h^{(1)}$ is indeed non-empty, i.e., $\exists\ \phi \in \mathbb{R}^P$, such that:

$$\phi^T\mathbf{h}_i > 0\ \forall\ i, \quad \phi^T(\mathbf{h}_j - \sum_{k=j+1}^{N}\mathbf{h}_k) > 0, j \in [b-1] \tag{3.6}$$

Define $\mathbf{u}_j := \mathbf{h}_j - \sum_{k=j+1}^{N} \mathbf{h}_k, j \in [b-1]$. Let $\mathscr{S} = \{\mathbf{h}_1, \mathbf{h}_2 \cdots, \mathbf{h}_N, \mathbf{u}_1, \mathbf{u}_2 \cdots, \mathbf{u}_{b-1}\}$, and consider its convex hull $\mathscr{A}_{\mathbf{H}} := \text{conv}(\mathscr{S})$ which is a (closed) polyhedral set. Observe that there exists a $\phi \in \mathbb{R}^P$ satisfying (3.6) if there exists a hyperplane $\phi^T \mathbf{x} = c$ ($c > 0$), which strictly separates the point $\mathbf{0}$ from $\mathscr{A}_{\mathbf{H}}$, i.e., $\phi^T \mathbf{x} > 0$ for all $\mathbf{x} \in \mathscr{A}_{\mathbf{H}}$. Since $\mathscr{A}_{\mathbf{H}}$ is a closed convex set, the strict hyperplane separation theorem will guarantee existence of the desired ϕ provided $\mathbf{0} \notin \mathscr{A}_{\mathbf{H}}$ [76, Prop 1.5.3]. We show $\mathbf{0} \notin \mathscr{A}_{\mathbf{H}}$ by contradiction. Suppose $\mathbf{0} \in \mathscr{A}_{\mathbf{H}}$. Then $\exists\ \alpha_i, \beta_j \geq 0$ satisfying $\sum_{i=1}^{N} \alpha_i + \sum_{j=1}^{b-1} \beta_j = 1$, and $\sum_{i=1}^{N} \alpha_i \mathbf{h}_i + \sum_{j=1}^{b-1} \beta_j \mathbf{u}_j = \mathbf{0}$ which can rearranged as $(\alpha_1 + \beta_1)\mathbf{h}_1 + \sum_{i=2}^{b-1} (\alpha_i + \beta_i - \sum_{j=1}^{i-1} \beta_j)\mathbf{h}_i + \sum_{i=b}^{N} (\alpha_i - \sum_{j=1}^{b-1} \beta_j)\mathbf{h}_i = \mathbf{0}$. Since $\text{rank}(\mathbf{H}) = N$, we must have

$$\begin{aligned} \alpha_1 + \beta_1 = 0, (\alpha_i + \beta_i - \sum_{j=1}^{i-1} \beta_j) = 0, \quad 2 \leq i \leq b-1, \\ (\alpha_i - \sum_{j=1}^{b-1} \beta_j) = 0, \quad b \leq i \leq N. \end{aligned} \tag{3.7}$$

Since α_i, β_i are also non-negative, it can be easily verified that (3.7) holds only if $\alpha_i = 0, \beta_j = 0$ for all i, j. This contradicts the fact that $\sum_{i=1}^{N} \alpha_i + \sum_{j=1}^{b-1} \beta_j = 1$. Hence $\mathbf{0} \notin \mathscr{A}_{\mathbf{H}}$, completing the proof.

□

Chapter 4

Harnessing The Benefits of Sparse Arrays in Sample-Starved Regime

In this chapter, we continue our investigation of sensing under stringent sampling budgets where the data exhibits spatio-temporal structure. A prominent example of such spatio-temporal data arises in array signal processing, where we sense the environment by placing multiple sensors/antennas in a certain geometry, and collect measurements over time. In recent times, antenna arrays with sparse geometry are being employed in automotive radars for enabling advanced driving assistance features [77–79]. They also feature in Massive MIMO communication systems for enabling low-complexity architectures that can form highly reliable directional links by hybrid beamforming at mmWave frequencies [80–83]. One of the most desirable properties of such antenna arrays is achieving higher resolution capabilities, without using additional sensing elements. A key question, therefore, is to design sparse array geometries (by non-uniform placement of sensors) that can provably enhance resolution using significantly fewer sensors, while also incurring minimal temporal sampling overhead.

Sparse array geometries such as nested arrays [84], and coprime arrays [85] have shown significant performance benefits over uniform linear arrays (ULAs) in terms of their ability to identify $K = O(P^2)$ uncorrelated sources with only P sensors [2, 84–90], their reduced Cramér-Rao lower bound (CRB) and higher resolution [86, 91, 92]. The enhanced spatial degrees of freedom (DOF) of sparse array is attributed to their large difference coarray, which can be of size

$\Theta(P^2)$. In the passive setting, these benefits are harnessed by computing the correlation between the received signal at different sensors. This has led to the belief that sparse arrays require a large number of temporal measurements to reliably estimate parameters of interest (DOAs) from these correlations, and therefore they may not be suitable in the sample-starved regime. Sample-starved regimes are typically encountered in automotive radar and joint communication and radar sensing applications, where the environment can be highly dynamic or the source signals may be coherent. As a result, the number of snapshots available for DOA estimation can be very scarce, in the extreme case, only one snapshot might be available. Therefore, there is a need for developing techniques that can harness the resolution benefits of sparse arrays while operating in the sample-starved regime.

In this chapter, we propose a new framework for leveraging the degrees of freedom in the difference set of a nested array in two different snapshot-limited regimes, where (i) few snapshots (of the order of the number of sources) are available and (ii) only a single snapshot (extreme-sample starved regime).

4.1 Prior Works

The enhanced DOF of sparse arrays is realized by obtaining an unbiased estimate of the *coarray covariance matrix* [84, 86, 93, 94] with a finite number of temporal snapshots. It becomes possible to directly apply subspace algorithms such as MUSIC, either on the contiguous ULA segment of the coarray [84–86, 93] or on an interpolated coarray [94–97]. Alternatively, estimators based on suitable regularization techniques that employ sparsity promoting l_1 norm [1, 2, 87, 98, 99], atomic norm [97, 100, 101], nuclear norm [102], and positive semidefinite constraints [89] have also been developed. However, these methods often require careful tuning of the regularization parameter. Furthermore, atomic norm-based methods require a strict separation condition on the source locations to ensure exact recovery in absence of noise [30, 35].

Coarray-based techniques typically require a large number (L) of temporal snapshots to accurately estimate the coarray covariance matrix. Hence, existing performance guarantees for coarray MUSIC-type algorithms on sparse arrays are primarily asymptotic in L [86, 103]. When the number of snapshots is limited, the performance of coarray based algorithms degrades significantly. It has been proved that when $K \geq P$, the CRB of sparse arrays saturates away from zero as SNR$\rightarrow \infty$ [86, 91]. It is also empirically observed (but not theoretically proven) that the MSE of coarray MUSIC exhibits a similar saturation when $1 < K < P$ [86, 88, 91]. [1] This raises the question: "Is it possible to harness the benefits offered by the difference coarray of nested arrays with a limited number of snapshots ($L \geq K$), while also avoiding saturation when $K < P$?"

Summary of Contributions We first prove that even for $K = 2$ sources and in absence of noise, coarray MUSIC will fail to recover the coarray subspace of nested arrays, unless the source signals are temporally orthogonal. In order to remedy this weakness of coarray MUSIC, we propose an alternative method (based on convex optimization) for coarray subspace estimation, called "Proxy Covariance" (Prox-Cov) estimation that *does not aim to estimate the true coarray covariance matrix, but only the coarray subspace.* We prove that (Prox-Cov) can exactly recover the coarray subspace in absence of noise with very few snapshots when there are fewer sources than sensors, thereby overcoming the saturation effect exhibited by coarray MUSIC in this regime. Our numerical results demonstrate the superior performance of (Prox-Cov) with limited snapshots when the number of sources exceeds the number of sensors, as well as when it is fewer.

4.1.1 Signal Model and Problem Formulation

Consider K far-field narrowband sources (with wavelength λ) impinging from directions $\theta_1, \theta_2, \cdots, \theta_K$ on a one-dimensional nested array with $P = 2M$ sensors located at $n\lambda/2$, $n \in \mathbb{S}_{\text{nst}}$, where $\mathbb{S}_{\text{nst}}$ is an integer set given by

$$\mathbb{S}_{\text{nst}} = \{m-1\}_{m=1}^{M} \bigcup \{m(M+1)-1\}_{m=1}^{M}.$$

[1] Only for $K = 1$, the MSE of coarray MUSIC provably decays to zero with SNR.

The signal received at the nested array is given by:

$$\mathbf{Y} = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)\mathbf{X} + \mathbf{N} \tag{4.1}$$

Here, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L] \in \mathbb{C}^{P\times L}$, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]^T \in \mathbb{C}^{K\times L}$ denote L snapshots of the received signal, and K unknown source signals respectively, and $\mathbf{N} \in \mathbb{C}^{P\times L}$ represents the additive noise. The normalized spatial frequencies are denoted by $\omega = [\omega_1, \omega_2, \cdots, \omega_K]^T$ where $\omega_i = \pi \sin(\theta_i)$. The nested array manifold matrix is given by $\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega) = [\mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_1), \mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_2), \ldots, \mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_K)] \in \mathbb{C}^{P\times K}$ where $\mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_i)$ is the steering vector corresponding to the normalized spatial frequency ω_i, and its elements are given by $[\mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_i)]_k = e^{j\omega_i d_k}, d_k \in \mathbb{S}_{\text{nst}}$. The difference coarray of $\mathbb{S}_{\text{nst}}$ is defined as $\mathbb{D}_{\mathbb{S}_{\text{nst}}} := \{n - m, m, n \in \mathbb{S}_{\text{nst}}\}$. For a nested array, it can be verified that [84]

$$\mathbb{D}_{\mathbb{S}_{\text{nst}}} = \big\{0, \pm 1, \pm 2, \ldots, \pm(N-1)\big\}, \quad N = M(M+1).$$

Let $\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega) = \Big[\mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega_1), \cdots, \mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega_K)\Big] \in \mathbb{C}^{N\times K}$ be the array manifold of a *virtual ULA* with N elements, whose sensor locations are given by the non-negative elements of the coarray $\mathbb{D}_{\mathbb{S}_{\text{nst}}}$. Assume that the source signals and noise are zero-mean random vectors and satisfying $\mathbb{E}[\mathbf{x}_l \mathbf{x}_l^H] = \text{diag}(\mathbf{p}),\ \mathbb{E}[\mathbf{n}_l \mathbf{n}_l^H] = \sigma_n^2 \mathbf{I}_P,\ \mathbb{E}[\mathbf{n}_l \mathbf{x}_m^H] = \mathbf{0}$ for all l and m. The vector $\mathbf{p} = [p_1, p_2, \cdots, p_K]^T$ denotes the source powers. It can be verified that the covariance matrix of the measurements, $\mathbf{R}_{\mathbf{yy}} := \mathbb{E}[\mathbf{y}\mathbf{y}^H]$ satisfies [84]

$$\mathbf{R}_{\mathbf{yy}} = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)\text{diag}(\mathbf{p})\mathbf{A}^H_{\mathbb{S}_{\text{nst}}}(\omega) + \sigma_n^2 \mathbf{I}_P = \mathbf{S}_{\text{nst}}\mathbf{T}\mathbf{S}^T_{\text{nst}}$$

where $\mathbf{S}_{\text{nst}} \in \mathbb{R}^{P\times N}$ is a (row) selection matrix which emulates the sensor locations of a nested array as follows:

$$[\mathbf{S}_{\text{nst}}]_{i,j} = \begin{cases} 1, & \text{if } d_i + 1 = j, d_i \in \mathbb{S}_{\text{nst}} \\ 0, & \text{otherwise} \end{cases}$$

The matrix $\mathbf{T} \in \mathbb{C}^{N \times N}$ is Toeplitz and represents *the coarray covariance matrix* corresponding to the *virtual coarray:*

$$\mathbf{T} = \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\text{diag}(\mathbf{p})\mathbf{A}^H_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega) + \sigma_n^2 \mathbf{I}_N \tag{4.2}$$

It is well-known that it is possible to uniquely identify $\mathbf{T}$ from $\mathbf{R}_{\mathbf{yy}}$ [104]. Coarray DOA estimation algorithms (such as coarray MUSIC [84–86, 93]) extract $\mathbf{T}$ from $\mathbf{R}_{\mathbf{yy}}$, and then apply standard subspace-based techniques on $\mathbf{T}$ to estimate ω, thereby leveraging the additional DOF provided by $\mathbb{D}_{\mathbb{S}_{\text{nst}}}$.

4.1.2 Inadequacy of Coarray MUSIC with Limited Snapshots

In practice, the ideal covariance matrix $\mathbf{R}_{\mathbf{yy}}$ is unavailable, and coarray based algorithms typically use the sample covariance matrix $\widehat{\mathbf{R}}_{\mathbf{yy}} = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H$ as its estimate. An estimate of $\mathbf{T}$ is obtained from $\widehat{\mathbf{R}}_{\mathbf{yy}}$ as [84, 86, 93, 98]

$$\hat{\mathbf{T}} = \mathscr{T}\left(\mathbf{F}\text{vec}\left(\widehat{\mathbf{R}}_{\mathbf{yy}}\right)\right) \in \mathbb{C}^{N \times N}. \tag{4.3}$$

Here, $\mathbf{F} \in \mathbb{R}^{(2N-1)\times P^2}$ is a redundancy averaging matrix whose m^{th} row $(1 \leq m \leq 2N-1)$ averages the elements of $\widehat{\mathbf{R}}_{\mathbf{yy}}$ corresponding to a lag of $m - N \in \mathbb{D}_{\mathbb{S}_{\text{nst}}}$, with a weight of $\Omega(|m-N|)$ denoting the number of pairs (d_k, d_l) with $d_k - d_l = m - N$ [84, 86]. Note that lags n and $-n$ have the same weight, given by $\Omega(|n|)$. By construction, $\mathbf{u} := \mathbf{F}\text{vec}(\widehat{\mathbf{R}}_{\mathbf{yy}}) \in \mathbb{C}^{2N-1}$ is a conjugate symmetric vector (i.e. $u_i = u^*_{2N-i}, 1 \leq i \leq N$), and the operator $\mathscr{T}(\cdot)$ returns a Toeplitz Hermitian matrix $\mathscr{T}(\mathbf{u})$ with $[\mathscr{T}(\mathbf{u})]_{i,j} = u_{N+(i-j)}$. Let $\mathbf{U} := [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K, \mathbf{u}_{K+1}, \ldots, \mathbf{u}_N]$ be a unitary matrix whose columns are eigenvectors of $\hat{\mathbf{T}}$, where $\mathbf{U}_1 := [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K]$ are the eigenvectors corresponding to K eigenvalues with largest magnitude (including repetitions, if any). Then $\mathscr{S} := \mathscr{R}(\mathbf{U}_1)$[2] is used as an estimate of the coarray subspace, which is subsequently utilized by subspace-based algorithms (such as coarray MUSIC) for DOA estimation.

[2] $\mathscr{R}(\mathbf{X})$ denotes the range space of a matrix $\mathbf{X}$.

Theorem 10. *Consider the measurement model (4.1) with $K = 2$ sources where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]^T \in \mathbb{R}^{2\times L}$ denotes the source signals. Suppose $\mathbf{x}_1^T \mathbf{x}_2 \neq 0$, $\omega_1 - \omega_2 \neq \frac{2m\pi}{N+M-1}$ for any integer m, and $M \geq 5$. Then, even in absence of noise ($\mathbf{N} = \mathbf{0}$), the coarray subspace cannot be identified from $\mathscr{S}$, i.e.*

$$\mathscr{S} \neq \mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{nst}}}(\omega)) \tag{4.4}$$

Proof. We establish (4.4) by contradiction. Suppose (4.4) does not hold, then we must have $\mathscr{R}(\mathbf{U}_1) = \mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega))$. Denote $\mathbf{U}_2 := [\mathbf{u}_{K+1}, \mathbf{u}_{K+2}, \ldots, \mathbf{u}_N]$. Since $\mathbf{U}$ is unitary, $\mathscr{R}(\mathbf{U}_2)$ is orthogonal to $\mathscr{R}(\mathbf{U}_1)$, and we have

$$\mathbf{U}_2^H \hat{\mathbf{T}} \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega) = \mathbf{0} \tag{4.5}$$

Denote $\text{vec}(\mathbf{X}\mathbf{X}^H) = [\sigma_1^2, \sigma_{12}, \sigma_{12}, \sigma_2^2]^T$ where $\sigma_i^2 = \|\mathbf{x}_i\|_2^2$, and $\sigma_{12} = \mathbf{x}_1^T \mathbf{x}_2$. We can decompose $\hat{\mathbf{T}}$ as:

$$\begin{aligned}\hat{\mathbf{T}} &= \frac{1}{L}\mathscr{T}\Big(\mathbf{F}\big(\big(\mathbf{A}_{\mathbb{S}_{\text{nst}}}^*(\omega) \otimes \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)\big)\,\text{vec}(\mathbf{X}\mathbf{X}^H)\big)\Big)\\ &= \frac{1}{L}\Big(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\text{diag}([\sigma_1^2, \sigma_2^2])\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}^H(\omega) + \sigma_{12}\mathscr{T}(\mathbf{v})\Big).\end{aligned}$$

Here $\mathbf{v} = \mathbf{F}(\mathbf{a}_{\mathbb{S}_{\text{nst}}}^*(\omega_1) \otimes \mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_2) + \mathbf{a}_{\mathbb{S}_{\text{nst}}}^*(\omega_2) \otimes \mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega_1))$. Since $\sigma_{12} \neq 0$, (4.5) is equivalent to $\mathbf{U}_2^H \mathscr{T}(\mathbf{v})\, \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega) = \mathbf{0}$. Let the elements of the conjugate symmetric vector $\mathbf{v}$ be denoted by $\mathbf{v} = [v_{N-1}^*,\ v_{N-2}^*,\ \ldots\ , v_0\ ,\ \ldots\ ,\ v_{N-2}, v_{N-1}]^T$, with

$$v_r = \frac{1}{\Omega(r)} \sum_{\substack{d_k, d_l \in \mathbb{S}_{\text{nst}} \\ r = d_k - d_l}} (e^{j(d_k\omega_2 - d_l\omega_1)} + e^{j(d_k\omega_1 - d_l\omega_2)}) \tag{4.6}$$

Now, $\mathbf{U}_2^H \mathscr{T}(\mathbf{v})\, \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega) = \mathbf{0}$ if and only if there exists $\mathbf{C} = [\mathbf{c}, \mathbf{c}'] \in \mathbb{C}^{2\times 2}$ such that

$\mathscr{T}(\mathbf{v})\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega) = \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\mathbf{C}$. In particular, we have $\mathscr{T}(\mathbf{v})\mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega_1) = \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\mathbf{c}$. Define

$$\begin{aligned}\alpha_k &:= e^{-j\omega_1}[\mathscr{T}(\mathbf{v})\mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega_1)]_k - [\mathscr{T}(\mathbf{v})\mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega_1)]_{k-1} \\ &= e^{-j\omega_1}v_{k-1} - e^{j\omega_1(N-1)}v^*_{N-k+1},\ N \geq k \geq 2\end{aligned}$$

We will choose an integer k_0 such that there exist two consecutive lags $k_0-1, k_0 \in \mathbb{D}_{\mathbb{S}_{\text{nst}}}$ whose weights satisfy (i) $\Omega(k_0-1) = \Omega(k_0) = 1$ and (ii) $\Omega(N-k_0+1) = \Omega(N-k_0) = 1$. Using the geometry of nested array, it can be seen that $k_0 = M+5$ will satisfy these conditions. Using (4.6) we have

$$\begin{aligned}\alpha_{M+5} &= e^{-j\omega_1}\big[e^{j\omega_2(2M+1)}e^{-j\omega_1(M-3)} + e^{j\omega_1(2M+1)}e^{-j\omega_2(M-3)}\big] \\ &\qquad - e^{j\omega_1(N-1)}\big[e^{-j(M^2-2)\omega_2}e^{j2\omega_1} + e^{-j(M^2-2)\omega_1}e^{j2\omega_2}\big] \qquad (4.7) \\ \alpha_{M+6} &= e^{-j\omega_1}\big[e^{j\omega_2(2M+1)}e^{-j\omega_1(M-4)} + e^{j\omega_1(2M+1)}e^{-j\omega_2(M-4)}\big] \\ &\qquad - e^{j\omega_1(N-1)}\big[e^{-j(M^2-2)\omega_2}e^{j3\omega_1} + e^{-j(M^2-2)\omega_1}e^{j3\omega_2}\big] \qquad (4.8)\end{aligned}$$

We also define $\beta_k := e^{-j\omega_1}[\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\mathbf{c}]_k - [\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\mathbf{c}]_{k-1}$, which can be simplified to $\beta_k = c_2 e^{j(k-2)\omega_2}(e^{j(\omega_2-\omega_1)}-1)$ where $\mathbf{c} = [c_1, c_2]^T$. It is clear that $e^{j\omega_2}\beta_k - \beta_{k+1} = 0$. From definition of α_k, β_k and due to the fact that $\mathscr{T}(\mathbf{v})\mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega_1) = \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\mathbf{c}$, it is also clear that $\alpha_k = \beta_k$. Therefore, we have $e^{j\omega_2}\alpha_k - \alpha_{k+1} = 0$. By substituting the values of $\alpha_{M+5}, \alpha_{M+6}$ from (4.7),(4.8),

$$\begin{aligned}&e^{j\omega_2}\alpha_{M+5} - \alpha_{M+6} = 0 \iff (e^{j\omega_2(2M+1)}e^{-j\omega_1(M-3)} - \\ &e^{-j\omega_2(M^2-2)}e^{j\omega_1(N+2)}) \times (e^{j(\omega_2-\omega_1)}-1) = 0\end{aligned}$$

Since $N = M(M+1)$ and $\omega_1 \neq \omega_2$, this implies that $e^{j\omega_2(N+M-1)} = e^{j\omega_1(N+M-1)}$ which can only happen if $\omega_2 - \omega_1 = \frac{2m\pi}{N+M-1}$ for some integer m. This contradicts the assumptions of Theorem 10, thereby proving our result. □

Theorem 10 proves that for non-orthogonal source signals (i.e. $\mathbf{x}_1^T\mathbf{x}_2 \neq 0$), coarray MUSIC will fail to identify the true coarray signal subspace $\mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega))$ in absence of noise, even when there are $K = 2$ sources. It provides the first theoretical validation of the empirical observation that with finite L, the MSE of coarray MUSIC always saturates away from 0 as SNR $\to \infty$, even when $1 < K < P$. In contrast, in absence of noise, direct MUSIC on a ULA can exactly recover the true signal subspace with only $L \geq K$ snapshots if $K < P$. This naturally leads to the question: *"With limited snapshots ($L \geq K$), can we harness the difference coarray of a nested array, while also ensuring exact recovery of the coarray signal subspace $\mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{nst}}}(\omega))$ in absence of noise, when $K \leq M$?"* In the next section, we provide a new algorithm for coarray subspace estimation that resolves this question.

4.1.3 Harnessing the Benefits of Virtual Coarray with Finite Snapshots

We present an alternative algorithm for coarray subspace estimation in the sample-starved regime, without sacrificing the enhanced DOF of nested arrays. We achieve this by *deviating away from existing methods for coarray subspace estimation [84, 86, 93], which primarily rely on the sample covariance matrix $\frac{1}{L}\mathbf{Y}\mathbf{Y}^H$.*

$$(\hat{\mathbf{T}}_{\text{ca}}, \hat{\mathbf{W}}) = \underset{\mathbf{T}_{\text{ca}},\mathbf{W}}{\arg\min} \left\|\mathbf{Y}\mathbf{W}\mathbf{Y}^H - \mathbf{S}_{\text{nst}}\mathbf{T}_{\text{ca}}\mathbf{S}_{\text{nst}}^H\right\|_F^2 \qquad \text{(Prox-Cov)}$$
$$\text{subject to} \quad \mathbf{T}_{\text{ca}} \succeq \mathbf{0}, \mathbf{T}_{\text{ca}} \text{ is Toeplitz}, \ \mathbf{W} \succeq \varepsilon \mathbf{I}_L$$

The main new idea here is to *jointly estimate a positive definite matrix* $\mathbf{W}$ *and a PSD Toeplitz matrix* $\mathbf{T}_{ca}$ such that $\mathbf{S}_{\text{nst}}\mathbf{T}_{\text{ca}}\mathbf{S}_{\text{nst}}^H$ best fits a weighted data covariance matrix $\mathbf{Y}\mathbf{W}\mathbf{Y}^H$. The constraint $\mathbf{W} \succeq \varepsilon \mathbf{I}_L$ for $\varepsilon > 0$ ensures that (Prox-Cov) does not return trivial solutions for $(\hat{\mathbf{T}}_{\text{ca}}, \hat{\mathbf{W}})$. Since we do not employ the sample covariance matrix $\frac{1}{L}\mathbf{Y}\mathbf{Y}^H$, $\hat{\mathbf{T}}_{\text{ca}}$ may no longer be an unbiased estimate of the true coarray covariance matrix, unlike the estimate $\hat{\mathbf{T}}$ in (4.3) used by coarray MUSIC. We call $\hat{\mathbf{T}}_{\text{ca}}$ as a "Proxy-Covariance" matrix, and the optimization problem as (Prox-Cov). Owing to the additional freedom available to design $\mathbf{W}$, and self regularization property of the PSD constraint $\mathbf{T}_{\text{ca}} \succeq \mathbf{0}$ [2, 89], (Proxy-Cov) will provide a superior estimate of the coarray subspace, especially with limited snapshots and low SNR (as will be shown in simulations). Most importantly, it overcomes a key drawback of coarray MUSIC (stated in Theorem 10) by ensuring *exact recovery of coarray signal subspace* in absence of noise, as long as $K \leq \min(M,L)$. The following theorem formally proves this.

