

COMPRESSIVE SYSTEM IDENTIFICATION (CSI): THEORY AND APPLICATIONS
OF EXPLOITING SPARSITY IN THE ANALYSIS OF HIGH-DIMENSIONAL
DYNAMICAL SYSTEMS

by

Borhan M. Sanandaji

© Copyright by Borhan M. Sanandaji, 2012

All Rights Reserved

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Electrical Engineering).

Golden, Colorado

Date _____

Signed: _____

Borhan M. Sanandaji

Signed: _____

Prof. Tyrone L. Vincent
Thesis Advisor

Signed: _____

Prof. Michael B. Wakin
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____

Prof. Randy L. Haupt
Professor and Head
Department of Electrical Engineering and Computer Science

ABSTRACT

The information content of many phenomena of practical interest is often much less than what is suggested by their actual size. As an inspiring example, one active research area in biology is to understand the relations between the genes. While the number of genes in a so-called gene network can be large, the number of contributing genes to each given gene in the network is usually small compared to the size of the network. In other words, the behavior of each gene can be expressed as a *sparse* combination of other genes.

The purpose of this thesis is to develop new theory and algorithms for exploiting this type of *simplicity* in the analysis of high-dimensional dynamical systems with a particular focus on system identification and estimation. In particular, we consider systems with a high-dimensional but sparse impulse response, large-scale interconnected dynamical systems when the associated graph has a sparse flow, linear time-varying systems with few piecewise-constant parameter changes, and systems with a high-dimensional but sparse initial state. We categorize all of these problems under the common theme of Compressive System Identification (CSI) in which one aims at identifying some facts (e.g., the impulse response of the system, the underlying topology of the interconnected graph, or the initial state of the system) about the system under study from the *smallest possible* number of observations.

Our work is inspired by the field of Compressive Sensing (CS) which is a recent paradigm in signal processing for sparse signal recovery. The CS recovery problem states that a sparse signal can be recovered from a small number of random linear measurements. Compared to the standard CS setup, however, we deal with structured sparse signals (e.g., block-sparse signals) and structured measurement matrices (e.g., Toeplitz matrices) where the structure is implied by the system under study.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF FIGURES	ix
ACKNOWLEDGMENTS	xiv
DEDICATION	xvi
CHAPTER 1 INTRODUCTION	1
1.1 Identification from Limited Data	1
1.2 High-dimensional but Sparse Systems	1
1.3 Exploiting Sparsity and CS	2
1.4 Overview and Contributions	3
CHAPTER 2 BACKGROUND	7
2.1 Mathematical Preliminaries and Notation	7
2.1.1 Signal Notation and Norms	7
2.1.2 Matrix Notation and Norms	8
2.1.3 Signal Models	9
2.2 Compressive Sensing (CS)	10
2.2.1 Recovery via ℓ_0 -minimization	11
2.2.2 Recovery via ℓ_1 -minimization	11
2.2.3 ℓ_0/ℓ_1 Equivalence and the Restricted Isometry Property (RIP)	11
2.2.4 CS Geometry	14
CHAPTER 3 COM INEQUALITIES FOR TOEPLITZ MATRICES	16

3.1	Introduction	16
3.2	Related Work	19
3.3	Contributions	20
3.4	Main Results	20
3.5	Proof of Theorem 3.7	27
3.6	Proof and Discussion of Theorem 3.15	31
3.6.1	Circulant Embedding	31
3.6.2	Deterministic Bound	33
3.6.3	Supporting Results	33
3.6.4	Completing the Proof of Theorem 3.15	35
3.7	Discussion	36
3.8	A Quadratic RIP Bound and Non-uniform Recovery	38
3.9	Compressive Binary Detection (CBD)	40
3.9.1	Problem Setup	40
3.9.2	Empirical Results and Receiver Operating Curves (ROCs)	42
CHAPTER 4 COMPRESSIVE TOPOLOGY IDENTIFICATION		45
4.1	Introduction	46
4.2	Interconnected Dynamical Systems	47
4.3	Network Tomography	49
4.4	Compressive Topology Identification	50
4.5	CTI via Block-Sparse Recovery	52
4.5.1	Block-Coherence and Sub-Coherence	53
4.5.2	Recovery Condition	54

4.5.3	Network Coherence	54
4.5.4	Simulations and Discussion on the Network Coherence	58
4.6	CTI via Clustered-Sparse Recovery	60
4.6.1	Clustered Orthogonal Matching Pursuit (COMP)	61
4.6.2	Numerical Simulations	64
CHAPTER 5 CSI OF LTI AND LTV ARX MODELS		67
5.1	Introduction	67
5.2	Auto Regressive with eXternal input (ARX) Models	68
5.3	CSI of Linear Time-Invariant (LTI) ARX Models	70
5.3.1	CSI of LTI Systems with Unknown Input Delays	70
5.3.2	Simulation Results	71
5.3.3	Bounds on Coherence	72
5.3.4	Reducing Coherence by Pre-Filtering	74
5.4	CSI of Linear Time-Variant (LTV) ARX Models	75
5.4.1	Piecewise-Constant $\theta(t)$ and Block-Sparse Recovery	76
5.4.2	Identifiability Issue	77
5.4.3	Sampling Approach for LTV System Identification	78
5.4.4	Simulation Results	79
5.5	Case Study: Harmonic Drive System	80
5.5.1	Acquisition of the Input-Output Data Set	81
5.5.2	Construction of the time-varying ARX Model	81
5.5.3	Identifiability Issue	83
5.5.4	BOMP for Noisy Data	84

5.5.5	Scaling and Column Normalization	85
5.5.6	Results	86
CHAPTER 6 OBSERVABILITY WITH RANDOM OBSERVATIONS		87
6.1	Introduction	88
6.1.1	Measurement Burdens in Observability Theory	88
6.1.2	Compressive Sensing and Randomized Measurements	89
6.1.3	Observability from Random, Compressive Measurements	91
6.1.4	Related Work	92
6.1.5	Chapter Organization	93
6.2	The RIP and the Observability Matrix	94
6.2.1	Bounding $\mathbf{E}[\delta_S]$	97
6.2.2	Tail Bound for δ_S	102
6.3	CoM Inequalities and the Observability Matrix	107
6.3.1	Independent Random Measurement Matrices	108
6.3.2	Unitary and Symmetric System Matrices	110
6.3.3	Identical Random Measurement Matrices	117
6.4	Case Study: Estimating the Initial State in a Diffusion Process	121
6.4.1	System Model	121
6.4.2	Diffusion and its Connections to Theorem 6.54	123
6.4.3	State Recovery from Compressive Measurements	124
CHAPTER 7 CONCLUSIONS		131
7.1	Sparse Recovery and CSI	132
7.2	Recovery Guarantees	132

7.2.1	Concentration of Measure Inequalities	132
7.2.2	Restricted Isometry Property	133
7.2.3	Coherence	134
7.3	Future Research Directions	134
	REFERENCES CITED	136
	APPENDIX - PROOFS	145
A.1	Proof of Lemma 3.48	145
A.2	Proof of Proposition 3.52	145
A.3	Proof of Lemma 4.23	146
A.4	Proof of Theorem 5.8	147
A.5	Proof of Theorem 5.11	149

LIST OF FIGURES

Figure 2.1	The Restricted Isometry Property (RIP) as a stable embedding. Each K -plane contains all K -sparse points in \mathbb{R}^N whose support (the locations of the non-zero entries) is the same.	12
Figure 2.2	ℓ_2 -minimization versus ℓ_1 -minimization. (a) ℓ_2 -minimization. The recovered solution is almost never sparse as the ℓ_2 -ball does not have sharp edges along the axes. (b) ℓ_1 -minimization. If the hyperplane \mathcal{H}_b does not intersect the ℓ_1 -ball of radius $\ \mathbf{x}^*\ _1$, we have true sparse recovery due to the geometry of the ℓ_1 -ball (sharp edges along the axes).	13
Figure 2.3	Incorrect recovery through ℓ_1 -minimization. (a) The hyperplane \mathcal{H}_b does intersect the ℓ_1 -ball of radius $\ \mathbf{x}^*\ _1$. (b) The ℓ_1 -ball intersects the shifted nullspace at a point different than the true sparse solution, ending up with incorrect recovery.	14
Figure 3.1	Illustrating the signal-dependency of the left-hand side of Concentration of Measure (CoM) inequality. With fixed $N = 1024$, $K = 64$ and $M = 512$, we consider two particular K -sparse signals, \mathbf{a}_1 , and \mathbf{a}_2 . Both \mathbf{a}_1 and \mathbf{a}_2 are normalized. We measure each of these signals with 1000 independent and identically distributed (i.i.d.) Gaussian $M \times N$ Toeplitz matrices. (a) The K non-zero entries of \mathbf{a}_1 have equal values and occur in the first K entries of the vector. (b) The K non-zero entries of \mathbf{a}_2 appear in randomly-selected locations with random signs and values. The two signals have different concentrations that can be upper bounded by the signal-dependent bound.	25
Figure 3.2	Sample mean of $\rho(\mathbf{a})$ in the time and frequency domains versus the expectation bounds, where $L = N + M - 1$. The sample mean $\bar{\rho}(\mathbf{a})$ is calculated by taking the mean over 1000 constructed signals \mathbf{a} . A logarithmic scale is used for the vertical axis. (a) Varying K with fixed $M = N = 256$. (b) Varying M with fixed $N = 512$ and $K = 32$	26
Figure 3.3	<i>Toeplitz matrix</i> $A \in \mathbb{R}^{L \times M}$ and its circulant counterpart $A_c \in \mathbb{R}^{L \times L}$ where $L = N + M - 1$	32
Figure 3.4	Empirical results. Sample mean of $\rho_c(\mathbf{a})$ in the time domain for full vectors $\mathbf{a} \in \mathbb{R}^N$ where $M = 256$ is fixed. Also plotted is $f(L) = \log(L)$, where $L = N + M - 1$	38

Figure 3.5	<i>Empirical results. (a) Simulation results vs. the conjectured bound</i> $g(K) = \frac{K}{c_1 K + c_2}$ with $c = 1$. (b) Linearity of $\frac{K}{\bar{\rho}_c(\mathbf{a})}$.	39
Figure 3.6	<i>Finite Impulse Response (FIR) filter of order N with impulse response $\{a_k\}_{k=1}^N$.</i>	40
Figure 3.7	ROCs for 1000 random matrices X for a fixed signal \mathbf{c} with $\rho(\mathbf{c}) = 45.6$. (a) Unstructured X . (b) Toeplitz X . The solid black curve is the average of 1000 curves.	42
Figure 3.8	Average ROCs over 1000 random matrices X for 6 different signals \mathbf{c} . (a) Unstructured X . All curves are overlapping. (b) Toeplitz X . The curves descend in the same order they appear in legend box.	43
Figure 4.1	(a) Network model of 6 interconnected nodes. Each edge of the directed graph (\mathbf{x}_i^j) represents an FIR filter of order n_i^j (i.e., $\mathbf{x}_i^j \in \mathbb{R}^{n_i^j}$). (b) Single node model. Each node is a summer whose inputs are the signals from the incoming edges, while the output of the summer is sent to the outgoing edges. In this illustration, node i sums the signals from nodes 3, 7, and 9 (i.e., $\mathcal{N}_i = \{3, 7, 9\}$) plus a node-specific input term $u_i(t)$.	48
Figure 4.2	Clustered sparsity versus block sparsity model. (a) A 5-block sparse signal with block size $m = 16$. (b) A (21, 5)-clustered sparse signal. Clusters have different sizes. Both signals have same length $N = 256$.	52
Figure 4.3	A simple network for our study of network coherence.	55
Figure 4.4	(a) Disconnected network. Node 1 has in-degree 2 but is disconnected from the rest of the network (i.e., out-degree zero). (b) Connected network. Node 1 has in-degree 2 and is connected to the rest of the network (i.e., in this example out-degree 5).	59
Figure 4.5	(a) Coherence metrics for the disconnected (dashed lines) and connected (solid lines) networks. The curves are averaged over 1000 realizations of the networks. Note that the coherence metrics approach a non-zero asymptote as the number of measurements M increases. (b) Coherence metrics for matrices with i.i.d. Gaussian entries (solid lines) and matrices which are block-concatenations of Toeplitz matrices with i.i.d. Gaussian entries. The curves are averaged over 1000 realizations of these types of matrices. Note that the coherence metrics approach zero as the number of measurements M increases.	60
Figure 4.6	Recovery rate comparison of node 1 (\mathbf{x}_1^2 and \mathbf{x}_1^3) for connected and disconnected networks. For each measurement, 1000 realizations of the network are carried out and the recovery rate is calculated.	61

Figure 4.7	A network of 32 interconnected nodes including trees, loops and self-loops. Each edge of the directed graph (\mathbf{x}_i^j) represents an FIR filter.	63
Figure 4.8	Recovery performance corresponding to node 10. (a) Signal \mathbf{x} corresponding to node 10 in the network graph. The cluster-sparsity level corresponds to the in-degree of node 10. (b) Recovery performance comparison between Clustered Orthogonal Matching Pursuit (COMP) and Block Orthogonal Matching Pursuit (BOMP) with different block sizes m . An initial value of $w = m = 8$ is chosen in COMP. The algorithm iterates by reducing w until stopping criteria are met. For comparison, BOMP is tested with three different block sizes ($m = \{2, 4, 8\}$). The success rate is calculated over 300 realizations of the network for a given number of measurements.	64
Figure 4.9	Recovery rate comparison of nodes 10, 23, and 32 in the network. An initial value of $w = m = 8$ is chosen in COMP. Nodes 23 and 32 have in-degree 2 and node 10 has in-degree 4. The success rate is calculated over 300 realizations of the network for a given number of measurements.	65
Figure 5.1	CSI results on a $\{2, 2, 40\}$ Single-Input Single-Output (SISO) Linear Time-Invariant (LTI) Auto Regressive with eXternal input (ARX) System. (a) In the recovery algorithm, m and d are unknown. The plot shows the recovery success rate over 1000 realizations of the system. (b) Averaged mutual coherence of Φ over 1000 realizations of the system (solid curve). Lower bound of Theorem 5.8 (dashed line).	72
Figure 5.2	CSI results on a $\{2, 2, \{60, 21, 10, 41\}\}$ Multi-Input Single-Output (MISO) LTI ARX system. In the recovery algorithm, m and $\{d_i\}_{i=1}^4$ are unknown. The plot shows the recovery success rate over 1000 realizations of the system.	73
Figure 5.3	Reducing coherence by pre-filtering. (a) Pre-filtering scheme. (b) For each α , the filter $G(z)$ is applied on the input/output signals and the limit of the expected value of coherence is calculated over 1000 realizations of system.	74
Figure 5.4	Random sampling scheme for $M = \{10, 30, 50\}$ measurements. Samples are chosen randomly according to a uniform distribution. System parameters are assumed to change at $t = 300$ and $t = 400$	79
Figure 5.5	A 3-model LTV ARX system. (a) Output of the model. System parameters change at $t = 300$ and $t = 400$. (b) Time-varying parameters of the 3-model system. At the time of change, all the system parameters change.	80

Figure 5.6	Recovery performance of 4 different systems. The plots show the recovery success rate over 1000 realizations of the system.	80
Figure 5.7	A DC Motor System consisting of a DC motor, a harmonic drive system, an inertial load, and an encoder.	81
Figure 5.8	The experimental input-output data of a harmonic drive system. A sinusoidal electric current $i(t)$ with 5.12[s] period is applied to the system, and the rotation angle is sampled with constant sampling interval 0.01[s] (i.e., $t_k = 0.01(k - 1)$). The angular velocity ($\triangleq \omega(t)$) and its derivative $\dot{\omega}(t)$ are calculated by backward and central difference, respectively.	82
Figure 5.9	The estimated ARX model parameters corresponding to the DC motor system. As can be seen, the identified parameters have a piecewise-constant behavior with only a few changes happening over the course of experiment.	83
Figure 5.10	A comparison between the measured experimental output $\dot{\omega}(t_k)$ and the estimated output $\hat{\dot{\omega}}(t_k)$ based on the identified piecewise-constant ARX model parameters.	84
Figure 6.1	One-dimensional diffusion process. At time zero, the concentration (the state) is non-zero only at a few locations of the path graph of $N = 100$ nodes.	123
Figure 6.2	Dense Measurements versus Line Measurements. The color of a node indicates the corresponding weight of that node. The darker the node color, the higher the weight. These weights are the entries of each row of each C_k . (a) Dense Measurements. The weights are drawn from a Gaussian distribution with mean zero and variance $\frac{1}{M}$. These values are random and change for each measurement. (b) Line Measurements. The weights are generated as a function of the perpendicular distances of all nodes of the grid to the line. The slope and the intercept of the line are random and change for each measurement.	124
Figure 6.3	Signal recovery from compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. The plots show the percent of trials (out of 300 trials in total) with perfect recovery of the initial state \mathbf{x}_0 versus the number of measurements M . (a) Recovery from compressive measurements at time $k = 0$. (b) Recovery from compressive measurements at time $k = 10$. . .	127

Figure 6.4	Signal recovery from compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. The plots show the percent of trials (out of 300 trials in total) with perfect recovery of the initial state \mathbf{x}_0 versus the number of measurements M taken at observation times $k = \{0, 1, 2, 8, 50, 100\}$. (a) Recovery from compressive Dense Measurements. (b) Recovery from compressive Line Measurements. . . .	128
Figure 6.5	Signal recovery from $M = 32$ compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. The plots show the recovery error of the initial state $\ \mathbf{e}\ _2 = \ \hat{\mathbf{x}}_0 - \mathbf{x}_0\ _2$ over 300 trials. (a) Recovery from compressive measurements at time $k = 2$. (b) Recovery from compressive measurements at time $k = 10$	129
Figure 6.6	Signal recovery from compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. A total of $KM = 32$ measurements are spread over $K = 4$ observation times while at each time, $M = 8$ measurements are taken. (a) Percent of trials (out of 300 trials in total) with perfect recovery of the initial state \mathbf{x}_0 are shown for different sample sets, Ω_i . (b) Recovery error of the initial state $\ \mathbf{e}\ _2 = \ \hat{\mathbf{x}}_0 - \mathbf{x}_0\ _2$ over 300 trials for set Ω_4	130

ACKNOWLEDGMENTS

During the last five years, I have been truly privileged and honored to have the chance to meet and work with my advisors, Tyrone Vincent and Michael Wakin. Without any doubt, it was their presence, support, patience, and boundless energy and enthusiasm that inspired me all the way through my PhD studies. It was due to them that I could develop as a better researcher, speaker, and writer. I thank them for providing me with a research environment in which I could freely explore new ideas and topics that are of my interest, yet challenging and novel. I thank Tyrone and Mike for being my friends before being my advisors and I will be forever thankful for this opportunity.

My deep gratitude extends to Kameshwar Poolla whose continued support and presence have significantly inspired me during the last few years. I would like to thank him for hosting me during Summer 2011 at the Berkeley Center for Control and Identification (BCCI) and also for being a member of my thesis committee. I would like to thank Kevin Moore for all of his advice and the opportunities he provided for me during my PhD studies and also for being a member of my thesis committee. I would like to thank Dinesh Mehta for kindly accepting to be my minor degree representative in mathematics and also for being a member of my thesis committee. I would like to thank Luis Tenorio for all of the time he spent helping me with my math questions, his insightful comments, and also for being a member of my thesis committee. I am very grateful to all of my thesis committee members and colleagues for the many ways they contributed to this work.

I am very thankful to Bob Kee, Neal Sullivan, and the Colorado Fuel Cell Center (CFCC) for all of the collaboration we had on the “SOFC System Identification and Control” research project during the first few years of my PhD studies. I particularly thank Bob for his collaboration and all of the fun times we had at our noon basketball playings.

I would like to thank Roland Tóth for his collaboration and his insightful comments on my work. My gratitude extends to Diego Regruto and Roland for their collaboration in organizing our CDC12 invited session on “Convex Relaxation in Identification and Control.” This thesis owes a great deal to the contributions of many great researchers: Tyrone Vincent (CSM), Mike Wakin (CSM), Kameshwar Poolla (UC Berkeley), Kevin Moore (CSM), Roland Tóth (TU Eindhoven), Ichiro Maruta (Kyoto University), Toshiharu Sugie (Kyoto University), Chris Rozell (Georgia Tech), Luis Tenorio (CSM), Armin Eftekhari, Han Lun Yap, and Alejandro Weinstein. I would also like to take this opportunity to thank all of my teachers and mentors, all the way from elementary school until my undergraduate school, Amirkabir U of Tech (Tehran Polytechnic), and my graduate school, Tehran Petroleum U of Tech.

From my first days at CSM, I have been truly lucky to meet Alejandro and his wife, Karem Tello, whose friendship has been priceless. I would like to express my deep gratitude to Alejandro for all of the time and energy he devoted to helping me and for always being there when I needed someone. I would like to extend my appreciation to all my officemates: Alejandro, Matt, Armin, Farshad, Andrew, Geraldine, Chang, Naveen, and many more for all the backgammon playings, Friday lunch gatherings, mountain biking, tennis and volleyball playings, etc. Thanks to all my friends whose presence has made this period more enjoyable: Mohsen G., Amir, Saeed, Sajjad, Soheil, Meysam, Navid, Mohsen M., Pegah, Abbas, Arshad, Neema, Azita, Little’s family, Bananeh, and everyone with whom I share good memories.

I have been truly privileged to have the presence of my “Uncle Nemat” Nemat Sanandaji and his family from the first moments of my arrival in Golden, CO. His great hospitality and kind support have significantly facilitated my life over the last five years.

My deepest appreciation belongs to my parents, Behnaz and Naser, my brother Borzu, my sister Negar, and my grandparents, Nezhat and Ali, for their unconditional love and endless support, all through my life. I have missed them all but their words and lessons have filled every second of my life. All I have achieved is a consequence of their efforts, lessons, guidance, love, and support. I owe them a lot and I dedicate this thesis to all of them.

To my grandfather, Ali

CHAPTER 1

INTRODUCTION

In this chapter, we present the motivation of our work and list our contributions.

1.1 Identification from Limited Data

Classical system identification approaches have limited performance in cases when the number of available data samples is small compared to the order of the system [1]. These approaches usually require a large data set in order to achieve a certain performance due to the asymptotic nature of their analysis. On the other hand, there are many application fields where only limited data points are available. Online estimation, Linear Time-Variant (LTV) system identification and setpoint-operated processes are examples of situations for which limited data samples are available. For some specific applications, the cost of the measuring process or the computational effort is also an issue. In such situations, it is *necessary* to perform the system identification from the smallest possible number of observations, although doing so leaves an apparently ill-conditioned identification problem.

1.2 High-dimensional but Sparse Systems

Many systems of practical interest are less complicated than what is suggested by the number of parameters in a standard model structure. They are often either low-order or can be represented in a suitable basis or formulation in which the number of parameters is small. The key element is that while the proper representation may be known, the particular elements within this representation that have non-zero coefficients are unknown. Thus, a particular system may be expressed by a coefficient vector with only a few non-zero elements, but this coefficient vector must be high-dimensional because we are not sure a priori which elements are non-zero. We term such a system a *sparse system*. In terms of a difference equation, for example, a sparse system may be a high-dimensional system with only a few non-zero

coefficients or it may be a system with an impulse response that is long but contains only a few non-zero terms. Multipath propagation [2–5], sparse channel estimation [6, 7], topology identification of interconnected systems [8–10] and sparse initial state estimation [11, 12] are examples involving systems that are high-order in terms of their ambient dimension but have a sparse (low-order) representation.

1.3 Exploiting Sparsity and Compressive Sensing (CS)

As mentioned in Section 1.1, performing the system identification from few observations leaves us with an apparently ill-conditioned set of linear equations. Solving such an ill-conditioned problem for a unique solution is impossible unless one has *extra* information about the true solution. CS, introduced by Candès, Romberg and Tao [13], and Donoho [14], however, is a powerful paradigm in signal processing which enables the recovery of an unknown vector from an underdetermined set of measurements under the assumption of sparsity of the signal and certain conditions on the measurement matrix. The CS recovery problem can be viewed as recovery of a K -sparse signal $\mathbf{x} \in \mathbb{R}^N$ from its observations $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ where $A \in \mathbb{R}^{M \times N}$ is the measurement matrix with $M < N$ (in many cases $M \ll N$). A K -sparse signal $\mathbf{x} \in \mathbb{R}^N$ is a signal of length N with K non-zero entries where $K < N$. The notation $K := \|\mathbf{x}\|_0$ denotes the sparsity level of \mathbf{x} . Since the null space of A is non-trivial, there are infinitely many candidate solutions to the equation $\mathbf{b} = A\mathbf{x}$; however, it has been shown that under certain conditions on the measurement matrix A , CS recovery algorithms can recover that unique solution if it is suitably sparse.

Inspired by CS, in this work we aim at performing system identification of sparse systems using a number of observations that is smaller than the ambient dimension. Although we look at different problems throughout this thesis, we categorize all of these problems as Compressive System Identification (CSI). CSI is beneficial in applications when only a limited data set is available. Moreover, CSI can help solve the issue of under and over parameterization, which is a common problem in parametric system identification. The chosen model structure should on one hand be rich enough to represent the behavior of the system

and on the other hand involve a minimal set of unknown parameters to minimize the variance of the parameter estimates. Under and over parameterization may have a considerable impact on the identification result, and choosing an optimal model structure is one of the primary challenges in system identification.

1.4 Overview and Contributions

We develop new theory and methods for exploiting underlying simplicity in the analysis of high-dimensional dynamical systems with a particular focus on system identification and estimation. In particular, we consider systems with a high-dimensional but sparse impulse response, large-scale interconnected dynamical systems when the associated graph has a sparse flow, linear time varying systems with few piecewise-constant parameter changes, and systems with a high-dimensional but sparse initial state. We outline the contributions of our work chapter-by-chapter.

We begin in Chapter 2 with a list of mathematical preliminaries and notation followed by an introduction to low-dimensional signal models, the signal dictionaries and representations. We continue by discussing sparse signal models and their recovery problems and algorithms. We also discuss the basics of CS.

In Chapter 3 we consider the problem of deriving Concentration of Measure (CoM) inequalities for randomized compressive Toeplitz matrices. Such inequalities are at the heart of the analysis of randomized compressive (having fewer rows than columns) linear operators. These inequalities show that the norm of a high-dimensional signal mapped by a compressive Toeplitz matrix to a low-dimensional space concentrates around its mean with a tail probability bound that decays exponentially in the dimension of the range space divided by a quantity which is a function of the signal. This implies that the CoM inequalities for compressive Toeplitz matrices are non-uniform and signal-dependent. To this end, we also consider analyzing the behavior of the introduced quantity. In particular, we show that for the class of *sparse* signals, the introduced quantity is bounded by the sparsity level of the signal.

Compressive Toeplitz matrices arise in problems involving the analysis of high-dimensional dynamical systems from few convolution-based measurements. As applications of the CoM inequalities, we consider Compressive Binary Detection (CBD) and discuss the implications of the CoM inequalities in such an application.