Theorem 11. *Consider the measurement model (4.1) with* $\mathbf{N} = \mathbf{0}$. *Assume* $K \leq \min(M,L)$ *and let rank*$(\mathbf{X}) = K$. *Then, for any* $\varepsilon > 0$, *every optimal solution* $\hat{\mathbf{T}}_{ca}$ *of the problem (Prox-Cov) satisfies,* $\mathscr{R}(\hat{\mathbf{T}}_{ca}) = \mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{nst}}}(\omega))$.

Proof. Since $\mathbf{N} = \mathbf{0}$, $\mathbf{Y} = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)\mathbf{X}$. For any $\varepsilon > 0$, there exists $\alpha > 0$ such that the pair $(\mathbf{T}^*, \mathbf{W}^*) = (\alpha \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega)\mathbf{A}^H_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega), \alpha \mathbf{X}^{\dagger}\mathbf{X}^{\dagger H})$ [3] is feasible, and attains the optimal value 0. Hence, any globally optimal solution of (Prox-Cov) must also attain the optimal value 0 and therefore satisfy

$$\mathbf{Y}\hat{\mathbf{W}}\mathbf{Y}^H = \mathbf{S}_{\text{nst}}\hat{\mathbf{T}}_{\text{ca}}\mathbf{S}^H_{\text{nst}}. \tag{4.9}$$

Due to the geometry of nested array, we can partition $\mathbf{a}_{\mathbb{S}_{\mathbf{nst}}}(\omega)$ as $\mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega) = [\mathbf{a}^T_{\text{in}}(\omega), \mathbf{a}^T_{\text{out}}(\omega)]^T$ where $\mathbf{a}_{\text{in}} \in \mathbb{C}^{M+1}$ and $\mathbf{a}_{\text{out}} \in \mathbb{C}^{M-1}$ are steering vectors corresponding to the inner and outer uniform linear subarrays of a nested array. Let $\mathbf{A}_{\text{in}}(\omega) = [\mathbf{a}_{\text{in}}(\omega_1), \ldots, \mathbf{a}_{\text{in}}(\omega_K)]$ be the array manifold corresponding to the inner ULA. Then $\mathbf{A}_{\text{in}}(\omega)$ is a Vandermonde matrix with rank K (since $K \leq M$). Since the rows of $\mathbf{A}_{\text{in}}(\omega)$ coincide with first $M+1$ rows of $\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)$, we have $\text{rank}(\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)) = \text{rank}(\mathbf{A}_{\text{in}}(\omega)) = K$. On the other hand, since $\text{rank}(\mathbf{X}) = K$ and $\hat{\mathbf{W}} \succ \mathbf{0}$, the following is true:

$$\text{rank}(\mathbf{Y}\hat{\mathbf{W}}\mathbf{Y}^H) = \text{rank}(\mathbf{Y}) = \text{rank}(\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)\mathbf{X}) = K \tag{4.10}$$

Now, since $\hat{\mathbf{T}}_{\text{ca}}$ is a PSD Toeplitz matrix, it has the following decomposition due to

[3] $\mathbf{X}^{\dagger}$ denotes a right inverse of $\mathbf{X}$, which exists since $\text{rank}(\mathbf{X}) = K$.

Caratheodory's Theorem, [105]

$$\hat{\mathbf{T}}_{\text{ca}} = \mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega')\Lambda\mathbf{A}^H_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega')$$

where $K' = \text{rank}(\hat{\mathbf{T}}_{\text{ca}})$, Λ is a diagonal matrix with positive entries and

$$\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega') = [\mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega'_1), \dots, \mathbf{a}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega'_{K'})] \in \mathbb{C}^{N \times K'}$$

. From (4.9),(4.10), we must have $K' \geq K$. We now show that $K' = K$ and $\omega = \omega'$, which will imply the desired result $\mathscr{R}(\hat{\mathbf{T}}_{\text{ca}}) = \mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{\text{nst}}}}(\omega))$. Suppose $\omega'_i \notin \{\omega_1, \dots, \omega_K\}$ for some i. Now from (4.9), we must have

$$\mathbf{a}_{\mathbb{S}_{\text{nst}}}(\omega'_i) \in \mathscr{R}(\mathbf{Y}) = \mathscr{R}(\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)) \Rightarrow \mathbf{a}_{\text{in}}(\omega'_i) \in \mathscr{R}(\mathbf{A}_{\text{in}}(\omega))$$

This implies that $\bar{\mathbf{A}} = [\mathbf{A}_{\text{in}}(\omega), \mathbf{a}_{\text{in}}(\omega'_i)] \in \mathbb{C}^{(M+1)\times(K+1)}$ is a column rank deficient matrix. However, $\bar{\mathbf{A}}$ is a Vandermonde matrix with $K+1 \leq M+1$ columns, and $\omega_1, \dots, \omega_K, \omega'_i$ are all distinct. Hence $\text{rank}(\bar{\mathbf{A}}) = K+1$, which leads to a contradiction. Therefore, we must have $\omega'_i \in \{\omega_1, \dots, \omega_K\}$ for every i. Since $K' \geq K$, this can only happen if $\omega = \omega'$. □

Remark 4. Theorem 11 ensures *that in absence of noise, (Prox-Cov) can exactly recover the coarray signal subspace $\mathscr{R}(\mathbf{A}_{\mathbb{D}_{\mathbb{S}_{nst}}}(\omega))$ with only $L \geq K$ snapshots. Moreover, Theorem 11 also shows that unlike coarray MUSIC, (Prox-Cov) does not require the sources to be uncorrelated.*[4]

One can now employ any subspace based DOA estimation algorithm on the estimate $\hat{\mathbf{T}}_{\text{ca}}$ produced by (Prox-Cov) and harness the benefits of using the full coarray aperture, while ensuring that the MSE does not saturate any more with SNR as long as $K \leq M$. Moreover, our simulations suggest that by moving away from sample covariance matrix, (Prox-Cov) can identify more sources than sensors with higher accuracy than coarray MUSIC, especially with

[4] Our proof uses the geometry of nested arrays and it is non-trivial to extend it to other sparse arrays.

limited snapshots.

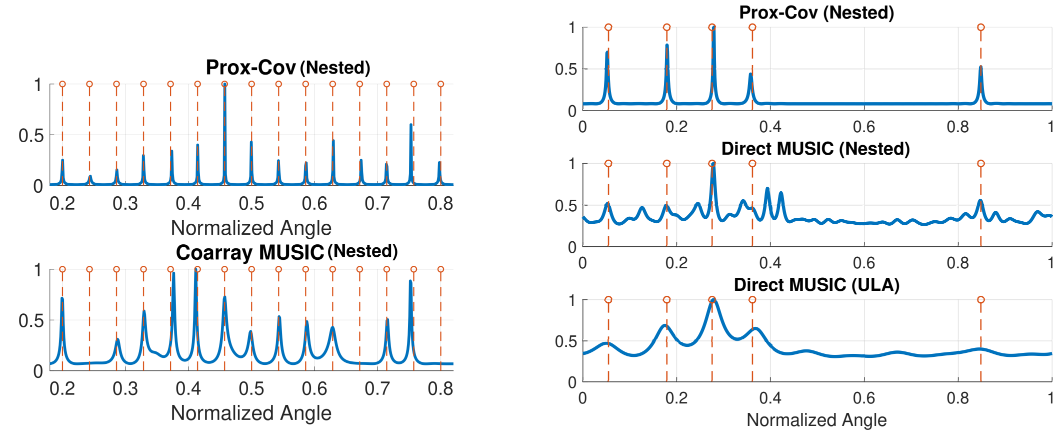


Figure 4.1. MUSIC spectra for (Left) (Prox-Cov) and coarray MUSIC at SNR = 20dB with 12 sensors and $L = 15$ snapshots, and (Right) (Prox-Cov) and Direct-MUSIC at SNR=5dB with 12 sensors and $L = 5$ snapshots. The true source locations are represented using red vertical dashed lines.

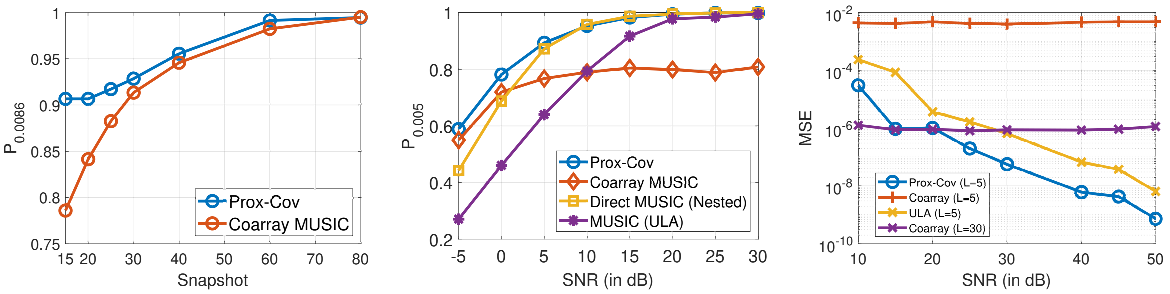
4.1.4 Simulations

Notice that (Prox-Cov) is a convex (specifically, conic) optimization problem. We implemented (Prox-Cov) using the SDPT3 solver in CVX package. In Figure 4.1, we compare the MUSIC spectrum generated using the output $\hat{\mathbf{T}}_{\text{ca}}$ of (Prox-Cov) against coarray MUSIC [84, 93], Direct MUSIC on nested array [106, 107] and MUSIC on ULA, all with 12 sensors. We define normalized DOAs $\bar{\theta} = \frac{\omega}{2\pi} \in [0, 1)$ [93]. The sources are assumed to be of unit power and the SNR is defined as $SNR = 10\log_{10}\left(1/\sigma_n^2\right)$. We consider two scenarios: (i) $K = 15(K > P)$ and (ii) $K = 5(K < P)$. In each case, we consider only $L = K$ snapshots. In both cases, (Prox-Cov) produces a sharper MUSIC spectrum where all sources are discernible. In Figure 4.1 (a), coarray MUSIC[5] fails to recover all the sources with only $L = 15$ snapshots. In Figure 4.1 (b), when $K < M$, direct MUSIC produces several small spurious peaks (for nested) and a flatter spectrum (for ULA).

In order to conduct a fair comparison of statistical performance of these algorithms against

[5] Recall that direct MUSIC on both nested and ULA fails when $K \geq P$

(Prox-Cov) in limited snapshot and/or low SNR regimes, we chose ”fraction of successfully recovered sources” as a more reliable metric [108]. We first generate K DOA estimates using root-MUSIC for the respective algorithms. Next, we choose a tolerance $r > 0$, and declare a source with DOA $\bar{\theta}$ to be successfully recovered if there exists an estimate $\hat{\theta}$ with $|\hat{\theta} - \bar{\theta}| < r$. We denote P_r as the fraction of successfully resolved sources. This metric also ensures that the MSE of the resolved sources is no larger than r. In Figure 4.2 (a), we consider $K = 15(K > P)$ sources located at $\bar{\theta}_n = 0.2 + 0.6(n-1)/14, n = 1, \dots, 15$ and compare (Prox-Cov) and coarray MUSIC. In this setting, coarray MUSIC requires almost twice the number of snapshots ($L = 30$) to achieve $P_{0.0086} > 0.9$ while (Prox-Cov) achieves with merely $L = 15$. On the other hand, for $K = 5(K < M)$ sources (with locations given by Figure 4.1 (b)), (Prox-Cov) can recover a larger fraction of sources with an error smaller than $r = 5 \times 10^{-3}$, especially when SNR is low. As predicted by Theorem 11, in Figure 4.2 (c), the MSE curve for (Prox-Cov) decreases sharply with SNR (similar to Direct MUSIC on ULA [109]), while coarray MUSIC saturates with SNR at different levels (depending on L). Overall, Figure 4.2 demonstrates that when L is small, (Prox-Cov) applied on nested arrays can recover more sources than sensors with higher accuracy than coarray MUSIC, while not exhibiting any saturation with respect to SNR when $K < M$.



(a) (b) (c)

Figure 4.2. Comparative study of DOA estimation performance of (Prox-Cov), coarray MUSIC and Direct MUSIC with 12 sensors. (a) P_r vs. L for $K = 15$ uncorrelated sources, SNR = 20 dB, $r = 8.6 \times 10^{-3}$. (b) P_r vs. SNR for $K = 5$ uncorrelated sources, $L = 5$ and $r = 5 \times 10^{-3}$ (c) Corresponding MSE for coarray MUSIC saturates with SNR, while MSE of (Prox-Cov) and ULA-MUSIC monotonically decay.

Finally in Figure 4.3, we compare the resolution of (Prox-Cov) and several state-of-the-art

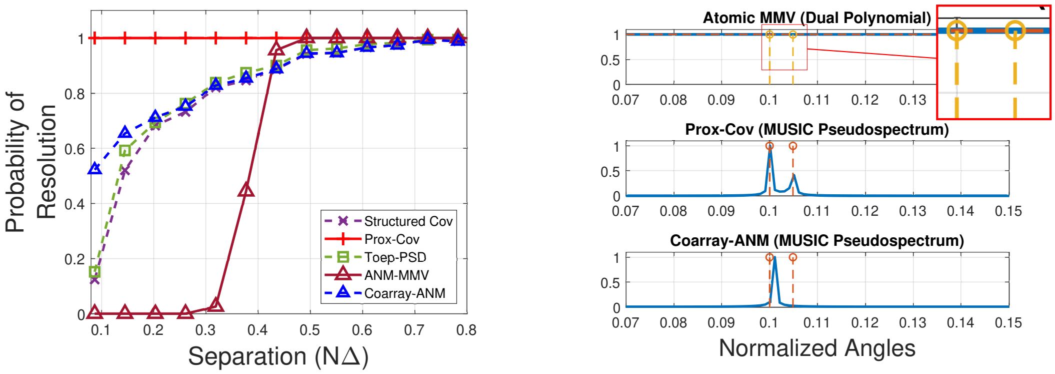


Figure 4.3. Comparison of resolution of (Prox-Cov) against gridless algorithms (Left) Probability of Resolution versus Separation for $K = 2$ sources in absence of noise with 12 sensors and $L = 2$ snapshots (Right) Comparison of DOA estimation of ANM-MMV, (Prox-Cov) and Coarray-ANM in absence of noise with $K = 2$ uncorrelated sources separated by $\Delta = 5\times 10^{-3}$ and $L = 2$ snapshots

gridless DOA estimation algorithms such as ANM-MMV [100], Toeplitz PSD [89], Structured Covariance [100], and Coarray-ANM [97], for resolving 2 sources with an angular separation of Δ in absence of noise. The sources are declared to be successfully resolved when the estimated DOAs satisfy $\max_i |\hat{\theta}_i - \bar{\theta}_i| \leq \Delta/2$.

As can be seen, when the separation is small ($\Delta = 5\times 10^{-3}$), ANM-MMV [100] and Coarray-ANM [97] (which use atomic-norm based regularizers) fail to resolve the sources, confirming the well-known fact that atomic-norm based methods can fail to resolve two sources even in absence of noise, if Δ violates the so-called "separation condition" [30, 35]. However, (Prox-Cov) succeeds for all values of Δ, which agrees with Theorem 11 where no restriction on the separation is imposed.

4.2 Resolving Coherent Sources with Sparse Arrays: An Interpolation Perspective

As discussed earlier, the large contiguous difference coarray of a well-designed sparse array is typically "synthesized" in the correlation domain, where the unobserved correlation values corresponding to missing sensors get *implicitly interpolated* by computing cross correlations between all pairs of sensor measurements. In the development of (Prox-Cov), although the number of snapshots is limited, it still leverages multiple snapshots generated by linearly independent source signals. However, this may still pose a challenge in applications where the sources/multi paths may be coherent and/or only a single snapshot might be available for source localization [110, 111].

In order to exploit the enhanced resolution of sparse arrays in sample-starved regimes, several algorithms have been developed for DOA estimation with a single (or limited) snapshot(s), using both off-grid and grid-based approaches [87, 110–119]. Another body of work aims at "completing a virtual ULA" with the same aperture as the sparse array, by estimating/interpolating the missing measurements with a single snapshot [115]. The virtual measurements can then be used for diverse tasks such as beamforming and source localization, aided by the enhanced resolution of the filled aperture of the virtual ULA [97, 110, 112, 115, 120]. A popular approach is to synthesize the virtual ULA measurements by using low-rank Toeplitz or Hankel matrix completion [112, 117, 120]. Indeed, the virtual measurements can be arranged in the form of a *low-rank* Hankel/Toeplitz matrix, and the measurements acquired by the sparse array only reveal certain entries of this matrix. In practice, for computational tractability, the rank constraint is often relaxed to a suitable convex surrogate, such as the nuclear norm or atomic norm [110, 112, 117, 120]. Although the aforementioned algorithms can also be applied for nested virtual array completion with only one snapshot, there is currently a disconnect between theory and practice. Existing guarantees for deterministic sparse array completion using nuclear norm minimization involve certain coherence conditions on the virtual Toeplitz/Hankel matrix and utilize specific graph-based array designs [110, 112] . On the other hand, theoretical guarantees for atomic norm minimization typically assume *randomized* sparse arrays, and require the source locations to satisfy a certain minimum separation even in the absence of noise [26, 35, 115]. These results therefore do not apply to deterministic spatial samplers such as nested arrays. Moreover, tight necessary and sufficient conditions remain an open question for single-snapshot virtual array completion via rank minimization.

Summary of our contributions: We address these open questions by providing the first necessary and sufficient conditions for rank-minimization to succeed in synthesizing the virtual array of a nested array with a single snapshot (Theorem 12). Since we consider the original rank-minimization framework, our results also reveal fundamental performance limits of any subsequent relaxation/approximation of the rank function. We guarantee exact interpolation (in absence of noise) regardless of the separation between sources, or coherence of the virtual Toeplitz matrix. Our converse results (necessary conditions) utilize the geometry of nested arrays in order to establish the existence of "ambiguous" source configurations (which we explicitly construct) for which rank-minimization will provably fail. **Notations:** Given a vector $\mathbf{z} \in \mathbb{C}^L$, the operator $\mathscr{T}_L(\mathbf{z})$ returns a $L \times L$ Hermitian Toeplitz matrix whose first column is given by $\mathbf{z}$. $\mathscr{R}(\mathbf{A})$ represents the range space of a given matrix $\mathbf{A}$. We denote $\mathbf{A}_{\mathbb{S}}(\omega) = [\mathbf{a}_{\mathbb{S}}(\omega_1), \mathbf{a}_{\mathbb{S}}(\omega_2), \dots, \mathbf{a}_{\mathbb{S}}(\omega_K)] \in \mathbb{C}^{P \times K}$ as the array manifold matrix of an array with sensors located at $n\lambda/2, n \in \mathbb{S} = \{d_1, d_2, \cdots, d_P\}$ (with source wavelength λ), and source frequencies are given by the set $\omega = \{\omega_1, \omega_2, \cdots, \omega_K\}$, with $[\mathbf{a}_{\mathbb{S}}(\omega_k)]_m = e^{j\omega_k d_m}$. We use $[\mathbf{v}]_{i_1:i_2}$ to denote a vector whose entries are given by those at indices $i_1, i_1+1, .., i_2$ of the vector $\mathbf{v}$.

4.2.1 Problem Formulation

Consider K far-field narrowband sources impinging from directions $\{\theta_i\}_{i=1}^K$ on a one-dimensional nested array with $P = 2M$ sensors whose locations are given by $\mathbb{S}_{\text{nst}}$, where $\mathbb{S}_{\text{nst}} := \mathbb{S}_1 \cup \mathbb{S}_2$ is the union of integer sets $\mathbb{S}_1 = \{m-1\}_{m=1}^{M+1}$ and $\mathbb{S}_2 = \{m(M+1)-1\}_{m=2}^{M}$. The signal received at the nested array is given by:

$$\mathbf{y}_{\text{nest}} = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega)\mathbf{x} + \mathbf{n}, \tag{4.11}$$

where $\mathbf{x} \in \mathbb{R}^K$ denotes real-valued [6] (deterministic) source signals and $\mathbf{n}$ is an additive noise term. The normalized spatial frequencies are denoted by $\omega = \{\omega_k\}_{k=1}^K$, with $\omega_i = \pi \sin(\theta_i)$.

The difference set $\mathbb{D}_{\mathbb{S}_{\text{nest}}}$ of $\mathbb{S}_{\text{nst}}$ is defined as $\mathbb{D}_{\mathbb{S}_{\text{nest}}} := \{m-n | m, n \in \mathbb{S}_{\text{nest}}\}$. It is well

[6] A similar setting with real source signals has been considered in [116]. In future, we will extend our theoretical results for the complex case.

known that the set of non-negative elements in $\mathbb{D}_{\mathbb{S}_{\text{nest}}}$ are given by $\mathbb{U} := \{0, 1, \cdots, N-1\}$ where $N{=}M(M+1)$ [84]. In the absence of noise, we can rewrite (4.11) as

$$\mathbf{y}_{\text{nest}} = \mathbf{S}_{\text{nest}}\mathbf{y}_{\text{full}}, \quad \mathbf{y}_{\text{full}} := \mathbf{A}_{\mathbb{U}}(\boldsymbol{\omega})\mathbf{x}, \tag{4.12}$$

where $\mathbf{S}_{\text{nest}} \in \mathbb{R}^{P \times N}$ is a row-selection matrix given by:

$$[\mathbf{S}_{\text{nest}}]_{i,j} = \begin{cases} 1, & \text{if } d_i + 1 = j, d_i \in \mathbb{S}_{\text{nst}} \\ 0, & \text{otherwise} \end{cases}$$

The vector $\mathbf{y}_{\text{full}}$ is a "*virtual measurement*", received by the virtual array $\mathbb{U}$, *with identical source configurations* (same DOAs $\boldsymbol{\omega}$ and source signal $\mathbf{x}$).

Key Question: We are interested in the problem of "sparse array interpolation" with only a *single snapshot*, where the goal is to estimate $\mathbf{y}_{\text{full}}$ from $\mathbf{y}_{\text{nest}}$. As discussed earlier, theoretical guarantees for matrix-completion or atomic norm mimimization based virtual array synthesis do not readily extend to nested arrays. This raises the open question: *What are the necessary and sufficient conditions under which rank-minimization with nested arrays leads to exact virtual array completion?*

Consider the noiseless measurement model (4.11) with $\mathbf{n} = \mathbf{0}$. From (4.12), it can be seen that when $\mathbf{x}$ is real, the matrix $\mathscr{T}_N(\mathbf{y}_{\text{full}})$ admits the following Vandermonde decomposition:

$$\mathscr{T}_N(\mathbf{y}_{\text{full}}) = \mathbf{A}_{\mathbb{U}}(\boldsymbol{\omega})\text{diag}(\mathbf{x})\mathbf{A}_{\mathbb{U}}^H(\boldsymbol{\omega}). \tag{4.13}$$

Consider the rank-minimization problem

$$\min_{\mathbf{u} \in \mathbb{C}^N} \text{rank}[\mathscr{T}_N(\mathbf{u})] \quad \text{subject to } \mathbf{S}_{\text{nest}}\mathbf{u} = \mathbf{y}_{\text{nest}}. \tag{P1}$$

The following theorem provides necessary and sufficient conditions under which perfect interpolation is possible (in absence of noise) by solving (P1).

Theorem 12. *Consider the measurement model* (4.11) *with* $\mathbf{n} = \mathbf{0}$*. If* $K \leq M$*, then* (P1) *has a unique solution* $\mathbf{u}^\star$ *satisfying* $\mathbf{u}^\star = \mathbf{y}_{full} = \mathbf{A}_{\mathbb{U}}(\omega)\mathbf{x}$*, for every* ω *and* $\mathbf{x}$*. Conversely if* $K > M$*, there exist source configurations with* K *source angles* $\omega_0 \in [-\pi, \pi)^K$*, and amplitudes* $\mathbf{x}_0 \in \mathbb{R}^K$*, such that one can find a vector* $\hat{\mathbf{y}}$*, with* $\hat{\mathbf{y}} \neq \mathbf{y}_{full}$ *(where* $\mathbf{y}_{full} = \mathbf{A}_{\mathbb{U}}(\omega_0)\mathbf{x}_0$*), satisfying*

$$\mathbf{S}_{nest}\hat{\mathbf{y}} = \mathbf{S}_{nest}\mathbf{y}_{full}, \quad rank(\mathscr{T}_N(\hat{\mathbf{y}})) \leq K \tag{4.14}$$

Proof. We first show that there exists no feasible point $\tilde{\mathbf{y}} \in \mathbb{C}^N$ of (P1) such that $\text{rank}(\mathscr{T}_N(\tilde{\mathbf{y}})) < K$. Consider a feasible point $\tilde{\mathbf{y}} \in \mathbb{C}^N$ and the following block partitioning of the matrix $\mathscr{T}_N(\tilde{\mathbf{y}})$:

$$\mathscr{T}_N(\tilde{\mathbf{y}}) = \begin{bmatrix} \mathbf{T}_1 & \mathbf{T}_2 \\ \mathrm{X} & \mathrm{Z} \end{bmatrix}, \tag{4.15}$$

where $\mathbf{T}_1 \in \mathbb{C}^{(M+1)\times(M+1)}, \mathbf{T}_2 \in \mathbb{C}^{M+1\times(N-M-1)}$. We also define a partitioning of the inner ULA manifold $\mathbf{A}_{\mathbb{S}_1}(\omega)$ as:

$$\mathbf{A}_{\mathbb{S}_1}(\omega) = \begin{bmatrix} \mathbf{1}^\top \\ \mathbf{B}(\omega) \end{bmatrix}, \tag{4.16}$$

where $\mathbf{B}(\omega) \in \mathbb{C}^{M\times K}$ is also a Vandermonde matrix due to the structure of the nested array. Since $\tilde{\mathbf{y}}$ is feasible, we have $\mathbf{S}_{\text{nest}}\tilde{\mathbf{y}} = \mathbf{y}_{\text{nest}}$, which implies

$$\mathbf{T}_1 = \mathscr{T}_{M+1}([\tilde{\mathbf{y}}]_{1:M+1}) = \mathscr{T}_{M+1}([\mathbf{y}_{\text{nest}}]_{1:M+1}) = \mathscr{T}_{M+1}(\mathbf{y}_{\mathbb{S}_1}).$$

where $\mathbf{y}_{\mathbb{S}_1} = \mathbf{A}_{\mathbb{S}_1}(\omega)\mathbf{x}$. Since $\mathbb{S}_1$ is a ULA, from (4.13), we have

$$\mathscr{T}_{M+1}(\mathbf{y}_{\mathbb{S}_1}) = \mathbf{A}_{\mathbb{S}_1}(\omega)\text{diag}(\mathbf{x})\mathbf{A}^H_{\mathbb{S}_1}(\omega). \tag{4.17}$$

Since $K \leq M$, $\text{rank}(\mathscr{T}_{M+1}(\mathbf{y}_{\mathbb{S}_1}))$=$K$. Hence, $\text{rank}(\mathscr{T}_N(\tilde{\mathbf{y}})) \geq K$, i.e., there exists no feasible point with rank strictly smaller than K.

Suppose $\text{rank}(\mathscr{T}_N(\tilde{\mathbf{y}})) = K$. We show that $\tilde{\mathbf{y}} = \mathbf{y}_{\text{full}}$ is the only feasible solution satisfying this property and this will prove that $\mathbf{y}_{\text{full}}$ is the unique solution to (P1). We need to show that $[\tilde{\mathbf{y}}]_i = [\mathbf{y}_{\text{full}}]_i$ for all $1 \leq i \leq N$. In other words, for every $j' = M+1, M+2, \cdots, N$, we will show that

$$[\tilde{\mathbf{y}}]_i = [\mathbf{y}_{\text{full}}]_i, \; \forall \; i \leq j'. \tag{4.18}$$

We establish this by induction on j'. The base case $j' = M+1$ follows because $\tilde{\mathbf{y}}$ is feasible and due to the structure of nested array, we also have $[\tilde{\mathbf{y}}]_i = [\mathbf{y}_{\mathbb{S}_1}]_i = [\mathbf{y}_{\text{full}}]_i, \; 1 \leq i \leq M+1$. Next, suppose (4.18) holds for $j' = j_0$ $(j_0 \geq M+1)$, and we will show that (4.18) also holds for $j_0 + 1$. Due to the induction hypothesis, showing (4.18) holds for $j' = j_0 + 1$ is equivalent to showing $[\tilde{\mathbf{y}}]_{j_0+1} = [\mathbf{y}_{\text{full}}]_{j_0+1}$. Denote $\bar{\mathbf{T}} := [\mathbf{T}_1 \quad \mathbf{T}_2] \in \mathbb{C}^{M+1 \times N}$. Due to the Toeplitz structure, the $(j_0+1)^{\text{th}}$ column of $\bar{\mathbf{T}}$ is given by:

$$\bar{\mathbf{t}}_{j_0+1} = \left[[\tilde{\mathbf{y}}]^*_{j_0+1}, [\tilde{\mathbf{y}}]^*_{j_0}, \ldots, [\tilde{\mathbf{y}}]^*_{j_0-M+1} \right]^\top \overset{(a)}{=} \left[[\tilde{\mathbf{y}}]^*_{j_0+1}, \bar{\mathbf{v}}^\top \right]^\top, \tag{4.19}$$

where $\bar{\mathbf{v}} = [[\mathbf{y}_{\text{full}}]^*_{j_0}, \cdots, [\mathbf{y}_{\text{full}}]^*_{j_0-M+1}]^\top$ and (a) follows from the induction hypothesis. From (4.12), for $i = 1, 2, \cdots, M$:

$$[\bar{\mathbf{v}}]_i = \sum_{k=1}^{K} e^{-j\omega_k(j_0-i)} x_k = \sum_{k=1}^{K} e^{j\omega_k i} e^{-j\omega_k j_0} x_k, \tag{4.20}$$

Define $\tilde{\mathbf{x}} \in \mathbb{C}^K$ as $[\tilde{\mathbf{x}}]_k = e^{-j\omega_k j_0} x_k$. From (4.16), we obtain

$$\bar{\mathbf{v}} = \mathbf{B}(\omega) \tilde{\mathbf{x}}. \tag{4.21}$$

Now, we use the fact that $\text{rank}(\mathscr{T}_N(\tilde{\mathbf{y}})) = K = \text{rank}(\mathbf{T}_1)$ which implies that $\text{rank}(\bar{\mathbf{T}}) = K$. Therefore, the $(j_0+1)^{\text{th}}$ column of $\bar{\mathbf{T}}$ $(\bar{\mathbf{t}}_{j_0+1})$ satisfies $\bar{\mathbf{t}}_{j_0+1} \in \mathscr{R}(\mathbf{T}_1)$. From the Vandermonde decomposition (4.17), it can be seen that $\mathbf{A}_{\mathbb{S}_1}(\boldsymbol{\omega})$ is a basis for $\mathscr{R}(\mathbf{T}_1)$, and hence there exists $\mathbf{c} \in \mathbb{C}^K$ such that

$$\bar{\mathbf{t}}_{j_0+1} = \mathbf{A}_{\mathbb{S}_1}(\boldsymbol{\omega})\mathbf{c} = \begin{bmatrix} \mathbf{1}^\top \\ \mathbf{B}(\boldsymbol{\omega}) \end{bmatrix} \mathbf{c}. \tag{4.22}$$

By combining (4.19), (4.21), and (4.22) we have the following:

$$\bar{\mathbf{t}}_{j_0+1} = \begin{bmatrix} [\tilde{\mathbf{y}}]^*_{j_0+1} \\ \bar{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^\top \mathbf{c} \\ \mathbf{B}(\boldsymbol{\omega})\mathbf{c} \end{bmatrix} \overset{(a)}{=} \begin{bmatrix} \mathbf{1}^\top \mathbf{c} \\ \mathbf{B}(\boldsymbol{\omega})\tilde{\mathbf{x}} \end{bmatrix}. \tag{4.23}$$

From the equality (a), we have $\mathbf{B}(\boldsymbol{\omega})\mathbf{c} = \mathbf{B}(\boldsymbol{\omega})\tilde{\mathbf{x}}$. Since $K \leq M$, $\mathbf{B}(\boldsymbol{\omega})$ is a Vandermonde matrix with full column rank and thus $\mathbf{c} = \tilde{\mathbf{x}}$. The proof is complete by plugging $\mathbf{c} = \tilde{\mathbf{x}}$ in (4.23) $[\tilde{\mathbf{y}}]^*_{j_0+1} = \sum_{k=1}^K [\tilde{\mathbf{x}}]_k = \sum_{k=1}^K e^{-j\omega_k j_0} x_k = [\mathbf{y}_{\text{full}}]^*_{j_0+1}$. For the converse results, we will show the existence of $\boldsymbol{\omega}_0, \mathbf{x}_0$ and $\hat{\mathbf{y}}$ with the desired properties by considering two cases (1) $2M+1 \leq K \leq N/2$ and (2) $M < K \leq 2M$:

1) ($2M+1 \leq K \leq N/2$): Consider any $2K$ distinct source angles denoted by the set $\Omega_{2K} := \{\omega_1, \omega_2, \cdots, \omega_{2K}\}$. We define a concatenated matrix $\mathbf{M}(\Omega_{2K}) \in \mathbb{R}^{4M \times 2K}$:

$$\mathbf{M}(\Omega_{2K}) = \begin{bmatrix} \operatorname{Re}(\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\Omega_{2K}))^\top & \operatorname{Im}(\mathbf{A}_{\mathbb{S}_{\text{nst}}}(\Omega_{2K}))^\top \end{bmatrix}^\top.$$

Since $K \geq 2M+1$, $\mathbf{M}(\Omega_{2K})$ has a non-trivial null space, i.e., there exists $\mathbf{v} \in \mathbb{R}^{2K}, \mathbf{v} \neq \mathbf{0}$ such that

$$\mathbf{M}(\Omega_{2K})\mathbf{v} = \mathbf{0}. \tag{4.24}$$

Suppose $\mathbf{v}$ has $L \leq 2K$ non-zero entries, and without loss of generality, let the indices of the non-zero elements be $\{1, 2, \cdots, L\}$ [7]. We select $\boldsymbol{\omega}_0$ as $\boldsymbol{\omega}_0 = \{\omega_1, \omega_2, \cdots, \omega_K\}$. Now, there can be two possibilities: either $L > K$, or $L \leq K$. Suppose $L > K$. In this case, let $\mathbf{x}_0 = -[\mathbf{v}]_{1:K} \in \mathbb{R}^K$ and construct $\hat{\mathbf{y}}$ as follows. Define $\bar{\boldsymbol{\omega}} := \{\omega_{K+1}, \cdots, \omega_L\}$ and $\bar{\mathbf{x}} := [\mathbf{v}]_{(K+1):L}$. Let $\hat{\mathbf{y}}$ be given

[7] The elements of the set Ω_{2K} can always be permuted to ensure this.

by $\hat{\mathbf{y}} = \mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}}$. In this case, since $[\mathbf{A}_{\mathbb{U}}(\bar{\omega}), \mathbf{A}_{\mathbb{U}}(\omega_0)]$ is a Vandermonde matrix with L distinct columns, it has full column-rank, since $L \leq 2K \leq N$. This implies that $\mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}} \neq \mathbf{A}_{\mathbb{U}}(\omega_0)\mathbf{x}_0$ for non-zero $\mathbf{x}_0, \bar{\mathbf{x}}$, and therefore $\hat{\mathbf{y}} \neq \mathbf{y}_{\text{full}}$ Next consider the case $L \leq K$. In this case, let $\mathbf{x}_0$ be given by $\mathbf{x}_0$=$[[\mathbf{v}]_{1:L}^{\top}, \mathbf{1}_{K-L}^{\top}]^{\top}$ (where $\mathbf{1}_{K-L} \in \mathbb{R}^{K-L}$ is a vector of all 1's), $\bar{\omega} := [\omega_{L+1}, \cdots, \omega_K]$, $\bar{\mathbf{x}}$=$\mathbf{1}_{K-L}$, and again construct $\hat{\mathbf{y}}$ as $\hat{\mathbf{y}} = \mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}}$. Once again, it can be verified that $\mathbf{y}_{\text{full}} \neq \hat{\mathbf{y}}$, otherwise it would imply (from the constructions of $\mathbf{x}_0$, $\bar{\mathbf{x}}$ and $\bar{\omega}$) that $\sum_{i=1}^{L} \mathbf{a}_{\mathbb{U}}(\omega_i)[\mathbf{x}_0]_i = 0$. This cannot happen since $\{\mathbf{a}_{\mathbb{U}}(\omega_i)\}_{i=1}^{L}$ are L distinct columns of a $N \times L$ Vandermonde matrix (with $L \leq N$), and are therefore linearly independent. Therefore, for each construction of $\hat{\mathbf{y}}$, we have $\mathbf{y}_{\text{full}} \neq \hat{\mathbf{y}}$, and (4.24) also implies that $\mathbf{S}_{\text{nest}}\mathbf{y}_{\text{full}} = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega_0)\mathbf{x}_0 = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\bar{\omega})\bar{\mathbf{x}} = \mathbf{S}_{\text{nest}}\mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}} = \mathbf{S}_{\text{nest}}\hat{\mathbf{y}}$. Since $\hat{\mathbf{y}} = \mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}}$, it also holds that $\text{rank}(\mathscr{T}_N(\hat{\mathbf{y}})) = \text{rank}\left(\mathbf{A}_{\mathbb{U}}(\bar{\omega})\text{diag}(\bar{\mathbf{x}})\mathbf{A}_{\mathbb{U}}^{H}(\bar{\omega})\right) = |K-L| \leq K$, as $L \leq 2K$.

2) ($M < K \leq 2M$): We begin by proving the following fact about the nested array. For every K in the range $M < K \leq 2M$, there is at most one $i \in \{2, \cdots, 2M\}$ (i.e. excluding the sensor at 0) such that d_i satisfies $\mod(d_i, K) = 0$. Suppose there exist two sensor locations d_l and d_m for which $\mod(d_l, K) = \mod(d_m, K) = 0$. Since $M < K < 2M+1$, and $d_i = (i-1), d_i \in \mathbb{S}_1$, this would imply that $d_l, d_m \in \mathbb{S}_2$. Therefore, there exist integers z_1, z_2 and $k_1, k_2 \in \{2, \cdots, M\}$, such that $k_1(M+1) - 1 = z_1 K$ and $k_2(M+1) - 1 = z_2 K$, which implies that $(k_2 - k_1)/K = z_2 k_1 - z_1 k_2$. Since $2 \leq k_1, k_2 \leq M$, we have $-(M-2) \leq k_2 - k_1 \leq M-2$. But we also have $M < K \leq 2M$. Hence, $(k_2 - k_1)/K \in \mathbb{Z}$ can be satisfied only if $k_1 = k_2$. This implies that $d_l = d_m$, and the statement is proved.

We now construct ω_0 as $\omega_0 = \{2\pi \frac{k}{K}\}_{k=0}^{K-1}$ [8] and let $\bar{\omega} = \omega_0 + 2\pi \frac{\alpha}{K}$, where α is chosen as follows. If there exists an integer $i_0 \in \{2, 3, \cdots 2M\}$ such that the sensor location $d_{i_0} = zK$ for some positive integer $z > 0$, then we choose $\alpha = \frac{1}{z}$. Else, α is chosen as an arbitrary real number satisfying $0 < \alpha < 1$. Redefine Ω_{2K} as $\Omega_{2K} := \omega_0 \cup \bar{\omega}$. We construct $\mathbf{w} \in \mathbb{R}^{2K}$ as follows: $[\mathbf{w}]_i = 1, [\mathbf{w}]_{K+i} = -1, \quad 1 \leq i \leq K$. Clearly, $[\mathbf{w}]_i \neq 0$ for all i. We will show that $\mathbf{w}$ satisfies

[8] Note that each $\omega_i = \pi \sin\theta_i$ maps to a unique angle in $[-\pi, \pi)$, which can again be uniquely mapped to $\omega_i \in [0, 2\pi)$.

$\mathbf{M}(\Omega_{2K})\mathbf{w} = \mathbf{0}$. Since the first sensor in nested array is assumed to be at the origin (i.e. $d_1 = 0$), we have $[\mathbf{M}(\Omega_{2K})\mathbf{w}]_1 = \mathbf{1}^\top \mathbf{w} = 0$, and $[\mathbf{M}(\Omega_{2K})\mathbf{w}]_{2M+1} = \mathbf{0}^\top \mathbf{w} = 0$. Consider any i in the range $2 \leq i \leq 2M$. First assume that the sensor location d_i satisfies $\mod (d_i, K) \neq 0$, implying that $\sin(\frac{\pi d_i}{K}) \neq 0$. Then,

$$\begin{aligned}[\mathbf{M}(\Omega_{2K})\mathbf{w}]_i &= \sum_{k=0}^{K-1} \cos(d_i \frac{2\pi k}{K}) - \sum_{k=0}^{K-1} \cos(d_i \frac{2\pi(k+\alpha)}{K}) \\ &= \frac{\sin(\pi d_i)}{\sin(\frac{\pi d_i}{K})}[\cos(\frac{\pi}{K} d_i (K-1)) - \cos(\frac{\pi}{K} d_i (K-1+2\alpha))] = 0\end{aligned}$$

since $\sin(\pi d_i) = 0$ for integer d_i. Similarly,

$$\begin{aligned}[\mathbf{M}(\Omega_{2K})\mathbf{w}]_{2M+i} &= \sum_{k=0}^{K-1} \sin(d_i \frac{2\pi k}{K}) - \sum_{k=0}^{K-1} \sin(d_i \frac{2\pi(k+\alpha)}{K}) \\ &= \frac{\sin(\pi d_i)}{\sin(\frac{\pi}{K} d_i)}[\sin(\frac{\pi}{K} d_i (K-1)) - \sin(\frac{\pi}{K} d_i (K-1+2\alpha))] = 0.\end{aligned}$$

Finally, suppose there exists i_0 such that d_{i_0}=zK. [9] Then, with the aforementioned choice of $\alpha = \frac{1}{z}$, we have $\cos(d_{i_0} \frac{2\pi k}{K}) = \cos(2\pi k z) = 1$ and $\cos(d_{i_0} \frac{2\pi(k+\alpha)}{K}) = \cos(2\pi k z + 2\pi) = 1$. This implies that $[\mathbf{M}(\Omega_{2K})\mathbf{w}]_{i_0} = 0$, as $\sum_i [\mathbf{w}]_i = 0$. Similarly, we have $\sin(d_{i_0} \frac{2\pi k}{K}) = \sin(2\pi k z) = 0$ and $\sin(d_{i_0} \frac{2\pi(k+\alpha)}{K}) = \sin(2\pi k z + 2\pi) = 0$, which implies that $[\mathbf{M}(\Omega_{2K})\mathbf{w}]_{i_0+2M} = 0$ as well. Combining the above results, we showed that $\mathbf{M}(\Omega_{2K})\mathbf{w} = \mathbf{0}$. Let $\mathbf{x}_0, \bar{\mathbf{x}} \in \mathbb{R}^K$ be defined as $[\mathbf{x}_0]_i = -[\mathbf{w}]_i, [\bar{\mathbf{x}}]_i = [\mathbf{w}]_{K+i}, 1 \leq i \leq K$. As before, construct $\hat{\mathbf{y}} = \mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}}$. Since $\mathbf{M}(\Omega_{2K})\mathbf{w} = \mathbf{0}$, we again have $\mathbf{S}_{\text{nest}}\mathbf{y}_{\text{full}} = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\omega_0)\mathbf{x}_0 = \mathbf{A}_{\mathbb{S}_{\text{nst}}}(\bar{\omega})\bar{\mathbf{x}} = \mathbf{S}_{\text{nest}}\mathbf{A}_{\mathbb{U}}(\bar{\omega})\bar{\mathbf{x}} = \mathbf{S}_{\text{nest}}\hat{\mathbf{y}}$. Since $K \leq N/2$, using a similar argument as the previous case, it can again be shown that $\hat{\mathbf{y}} \neq \mathbf{y}_{\text{full}}$. Furthermore, $\text{rank}(\mathscr{T}_N(\hat{\mathbf{y}})) = \text{rank}\left(\mathbf{A}_{\mathbb{U}}(\bar{\omega})\text{diag}(\bar{\mathbf{x}})\mathbf{A}_{\mathbb{U}}^H(\bar{\omega})\right) = K$. This concludes the proof. □

Theorem 12 guarantees that when $K \leq M$, it is possible to perfectly interpolate the missing sensors in a nested array, by solving the rank-minimization problem (P1) *regardless of the separation* between the sources. In fact, it can be shown that the sufficient condition broadly

[9] From our previous argument, there can be at most one such sensor.

applies to any sparse array with a ULA segment of length at least K. It also shows that, even with a single snapshot, a nested array can identify $O(M)$ sources (by applying any subspace based technique on the output of (P1)). While an ULA can also identify $K \leq M$ sources using single-snapshot MUSIC (SS-MUSIC) [20], an interpolated nested array can resolve sources with much smaller separation, especially in presence of noise. This is demonstrated in Figure 4.5. In the future, we will analyze how the outer ULA $\mathbb{S}_2$ controls this noisy interpolation error.

Theoretical conditions under which nested arrays will provably fail to identify sources with one snapshot were also unavailable.[10] An important contribution of Theorem 12 is to settle this question by showing that the sufficient condition is also necessary. This is done by constructing ambiguous source configurations that exploit the nested geometry.

4.2.2 Simulations

We solve a relaxed version of (P1) by replacing the rank with nuclear norm[11] and the equality constraint by a norm constraint $||\mathbf{S}_{\text{nest}}\mathbf{u} - \mathbf{y}_{\text{nest}}||_2 \leq \varepsilon$, assuming that the noise is bounded as $||\mathbf{n}||_2 \leq \varepsilon$. We call this approach Toeplitz Completion (TC). We first show the benefits of interpolation in beamforming with nested arrays using a single snapshot. We consider noiseless measurements acquired by a nested array with $P = 14$ sensors, comprised of three sources with amplitudes $\mathbf{x} = [1, -1, 1]$. In Figure 4.4, we plot the beam pattern obtained by interpolating the nested array measurements using TC, and then beamforming with the interpolated measurements for two different DOA configurations. For comparison, we plot the beam pattern obtained from beamforming with the physical nested array (without interpolation). We also perform interpolation with Atomic Norm Minimization (ANM) [115] and plot the resulting beam pattern in Figure 4.4 (last row). Due to large side lobes of the nested array, the source locations are not distinguishable when using the physical measurements. On the other hand, using the interpolated signal produced by TC, we can identify three closely spaced sources. Beamforming with the

[10] Most existing works focus on the multi-snapshot setting [84, 103, 119]

[11] Nuclear norm is just one among many approaches to replace “rank” by a suitable surrogate in order to attain computational tractability.