In Chapter 4 we consider the problem of identifying the topology of large-scale interconnected dynamical systems. We assume that the system topology under study has the structure of a directed graph of P nodes. Each edge of the directed graph represents a Finite Impulse Response (FIR) filter. Each node is a summer, whose inputs are the signals from the incoming edges, while the output of the summer is sent to outgoing edges. Both the graph topology and the impulse response of the FIR filters are unknown. We aim to perform the topology identification from the *smallest possible* number of node observations when there is limited data available and for this reason, we call this problem Compressive Topology Identification (CTI). Inspired by CS we show that in cases where the network interconnections are suitably *sparse* (i.e., the network contains sufficiently few links), it is possible to perfectly identify the network topology along with the filter impulse responses and the delays from small numbers of node observations, even though this leaves an apparently ill-conditioned identification problem.

If all the filters in the graph share the same order, we show that CTI can be cast as recovery of a *block-sparse* signal $\mathbf{x} \in \mathbb{R}^N$ from observations $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ with $M < N$, where the matrix A is a block-concatenation of P Toeplitz matrices. We use block-sparse recovery algorithms from the CS literature such as Block Orthogonal Matching Pursuit (BOMP) [15–17] to perform CTI, discuss identification guarantees, introduce the notion of *network coherence* for the analysis of interconnected networks, and support our discussions with illustrative simulations. In a more general scenario, and when the filters in the graph can be of different order (unknown) with possible transport delay (unknown), we show that the identification problem can be cast as the recovery of a *clustered-sparse* signal $\mathbf{x} \in \mathbb{R}^N$ from the measurements $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ with $M < N$, where the matrix A is a block-concatenation of

Toeplitz matrices. To this end, we introduce a greedy algorithm called Clustered Orthogonal Matching Pursuit (COMP) that tackles the problem of recovering clustered-sparse signals from few measurements. In a clustered-sparse model, in contrast to block-sparse models, there is no prior knowledge of the locations or the sizes of the clusters. We discuss the COMP algorithm and support our discussions with simulations.

In Chapter 5 we consider the problem of identifying Auto Regressive with eXternal input (ARX) models for both Linear Time-Invariant (LTI) and LTV systems. In the case of LTI ARX systems, a system with a large number of inputs and unknown input delays on each channel can require a model structure with a large number of parameters, unless input delay estimation is performed. Since the complexity of input delay estimation increases exponentially in the number of inputs, this can be difficult for high dimensional systems. We show that in cases where the LTI system has possibly many inputs with different unknown delays, simultaneous ARX identification and input delay estimation is possible from few observations, even though this leaves an apparently ill-conditioned identification problem. We discuss identification guarantees and support our proposed method with simulations.

We also consider the problem of identifying LTV ARX models. In particular, we consider systems with parameters that change only at a few time instants in a piecewise-constant manner where neither the change moments nor the number of changes is known a priori. The main technical novelty of our approach is in casting the identification problem as the recovery of a block-sparse signal from an underdetermined set of linear equations. We suggest a random sampling approach for LTV identification, address the issue of identifiability and again support our approach with illustrative simulations.

As an application of the proposed method for LTV identification, we apply the proposed method to experimental noisy input-output data from a DC motor system to demonstrate the effectiveness of the proposed identification method. From the data, we construct a multi-mode linear regression model where the mode-dependent coefficients correspond to torque, viscous friction, and static friction.

In Chapter 6 we consider the problem of characterizing the observability of linear systems with high-dimensional but sparse initial states. Recovering or estimating the initial state of a high-dimensional system can require a large number of measurements. In this chapter, we explain how this burden can be significantly reduced for certain linear systems when randomized measurement operators are employed. Our work builds upon recent results from CS regarding block-diagonal matrices.

A sufficient condition for stable recovery of a sparse high-dimensional vector is the Restricted Isometry Property (RIP). We show that the observability matrix satisfies the RIP under certain conditions on the state transition matrix. For example, we show that if the state transition matrix is unitary, and if independent, randomly-populated measurement matrices are employed, then it is possible to uniquely recover a sparse high-dimensional initial state when the total number of measurements scales *linearly* in the sparsity level of the initial state. We support our analysis with a case study of a diffusion system.

We then derive CoM inequalities for the observability matrix and explain how the interaction between the state transition matrix and the initial state affect the concentration bounds. The concentration results cover a larger class of systems (e.g., not necessarily unitary) and initial states (not necessarily sparse). Aside from guaranteeing recovery of sparse initial states, the CoM results have potential applications in solving inference problems such as detection and classification of more general initial states and systems.

We conclude in Chapter 7 with a final discussion and directions for future research. This thesis is a reflection of a series of inspiring collaborations. Where appropriate, the first page of each chapter includes a list of primary collaborators, who share the credit for this work.

CHAPTER 2

BACKGROUND

In this chapter, we first provide and list some of the general mathematical tools that we use throughout this thesis. We then summarize some of the basic concepts, definitions, and theorems in the field of Compressive Sensing (CS) which are useful throughout this thesis.

2.1 Mathematical Preliminaries and Notation

We start by providing some mathematical notation involving signals and matrices.

2.1.1 Signal Notation and Norms

Signal Notation

Most of the signals that we encounter in this thesis are real-valued and have discrete and finite domains. They are represented equivalently as either the function of integer time $x(t)$ for $t = 1, 2, \dots, N$, or grouped into a vector form using the boldface $\mathbf{x} \in \mathbb{R}^N$. In the case of complex-valued signals, we denote $\mathbf{x} \in \mathbb{C}^N$. Each of these assumptions will be made clear in the particular chapter or section.

Signal Norms

The p norm of a discrete signal \mathbf{x} of dimension N is denoted as $\|\mathbf{x}\|_p$ and is defined as

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^N |x(i)|^p \right)^{\frac{1}{p}}, \quad p \in [1, \infty). \quad (2.1)$$

The $\|\cdot\|_0$ and $\|\cdot\|_p$ for $p < 1$ do not meet the technical criteria for norms (e.g., triangle inequality). However, we usually refer to them as signal norms. The $\|\cdot\|_0$ and $\|\cdot\|_\infty$ norms do not follow the definition (2.1) and are defined as

$$\|\mathbf{x}\|_0 := \sum_{i=1}^N \mathbf{1}_{x(i) \neq 0},$$

where $\mathbf{1}$ denotes the indicator function and

$$\|\mathbf{x}\|_\infty := \max_i |x(i)|.$$

For any signal $\mathbf{x} \in \mathbb{R}^N$, the following inequality holds between these signal norms [18]:

$$\|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p \leq N^{(\frac{1}{p}-\frac{1}{q})} \|\mathbf{x}\|_q, \quad 0 < p \leq q < \infty.$$

2.1.2 Matrix Notation and Norms

Matrix Notation

We denote a real-valued $M \times N$ matrix using a capital letter, such as $A \in \mathbb{R}^{M \times N}$. A similar notation is used for complex-valued matrices such as $A \in \mathbb{C}^{M \times N}$. For a real-valued matrix A , we denote the transpose of A as A^T while for a complex-valued A , A^* denotes the conjugate transpose of A . Unless specified, we consider a matrix A to be real-valued.

Matrix Norms

There exist three important matrix norms that we encounter throughout this thesis, $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_*$. The *operator norm* of A (or induced ℓ_2 -norm) is denoted as $\|A\|_2$ and is equal to its largest singular value. Equivalently, $\|A\|_2$ is represented in terms of the singular values of A as $\|A\|_2 = \|\boldsymbol{\delta}\|_\infty$ where $\boldsymbol{\delta}$ is a vector containing the singular values of A . Similarly, the *Frobenius norm* of A is defined as $\|A\|_F = \|\boldsymbol{\delta}\|_2$, and the *nuclear norm* of A is defined as $\|A\|_* = \|\boldsymbol{\delta}\|_1$. For any matrix A , the following inequality holds between these matrix norms [18]:

$$\|A\|_2 \leq \|A\|_F \leq \|A\|_* \leq \sqrt{r} \|A\|_F \leq r \|A\|_2,$$

where r is the rank of A .

Matrix Representations

We can look at a matrix $A \in \mathbb{R}^{M \times N}$ as a mapping from \mathbb{R}^N into \mathbb{R}^M . We denote the *nullspace* of A as $\mathcal{N}(A)$ which is defined as $\mathcal{N}(A) := \{\mathbf{x} \in \mathbb{R}^N \mid A\mathbf{x} = \mathbf{0}\}$. The *rangespace* of A is defined as $\mathcal{R}(A) := \{\mathbf{b} \in \mathbb{R}^M \mid A\mathbf{x} = \mathbf{b}, \forall \mathbf{x} \in \mathbb{R}^N\}$. We can also view an $N \times P$ matrix $\Psi \in \mathbb{R}^{N \times P}$ as a *dictionary*¹ whose *words* are column vectors in \mathbb{R}^N . Given Ψ , a

¹A dictionary Ψ is simply a collection of words or elements. A combination of the elements of a given dictionary can be used to represent a given element. For example, if the elements of a dictionary are drawn

signal $\mathbf{s} \in \mathbb{R}^N$ can be represented as a linear combination of the columns of Ψ , denoted as $\boldsymbol{\psi}_i^\downarrow \in \mathbb{R}^N$. Therefore, we can write $\mathbf{s} = \sum_{i=1}^P x(i)\boldsymbol{\psi}_i^\downarrow = \Psi\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^P$. When a matrix (a dictionary) contains exactly N linearly independent columns, it is called a *basis*. Given a basis $\Psi \in \mathbb{R}^{N \times N}$, each signal $\mathbf{s} \in \mathbb{R}^N$ has a *unique* vector of expansion coefficients $\mathbf{x} = \Psi^{-1}\mathbf{s}$ (Ψ is a full-rank matrix, so it is invertible). If the columns of a basis are orthogonal (i.e., $\langle \boldsymbol{\psi}_i^\downarrow, \boldsymbol{\psi}_j^\downarrow \rangle = 0, \forall i, j, i \neq j$) of unit ℓ_2 norm, then that basis is called an *orthobasis*.

Among different bases, the $N \times N$ identity matrix (i.e., $\Psi = I_N$) (this is also known as the *canonical orthobasis*) is the most standard basis for representing a signal. Signals are often said to be in the *time domain* if they are represented in the canonical basis. Another standard domain to represent signals is the *frequency domain*. A signal $\mathbf{s} \in \mathbb{C}^N$ has a Discrete Fourier Transform (DFT) when it is represented in the *Fourier orthobasis* $\Psi = F_N \in \mathbb{C}^N$ as

$$F_N = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & \dots & w^{N-1} \\ 1 & w^2 & w^4 & w^6 & \dots & w^{2(N-1)} \\ 1 & w^3 & w^6 & w^9 & \dots & w^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & w^{3(N-1)} & \dots & w^{(N-1)^2} \end{bmatrix}, \quad (2.2)$$

with entries $F_N(\ell, m) = \frac{1}{\sqrt{N}}w^{(\ell-1)(m-1)}$, where $w = e^{-\frac{2\pi j}{N}}$ ($\sum_{k=0}^{N-1} w^k = 0$) is the N th root of unity. We can also consider a real-valued version of the Fourier orthobasis (2.2). Without loss of generality, suppose N is even. The real Fourier orthobasis R_N is constructed as follows. The first column of R_N equals the first column of F_N . Then $R_{\{2, \dots, \frac{N}{2}\}} = \text{Real}(\sqrt{2}F_{\{2, \dots, \frac{N}{2}\}})$ and $R_{\{\frac{N}{2}+1, \dots, N-1\}} = \text{Imaginary}(\sqrt{2}F_{\{2, \dots, \frac{N}{2}\}})$. The last column of R_N is equal to the $(\frac{N}{2} + 1)$ th column of F_N . Similar steps can be taken to construct a real Fourier orthobasis when N is odd. There exists a wide variety of other dictionaries which we will explain when needed.

2.1.3 Signal Models

In this section, we provide an overview of linear and nonlinear signal models. We start with linear signal models which are the most standard concepts in linear algebra and signal

from a signal space, the dictionary can be used to approximate signals in that space [19].

processing. For a given matrix $A \in \mathbb{R}^{M \times N}$, the nullspace of A , or $\mathcal{N}(A)$, is the set of signals $\mathbf{x} \in \mathbb{R}^N$ that satisfy a *linear* model $A\mathbf{x} = \mathbf{0}$ (a set of linear equalities). Similarly, we can consider the set of signals $\mathbf{x} \in \mathbb{R}^N$ that satisfy an *affine* model $A\mathbf{x} = \mathbf{b}$. In fact, this class of signals lives in a shifted nullspace $\mathbf{x}^* + \mathcal{N}(A)$, where \mathbf{x}^* is any solution to $A\mathbf{x}^* = \mathbf{b}$. Given A , another important linear model is the set of signals that are constructed using linear combinations of K *specific* columns of A . This set of signals forms a K -dimensional hyperplane in the ambient signal space. As a generalization of this model, one can consider the set of signals that are constructed using linear combinations of K columns of A , while the constructing columns (the K columns of A that are involved in the construction of a given signal) may change from one signal to another signal. In other words, such signals form a *union* of K -dimensional hyperplanes. One should note that for a matrix A with N columns, there are $\binom{N}{K}$ such hyperplanes (also known as K -planes), each spanned by K columns of A .

2.2 Compressive Sensing (CS)

CS, introduced by Candès, Romberg and Tao [13] and Donoho [14], is a powerful paradigm in signal processing which enables the recovery of an unknown vector from an underdetermined set of measurements under the assumption of sparsity of the signal and under certain conditions on the measurement matrix. The CS recovery problem can be viewed as recovery of a K -sparse signal $\mathbf{x} \in \mathbb{R}^N$ from its observations $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ where $A \in \mathbb{R}^{M \times N}$ is the measurement matrix with $M < N$ (in many cases $M \ll N$). A K -sparse signal $\mathbf{x} \in \mathbb{R}^N$ is a signal of length N with K non-zero entries where $K \ll N$. The notation $K := \|\mathbf{x}\|_0$ denotes the sparsity level of \mathbf{x} . Since the matrix $A \in \mathbb{R}^{M \times N}$ has a non-trivial nullspace when $M < N$, there exist infinitely many solutions to the equation $\mathbf{b} = A\mathbf{x}$, given \mathbf{b} . However, recovery of \mathbf{x} is indeed possible from CS measurements if the true signal is known to be sparse. Recovery of the true signal can be accomplished by seeking a sparse solution among these candidates. In the following, we review some of the algorithms for sparse recovery.

2.2.1 Recovery via ℓ_0 -minimization

Assume $\mathbf{x} \in \mathbb{R}^N$ is exactly K -sparse, i.e., out of its N entries only K are non-zero. Then, recovery of \mathbf{x} from measurements \mathbf{b} can be formulated as the ℓ_0 -minimization

$$\hat{\mathbf{x}} = \arg \min \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{b} = A\mathbf{x}. \quad (2.3)$$

It has been shown that [19, Theorem 2.1],[14] when A is populated with random entries (e.g., independent and identically distributed (i.i.d.) Gaussian random variables) and given some technical conditions on A , then with high probability the ℓ_0 -minimization problem (2.3) returns the proper K -sparse solution \mathbf{x} , even from an underdetermined set of linear equations.

2.2.2 Recovery via ℓ_1 -minimization

Solving the ℓ_0 -minimization problem (2.3) is known to be NP-hard. Thanks to the results of CS in regards to sparse signal recovery, however, it has been discovered that a much tractable minimization problem can be solved which often (under some conditions on the measurement matrix A) yields an equivalent solution. In this minimization problem, we seek the K -sparse solution by minimizing the ℓ_1 -norm of all candidate signals \mathbf{x} that satisfy the equality constraint $\mathbf{b} = A\mathbf{x}$ by solving [13, 14, 20–23]

$$\hat{\mathbf{x}} = \arg \min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{b} = A\mathbf{x}. \quad (2.4)$$

The ℓ_1 -minimization problem (2.4) is also known as Basis Pursuit (BP) [24] and can be solved via linear programming. In the following, we review some of the conditions under which the solutions to the ℓ_0 -minimization and ℓ_1 -minimization problems are equivalent.

2.2.3 ℓ_0/ℓ_1 Equivalence and the Restricted Isometry Property (RIP)

The Restricted Isometry Property (RIP), introduced by Candès and Tao [22], is one of the most fundamental recovery conditions that has been studied in the CS literature [25]. Establishing the RIP for a given matrix A guarantees that the solution to the ℓ_1 -minimization problem (2.4) is equivalent to the solution to the ℓ_0 -minimization problem (2.3). In the following, we provide the definition of the RIP.

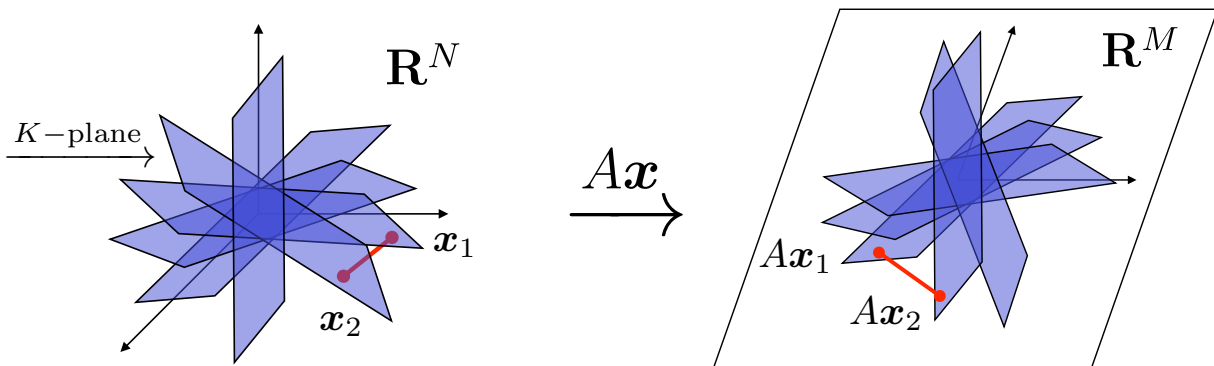


Figure 2.1: The RIP as a stable embedding. Each K -plane contains all K -sparse points in \mathbb{R}^N whose support (the locations of the non-zero entries) is the same.

Definition 2.5 A matrix $A \in \mathbb{R}^{M \times N}$ ($M < N$) is said to satisfy the RIP of order K with isometry constant $\delta_K \in (0, 1)$ if

$$(1 - \delta_K) \|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \delta_K) \|\mathbf{x}\|_2^2 \quad (2.6)$$

holds for all K -sparse signals $\mathbf{x} \in \mathbb{R}^N$ (i.e., $\forall \mathbf{x} \in \mathbb{R}^N$ with $\|\mathbf{x}\|_0 \leq K$).

If A satisfies the RIP of order $2K$ for a sufficiently small isometry constant δ_{2K} , then it is possible to *uniquely* recover any K -sparse signal \mathbf{x} from the measurements $\mathbf{b} = A\mathbf{x}$ using a tractable convex optimization program (the ℓ_1 -minimization problem) [14, 22, 26]. The RIP also ensures that the recovery process is robust to noise and stable in cases where \mathbf{x} is not precisely sparse [27]. Similar statements can be made for recovery using various iterative greedy algorithms [28–31]. A useful interpretation of this condition is in terms of the singular values. Establishing the RIP of order K with isometry constant δ_K (namely, $\text{RIP}(K, \delta_K)$) is equivalent to restricting all eigenvalues of all submatrices $A_{\mathcal{S}}^T A_{\mathcal{S}}$ to the interval $(1 - \delta_K, 1 + \delta_K)$. In this notation, $A_{\mathcal{S}}$ is an $M \times K$ submatrix of A whose columns are those columns of A indexed by the set \mathcal{S} with cardinality $|\mathcal{S}| \leq K$. Figure 2.1 illustrates the RIP as a stable embedding of the K -sparse points from \mathbb{R}^N into \mathbb{R}^M . Each K -plane contains all K -sparse points in \mathbb{R}^N whose support (the locations of the non-zero entries) is the same. Note that by definition of the RIP, one needs to show that the inequality (2.6)

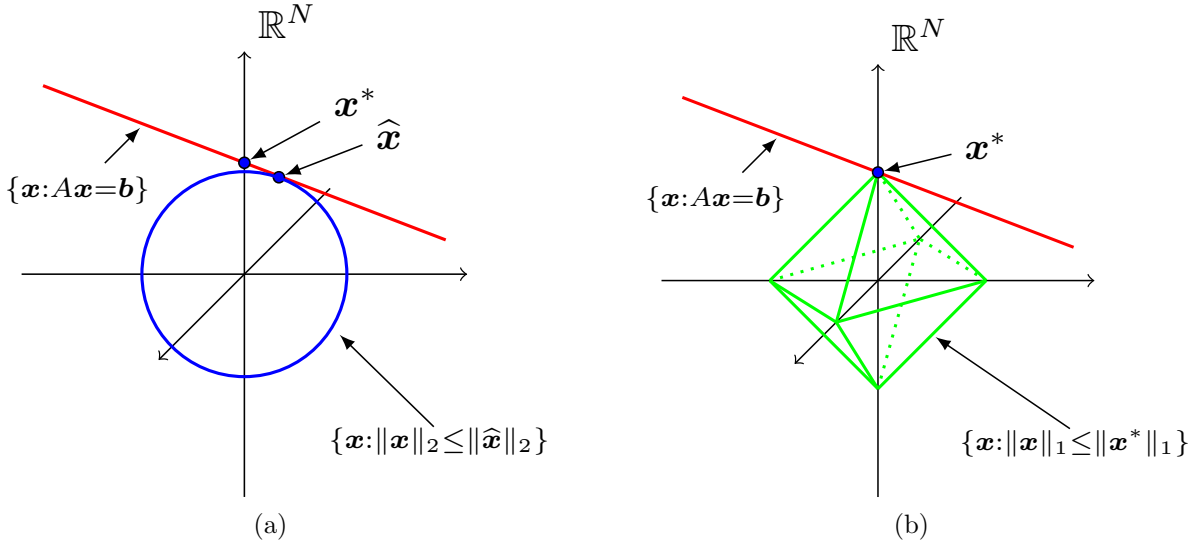


Figure 2.2: ℓ_2 -minimization versus ℓ_1 -minimization. (a) ℓ_2 -minimization. The recovered solution is almost never sparse as the ℓ_2 -ball does not have sharp edges along the axes. (b) ℓ_1 -minimization. If the hyperplane \mathcal{H}_b does not intersect the ℓ_1 -ball of radius $\|\mathbf{x}^*\|_1$, we have true sparse recovery due to the geometry of the ℓ_1 -ball (sharp edges along the axes).

holds for all the $\binom{N}{K}$ submatrices $A_S^T A_S$ [6, 32, 33]. Therefore, establishing the RIP for a given matrix is a combinatorial task. However, it has been shown that some specific matrix ensembles satisfy the RIP. A common way to establish the RIP for Φ is by populating Φ with random entries [32]. If Φ is populated with i.i.d. Gaussian random variables having mean zero and variance $\frac{1}{M}$, for example, then Φ satisfies the RIP of order S with very high probability when $\widetilde{M} \geq \delta_S^{-2} S \log \frac{N}{S}$. This result is significant because it indicates that the number of measurements sufficient for correct recovery scales *linearly* in the sparsity level and only *logarithmically* in the ambient dimension. Other random distributions may also be considered, including matrices with uniform entries of random signs. Consequently, a number of new sensing hardware architectures, from analog-to-digital converters to digital cameras, are being developed to take advantage of the benefits of random measurements [34–37]. There exist other recovery guarantees including the Exact Recovery Condition (ERC) [38] and the mutual coherence [39, 40] that have been proposed in the CS literature.

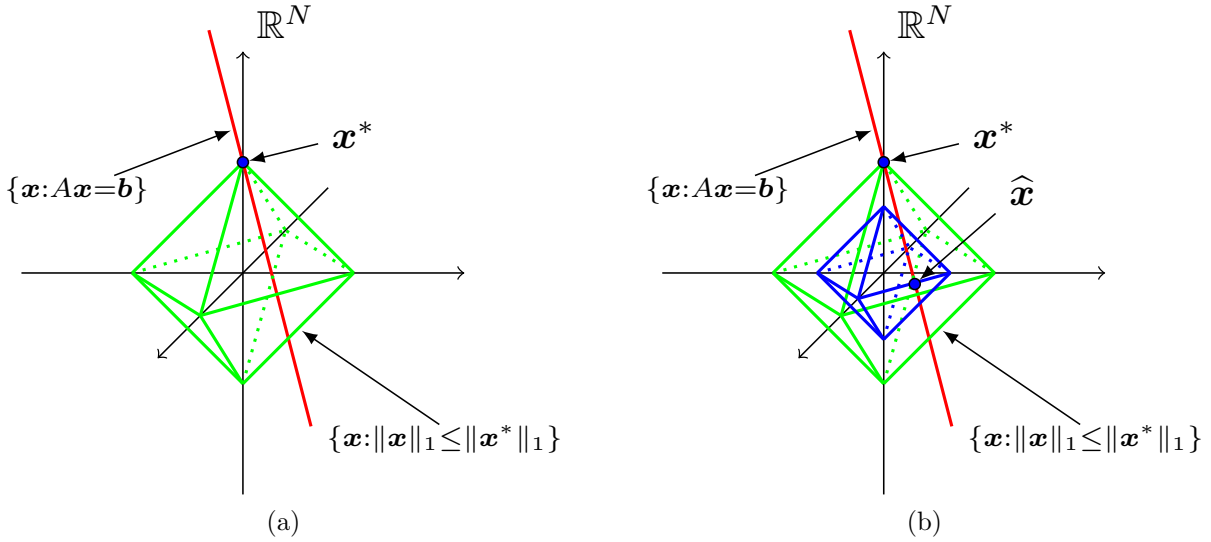


Figure 2.3: Incorrect recovery through ℓ_1 -minimization. (a) The hyperplane \mathcal{H}_b does intersect the ℓ_1 -ball of radius $\|x^*\|_1$. (b) The ℓ_1 -ball intersects the shifted nullspace at a point different than the true sparse solution, ending up with incorrect recovery.

2.2.4 CS Geometry

An interesting way of understanding ℓ_0/ℓ_1 equivalence is by looking at the geometry of such problems. Let us start by looking at the geometry of the ℓ_2 -minimization problem in which one seeks the solution with the smallest ℓ_2 -norm among other candidates. Figure 2.2(a) depicts such a problem. As can be seen, the recovered solution \hat{x} is not the true sparse solution x^* . One simple way to understand this is to observe that the ℓ_2 -ball does not have sharp edges along the axes, so the point where the ℓ_2 -ball intersects the shifted nullspace of A is usually not on one of the axes. On the other hand, the ℓ_1 -ball as shown in Figure 2.2(b) has sharp edges along the axes. In particular, given the measurements $\mathbf{b} = A\mathbf{x}^*$ where $A \in \mathbb{R}^{M \times N}$ with $M < N$, there exists an $(N - M)$ -dimensional hyperplane $\mathcal{H}_b = \{x \in \mathbb{R}^N : \mathbf{b} = Ax\} = \mathcal{N}(A) + \mathbf{x}^*$ of feasible signals as $A(\mathcal{N}(A) + \mathbf{x}^*) = A\mathbf{x}^* = \mathbf{b}$.