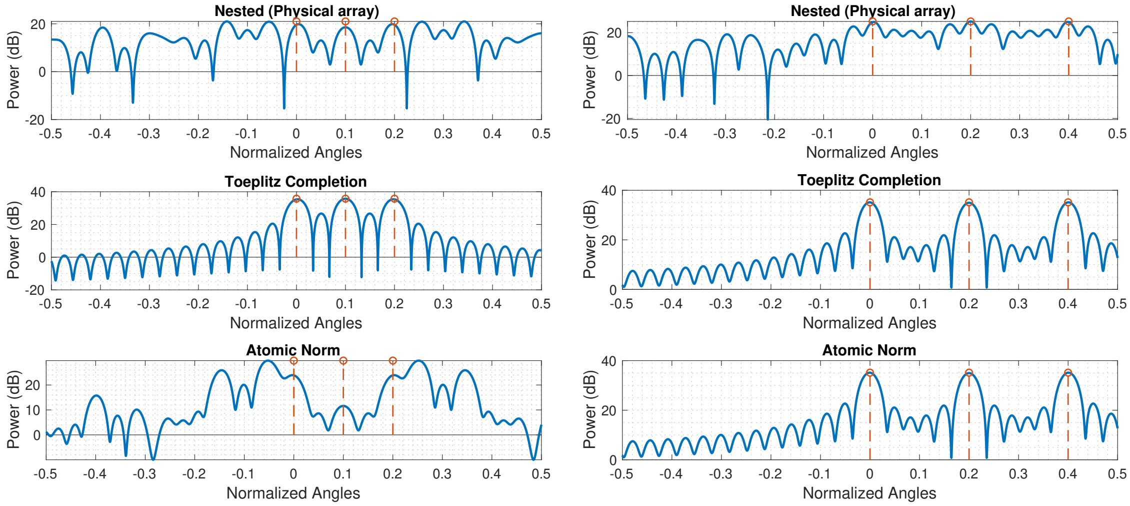


Figure 4.4. Comparison of beamforming on the nested array and interpolated virtual array with $K = 3$ sources located at (left) $\omega = \{0.0, 0.1, 0.2\}$ and (right) $\omega = \{0.0, 0.2, 0.4\}$. The total number of sensors is $P = 14$ and the interpolation was performed up to $N = 56$ sensors corresponding to the aperture of the nested array.

interpolated measurements using ANM fails to resolve sources with small separation (left), and succeeds only when the separation is large enough (right). We next study the DOA estimation error and the interpolation error in presence of noise. We consider a nested array with $P = 12$ sensors and $K = 5$ sources with spatial frequencies $\omega = \{\pi/20 + 0.1k\}_{k=0}^{4}$, and fixed amplitudes $\mathbf{x} = [1, -1, 1, 1, -1]^\top$. The additive noise is generated as $\mathbf{n} \sim \mathscr{U}(-\sigma/2, \sigma/2)$ and

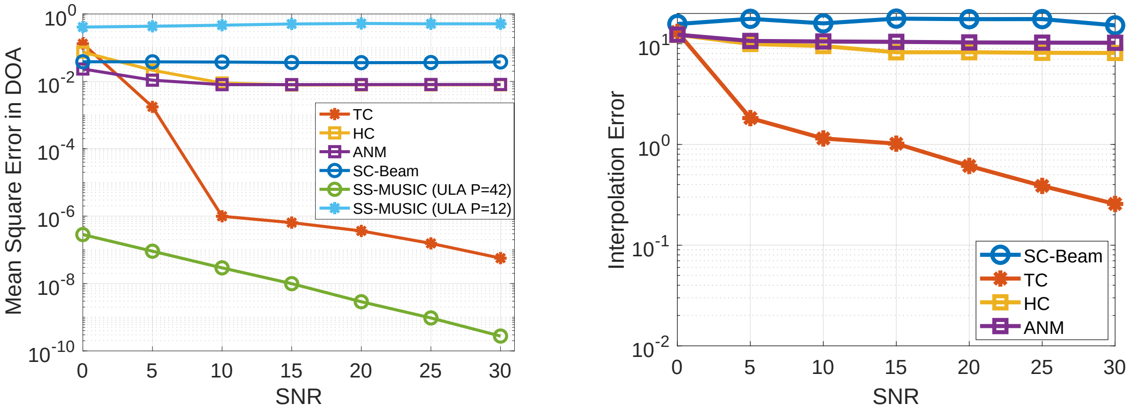


Figure 4.5. (Left) MSE in DOA vs. SNR for $K = 5$ sources with a nested array with $P = 12$ sensors and ULA (with $P = 12$ and $P = 42$). (Right) Normalized interpolation error vs. SNR for interpolating up to $N = 42$ sensors corresponding to the aperture of the $P = 12$ sensor nested array.

the SNR=$10\log(1/\sigma^2)$ is controlled by varying σ. In Figure 4.5 (left) we plot the MSE of DOA estimates (computed over 200 trials) as a function of SNR, by performing Root-MUSIC [121] on the output of TC. We also compare against Successive cancellation beamforming (SC-Beam) [113], ANM [29], and Hankel Completion (HC) [112], all of which permit single-snapshot DOA estimation with (arbitrary) sparse arrays, although their performances vary. We also compare the performance of SS-MUSIC on ULA (with 12 and 42 sensors). The 12-element nested array outperforms the ULA with 12 sensors and comes close to the performance of the 42-element ULA.

In Figure 4.5 (right) we also plot the interpolation error $\|\hat{\mathbf{y}}_{\text{full}} - \mathbf{y}_{\text{full}}\|_2/N$ versus SNR where $\hat{\mathbf{y}}_{\text{full}}$ is the estimated virtual measurement. For algorithms such as SC-Beam that does not perform explicit interpolation, we generate the interpolated signal using the DOA and source amplitude estimates as $\hat{\mathbf{y}}_{\text{full}} = \mathbf{A}_{\mathbb{U}}(\hat{\omega})\hat{\mathbf{x}}$. In both plots, we observe that the MSE of TC decays sharply with SNR while the other algorithms exhibit saturation. It is to be noted that the theoretical guarantees for these algorithms (if available) do not necessarily apply to deterministic sampling patterns such as nested arrays. Therefore, these techniques may fail to correctly identify all K sources (especially with small separation) with nested arrays, leading to saturation. The steady decay in the MSE of TC with increasing SNR is consistent with Theorem 12, which guarantees that exact interpolation is possible with nested arrays with $K \leq M$ sources by seeking low-rank solutions.

4.3 Conclusion

In the first half of the chapter, we proposed a new convex optimization framework called (Prox-Cov) to estimate the virtual coarray subspace of nested arrays with limited snapshots. When $K \leq \min(M,L)$, (Prox-Cov) provably recovers the true coarray subspace in the noiseless setting, while coarray MUSIC provably fails to do so. In numerical simulations, (Prox-Cov) outperforms coarray MUSIC as well as state-of-the-art gridless DOA estimation algorithms.

Extending the guarantees of (Prox-Cov) for other kinds of sparse arrays, and analyzing the performance of subspace algorithms [86, 103, 122] on the output of (Prox-Cov), are promising directions for future research.

In the later half of the chapter, we provided necessary and sufficient conditions for rank-minimization based techniques to successfully complete the virtual array of nested arrays from a single snapshot in the absence of noise. We showed that if $K \leq M$, one can exactly recover the missing measurements (and synthesize the virtual array) for any source configuration, by solving a Toeplitz matrix completion problem via rank-minimization. In contrast, when $K > M$, there exist source configurations (which depend on the nested geometry) where the recovery will provably fail. Our results indicate that the unique geometry of the nested array allows it to leverage the enhanced resolution of the virtual coarray (via non-linear interpolation), even with a single snapshot. In numerical simulations, the Toeplitz completion approach is observed to be robust to noise and outperforms other single snapshot source localization methods with nested arrays.

Chapter 4, in part, is a reprint of the material as it appears in the following papers:

- P. Sarangi, M. C. Hücümenoglu and P. Pal, “Beyond Coarray MUSIC: Harnessing the Difference Sets of Nested Arrays With Limited Snapshots,” in IEEE Signal Processing Letters, vol. 28, pp. 2172-2176, 2021.

- P. Sarangi, M. C. Hücümenoglu and P. Pal, “Single-Snapshot Nested Virtual Array Completion: Necessary and Sufficient Conditions,” IEEE Signal Processing Letters, vol. 29, pp. 2113-2117, 2022.

The dissertation author was one of the primary investigator and author of these papers.

Chapter 5

Bi-Linear and Quadratic Inverse Problems

5.1 Introduction: Non-Linear Inverse Problems

In Chapters 2 and 3, we focused on linear inverse problem. However, in many applications, the acquired measurements are non-linear functions of the parameters of interest. For example in optical imaging, CCD detectors are capable of only capturing the intensity and cannot record the phase information [123]. Consequently, recently efforts have been geared towards understanding the identifiability properties of certain classes of non-linear inverse problems, and developing algorithms with provable guarantees. In the first part of this chapter, we study a type of underdetermined bilinear inverse problem called blind deconvolution, where the measurements are a bilinear function of two unknown inputs. In the second part, we study a quadratic inverse problem known as phase retrieval, where we are interested in recovering the underlying structured signal from the magnitude of its linear/affine transformation (with a known operator). The results are developed for two kinds of compressive measurement operators. These problems are of great interest in signal processing and machine learning, as they are considered prototypical problems for understanding non-convex optimization.

5.2 Non-Negative Parametric Blind Deconvolution

The problem of recovering an unknown signal from its convolution with an unknown kernel (or filter) is known as blind deconvolution and it arises in a wide range of applications such as fluorescence microscopy, image deblurring, seismic imaging, neural spike detection, and communication [124–126]. Blind deconvolution is an inherently ill-posed bilinear inverse problem. However, it is possible to identify the signals (upto trivial scaling ambiguities) when suitable structural priors such as non-negativity, sparsity and subspace constraints are imposed on the underlying signals [127–129].

Blind deconvolution has received significant attention in recent times, largely because it is possible to develop probabilistic guarantees using convex and non-convex algorithms [127, 130–132]. Majority of these works typically assume the signal and/or kernel belong to low-dimensional subspaces [127, 130], exhibit sparsity over random dictionaries [128, 129], or consider deterministic short kernels being convolved with long and sparse signals [133, 134]. A common feature of these lines of work is that the unknown kernel is finite-length (or FIR filter) often with additional constraints.

The widely used assumptions for obtaining guarantees in blind deconvolution such as random subspace or sparsity over a random dictionary are often not applicable to naturally occurring signals. Instead, in several practical scenarios, it is possible to obtain a parametric representation for the kernel. One motivating example is the problem of neural spike deconvolution from calcium imaging data, which is already introduced in Chapter 2. As discussed earlier, in this problem, the underlying kernel is exponential decaying and can be well approximated by a stable first order autoregressive (AR(1)) process [10, 11, 135–137]. In contrast to recent works, the unknown kernel in this case is an infinite impulse response (IIR) filter and not bandlimited. In the neuroscience community, there have been several approaches to tackle this spike deconvolution problem problem such as template matching [52], probabilistic methods [10, 11, 135], convex formulations with imposition of priors such as sparsity, non-negativity [136–138] and using finite rate of innovation approach [38]. However,these algorithms either assume the kernel (model) parameters to be known apriori or separately estimated. None of these algorithms study the effect of subsampling on the kernel parameter and spike estimation which naturally originates in the problem. Estimation of Autoregressive parameter is a classical problem which has been studied in [139–141] using higher order moments and empirical spectral methods. However, these classical techniques do not consider undersampled measurements, or use of explicit priors such as non-negativity and sparsity for parameter estimation, and therefore cannot provide finite sample probabilistic bounds.

Summary of Contributions: In this chapter, we consider the problem of blind deconvolution of sparse signals using an unknown AR(1) kernel. We will specifically consider the case of uniform subsampling instead of randomized compressive sampling, since the former is more practical in the context of calcium imaging. However, obtaining guarantees for blind deconvolution from such uniformly subsampled measurements is also considerably more challenging. In [142], the authors show that certain deterministic compressive samplers for wide sense stationary signals can outperform randomized samplers by exploiting positivity constraint and Toeplitz structure. One of the key contribution of this chapter is to show that it is possible to uniquely recover the sparse input signal (representing neural spikes) from its subsampled convolution by exploiting the structure of the autoregressive model and by imposing non-negative constraints on the signal and the kernel. Assuming that the underlying sparse signal is generated by a Bernoulli model, we show that it is possible to uniquely recover the sparse signal and the AR(1) parameter with $O(s)$ (s being the expected sparsity) measurements with very high probability.

5.2.1 Problem Formulation

Let y_n represent the output of a first-order Autoregressive AR(1) filter with parameter α, $0 < \alpha < 1$, driven by the input signal s_n

$$y_n = \alpha y_{n-1} + s_n \tag{5.1}$$

Assuming that system to be at rest initially, i.e. $y_n = 0, n < 0$, we can rewrite $N+1$ consecutive samples of (5.1) in matrix-vector form as

$$y = \mathbf{G}_\alpha \mathbf{s} \tag{5.2}$$

where $y = [y_0, y_1, \cdots, y_N]^T, \mathbf{s} = [s_0, s_1, \cdots, s_N]^T$ and $\mathbf{G}_\alpha \in \mathbb{R}^{(N+1)\times(N+1)}$ is a Toeplitz matrix given by:

$$\mathbf{G}_\alpha = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^N & \alpha^{N-1} & \cdots & 1 \end{bmatrix}$$

The input signal $\mathbf{s}$ is assumed to be sparse with $||\mathbf{s}||_0 \ll N+1$ non-zero elements. In this chapter, we consider acquiring compressive measurements z_m by *uniformly subsampling* y by a factor of P, i.e. $z_m = y_{mP}$. Defining $M = \lceil \frac{N+1}{P} \rceil$, and $z = [z_0, z_1, \cdots, z_{M-1}]^T$ we obtain our main measurement model

$$z = \mathbf{D}y = \mathbf{D}\mathbf{G}_\alpha \mathbf{s} \tag{5.3}$$

where $\mathbf{D} \in \mathbb{R}^{M\times(N+1)}$ is a row-selection matrix representing the uniform downsampling operation described above. In this chapter, we assume that both α and the sparse input $\mathbf{s}$ are unknown quantities, and our goal is to perform blind deconvolution and identify $\alpha, \mathbf{s}$ from the uniformly subsampled measurements z.

Ambiguities in Blind AR(1) deconvolution: If α is known, and in absence of downsampling (i.e. $P = 1$, equivalently $\mathbf{D} = \mathbf{I}$), it is easy to see that $\mathbf{s}$ can be exactly recovered from $\mathbf{z}$ since $\mathbf{G}_\alpha$ is invertible. However, when α is unknown, the problem of recovering $\mathbf{s}$ from $\mathbf{z}$ is ill-posed, even in absence of downsampling. Specifically, even when $P = 1$, any input $\bar{\mathbf{s}} = [\bar{s}_0, \cdots, \bar{s}_N]^T$

constructed as

$$\bar{s}_0 = s_0, \quad \bar{s}_i = s_i + \beta z_{i-1}, i = 1, 2, \cdots, N-1$$

satisfies $z = \mathbf{G}_{\bar{\alpha}} \bar{\mathbf{s}}$ with $\bar{\alpha} = \alpha - \beta$ (choose β such that $\bar{\alpha} < 1$). The problem becomes severely ill-posed with downsampled measurements ($P > 1$). This necessitates imposing appropriate constraints on **s** to uniquely identify it. In the following section, we will assume that **s** is a *non-negative sparse* signal and exploit its non-negativity to obtain exact recovery guarantees. [1]

5.2.2 Identification of AR(1) Parameter and Sparse Signal

We begin by reformulating our measurements z_m by using the AR(1) model for y_n. First notice that using (5.1), for any $m \geq 1$, y_{mP} can be written as

$$y_{mP} = \alpha^P y_{(m-1)P} + \sum_{i=1}^{P} \alpha^{P-i} s_{(m-1)P+i}$$

. This implies that z_m satisfies

$$\begin{aligned} z_0 &= s_0 \\ z_m &= \alpha^P z_{m-1} + c_m, 1 \leq m \leq M-1 \end{aligned}$$

where $c_m := \sum_{i=1}^{P} \alpha^{P-i} s_{(m-1)P+i}$ for $1 \leq m \leq M-1$. These M equations can be re-arranged to obtain

$$\mathbf{A_z} \begin{bmatrix} \mathbf{c} \\ \alpha^P \end{bmatrix} = z \tag{5.4}$$

[1] The non-negativity of **s** is a natural assumption in applications such as neural spike deconvolution [11, 135]

where $\mathbf{c} = [s_0, c_1, c_2, \cdots, c_{M-1}]^T$ and $\mathbf{A_z} \in \mathbb{R}^{M\times(M+1)}$ is given by $\mathbf{A_z} = \begin{bmatrix} \mathbf{I}_M & z_s \end{bmatrix}$ and $z_s = [0, z_0, \cdots, z_{M-2}]^T$. Notice that c is related to the underlying sparse signal $\mathbf{s}$ via

$$c = \mathbf{H}(\alpha)\mathbf{s} \tag{5.5}$$

where $\mathbf{H}(\alpha) \in \mathbb{R}^{M\times(N+1)}$ is given by

$$[\mathbf{H}(\alpha)]_{m,n} = \begin{cases} \alpha^{mP-n}, & m \geq 1, (m-1)P+1 \leq n \leq mP, \\ 1, & m=0, n=0 \\ 0, & \text{otherwise} \end{cases}$$

We first aim to recover α by solving the linear system of equations (5.4). Notice that the matrix $\mathbf{A_z}$ also depends on the measurements $\mathbf{z}$. Moreover, it is easily seen that the dimension of null space of $\mathbf{A_z}$ is 1 and hence (5.4) has many solutions. We propose to seek a *non negative solution with minimum l_1 norm* by solving the following convex problem

$$\{\mathbf{c}^\star, \kappa^\star\} = \arg \min_{\mathbf{v}\in\mathbb{R}^M, \kappa\in\mathbb{R}} \|\mathbf{v}\|_1 \qquad (P1)$$
$$\text{subject to } \mathbf{A_z}\begin{bmatrix} \mathbf{v} \\ \kappa \end{bmatrix} = z, \quad \begin{bmatrix} \mathbf{v} \\ \kappa \end{bmatrix} \geq \mathbf{0}$$

Using the estimates $\mathbf{c}^\star$ and $\kappa^\star$ in (5.5), one may attempt to solve for the true spike signal $\mathbf{s}$. We develop conditions under which the true $\mathbf{s}$ can be the only solution to (5.5). In order to obtain our guarantees, we assume that the sparse signal $\mathbf{s}$ is generated according to a Bernoulli distribution (which naturally promotes sparsity [134, 135]) with parameter θ. Specifically, we assume that s_i are i.i.d Bernoulli variable with

$$\mathbb{P}(s_i = 1) = \theta$$

It is easy to see that the expected sparsity of $\mathbf{s}$ is $(N+1)\theta$, i.e., $E(\|\mathbf{s}\|_0) = (N+1)\theta$. Our main result is given by the following theorem.

Theorem 13. *Suppose $s_i, 0 \le i \le N$ are i.i.d Bernoulli random variables with $\mathbb{P}(s_i = 1) = \theta$. Let $\theta' = 1-(1-\theta)^P$ then for almost all α in the range $0 < \alpha < 1$ (except when α belongs to subsets of measure 0), the following hold with probability at least $1-p_e$*

$$p_e = (1-\theta)\frac{(1-\theta')^{M-1} - \theta'^{M-1}}{1-2\theta'}\theta' + \theta\theta'^{M-1}$$

(i) The solution to $(P1)$ satisfies $\mathbf{c}^\star = \mathbf{c}$ and $\kappa^\star = \alpha^P$

(ii) The true signal $\mathbf{s}$ is the only vector in $\{0,1\}^{N+1}$ that is solution to the following system of equations in $\mathbf{x}$:

$$c^\star = \mathbf{H}(\alpha^\star)\mathbf{x}$$

Proof. Let us define a partition for a signal $\mathbf{x} \in \mathbb{R}^{N+1}$

$$\mathbf{x} = [\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1^T, \cdots, \tilde{\mathbf{x}}_{M-1}^T, \tilde{\mathbf{x}}_r^T]^T \tag{5.6}$$

where $\tilde{\mathbf{x}}_i \in \mathbb{R}^P$, $\tilde{\mathbf{x}}_0 \in \mathbb{R}$ and $\tilde{\mathbf{x}}_r \in \mathbb{R}^{N-(M-1)P}$. Based on this partition we define a set of signals:

$$\mathscr{A} = \left\{\mathbf{x} \in \mathbb{R}^{N+1} | \exists i \; 0 \le i \le M-2 \;\; \text{s.t} \;\; \tilde{\mathbf{x}}_i \neq \mathbf{0} \text{ and } \tilde{\mathbf{x}}_{i+1} = \mathbf{0}\right\}$$

Step 1: Recovery of α and s: We first show that if the true input signal $\mathbf{s} \in \mathscr{A}$, then the solution to $(P1)$ satisfies $\mathbf{c}^\star = \mathbf{c}$ and $\kappa^\star = \alpha^P$. Since $\mathbf{s} \in \mathscr{A}$, it can be verified that there exists a k such that $c_k > 0$ and $c_{k+1} = 0$. Let the support of $\mathbf{c}$ and $\mathbf{c}^\star$ be S_0 and S^* respectively. We show that $S^* \subseteq S_0$. Suppose this is not true and $\exists \; i \ge 1, i \in S^\star$ and $i \notin S_0$. Let

$$v = \beta[0, z_0, z_1, \cdots, z_{M-1}, -1]^T$$

be the null space vector of $\mathbf{A}_z$ for some β. Since $\mathbf{z}_{i-1} > 0$ we have:

$$c_i^\star = c_i + v_i > 0 \Rightarrow \beta > 0$$
$$\Rightarrow v_j = \beta z_{j-1} \geq 0 \; \forall \; 1 \leq j \leq M-1, j \neq i \text{ and } v_i > 0$$

The value of the objective at the optimal $\mathbf{c}^\star$ is given by:

$$\|\mathbf{c}^\star\|_1 = \|\mathbf{c}\|_1 + \sum_{i=0}^{M-1} v_i > \|\mathbf{c}\|_1$$

This contradicts the fact that $\mathbf{c}^\star$ minimizes $(P1)$ and therefore $S^* \subseteq S_0$. For any $j \notin S_0$, it implies

$$c_j^\star = c_j + v_j = 0$$

Recall since $\mathbf{s} \in \mathscr{A}$, $\exists k$ such that $c_k > 0$ and $c_{k+1} = 0$ which implies $k+1 \notin S_0$ and $z_k > 0$. For any $\mathbf{c}^\star$ we have

$$c_{k+1}^\star = c_{k+1} + \beta z_k = 0 \Rightarrow \beta = 0$$

Therefore, $\mathbf{c}^\star = \mathbf{c}$ and $\kappa^* = \alpha^P$ are the optimal solution of (P1) provided the ground truth $\mathbf{s} \in \mathscr{A}$.

Step 2: Unique Reconstruction of Spike:

Now, we show that (ii) holds for almost all α in $0 < \alpha < 1$ except for the measure zero set defined as $\mathbb{S} = \bigcup_{i=1}^{3^P} \mathbf{Z}_i$ where the set $\mathbf{Z}_i$ is given by

$$\mathbf{Z}_i = \{x| \sum_{k=1}^{P} v_k^{(i)} x^{P-k} = 0\}.$$

Each set $\mathbf{Z}_i$ is defined by a unique vector $\mathbf{v}^{(i)} \in \mathbb{R}^P$ where $v_k^{(i)} \in \{-1,0,1\} \; \forall k \; 1 \leq k \leq P$. It can be verified that $|\mathbb{S}| \leq (P-1)3^P$.

Consider a fixed $\alpha \in (0,1)$. Let $\mathbf{u}, \mathbf{v} \in \{0,1\}^{N+1}$ be two distinct vectors ($\mathbf{u} \neq \mathbf{v}$) which satisfy $\mathbf{H}(\alpha)\mathbf{u} = \mathbf{H}(\alpha)\mathbf{v}$

$$\mathbf{H}(\alpha)\mathbf{u} - \mathbf{H}(\alpha)\mathbf{v} = \mathbf{0} \Rightarrow (\sum_{i=1}^{P}(u_i - v_i)\alpha^{P-i}) = 0. \tag{5.7}$$

Since, $u_i - v_i \in \{0, 1, -1\}, 1 \leq i \leq P$, equation (5.7) can be satisfied only if $\alpha \in \mathbb{S}$. If $\alpha \in (0,1) \setminus \mathbb{S}$ then $\mathbf{H}(\alpha)$ is one-to-one over the domain $\{0,1\}^{N+1}$. The probability $\mathbb{P}(\alpha \in \mathbb{S}) = 0$ since the set $\mathbb{S}$ is countable with cardinality at most $(m-1)3^m$ implying it has measure zero. Therefore, for almost all α, the mapping $\mathbf{H}(\alpha)$ is one-to-one over the domain $\{0,1\}^{N+1}$. If we exactly recover $\mathbf{c}$ and α from (P1) and $\alpha \in (0,1) \setminus \mathbb{S}$ then we can exactly identify $\mathbf{s}$.