Assume that the true solution $x^* \in \mathbb{R}^N$ is K -sparse. The ℓ_1 -minimization problem will recover the true solution if and only if $\|x\|_1 > \|x^*\|_1$ for every other signal $x \in \mathcal{H}_b$ on the hyperplane. Geometrically, one should observe that when the hyperplane \mathcal{H}_b does not

intersect the ℓ_1 -ball of radius $\|\boldsymbol{x}^*\|_1$, then the ℓ_1 -minimization recovers the solution. This is shown in Figure 2.2(b). Similarly, if the hyperplane intersects the ℓ_1 -ball, we end up with incorrect recovery. This event is shown in Figure 2.3. For precise geometric arguments of this kind, one should refer to the works of Donoho and Tanner [41–43].

CHAPTER 3

COM INEQUALITIES FOR TOEPLITZ MATRICES

In this chapter² we derive Concentration of Measure (CoM) inequalities for randomized Toeplitz matrices. These inequalities show that the norm of a high-dimensional signal mapped by a Toeplitz matrix to a low-dimensional space concentrates around its mean with a tail probability bound that decays exponentially in the dimension of the range space divided by a quantity which is a function of the signal. For the class of *sparse* signals, the introduced quantity is bounded by the sparsity level of the signal. However, we observe that this bound is highly pessimistic for most sparse signals and we show that if a random distribution is imposed on the non-zero entries of the signal, the typical value of the quantity is bounded by a term that scales logarithmically in the ambient dimension. As an application of the CoM inequalities, we consider Compressive Binary Detection (CBD).

3.1 Introduction

Motivated to reduce the burdens of acquiring, transmitting, storing, and analyzing vast quantities of data, signal processing researchers have over the last few decades developed a variety of techniques for data compression and dimensionality reduction. Unfortunately, many of these techniques require a raw, high-dimensional data set to be acquired before its essential low-dimensional structure can be identified, extracted, and exploited. In contrast, what would be truly desirable are sensors/operators that require fewer raw measurements yet still capture the essential information in a data set. These operators can be called *compressive* in the sense that they act as mappings from a high-dimensional to a low-dimensional space, e.g., $X : \mathbb{R}^N \rightarrow \mathbb{R}^M$ where $M < N$. Linear compressive operators correspond to matrices having fewer rows than columns. Although such matrices can have arbitrary/deterministic entries, *randomized* matrices (those with entries drawn from a random distribution)

²This work is in collaboration with Tyrone L. Vincent and Michael B. Wakin [3–5].

have attracted the attention of researchers due to their universality and ease of analysis. Utilizing such compressive operators to achieve *information-preserving embeddings* of high-dimensional (but compressible) data sets into low-dimensional spaces can drastically simplify the acquisition process, reduce the needed amount of storage space, and decrease the computational demands of data processing.

CoM inequalities are one of the leading techniques used in the theoretical analysis of randomized compressive linear operators [44–46]. These inequalities quantify how well a random matrix will preserve the norm of a high-dimensional signal when mapping it to a low-dimensional space. A typical CoM inequality takes the following form. For any fixed signal $\mathbf{a} \in \mathbb{R}^N$, and a suitable random $M \times N$ matrix X , the random variable $\|X\mathbf{a}\|_2^2$ will be highly concentrated around its expected value, $\mathbf{E} [\|X\mathbf{a}\|_2^2]$, with high probability. Formally, there exist constants c_1 and c_2 such that for any fixed $\mathbf{a} \in \mathbb{R}^N$,

$$\mathbf{P} \{ \left| \|X\mathbf{a}\|_2^2 - \mathbf{E} [\|X\mathbf{a}\|_2^2] \right| \geq \epsilon \mathbf{E} [\|X\mathbf{a}\|_2^2] \} \leq c_1 e^{-c_2 M c_0(\epsilon)}, \quad (3.1)$$

where $c_0(\epsilon)$ is a positive constant that depends on $\epsilon \in (0, 1)$.

CoM inequalities for random operators have been shown to have important implications in signal processing and machine learning. One of the most prominent results in this area is the Johnson-Lindenstrauss (JL) lemma, which concerns embedding a finite set of points in a lower dimensional space using a distance preserving mapping [47]. Dasgupta et al. [45] and Achlioptas [46] showed how a CoM inequality of the form (3.1) could establish that with high probability, an independent and identically distributed (i.i.d.) random compressive operator $X \in \mathbb{R}^{M \times N}$ ($M < N$) provides a JL-embedding. Specifically, for a given $\epsilon \in (0, 1)$, for any fixed point set $Q \subseteq \mathbb{R}^N$,

$$(1 - \epsilon) \|\mathbf{a} - \mathbf{b}\|_2^2 \leq \|X\mathbf{a} - X\mathbf{b}\|_2^2 \leq (1 + \epsilon) \|\mathbf{a} - \mathbf{b}\|_2^2 \quad (3.2)$$

holds with high probability for all $\mathbf{a}, \mathbf{b} \in Q$ if $M = \mathcal{O}(\epsilon^{-2} \log(|Q|))$. One of the other significant consequences of CoM inequalities is in the context of Compressive Sensing (CS) [25] and the Restricted Isometry Property (RIP) [22]. If a matrix X satisfies (3.2) for all pairs

\mathbf{a} , \mathbf{b} of K -sparse signals in \mathbb{R}^N , then X is said to satisfy the RIP of order $2K$ with isometry constant ϵ . Establishing the RIP of order $2K$ for a given compressive matrix X leads to understanding the number of measurements required to have exact recovery for any K -sparse signal $\mathbf{a} \in \mathbb{R}^N$. Baraniuk et al. [32] and Mendelson et al. [48] showed that CoM inequalities can be used to prove the RIP for random compressive matrices.

CoM inequalities have been well-studied and derived for *unstructured* random compressive matrices, populated with i.i.d. random entries [45, 46]. However, in many practical applications, measurement matrices possess a certain structure. In particular, when linear dynamical systems are involved, Toeplitz and circulant matrices appear due to the convolution process [2, 3, 6, 49]. Specifically, consider the Linear Time-Invariant (LTI) dynamical system with system finite impulse response $\mathbf{a} = \{a_k\}_{k=1}^N$. Let $\mathbf{x} = \{x_k\}_{k=1}^{N+M-1}$ be the applied input sequence. Then the corresponding output is calculated from the time-domain convolution of \mathbf{a} and \mathbf{x} . Supposing the x_k and a_k sequences are zero-padded from both sides, each output sample y_k can be written as

$$y_k = \sum_{j=1}^N a_j x_{k-j}. \quad (3.3)$$

If we keep only M consecutive observations of the system, $\mathbf{y} = \{y_k\}_{k=N+1}^{N+M}$, then (3.3) can be written in matrix-vector multiplication format as

$$\mathbf{y} = X\mathbf{a}, \quad (3.4)$$

where

$$X = \begin{bmatrix} x_N & x_{N-1} & \cdots & x_1 \\ x_{N+1} & x_N & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{N+M-1} & x_{N+M-2} & \cdots & x_M \end{bmatrix} \quad (3.5)$$

is an $M \times N$ Toeplitz matrix. If the entries of X are generated randomly, we say X is a randomized Toeplitz matrix. Other types of *structured* random matrices also arise when dynamical systems are involved. For example block-diagonal matrices appear in applications such as distributed sensing systems [50] and initial state estimation (observability) of linear

systems [11].

In this chapter, we consider compressive randomized Toeplitz matrices, derive CoM inequalities, and discuss their implications in applications such as sparse impulse response recovery [6, 51, 52]. We also consider the problem of detecting a deviation in a system's behavior. We show that by characterizing the deviation using a particular measure that appears in our CoM inequality, the detector performance can be correctly predicted.

3.2 Related Work

Compressive Toeplitz (and circulant) matrices have been previously studied in the context of CS [2, 6, 49, 51, 53, 54], with applications involving channel estimation, synthetic aperture radar, etc. Tropp et al. [53] originally considered compressive Toeplitz matrices in an early CS paper that proposed an efficient measurement mechanism involving a Finite Impulse Response (FIR) filter with random taps. Motivated by applications related to sparse channel estimation, Bajwa et al. [54] studied such matrices more formally in the case where the matrix entries are drawn from a symmetric Bernoulli distribution. Later they extended this study to random matrices whose entries are bounded or Gaussian-distributed and showed that with high probability, $M \geq \mathcal{O}\left(K^2 \log\left(\frac{N}{K}\right)\right)$ measurements are sufficient to establish the RIP of order $2K$ for vectors sparse in the time domain [6, 51]. (It should be noted that the quadratic RIP result can also be achieved using other methods such as a coherence argument [49, 55].) Recently, using more complicated mathematical tools such as Dudley's inequality for chaos and generic chaining, Rauhut et al. [33] showed that with $M \geq \mathcal{O}\left(K^{1.5} \log(N)^{1.5}\right)$ measurements the RIP of order $2K$ will hold.³ Note that these bounds compare to $M \geq \mathcal{O}\left(K \log\left(\frac{N}{K}\right)\right)$ measurements which are known to suffice when X is unstructured [32].

In this chapter, we derive CoM inequalities for Toeplitz matrices and show how these inequalities reveal non-uniformity and signal-dependency of the mappings. As one conse-

³In their very recent work, Krahmer et al. [56] showed that the minimal required number of measurements scales linearly with K , or formally $M \geq \mathcal{O}\left(K \log(K)^2 \log(N)^2\right)$ measurements are sufficient to establish the RIP of order $2K$. The recent linear RIP result confirms what is suggested by simulations.

quence of these CoM inequalities, one could use them (along with standard covering number estimates) to prove the RIP for compressive Toeplitz matrices. Although the estimate of the required number of measurements would be quadratic in terms of sparsity (i.e., $M \sim K^2$) and fall short of the best known estimates described above, studying concentration inequalities for Toeplitz matrices is of its own interest and gives insight to other applications such as the binary detection problem.

There also exist CoM analyses for other types of structured matrices. For example, Park et al. [50] derived concentration bounds for two types of block diagonal compressive matrices, one in which the blocks along the diagonal are random and independent, and one in which the blocks are random but equal.⁴ We subsequently extended these CoM results for block diagonal matrices to the observability matrices that arise in the analysis of linear dynamical systems [11]. We will address the observability problem in details in Chapter 6 of this thesis.

3.3 Contributions

In summary, we derive CoM inequalities for randomized Toeplitz matrices. The derived bounds in the inequalities are non-uniform and depend on a quantity which is a function of the signal. For the class of *sparse* signals, the introduced quantity is bounded by the sparsity level of the signal while if a random distribution is imposed on the non-zero entries of the signal, the typical value of the quantity is bounded by a term that scales logarithmically in the ambient dimension. As an application of the CoM inequalities, we consider CBD.

3.4 Main Results

In this work, we derive CoM bounds for compressive Toeplitz matrices as given in (3.5) with entries $\{x_k\}_{k=1}^{N+M-1}$ drawn from an i.i.d. Gaussian random sequence. Our first main result, detailed in Theorem 3.7, states that the upper and lower tail probability bounds

⁴Shortly after our own development of CoM inequalities for compressive Toeplitz matrices (a preliminary version of Theorem 3.7 appeared in [3]), Yap and Rozell [57] showed that similar inequalities can be derived by extending the CoM results for block diagonal matrices. Our Theorem 3.15 and the associated discussion, however, is unique to this work.

depend on the number of measurements M and on the eigenvalues of the covariance matrix of the vector \mathbf{a} defined as

$$P(\mathbf{a}) = \begin{bmatrix} \mathcal{R}_{\mathbf{a}}(0) & \mathcal{R}_{\mathbf{a}}(1) & \cdots & \mathcal{R}_{\mathbf{a}}(M-1) \\ \mathcal{R}_{\mathbf{a}}(1) & \mathcal{R}_{\mathbf{a}}(0) & \cdots & \mathcal{R}_{\mathbf{a}}(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{R}_{\mathbf{a}}(M-1) & \mathcal{R}_{\mathbf{a}}(M-2) & \cdots & \mathcal{R}_{\mathbf{a}}(0) \end{bmatrix}, \quad (3.6)$$

where $\mathcal{R}_{\mathbf{a}}(\tau) := \sum_{i=1}^{N-\tau} a_i a_{i+\tau}$ denotes the un-normalized sample autocorrelation function of $\mathbf{a} \in \mathbb{R}^N$.

Theorem 3.7 *Let $\mathbf{a} \in \mathbb{R}^N$ be fixed. Define two quantities $\rho(\mathbf{a})$ and $\mu(\mathbf{a})$ associated with the eigenvalues of the covariance matrix $P(\mathbf{a})$ as $\rho(\mathbf{a}) := \frac{\max_i \lambda_i}{\|\mathbf{a}\|_2^2}$ and $\mu(\mathbf{a}) := \frac{\sum_{i=1}^M \lambda_i^2}{M\|\mathbf{a}\|_2^4}$, where λ_i is the i -th eigenvalue of $P(\mathbf{a})$. Let $\mathbf{y} = X\mathbf{a}$, where X is a random compressive Toeplitz matrix with i.i.d. Gaussian entries having zero mean and unit variance. Noting that $\mathbf{E}[\|\mathbf{y}\|_2^2] = M\|\mathbf{a}\|_2^2$, then for any $\epsilon \in (0, 1)$, the upper tail probability bound is*

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 - M\|\mathbf{a}\|_2^2 \geq \epsilon M\|\mathbf{a}\|_2^2 \right\} \leq e^{-\frac{\epsilon^2 M}{8\rho(\mathbf{a})}} \quad (3.8)$$

and the lower tail probability bound is

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 - M\|\mathbf{a}\|_2^2 \leq -\epsilon M\|\mathbf{a}\|_2^2 \right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\mathbf{a})}}. \quad (3.9)$$

Theorem 3.7 provides CoM inequalities for *any* (not necessarily sparse) signal $\mathbf{a} \in \mathbb{R}^N$. The significance of these results comes from the fact that the tail probability bounds are functions of the signal \mathbf{a} , where the dependency is captured in the quantities $\rho(\mathbf{a})$ and $\mu(\mathbf{a})$. This is not the case when X is unstructured. Indeed, allowing X to have $M \times N$ i.i.d. Gaussian entries with zero mean and unit variance (and thus, no Toeplitz structure) would result in the concentration inequality (see, e.g., [46])

$$\mathbf{P} \left\{ \left| \|\mathbf{y}\|_2^2 - M\|\mathbf{a}\|_2^2 \right| \geq \epsilon M\|\mathbf{a}\|_2^2 \right\} \leq 2e^{-\frac{\epsilon^2 M}{4}}. \quad (3.10)$$

Thus, comparing the bound in (3.10) with the ones in (3.8) and (3.9), one could conclude that achieving the same probability bound for Toeplitz matrices requires choosing M larger by a

factor of $2\rho(\mathbf{a})$ or $2\mu(\mathbf{a})$. Typically, when using CoM inequalities such as (3.8) and (3.9), we must set M large enough so that both bounds are sufficiently small over all signals \mathbf{a} belonging to some class of interest. For example, we are often interested in signals that have a *sparse* representation. Because we generally wish to keep M as small as possible, it is interesting to try to obtain an upper bound for the important quantities $\rho(\mathbf{a})$ and $\mu(\mathbf{a})$ over the class of signals of interest. It is easy to show that for all $\mathbf{a} \in \mathbb{R}^N$, $\mu(\mathbf{a}) \leq \rho(\mathbf{a})$. Thus, we limit our analysis to finding the sharpest upper bound for $\rho(\mathbf{a})$ when \mathbf{a} is K -sparse. For the sake of generality, we allow the signal to be sparse in an arbitrary orthobasis.

Definition 3.11 *A signal $\mathbf{a} \in \mathbb{R}^N$ is called K -sparse in an orthobasis $G \in \mathbb{R}^{N \times N}$ if it can be represented as $\mathbf{a} = G\mathbf{q}$, where $\mathbf{q} \in \mathbb{R}^N$ is K -sparse (a vector with $K < N$ non-zero entries).*

We also introduce the notion of *K -sparse Fourier coherence* of the orthobasis G . This measures how strongly the columns of G are correlated with the length L Fourier basis, $F_L \in \mathbb{C}^{L \times L}$, which has entries $F_L(\ell, m) = \frac{1}{\sqrt{L}}w^{(\ell-1)(m-1)}$, where $w = e^{-\frac{2\pi j}{L}}$.

Definition 3.12 *Given an orthobasis $G \in \mathbb{R}^{N \times N}$ and measurement length M , let $L = N + M - 1$. The K -sparse Fourier coherence of G , denoted $\nu_K(G)$, is defined as*

$$\nu_K(G) := \max_{i,S} \|F_{1:N}^{i \rightarrow} G_S\|_2, \quad (3.13)$$

where $S \subseteq \{1, 2, \dots, N\}$ is the support set and varies over all possible sets with cardinality $|S| = K$, $G_S \in \mathbb{R}^{N \times K}$ is a matrix containing the columns of $G \in \mathbb{R}^{N \times N}$ indexed by the support set S , and $F_{1:N}^{i \rightarrow} \in \mathbb{C}^N$ is a row vector containing the first N entries of the i -th row of the Fourier orthobasis $F_L \in \mathbb{C}^{L \times L}$. Observe that for a given orthobasis G , $\nu_K(G)$ depends on K .

Using the notion of Fourier coherence, we show in Section 3.6 that for all vectors $\mathbf{a} \in \mathbb{R}^N$ that are K -sparse in an orthobasis $G \in \mathbb{R}^{N \times N}$,

$$\rho(\mathbf{a}) \leq L\nu_K^2(G), \quad (3.14)$$

where, as above, $L = N + M - 1$. This bound, however, appears to be highly pessimistic for most K -sparse signals. As a step towards better understanding the behavior of $\rho(\mathbf{a})$, we consider a random model for \mathbf{a} . In particular, we consider a fixed K -sparse support set, and on this set we suppose the K non-zero entries of the coefficient vector \mathbf{q} are drawn from a random distribution. Based on this model, we derive an upper bound for $\mathbf{E}[\rho(\mathbf{a})]$.

Theorem 3.15 (Upper Bound on $\mathbf{E}[\rho(\mathbf{a})]$) *Let $\mathbf{q} \in \mathbb{R}^N$ be a random K -sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \frac{1}{K})$. Select the measurement length M , which corresponds to the dimension of $P(\mathbf{a})$, and set $L = N + M - 1$. Let $\mathbf{a} = G\mathbf{q}$ where $G \in \mathbb{R}^{N \times N}$ is an orthobasis. Then*

$$\mathbf{E}[\rho(\mathbf{a})] \leq \frac{8L\nu_K^2(G)}{K} (\log 2L + 2). \quad (3.16)$$

The K -sparse Fourier coherence $\nu_K(G)$ and consequently the bounds (3.14) and (3.16) can be explicitly evaluated for some specific orthobases G . For example, letting $G = I_N$ (the $N \times N$ identity matrix), we can consider signals that are sparse in the time domain. With this choice of G , one can show that $\nu_K(I_N) = \sqrt{\frac{K}{L}}$. As another example, we can consider signals that are sparse in the frequency domain. To do this, we set G equal to a real-valued version of the Fourier orthobasis. (Construction of a real-valued Discrete Fourier Transform (DFT) matrix is explained in Chapter 2 of this thesis.) With this choice of G , one can show that $\nu_K(R_N) \leq \sqrt{\frac{N}{L}}$. Using these upper bounds on the Fourier coherence, we have the following deterministic bounds on $\rho(\mathbf{a})$ in the time and frequency domains:

$$\rho(\mathbf{a}) \leq K \quad (\text{time domain sparsity}) \quad \text{and} \quad (3.17)$$

$$\rho(\mathbf{a}) \leq N \quad (\text{frequency domain sparsity}). \quad (3.18)$$

We also obtain bounds on the expected value of $\rho(\mathbf{a})$ under the random signal model as:

$$\mathbf{E}[\rho(\mathbf{a})] \leq 8(\log 2L + 2) \quad (\text{time domain sparsity}) \quad \text{and} \quad (3.19)$$

$$\mathbf{E}[\rho(\mathbf{a})] \leq \frac{8N}{K}(\log 2L + 2) \quad (\text{frequency domain sparsity}). \quad (3.20)$$

We offer a brief interpretation and analysis of these bounds in this paragraph and several examples that follow. First, because $K \leq N$, the deterministic and expectation bounds on $\rho(\mathbf{a})$ are smaller for signals that are sparse in the time domain than for signals that are sparse in the frequency domain. The simulations described in Examples 3.22 and 3.23 below confirm that, on average, $\rho(\mathbf{a})$ does indeed tend to be smaller under the model of time domain sparsity. Second, these bounds exhibit varying dependencies on the sparsity level K : (3.17) increases with K and (3.20) decreases with K , while (3.18) and (3.19) are agnostic to K . The simulation described in Example 3.22 below confirms that, on average, $\rho(\mathbf{a})$ increases with K for signals that are sparse in the time domain but decreases with K for signals that are sparse in the frequency domain. This actually reveals a looseness in (3.19); however, in Section 3.8, we conjecture a sparsity-dependent expectation bound that closely matches the empirical results for signals that are sparse in the time domain. Third, under both models of sparsity, and assuming $8(\log 2L + 2) \ll K$ for signals of practical interest, the expectation bounds on $\rho(\mathbf{a})$ are qualitatively lower than the deterministic bounds. This raises the question of whether the deterministic bounds are sharp. We confirm that this is the case in Example 3.24 below.

Example 3.21 (*Illustrating the signal-dependency of the left-hand side of CoM inequalities (3.8) and (3.9)*) *In this example, we illustrate that the CoM behavior for randomized Toeplitz matrices is indeed signal-dependent. We consider inequality (3.8) while a similar analysis can be made for (3.9). We consider two particular K -sparse ($K = 64$) signals, \mathbf{a}_1 and \mathbf{a}_2 both in \mathbb{R}^N ($N = 1024$) where the K non-zero entries of \mathbf{a}_1 have equal values and occur in the first K entries of the vector ($\rho(\mathbf{a}_1) = 63.26$), while the K non-zero entries of \mathbf{a}_2 appear in a randomly-selected locations with random signs and values ($\rho(\mathbf{a}_2) = 5.47$). Both*

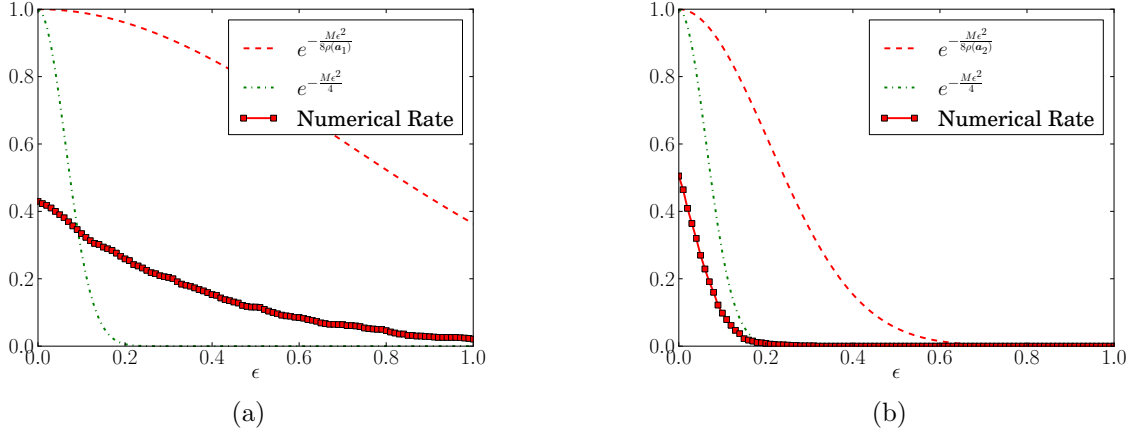


Figure 3.1: Illustrating the signal-dependency of the left-hand side of CoM inequality. With fixed $N = 1024$, $K = 64$ and $M = 512$, we consider two particular K -sparse signals, \mathbf{a}_1 , and \mathbf{a}_2 . Both \mathbf{a}_1 and \mathbf{a}_2 are normalized. We measure each of these signals with 1000 i.i.d. Gaussian $M \times N$ Toeplitz matrices. (a) The K non-zero entries of \mathbf{a}_1 have equal values and occur in the first K entries of the vector. (b) The K non-zero entries of \mathbf{a}_2 appear in randomly-selected locations with random signs and values. The two signals have different concentrations that can be upper bounded by the signal-dependent bound.

\mathbf{a}_1 and \mathbf{a}_2 are normalized. For a fixed $M = 512$, we measure each of these signals with 1000 i.i.d. Gaussian $M \times N$ Toeplitz matrices.

Figure 3.1 depicts the numerically determined rate of occurrence of the event $\|\mathbf{y}\|_2^2 - M\|\mathbf{a}\|_2^2 \geq \epsilon M\|\mathbf{a}\|_2^2$ over 1000 trials versus $\epsilon \in (0, 1)$. For comparison, the derived analytical bound in (3.8) (for Toeplitz X) as well as the bound in (3.10) (for unstructured X) is depicted. As can be seen the two signals have different concentrations. In particular, \mathbf{a}_1 (Figure 3.1(a)) has worse concentration compared to \mathbf{a}_2 (Figure 3.1(b)) when measured by Toeplitz matrices. This signal-dependency of the concentration does not happen when these signals are measured by unstructured Gaussian random matrices. Moreover, the derived signal-dependent bound (3.8) successfully upper bounds the numerical event rate of occurrence for each signal while the bound (3.10) for unstructured X fails to do so. Also observe that the analytical bound $e^{-\frac{M\epsilon^2}{8\rho(\mathbf{a}_2)}}$ in Figure 3.1(b) can not bound the numerical event rate of occurrence for \mathbf{a}_1 as depicted in Figure 3.1(a).

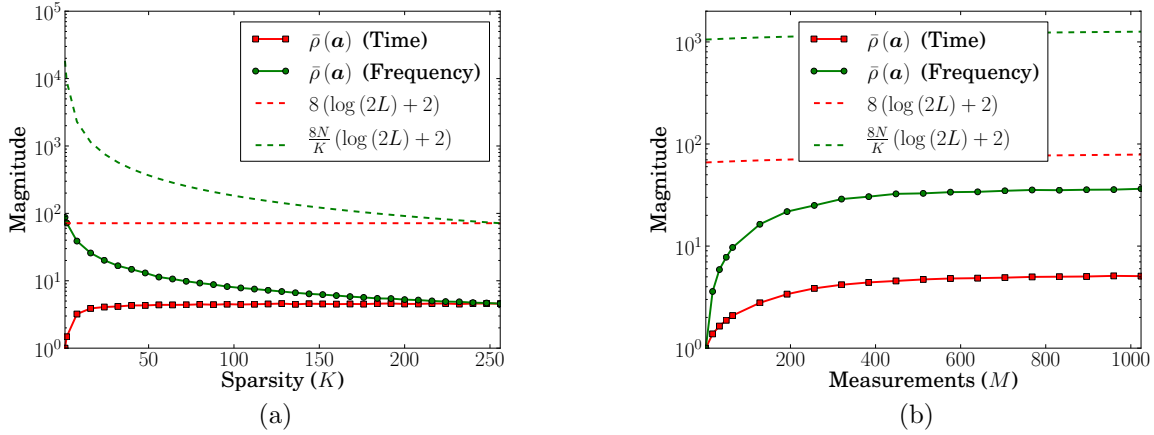


Figure 3.2: Sample mean of $\rho(\mathbf{a})$ in the time and frequency domains versus the expectation bounds, where $L = N + M - 1$. The sample mean $\bar{\rho}(\mathbf{a})$ is calculated by taking the mean over 1000 constructed signals \mathbf{a} . A logarithmic scale is used for the vertical axis. (a) Varying K with fixed $M = N = 256$. (b) Varying M with fixed $N = 512$ and $K = 32$.

Example 3.22 (*Varying K and comparing the time and frequency domains*) In this experiment, we fix M and N . For each value of K and each sparse basis $G = I_N$ and $G = R_N$, we construct 1000 random sparse vectors $\mathbf{q} \in \mathbb{R}^N$ with random support and having K non-zero entries drawn from a Gaussian distribution with mean zero and variance $\frac{1}{K}$. For each vector, we compute $\mathbf{a} = G\mathbf{q}$, and we then let $\bar{\rho}(\mathbf{a})$ denote the sample mean of $\rho(\mathbf{a})$ across these 1000 signals. The results, as a function of K , are plotted in Figure 3.2(a). As anticipated, signals that are sparse in the frequency domain have a larger value of $\bar{\rho}(\mathbf{a})$ than signals that are sparse in the time domain. Moreover, $\bar{\rho}(\mathbf{a})$ decreases with K in the frequency domain but increases with K in the time domain. Overall, the empirical behavior of $\bar{\rho}(\mathbf{a})$ is mostly consistent with the bounds (3.19) and (3.20), although our constants may be larger than necessary.

Example 3.23 (*Varying M and comparing the time and frequency domains*) This experiment is identical to the one in Example 3.22, except that we vary M while keeping K and N fixed. The results are plotted in Figure 3.2(b). Once again, signals that are sparse in the frequency domain have a larger value of $\bar{\rho}(\mathbf{a})$ than signals that are sparse in the time domain.

Moreover, in both cases $\bar{\rho}(\mathbf{a})$ appears to increase logarithmically with M as predicted by the bounds in (3.19) and (3.20), although our constants may be larger than necessary.

Example 3.24 (*Confirming the tightness of the deterministic bounds*) We fix N , consider a vector $\mathbf{a} \in \mathbb{R}^N$ that is K -sparse in the time domain, and suppose the K non-zero entries of \mathbf{a} take equal values and occur in the first K entries of the vector. For such a vector one can derive a lower bound on $\rho(\mathbf{a})$ by embedding $P(\mathbf{a})$ inside a circulant matrix, applying the Cauchy Interlacing Theorem [58] (we describe these steps more fully in Section 3.6.1), and then performing further computations that we omit for the sake of space. With these steps, one concludes that for this specific vector $\mathbf{a} \in \mathbb{R}^N$,

$$\rho(\mathbf{a}) \geq K \left(1 - \frac{\pi^2}{24} \left(\frac{K^2}{M + K - 1} \right)^2 \right)^2. \quad (3.25)$$

When $M \gg K^2$, the right-hand side of (3.25) approaches K . This confirms that (3.17) is sharp for large M .

In the remainder of this chapter, the proofs of Theorem 3.7 and Theorem 3.15 are presented, followed by additional discussions concerning the relevance of the main results. Our results have important consequences in the analysis of high-dimensional dynamical systems. We expound on this fact by exploring a CBD problem in Section 3.9.

3.5 Proof of Theorem 3.7

The proofs of the upper and lower bounds of Theorem 3.7 are given separately in Lemmas 3.31 and 3.35 below. The proofs utilize Markov's inequality along with a suitable bound on the moment generating function of $\|\mathbf{y}\|_2^2$, given in Lemma 3.26. Observe that for a fixed vector $\mathbf{a} \in \mathbb{R}^N$ and a random Gaussian Toeplitz matrix $X \in \mathbb{R}^{M \times N}$, the vector $\mathbf{y} = X\mathbf{a} \in \mathbb{R}^M$ will be a Gaussian random vector with zero mean and $M \times M$ covariance matrix $P(\mathbf{a})$ given in (3.6).

Lemma 3.26 *If $\mathbf{y} \in \mathbb{R}^M$ is a zero mean Gaussian random vector with covariance matrix P , then*

$$\mathbf{E} \left[e^{t\mathbf{y}^T \mathbf{y}} \right] = \frac{1}{\sqrt{\det(I_M - 2tP)}} \quad (3.27)$$

holds for all $t \in (-\infty, \frac{1}{2\lambda_{\max}(P)})$.

Proof

$$\begin{aligned} \mathbf{E} \left[e^{t\mathbf{y}^T \mathbf{y}} \right] &= \int \frac{1}{(2\pi)^{\frac{M}{2}} \det^{\frac{1}{2}}(P)} e^{t\mathbf{y}^T \mathbf{y}} e^{-\frac{1}{2}\mathbf{y}^T P^{-1} \mathbf{y}} d\mathbf{y} \\ &= \int \frac{1}{(2\pi)^{\frac{M}{2}} \det^{\frac{1}{2}}(P)} e^{-\frac{1}{2}\mathbf{y}^T (P^{-1} - 2tI_M) \mathbf{y}} d\mathbf{y} = \frac{\det^{\frac{1}{2}}((P^{-1} - 2tI_M)^{-1})}{\det^{\frac{1}{2}}(P)} \\ &= \frac{1}{(\det(P^{-1} - 2tI_M) \det P)^{\frac{1}{2}}} = \frac{1}{\sqrt{\det(I_M - 2tP)}}. \end{aligned}$$

■

Observe that as a special case of Lemma 3.26, if $y \in \mathbb{R}$ is a scalar Gaussian random variable of unit variance, then we obtain the well known result of $\mathbf{E} \left[e^{ty^2} \right] = \frac{1}{\sqrt{1-2t}}$, for $t \in (-\infty, \frac{1}{2})$. Based on Lemma 3.26, we use Chernoff's bounding method [59] for computing the upper and lower tail probability bounds. In particular, we are interested in finding bounds for the tail probabilities

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \geq M\|\mathbf{a}\|_2^2(1 + \epsilon) \right\} \quad (3.28a)$$

and

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \leq M\|\mathbf{a}\|_2^2(1 - \epsilon) \right\}. \quad (3.28b)$$

Observe that in (3.28a) and (3.28b) concentration behavior is sought around $\mathbf{E} [\|\mathbf{y}\|_2^2] = M\|\mathbf{a}\|_2^2$. For a random variable z , and all $t > 0$,

$$\mathbf{P} \{ z \geq \epsilon \} = \mathbf{P} \{ e^{tz} \geq e^{t\epsilon} \} \leq \frac{\mathbf{E} [e^{tz}]}{e^{t\epsilon}} \quad (3.29)$$

(see, e.g., [59]). Applying (3.29) to (3.28a), for example, and then applying Lemma 3.26 yields

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 > M\|\mathbf{a}\|_2^2 (1 + \epsilon) \right\} \leq \frac{\mathbf{E} \left[e^{t\mathbf{y}^T \mathbf{y}} \right]}{e^{M\|\mathbf{a}\|_2^2(1+\epsilon)t}} = (\det (I_M - 2tP))^{-\frac{1}{2}} e^{-M\|\mathbf{a}\|_2^2(1+\epsilon)t}. \quad (3.30)$$

In (3.30), $t \in (-\infty, \frac{1}{2(\max_i \lambda_i)})$ is a free variable which—as in a Chernoff bound—can be varied to make the right-hand side as small as possible. Though not necessarily optimal, we propose to use $t = \frac{\epsilon}{2(1+\epsilon)f(\mathbf{a})\|\mathbf{a}\|_2^2}$, where f is a function of \mathbf{a} that we specify below. We state the upper tail probability bound in Lemma 3.31 and the lower tail probability bound in Lemma 3.35.

Lemma 3.31 *Let $\mathbf{a} \in \mathbb{R}^N$ be fixed, let $P = P(\mathbf{a})$ be as given in (3.6), and let $\mathbf{y} \in \mathbb{R}^M$ be a zero mean Gaussian random vector with covariance matrix P . Then, for any $\epsilon \in (0, 1)$,*

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \geq M\|\mathbf{a}\|_2^2 (1 + \epsilon) \right\} \leq e^{-\frac{\epsilon^2 M}{8\rho(\mathbf{a})}}. \quad (3.32)$$

Proof Choosing t as

$$t = \frac{\epsilon}{2(1+\epsilon)\rho(\mathbf{a})\|\mathbf{a}\|_2^2}$$

and noting that $t \in (-\infty, \frac{1}{2\max_i \lambda_i})$, the right-hand side of (3.30) can be written as

$$\left(\left(\det \left(I_M - \frac{\epsilon}{(1+\epsilon)} \frac{P}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2} \right) \right)^{-\frac{1}{M}} e^{-\frac{\epsilon}{\rho(\mathbf{a})}} \right)^{\frac{M}{2}}. \quad (3.33)$$

This expression can be simplified. Note that

$$\begin{aligned} \det \left(I_M - \frac{\epsilon}{(1+\epsilon)} \frac{P}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2} \right) &= \prod_{i=1}^M \left(1 - \frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2} \right) \\ &= e^{\sum_{i=1}^M \log \left(1 - \frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2} \right)}. \end{aligned}$$

Using the facts that $\log(1 - c_1 c_2) \geq c_2 \log(1 - c_1)$ for any $c_1, c_2 \in [0, 1]$ and $\text{Tr}(P) = M\|\mathbf{a}\|_2^2$, we have

$$\begin{aligned}
e^{\sum_{i=1}^M \log\left(1 - \frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2}\right)} &\geq e^{\sum_{i=1}^M \frac{\lambda_i}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2} \log\left(1 - \frac{\epsilon}{1+\epsilon}\right)} \\
&= e^{\frac{\text{Tr}(P)}{\rho(\mathbf{a})\|\mathbf{a}\|_2^2} \log\left(\frac{1}{1+\epsilon}\right)} = e^{\frac{M}{\rho(\mathbf{a})} \log\left(\frac{1}{1+\epsilon}\right)} = \left(\frac{1}{1+\epsilon}\right)^{\frac{M}{\rho(\mathbf{a})}}.
\end{aligned} \tag{3.34}$$

Combining (3.30), (3.33), and (3.34) gives us

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 > M\|\mathbf{a}\|_2^2(1+\epsilon) \right\} \leq \left(\left(\frac{1}{1+\epsilon}\right)^{-\frac{1}{\rho(\mathbf{a})}} e^{-\frac{\epsilon}{\rho(\mathbf{a})}} \right)^{\frac{M}{2}} = ((1+\epsilon)e^{-\epsilon})^{\frac{M}{2\rho(\mathbf{a})}}.$$

The final bound comes by noting that $(1+\epsilon)e^{-\epsilon} \leq e^{-\epsilon^2/4}$. ■

Lemma 3.35 *Using the same assumptions as in Lemma 3.31, for any $\epsilon \in (0, 1)$,*

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \leq M\|\mathbf{a}\|_2^2(1-\epsilon) \right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\mathbf{a})}}.$$

Proof Applying Markov's inequality to (3.28b), we obtain

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \leq M\|\mathbf{a}\|_2^2(1-\epsilon) \right\} = \mathbf{P} \left\{ -\|\mathbf{y}\|_2^2 \geq -M\|\mathbf{a}\|_2^2(1-\epsilon) \right\} \leq \frac{\mathbf{E} \left[e^{-t\mathbf{y}^T \mathbf{y}} \right]}{e^{-M\|\mathbf{a}\|_2^2(1-\epsilon)t}}. \tag{3.36}$$

Using Lemma 3.26, this implies that

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \leq M\|\mathbf{a}\|_2^2(1-\epsilon) \right\} \leq \left((\det(I_M + 2tP))^{-\frac{1}{M}} e^{2\|\mathbf{a}\|_2^2(1-\epsilon)t} \right)^{\frac{M}{2}}. \tag{3.37}$$

In this case, we choose

$$t = \frac{\epsilon}{2(1+\epsilon)\mu(\mathbf{a})\|\mathbf{a}\|_2^2},$$

and note that $t > 0$. Plugging t into (3.37) and following similar steps as for the upper tail bound, we get

$$\det(I_M + 2tP) = \prod_{i=1}^M \left(1 + \frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\mu(\mathbf{a})\|\mathbf{a}\|_2^2} \right) = e^{\sum_{i=1}^M \log\left(1 + \frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\mu(\mathbf{a})\|\mathbf{a}\|_2^2}\right)}. \tag{3.38}$$

Since $\log(1+c) \geq c - \frac{c^2}{2}$ for $c > 0$,

$$\sum_{i=1}^M \log \left(1 + \frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\mu(\mathbf{a}) \|\mathbf{a}\|_2^2} \right) \quad (3.39)$$

$$\begin{aligned} &\geq \sum_{i=1}^M \left(\frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\mu(\mathbf{a}) \|\mathbf{a}\|_2^2} - \frac{1}{2} \left(\frac{\epsilon}{(1+\epsilon)} \frac{\lambda_i}{\mu(\mathbf{a}) \|\mathbf{a}\|_2^2} \right)^2 \right) \\ &= \frac{\epsilon}{(1+\epsilon)} \frac{\sum_{i=1}^M \lambda_i}{\mu(\mathbf{a}) \|\mathbf{a}\|_2^2} - \frac{1}{2} \left(\frac{\epsilon}{(1+\epsilon)} \frac{\sum_{i=1}^M \lambda_i}{\mu(\mathbf{a}) \|\mathbf{a}\|_2^2} \right)^2 \\ &= \frac{\epsilon}{(1+\epsilon)} \frac{M}{\mu(\mathbf{a})} - \frac{1}{2} \left(\frac{\epsilon}{1+\epsilon} \right)^2 \frac{M}{\mu(\mathbf{a})} = \frac{M}{\mu(\mathbf{a})} \left(\frac{\epsilon^2 + 2\epsilon}{2(1+\epsilon)^2} \right). \end{aligned} \quad (3.40)$$

Combining (3.38) and (3.40) gives the bound

$$\det(I_M + 2tP) \geq e^{\frac{M}{\mu(\mathbf{a})} \left(\frac{\epsilon^2 + 2\epsilon}{2(1+\epsilon)^2} \right)} = \left(e^{\frac{\epsilon^2 + 2\epsilon}{2(1+\epsilon)^2}} \right)^{\frac{M}{\mu(\mathbf{a})}}. \quad (3.41)$$

By substituting (3.41) into (3.37), we obtain

$$\mathbf{P} \left\{ \|\mathbf{y}\|_2^2 \leq M \|\mathbf{a}\|_2^2 (1 - \epsilon) \right\} \leq \left(e^{\frac{-\epsilon^2 - 2\epsilon}{2(1+\epsilon)^2}} e^{\frac{\epsilon(1-\epsilon)}{1+\epsilon}} \right)^{\frac{M}{2\mu(\mathbf{a})}} = \left(e^{\frac{-2\epsilon^3 - \epsilon^2}{2(1+\epsilon)^2}} \right)^{\frac{M}{2\mu(\mathbf{a})}}.$$

The final bound comes by noting that $e^{\frac{-2\epsilon^3 - \epsilon^2}{2(1+\epsilon)^2}} \leq e^{-\epsilon^2/4}$. ■

3.6 Proof and Discussion of Theorem 3.15

In this section, we first provide the proof of Theorem 3.15. We then discuss some of the implications of the result. An important step towards the proof is based on the so-called ‘‘circulant embedding’’ explained in the next section.

3.6.1 Circulant Embedding

The covariance matrix $P(\mathbf{a})$ described in (3.6) is an $M \times M$ symmetric Toeplitz matrix which can be decomposed as $P(\mathbf{a}) = A^T A$, where A is an $(N + M - 1) \times M$ Toeplitz matrix (as shown in Figure 3.3) and A^T is the transpose of A . In order to derive an upper bound on the maximum eigenvalue of $P(\mathbf{a})$, we embed the matrix A inside its $(N + M - 1) \times (N + M - 1)$

$$A = \begin{bmatrix} a_1 & 0 & \dots & 0 \\ a_2 & \ddots & & (0) \\ \vdots & \ddots & a_1 & \vdots \\ a_N & & a_2 & \ddots & 0 \\ 0 & \ddots & \vdots & \ddots & a_1 \\ \vdots & & a_N & a_2 & \\ & (0) & & \ddots & \vdots \\ 0 & & \dots & 0 & a_N \end{bmatrix} \quad A_c = \begin{bmatrix} a_1 & 0 & \dots & 0 & a_N & \dots & a_2 \\ a_2 & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & a_1 & (0) & & \ddots & a_N \\ a_N & & a_2 & \ddots & & & 0 \\ 0 & \ddots & \vdots & \ddots & a_1 & & \\ \vdots & & a_N & a_2 & \ddots & \ddots & \vdots \\ & (0) & & \ddots & \vdots & \ddots & \ddots & 0 \\ 0 & & \dots & 0 & a_N & \dots & a_2 & a_1 \end{bmatrix}$$

Figure 3.3: Toeplitz matrix $A \in \mathbb{R}^{L \times M}$ and its circulant counterpart $A_c \in \mathbb{R}^{L \times L}$ where $L = N + M - 1$.

circulant counterpart A_c where each column of A_c is a cyclic downward shifted version of the previous column. Thus, A_c is uniquely determined by its first column, which we denote by

$$\tilde{\mathbf{a}} = \left[\underbrace{a_1 \cdots a_N}_{\mathbf{a}^T} \quad \underbrace{0 \cdots 0}_{(M-1) \text{ zeros}} \right]^T \in \mathbb{R}^L,$$

where $L = N + M - 1$. Observe that the circulant matrix $A_c \in \mathbb{R}^{L \times L}$ contains the Toeplitz matrix $A \in \mathbb{R}^{L \times M}$ in its first M columns. Because of this embedding, the Cauchy Interlacing Theorem [58] implies that $\max_{i \leq M} \lambda_i(A^T A) \leq \max_{i \leq L} \lambda_i(A_c^T A_c)$. Therefore, we have

$$\rho(\mathbf{a}) = \frac{\max_i \lambda_i(P(\mathbf{a}))}{\|\mathbf{a}\|_2^2} = \frac{\max_i \lambda_i(A^T A)}{\|\mathbf{a}\|_2^2} \leq \frac{\max_i \lambda_i(A_c^T A_c)}{\|\mathbf{a}\|_2^2} = \frac{\max_i |\lambda_i(A_c^T)|^2}{\|\mathbf{a}\|_2^2} =: \rho_c(\mathbf{a}). \quad (3.42)$$

Thus, an upper bound for $\rho(\mathbf{a})$ can be achieved by bounding the maximum absolute eigenvalue of A_c^T . Since A_c^T is circulant, its eigenvalues are given by the un-normalized length- L DFT of the first row of A_c^T (the first column of A_c). Specifically, for $i = 1, 2, \dots, L$,

$$\lambda_i(A_c^T) = \sum_{k=1}^L \tilde{a}_k e^{-\frac{2\pi j}{L}(i-1)(k-1)} = \sum_{k=1}^N a_k e^{-\frac{2\pi j}{L}(i-1)(k-1)}. \quad (3.43)$$

Recall that $F_L \in \mathbb{C}^{L \times L}$ is the Fourier orthobasis with entries $F_L(\ell, m) = \frac{1}{\sqrt{L}} w^{(\ell-1)(m-1)}$ where $w = e^{-\frac{2\pi j}{L}}$, and let $F_L^{i \rightarrow} \in \mathbb{C}^L$ be the i -th row of F_L . Using matrix-vector notation, (3.43) can be written as

$$\lambda_i(A_c^T) = \sqrt{L}F_L^{i\rightarrow}\tilde{\mathbf{a}} = \sqrt{L}F_{1:N}^{i\rightarrow}\mathbf{a} = \sqrt{L}F_{1:N}^{i\rightarrow}G\mathbf{q} = \sqrt{L}F_{1:N}^{i\rightarrow}G_S\mathbf{q}_S, \quad (3.44)$$

where $F_{1:N}^{i\rightarrow} \in \mathbb{C}^N$ is a row vector containing the first N entries of $F_L^{i\rightarrow}$, $\mathbf{q}_S \in \mathbb{R}^K$ is the part of $\mathbf{q} \in \mathbb{R}^N$ restricted to the support S (the location of the non-zero entries of \mathbf{q}) with cardinality $|S| = K$, and $G_S \in \mathbb{R}^{N \times K}$ contains the columns of $G \in \mathbb{R}^{N \times N}$ indexed by the support S .

3.6.2 Deterministic Bound

We can bound $\rho(\mathbf{a})$ over all sparse \mathbf{a} using the Cauchy-Schwarz inequality. From (3.44), it follows for any $i \in \{1, 2, \dots, L\}$ that

$$|\lambda_i(A_c^T)| = |\sqrt{L}F_{1:N}^{i\rightarrow}G_S\mathbf{q}_S| \leq \sqrt{L}\|F_{1:N}^{i\rightarrow}G_S\|_2\|\mathbf{q}_S\|_2 = \sqrt{L}\|F_{1:N}^{i\rightarrow}G_S\|_2\|\mathbf{a}\|_2. \quad (3.45)$$

By combining Definition 3.12, (3.42), and (3.45), we arrive at the deterministic bound (3.14). This bound appears to be highly pessimistic for *most* sparse vectors \mathbf{a} . In other words, although in Example 3.24 we illustrate that for a specific signal \mathbf{a} , the deterministic bound (3.14) is tight when $M \gg K$, we observe that for many other classes of sparse signals \mathbf{a} , the bound is pessimistic. In particular, if a random model is imposed on the non-zero entries of \mathbf{a} , an upper bound on the typical value of $\rho(\mathbf{a})$ derived in (3.19) scales logarithmically in the ambient dimension L which is qualitatively smaller than K . We show this analysis in the proof of Theorem 3.15. In order to make this proof self-contained, we first list some results that we will draw from.

3.6.3 Supporting Results

We utilize the following propositions.

Lemma 3.46 [55] *Let z be any random variable. Then*

$$\mathbf{E}[|z|] = \int_0^\infty \mathbf{P}\{|z| \geq x\} dx. \quad (3.47)$$

Lemma 3.48 *Let z_1 and z_2 be positive random variables. Then for any U ,*

$$\mathbf{P} \{z_1 + z_2 \geq U\} \leq \mathbf{P} \left\{ z_1 \geq \frac{U}{2} \right\} + \mathbf{P} \left\{ z_2 \geq \frac{U}{2} \right\}, \quad (3.49)$$

and for any U_1 and U_2 ,

$$\mathbf{P} \left\{ \frac{z_1}{z_2} \geq \frac{U_1}{U_2} \right\} \leq \mathbf{P} \{z_1 \geq U_1\} + \mathbf{P} \{z_2 \leq U_2\}. \quad (3.50)$$

Proof See Appendix A.1. ■

Proposition 3.51 [46] (Concentration Inequality for Sums of Squared Gaussian Random Variables) *Let $\mathbf{q} \in \mathbb{R}^N$ be a random K -sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma^2)$. Then for any $\epsilon > 0$,*

$$\mathbf{P} \{ \|\mathbf{q}\|_2^2 \leq K\sigma^2(1 - \epsilon) \} \leq e^{-\frac{K\epsilon^2}{4}}.$$

Proposition 3.52 (Hoeffding's Inequality for Complex-Valued Gaussian Sums) *Let $\mathbf{b} \in \mathbb{C}^N$ be fixed, and let $\epsilon \in \mathbb{R}^N$ be a random vector whose N entries are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma^2)$. Then, for any $u > 0$,*

$$\mathbf{P} \left\{ \left| \sum_{i=1}^N \epsilon_i b_i \right| \geq u \right\} \leq 2e^{-\frac{u^2}{4\sigma^2 \|\mathbf{b}\|_2^2}}.$$

Proof See Appendix A.2. ■

In order to prove Theorem 3.15, we also require a tail probability bound for the eigenvalues of A_c^T .

Proposition 3.53 *Let $\mathbf{q} \in \mathbb{R}^N$ be a random K -sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with*

$\mathcal{N}(0, \frac{1}{K})$. Let $\mathbf{a} = G\mathbf{q}$ where $G \in \mathbb{R}^{N \times N}$ is an orthobasis, and let A_c be an $L \times L$ circulant matrix, where the first N entries of the first column of A_c are given by \mathbf{a} . Then for any $u > 0$, and for $i = 1, 2, \dots, L$,

$$\mathbf{P} \{ |\lambda_i(A_c)| \geq u \} \leq 2e^{-\frac{u^2 K}{4L\nu_K^2(G)}}. \quad (3.54)$$

Proof Define the row vector $\mathbf{b} = \sqrt{L}F_{1:N}^{i \rightarrow} G_S \in \mathbb{C}^K$. From (3.44) and the Cauchy-Schwarz inequality, it follows that $|\lambda_i(A_c)| = |\lambda_i(A_c^T)| = |\sqrt{L}F_{1:N}^{i \rightarrow} G_S \mathbf{q}_S| = |\sum_{i=1}^K \epsilon_i b_i|$, where $\epsilon_i = (\mathbf{q}_S)_i$. From Definition 3.12, we have $\|\mathbf{b}\|_2 \leq \sqrt{L}\nu_K(G)$. The tail probability bound (3.54) follows from applying Proposition 3.52. \blacksquare

3.6.4 Completing the Proof of Theorem 3.15

From (3.42), we have

$$\begin{aligned} \mathbf{E}[\rho(\mathbf{a})] &\leq \mathbf{E}[\rho_c(\mathbf{a})] = \mathbf{E} \left[\frac{\max_i |\lambda_i(A_c^T)|^2}{\|\mathbf{a}\|_2^2} \right] = \int_0^\infty \mathbf{P} \left\{ \frac{\max_i |\lambda_i(A_c^T)|^2}{\|\mathbf{a}\|_2^2} \geq x \right\} dx \\ &= \int_0^{L\nu_K^2(G)} \mathbf{P} \left\{ \frac{\max_i |\lambda_i(A_c^T)|^2}{\|\mathbf{a}\|_2^2} \geq x \right\} dx, \end{aligned}$$

where the last equality comes from the deterministic upper bound

$|\lambda_i(A_c^T)| \leq \sqrt{L}\|F_{1:N}^{i \rightarrow} G_S\|_2 \|\mathbf{a}\|_2 \leq \sqrt{L}\nu_K(G) \|\mathbf{a}\|_2$. Using a union bound, for any $t > 0$ we have

$$\begin{aligned} \int_0^{L\nu_K^2(G)} \mathbf{P} \left\{ \frac{\max_i |\lambda_i(A_c^T)|^2}{\|\mathbf{a}\|_2^2} \geq x \right\} dx &= \int_0^{L\nu_K^2(G)} \mathbf{P} \left\{ \frac{\max_i |\lambda_i(A_c^T)|^2}{\|\mathbf{a}\|_2^2} \geq \frac{tx}{t} \right\} dx \\ &\leq \int_0^{L\nu_K^2(G)} \mathbf{P} \left\{ \max_i |\lambda_i(A_c^T)|^2 \geq tx \right\} dx \\ &\quad + \int_0^{L\nu_K^2(G)} \mathbf{P} \left\{ \|\mathbf{a}\|_2^2 \leq t \right\} dx. \end{aligned} \quad (3.55)$$

The first term in the right hand side of (3.55) can be bounded as follows. For every $\delta \geq 0$, by partitioning the range of integration [55, 60], we obtain

$$\begin{aligned}
\int_0^{L\nu_K^2(G)} \mathbf{P} \left\{ \max_i |\lambda_i(A_c^T)|^2 \geq tx \right\} dx &\leq \int_0^\infty \mathbf{P} \left\{ \max_i |\lambda_i(A_c^T)|^2 \geq tx \right\} dx \\
&\leq \delta + \int_\delta^\infty \mathbf{P} \left\{ \max_i |\lambda_i(A_c)|^2 \geq tx \right\} dx \\
&\leq \delta + \int_\delta^\infty \sum_{i=1}^L \mathbf{P} \left\{ |\lambda_i(A_c)|^2 \geq tx \right\} dx \\
&\leq \delta + \int_\delta^\infty \sum_{i=1}^L 2e^{-\frac{Ktx}{4L\nu_K^2(G)}} dx \\
&= \delta + 2L \int_\delta^\infty e^{-\frac{Ktx}{4L\nu_K^2(G)}} dx \\
&= \delta + \frac{8L^2\nu_K^2(G)}{Kt} e^{-\frac{Kt\delta}{4L\nu_K^2(G)}},
\end{aligned}$$

where we used Proposition 3.53 in the last inequality. The second term in (3.55) can be bounded using the concentration inequality of Proposition 3.51. We have for $0 < t \leq 1$, $\mathbf{P} \{ \|\mathbf{a}\|_2^2 \leq t \} \leq e^{-\frac{K(1-t)^2}{4}}$. Putting together the bounds for the two terms of inequality (3.55), we have

$$\mathbf{E} [\rho(\mathbf{a})] \leq \mathbf{E} [\rho_c(\mathbf{a})] \leq \delta + \frac{8L^2\nu_K^2(G)}{Kt} e^{-\frac{Kt\delta}{4L\nu_K^2(G)}} + L\nu_K^2(G) e^{-\frac{K(1-t)^2}{4}}. \quad (3.56)$$

Now we pick δ to minimize the upper bound in (3.56). Using the minimizer $\delta^* = \frac{4L\nu_K^2(G) \log 2L}{Kt}$ yields

$$\mathbf{E} [\rho(\mathbf{a})] \leq \mathbf{E} [\rho_c(\mathbf{a})] \leq \frac{4L\nu_K^2(G)}{Kt} \left(\log 2L + 1 + \frac{Kt}{4} e^{-\frac{K(1-t)^2}{4}} \right). \quad (3.57)$$

Let $g(K, t) := \frac{Kt}{4} e^{-\frac{K(1-t)^2}{4}}$. It is trivial to show that $g(K, 0.5) \leq 1$ for all K (for $t = 0.5$, $\max_K g(K, 0.5) = \frac{2}{e}$). Therefore, $\mathbf{E} [\rho(\mathbf{a})] \leq \frac{8L\nu_K^2(G)}{K} (\log 2L + 2)$, which completes the proof. \blacksquare

3.7 Discussion

Remark 3.58 *In Theorem 3.15, we derived an upper bound on $\mathbf{E} [\rho(\mathbf{a})]$ by finding an upper bound on $\mathbf{E} [\rho_c(\mathbf{a})]$ and using the fact that for all vectors \mathbf{a} , we have $\rho(\mathbf{a}) \leq \rho_c(\mathbf{a})$. However, we should note that this inequality gets tighter as M (the number of columns of A) increases.*

For small M the interlacing technique results in a looser bound.