Step 3: Probabilistic Characterization:

We have established that for almost all α the event $\mathbf{s} \in \mathscr{A}$ is a sufficient condition for (i) and (ii). Therefore, (i) and (ii) will hold with a probability at least $P(\mathbf{s} \in \mathscr{A})$. Now, we characterize the probability that $\mathbf{s} \in \mathscr{A}$ under the Bernoulli model. Let $\mathbb{B}_i$,$0 \leq i \leq M-1$ be a set defined as:

$$\mathbb{B}_i = \{\mathbf{x} \in \mathbb{R}^{N+1} | \tilde{\mathbf{x}}_j = \mathbf{0}, j < i \text{ and } \tilde{\mathbf{x}}_j \neq \mathbf{0}, j \geq i\}$$

Then for the complement set $\mathscr{A}^c = \bigcup_{i=0}^{M-1} \mathbb{B}_i$ the probability is

$$P(\mathbf{s} \in \mathscr{A}^c) = \sum_{i=0}^{M-1} P(\mathbf{s} \in \mathbb{B}_i) \tag{5.8}$$

To calculate $P(\mathbf{s} \in \mathbb{B}_i)$, we first evaluate $P(\tilde{\mathbf{s}}_j = \mathbf{0}), j \geq 1$ as follows:

$$\begin{aligned} P(\tilde{\mathbf{s}}_j = \mathbf{0}) &= P(s_i = 0,\ (j-1)P + 1 \leq i \leq jP) \\ &= (1-\theta)^P = 1 - \theta' \end{aligned}$$

Based on this, we can now compute the $P(\mathbf{s} \in \mathbb{B}_i)$ as follows:

$$
\begin{aligned}
P(\mathbf{s} \in \mathbb{B}_i) &= (1-\theta)\theta'^{(M-i)}(1-\theta')^{i-1}, i \geq 1 \\
P(\mathbf{s} \in \mathbb{B}_0) &= \theta\theta'^{M-1}
\end{aligned}
$$

Now, we can simplify (5.8) as follows:

$$
\begin{aligned}
P(\mathbf{s} \in \mathscr{A}^c) &= \theta\theta'^{M-1} + (1-\theta)\sum_{i=1}^{M-1} \theta'^{(M-i)}(1-\theta')^{i-1} \\
&= \theta\theta'^{M-1} + (1-\theta)\frac{(1-\theta')^{M-1} - \theta'^{M-1}}{1-2\theta'}\theta'
\end{aligned}
$$

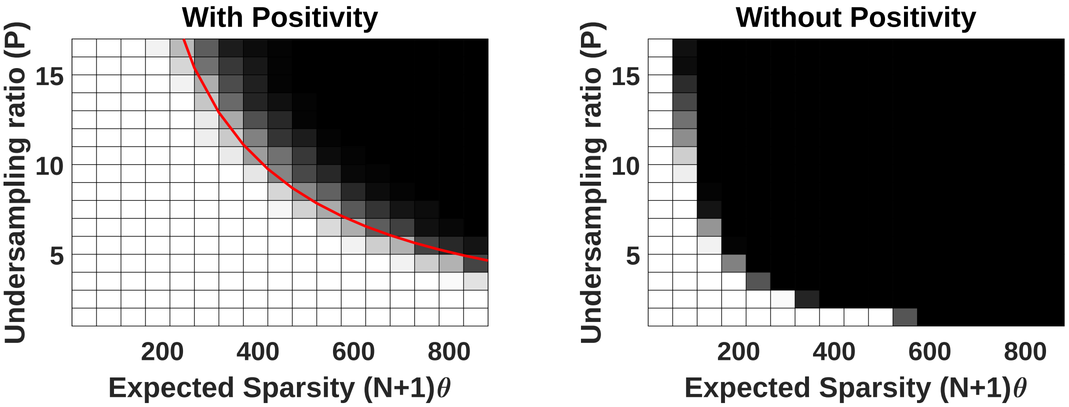
Therefore, we have $P(\mathbf{s} \in \mathscr{A}) = 1 - P(\mathbf{s} \in \mathscr{A}^c)$. Finally, we show that when $M \geq (N+1)\theta$ then asymptotically as $M, N \to \infty$ the probability of error goes to 0. It can be verified that when $M \geq (N+1)\theta$, both $\theta', (1-\theta') < 1$ and therefore the probability of error:

$$
p_e = \theta\theta'^{M-1} + (1-\theta)\sum_{i=1}^{M-1} \theta'^{(M-i)}(1-\theta')^{i-1} \to 0
$$

□

Remark 1. *Theorem 13 shows that as long as $M \geq (N+1)\theta$, it is possible to recover the AR parameter α with probability $\geq 1 - Me^{-cM}$, that exponentially increases to 1 in M. Notice that our sample complexity is proportional to expected sparsity (since $E(\|\mathbf{s}\|_0) = (N+1)K$) and hence optimal in this regime of sparsity.*

Remark 2. *Theorem 13 also shows that with the same sample complexity, the true sparse signal $\mathbf{s}$ can be uniquely identified as the only binary solution to (5.5), although we employed a naive uniform downsampling operation (instead of randomized measurements). This has very interesting connections to neural spike deconvolution algorithms (such as MLspike [11]) that treats the unknown spike signal as binary sequences, and develops decoding algorithms inspired* *from Viterbi decoding. Our result can provide performance guarantees for these techniques, and also lead to design of efficient decoding algorithms for recovering* **s***.*



(a)

(b)

Figure 5.1. Phase Transition Plots for successful recovery of **s** and α by solving $(P1)$ (a) with non-negative constraints (b) without non-negative constraints. White pixels indicate probability of success being 1 and black pixels denote zero probability of success.

5.2.3 Simulation

In the first experiment, we generate phase transition plots to characterize the empirical probability of recovering the sparse signal **s** and parameter α for different values of expected sparsity (by varying the Bernoulli parameter θ) and undersampling ratio P. In the phase transition plot, white color represents exact recovery whereas black denotes complete failure. We fix $N+1 = 1024$, $\alpha = 0.9$, and generate **s** according to the Bernoulli model described in Theorem 13. The estimates $\alpha^\star$ and $\mathbf{c}^\star$ are obtained by solving (P1). Using $\alpha^\star$ and $\mathbf{c}^\star$ in (5.5), we obtain an estimate of **s** by searching for a binary vector that solves (5.5). Figure 5.1 (a) shows the phase transition plot whose boundary corresponds to the red curve $P = 4/\theta$. This agrees with the observation regarding sample complexity made in Remark 1. Figure 5.1 (b) shows a second phase transition plot which is obtained by solving $(P1)$ without imposing the non-negative constraint. As can be seen, the performance degrades much quicker, showing the significance of the non-negative constraint.

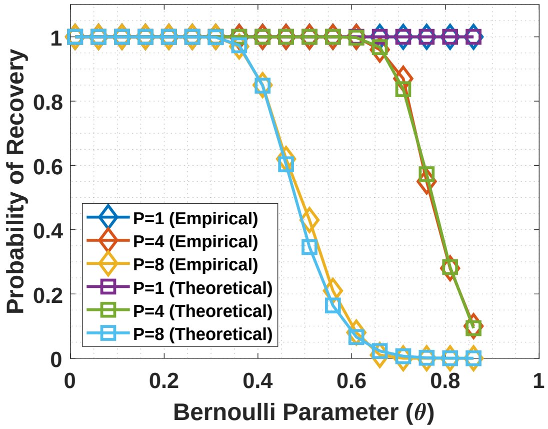


Figure 5.2. Comparison of theoretical and empirical probabilities of exact recovery of $\mathbf{s}$ and α (specified in Theorem 13) as a function of the Bernoulli parameters θ.

In the second experiment, we compare the theoretical recovery probability $1 - p_e$ specified in Theorem 13, with its empirically computed value, as a function of sparsity. We consider identical settings as the first experiment. Figure 5.2 shows the theoretical and empirical probability for exact recovery of α and $\mathbf{s}$ for different undersampling ratios as the Bernoulli parameter θ varies. We see that the theoretical and empirical recovery probabilities almost overlay each other, validating the main result from Theorem 13.

5.3 Sparse Phase Retrieval

In this section, we consider the problem of recovering a signal $\mathbf{x}_0 \in \mathbb{R}^n$ from quadratic measurements of the form

$$\mathbf{y}_i = |\mathbf{a}_i^T \mathbf{x}_0|^2, i = 1, 2, \cdots, m. \tag{5.9}$$

This problem is a generalization of the celebrated “phase retrieval” problem from optical imaging, and it features widely across a large number of imaging applications [143–145]. The problem has also been studied to develop efficient algorithms and analytical tools for performance.Beyond imaging, it is also closely related to other widely-studied non-convex problems such as low-rank matrix recovery [146] which are of great interest in machine learning and statistical signal processing.

It is well-known that an unstructured $\mathbf{x}_0$ can be provably recovered from (5.9) with $m = Cn, (C > 1)$ Gaussian measurements using convex algorithms in either lifted dimension [147–150] or in the original dimension [151, 152]. Non-convex algorithms are also known to provide exact recovery guarantees with similar sample complexity [153–156]. However, real world signals often possess lower-dimensional structure (such as sparsity). It is therefore crucial to understand how to exploit such structural assumptions to optimally reduce the sample complexity and yet ensure exact recovery with fewer measurements than the signal dimension n.

In this chapter, we will focus on the sparse phase retrieval problem where $\mathbf{x}_0$ is assumed to be sparse, i.e., it has only a few ($s < n$) non-zero entries. Lifting based convex relaxations have been modified to incorporate an additional sparsity promoting penalty, but they require $O(s^2 \log n)$ measurements, which is sub-optimal. In fact, the authors in [157, 158] show that such convex relaxation cannot achieve the desired sample complexity of $O(s \log n)$. It has also been shown that a different convex formulation in the original dimension can attain a sample complexity of $O(s \log n/s)$, provided it is initialized close to the true signal [159, 160]. Another class of algorithms directly solves the non-convex problem by first designing an initialization scheme, followed by an iterative minimization of either an intensity or amplitude loss function [161–163]. However, these non-convex algorithms also require $O(s^2 \log n)$ measurements to ensure exact recovery. Recently, [164] showed that it is possible to achieve a sample complexity of $O(s \log n/s)$ when optimizing the amplitude loss-function using the idea of *projected* gradient descent, provided the algorithm is initialized in the neighbourhood of the ground truth. The best known initialization scheme currently requires $O(s^2 \log n)$ measurements [162]. Therefore, for both convex and non-convex sparse phase retrieval algorithms, there is still a quadratic gap between the achievable sample complexity and the information theoretic lower limit of $m = \Omega(s \log(n/s))$ measurements. [2]

[2] In a separate line of work, the authors in [165, 166] have shown that it is possible to achieve a sample complexity

Summary of Contributions: One of the main contributions of this chapter is to show that it is possible to attain the optimal sample complexity of $\Omega(s\log(n/s))$ for sparse phase retrieval with Gaussian measurements in the *lifted space* without imposing any low-rank constraints. Exact recovery of the sparse signal is possible by only imposing positive semi-definite (PSD) constraint and sparsity constraints on the lifted variable. Our result is a generalization of a similar observation made by the authors in [149, 167] regarding unique optimization-free recovery of an unstructured signal in the lifted space with $O(n)$ measurements. Although our result is primarily of theoretical importance, it also provided us with crucial insights to develop a reweighted l_1 minimization algorithm in the lifted space that empirically outperforms competing sparse phase retrieval algorithms.

5.3.1 Problem Formulation

The goal of "generalized phase retrieval problem" is to recover a signal $\mathbf{x}_0 \in \mathbb{R}^n$ from m quadratic measurements of the form

$$\mathbf{y}_i = |\mathbf{a}_i^T \mathbf{x}_0|^2, \quad i = 1, 2, \cdots, m. \tag{5.10}$$

When $\mathbf{x}_0$ is sparse with $||\mathbf{x}_0||_0 = s$ non-zero elements, it has been shown [157] that if $m \geq 4s - 1$ and $\mathbf{a}_i \in \mathbb{R}^n, i = 1, 2, \cdots m$ are generic measurements vectors, the following problem has a unique solution (up to global sign ambiguity), coinciding with $\mathbf{x}_0$ [3]

$$\text{find } \mathbf{x}, \quad \mathbf{y}_i = |\mathbf{a}_i^T \mathbf{x}|^2, 1 \leq i \leq m, ||\mathbf{x}||_0 = s \tag{5.11}$$

Several approaches [157, 159–161, 163, 164] have been proposed to solve this sparse phase retrieval problem. However, to ensure recovery of $\mathbf{x}_0$ (with high probability), all these techniques

of $O(s\log n/s)$ using an initialization free approach, by restricting the measurement vectors $\mathbf{a}_i$'s to belong to a known subspace. However, these results are not applicable for Gaussian measurements.

[3] This result was strengthened to show that stable recovery of $\mathbf{x}_0$ is also possible with high probability, provided $m > c_1 s\log(en/s))$ (c_1 being a constant) and $\mathbf{a}_{i,j}$ are i.i.d standard Gaussian random variables [168]

require $m = \Omega(s^2 \log n)$ measurements. This indicates a critical quadratic gap between the number of measurements needed for exact recovery, and the sample complexity of these techniques [157, 158, 164]. Recent analysis of iterative algorithms such as PhaseMax [159, 160] and amplitude-based projected gradient descent [164] has shown that they can break the $O(s^2 \log(n))$ barrier, *provided one initializes these algorithms carefully*. However, the overall sample complexity of these algorithms still remains unknown (since it is unclear if such careful initialization itself will require $\Omega(s^2 \log n)$ measurements) [160].

A well-known technique to *linearize* the quadratic constraints (5.10) on $\mathbf{x}$ is to *lift* the vector $\mathbf{x} \in \mathbb{R}^n$ to a matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ [147, 148]. In particular, we solve

$$\min_{\mathbf{X}} \operatorname{rank}(\mathbf{X}), \quad \mathscr{A}(\mathbf{X}) = \mathbf{y}, \mathbf{X} \succeq \mathbf{0} \tag{5.12}$$

Here, $\mathscr{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ is a linear map such that $[\mathscr{A}(\mathbf{X})]_i \triangleq \mathbf{a}_i^T \mathbf{X} \mathbf{a}_i$. When $\mathbf{x}_0$ is sparse, the lifted matrix $\mathbf{X}_0 \triangleq \mathbf{x}_0 \mathbf{x}_0^T$ is simultaneously sparse and low-rank. Lifting-based convex approaches for sparse phase retrieval typically aim to recover $\mathbf{X}_0$ by simultaneously minimizing a linear combination of its l_1 norm and nuclear norm (which are convex surrogates for sparsity and rank respectively) [157, 158]

$$\begin{aligned} \text{minimize} \quad & \operatorname{Trace}(\mathbf{X}) + \lambda \|\operatorname{vec}(\mathbf{X})\|_1 \quad \text{(P2)} \\ \text{subject to } & \mathscr{A}(\mathbf{X}) = \mathbf{y} \quad \mathbf{X} \succeq \mathbf{0} \end{aligned}$$

When $\mathbf{a}_i$ are i.i.d standard normal vectors, the solution of (P2) recovers $\mathbf{X}_0$ with high probability, provided we acquire $m \geq c_3 s^2 \log(n)$ measurements, which again points to a quadratic barrier in sample complexity [157, 158].

Redundancy of Rank/Trace Minimization: A remarkable result in [167] shows that when $\mathbf{a}_i$

are independent vectors distributed uniformly on the unit sphere of $\mathbb{R}^n$, and $m \geq c_4 n \log n$, the set

$$\mathscr{F} \triangleq \{\mathbf{X} \in \mathbb{R}^{n \times n}, \text{ s.t. } \mathbf{X} \succeq \mathbf{0}, \mathscr{A}(\mathbf{X}) = \mathbf{y}\} \tag{5.13}$$

is a singleton (containing only $\mathbf{X}_0$) with high probability. Hence trace or rank minimization becomes unnecessary, and it is enough to simply solve a (convex) feasibility problem of the form "find $\mathbf{X}, \mathbf{X} \in \mathscr{F}$". Motivated by this result, we investigate an analogous question for sparse phase retrieval:

(Q): "When $\mathbf{x}_0$ is s-sparse, is it still possible to eliminate need for trace minimization in the lifted space and yet exactly recovery $\mathbf{X_0}$ with $m = \Omega\left(s \log(n/s)\right)$ measurements? "

5.3.2 Sparse phase retrieval With Optimum Sample Complexity via Lifting

In this chapter, we show that the answer to the above question is affirmative, provided we impose a *suitable sparsity penalty* along with the geometry of PSD cone. We begin by stating a simple property of sparse positive semi-definite matrices.

Lemma 11. *If $\mathbf{Z} \in \mathbb{R}^{n \times n}$ is a symmetric positive semi-definite matrix with $[\mathbf{Z}]_{ii} = 0$ for some i then $[\mathbf{Z}]_{ij} = 0$ and $[\mathbf{Z}]_{ji} = 0 \;\forall j$.*

Before understanding the implications of Lemma 11, we introduce some notations. Given an ordered index set $\mathbb{T} = \{i_1, i_2, \cdots, i_{|\mathbb{T}|}\} \subset [n]$ with $(i_1 < i_2 < \cdots < i_{|\mathbb{T}|})$, define the map $\mathscr{A}_{\mathbb{T}} : \mathbb{R}^{|\mathbb{T}| \times |\mathbb{T}|} \to \mathbb{R}^m$ as

$$[\mathscr{A}_{\mathbb{T}}(\mathbf{Z})]_i \overset{\Delta}{=} \mathbf{a}_{i\mathbb{T}}^T \mathbf{Z} \mathbf{a}_{i\mathbb{T}}, \quad 1 \leq i \leq m \tag{5.14}$$

where $\mathbf{a}_{i\mathbb{T}} \in \mathbb{R}^{|\mathbb{T}|}$ denotes the sub-vector of $\mathbf{a}_i$ indexed by $\mathbb{T}$. Moreover, for any matrix $\mathbf{Z} \in \mathbb{R}^{n \times n}$, we define $\mathbf{Z}_{\mathbb{T}} \in \mathbb{R}^{|\mathbb{T}|\times|\mathbb{T}|}$ to be the principal submatrix of $\mathbf{Z}$ corresponding to the index set $\mathbb{T}$, i.e.,

$$[\mathbf{Z}_{\mathbb{T}}]_{m,n} = [\mathbf{Z}]_{i_m,i_n}, \quad i_m, i_n \in \mathbb{T} \tag{5.15}$$

With these notations in place, we have the following result implied by Lemma 11. Suppose $\mathbb{S}$ denotes the set of indices of the non-zero diagonal elements of a positive semi-definite matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$. Lemma 11 then implies that, for all $\mathbb{T} \supset \mathbb{S}$, we have

$$\mathscr{A}(\mathbf{Z}) = \mathscr{A}_{\mathbb{S}}(\mathbf{Z}_{\mathbb{S}}) = \mathscr{A}_{\mathbb{T}}(\mathbf{Z}_{\mathbb{T}}) \tag{5.16}$$

Main Result

Given any $\mathbf{Z} \in \mathbb{R}^{n\times n}$, let $\operatorname{diag}(\mathbf{Z}) = [\mathbf{Z}_{1,1}, \cdots, \mathbf{Z}_{n,n}]^T \in \mathbb{R}^n$ be a vector consisting of the diagonal elements of the matrix $\mathbf{Z}$. Instead of simultaneously minimizing a linear combination of the rank and sparsity in the lifted space, we propose solving the following problem in order to recover $\mathbf{X}_0$ via lifting,

$$\begin{gathered} \min\nolimits_{\mathbf{X}\in\mathbb{R}^{n\times n}} \|\operatorname{diag}(\mathbf{X})\|_0 \\ \text{subject to } \mathbf{X} \succeq \mathbf{0}, \quad \mathscr{A}(\mathbf{X}) = \mathscr{A}(\mathbf{X}_0) \end{gathered} \tag{5.17}$$

Notice that we *do not* perform rank or trace minimization, but only exploit the fact that $\mathbf{X}_0$ is sparse. However, unlike existing methods that minimize the l_0 norm (or l_1 norm) of the *entire matrix in the lifted space*, we only seek the one with the *sparsest diagonal.* This specific choice of sparsity penalty is crucial to prove our main result that overcomes the quadratic barrier in sample complexity. Notice that since $\|\mathbf{X}_0\|_0 = s$, any solution to (5.17) belongs to the following set

$$\mathbb{C}_{\mathscr{A}} = \{\mathbf{Z} \in \mathbb{R}^{n\times n} | \mathscr{A}(\mathbf{Z}) = \mathscr{A}(\mathbf{X}_0), \|\operatorname{diag}(\mathbf{Z})\|_0 \leq s, \mathbf{Z} \succeq \mathbf{0}\}.$$

Our main result (stated below) is to show $\mathbb{C}_{\mathscr{A}}$ is singleton (i.e. $\mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}$) with high probability, provided $m \geq k_1 s \log(n/s)$.

Theorem 14. *Suppose $\mathbf{a}_i$ are i.i.d Gaussian vectors with independent $\mathcal{N}(0,1)$ entries. There exist constants c_1, c_2 such that with probability at least $1 - 3e^{-c_1\left(m - c_2 s \log(\frac{ne}{2s})\right)}$, it holds that $\mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}$.*

Proof. Our proof utilizes the concept of Frobenius-robust Rank Null Space Property (FRRNSP) of $\mathscr{A}$ from [146] along with the fact that $\mathbb{C}_{\mathscr{A}}$ consists of PSD matrices. We provide a sketch of proof. We first review the definition of FRRNSP from [146]

Definition 3. *We say that $\mathscr{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ satisfy the Frobenius-Robust Rank Null Space Property (FRRNSP) with respect to l_2 of order r with constants $0 < \rho < 1$ and $\tau > 0$ if for all $\mathbf{M} \in \mathbb{C}^{n_1 \times n_2}$, the singular values of $\mathbf{M}$ satisfy*

$$\|\mathbf{M}_r\|_2 \leq \frac{\rho}{\sqrt{r}} \|\mathbf{M}_c\|_1 + \tau \|\mathscr{A}(\mathbf{M})\|_2.$$

Here $\|\mathbf{M}_r\|_2 = (\sum_{i=1}^{r} \sigma_i^2)^{1/2}$ and $\|\mathbf{M}_c\|_1 = \sum_{i=r+1}^{\min(n_1,n_2)} \sigma_i$ and σ_i denotes the ith singular value of $\mathbf{M}$.

Given any index set $\mathbb{T} \subset [n]$, we define the set $\mathbb{G}_{\mathbb{T}}$ as[4]

$$\begin{aligned} \mathbb{G}_{\mathbb{T}} = \{ & \mathbf{A}, \text{ s.t. } \mathbf{W}_{\mathbb{T}} \triangleq \frac{1}{m} \sum_{j=1}^{m} \mathbf{a}_{j\mathbb{T}} \mathbf{a}_{j\mathbb{T}}^T \succ \mathbf{0}, \text{ and} \\ & \mathscr{A}_{\mathbb{T}} \text{ satisfies FRRNSP with } 0 < \rho < \frac{1}{\kappa(\mathbf{W}_{\mathbb{T}})} \} \end{aligned}$$

where $\kappa(\mathbf{X})$ denotes the condition number of a matrix $\mathbf{X}$. Next, we define the following sets

$$\mathbb{G} = \bigcap_{|\mathbb{T}|=2s} \mathbb{G}_{\mathbb{T}}, \quad \mathbb{E} = \{\mathbf{A}, \text{ s.t. } \mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}\}$$

[4] $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_m] \in \mathbb{R}^{n \times m}$

Consider a fixed $\mathbf{A} \in \mathbb{G}$. Suppose $\mathbf{Z}^* \in \mathbb{C}_{\mathscr{A}}$ is a solution to (5.17) with $\mathbb{S}^* \triangleq \mathrm{Supp}(\mathrm{diag}(\mathbf{Z}^*))$. Let $\mathbb{L} \subset [n]$ be any set with $\|\mathbb{L}\|_0 = 2s$ and $\mathbb{S}^* \cup \mathbb{S}_0 \subseteq \mathbb{L}$. Using the fact that $\mathbf{A} \in \mathbb{G}$ and $\mathbf{Z}^*_{\mathbb{L}}$ and $\mathbf{X}_{0\mathbb{L}}$ are PSD matrices satisfying (5.16), it can be shown that [146]

$$\|\mathbf{Z}^*_{\mathbb{L}} - \mathbf{X}_{0\mathbb{L}}\|_2 \leq \alpha_1 \|\mathscr{A}_{\mathbb{L}}(\mathbf{Z}^*_{\mathbb{L}}) - \mathscr{A}_{\mathbb{L}}(\mathbf{X}_{0\mathbb{L}})\| = 0$$

This implies that $\mathbf{Z}^* = \mathbf{X}_0$ and hence $\mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}$. Hence, if $\mathbf{A} \in \mathbb{G}$, we also have $\mathbf{A} \in \mathbb{E}$, implying that $\mathbb{G} \subset \mathbb{E}$. Hence we have

$$\begin{aligned} \mathbb{P}(\mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}) &= \mathbb{P}(\mathbf{A} \in \mathbb{E}) \geq \mathbb{P}(\mathbf{A} \in \mathbb{G}) \\ &\geq 1 - \textstyle\sum_{\|\mathbb{T}\|=2s} \mathbb{P}\left(\mathbf{A} \in \mathbb{G}^c_{\mathbb{T}}\right) \end{aligned} \tag{5.18}$$

Using the fact that $\mathbf{A}$ has i.i.d standard Normal entries, we can use results from [146] to show that

$$\mathbb{P}(\mathbf{A} \in \mathbb{G}^c_{\mathbb{T}}) \leq 3e^{-c_1 m} \tag{5.19}$$

whenever $m \geq c' s$ for some constants c_1, c'. Substituting (5.19) into (5.18) finally yields

$$\mathbb{P}(\mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}) \geq 1 - 3e^{-c_1(m - sc_2 \log(\frac{ne}{2s}))}$$

□

A direct consequence of Theorem 14 is that there exist constants k_1, k_2 such that whenever $m \geq k_1 s \log(n/s)$, $\mathbb{C}_{\mathscr{A}} = \{\mathbf{X}_0\}$ with probability at least $1 - e^{-k_2 m}$. Hence our result shows that it is possible to attain $O(s\log(n/s))$ sample complexity for sparse phase retrieval in the lifted dimension even without explicitly enforcing low-rank constraints.

Iterative Reweighted Algorithm To Solve (5.17)

Theorem 14 shows that (5.17) has a unique solution with high probability if $m \geq k_1 s \log(n/s)$. We now propose an algorithm to approximate this non-convex l_0 minimization problem by solving a sequence of convex problems. In each iteration, we solve a re-weighted trace minimization problem (P_w), the details of which is provided in Algorithm 5.

$$\begin{aligned} &\text{minimize} \quad \text{trace}(\mathbf{WZ}) \quad (P_w) \\ &\text{subject to } \mathscr{A}(\mathbf{Z}) = \mathscr{A}(\mathbf{X}_0), \mathbf{Z} \succeq \mathbf{0} \end{aligned}$$

Notice that our weighting scheme is different from [148], where the weights are updated to promote sparsity of the entire matrix. In contrast, we only update the weights to enforce sparsity of the diagonal entries since the positive semi-definite constraint will implicitly promote sparsity of the remaining matrix.

Algorithm 5. PhaseLift with Diagonal Sparsity Enforcing Reweighted l_1 minimization

```
**Input:** Quadratic measurements $\mathbf{y} \in \mathbb{R}^m$
**Output:** Estimate $\hat{\mathbf{x}}$ of the sparse signal $\mathbf{x}_0$
Initialize with $\mathbf{W}^{(0)} = \mathbf{I}$, $k \leftarrow 0$, and a sequence of non-increasing numbers $\{\varepsilon_k\}$ satisfying $\lim_{k\to\infty} \varepsilon_k \geq 0$.
**Repeat**
  1. Obtain $\mathbf{X}^{(k+1)}$ as the solution to $(P_w)$ with $\mathbf{W} = \mathbf{W}^{(k)}$
  2. Update the weights $\mathbf{W}^{(k+1)}_{i,j} = \begin{cases} 0, i \neq j \\ \frac{1}{\mathbf{X}^{(k+1)}_{i,i} + \varepsilon_k}, i = j \end{cases}$
  3. $k \leftarrow k+1$
**until convergence** $\|\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}\|_F < \tilde{\varepsilon}_1$
After convergence, let $\hat{\mathbf{X}}$ be the best rank-1 approximation to $\mathbf{X}^{(k)}$. Obtain $\hat{\mathbf{x}}$ as the top singular vector of $\hat{\mathbf{X}}$, i.e. $\hat{\mathbf{X}} = \hat{\mathbf{x}}\hat{\mathbf{x}}^T$
```

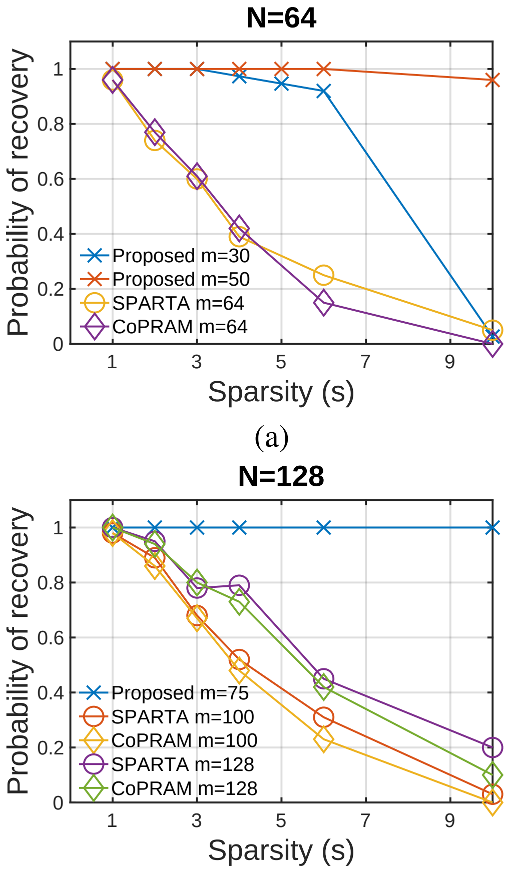
5.3.3 Simulations

In this section, we demonstrate the effectiveness of the proposed reweighted l_1 minimization algorithm inspired by the insights from Theorem 14. We compare our algorithm with SPARTA [161] and CoPRAM [163], which are effective for sparse phase retrieval with Gaussian measurements. For each iteration, we generate a sparse signal $\mathbf{x}_0$ of length N and sparsity s, whose non-zero entries are uniformly distributed in the range $[1,4]$. We evaluate the performance of each algorithm for a given sparsity s in terms of successful recovery rate computed over 100 iterations. We consider a recovery successful if the relative error in the lifted space satisfies $\frac{\|\mathbf{Z}-\mathbf{x}\mathbf{x}^T\|_F}{\|\mathbf{x}\mathbf{x}^T\|_F} \leq 10^{-3}$. In the first set of experiments, we consider two different values of N and plot the probability of success as a function of sparsity s in Figure 5.3 (a). For $N = 64$, we provide $m = 64$ measurements to both SPARTA and CoPRAM and compare against our algorithm which is given $m = 30$ and $m = 50$ measurements. For $N = 128$, we test SPARTA and CoPRAM with $m = 100$ and $m = 128$ measurements, whereas we use $m = 75$ measurements for our algorithm (shown inFigure 5.3 (b)). In both cases, we see that our algorithm significantly outperforms SPARTA and CoPRAM even when it uses fewer number of measurements.

In the next experiment, we fix $N = 256$, $s = 5$ and compute the probability of success as a function of the number of measurements. Figure 5.4, shows that the proposed algorithm requires only 160 measurements to recover the signal with probability 1, whereas the other two algorithms can only achieve a success rate of 0.5.

5.4 Interferometric Phase Retrieval

Optical Coherence Tomography (OCT) is an interferometic imaging technique which is widely used for imaging of biological tissues (such as retina, skin and coronary arteries) and capturing microstructure in materials. The two widely used approaches of performing OCT are Time Domain OCT (TDOCT) and Frequency Domain OCT (FDOCT) The setup of both are very similar but FDOCT offers distinct advantages over TDOCT since it allows faster acquisition



(a)

(b)

Figure 5.3. Comparison of probability of success versus sparsity. Here (a) N=64 (b) N=128.

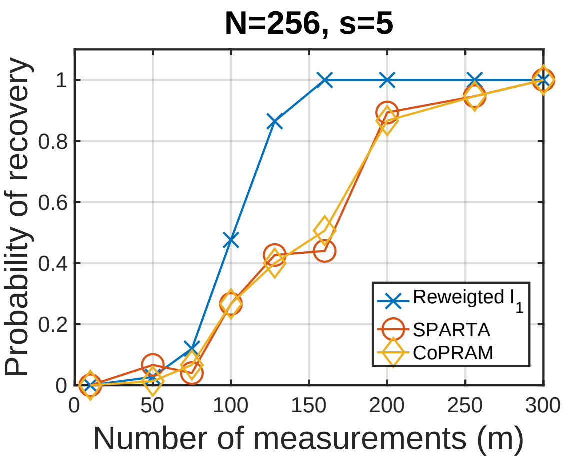


Figure 5.4. Comparison of probability of success vs number of measurements m. Here, $N = 256$ and $s = 5$.

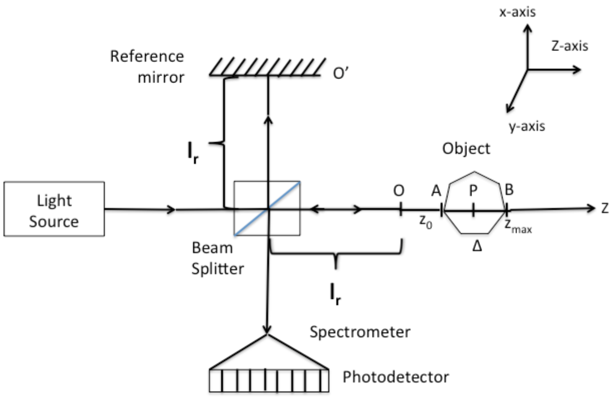


Figure 5.5. Michelson Interferometric Setup for FDOCT

of the entire 1-D image without any mechanical scanning parts, and has better dynamic range compared to TDOCT [169]. The central goal of FDOCT is to reconstruct the underlying scattering characteristic of the object only from the *magnitude of Fourier measurements* recorded by the FDOCT detector. This makes FDOCT a classical Phase Retrieval (PR) problem which arises in a wide range of imaging applications such as crystallography [170], holography [171], and electron microscopy [172]. The problem of phase retrieval originated in optics, and early works by [173], [174] proposed iterative algorithms to recover the phase by imposing suitable constraints. There has been growing interest in phase retrieval problems in recent times, driven by the success of compressed sensing and sparsity enforcing algorithms. Since phase retrieval is a non-linear problem, a significant body of work attempts to linearize it by using the so-called "lifting" technique, and cast it as a low rank matrix recovery problem that can be solved using Semi-Definite Programming (SDP) [147] [175] [176]. Since lifting increases the dimension of the problem, the number of measurements are typically suboptimal and one requires $O(s^2 \log N)$ measurements to recover an N-dimensional signal with s non-zero elements. On the other hand, fast algorithms such as GESPAR [177] have been proposed, which use a greedy approach to solve a quadratic least square problem with sparsity constraints. However, the number of measurements required by GESPAR for recovering is $O(N)$, which is again suboptimal for sparse signals. Another popular approach for solving the Fourier phase retrieval problem is to generate additional sets of measurements using masks [178] [179], optical gratings [180], oblique illumination [181], and Short Time Fourier Transform (STFT) which allows overlap between two consecutive windows to be able to recover the signal. However, the measurement models for these methods cannot be directly applied to FDOCT.

The problem of phase retrieval specific to FDOCT has been studied in [182], [183]. To further improve the performance of FDOCT, a sparsity driven Fienup-type iterative algorithm was proposed in [184]. However, there are no theoretical results that specify the number of measurements needed for this algorithm to succeed.

Summary of Contribution: In this chapter, for the first time, we will develop a sparse phase retrieval algorithm for FDOCT that *provably* recovers the desired signal using minimal number of measurements. Our approach uses the differential FDOCT (or dFDOCT) [185] setup, and acquires compressive measurements to reduce the number of samples. We then develop and analyze an l_1 minimization based reconstruction algorithm for dFDOCT which allows perfect recovery of the underlying signal with $O(s \operatorname{poly}\log N)$ measurements. Simulation results show that our algorithm outperforms existing sparse Phase Retrieval algorithms even for large signal length (N) and sparsity (s).

5.4.1 Review of FDOCT: Measurement Model and Reconstruction Techniques

Measurement Model

In FDOCT, the measurements are acquired using a Michelson Interferometer as shown in Figure 5.5 [183]. The interferometer consists of a *reference-arm* containing a broadband mirror, and an *object-arm* which captures light scattered from the object of interest. Light from a broadband source is split into two beams and channeled towards the reference and object arms. The light reflected from the mirror serves as the reference signal. It is coupled with the light from the object arm using a fiber coupler, and the combined signal is analyzed by a spectrometer. The signal corresponding to the object consists of many elementary waves scattered from different depths of the object along the z-axis axis [186]. Let $a(z)$ denote the amplitude of the light field scattered by object as a function of the depth z. The spectrometer measurement $I(k)$, as a function of the wavenumber $k = 2\pi/\lambda$ (where λ is the wavelength), is given by [187]

$$I(k) = S(k)\left| a_R e^{j2kl_r} + \int_{-\infty}^{\infty} a(z) e^{j2k(l_r + n(z)z)} dz \right|^2 \tag{5.20}$$

Here, $S(k)$ is the power spectrum of the incident light, a_R is the amplitude of the light reflected from the mirror, $2l_r$ is the path length for the reference arm, and $n(z)$ is the refractive index of the object as a function of depth z. The quantity $S(k)$ is typically known apriori [182]. We can also assume that $a_R = 1$ without loss of generality, and simplify (5.20) by approximating $n(z)$ as $n(z) = n$ (also known as a zeroth-order approximation, used for broadband light source) [183]. This yields

$$I(k) = \left| 1 + \int_{-\infty}^{\infty} a(z) e^{j2knz} dz \right|^2 \tag{5.21}$$

Brief Review of Reconstruction Techniques

The central goal in FDOCT is to reconstruct $a(z)$ given the spectral measurements $I(k)$. Notice that $I(k)$ can be rewritten as $I(k) = \left|1 + A(k)\right|^2$ where $A(k) = \int_{-\infty}^{\infty} a(z) e^{j2knz} dz$ is the Fourier transform of the scattering amplitude of the object, also known as the Müller fringe. Traditional techniques attempt to reconstruct $a(z)$ by computing the Fourier transform of $I(k)$, denoted as $\hat{I}(z)$. In this case, $\hat{I}(z)$ consists of three terms

$$\hat{I}(z) = \delta(z) + a(z) + a^*(-z) + r_{aa}(z)$$

where $r_{aa}(z)$ is the autocorrelation function of $a(z)$. Hence, a direct Fourier inversion creates artifacts due to superposition of the autocorrelation $r_{aa}(z)$ on the signal of interest $a(z)$. To address this issue, a common assumption used in existing literature is that the light scattered from the object is sufficiently weaker than the reference. This allows one to effectively ignore $r_{aa}(z)$ and ensure exact reconstruction using different techniques based on homomorphic signal processing, and finite rate of innovation [183], [182]. Alternatively, the authors in [185] proposed a differential Fourier domain method, called dFDOCT, which can completely remove the effect of $r_{aa}(z)$ without any assumptions on the strength of $a(z)$.

Differential FDOCT

The key idea in dFDOCT is to acquire an additional set of measurements $I'(k)$ by adding a phase difference of π in the reference path (e.g. by using a phase modulator). In particular,

$$I'(k) = \left| -1 + \int_{-\infty}^{\infty} a(z) e^{j2knz} dz \right|^2 \tag{5.22}$$

In a practical scenario, $a(z)$ can safely be assumed to be real valued as it's the amplitude of the scattered light field. Hence, In the rest of the chapter we would treat $a(z)$ to be real. After making a substitution $\omega = -2kn$ and subtracting (5.22) from (5.21), we obtain the differential measurement

$$\Delta I(\omega) \triangleq I(\omega) - I'(\omega) = 2\Big(\int_{-\infty}^{\infty} \Big(a(z) + a(-z)\Big) e^{-j\omega z} dz\Big) \tag{5.23}$$

It can be readily seen that the $\Delta I(\omega)$ does not contain the autocorrelation term, and is simply the scaled Fourier transform of the signal $a(z) + a(-z)$. Since $a(z)$ is causal (i.e. $a(z) = 0, z < 0$), we can exactly reconstruct $a(z)$ simply by computing the Inverse Fourier transform of $\Delta I(\omega)$. Thus, dFDOCT offers an elegant way to image the object free from autocorrelation-induced artifacts, without any compromise on accessible depth or resolution requirement [185]. However, the main disadvantage is that we require twice the number of measurements compared to standard FDOCT. In this chapter, we will overcome this shortcoming by exploiting sparsity of the desired image.

5.4.2 Compressive Differential FDOCT with minimal measurements

The role of sparsity in Fourier based phase retrieval is an active area of research. However, sparsity enforcing algorithms are either sub-optimal in terms of the number of required measurements, or they require additional assumptions on the signal and measurement system, which may not be easy to enforce on the signal acquisition setup for FDOCT. On the other hand, iterative algorithms for reconstructing sparse signals from their FDOCT measurements, were developed in [184]. However, no guarantees exist in terms of the required number of measurements.

We will now show how the measurement model for differential FDOCT can be used to exploit the sparsity of $a(z)$, and develop algorithms that can provably reconstruct it with minimal number of measurements. Consider discrete measurements $\Delta I(\omega)$, denoted by $\Delta I[n] = \Delta I(n\Delta\omega)$ where $\Delta\omega$ is the sampling step size. Assuming $a(z)$ is compactly supported, that is, $a(z) =$

$0 \; \forall \; z < 0 \; \& \; z > z_{max}$, we can choose $\Delta\omega(\leq \frac{\pi}{z_{max}})$ such that

$$\sum_{n=-\infty}^{\infty} \Delta I[n] e^{jz\Delta\omega n} = 2\tilde{a}(z)$$

where

$$\tilde{a}(z) = \begin{cases} 2a[0] & z = 0 \\ a(z) & 0 < z \leq \pi/\Delta\omega \\ a(2\pi/\Delta\omega - z) & \pi/\Delta\omega \leq z < 2\pi/\Delta\omega \end{cases}$$

Consider $2N-1$ samples $\Delta I[n], n = 0, 1, \cdots, 2N-2$. The samples of $\tilde{a}(z)$, denoted by $\tilde{a}[k] = \tilde{a}(2\pi k/((2N-1)\Delta\omega))$ are related to $\Delta I[n]$ as

$$\sum_{k=0}^{2N-2} 2\tilde{a}[k] e^{-j2\pi kn/(2N-1)} = \sum_{p=-\infty}^{\infty} \Delta I[n + pN]$$

Assuming that N is large enough so that $\Delta I[n]$ is negligible for $n \geq 2N-1$, we have, for $0 \leq n \leq 2N-2$,

$$\Delta I[n] = \sum_{k=0}^{2N-2} 2\tilde{a}[k] e^{-j2\pi kn/(2N-1)} \tag{5.24}$$

Hence, in order to reconstruct N samples of $a(z)$ at a resolution of $2\pi/\Delta\omega$, we need to acquire $2N-1$ differential measurements $\Delta I[n]$ (which implies a total of $4N-2$ spectral measurements), and only retain the first N values of its IDFT.

Compressive d-FDOCT exploiting sparsity of $a(z)$

Since most images are sparse over suitable basis, it is natural to assume that $\mathbf{a} = \Big[a[0], a[1], \cdots, a[N-1]\Big]^T$ has a sparse representation over a basis $\Psi \in \mathbb{C}^{N \times N}$ as

$$\mathbf{a} = \Psi \mathbf{x}_0$$

where $\mathbf{x}_0 \in \mathbb{C}^N$ is a sparse vector with $s_0 \ll N$ non zero elements. This is a widely used assumption in sparse phase retrieval [165, 177, 184] where the object being imaged by the optical setup can be assumed to have sparse representation over a suitable choice of finite basis (DCT, Wavelets etc.). Defining $2\mathbf{y} = \Big[\Delta I[0], \Delta I[1], \cdots, \Delta I[2N-1]\Big]$, we can rewrite (5.24) as

$$\mathbf{y} = \mathbf{W}\bar{\Psi}\mathbf{x}_0 \tag{5.25}$$

Here $\mathbf{W} \in \mathbb{C}^{(2N-1)\times(2N-1)}$ is a $(2N-1)$ point DFT matrix, and $\bar{\Psi} \in \mathbb{C}^{(2N-1)\times N}$ is given by

$$\bar{\Psi} = [2\psi_0, \psi_1, \cdots, \psi_{N-1}, \psi_{N-1}, \psi_{N-2}, \cdots \psi_1]^T$$

where $\psi_m^T \in \mathbb{C}^{1\times N}$ denotes the mth row of Ψ. Instead of directly computing the inverse DFT of $\mathbf{y}$ to recover $\mathbf{a}$ (which, as shown earlier, would require $2N-1$ differential measurements), we propose to exploit the sparsity of $\mathbf{x}$ to significantly reduce the number of differential measurements. We can directly sample the signals $I(\omega)$ and $I'(\omega)$ at the output of the spectrometer in differential FDOCT, to obtain $2M$ measurements $I[k], k = 0, 1, \cdots M-1$ and $I'[k]$ where

$$I[k] = I(m_k 2\pi/\Delta\omega), \quad I'[k] = I'(m_k 2\pi/\Delta\omega), \quad k = 0, 1, \cdots, M-1$$

Here, m_k denote the sampling locations (using a step size of $2\pi/\Delta\omega$), which are integers in the range $0 \leq m_k \leq 2N-2$. Let $\Omega = \{m_k, 0 \leq k \leq M-1\}$ denote the set of sampling indices. Using 5.25, the differential measurements $\Delta I[k], k = 0, 1, \cdots, M-1$ can be equivalently represented as

$$\mathbf{y}_\Omega = \mathbf{W}_\Omega \bar{\Psi} \mathbf{x}_0 \tag{5.26}$$

where $\mathbf{W}_\Omega \in \mathbb{C}^{M \times (2N-1)}$ represents a subset of M rows of $\mathbf{W}$, indexed by Ω. The goal is to ensure reconstruction of sparse $\mathbf{x}$ with far fewer measurements than its ambient dimension N. To this end, we propose to recover $\mathbf{x}_0$ by solving the following l_1 minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{x}\|_1 \quad (P1) \tag{5.27}$$

$$\text{subject to} \qquad \mathbf{y}_\Omega = \mathbf{W}_\Omega \bar{\Psi} \mathbf{x} \tag{5.28}$$

If $\mathbf{x}_0^*$ is the solution to (P1), the final reconstructed signal is given by $\mathbf{a}^* = \Psi \mathbf{x}_0^*$

Exact Recovery with Minimal Number of Measurements

The number of measurements M used in our proposed dFDCT based approach can indeed be made significantly smaller than N, especially when $\mathbf{x}$ is sufficiently sparse, without compromising the performance. To prove this, we invoke the following theorem from [8]

Theorem 15. *Let $\mathbf{U} \in \mathbb{C}^{N \times N}$ be a unitary matrix bounded entries satisfying $|\mathbf{U}_{m,n}| \leq K/\sqrt{N}$ where K is a constant independent of N. Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ be a submatrix of U obtained by selecting a subset of M rows uniformly at random. If*

$$M \geq CK^2 s \Big(\log(N) \Big)^4 \tag{5.29}$$

where C is a universal constant, then, with probability at least $1 - N^{-log^3(N)}$, every sparse vector $\mathbf{x}_0 \in \mathbb{C}^N$ with s non zero elements, is the unique minimizer of the problem:

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1, \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}_0.$$

When the vector $\mathbf{a}$ is naturally sparse, i.e. $\Psi = \mathbf{I}$, Theorem 15 can be directly used to determine the minimum number of measurements that ensures perfect reconstruction with high probability (that tends to 1 exponentially with N) for our approach. In this case, the problem (P1) is equivalent to

$$\min_{\mathbf{x} \in \mathbb{R}^{2N-1}} \|\mathbf{x}\|_1, \quad \text{s. t. } \mathbf{A}\bar{\mathbf{x}}_{\mathbf{0}} = \mathbf{A}\mathbf{x} \quad (P2) \tag{5.30}$$

where $\mathbf{A} = \frac{1}{\sqrt{2N-1}}\mathbf{W}_{\Omega}$ consists of M rows of $\mathbf{U} = \frac{1}{\sqrt{2N-1}}\mathbf{W}$ indexed by Ω. The matrix $\mathbf{U}$ is unitary and satisfies $|\mathbf{U}_{m,n}| \leq \frac{1}{\sqrt{2N-1}}$. The vector $\bar{\mathbf{x}}_0 = [2\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_{N-1}, \mathbf{a}_{N-1}, \mathbf{a}_{N-2}, \cdots, \mathbf{a}_1]^T$ is sparse with no more than $2s_0$ non-zero elements. It is clear that $\mathbf{a}$ can be reconstructed once $\bar{\mathbf{x}}_0$ has been recovered by solving (P2) (or equivalently, P1). The following Corollary to Theorem 15 provides theoretical guarantees for our proposed approach in terms of the number of measurements M:

Corollary 1. *Consider the proposed measurement model (5.26) where the entries of Ω are selected uniformly at random from $[0, 2N-2]$, and the signal $\mathbf{a}$ is naturally sparse, i.e., $\Psi = \mathbf{I}$. If*

$$M \geq 2Cs_0\Big(\log(2N-1)\Big)^4$$

then, with probability at least $1 - (2N-1)^{-log^3(2N-1)}$, $\mathbf{a}$ can be uniquely recovered by solving (P1).