Remark 3.59 By taking $G = I_N$ and noting that $\nu_K(I_N) = \sqrt{\frac{K}{L}}$, (3.57) leads to an upper bound on $\mathbf{E}[\rho_c(\mathbf{a})]$ when the signal \mathbf{a} is K -sparse in the time domain (specifically, $\mathbf{E}[\rho_c(\mathbf{a})] \leq 8(\log 2L + 2)$). Although this bound scales logarithmically in the ambient dimension L , it does not show a dependency on the sparsity level K of the vector \mathbf{a} . Over multiple simulations where we have computed the sample mean $\bar{\rho}_c(\mathbf{a})$, we have observed a linear behavior of the quantity $\frac{K}{\bar{\rho}_c(\mathbf{a})}$ as K increases, and this leads us to the conjecture below. Although at this point we are not able to prove the conjecture, the proposed bound matches closely with empirical data.

Conjecture 3.60 Fix N and M . Let $\mathbf{a} \in \mathbb{R}^N$ be a random K -sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \frac{1}{K})$. Then

$$\mathbf{E}[\rho_c(\mathbf{a})] \sim \frac{K}{c_1 K + c_2},$$

where $c_1 = \frac{1}{c \log L}$ for some constant c , and $c_2 = 1 - c_1$.

The conjecture follows from our empirical observation that $\frac{K}{\bar{\rho}_c(\mathbf{a})} \sim c_1 K + c_2$ for some constants c_1 and c_2 , the fact that $\rho_c(\mathbf{a}) = 1$ for $K = 1$, and the observation that $\bar{\rho}_c(\mathbf{a}) \sim c \log L$ when $K = N$ for large N . In the following examples, we illustrate these points and show how the conjectured bound can sharply approximate the empirical mean of $\rho_c(\mathbf{a})$.

Example 3.61 In this experiment, we fix $M = 256$ and take $G = I_N$. For each value of N , we construct 1000 random non-sparse vectors $\mathbf{a} \in \mathbb{R}^N$ whose N entries are drawn from a Gaussian distribution with mean zero and variance $\frac{1}{N}$. We let $\bar{\rho}_c(\mathbf{a})$ denote the sample mean of $\rho_c(\mathbf{a})$ across these 1000 signals. The results, as a function of N , are plotted in Figure 3.4. Also plotted is the function $f(L) = \log(L)$ where $L = N + M - 1$; this closely approximates the empirical data.

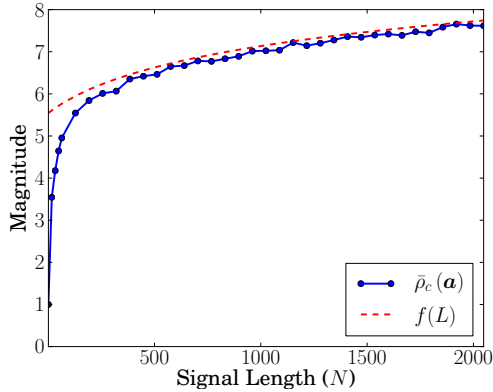


Figure 3.4: Empirical results. Sample mean of $\rho_c(\mathbf{a})$ in the time domain for full vectors $\mathbf{a} \in \mathbb{R}^N$ where $M = 256$ is fixed. Also plotted is $f(L) = \log(L)$, where $L = N + M - 1$.

Example 3.62 In this experiment, we fix $N = 1024$. For each value of K , we construct 1000 random sparse vectors $\mathbf{a} \in \mathbb{R}^N$ with random support and having K non-zero entries drawn from a Gaussian distribution with mean zero and variance $\frac{1}{K}$. We let $\bar{\rho}_c(\mathbf{a})$ denote the sample mean of $\rho_c(\mathbf{a})$ across these 1000 signals. The results, as a function of K for two fixed values $M = 1$ and $M = 1024$, are plotted in Figure 3.5.

Remark 3.63 As a final note in this section, the result of Theorem 3.15 can be easily extended to the case when $G \in \mathbb{C}^{N \times N}$ is a complex orthobasis and \mathbf{q} and \mathbf{a} are complex vectors. The bounds can be derived in a similar way.

3.8 A Quadratic RIP Bound and Non-uniform Recovery

An approach identical to the one taken by Baraniuk et al. [32] can be used to establish the RIP for Toeplitz matrices X based on the CoM inequalities given in Theorem 3.7. As mentioned in Section 3.4, the bounds of the CoM inequalities for Toeplitz matrices are looser by a factor of $2\rho(\mathbf{a})$ or $2\mu(\mathbf{a})$ as compared to the ones for unstructured X . Since $\rho(\mathbf{a})$ is bounded by K for all K -sparse signals in the time domain (the deterministic bound), with straightforward calculations a quadratic estimate of the number of measurements in terms of sparsity ($M \sim K^2$) can be achieved for Toeplitz matrices. As mentioned earlier, on the

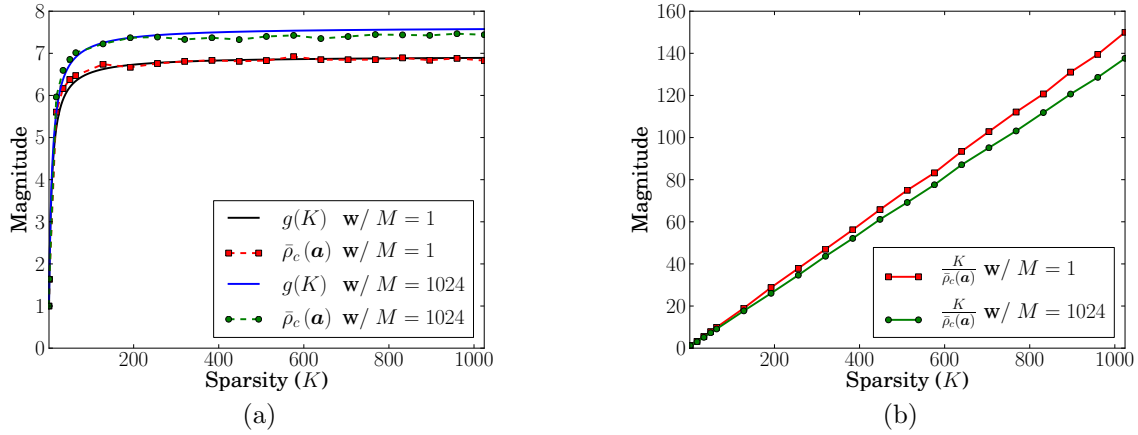


Figure 3.5: *Empirical results. (a) Simulation results vs. the conjectured bound $g(K) = \frac{K}{c_1K+c_2}$ with $c = 1$. (b) Linearity of $\frac{K}{\bar{\rho}_c(\mathbf{a})}$.*

other hand, there exists an extremely non-uniform distribution of $\rho(\mathbf{a})$ over the set of all K -sparse signals \mathbf{a} , for as Theorem 3.15 states, if a random model is imposed on \mathbf{a} , an upper bound on the typical value of $\rho(\mathbf{a})$ scales logarithmically in the ambient dimension L . This suggests that for most K -sparse signals \mathbf{a} the value of $\rho(\mathbf{a})$ is much smaller than K (observe that $8(\log 2L + 2) \ll K$ for many signals of practical interest). Only for a very small set of signals does the value of $\rho(\mathbf{a})$ approach the deterministic bound of K . One can show, for example, that for any K -sparse signal whose K non-zero entries are all the same, we have $\rho(\mathbf{a}) \leq \rho_c(\mathbf{a}) = K$ (Example 3.24). This non-uniformity of $\rho(\mathbf{a})$ over the set of sparse signals may be useful for proving a non-uniform recovery bound or for strengthening the RIP result; our work on these fronts remains in progress. Using different techniques than pursued here (non-commutative Khintchine type inequalities), a non-uniform recovery result with a linear estimate of M in K up to log-factors has been proven by Rauhut [49]. For a detailed description of non-uniform recovery and its comparison to uniform recovery, one could refer to a paper by Rauhut [Sections 3.1 and 4.2, [55]]. The behavior of $\rho(\mathbf{a})$ also has important implications in the binary detection problem which we discuss in the next section.

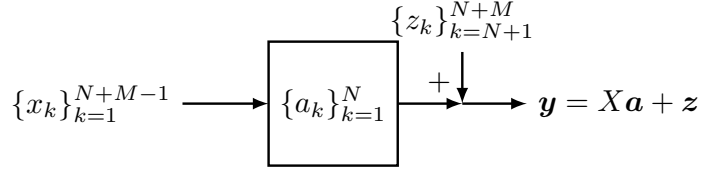


Figure 3.6: *FIR filter of order N with impulse response $\{a_k\}_{k=1}^N$.*

3.9 Compressive Binary Detection (CBD)

In this section, as an application of the CoM inequalities for Toeplitz matrices, we consider a CBD problem.

3.9.1 Problem Setup

In this section, we address the problem of detecting a change in the dynamics of a linear system. We aim to perform the detection from the smallest number of observations, and for this reason, we call this problem Compressive Binary Detection (CBD).

We consider an FIR filter with a known impulse response $\mathbf{a} = \{a_k\}_{k=1}^N$. The response of this filter to a test signal $\mathbf{x} = \{x_k\}_{k=1}^{N+M-1}$ is described in (3.3). We suppose the output of this filter is corrupted by random additive measurement noise \mathbf{z} . Figure 3.6 shows the schematic of this measurement process.

From a collection of M measurements \mathbf{y} with $M < N$, our specific goal is to detect whether the dynamics of the system have changed to a different impulse response $\mathbf{b} = \{b_k\}_{k=1}^N$, which we also assume to be known. Since the nominal impulse response \mathbf{a} is known, the expected response $X\mathbf{a}$ can be subtracted off from \mathbf{y} , and thus without loss of generality, our detection problem can be stated as follows [61]: Distinguish between two events which we define as $\mathcal{E}_0 \triangleq \{\mathbf{y} = \mathbf{z}\}$ and $\mathcal{E}_1 \triangleq \{\mathbf{y} = X\mathbf{c} + \mathbf{z}\}$, where $\mathbf{c} = \mathbf{b} - \mathbf{a}$ and \mathbf{z} is a vector of i.i.d. Gaussian noise with variance σ^2 .

For any detection algorithm, one can define the false-alarm probability as $P_{FA} \triangleq \mathbf{P}\{(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_0)\}$ and the detection probability $P_D \triangleq \mathbf{P}\{(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_1)\}$. A Receiver Operating Curve (ROC) is a plot of P_D as a function of P_{FA} . A Neyman-Pearson

(NP) detector maximizes P_D for a given limit on the failure probability, $P_{FA} \leq \alpha$. The NP test for our problem can be written as $\mathbf{y}^T X \mathbf{c} \underset{\mathcal{E}_0}{\overset{\mathcal{E}_1}{\geq}} \gamma$, where the threshold γ is chosen to meet the constraint $P_{FA} \leq \alpha$. Consequently, we consider the detection statistic $d := \mathbf{y}^T X \mathbf{c}$. By evaluating d and comparing to the threshold γ , we are now able to decide between the two events \mathcal{E}_0 and \mathcal{E}_1 . To fix the failure limit, we set $P_{FA} = \alpha$ which leads to

$$P_D(\alpha) = Q\left(Q^{-1}(\alpha) - \frac{\|X\mathbf{c}\|_2}{\sigma}\right), \quad (3.64)$$

where $Q(q) = \frac{1}{\sqrt{2\pi}} \int_q^\infty e^{-\frac{u^2}{2}} du$. As is evident from (3.64), for a given α , $P_D(\alpha)$ directly depends on $\|X\mathbf{c}\|_2$. On the other hand, because X is a compressive random Toeplitz matrix, Theorem 3.7 suggests that $\|X\mathbf{c}\|_2$ is concentrated around its expected value with high probability and with a tail probability bound that decays exponentially in M divided by $\rho(\mathbf{c})$. Consequently, one could conclude that for fixed M , the behavior of $\rho(\mathbf{c})$ affects the behavior of $P_D(\alpha)$ over α . The following example illustrates this dependency.

Example 3.65 (*Detector Performance*) Assume with a failure probability of $P_{FA} = \alpha = 0.05$, a detection probability of $P_D(\alpha) = 0.95$ is desired. Assume $\sigma = 0.3$. From (3.64) and noting that $Q(-1.6449) = 0.95$ and $Q^{-1}(0.05) = 1.6449$, one concludes that in order to achieve the desired detection, $\|X\mathbf{c}\|_2$ should exceed $0.3 \times 2 \times 1.6449 = 0.9869$ (i.e., $\|X\mathbf{c}\|_2^2 \geq 0.9741$). On the other hand, for a Toeplitz X with i.i.d. entries drawn from $\mathcal{N}(0, \frac{1}{M})$, $\mathbf{E}[\|X\mathbf{c}\|_2^2] = \|\mathbf{c}\|_2^2$. Assume without loss of generality, $\|\mathbf{c}\|_2 = 1$. Thus, from Theorem 3.7 and the bound in (3.9), we have for $\epsilon \in (0, 1)$

$$\mathbf{P}\left\{\|X\mathbf{c}\|_2^2 - 1 \leq -\epsilon\right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\mathbf{c})}} \leq e^{-\frac{\epsilon^2 M}{8\rho(\mathbf{c})}}. \quad (3.66)$$

Therefore, for a choice of $\epsilon = 1 - 0.9741 = 0.0259$ and from (3.66), one could conclude that

$$\mathbf{P}\left\{\|X\mathbf{c}\|_2^2 \leq 0.9741\right\} \leq e^{-\frac{6.7 \times 10^{-4} M}{8\rho(\mathbf{c})}}.$$

Consequently, for $\zeta \in (0, 1)$, if $M \geq \frac{16\rho(\mathbf{c})}{6.7 \times 10^{-4}} \log \zeta^{-1}$, then with probability at least $1 - \zeta^2$, $\|X\mathbf{c}\|_2^2$ exceeds 0.9741, achieving the desired detection performance. Apparently, M depends on $\rho(\mathbf{c})$ and qualitatively, one could conclude that for a fixed M , a signal \mathbf{c} with small $\rho(\mathbf{c})$

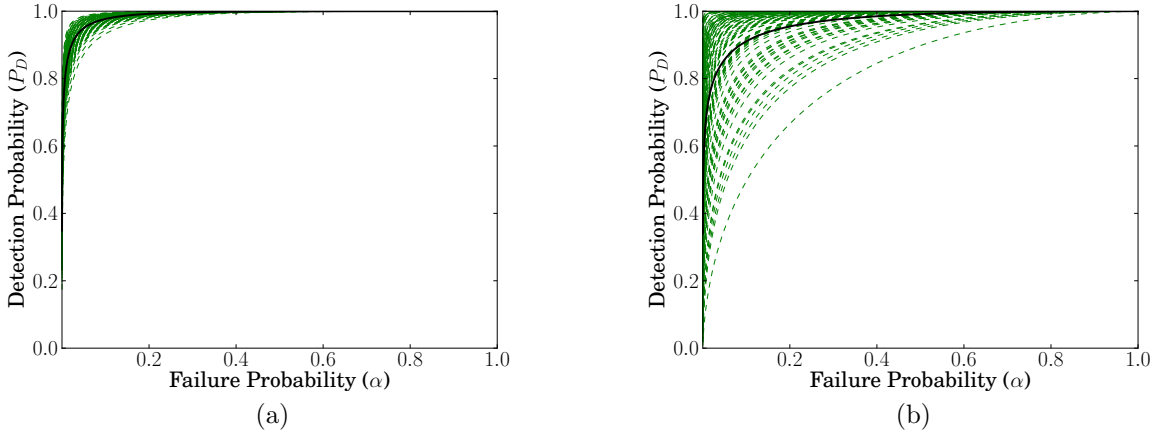


Figure 3.7: ROCs for 1000 random matrices X for a fixed signal \mathbf{c} with $\rho(\mathbf{c}) = 45.6$. (a) Unstructured X . (b) Toeplitz X . The solid black curve is the average of 1000 curves.

leads to better detection (*i.e.*, maximized $P_D(\alpha)$ over α). Similarly, a signal \mathbf{c} with large $\rho(\mathbf{c})$ is more difficult to reliably detect.

In the next section, we examine signals of different $\rho(\mathbf{c})$ values and show how their ROCs change. It is interesting to note that this dependence would not occur if the matrix X were unstructured (which, of course, would not apply to the convolution-based measurement scenario considered here but is a useful comparison) as the CoM behavior of unstructured Gaussian matrices is agnostic to the signal \mathbf{c} .

3.9.2 Empirical Results and ROCs

In several simulations, we examine the impact of $\rho(\mathbf{c})$ on the detector performance. To begin, we fix a signal $\mathbf{c} \in \mathbb{R}^{256}$ with 50 non-zero entries all taking the same value; this signal has $\|\mathbf{c}\|_2 = 1$ and $\rho(\mathbf{c}) = 45.6$ with our choice of $M = 128$. We generate 1000 random unstructured and Toeplitz matrices X with i.i.d. entries drawn from $\mathcal{N}(0, \frac{1}{M})$. For each matrix X , we compute a curve of P_D over P_{FA} using (3.64); we set $\sigma = 0.3$. Figure 3.7(a) and Figure 3.7(b) show the ROCs resulting from the unstructured and Toeplitz matrices, respectively. As can be seen, the ROCs associated with Toeplitz matrices are more scattered than the ROCs associated with unstructured matrices. This is in fact due to the weaker

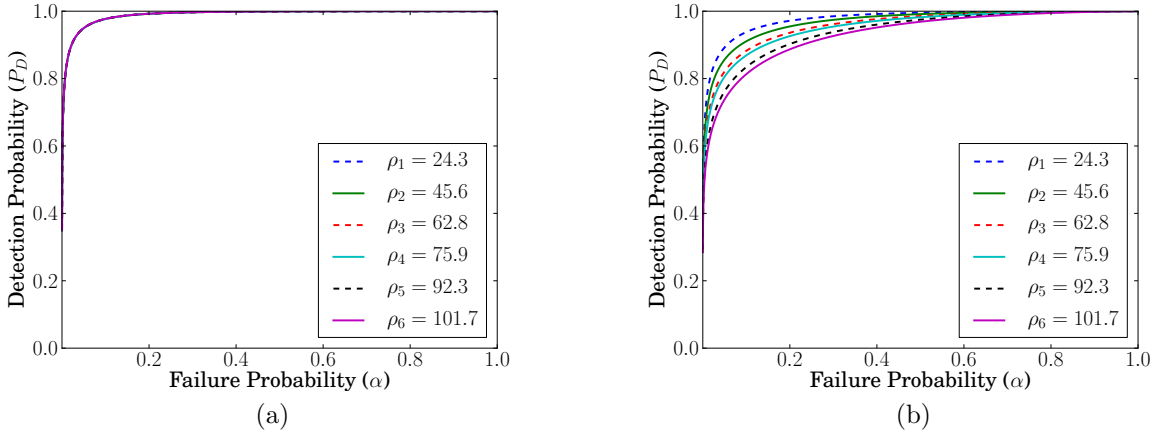


Figure 3.8: Average ROCs over 1000 random matrices X for 6 different signals \mathbf{c} . (a) Unstructured X . All curves are overlapping. (b) Toeplitz X . The curves descend in the same order they appear in legend box.

concentration of $\|X\mathbf{c}\|_2$ around its expected value for Toeplitz X (recall (3.8) and (3.9)) as compared to unstructured X (recall (3.10)).

To compare the ROCs among signals having different $\rho(\mathbf{c})$ values, we design a simulation with 6 different signals. Each signal again has $\|\mathbf{c}\|_2 = 1$, and we take $\sigma = 0.3$ as above. Figure 3.8(a) and Figure 3.8(b) plot the average ROC for each signal over 1000 random unstructured and Toeplitz matrices, respectively. Two things are evident from these plots. First, the plots associated with Toeplitz matrices show a signal dependency while the ones associated with unstructured matrices are signal-agnostic. Second, with regards to the plots associated with Toeplitz X , we see a decrease in the curves (i.e., inferior detector performance) for signals with larger values of $\rho(\mathbf{c})$.

In summary, our theory suggests and our simulations confirm that the value of $\rho(\mathbf{c})$ has a direct influence on the detector performance. From a systems perspective, as an example, this means that detecting changes in systems having a sparse impulse response in the time domain (e.g., communication channels with multipath propagation) will be easier than doing so for systems having a sparse impulse response in the frequency domain (e.g., certain resonant systems). It is worth mentioning that while detection analysis of systems

with sparse impulse response is interesting, our analysis can be applied to situations where neither the impulse responses \mathbf{a} and \mathbf{b} nor the change \mathbf{c} are sparse.

CHAPTER 4

COMPRESSIVE TOPOLOGY IDENTIFICATION

Structure identification of large-scale but sparse-flow interconnected dynamical systems from limited data has recently gained much attention in the control and signal processing communities. In this chapter⁵ we consider the topology identification of such systems.

In our model, the system topology under study has the structure of a directed graph of P nodes. Each edge of the directed graph represents a Finite Impulse Response (FIR) filter. Each node is a summer, whose inputs are the signals from the incoming edges, while the output of the summer is sent to outgoing edges. Both the graph topology and the impulse response of the FIR filters are unknown. The goal is to perform the topology identification using the *smallest possible* number of node observations when there is limited data available and for this reason, we call this problem Compressive Topology Identification (CTI). Inspired by Compressive Sensing (CS) we show that in cases where the network interconnections are suitably *sparse* (i.e., the network contains sufficiently few links), it is possible to perfectly identify the network topology along with the filter impulse responses from small numbers of node observations, although this leaves an apparently ill-conditioned identification problem.

If all filters in the graph share the same order, we show that CTI can be cast as the recovery of a *block-sparse* signal $\mathbf{x} \in \mathbb{R}^N$ from observations $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ with $M < N$, where matrix A is a block-concatenation of P Toeplitz matrices. We use block-sparse recovery algorithms from the CS literature such as Block Orthogonal Matching Pursuit (BOMP) [15–17] in order to perform CTI, discuss identification guarantees, introduce the notion of *network coherence* for the analysis of interconnected networks, and support the discussions with illustrative simulations. In a more general scenario, and when the filters in the graph can be of different orders (unknown), we show that the identification problem can be cast as the

⁵This work is in collaboration with Tyrone L. Vincent and Michael B. Wakin [8–10].

recovery of a *clustered-sparse* signal $\mathbf{x} \in \mathbb{R}^N$ from the measurements $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ with $M < N$, where the matrix A is a block-concatenation of P Toeplitz matrices. To this end, we introduce a greedy algorithm called Clustered Orthogonal Matching Pursuit (COMP) that tackles the problem of recovering clustered-sparse signals from few measurements. In a clustered-sparse model, in contrast to block-sparse models, there is no prior knowledge of the locations or the sizes of the clusters. We discuss the COMP algorithm and support the discussions with simulations.

4.1 Introduction

System identification is usually concerned with developing a model of a dynamical system from data for use in control design or for prediction. Large scale systems, which we define as systems with a large number of observable signals, present particular challenges for identification, particularly in the choice of model structure and the potentially large number of parameters that characterize this structure. One could attempt to meet this challenge by performing the identification with different model structures, and evaluating the prediction error using cross-validation or a prediction-error criterion that includes a penalty for model complexity, such as the AIC [1]. However, when the model structure is truly unknown, this could require an unacceptably large number of identification problems to be solved.

Many problems of current interest to the controls community involve large-scale interconnected dynamical systems. In this work, we focus on systems with a large number of observable variables, where the relations between these variables can be described by a signal flow graph with nodes of low maximum in-degree. Examples of such systems come from thermal modeling of buildings [62, 63], biological systems [64], and economics [65]. While there has been quite a bit of work to date on the analysis and control of networked systems (see e.g., [66–68]), such analysis typically requires knowledge of the network topology, which is not always available a priori. Thus, there is a need for effective “topological identification” procedures [69–71] which, given measurements of the nodes of an interconnected dynamical system over a finite time interval, can determine the correct interconnection topology.