The above result has the following implications, both for FDOCT as well as for the more general problem of sparse phase retrieval:

1. **Exact FDOCT with minimal measurements:** Our results guarantee exact reconstruction of the desired image using the FDOCT experimental setup, without any approximations (such as neglecting the autocorrelation function $r_{aa}(z)$) or introducing an offset of the zero-phase plane [183]. It also significantly improves the performance of dFDOCT by

enabling exact reconstruction without doubling the number of measurements. Finally, unlike many existing FDOCT algorithms, we are able to provide the exact number of measurements (as a function of N and sparsity s_0) that can provably recover the desired image.

2. **Sparse Fourier phase retrieval with $O(s \text{ poly} \log N)$ measurements:** Although powerful algorithms such as GESPAR and those based on convex relaxations (using the idea of lifting) have proved to be effective for Fourier phase retrieval, they typically require much larger number of measurements (which can be $O(N)$ or $O(s^2 \log N)$) and cannot ensure perfect reconstruction with $O(s \log N)$ measurements. The authors in [165] have shown that perfect recovery is possible with just $O(s \log(N/s))$ by restricting the measurement vectors to an incoherent subspace. However, such measurement schemes may not be physically realizable by an optical setup. Recently, [166] also shows that $O(s \log(N/s))$ measurements are sufficient for PR using an appropriate two-step algorithm. In contrast, our proposed approach with the differential measurement shows that it is possible to perform *Fourier phase retrieval with $O(s$ poly $\log N)$* measurements using l1 minimization and is also physically realizable with FDOCT experimental setup.

5.4.3 Simulations

In order to show the effectiveness of the proposed algorithm (PA) for Sparse FDOCT, we compare it with two other sparsity PR algorithms: GESPAR [177] and Max-K algorithm [184] where the latter was specifically developed for FDOCT. We compare the performance of these three algorithms in terms of the probability of successful recovery for various sparsity levels, signal length and also study the effect of exact knowledge of sparsity. In the first experiment, we generate a $k-$sparse signal $\mathbf{x}$ whose non-zero entries are sampled from a zero mean and unit variance Gaussian distribution. For each algorithm, we compute the probability of recovering the sparse signal $\mathbf{x}$. The recovery is declared successful if the normalized mean squared error (NMSE) satisfies $\frac{||\hat{x}-x||_2}{||x||_2} \leq 0.001$. We choose $N = 256$ and study the probability of successful

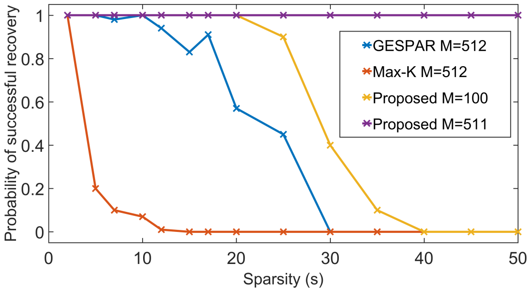


Figure 5.6. Probability of successful recovery vs Sparsity for the PA, GESPAR and Max-K. Here, $N = 256$ and both GESPAR and Max-K use $M = 2N = 512$ measurements.

recovery as a function of the sparsity s. For a fixed sparsity s, we randomly generate a signal x 100 times, each time with a random support and populated by random values. The length of this signal x is set to $N = 256$ and we record the number of times this signal is recovered successfully out of the 100 trials. We use $M = 2N = 512$ measurements for GESPAR and Max-K, whereas we test the PA for two values of $M : 511$ and 100. Figure 5.6 shows the probability of successful recovery for all the three algorithms. It is clear that the proposed algorithm outperforms the other two methods. Even when it uses fewer measurements ($M = 100$) compared to GESPAR and Max-K (each of which uses $M = 512$ measurements), the PA allows perfect reconstruction upto a sparsity of $s = 25$ whereas the performance of GESPAR and Max-K starts to deteriorate at $s = 17$ *and* 5 respectively. It is to be noted that the exact knowledge of s was provided to both GESPAR and Max-K algorithms since they are sensitive to the knowledge of sparsity. However, our algorithm does not require to know s apriori. Also, the solutions of both GESPAR and Max-K algorithms have trivial ambiguities such circular shift, mirroring and sign reversal. The NMSE for both algorithms is computed *after compensating for these ambiguities* by searching for the minimum NMSE over all possible transformations of the recovered signal. However, our algorithm is free from such ambiguities and does not require any post-processing.

In the second experiment, we compare the quality of reconstruction of the three algorithms for a 2-D synthetic sparse image which is created following the same approach as [177]. Figure 5.7 shows the images reconstructed by the three algorithms under different settings. It

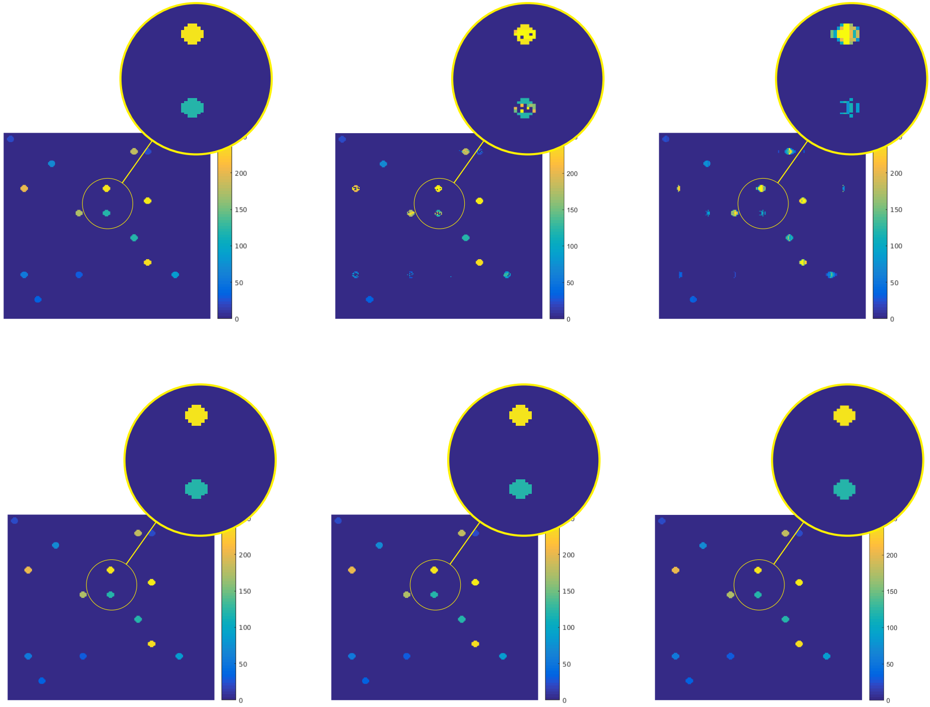


Figure 5.7. (Top Left) Ground Truth $2D$ synthetic image (Top Middle) GESPAR with upper limit of $s = 10$ and $M = 512$ (Top Right) Max-K, with $M = 512$ (Bottom Left) PA with $M = 511$ (Bottom Middle) GESPAR with knowledge of exact sparsity (Bottom right) PA with $M = 200$

can be seen that both the PA & GESPAR can exactly recover the image but the image recovered by the Max-K algorithm has some undesirable artifacts. On the other hand, if we only provide an upper limit on the sparsity s to GESPAR, it also exhibits artifacts which is also evident from an increase in the NMSE from 2.4×10^{-4} to 0.3227. However, the most important metric where the PA significantly outperforms both these algorithms is the number of measurements. In spite of reducing the number of measurements to only $M = 200$ the PA still maintains NMSE of 8.45×10^{-9}. GESPAR and Max-K both use $2N = 512$ measurements but our algorithm is able to recover the image with a total of $M = 200$ measurements, with no deterioration in the quality of reconstruction.

5.4.4 Conclusion

In the first half of the chapter, we considered the problem of blind deconvolution with autoregressive filters when we have access to downsampled measurements at the output of the filter. We leverage positivity constraint on the input signal and the structure of the filter to derive guarantees for unique identification of the signal and the filter. Our simulations demonstrate that non-negative constraints can significantly improve the ability to recover signals with larger sparsity. In future, it will be an interesting direction to extend our analysis for noisy measurements and use efficient decoding strategies for recovering the spikes after identifying the kernel.

In the second chapter, first we show that it is possible to achieve optimal sample complexity of $\Omega(s \log n/s)$ for sparse phase retrieval in the lifted space using only sparsity constraint and eliminating low rank constraint, which is the first result of its kind. Inspired by the power of the diagonal sparsity constraint, we propose an iterative reweighted algorithm based on Phaselift which is also initialization free. The numerical experiments demonstrate that our algorithm significantly outperforms existing algorithms for sparse phase retrieval.

Finally, we proposed and analyzed a new compressive (Fourier) phase retrieval approach based on differential FDOCT (dFDOCT) that can provably recover sparse signals from phaseless measurements with minimal number of measurements. Our simulations establish superior performance of our algorithm compared to existing sparse phase retrieval techniques such as GESPAR and Max-K algorithm. Our method also overcomes an apparent drawback of dFDOCT by posing the recovery problem in a compressed setting, and significantly reducing the number of required measurements.

Chapter 5, in part, is a reprint of the material as it appears in the following papers:

- P. Sarangi, M. C. Hücümenoglu and P. Pal, “Effect of Undersampling on Non-Negative Blind Deconvolution with Autoregressive Filters,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 5725-5729.

- P. Sarangi, M. C. Hücümenoglu and P. Pal, ”Understanding Sample Complexities for Structured Signal Recovery from Non-Linear Measurements,” 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Le gosier, Guadeloupe, 2019, pp. 81-85.

- P. Sarangi, H. Qiao and P. Pal, “On the role of sampling and sparsity in phase retrieval for optical coherence tomography,” 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Curacao, 2017, pp. 1-5.

The dissertation author was one of the primary investigator and author of these papers.

Chapter 6

Open Questions and Emerging Directions

In this chapter, we discuss connections between the ideas explored in this thesis to problems in other application domains, and indicate future research directions.

6.1 Binary Super-resolution: Interference Channel Perspective

An important class of channel model known as the "K-user interference channel" has been widely studied in wireless communication [188,189]. Mathematically, a K-user interference channel consisting of K single antenna transmitters and M single antenna receivers is given by:

$$y_m = \sum_{k=1}^{K} h_{m,k} x_k + w_m, \quad m = 1, 2, \cdots, M \tag{6.1}$$

Here $h_{m,k}$ is the channel between the k^{th} user and m^{th} receiver and w_m is an additive noise term. In communication systems, it is common to restrict the inputs $\{x_k\}_{k=1}^{K}$ to finite sets (constellation). Typically $M = K$ and the k^{th} receiver is interested in only decoding the input x_k (from the k^{th} user). There has been tremendous efforts in understanding the information theoretic properties such as capacity (by characterizing the DOFs) of these interference channels [188]. The "interference alignment" scheme has emerged as a practical method that is capable of achieving the available DOF in the high SNR regime [188, 189].

Recall from Chapter 2 (Section 2.5), the fundamental problem in Binary super-resolution is of the form:

$$y_m = \sum_{k=1}^{\mathrm{D}} \alpha^{\mathrm{D}-k} x_k + w_m$$

This model closely resembles the interference channel introduced earlier, however, the key distinction lies in the fact that the goal here is to decode (from the single measurement y_m) the messages sent by all D-users, as opposed to only the m^{th} message. Due to the finite-valued nature of the input, the decoding strategies proposed in this thesis can be useful for decoding all the messages simultaneously, especially when the number of receivers is significantly smaller than the number of transmitting users. An interesting future direction would be to explore the similarities with the interference channel to characterize the information-theoretic properties of the channel for binary super-resolution. This analysis can reveal insights into channels that are more benign for the task of super-resolution. In the context of the calcium imaging application discussed in Chapter 2, this can translate into design of new calcium indicators which determine the "effective" channel parameters.

6.2 Characterizing Minimum Distance: Diophantine Approximation Problem

In Chapter 2 (Section 2.4.2), characterizing the robustness performance of the super-resolution problem involved computing a certain "minimum-distance" ($\Delta\theta_{\min}(\alpha, \mathrm{D})$). A question that remains to be answered is the full characterization of the behavior of this minimum distance as a function of the filter parameter α and downsampling factor D. For a fixed D, there are a total of $3^{\mathrm{D}} - 1$ ternary polynomials $p_i(\alpha)$. We can characterize $\Delta\theta_{\min}(\alpha, \mathrm{D})$ by alternatively viewing it as finding the minimum out of these $3^{\mathrm{D}} - 1$ polynomials (in variable α):

$$\Delta\theta_{\min}(\alpha, \mathrm{D}) = \min_{1 \leq i \leq 3^{\mathrm{D}}-1} |p_i(\alpha)| = \min_{\mathbf{v}_i \in \{-1,0,1\}^{\mathrm{D}} \setminus \{\mathbf{0}\}} |\mathbf{h}_\alpha^\top \mathbf{v}_i|, \text{ where } \mathbf{h}_\alpha^\top = [\alpha^{\mathrm{D}-1}, \alpha^{\mathrm{D}-2}, \cdots, 1]$$

Note that for a fixed α, any one of those $3^{\mathrm{D}-1}$ polynomial can attain the minima, the key challenge lies in identifying which polynomial p_i is the minimizer at a given α. In the regime $\alpha \leq 0.5$, we were able to analytically characterize this minimum distance. It turns out the polynomial with coefficient vector $[1,0,0,\cdots,0]$ always attains the minima in this regime. However, an analytical solution for this minimization problem for $0.5 < \alpha \leq 1$ seems to be more challenging. Instead, we can aim to obtain insightful lower bounds as a function of α, D to analyze the noise robustness of decoding. It turns out that this question has close ties to the problem of *"Diophantine Approximation"*, which deals with the question of approximating real numbers using rational numbers. A potential tool for addressing this problem is leveraging results in the Diophantine Approximation literature concerned with characterizing the closeness of points to rational hyperplanes. Let $\psi : \mathbb{N} \to \mathbb{R}^+$. Given $\mathbf{q} \in \mathbb{Z}^n$ and $p \in \mathbb{Z}$, the system of equations $\mathbf{q}^\top \mathbf{x} = p$ is a rational hyperplane. A point $\mathbf{z} \in \mathbb{R}^n$ is called "dually ψ-approximable" if the following inequality

$$|\mathbf{q}^\top \mathbf{z} - p| < \psi(\|\mathbf{q}\|_\infty) \tag{6.2}$$

holds for infinitely many $(p, \mathbf{q}) \in \mathbb{Z} \times \mathbb{Z}^n$. The properties of $\psi, \mathbf{z}$ under which such approximation is possible is given by the Khintchine's theorem [190] and its extension by Groshev [191]. However, in our case, we are not interested in a generic point $\mathbf{z}$, but rather points specifically of the form $\mathbf{z} = (1, \alpha, \alpha^2, \cdots, \alpha^{n-1})$.

6.3 From Measurement-Algorithm Co-Design to Codebook Design for Unsourced Random Access

The current and future generations of wireless networks are expected to support massive machine-type communications due to the growing number of Internet of Things (IoT) devices. This has resulted in the "unsourced massive random access" channel model where a large number of total users communicate using a single shared codebook and only a subset of the users are assumed to be active in a given time-slot [18]. This problem can be modeled as follows:

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{a}_i q_i + \mathbf{n}$$

where $\mathbf{a}_i \in \mathscr{C} \subset \mathbb{R}^M$ is a message from the shared codebook $\mathscr{C}$, $q_i \in \{0,1\}$ is a binary signal that indicates whether the message $\mathbf{a}_i$ is transmitted by any of the active users and $\mathbf{n}$ is an additive noise term. Given the received signal $\mathbf{y}$, the goal of the decoder is to return a list of messages (columns of the codebook) that were transmitted from the received signal . This amounts to recovering a binary vector $\mathbf{q}$ from (noisy) observation $\mathbf{y}$. Recently, sparse superposition codes or sparse regression codes (SPARCs) [192–194] have been used along with either tree-based algorithms or the Approximate Message Passing (AMP) algorithm. The insights from Chapter 3 can be leveraged to design a shared structured codebook (with a partial convolutional structure) using the proposed measurement-algorithm co-design framework. Such a codebook would allow us to deploy a low-complexity sequential decoding algorithm. Furthermore, the ability to operate in the extreme compression regime directly translates to the ability to *support many more users* using a constant number of channel uses, i.e., $M = \Omega(1)$.

6.4 Binary Priors for Non-Linear Measurement Model: Finding Quantized Neural Networks

Deep neural networks have become widely used for a variety of machine intelligence tasks such as object detection, speech recognition and many more. Deploying neural networks in low-power hardware platforms, such as mobile devices, has been challenging due to the large number of parameters and massive number of multiply-accumulate operations required even in the inference stages. This has led to research on deploying of neural networks using low-precision parameters without significantly sacrificing performance. Such solutions become desirable due to their lower memory and computational footprint compared to full-precision networks. A common approach is to quantize a pre-trained model (with full-precision) followed by a refinement stage [195–197]. The benefits of finite-valued priors for linear inverse problems advocated in this thesis opens up research problems for transferring the recovery techniques for linear inverse problems to efficiently solve a non-linear inverse problem such as training a quantized neural network. In the training phase, our goal is to find the parameters θ of the neural network, which belongs to a finite set $\mathscr{A}$ whose cardinality is determined by the desired level of quantization. Given training data points $\{\mathbf{x}_i, y_i\}_{i=1}^{M}$, the objective is to minimize the following combinatorial optimization problem:

$$\min_{\theta \in \mathscr{A}^K} \frac{1}{M} \sum_{i=1}^{M} (y_i - f_\theta(\mathbf{x}_i))^2 \tag{6.3}$$

An interesting question would be to formulate a sequential (greedy) algorithm to solve the above optimization problem by leveraging structure of the neural network architecture and choice of non-linearity.

6.5 Biased Subspace Estimation: Data-Starved Regime

In chapter 4, we addressed the problem of DOA estimation using deterministic sparse arrays when the number of snapshots is limited. The central idea was to move away from using the sample covariance matrix and instead use the available snapshots to obtain a biased estimate of the covariance matrix preserving the subspace information. We can generalize this idea to adopt a data-driven framework that can learn the biased estimate specific to the task and scenario (SNR and number of snapshots that are available) at hand. This gives rise to several interesting questions regarding how to formulate the “bias learning” problem in an effective manner. Ideally, the estimator should also be asymptotically unbiased in the number of snapshots, i.e., the bias should go to zero as we have more and more snapshots.

6.6 Robustness Guarantees for Interpolation

In chapter 4, we also provided guarantees for noiseless interpolation using the deterministic nested array. A natural question would be to analyze the robustness properties of this interpolation framework and characterize the resolution limits. In particular, how the resolution is affected by the array geometry and the SNR. The considered interpolation framework attempts to leverage single-snapshot DOA estimation schemes that are already known for the ULA, such as spatial-smoothing. However, synthesizing the virtual array using rank-minimization based interpolation can be computationally cumbersome. As a result, an important research direction that emerges is to explore schemes that have a lower computational complexity. A potential way to achieve this is by designing algorithms that work on the sparse array measurements directly and possibly avoid interpolation altogether.

Bibliography

[1] P. Pal and P. Vaidyanathan, “Pushing the limits of sparse support recovery using correlation information,” *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 711–726, 2014.

[2] H. Qiao and P. Pal, “Guaranteed localization of more sources than sensors with finite snapshots in multiple measurement vector models using difference co-arrays,” *IEEE Transactions on Signal Processing*, vol. 67, no. 22, pp. 5715–5729, 2019.

[3] A. Koochakzadeh, H. Qiao, and P. Pal, “On fundamental limits of joint sparse support recovery using certain correlation priors,” *IEEE Transactions on Signal Processing*, vol. 66, no. 17, pp. 4612–4625, 2018.

[4] P. K. Kota, D. LeJeune, R. A. Drezek, and R. G. Baraniuk, “Extreme compressed sensing of poisson rates from multiple measurements,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 2388–2401, 2022.

[5] A. Rényi, “Representations for real numbers and their ergodic properties,” *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 8, no. 3-4, pp. 477–493, 1957.

[6] P. Glendinning and N. Sidorov, “Unique representations of real numbers in non-integer bases,” *Mathematical Research Letters*, vol. 8, no. 4, pp. 535–543, 2001.

[7] N. Sidorov, “Almost every number has a continuum of β-expansions,” *The American Mathematical Monthly*, vol. 110, no. 9, pp. 838–842, 2003.

[8] S. Foucart, H. Rauhut, S. Foucart, and H. Rauhut, *A Mathematical Introduction to compressive sensing*. Springer, 2013.

[9] R. Brette and A. Destexhe, *Handbook of neural activity measurement*. Cambridge University Press, 2012.

[10] J. T. Vogelstein, B. O. Watson, A. M. Packer, R. Yuste, B. Jedynak, and L. Paninski, “Spike inference from calcium imaging using sequential monte carlo methods,” *Biophysical journal*, vol. 97, no. 2, pp. 636–655, 2009.

[11] T. Deneux, A. Kaszas, G. Szalay, G. Katona, T. Lakner, A. Grinvald, B. Rózsa, and I. Vanzetta, “Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo,” *Nature communications*, vol. 7, p. 12190, 2016.

[12] S. Yang and L. Hanzo, “Fifty years of mimo detection: The road to large-scale mimos,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 1941–1988, 2015.

[13] W. Dai and O. Milenkovic, “Weighted superimposed codes and constrained integer compressed sensing,” *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2215–2229, 2009.

[14] J. Nemeth and P. Balazs, “Restoration of blurred binary images using discrete tomography,” in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2013, pp. 80–90.

[15] K. J. Batenburg and J. Sijbers, “Dart: a practical reconstruction algorithm for discrete tomography,” *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2542–2553, 2011.

[16] K.-K. Wong, A. Paulraj, and R. D. Murch, “Efficient high-performance decoding for overloaded mimo antenna systems,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1833–1843, 2007.

[17] R. Hayakawa and K. Hayashi, “Convex optimization-based signal detection for massive overloaded mimo systems,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7080–7091, 2017.

[18] Y. Polyanskiy, “A perspective on massive random-access,” in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2523–2527.

[19] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[20] W. Liao and A. Fannjiang, “Music for single-snapshot spectral estimation: Stability and super-resolution,” *Applied and Computational Harmonic Analysis*, vol. 40, no. 1, pp. 33–67, 2016.

[21] R. Roy and T. Kailath, “Esprit-estimation of signal parameters via rotational invariance techniques,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[22] W. Li, W. Liao, and A. Fannjiang, “Super-resolution limit of the esprit algorithm,” *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4593–4608, 2020.

[23] Y. Hua and T. K. Sarkar, “Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 5, pp. 814–824, 1990.

[24] D. Batenkov, G. Goldman, and Y. Yomdin, “Super-resolution of near-colliding point sources,” *Information and Inference: A Journal of the IMA*, vol. 10, no. 2, pp. 515–572, 2021.

[25] D. L. Donoho, “Superresolution via sparsity constraints,” *SIAM journal on mathematical analysis*, vol. 23, no. 5, pp. 1309–1331, 1992.

[26] E. J. Candès and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” *Communications on pure and applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[27] B. Bernstein and C. Fernandez-Granda, “Deconvolution of point sources: a sampling theorem and robustness guarantees,” *Communications on Pure and Applied Mathematics*, vol. 72, no. 6, pp. 1152–1230, 2019.

[28] A. Koulouri, P. Heins, and M. Burger, “Adaptive superresolution in deconvolution of sparse peaks,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 165–178, 2020.

[29] B. N. Bhaskar, G. Tang, and B. Recht, “Atomic norm denoising with applications to line spectral estimation,” *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5987–5999, 2013.

[30] Y. Chi and M. F. Da Costa, “Harnessing sparsity over the continuum: Atomic norm minimization for superresolution,” *IEEE Signal Processing Magazine*, vol. 37, no. 2, pp. 39–57, 2020.

[31] T. Bendory, “Robust recovery of positive stream of pulses,” *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 2114–2122, 2017.

[32] G. Schiebinger, E. Robeva, and B. Recht, “Superresolution without separation,” *Information and Inference: A Journal of the IMA*, vol. 7, no. 1, pp. 1–30, 2017.

[33] V. I. Morgenshtern and E. J. Candes, “Super-resolution of positive sources: The discrete setup,” *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 412–444, 2016.

[34] D. Batenkov, A. Bhandari, and T. Blu, “Rethinking super-resolution: the bandwidth selection problem,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5087–5091.

[35] M. F. Da Costa and W. Dai, “A tight converse to the spectral resolution limit via convex programming,” in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 901–905.

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

[37] J. A. Urigüen, T. Blu, and P. L. Dragotti, “Fri sampling with arbitrary kernels,” *IEEE Transactions on Signal Processing*, vol. 61, no. 21, pp. 5310–5323, 2013.

[38] J. Onativia, S. R. Schultz, and P. L. Dragotti, “A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging,” *Journal of neural engineering*, vol. 10, no. 4, p. 046017, 2013.