The topology identification problem has been addressed by Materassi and Innocenti [71] in the case that the interconnection graph has a tree structure and enough data is available to form reliable estimates of cross-power spectral densities. In this chapter, we consider a more general setting, allowing arbitrary interconnections (including trees, loops, and self-loops) between nodes in the network, but we assume that the interconnection graph is sparse in the sense that each node has a relatively low in-degree. For the types of interconnected systems of interest here, we assume that each observable signal is linearly and causally dependent on a small subset of the remaining observable signals, plus an independent signal (i.e., input) that may or may not be measured. A more recent work by Bolstad et al. [72] also considered a causal network inference problem and derived conditions under which the recovery algorithm consistently estimates the sparse network structure. Defining the *false connection score*, they show that when the network size and the number of node observations grow to infinity (when the sample size can grow much slower compared to the network size), then the network topology can be inferred with high probability in the limit.

4.2 Interconnected Dynamical Systems

In order to provide an analysis, we assume that the interconnected dynamical system is a realization of a specific model structure. In what follows, all signals are discrete time, defined over a finite non-negative time interval, and represented equivalently as either the function of integer time $a(t)$ or grouped into vector form using the boldface \mathbf{a} . The model structure can be defined using a *directed graph*, such as shown in Figure 4.1(a). Each measured signal is represented by a node in the graph. Edges represent filters that operate on the node signal, and are assumed to be FIR filters of different orders. Any types of interconnections between nodes—such as trees, loops, and self-loops—are allowed.

Given an interconnected network of P nodes, let the time series $a_i(t)$, $t = 1, 2, \dots, M$, denote the output of node i . Each edge in the graph, labeled $\mathbf{x}_i^j \in \mathbb{R}^{n_i^j}$, filters the signal $a_j(t)$ at the tail of the edge, and passes the result, $y_i^j(t) = \sum_{s=1}^{n_i^j} x_i^j(s) a_j(t-s)$, to the node at the head of the edge. The signal $x_i^j(t)$ is the impulse response of the filter for the edge from

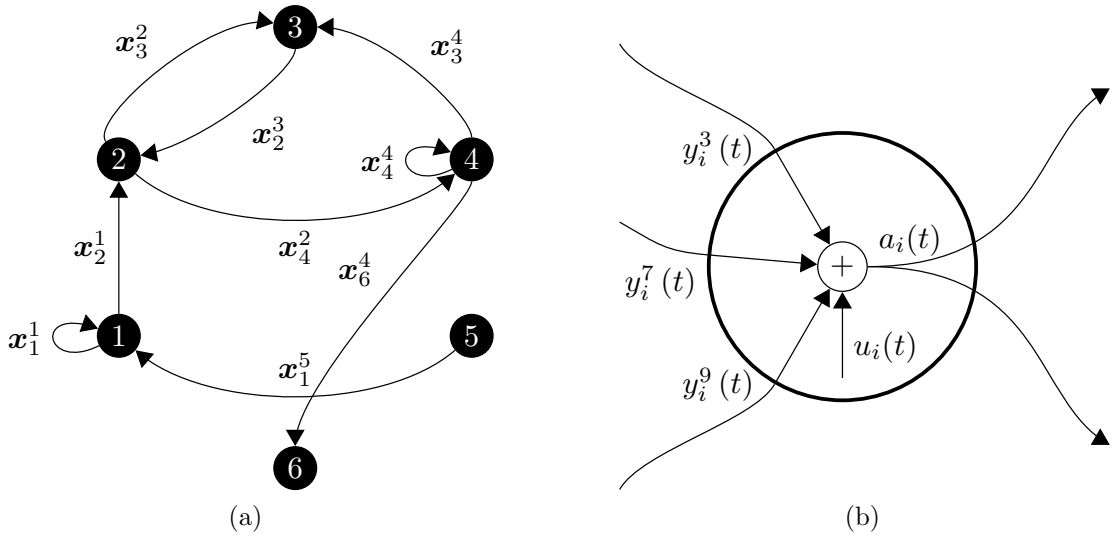


Figure 4.1: (a) Network model of 6 interconnected nodes. Each edge of the directed graph (\mathbf{x}_i^j) represents an FIR filter of order n_i^j (i.e., $\mathbf{x}_i^j \in \mathbb{R}^{n_i^j}$). (b) Single node model. Each node is a summer whose inputs are the signals from the incoming edges, while the output of the summer is sent to the outgoing edges. In this illustration, node i sums the signals from nodes 3, 7, and 9 (i.e., $\mathcal{N}_i = \{3, 7, 9\}$) plus a node-specific input term $u_i(t)$.

node j to node i . Note that we have assumed no feedthrough term. Let \mathcal{N}_i denote the set of nodes whose outputs are processed and fed to node i . As shown in Figure 4.1(b), we assume that each node i simply sums the signals that terminate upon it and adds a node-specific input term $u_i(t)$ plus measurement noise $e_i(t)$. In other words, the output of node i for $t = 1, 2, \dots, M$, is given by

$$\mathbf{a}_i(t) = \sum_{j \in \mathcal{N}_i} \mathbf{y}_i^j(t) + u_i(t) + e_i(t), \quad (4.1)$$

for $i = 1, 2, \dots, P$. Using a more convenient matrix-vector notation, the output of each node $\mathbf{a}_i \in \mathbb{R}^M$ for $i = 1, 2, \dots, P$, can be written as

$$\mathbf{a}_i = \sum_{j \in \mathcal{N}_i} A_j \mathbf{x}_i^j + \mathbf{u}_i + \mathbf{e}_i. \quad (4.2)$$

For uniformity, we take $m = \max_{i,j}(n_i^j)$ and represent $\mathbf{x}_i^j \in \mathbb{R}^m$ with trailing zeros as necessary. Observe that with this assumption, $\{\mathbf{x}_i^j\}_{i,j=1}^P$ may have different support patterns due to different possible unknown transport delays and different numbers of trailing zeros. In (4.2), A_j is an $M \times m$ Toeplitz matrix, $\mathbf{u}_i \in \mathbb{R}^M$, and $\mathbf{e}_i \in \mathbb{R}^M$. We use the notation

$A_j = \mathcal{T}(\mathbf{a}_j)_M^m$ where $\mathcal{T}(\mathbf{a})_M^m$ defines a mapping from a finite sequence \mathbf{a} to a Toeplitz matrix.

$$\mathcal{T}(\mathbf{a})_M^m := \begin{bmatrix} a(0) & 0 & \dots & 0 \\ a(1) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a(M-m) & & \ddots & a(0) \\ \vdots & \ddots & & a(1) \\ & & \ddots & \vdots \\ a(M-1) & \dots & a(M-m) & \end{bmatrix},$$

for $M \geq m$ where zeros are applied if the index goes outside the defined range of \mathbf{a} . A matrix with the same entries along all its diagonals is called Toeplitz. Setting $\mathbf{x}_i^j = \mathbf{0}$ for $\forall j \notin \mathcal{N}_i$, (4.2) can be rewritten as

$$\mathbf{a}_i = \sum_{j=1}^P A_j \mathbf{x}_i^j + \mathbf{u}_i + \mathbf{e}_i, \quad (4.3)$$

which can be expanded as

$$\mathbf{a}_i = \underbrace{\begin{bmatrix} A_1 & \dots & A_j & \dots & A_P \end{bmatrix}}_A \underbrace{\begin{bmatrix} \mathbf{x}_i^1 \\ \vdots \\ \mathbf{x}_i^j \\ \vdots \\ \mathbf{x}_i^P \end{bmatrix}}_{\mathbf{x}_i} + \mathbf{u}_i + \mathbf{e}_i, \quad (4.4)$$

or equivalently as

$$\mathbf{a}_i = A\mathbf{x}_i + \mathbf{u}_i + \mathbf{e}_i, \quad (4.5)$$

where $\mathbf{a}_i \in \mathbb{R}^M$, $\mathbf{x}_i \in \mathbb{R}^{Pm}$, and $A \in \mathbb{R}^{M \times Pm}$ is a matrix formed by the concatenation of P Toeplitz matrices.

4.3 Network Tomography

Given an interconnected graph of P nodes, the topology identification problem can be viewed as recovering the set of interconnected links (\mathcal{N}_i) for each node i in the graph. The links include unknown FIR filters of different orders, including unknown transport delays. The assumed a priori knowledge is the total number of nodes P in the network and the max-

imum possible degree m of each link. By the formulation given in the previous section, the topology identification problem is equivalent to recovering $\{\mathbf{x}_i\}_{i=1}^P$ given node observations. In particular, the measurements available to us consist of all node inputs $\{\mathbf{u}_i\}_{i=1}^P$ and all node outputs $\{\mathbf{a}_i\}_{i=1}^P$. One possible approach to this problem would be to solve

$$\min_{\{\mathbf{x}_i\}_{i=1}^P} \sum_{i=1}^P \|(\mathbf{a}_i - \mathbf{u}_i) - A\mathbf{x}_i\|_2^2. \quad (4.6)$$

The objective function in (4.6) can be minimized by solving

$$\min_{\mathbf{x}_i} \|(\mathbf{a}_i - \mathbf{u}_i) - A\mathbf{x}_i\|_2^2 \quad (4.7)$$

separately for each node i in the network. Observe that the same matrix A is used for recovery of all \mathbf{x}_i . For simplicity and without loss of generality, we will suppose henceforth that $\mathbf{a}_i - \mathbf{u}_i = \mathbf{b}$ and $\mathbf{x}_i = \mathbf{x}$ for each given node and focus on the task of solving

$$\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|_2^2, \quad (4.8)$$

where by letting $N = Pm$, we have $\mathbf{b} \in \mathbb{R}^M$ and $\mathbf{x} \in \mathbb{R}^N$, and $A \in \mathbb{R}^{M \times N}$ is a matrix consisting of a concatenation of P Toeplitz matrices.

4.4 Compressive Topology Identification

The optimization problem (4.8) has a unique solution if we collect $M \geq N$ measurements and if the matrix A is full rank. From standard linear algebra, we would know that exact recovery of \mathbf{x} when $\mathbf{e}_i = \mathbf{0}$ is possible from $\mathbf{x}^* = A^\dagger \mathbf{b}$, where $A^\dagger = (A^T A)^{-1} A^T$ is the Moore-Penrose pseudoinverse of A . However, since N depends linearly on P , the condition $M \geq N$ requires large M for topology identification for large-scale interconnected networks (i.e., large P). On the other hand, if the system topology was known a priori, then only the parameters for the links that actually exist would need to be identified. The number of such parameters is independent of the total number of nodes, P .

In order to move towards a data requirement closer to the case of a priori knowledge of the topology, one can attempt to apply additional information or assumptions that can add additional constraints and reduce the effective degrees of freedom. In the case of intercon-

nected dynamical systems, one assumption that is valid for a large number of applications of interest is low node in-degree (i.e., each node has only a small number of incoming edges). With this assumption, there will be a distinct structure to the solutions \mathbf{x} that we are searching for. In particular, a typical vector \mathbf{x} under our model assumptions will have very few non-zero entries, and these non-zero entries will be grouped in a few locations. The number of groups corresponds to the number of links that contribute to the output of the current node of interest (i.e., the cardinality of the set \mathcal{N}_i for node i), while the size of each group depends on the order and structure of the corresponding FIR filter connected to node i .

If all links in the network have the same order with no transport delay (i.e., $\forall i, j, n_i^j = m$), the non-zero entries of \mathbf{x} appear in locations of the same size. In the CS literature, such a structure is known as *block-sparsity*, defined as the following.

Definition 4.9 ([16]) *Let $\mathbf{x} \in \mathbb{R}^N$ be a concatenation of P vector-blocks $\mathbf{x}^j \in \mathbb{R}^m$ of the same length where $N = Pm$, i.e., $\mathbf{x} = [\mathbf{x}^{1T} \dots \mathbf{x}^{jT} \dots \mathbf{x}^{PT}]^T$. The vector $\mathbf{x} \in \mathbb{R}^N$ is called K -block sparse if it has $K < P$ non-zero blocks.*

On the other hand, if links are allowed to have different orders and different unknown transport delays, the vector \mathbf{x} will no longer have a block-sparse structure. Instead, \mathbf{x} has a *clustered-sparse* [73] structure, defined as the following.

Definition 4.10 ([73]) *A signal $\mathbf{x} \in \mathbb{R}^N$ is called (K, C) -clustered sparse if it contains a total of K non-zero coefficients, spread among at most C disjoint clusters of unknown size and location.*

Figure 4.2 shows a comparison between a block-sparse signal (Figure 4.2(a)) and a clustered-sparse signal (Figure 4.2(b)). In a block-sparse structure, the non-zero entries appear in blocks of the same size while in a clustered-sparse structure, they appear in clusters of unknown size and location.

In the following sections, we consider block-sparse and clustered-sparse structures in the context of CTI with more details.

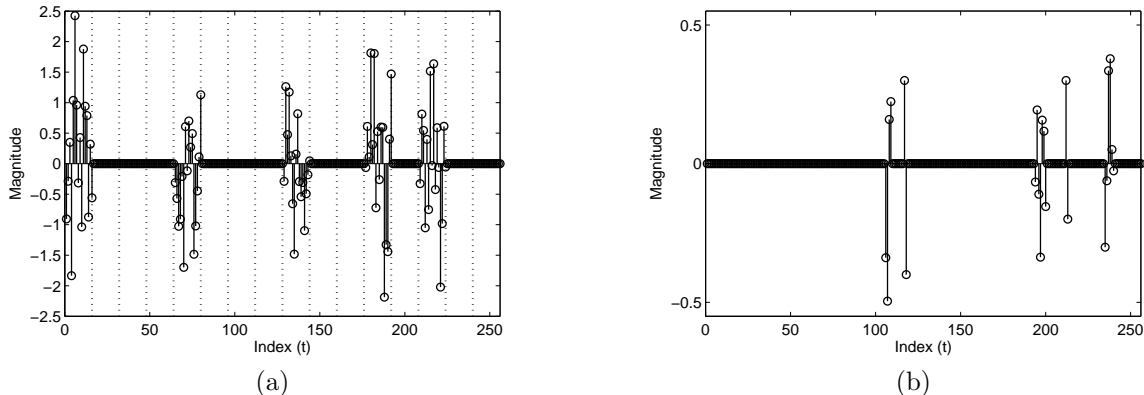


Figure 4.2: Clustered sparsity versus block sparsity model. (a) A 5-block sparse signal with block size $m = 16$. (b) A (21, 5)-clustered sparse signal. Clusters have different sizes. Both signals have same length $N = 256$.

4.5 CTI via Block-Sparse Recovery

In this section we assume that all the links in the network share the same order with no additional transport delay. As mentioned earlier, in this situation CTI boils down to recovery of a block-sparse signal from few measurements.

Several extensions of the standard CS recovery algorithms have been proposed to account for additional structure in the sparse signal to be recovered [16, 74]. Among these, the BOMP (Block Orthogonal Matching Pursuit (OMP)) algorithm [15–17] is designed to exploit block sparsity. We will consider BOMP for the topology identification problem due to its ease of implementation and its flexibility in recovering block-sparse signals of different sparsity levels. To find a block-sparse solution to the equation $\mathbf{b} = A\mathbf{x}$, the formal steps of the BOMP algorithm are listed in Algorithm 1. The basic intuition behind BOMP is as follows.

Due to the block sparsity of \mathbf{x} , the vector of observations \mathbf{b} can be written as a succinct linear combination of the columns of A , with the selections of columns occurring in clusters due to the block structure of the sparsity pattern in \mathbf{x} . BOMP attempts to identify the participating indices by correlating the measurements \mathbf{b} against the columns of A and comparing the correlation statistics among different blocks. Once a significant block has been

Algorithm 1 The BOMP algorithm for recovery of block-sparse signals

Require: matrix A , measurements \mathbf{b} , block size m , stopping criteria

Ensure: $\mathbf{r}^0 = \mathbf{b}$, $\mathbf{x}^0 = \mathbf{0}$, $\Lambda^0 = \emptyset$, $\ell = 0$

repeat

1. **match:** $\mathbf{h}_i = A_i^T \mathbf{r}^\ell$, $i = 1, 2, \dots, P$
2. **identify support:** $\lambda = \arg \max_i \|\mathbf{h}_i\|_2$
3. **update the support:** $\Lambda^{\ell+1} = \Lambda^\ell \cup \lambda$
4. **update signal estimate:** $\mathbf{x}^{\ell+1} = \arg \min_{\mathbf{s}: \text{supp}(\mathbf{s}) \subseteq \Lambda^{\ell+1}} \|\mathbf{b} - A\mathbf{s}\|_2$,
where $\text{supp}(\mathbf{s})$ indicates the blocks on which \mathbf{s} may be non-zero
5. **update residual estimate:** $\mathbf{r}^{\ell+1} = \mathbf{b} - A\mathbf{x}^{\ell+1}$
6. **increase index ℓ by 1**

until stopping criteria true

output: $\hat{\mathbf{x}} = \mathbf{x}^\ell$

identified, its influence is removed from the measurements \mathbf{b} via an orthogonal projection, and the correlation statistics are recomputed for the remaining blocks. This process repeats until convergence.

Eldar et al. [16] proposed a sufficient condition for BOMP to recover any sufficiently concise block-sparse signal \mathbf{x} from compressive measurements. This condition depends on the properties of A , as described in the next section.

4.5.1 Block-Coherence and Sub-Coherence

Let $A \in \mathbb{R}^{M \times N}$ be a concatenation of P matrix-blocks $A_i \in \mathbb{R}^{M \times m}$ as

$$A = [A_1 \ \cdots \ A_i \ \cdots \ A_P]. \quad (4.11)$$

We assume that there is a unique K -block sparse signal \mathbf{x} that satisfies $\mathbf{b} = A\mathbf{x}$. Assume for the moment that matrix A has columns of unit norm. The *block-coherence* [15–17] of A is defined as

$$\mu_{\text{block}}(A) := \max_{i,j \neq i} \frac{1}{m} \|(A_i^T A_j)\|_2, \quad (4.12)$$

where $\|A\|_2$ is the spectral norm of A . In the case where $m = 1$, this matches the conventional definition of coherence [39, 40],

$$\mu(A) := \max_{i,j \neq i} |\mathbf{a}_i^T \mathbf{a}_j|, \quad (4.13)$$

where $\{\mathbf{a}_i\}_{i=1}^P$ are the columns of matrix A . While μ_{block} characterizes the intra-block relationships within matrix A , the inter-block properties can be quantified by the *sub-coherence* [15–17] of A as

$$\mu_{\text{sub-block}}(A) := \max_k \max_{i,j \neq i} |\mathbf{a}_{k_i}^T \mathbf{a}_{k_j}|, \quad (4.14)$$

where $\mathbf{a}_{k_i}, \mathbf{a}_{k_j}$ are columns of the matrix-block A_k .

4.5.2 Recovery Condition

In [16, Theorem 3], a sufficient condition is provided that guarantees recovery of any K -block sparse signal \mathbf{x} from the measurements $\mathbf{b} = A\mathbf{x}$ via BOMP. This condition is stated in terms of the block-coherence metrics, μ_{block} and $\mu_{\text{sub-block}}$ of the matrix A .

Theorem 4.15 [16] *If \mathbf{x} is block-sparse with K non-zero blocks of length n , then BOMP will recover \mathbf{x} from the measurements $\mathbf{b} = A\mathbf{x}$ if*

$$Km < \mu_T, \quad (4.16)$$

where

$$\mu_T = \frac{1}{2} \left(\mu_{\text{block}}^{-1} + m - (m-1) \frac{\mu_{\text{sub-block}}}{\mu_{\text{block}}} \right). \quad (4.17)$$

When $m = 1$, condition (4.16) is equivalent to the exact recovery condition using OMP [40], namely, $K < \frac{1}{2}(\mu^{-1} + 1)$. What Theorem 4.15 tells us is that, for a given matrix A with certain block-coherence metrics (μ_{block} and $\mu_{\text{sub-block}}$), BOMP is guaranteed exact recovery of block-sparse signals of a limited sparsity level. The smaller the value of μ_T , the higher the permitted value of K , and the broader the class of signals that can be recovered via BOMP.

4.5.3 Network Coherence

In previous sections, we explained how we can cast the topology identification of a large-scale interconnected network as a CS recovery problem where the signal to be identified has

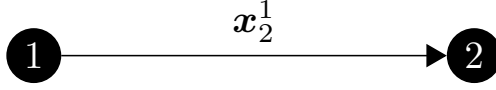


Figure 4.3: A simple network for our study of network coherence.

a block-sparse structure and the measurement matrix is a block-concatenation of Toeplitz matrices. Since the block-coherence metrics (4.12) and (4.14) give a sufficient condition for recovery via BOMP, it is of interest to examine the particular effect of the network interconnection structure on the block-coherence metrics of A . To highlight the important role that these metrics play in the context of our topology identification problem, where the coupling between node outputs is based on a discrete-time convolution process, we will collectively refer to μ_{block} and $\mu_{\text{sub-block}}$ as the *network coherence* metrics. In order to give some insight into how the network coherence relates to the network topology, we focus in this section on networks with very simple interconnection structures.

To begin, let us consider the simple network shown in Figure 4.3 and assume that the input $d_i(t) =: u_i(t) + e_i(t)$ is a zero mean Gaussian sequence with unit variance. We would like to estimate the block-coherence and sub-coherence of the matrix A associated with this network. For this network configuration, we can write the output of each node as

$$\mathbf{a}_1 = \mathbf{d}_1 \quad \text{and} \quad \mathbf{a}_2 = A_1 \mathbf{x}_2^1 + \mathbf{d}_2. \quad (4.18)$$

It is easy to see that \mathbf{a}_2 can be rewritten as

$$\mathbf{a}_2 = G_2^1 \mathbf{d}_1 + \mathbf{d}_2, \quad (4.19)$$

where $G_2^1 =: \mathcal{T}(\mathbf{x}_2^1)_M^M$. Using the down-shift operator $S_M \in \mathbb{R}^{M \times M}$ defined as

$$S_M = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix},$$

$G_2^1 \in \mathbb{R}^{M \times M}$ can be rewritten as

$$G_2^1 = x_2^1(1) S_M + x_2^1(2) S_M^2 + \cdots + x_2^1(m) S_M^m. \quad (4.20)$$

Similarly, the matrix $A \in \mathbb{R}^{M \times 2m}$ can be written as

$$A = [A_1 \mid A_2] = [S_M \mathbf{a}_1 \ S_M^2 \mathbf{a}_1 \ \cdots \ S_M^m \mathbf{a}_1 \mid S_M \mathbf{a}_2 \ S_M^2 \mathbf{a}_2 \ \cdots \ S_M^m \mathbf{a}_2]. \quad (4.21)$$

Note that $\mathbf{E}[\|\mathbf{a}_1\|_2^2] = M$ and $\mathbf{E}[\|\mathbf{a}_2\|_2^2] = M(1 + \|\mathbf{x}_2^1\|_2^2)$. Using concentration of measure inequalities, it can be shown that as $M \rightarrow \infty$, $\|\mathbf{a}_1\|_2^2$ and $\|\mathbf{a}_2\|_2^2$ are highly concentrated around their expected values [3, 4]. Normalizing by these expected column norms, let \widehat{A}_1 and \widehat{A}_2 be A_1 and A_2 with approximately normalized columns, and let $\widehat{A} = [\widehat{A}_1 \mid \widehat{A}_2]$. Therefore, a reasonable estimate of the block-coherence $\mu_{\text{block}}(A)$ is simply given by the spectral norm of $\widehat{A}_1^T \widehat{A}_2$. We define such an estimate:

$$\widetilde{\mu}_{\text{block}}(A) := \mu_{\text{block}}(\widehat{A}) = \frac{1}{m} \|\widehat{A}_1^T \widehat{A}_2\|_2. \quad (4.22)$$

In order to derive a lower bound on $\mathbf{E}[\widetilde{\mu}_{\text{block}}(A)]$, we use the result of Lemma 4.23 which states lower and upper bounds on $\|\mathbf{E}[\widehat{A}_1^T \widehat{A}_2]\|_2$.

Lemma 4.23 *Assume $M > m$. Considering the configuration of the network shown in Figure 4.3, we have*

$$\frac{\|\mathbf{x}_2^1\|_2}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}} \leq \|\mathbf{E}[\widehat{A}_1^T \widehat{A}_2]\|_2 \leq \frac{\|\mathbf{x}_2^1\|_1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}}. \quad (4.24)$$

Proof See Appendix A.3. ■

We are particularly interested in deriving lower bounds on the expected value of the network coherence metrics. However, upper bounds would also be of interest. Using Lemma 4.23, we can state the following theorem.

Theorem 4.25 *For the network Figure 4.3, $\mathbf{E}[\widetilde{\mu}_{\text{block}}(A)]$ is bounded from below as*

$$\mathbf{E}[\widetilde{\mu}_{\text{block}}(A)] \geq \frac{\|\mathbf{x}_2^1\|_2}{m\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}}. \quad (4.26)$$

Proof From Jensen's inequality applied for convex functions and (4.22), we have the following lower bound for $\mathbf{E} [\tilde{\mu}_{\text{block}}(A)]$ as

$$\mathbf{E} [\tilde{\mu}_{\text{block}}(A)] = \frac{1}{m} \mathbf{E} \left[\|\widehat{A}_1^T \widehat{A}_2\|_2 \right] \geq \frac{1}{m} \|\mathbf{E} [\widehat{A}_1^T \widehat{A}_2]\|_2, \quad (4.27)$$

where we use the fact that the spectral norm of a matrix $\|\cdot\|_2$ is a convex function. Combining (4.27) and (4.24), we have

$$\mathbf{E} [\tilde{\mu}_{\text{block}}(A)] \geq \frac{\|\mathbf{x}_2^1\|_2}{m\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}}. \quad (4.28)$$

■

A similar approach can be carried out for the analysis of other types of network elements. For example, we can show that the network coherence of Figure 4.4(a) is bounded by

$$\mathbf{E} [\tilde{\mu}_{\text{block}}(A)] \geq \frac{1}{m} \max \left\{ \frac{\|\mathbf{x}_1^2\|_2}{\sqrt{1 + \|\mathbf{x}_1^2\|_2^2}}, \frac{\|\mathbf{x}_1^3\|_2}{\sqrt{1 + \|\mathbf{x}_1^3\|_2^2}} \right\}. \quad (4.29)$$

We can follow the same steps and derive a bound for the sub-coherence of the simple network of Figure 4.3. We simply state the result here. Letting $\tilde{\mu}_{\text{sub-block}}(A) := \mu_{\text{sub-block}}(\widehat{A})$, we have

$$\mathbf{E} [\tilde{\mu}_{\text{sub-block}}(A)] \geq \|\mathfrak{R}_{\mathbf{x}_2^1}(1) \cdots \mathfrak{R}_{\mathbf{x}_2^1}(m-1)\|_\infty^T, \quad (4.30)$$

where

$$\mathfrak{R}_{\mathbf{x}_2^1}(\tau) := \sum_{i=1}^{m-\tau} x_2^1(i)x_2^1(i+\tau)$$

denotes the un-normalized sample autocorrelation function of $\mathbf{x}_2^1 \in \mathbb{R}^m$. While finding network coherence bounds for more complicated interconnected networks (e.g., networks with loops and nodes of high out-degree) is a harder task, we observe the following important characteristics:

1. In the limit, the network coherence metrics are independent of the number of measurements.
2. The network coherence metrics are bounded below by a non-zero value that depends on the link impulse responses.

The latter phenomenon may suggest an ultimate limitation of the coherence metrics in the analysis of interconnected dynamical networks. Nevertheless, our simulations in the network topology problem do indicate that as the number of measurements M increases, recovery remains possible for a range of interesting problem sizes. The asymptotic behavior of the network coherence metrics is contrary to the conventional behavior in CS, in which increasing the number of rows of a dense matrix (number of measurements M) populated with independent and identically distributed (i.i.d.) Gaussian random variables will make the coherence approach a zero value, guaranteeing the recovery of signals with more and more non-zero coefficients.

4.5.4 Simulations and Discussion on the Network Coherence

In this section, we examine CTI for recovering the topology of a dynamical interconnected network based on compressive observations with random but known inputs, and we observe how the probability of successful recovery changes for different nodes in the network based on the local sparsity. In all of these simulations, we consider a network of 32 nodes with second order ($m = 2$) interconnecting links.

To begin, and in order to highlight the influence of network coherence in the recovery success rate of BOMP, consider the networks illustrated in Figure 4.4.

For simulation purposes, we consider two networks with the same topology structure as the networks shown in Figure 4.4 but within a 32-node network. We consider two possible scenarios: in one case shown in Figure 4.4(a), node 1 has no subsequent connections to other nodes in the network (we call this the “disconnected” case), while in the other, “connected” case, node 1 has subsequent connections to all other 29 nodes in the network. In both scenarios, the in-degree of node 1 is 2, while its out-degree is 0 in the disconnected case and 29 in the connected case.

In either scenario, we are interested in recovering the incoming links that contribute to node 1. As a function of the number of measurements M , Figure 4.5(a) plots the coherence measures for the two cases where curves are averaged over 1000 realizations of the network.

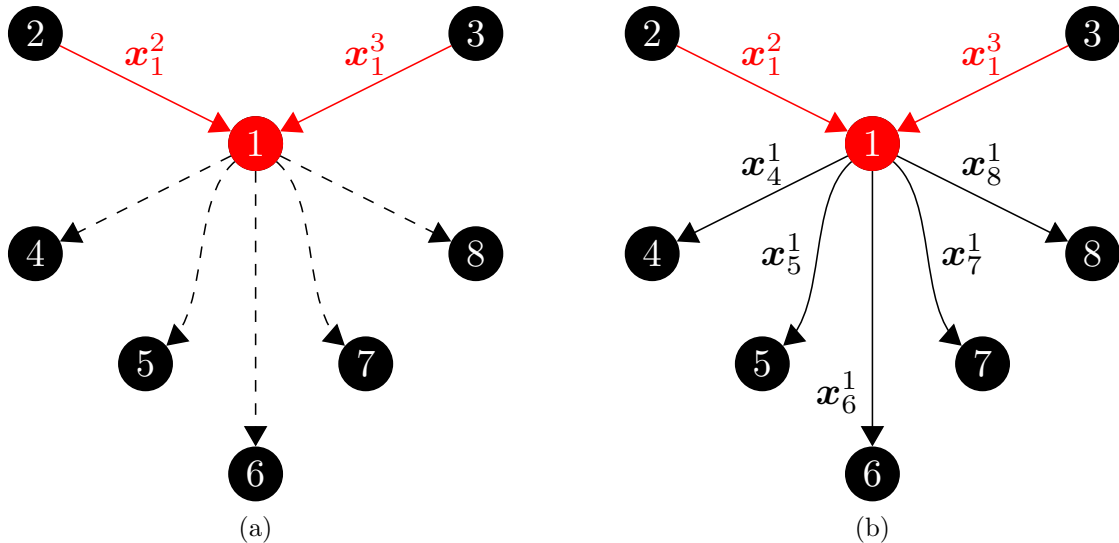


Figure 4.4: (a) Disconnected network. Node 1 has in-degree 2 but is disconnected from the rest of the network (i.e., out-degree zero). (b) Connected network. Node 1 has in-degree 2 and is connected to the rest of the network (i.e., in this example out-degree 5).

As would be expected from our analysis on the network coherence in Section 4.5.3, the coherence metrics are bounded from below by a non-zero value that depends on the link impulse responses, namely expressed in (4.29) for the disconnected network. The connected network, however, has higher typical block- and sub-coherence measures. Although coherence is only a sufficient condition for recovery, simulation results do show weaker recovery performance for the connected network, as shown in Figure 4.6. For each value of M , 1000 realizations of the network are carried out and the recovery rate is calculated.

For the sake of comparison, we compute the same coherence metrics for matrices populated with random Gaussian entries in either an unstructured format or in a Toeplitz block format. The results in Figure 4.5(b) show that for $A \in \mathbb{R}^{M \times N}$, the coherence measures approach zero as M increases. In contrast, as we have seen from Figure 4.5(a), in an interconnected network of dynamical systems, the coherence measures have an asymptotic behavior. This prevents the predicted recovery performance from growing as M increases (see the plot of μ_T in Figure 4.5(a)).

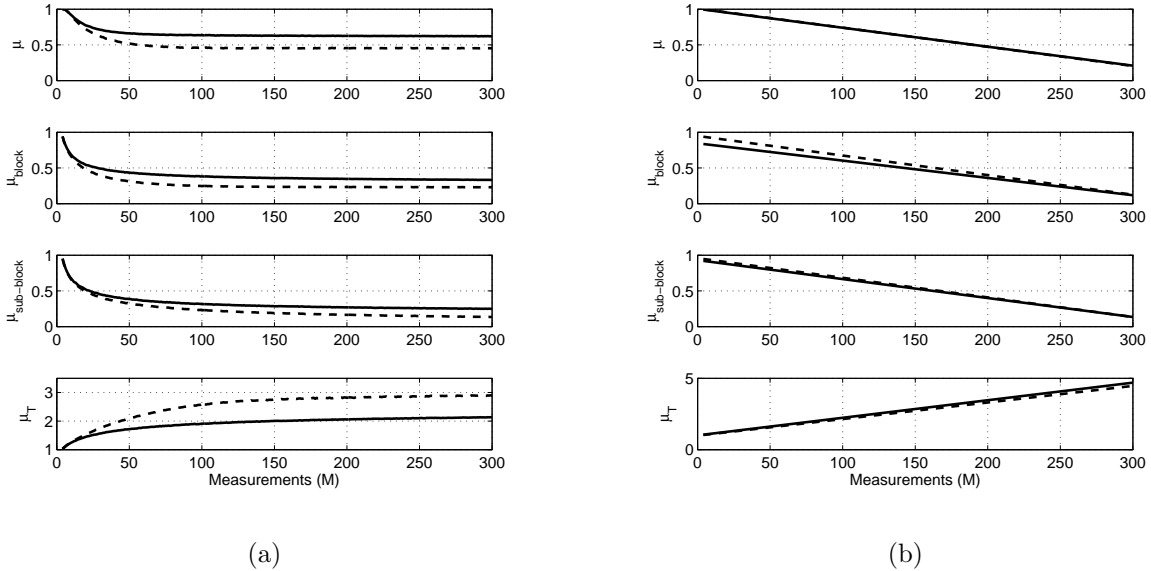


Figure 4.5: (a) Coherence metrics for the disconnected (dashed lines) and connected (solid lines) networks. The curves are averaged over 1000 realizations of the networks. Note that the coherence metrics approach a non-zero asymptote as the number of measurements M increases. (b) Coherence metrics for matrices with i.i.d. Gaussian entries (solid lines) and matrices which are block-concatenations of Toeplitz matrices with i.i.d. Gaussian entries. The curves are averaged over 1000 realizations of these types of matrices. Note that the coherence metrics approach zero as the number of measurements M increases.

4.6 CTI via Clustered-Sparse Recovery

In a more general setting, in this section we assume that the links in the interconnected network can have impulse responses of different order and a transport delay. Neither the filter orders nor the delays are known to the identification algorithm. Based on these assumptions, we show that the CTI problem boils down to finding a clustered-sparse solution to an apparently ill-conditioned set of linear equations. In the next section, we introduce a greedy algorithm called Clustered Orthogonal Matching Pursuit (COMP) which is designed for recovery of clustered signals from few measurements.

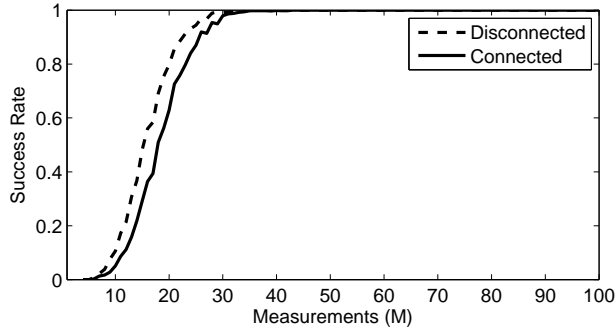


Figure 4.6: Recovery rate comparison of node 1 (\mathbf{x}_1^2 and \mathbf{x}_1^3) for connected and disconnected networks. For each measurement, 1000 realizations of the network are carried out and the recovery rate is calculated.

4.6.1 Clustered Orthogonal Matching Pursuit (COMP)

In a block-sparse structure as mentioned in Definition 4.9, the non-zero coefficients appear in blocks of the same length m . The BOMP algorithm is designed for recovering such block-sparse signals. As mentioned in Algorithm 1, the block size m is assumed to be known as one of the inputs to the algorithm. However, in this section we are interested in recovering signals whose non-zero entries appear in clusters of different sizes. The only assumption is on the maximum cluster length. In CS, such signals are called *clustered-sparse*.

In this section, we provide an algorithm that can be used for recovering clustered-sparse signals. The proposed method is an iterative greedy algorithm that is based on the well-known OMP algorithm. Its idea is intuitive and simple and also easy to implement.

The idea behind COMP is to exploit the knowledge that the non-zero entries of the signal appear in clusters, although of an arbitrary size and location. We modify the iterations of OMP in a way that exploits the clustered-sparsity pattern of the signal. The steps of the COMP algorithm are listed in Algorithm 2. The first two steps of COMP are the same as the first two steps of OMP. The outcome of step 2 at each iteration is a candidate for the true support. Let λ^ℓ denote the support candidate at iteration ℓ of the algorithm. If λ^0 is a valid candidate, i.e., $\lambda^0 \in T$ where T is the true support, then we can use our extra knowledge about the clustered-sparsity of the signal. In fact, we can use λ^0 as an

Algorithm 2 The COMP algorithm for recovery of clustered-sparse signals

Require: matrix A , measurements \mathbf{b} , maximum cluster size m , stopping criteria

Ensure: $\mathbf{r}^0 = \mathbf{b}$, $\mathbf{x}^0 = \mathbf{0}$, $\Lambda^0 = \emptyset$, $\ell = 0$, $w = m$

repeat

1. **match:** $\mathbf{h}^\ell = A^T \mathbf{r}^\ell$
2. **identify support indicator:** $\lambda^\ell = \arg \max_j |h^\ell(j)|$
3. **extend support:** $\widehat{\Lambda}^\ell = \{\lambda^\ell - w + 1, \dots, \lambda^\ell, \dots, \lambda^\ell + w - 1\}$
4. **update the support:** $\Lambda^{\ell+1} = \Lambda^\ell \cup \widehat{\Lambda}^\ell$
5. **update signal estimate:** $\mathbf{x}^{\ell+1} = \arg \min_{\mathbf{s}: \text{supp}(\mathbf{s}) \subseteq \Lambda^{\ell+1}} \|\mathbf{b} - A\mathbf{s}\|_2$,
where $\text{supp}(\mathbf{s})$ indicates the indices on which \mathbf{s} may be non-zero
6. **update residual estimate:** $\mathbf{r}^{\ell+1} = \mathbf{b} - A\mathbf{x}^{\ell+1}$
7. **increase index ℓ by 1**

until stopping criteria true

output: $\widehat{\mathbf{x}} = \mathbf{x}^\ell = \arg \min_{\mathbf{s}: \text{supp}(\mathbf{s}) \subseteq \Lambda^\ell} \|\mathbf{b} - A\mathbf{s}\|_2$

indicator for the location of one of the clusters in the signal. Therefore, if we consider a window with proper length centered around λ^0 , the extended support candidate is the window $\Lambda^1 = \widehat{\Lambda}^0 = \{\lambda^0 - w + 1, \dots, \lambda^0, \dots, \lambda^0 + w - 1\}$ with window size $2w - 1$. Because the algorithm does not know where exactly λ^0 is located in the true cluster, the window length $2w - 1$ should be large enough such that the true cluster which by assumption is at most of size m , will be contained in the extended support candidate Λ^1 . Apparently, the most conservative value for w is m . In the next step, the algorithm updates the signal estimate on the extended support candidate Λ^1 . Having this estimate, the algorithm continues by updating the residual estimate. In the next iteration of COMP, the algorithm finds the column that is most correlated with the current residual (steps 1 and 2). The new support candidate λ^1 will not be one of the already chosen indices due to orthogonal projection properties, i.e., $\lambda^1 \notin \Lambda^1$. Again the algorithm considers a window of length $2w - 1$ centered around λ^1 and combines it with the previous support, i.e., $\Lambda^2 = \Lambda^1 \cup \{\lambda^1 - w + 1, \dots, \lambda^1, \dots, \lambda^1 + w - 1\}$. COMP continues until stopping criteria are met.

Note that Λ^f (the final support candidate found by COMP) should contain the true support, i.e., $T \subset \Lambda^f$, while the reverse $\Lambda^f \subset T$ is not a necessity. In addition, the cardinality of Λ^f should be smaller than the number of measurements M in order to have a unique least-

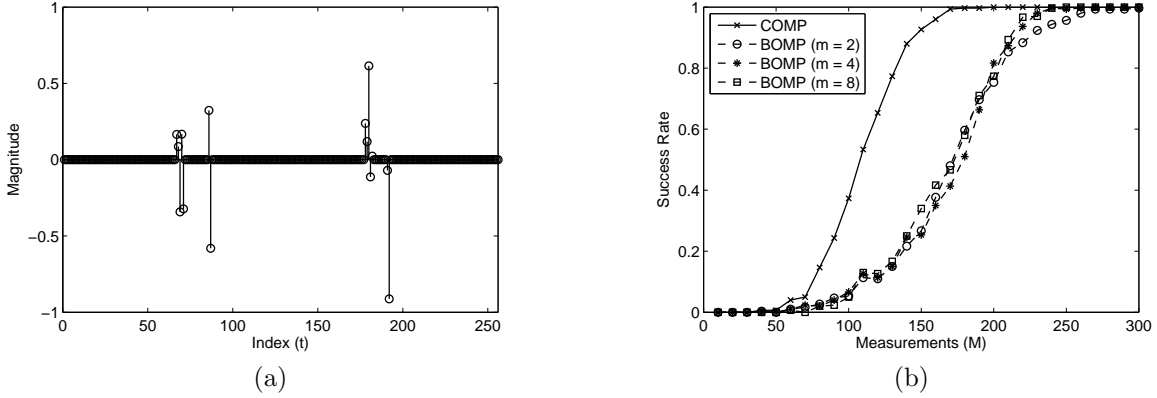


Figure 4.8: Recovery performance corresponding to node 10. (a) Signal \mathbf{x} corresponding to node 10 in the network graph. The cluster-sparsity level corresponds to the in-degree of node 10. (b) Recovery performance comparison between COMP and BOMP with different block sizes m . An initial value of $w = m = 8$ is chosen in COMP. The algorithm iterates by reducing w until stopping criteria are met. For comparison, BOMP is tested with three different block sizes ($m = \{2, 4, 8\}$). The success rate is calculated over 300 realizations of the network for a given number of measurements.

4.6.2 Numerical Simulations

We evaluate the performance of the COMP algorithm in CTI of an interconnected network. As explained earlier, we cast the topology identification problem as recovery of a clustered-sparse signal whose few non-zero coefficients appear in clustered locations. The clusters are of arbitrary sizes. The only knowledge is on the maximum cluster size m . In order to compare the performance of the COMP algorithm, we also consider recovery using the BOMP algorithm. Moreover, in order to make a fair comparison between the two algorithms, we consider recovery using the BOMP algorithm with several block sizes n .

Figure 4.7 shows a network of $P = 32$ nodes. Each edge of the directed network graph represents an FIR filter with possible transport delay. Each node is a summer, whose inputs are the signals from the incoming edges, while the output of the summer is sent to outgoing edges. Edges of the graph can be of different unknown orders and delays. Both the graph topology and the FIR filters and delays that make up the edges are unknown. The only knowledge is on the maximum cluster size $m = 8$. Therefore, for each node, the signal \mathbf{z} has

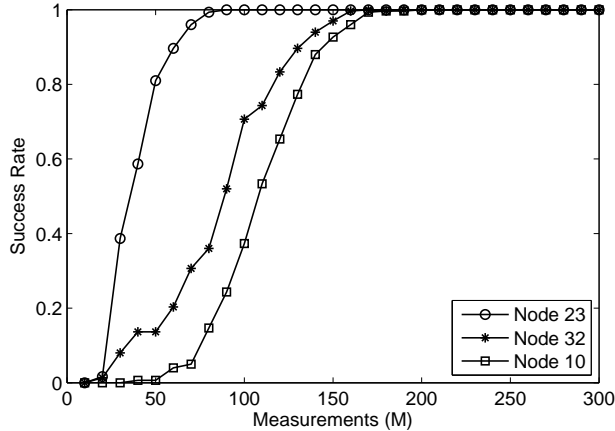


Figure 4.9: Recovery rate comparison of nodes 10, 23, and 32 in the network. An initial value of $w = m = 8$ is chosen in COMP. Nodes 23 and 32 have in-degree 2 and node 10 has in-degree 4. The success rate is calculated over 300 realizations of the network for a given number of measurements.

length $N = Pm = 256$.

Figure 4.8 shows the recovery performance corresponding to node 10 of the network graph of Figure 4.7. The corresponding signal \mathbf{x} to be recovered is shown in Figure 4.8(a). As can be seen the signal has a clustered-sparse structure with 4 clusters of different sizes. The number of clusters corresponds to the in-degree of node 10 while the size of each cluster depends on the order of the FIR filter of incoming edges. Figure 4.8(b) shows the recovery performance comparison between COMP and BOMP with different block sizes m . An initial value of $w = m = 8$ is chosen in COMP. The algorithm iterates by reducing w until stopping criteria are met. For comparison, BOMP is tested with three different block sizes ($m = \{2, 4, 8\}$). The success rate is calculated over 300 realizations of the network for a given number of measurements. As can be seen, the COMP algorithm outperforms the BOMP algorithm. For this signal, the recovery performance of BOMP does not significantly improve by changing the block size m .

Figure 4.9 shows the recovery rate comparison of nodes 10, 23, and 32 in the network of Figure 4.7. The success rate is calculated over 300 realizations of the network for a given number of measurements. Node 10 has in-degree 4 and nodes 23 and 32 have in-degree

2. We observe how the probability of successful recovery changes for different nodes in the network based on the local sparsity and the type of interconnection. For example, node 10 which has in-degree 4 requires more measurements compared to nodes 23 and 32 which have in-degree 2. In addition to the local sparsity of each node, we observe that nodes of same in-degree have different recovery performance. For example, nodes 23 and 32 both have in-degree 2. However, node 32 is much easier to recover with the COMP algorithm, i.e., it requires a smaller number of measurements for perfect recovery as compared to node 23. This difference may be related to the type of incoming interconnections to each node. The incoming edges to node 32 have a tree structure while the incoming edges to node 23 include a loop.

CHAPTER 5

CSI OF LTI AND LTV ARX MODELS

In this chapter⁶ we consider Compressive System Identification (CSI) (identification from few measurements) of Auto Regressive with eXternal input (ARX) models for Linear Time-Invariant (LTI) and Linear Time-Variant (LTV) systems.

In the case of LTI ARX systems, a system with a large number of inputs and unknown input delays on each channel can require a model structure with a large number of parameters, unless input delay estimation is performed. Since the complexity of input delay estimation increases exponentially in the number of inputs, this can be difficult for high dimensional systems. We show that in cases where the LTI system has possibly many inputs with different unknown delays, simultaneous ARX identification and input delay estimation is possible from few observations, even though this leaves an apparently ill-conditioned identification problem. We discuss identification guarantees and provide illustrative simulations.

We also consider identifying LTV ARX models. In particular, we consider systems with parameters that change only at a few time instants in a piecewise-constant manner where neither the change moments nor the number of changes is known a priori. The main technical novelty of our approach is in casting the identification problem as recovery of a block-sparse signal from an underdetermined set of linear equations. We suggest a random sampling approach for LTV identification, address the issue of identifiability, and support our proposed methods with simulations.

5.1 Introduction

As mentioned earlier in Chapter 1, under and over parameterization may have a considerable impact on the identification result, and choosing an optimal model structure is one of

⁶This work is in collaboration with Tyrone L. Vincent, Michael B. Wakin, Roland Tóth, and Kameshwar Poolla [7, 52, 75].

the primary challenges in system identification. Specifically, it can be a more problematic issue when the actual system to be identified (a) is sparse and/or (b) is multivariable (Multi-Input Single-Output (MISO) or Multi-Input Multi-Output (MIMO)) with I/O channels of different orders and unknown (possibly large) input delays. Finding an optimal choice of the model structure for such systems is less likely to happen from cross-validation approaches. In this chapter we consider CSI of ARX models for both LTI and LTV systems. We examine parameter estimation in the context of Compressive Sensing (CS) and formulate the identification problem as recovery of a *block-sparse* signal from an underdetermined set of linear equations. We discuss required measurements in terms of recovery conditions, derive bounds for such guarantees, and support our approach with simulations.

Related works include regularization techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm [76] and the Non-Negative Garrote (NNG) method [77]. These methods were first introduced for linear regression models in statistics. There also exist some results on the application of these methods to LTI ARX identification [78]. However, most of these results concern the stochastic properties of the parameter estimates in an asymptotic sense, with few results considering the limited data case. There is also some recent work on regularization of ARX parameters for LTV systems [79, 80].

5.2 Auto Regressive with eXternal input (ARX) Models

In this section, we introduce the ARX models and establish our notation. For simplicity we consider Single-Input Single-Output (SISO) systems; the formulations can be easily extended for multivariable systems.

An LTI SISO ARX model [1] with parameters $\{n, m, d\}$ is given by the difference equation

$$y(t) + a_1y(t-1) + \cdots + a_ny(t-n) = b_1u(t-d-1) + \cdots + b_mu(t-d-m) + e(t), \quad (5.1)$$

where $y(t) \in \mathbb{R}$ is the output at time instant t , $u(t) \in \mathbb{R}$ is the input, d is the input delay, and $e(t)$ is a zero mean stochastic noise process. Assuming $d + m \leq p$, where p is the input

maximum length (including delays), (5.1) can be written compactly as

$$y(t) = \boldsymbol{\phi}^T(t)\boldsymbol{\theta} + e(t) \quad (5.2)$$

where

$$\boldsymbol{\phi}(t) = \begin{bmatrix} -y(t-1) \\ \vdots \\ -y(t-n) \\ u(t-1) \\ \vdots \\ u(t-d) \\ u(t-d-1) \\ \vdots \\ u(t-d-m) \\ u(t-d-m-1) \\ \vdots \\ u(t-p) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\theta} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ 0 \\ \vdots \\ 0 \\ b_1 \\ \vdots \\ b_m \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$\boldsymbol{\phi}(t) \in \mathbb{R}^{n+p}$ is the data vector containing input-output measurements, and $\boldsymbol{\theta} \in \mathbb{R}^{n+p}$ is the parameter vector. The goal of the system identification problem is to estimate the parameter vector $\boldsymbol{\theta}$ from M observations of the system. Taking M consecutive measurements and putting them in a regression form, we have

$$\underbrace{\begin{bmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+M-1) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \boldsymbol{\phi}^T(t) \\ \boldsymbol{\phi}^T(t+1) \\ \vdots \\ \boldsymbol{\phi}^T(t+M-1) \end{bmatrix}}_{\boldsymbol{\Phi}} \boldsymbol{\theta} + \underbrace{\begin{bmatrix} e(t) \\ e(t+1) \\ \vdots \\ e(t+M-1) \end{bmatrix}}_{\mathbf{e}}$$

or equivalently

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \mathbf{e}. \quad (5.3)$$

In a noiseless scenario (i.e., $\mathbf{e} = \mathbf{0}$), from standard arguments in linear algebra, $\boldsymbol{\theta}$ can be exactly recovered from $M > n + p$ observations under the assumption of a persistently exciting input. Note that $\boldsymbol{\Phi}$ in (5.3) is a concatenation of 2 blocks

$$\boldsymbol{\Phi} = [\boldsymbol{\Phi}_y \mid \boldsymbol{\Phi}_u], \quad (5.4)$$

where $\Phi_y \in \mathbb{R}^{M \times n}$ and $\Phi_u \in \mathbb{R}^{M \times p}$ are Toeplitz matrices. Equation (5.4) can be extended for MISO systems with ℓ inputs as

$$\Phi = [\Phi_y \mid \Phi_{u_1} \mid \Phi_{u_2} \mid \cdots \mid \Phi_{u_\ell}], \quad (5.5)$$

where the Φ_{u_i} 's are Toeplitz matrices, each containing regression over one of the inputs. The ARX model in (5.1) can be also represented as

$$\mathcal{A}(q^{-1})y(t) = q^{-d}\mathcal{B}(q^{-1})u(t), \quad (5.6)$$

where q^{-1} is the backward time-shift operator, e.g., $q^{-1}y(t) = y(t-1)$, and $\mathcal{A}(q^{-1})$ and $\mathcal{B}(q^{-1})$ are vector polynomials defined as $\mathcal{A}(q^{-1}) := [1 \ a_1q^{-1} \ \cdots \ a_nq^{-n}]$, and $\mathcal{B}(q^{-1}) := [b_1q^{-1} \ \cdots \ b_mq^{-m}]$. For a MISO system with ℓ inputs, (5.6) extends to

$$\mathcal{A}(q^{-1})y(t) = q^{-d_1}\mathcal{B}_1(q^{-1})u_1(t) + \cdots + q^{-d_\ell}\mathcal{B}_\ell(q^{-1})u_\ell(t), \quad (5.7)$$

where $\mathcal{B}_i(q^{-1}), i = 1, 2, \dots, \ell$, are low-order polynomials.

5.3 CSI of Linear Time-Invariant (LTI) ARX Models

Identification of LTI ARX models in both SISO and MISO cases is considered in this section. As a first step towards CSI and for the sake of simplicity we consider the noiseless case. Inspired by CS, we show that in cases where the LTI system has a *sparse* impulse response, simultaneous ARX model identification and input delay estimation is possible from a small number of observations, even though this leaves the aforementioned linear equations highly underdetermined. We discuss the required number of measurements in terms of metrics that guarantee exact identification, derive bounds on such metrics, and suggest a pre-filtering scheme by which these metrics can be reduced.