[39] R. Tur, Y. C. Eldar, and Z. Friedman, "Innovation rate sampling of pulse streams with application to ultrasound imaging," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1827–1842, 2011.

[40] S. Rudresh and C. S. Seelamantula, "Finite-rate-of-innovation-sampling-based super-resolution radar imaging," *IEEE Transactions on Signal Processing*, vol. 65, no. 19, pp. 5021–5033, 2017.

[41] M. Stojnic, "Recovery thresholds for l_1 optimization in binary compressed sensing," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1593–1597.

[42] S. Keiper, G. Kutyniok, D. G. Lee, and G. E. Pfander, "Compressed sensing for finite-valued signals," *Linear Algebra and its Applications*, vol. 532, pp. 570–613, 2017.

[43] A. Flinth and S. Keiper, "Recovery of binary sparse signals with biased measurement matrices," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8084–8094, 2019.

[44] S. M. Fosson and M. Abuabiah, "Recovery of binary sparse signals from compressed linear measurements via polynomial optimization," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1070–1074, 2019.

[45] Z. Tian, G. Leus, and V. Lottici, "Detection of sparse signals under finite-alphabet constraints," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 2349–2352.

[46] P. Sarangi and P. Pal, "No relaxation: Guaranteed recovery of finite-valued signals from undersampled measurements," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5440–5444.

[47] ——, "Measurement matrix design for sample-efficient binary compressed sensing," *IEEE Signal Processing Letters*, 2022.

[48] S. Razavikia, A. Amini, and S. Daei, "Reconstruction of binary shapes from blurred images via hankel-structured low-rank matrix recovery," *IEEE Transactions on Image Processing*, vol. 29, pp. 2452–2462, 2019.

[49] J.-H. Lange, M. E. Pfetsch, B. M. Seib, and A. M. Tillmann, "Sparse recovery with integrality constraints," *Discrete Applied Mathematics*, vol. 283, pp. 346–366, 2020.

[50] L. Fukshansky, D. Needell, and B. Sudakov, "An algebraic perspective on integer sparse recovery," *Applied Mathematics and Computation*, vol. 340, pp. 31–42, 2019.

[51] J. Friedrich, P. Zhou, and L. Paninski, "Fast online deconvolution of calcium imaging data," *PLoS computational biology*, vol. 13, no. 3, p. e1005423, 2017.

[52] B. F. Grewe, D. Langer, H. Kasper, B. M. Kampa, and F. Helmchen, "High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision," *Nature methods*, vol. 7, no. 5, p. 399, 2010.

[53] S. W. Jewell, T. D. Hocking, P. Fearnhead, and D. M. Witten, “Fast nonconvex deconvolution of calcium imaging data,” *Biostatistics*, vol. 21, no. 4, pp. 709–726, 2020.

[54] P. Sarangi, M. C. Hücümenoğlu, and P. Pal, “Effect of undersampling on non-negative blind deconvolution with autoregressive filters,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5725–5729.

[55] A. Rupasinghe and B. Babadi, “Robust inference of neuronal correlations from blurred and noisy spiking observations,” in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2020, pp. 1–5.

[56] C. Frougny and B. Solomyak, “Finite beta-expansions,” *Ergodic Theory Dynam. Systems*, vol. 12, no. 4, pp. 713–723, 1992.

[57] C. Poon, “On the role of total variation in compressed sensing,” *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 682–720, 2015.

[58] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, “An introduction to total variation for image analysis,” *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, no. 263-340, p. 227, 2010.

[59] V. Komornik and P. Loreti, “Expansions in noninteger bases.” *Integers*, vol. 11, no. A9, p. 30, 2011.

[60] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.

[61] T.-W. Chen, T. J. Wardill, Y. Sun, S. R. Pulver, S. L. Renninger, A. Baohan, E. R. Schreiter, R. A. Kerr, M. B. Orger, V. Jayaraman *et al.*, “Ultrasensitive fluorescent proteins for imaging neuronal activity,” *Nature*, vol. 499, no. 7458, pp. 295–300, 2013.

[62] H. K. S. c. GENIE Project, Janelia Farm Campus, “Simultaneous imaging and loose-seal cell-attached electrical recordings from neurons expressing a variety of genetically encoded calcium indicators,” *CRCNS. org*, 2015.

[63] T. Zhang, J. M. Pauly, and I. R. Levesque, “Accelerating parameter mapping with a locally low rank constraint,” *Magnetic resonance in medicine*, vol. 73, no. 2, pp. 655–661, 2015.

[64] S. Razavikia, H. Zamani, and A. Amini, “Sampling and recovery of binary shapes via low-rank structures,” in *2019 13th International conference on Sampling Theory and Applications (SampTA)*. IEEE, 2019, pp. 1–4.

[65] N. Alajlan, M. S. Kamel, and G. H. Freeman, “Geometry-based image retrieval in binary image databases,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 6, pp. 1003–1013, 2008.

[66] A. Flinth and G. Kutyniok, “Promp: A sparse recovery approach to lattice-valued signals,” *Applied and Computational Harmonic Analysis*, vol. 45, no. 3, pp. 668–708, 2018.

[67] U. Nakarmi and N. Rahnavard, “Bcs: Compressive sensing for binary sparse signals,” in *MILCOM 2012-2012 IEEE Military Communications Conference*. IEEE, 2012, pp. 1–5.

[68] O. L. Mangasarian and B. Recht, “Probability of unique integer solution to a system of linear equations,” *European Journal of Operational Research*, vol. 214, no. 1, pp. 27–30, 2011.

[69] S. Sparrer and R. F. Fischer, “Soft-feedback omp for the recovery of discrete-valued sparse signals,” in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 1461–1465.

[70] H. Zhu and G. B. Giannakis, “Exploiting sparse user activity in multiuser detection,” *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 454–465, 2010.

[71] S. M. Fosson, “Non-convex approach to binary compressed sensing,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 1959–1963.

[72] S. Verdu *et al.*, *Multiuser detection*. Cambridge university press, 1998.

[73] E. Arikan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Transactions on information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[74] A. Alamdar-Yazdi and F. R. Kschischang, “A simplified successive-cancellation decoder for polar codes,” *IEEE communications letters*, vol. 15, no. 12, pp. 1378–1380, 2011.

[75] M. Grant and S. Boyd, “Cvx: Matlab software for disciplined convex programming, version 2.1,” 2014.

[76] D. Bertsekas, *Convex optimization theory*. Athena Scientific, 2009, vol. 1.

[77] S. M. Patole, M. Torlak, D. Wang, and M. Ali, “Automotive radars: A review of signal processing techniques,” *IEEE Signal Processing Magazine*, vol. 34, no. 2, pp. 22–35, 2017.

[78] S. Sun, A. P. Petropulu, and H. V. Poor, “Mimo radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges,” *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 98–117, 2020.

[79] I. Bilik, O. Bialer, S. Villeval, H. Sharifi, K. Kona, M. Pan, D. Persechini, M. Musni, and K. Geary, “Automotive mimo radar for urban environments,” in *2016 IEEE Radar Conference (RadarConf)*. IEEE, 2016, pp. 1–6.

[80] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE journal of selected topics in signal processing*, vol. 8, no. 5, pp. 831–846, 2014.

[81] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, “Hybrid mimo architectures for millimeter wave communications: Phase shifters or switches?” *IEEE access*, vol. 4, pp. 247–267, 2016.

[82] S. Haghighatshoar and G. Caire, “Massive mimo channel subspace estimation from low-dimensional projections,” *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 303–318, 2016.

[83] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, “Spatial-and frequency-wideband effects in millimeter-wave massive mimo systems,” *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3393–3406, 2018.

[84] P. Pal and P. P. Vaidyanathan, “Nested arrays: A novel approach to array processing with enhanced degrees of freedom,” *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4167–4181, 2010.

[85] ——, “Coprime sampling and the music algorithm,” in *2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*. IEEE, 2011, pp. 289–294.

[86] M. Wang and A. Nehorai, “Coarrays, music, and the cramér–rao bound,” *IEEE Transactions on Signal Processing*, vol. 65, no. 4, pp. 933–946, 2016.

[87] Y. D. Zhang, M. G. Amin, and B. Himed, “Sparsity-based doa estimation using co-prime arrays,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3967–3971.

[88] C. Zhou, Z. Shi, Y. Gu, and Y. D. Zhang, “Coarray interpolation-based coprime array doa estimation via covariance matrix reconstruction,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3479–3483.

[89] H. Qiao and P. Pal, “Gridless line spectrum estimation and low-rank toeplitz matrix compression using structured samplers: A regularization-free approach,” *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2221–2236, 2017.

[90] R. Rajamäki and V. Koivunen, “Sparse symmetric linear arrays with low redundancy and a contiguous sum co-array,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 1697–1712, 2021.

[91] C.-L. Liu and P. Vaidyanathan, “Cramér–rao bounds for coprime and other sparse arrays, which find more sources than sensors,” *Digital Signal Processing*, vol. 61, pp. 43–61, 2017.

[92] A. Koochakzadeh and P. Pal, “Cramér–rao bounds for underdetermined source localization,” *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 919–923, 2016.

[93] C.-L. Liu and P. Vaidyanathan, “Remarks on the spatial smoothing step in coarray music,” *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1438–1442, 2015.

[94] Y. I. Abramovich, D. A. Gray, A. Y. Gorokhov, and N. K. Spencer, “Positive-definite toeplitz completion in doa estimation for nonuniform linear antenna arrays. i. fully augmentable arrays,” *IEEE Transactions on Signal Processing*, vol. 46, no. 9, pp. 2458–2471, 1998.

[95] C.-L. Liu, P. Vaidyanathan, and P. Pal, “Coprime coarray interpolation for doa estimation via nuclear norm minimization,” in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2016, pp. 2639–2642.

[96] H. Qiao and P. Pal, “Unified analysis of co-array interpolation for direction-of-arrival estimation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 3056–3060.

[97] C. Zhou, Y. Gu, X. Fan, Z. Shi, G. Mao, and Y. D. Zhang, “Direction-of-arrival estimation for coprime array via virtual array interpolation,” *IEEE Transactions on Signal Processing*, vol. 66, no. 22, pp. 5956–5971, 2018.

[98] S. Qin, Y. D. Zhang, and M. G. Amin, “Generalized coprime array configurations for direction-of-arrival estimation,” *IEEE Transactions on Signal Processing*, vol. 63, no. 6, pp. 1377–1390, 2015.

[99] Y. D. Zhang, S. Qin, and M. G. Amin, “Doa estimation exploiting coprime arrays with sparse sensor spacing,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2267–2271.

[100] Y. Li and Y. Chi, “Off-the-grid line spectrum denoising and estimation with multiple measurement vectors,” *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1257–1269, 2015.

[101] Z. Tan, Y. C. Eldar, and A. Nehorai, “Direction of arrival estimation using co-prime arrays: A super resolution viewpoint,” *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5565–5576, 2014.

[102] P. Pal and P. P. Vaidyanathan, “A grid-less approach to underdetermined direction of arrival estimation via low rank matrix denoising,” *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 737–741, 2014.

[103] M. Wang, Z. Zhang, and A. Nehorai, “Performance analysis of coarray-based music and the cramér-rao bound,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 3061–3065.

[104] H. Qiao and P. Pal, “Generalized nested sampling for compressing low rank toeplitz matrices,” *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1844–1848, 2015.

[105] G. Cybenko, “Moment problems and low rank toeplitz approximations,” *Circuits, Systems and Signal Processing*, vol. 1, no. 3, pp. 345–366, 1982.

[106] P. Vaidyanathan and P. Pal, “Direct-music on sparse arrays,” in *2012 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2012, pp. 1–5.

[107] ——, “Why does direct-music on sparse-arrays work?” in *2013 Asilomar Conference on Signals, Systems and Computers*. IEEE, 2013, pp. 2007–2011.

[108] G. Schiebinger, E. Robeva, and B. Recht, “Superresolution without separation,” *Information and Inference: A Journal of the IMA*, vol. 7, no. 1, pp. 1–30, 2018.

[109] P. Stoica and A. Nehorai, “Music, maximum likelihood, and cramer-rao bound: further results and comparisons,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2140–2150, 1990.

[110] S. Sun and Y. D. Zhang, “4D Automotive Radar sensing for autonomous vehicles: A sparsity-oriented approach,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 879–891, 2021.

[111] S. Fortunati, R. Grasso, F. Gini, M. S. Greco, and K. LePage, “Single-snapshot DOA estimation by using compressed sensing,” *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–17, 2014.

[112] S. Sun and A. P. Petropulu, “A sparse linear array approach in automotive radars using matrix completion,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8614–8618.

[113] Y. Ma, X. Cao, X. Wang, M. S. Greco, and F. Gini, “Multi-source off-grid DOA estimation with single snapshot using non-uniform linear arrays,” *Signal Processing*, vol. 189, p. 108238, 2021.

[114] H. Huang, H. C. So, and A. M. Zoubir, “Off-grid direction-of-arrival estimation using second-order Taylor approximation,” *Signal Processing*, vol. 196, p. 108513, 2022.

[115] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, “Compressed sensing off the grid,” *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[116] Y. Ma, X. Cao, and X. Wang, “Off-grid DOA estimation with arbitrary-spaced linear array using single snapshot,” in *2019 IEEE Radar Conference (RadarConf)*. IEEE, 2019, pp. 1–6.

[117] Y. Chen and Y. Chi, “Spectral compressed sensing via structured matrix completion,” in *International Conference on Machine Learning*. PMLR, 2013, pp. 414–422.

[118] A. G. Raj and J. H. McClellan, “Single snapshot super-resolution DOA estimation for arbitrary array geometries,” *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 119–123, 2018.

[119] P. Sarangi, M. C. Hücümenoğlu, and P. Pal, “Beyond coarray MUSIC: Harnessing the difference sets of nested arrays with limited snapshots,” *IEEE Signal Processing Letters*, vol. 28, pp. 2172–2176, 2021.

[120] S. Liu, Z. Mao, Y. D. Zhang, and Y. Huang, "Rank minimization-based toeplitz reconstruction for doa estimation using coprime array," *IEEE Communications Letters*, 2021.

[121] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8. IEEE, 1983, pp. 336–339.

[122] D. Ciuonzo, G. Romano, and R. Solimene, "Performance analysis of time-reversal music," *IEEE Transactions on Signal Processing*, vol. 63, no. 10, pp. 2650–2662, 2015.

[123] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," *Optical Compressive Imaging*, pp. 279–312, 2016.

[124] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[125] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding blind deconvolution algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2354–2367, 2011.

[126] L. Borcea, G. Papanicolaou, C. Tsogka, and J. Berryman, "Imaging and time reversal in random media," *Inverse Problems*, vol. 18, no. 5, p. 1247, 2002.

[127] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Transactions on Information Theory*, vol. 60, no. 3, pp. 1711–1732, 2013.

[128] Y. Li, K. Lee, and Y. Bresler, "Identifiability in blind deconvolution with subspace or sparsity constraints," *IEEE Transactions on information Theory*, vol. 62, no. 7, pp. 4266–4275, 2016.

[129] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[130] Y. Chi, "Guaranteed blind sparse spikes deconvolution via lifting and convex optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 782–794, 2016.

[131] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, vol. 47, no. 3, pp. 893–934, 2019.

[132] Y. Zhang, Y. Lau, H.-w. Kuo, S. Cheung, A. Pasupathy, and J. Wright, "On the global geometry of sphere-constrained sparse blind deconvolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4894–4902.

[133] H.-W. Kuo, Y. Lau, Y. Zhang, and J. Wright, "Geometry and symmetry in short-and-sparse deconvolution," *arXiv preprint arXiv:1901.00256*, 2019.

[134] Y. Lau, Q. Qu, H.-W. Kuo, P. Zhou, Y. Zhang, and J. Wright, "Short-and-sparse deconvolution–a geometric approach," *arXiv preprint arXiv:1908.10959*, 2019.

[135] E. A. Pnevmatikakis, J. Merel, A. Pakman, and L. Paninski, "Bayesian spike inference from calcium imaging data," in *2013 Asilomar Conference on Signals, Systems and Computers*. IEEE, 2013, pp. 349–353.

[136] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang *et al.*, "Simultaneous denoising, deconvolution, and demixing of calcium imaging data," *Neuron*, vol. 89, no. 2, pp. 285–299, 2016.

[137] J. Friedrich and L. Paninski, "Fast active set methods for online spike inference from calcium imaging," in *Advances In Neural Information Processing Systems*, 2016, pp. 1984–1992.

[138] J. T. Vogelstein, A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski, "Fast nonnegative deconvolution for spike train inference from population calcium imaging," *Journal of neurophysiology*, vol. 104, no. 6, pp. 3691–3704, 2010.

[139] M. Rosenblatt, "Some simple remarks on an autoregressive scheme and an implied problem," *Journal of Theoretical Probability*, vol. 10, no. 2, pp. 295–305, 1997.

[140] T.-H. Li *et al.*, "Blind deconvolution of linear systems with multilevel nonstationary inputs," *The Annals of Statistics*, vol. 23, no. 2, pp. 690–704, 1995.

[141] F. Gamboa, E. Gassiat *et al.*, "Blind deconvolution of discrete linear systems," *The Annals of Statistics*, vol. 24, no. 5, pp. 1964–1981, 1996.

[142] H. Qiao and P. Pal, "Gridless line spectrum estimation and low-rank toeplitz matrix compression using structured samplers: A regularization-free approach," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2221–2236, May 2017.

[143] J. Dainty and J. Fienup, "Image recovery: Theory and application, edited by h. stark," 1987.

[144] A. M. Maiden and J. M. Rodenburg, "An improved ptychographical phase retrieval algorithm for diffractive imaging," *Ultramicroscopy*, vol. 109, no. 10, pp. 1256–1262, 2009.

[145] D. Gabor, "A new microscopic principle," 1948.

[146] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," *Information and Inference: A Journal of the IMA*, vol. 5, no. 4, pp. 405–441, 2016.

[147] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[148] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[149] E. J. Candès and X. Li, "Solving quadratic equations via phaselift when there are about as many equations as unknowns," *Foundations of Computational Mathematics*, vol. 14, no. 5, pp. 1017–1026, 2014.

[150] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.

[151] T. Goldstein and C. Studer, "Phasemax: Convex phase retrieval via basis pursuit," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2675–2689, 2018.

[152] S. Bahmani and J. Romberg, "Phase retrieval meets statistical learning theory: A flexible convex relaxation," *arXiv preprint arXiv:1610.04210*, 2016.

[153] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[154] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[155] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 773–794, 2017.

[156] H. Zhang and Y. Liang, "Reshaped wirtinger flow for solving quadratic system of equations," in *Advances in Neural Information Processing Systems*, 2016, pp. 2622–2630.

[157] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM Journal on Mathematical Analysis*, vol. 45, no. 5, pp. 3019–3033, 2013.

[158] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[159] P. Hand and V. Voroninski, "Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space," *arXiv preprint arXiv:1611.05985*, 2016.

[160] F. Salehi, E. Abbasi, and B. Hassibi, "Learning without the phase: Regularized phasemax achieves optimal sample complexity," in *Advances in Neural Information Processing Systems*, 2018, pp. 8641–8652.

[161] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen, “Sparse phase retrieval via truncated amplitude flow,” *IEEE Transactions on Signal Processing*, vol. 66, no. 2, pp. 479–491, 2017.

[162] T. T. Cai, X. Li, Z. Ma *et al.*, “Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow,” *The Annals of Statistics*, vol. 44, no. 5, pp. 2221–2251, 2016.

[163] G. Jagatap and C. Hegde, “Fast, sample-efficient algorithms for structured phase retrieval,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4917–4927.

[164] M. Soltanolkotabi, “Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization,” *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2374–2400, 2019.

[165] S. Bahmani and J. Romberg, “Efficient compressive phase retrieval with constrained sensing vectors,” in *Advances in Neural Information Processing Systems*, 2015, pp. 523–531.

[166] M. Iwen, A. Viswanathan, and Y. Wang, “Robust sparse phase retrieval made easy,” *Applied and Computational Harmonic Analysis*, vol. 42, no. 1, pp. 135–142, 2017.

[167] L. Demanet and P. Hand, “Stable optimizationless recovery from phaseless linear measurements,” *Journal of Fourier Analysis and Applications*, vol. 20, no. 1, pp. 199–221, 2014.

[168] Y. C. Eldar and S. Mendelson, “Phase retrieval: Stability and recovery guarantees,” *Applied and Computational Harmonic Analysis*, vol. 36, no. 3, pp. 473–494, 2014.

[169] R. Leitgeb, C. Hitzenberger, and A. F. Fercher, “Performance of fourier domain vs. time domain optical coherence tomography,” *Optics express*, vol. 11, no. 8, pp. 889–894, 2003.

[170] R. P. Millane, “Phase retrieval in crystallography and optics,” *JOSA A*, vol. 7, no. 3, pp. 394–411, 1990.

[171] A. Szöke, “Holographic microscopy with a complicated reference,” *Journal of Imaging Science and Technology*, vol. 41, no. 4, pp. 332–341, 1997.

[172] A. Drenth, A. Huiser, and H. Ferwerda, “The problem of phase retrieval in light and electron microscopy of strong objects,” *Optica Acta: International Journal of Optics*, vol. 22, no. 7, pp. 615–628, 1975.

[173] J. R. Fienup, “Phase retrieval algorithms: a comparison,” *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[174] R. W. Gerchberg, “Phase determination from image and diffraction plane pictures in the electron microscope,” *Optik*, vol. 34, pp. 275–284, 1971.

[175] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, pp. 47–81, 2015.

[176] K. Jaganathan, S. Oymak, and B. Hassibi, "Recovery of sparse 1-d signals from the magnitudes of their fourier transform," in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 1473–1477.

[177] Y. Shechtman, A. Beck, and Y. C. Eldar, "Gespar: Efficient phase retrieval of sparse signals," *IEEE transactions on signal processing*, vol. 62, no. 4, pp. 928–938, 2014.

[178] K. Jaganathan, Y. Eldar, and B. Hassibi, "Phase retrieval with masks using convex optimization," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1655–1659.

[179] A. S. Bandeira, Y. Chen, and D. G. Mixon, "Phase retrieval from power spectra of masked signals," *Information and Inference: a Journal of the IMA*, vol. 3, no. 2, pp. 83–102, 2014.

[180] E. G. Loewen and E. Popov, *Diffraction gratings and applications*. CRC Press, 2018.

[181] A. Faridian, D. Hopp, G. Pedrini, U. Eigenthaler, M. Hirscher, and W. Osten, "Nanoscale imaging using deep ultraviolet digital holographic microscopy," *Optics express*, vol. 18, no. 13, pp. 14 159–14 164, 2010.

[182] S. C. Sekhar, H. Nazkani, T. Blu, and M. Unser, "A new technique for high-resolution frequency domain optical coherence tomography," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1. IEEE, 2007, pp. I–425.

[183] C. S. Seelamantula, M. L. Villiger, R. A. Leitgeb, and M. Unser, "Exact and efficient signal reconstruction in frequency-domain optical-coherence tomography," *JOSA A*, vol. 25, no. 7, pp. 1762–1771, 2008.

[184] S. Mukherjee and C. S. Seelamantula, "Fienup algorithm with sparsity constraints: Application to frequency-domain optical-coherence tomography," *IEEE Transactions on Signal Processing*, vol. 62, no. 18, pp. 4659–4672, 2014.

[185] M. Wojtkowski, R. Leitgeb, A. Kowalczyk, T. Bajraszewski, and A. F. Fercher, "In vivo human retinal imaging by fourier domain optical coherence tomography," *Journal of biomedical optics*, vol. 7, no. 3, pp. 457–463, 2002.

[186] J. A. Izatt and M. A. Choma, "Theory of optical coherence tomography," *Optical Coherence Tomography: Technology and Applications*, pp. 47–72, 2008.

[187] B. Bouma, "Handbook of optical coherence tomography," 2001.

[188] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the *k*-user interference channel," *IEEE transactions on information theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[189] A. S. Motahari, S. Oveis-Gharan, M.-A. Maddah-Ali, and A. K. Khandani, "Real interference alignment: Exploiting the potential of single antenna systems," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4799–4810, 2014.

[190] A. Khintchine, "Einige sätze über kettenbrüche, mit anwendungen auf die theorie der diophantischen approximationen," *Mathematische Annalen*, vol. 92, no. 1-2, pp. 115–125, 1924.

[191] V. G. Sprindzhuk, *Metric theory of Diophantine approximations*. VH Winston, 1979.

[192] A. Fengler, P. Jung, and G. Caire, "Sparcs for unsourced random access," *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6894–6915, 2021.

[193] V. K. Amalladinne, J.-F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for unsourced multiple access," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6509–6533, 2020.

[194] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive mimo receiver," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2925–2951, 2021.

[195] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.

[196] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[197] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.