5.3.1 CSI of LTI Systems with Unknown Input Delays

Input delay estimation can be challenging, especially for large-scale multivariable (MISO or MIMO) systems when there exist several inputs with different unknown (possibly large) delays. Identification of such systems requires estimating (or guessing) the proper value

of the d_i 's separately. Typically this is done via model complexity metrics such as the AIC or BIC, or via cross-validation by splitting the available data into an identification set and a validation set and estimating the parameters on the identification set for a fixed set of parameters $\{d_i\}$. This procedure continues by fixing another set of parameters, and finishes by selecting the parameters that give the best fit on the validation set. However, complete delay estimation would require estimation and cross validation with all possible delay combinations, which can grow quickly with the number of inputs. For instance, with 5 inputs, checking for delays in each channel between 1 and 10 samples requires solving 10^5 least-squares problems. A review of other time-delay estimation techniques is given in [81]. For a sufficiently large number of inputs with possibly large delays, we will show that by using the tools in CS, it is possible to implicitly estimate the delays by favoring block-sparse solutions for $\boldsymbol{\theta}$.

Letting m_i be the length of \mathcal{B}_i and bounding the maximum length (including delays) for all inputs by p ($\max_i(d_i + m_i) \leq p$), we build the regression matrix with each $\Phi_{u_i} \in \mathbb{R}^{M \times p}$ to be a Toeplitz matrix associated with one input. This results in an $M \times (n + lp)$ matrix Φ . However, considering a low-order polynomial for each input ($\max_i m_i \leq m$) for some m , the corresponding parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{n+lp}$ has at most $n + lm$ non-zero entries. Assuming $m < p$, this formulation suggests *sparsity* of the parameter vector $\boldsymbol{\theta}$ and encourages us to use the tools in CS for recovery. Moreover, this allows us to do the identification from an underdetermined set of equations Φ where $M < n + lp$.

5.3.2 Simulation Results

Figure 5.1(a) illustrates the recovery of a $\{2, 2, 40\}$ SISO LTI ARX model where m and d are unknown. The only knowledge is of $p = 62$. For each system realization, the input is generated as an independent and identically distributed (i.i.d.) Gaussian random sequence. Assuming at least d iterations of the simulation have passed, M consecutive samples of the output are taken. As n is known, we modify the Block Orthogonal Matching Pursuit (BOMP) algorithm to include the first n locations as part of the support of $\boldsymbol{\theta}$.

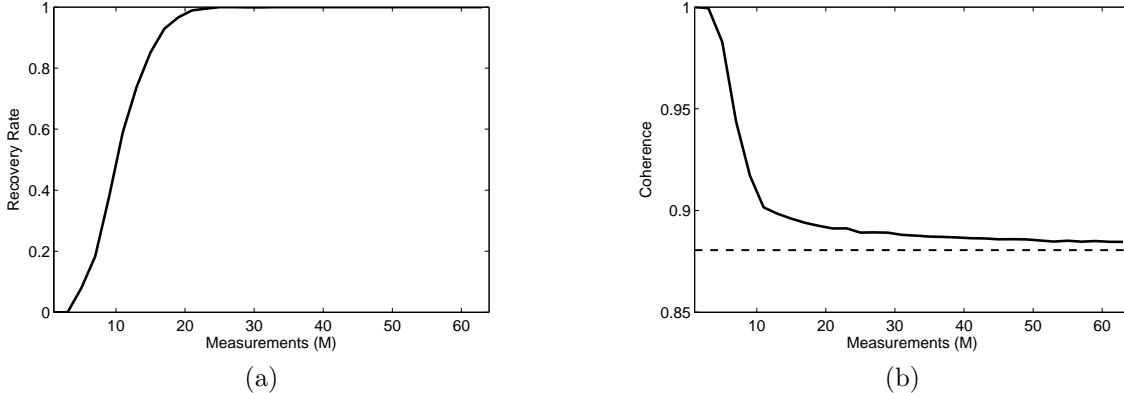


Figure 5.1: CSI results on a $\{2, 2, 40\}$ SISO LTI ARX System. (a) In the recovery algorithm, m and d are unknown. The plot shows the recovery success rate over 1000 realizations of the system. (b) Averaged mutual coherence of Φ over 1000 realizations of the system (solid curve). Lower bound of Theorem 5.8 (dashed line).

The plot shows the recovery success rate over 1000 realizations of the system. As shown in Figure 5.1(a), with 25 measurements, the system is perfectly identified in 100% of the trials. The average coherence value is also depicted in Figure 5.1(b) (solid curve). After taking a certain number of measurements, the average coherence converges to a constant value (dashed line). We will address this in detail in the next section.

Identification of a MISO system is shown in Figure 5.2 where the actual system has parameters $n = 2$, $m = 2$ for all inputs, and $d_1 = 60, d_2 = 21, d_3 = 10, d_4 = 41$. Assuming $p = 64$, the parameter vector θ has 258 entries, only 10 of which are non-zero. Applying the BOMP algorithm with n given and m and $\{d_i\}$ unknown, implicit input delay estimation and parameter identification is possible in 100% of the trials by taking $M = 150$ measurements.

5.3.3 Bounds on Coherence

As depicted in Figure 5.1(b), the typical coherence $\mu(\Phi)$ has an asymptotic behavior. In this section, we derive a lower bound on the typical value of $\mu(\Phi)$ for SISO LTI ARX models. Specifically, for a given system excited by a random i.i.d. Gaussian input, we are interested in finding $\mathbf{E}[\mu(\Phi)]$ where Φ is as in (5.3).

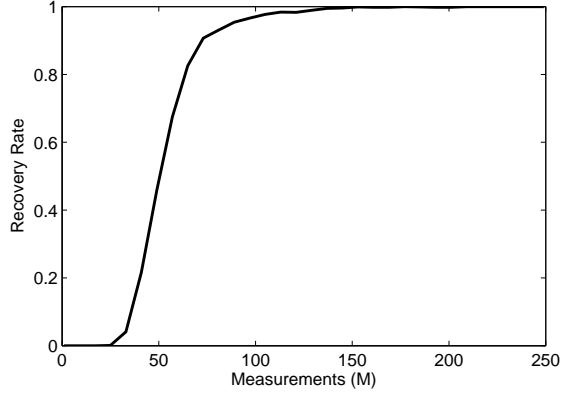


Figure 5.2: CSI results on a $\{2, 2, \{60, 21, 10, 41\}\}$ MISO LTI ARX system. In the recovery algorithm, m and $\{d_i\}_{i=1}^4$ are unknown. The plot shows the recovery success rate over 1000 realizations of the system.

Theorem 5.8 *Suppose the system described by difference equation in (5.1) (ARX model $\{n, m, d\}$) is characterized by its impulse response $h(k)$ in a convolution form as*

$$y(t) = \sum_{k=-\infty}^{\infty} h(k)u(t-k). \quad (5.9)$$

Then, for a zero mean, unit variance i.i.d. Gaussian input,

$$\lim_{M \rightarrow \infty} \mathbf{E}[\mu(\Phi)] \geq \max_{s \neq 0} \left\{ \frac{|\mathcal{H}(s)|}{\|h\|_2^2}, \frac{|h(s)|}{\|h\|_2} \right\} \quad (5.10)$$

where $\mathcal{H}(s) = \sum_{k=-\infty}^{\infty} h(k)h(k+s)$.

Proof See Appendix A.4. ■

Discussion:

As Theorem 5.8 suggests, the typical coherence of Φ is bounded below by a non-zero value that depends on the impulse response of the system and it has an asymptotic behavior. For example, for the system given in Figure 5.1, the typical coherence does not get lower than 0.88 even for large M . With this value of coherence, the analytical recovery guarantees for the BOMP algorithm [16], which can be reasonably represented by mutual coherence defined in (4.13), do not guarantee recovery of any one-block sparse signals. However, as can be seen in Figure 5.1(a), perfect recovery is possible. This indicates a gap between the

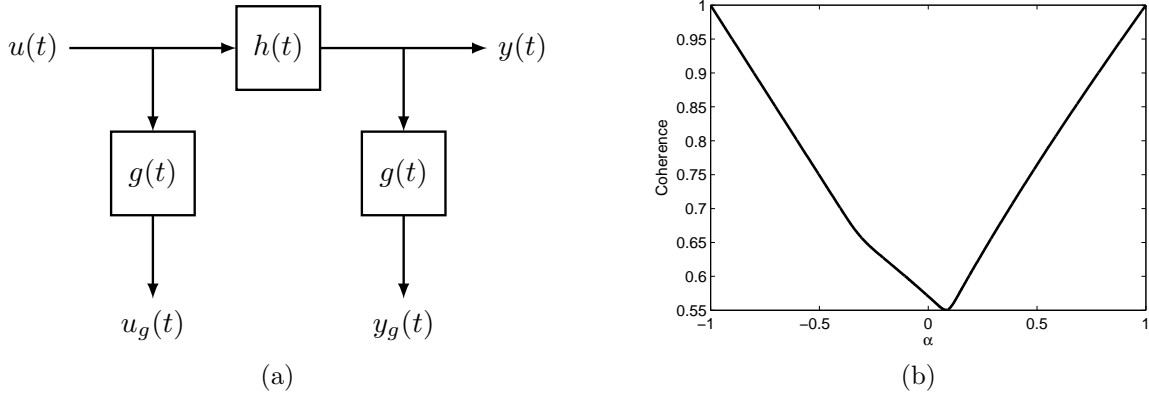


Figure 5.3: Reducing coherence by pre-filtering. (a) Pre-filtering scheme. (b) For each α , the filter $G(z)$ is applied on the input/output signals and the limit of the expected value of coherence is calculated over 1000 realizations of system.

available analytical guarantee and the true recovery performance for ARX systems. This suggests that coherence-based performance guarantees for matrices that appear in ARX identification are not sharp tools as they only reflect the worst correlations in the matrix. As a first step towards investigating this gap, we suggest a pre-filtering scheme by which the coherence of such matrices can be reduced.

5.3.4 Reducing Coherence by Pre-Filtering

In this section, we show that we can reduce the coherence by designing a pre-filter g applied on u and y .

Theorem 5.11 *Assume the system described as in Theorem 5.8. Given a filter g , define $u_g = u * g$ and $y_g = y * g$. Build the regression matrix Φ_g from u_g and y_g as in (5.3). The pre-filtering scheme is shown in Figure 5.3(a). Then we have*

$$\lim_{M \rightarrow \infty} \mathbf{E} [\mu(\Phi_g)] \geq \max_{s \neq 0} \left\{ \frac{|\mathcal{G}(s)|}{\|g\|_2^2}, \frac{|\mathcal{F}(s)|}{\|f\|_2^2}, \frac{|\mathcal{GF}(s)|}{\|g\|_2 \|f\|_2} \right\},$$

where $f = g * h$, $\mathcal{G}(s) = \sum_{k=-\infty}^{\infty} g(k)g(k+s)$, $\mathcal{F}(s) = \sum_{k=-\infty}^{\infty} f(k)f(k+s)$, and $\mathcal{GF}(s) = \sum_{k=-\infty}^{\infty} g(k)f(k+s)$.

Proof See Appendix A.5. ■

Theorem 5.11 suggests that by choosing an appropriate filter $g(t)$, the typical coherence can possibly be reduced, although it is bounded below by a non-zero value. We follow the discussion by showing how the coherence of Φ can be reduced by pre-filtering within an illustrative example. Consider a SISO system characterized by the transfer function

$$H(z) = \frac{z - 0.4}{(z + 0.9)(z + 0.2)}. \quad (5.12)$$

Using the bound given in Theorem 5.8, for large M , $\mathbf{E}[\mu\Phi] \geq 0.95$ which indicates a highly correlated matrix Φ . However, using the analysis given in Theorem 5.11, we can design a filter $G(z)$ such that the coherence of the resulting matrix Φ_g is reduced almost by half. For example, consider a notch filter $G(z)$ given by

$$G(z) = \frac{z + 0.9}{(z + \alpha)}, \quad (5.13)$$

where α is a parameter to be chosen. For a given α , the filter $G(z)$ is applied on the input/output data as illustrated in Figure 5.3(a) and the average coherence of Φ_g is calculated. The result of this pre-filtering and its effect on the coherence is shown in Figure 5.3(b). The results indicate that actual performance of Φ may actually be better than what $\mu(\Phi)$ suggests. As can be seen, for α around 0.1, the coherence is reduced to 0.55 which is almost half of the primary coherence.

5.4 CSI of Linear Time-Variant (LTV) ARX Models

In (5.2) the parameters are assumed to be fixed over time. In this section, we study ARX models where the parameter vector $\boldsymbol{\theta}(t)$ is varying over time. As an extension of (5.2), for time-varying systems, we have

$$y(t) = \boldsymbol{\phi}^T(t) \boldsymbol{\theta}(t) + e(t).$$

Collecting M consecutive measurements of such a system and following similar steps, for a SISO LTV ARX model we can formulate the parameter estimation problem as

$$\underbrace{\begin{bmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+M-1) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \phi^T(t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \phi^T(t+1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \phi^T(t+M-1) \end{bmatrix}}_{\Omega} \underbrace{\begin{bmatrix} \boldsymbol{\theta}(t) \\ \boldsymbol{\theta}(t+1) \\ \vdots \\ \boldsymbol{\theta}(t+M-1) \end{bmatrix}}_{\boldsymbol{\vartheta}} + \mathbf{e}$$

or equivalently

$$\mathbf{y} = \Omega \boldsymbol{\vartheta} + \mathbf{e}, \quad (5.14)$$

where for simplicity $d = 0$, $p = m$, $\mathbf{y} \in \mathbb{R}^M$, $\Omega \in \mathbb{R}^{M \times M(n+m)}$ and $\boldsymbol{\vartheta} \in \mathbb{R}^{M(n+m)}$. The goal is to solve (5.14) for $\boldsymbol{\vartheta}$ given \mathbf{y} and Ω . Typical estimation is via

$$\min_{\boldsymbol{\vartheta}} \|\mathbf{y} - \Omega \boldsymbol{\vartheta}\|_2^2. \quad (5.15)$$

However, the minimization problem in (5.15) contains an underdetermined set of equations ($M < M(n+m)$) and therefore has many solutions.

5.4.1 Piecewise-Constant $\boldsymbol{\theta}(t)$ and Block-Sparse Recovery

Assuming $\boldsymbol{\theta}(t)$ is piecewise-constant, we show how the LTV ARX identification problem can be formulated as the recovery of a block-sparse signal. Using the developed tools in CS we show the identification of such systems can be done from relatively few measurements. Assume that $\mathbf{e} = \mathbf{0}$ and that $\boldsymbol{\theta}(t)$ changes only at a few time instants $t_i \in \mathcal{C}$ where $\mathcal{C} \triangleq \{t_1, t_2, \dots\}$ with $|\mathcal{C}| \ll M$, i.e.,

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t_i), \quad t_i \leq t < t_{i+1}. \quad (5.16)$$

Note that neither the change moments t_i nor the number of changes is known a priori to the identification algorithm. An example of $\boldsymbol{\vartheta}$ would be

$$\boldsymbol{\vartheta} = [\boldsymbol{\theta}^T(t_1) \ \cdots \ \boldsymbol{\theta}^T(t_1) \ \boldsymbol{\theta}^T(t_2) \ \cdots \ \boldsymbol{\theta}^T(t_2)]^T \quad (5.17)$$

which has 2 different constant pieces, i.e., $\mathcal{C} = \{t_1, t_2\}$. In order to exploit the existing sparsity pattern in $\boldsymbol{\vartheta}$, define the differencing operator

$$\Delta = \begin{bmatrix} -I_{n+m} & 0_{n+m} & \cdots & \cdots & 0_{n+m} \\ I_{n+m} & -I_{n+m} & \ddots & \ddots & \vdots \\ 0_{n+m} & I_{n+m} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0_{n+m} \\ 0_{n+m} & \cdots & 0_{n+m} & I_{n+m} & -I_{n+m} \end{bmatrix}.$$

Applying Δ to $\boldsymbol{\vartheta}$, we define $\boldsymbol{\vartheta}_\delta$ as

$$\boldsymbol{\vartheta}_\delta = \Delta \boldsymbol{\vartheta}, \quad (5.18)$$

which has a block-sparse structure. For the given example in (5.17), we have

$$\boldsymbol{\vartheta}_\delta = [-\boldsymbol{\theta}^T(t_1) \mathbf{0} \cdots \mathbf{0} \boldsymbol{\theta}^T(t_1) - \boldsymbol{\theta}^T(t_2) \mathbf{0} \cdots \mathbf{0}]^T. \quad (5.19)$$

The vector $\boldsymbol{\vartheta}_\delta \in \mathbb{R}^{M(n+m)}$ in (5.19) now has a block-sparse structure: out of its $M(n+m)$ entries, grouped in M blocks of length $n+m$, only a few of them are non-zero and they appear in block locations. The number of non-zero blocks corresponds to the number of different levels of $\boldsymbol{\theta}(t)$. In the example given in (5.17), $\boldsymbol{\theta}(t)$ takes 2 different levels over time and thus, $\boldsymbol{\vartheta}_\delta$ has a block-sparsity level of 2 with each block size of $n+m$. Note that Δ^{-1} has the form

$$\Delta^{-1} = \begin{bmatrix} -I_{n+m} & 0_{n+m} & \cdots & \cdots & 0_{n+m} \\ -I_{n+m} & -I_{n+m} & \ddots & \ddots & \vdots \\ -I_{n+m} & -I_{n+m} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0_{n+m} \\ -I_{n+m} & \cdots & -I_{n+m} & -I_{n+m} & -I_{n+m} \end{bmatrix}.$$

By this formulation, the parameter estimation of LTV ARX models with piecewise-constant parameter changes can be cast as recovering a block-sparse signal $\boldsymbol{\vartheta}_\delta$ from measurements

$$\mathbf{y} = \Omega_\delta \boldsymbol{\vartheta}_\delta, \quad (5.20)$$

where $\Omega_\delta = \Omega \Delta^{-1}$.

5.4.2 Identifiability Issue

Before presenting the simulation results, we address the identifiability issue faced in the LTV case. The matrix Ω_δ has the following structure.

$$\Omega_\delta = \begin{bmatrix} -\boldsymbol{\phi}^T(t) & \mathbf{0} & \mathbf{0} & \cdots \\ -\boldsymbol{\phi}^T(t+1) & -\boldsymbol{\phi}^T(t+1) & \mathbf{0} & \cdots \\ -\boldsymbol{\phi}^T(t+2) & -\boldsymbol{\phi}^T(t+2) & -\boldsymbol{\phi}^T(t+2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

If the change in the system actually happens at time instant $t+2$, the corresponding solution to (5.20) has the form

$$\boldsymbol{\vartheta}_\delta = [-\boldsymbol{\theta}^T(t_1) \quad \mathbf{0} \quad \boldsymbol{\theta}^T(t_1) - \boldsymbol{\theta}^T(t_2) \quad \mathbf{0} \quad \cdots]^T.$$

However, due to the special structure of the matrix Ω_δ , there exist other solutions to this problem. For example

$$\widehat{\boldsymbol{\vartheta}}_\delta = [-\boldsymbol{\theta}^T(t_1) \quad \mathbf{0} \quad \boldsymbol{\theta}^T(t_1) - \boldsymbol{\theta}^T(t_2) + \boldsymbol{\gamma}^T \quad -\boldsymbol{\gamma}^T \quad \cdots]^T$$

is another solution where $\boldsymbol{\gamma}$ is a vector in the null space of $\boldsymbol{\phi}^T(t)$, i.e., $\boldsymbol{\phi}^T(t)\boldsymbol{\gamma} = 0$. However, this only results in a small ambiguity in the solution around the transition point. Therefore, $\widehat{\boldsymbol{\vartheta}}_\delta$ can be considered as an acceptable solution as $\widehat{\boldsymbol{\vartheta}} = \Delta^{-1}\widehat{\boldsymbol{\vartheta}}_\delta$ is exactly equal to the true parameter vector $\boldsymbol{\vartheta}$ except at very few time instants around the transition point. In the next section, we consider $\widehat{\boldsymbol{\vartheta}}_\delta$ as a valid solution.

5.4.3 Sampling Approach for LTV System Identification

In this section, we suggest a sampling scheme for identifying LTV systems. Note that in a noiseless scenario, the LTV identification can be performed by taking consecutive observations in a frame, identifying the system on that frame, and then moving the frame forward until we identify a change in the system. Of course, this can be very inefficient when the time instants at which the changes happen are unknown to us beforehand as we end up taking many unnecessary measurements. As an alternative, we suggest a *random* sampling scheme (as compared to consecutive sampling) for identifying such LTV systems. Figure 5.4 shows examples of this sampling approach for $M = 10$, $M = 30$ and $M = 50$ measurements. As can be seen, the samples are chosen randomly according to a uniform distribution. Note that these samples are not necessarily consecutive. By this approach, we can dramatically

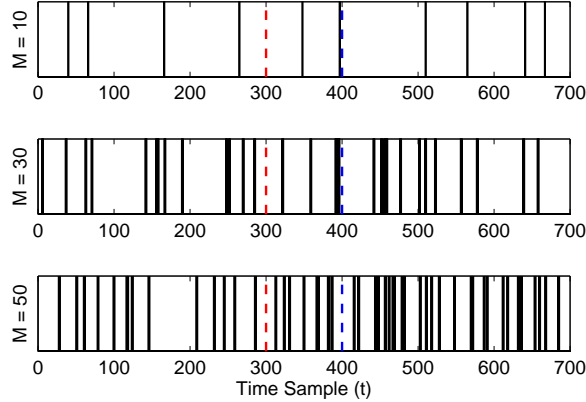


Figure 5.4: Random sampling scheme for $M = \{10, 30, 50\}$ measurements. Samples are chosen randomly according to a uniform distribution. System parameters are assumed to change at $t = 300$ and $t = 400$.

reduce the required number of measurements for LTV system identification.

5.4.4 Simulation Results

Consider a system described by its $\{2, 2, 0\}$ ARX model

$$y(t) + a_1y(t - 1) + a_2y(t - 2) = b_1u(t - 1) + b_2u(t - 2) \quad (5.21)$$

with i.i.d. Gaussian input $u(t) \sim \mathcal{N}(0, 1)$. Figure 5.5(a) shows one realization of the output of this system whose parameters are changing over time as shown in Figure 5.5(b). As can be seen, the parameters change in a piecewise-constant manner over 700 time instants at $t = 300$ and $t = 400$. The goal of the identification is to identify the parameters of this time-variant system along with the location of the changes. Figure 5.6 illustrates the recovery performance of 4 LTV systems, each with a different number of changes over time. For each measurement sequence (randomly selected), 1000 realizations of the system are carried out. We highlight two points about this plot. First, we are able to identify a system (up to the ambiguity around the time of change as discussed in Section 5.4.2) which changes 3 times over 700 time instants by taking only 50 measurements without knowing the location of the changes. Second, the number of measurements sufficient for correct recovery scales with the

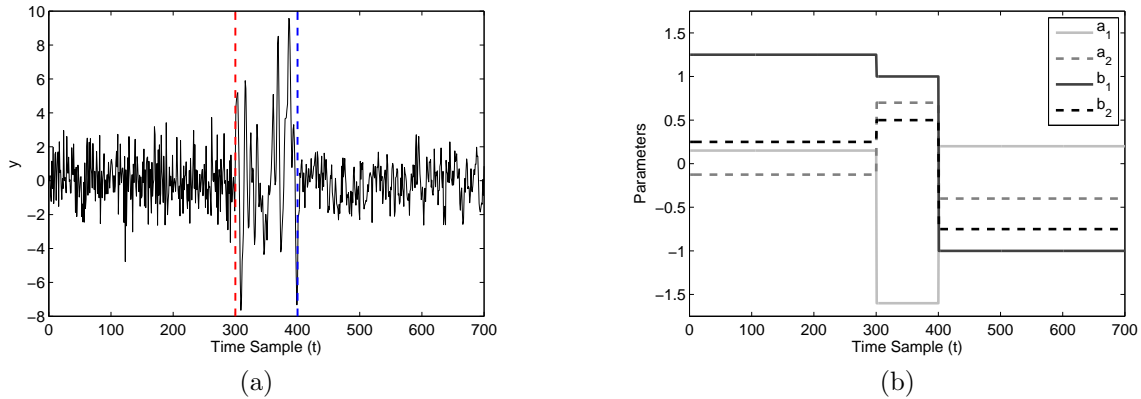


Figure 5.5: A 3-model LTV ARX system. (a) Output of the model. System parameters change at $t = 300$ and $t = 400$. (b) Time-varying parameters of the 3-model system. At the time of change, all the system parameters change.

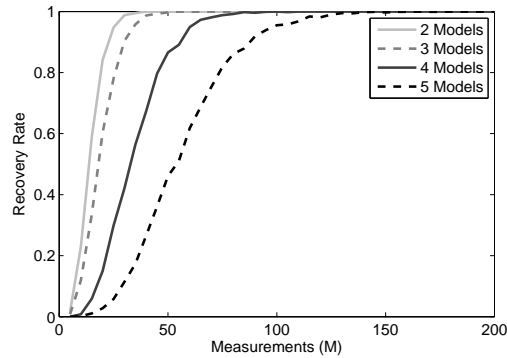


Figure 5.6: Recovery performance of 4 different systems. The plots show the recovery success rate over 1000 realizations of the system.

number of changes that a system undergoes over the course of identification. Systems with more changes require more measurements for correct recovery and identification.

5.5 Case Study: Harmonic Drive System

In this section, we apply the proposed identification scheme to a set of noisy experimental input-output data from a DC motor system. In this case study, the DC motor system shown in Figure 5.7 is used as the data generating system. The system under study consists of a DC motor with a current input ($\triangleq i(t)$), a harmonic drive system, an inertial load,

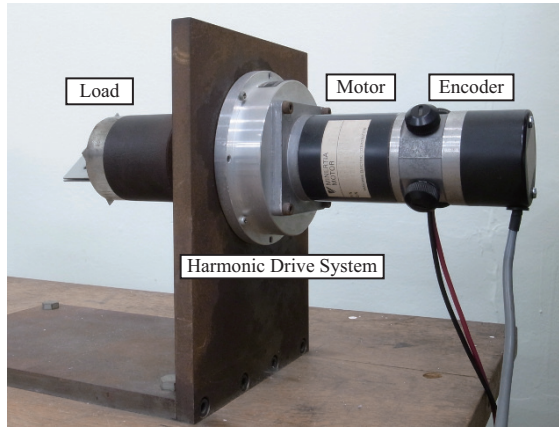


Figure 5.7: A DC Motor System consisting of a DC motor, a harmonic drive system, an inertial load, and an encoder.

and an encoder which outputs the rotation angle. The harmonic drive system is a torque transmission system, which has widespread industrial applications and complex dynamic behavior [82].

5.5.1 Acquisition of the Input-Output Data Set

In order to obtain the required input-output data for performing the identification, a sinusoidal electric current $i(t)$ with 5.12[s] period is applied to the system, and the rotation angle is sampled with constant sampling interval 0.01[s].

Then, the angular velocity ($\triangleq \omega(t)$) and its derivative $\dot{\omega}(t)$ are calculated by backward and central difference, respectively. Since backward and central difference produce large noise components, an input-output data set averaged over 1000 cycles is used in the identification process. The resulting input-output data set is shown in Figure 5.8. As can be seen, the response of the angular velocity $\omega(t)$ is clearly distorted by the non-linearity of the system.

5.5.2 Construction of the time-varying ARX Model

The relation between the input-output variables in a DC motor system can be represented as

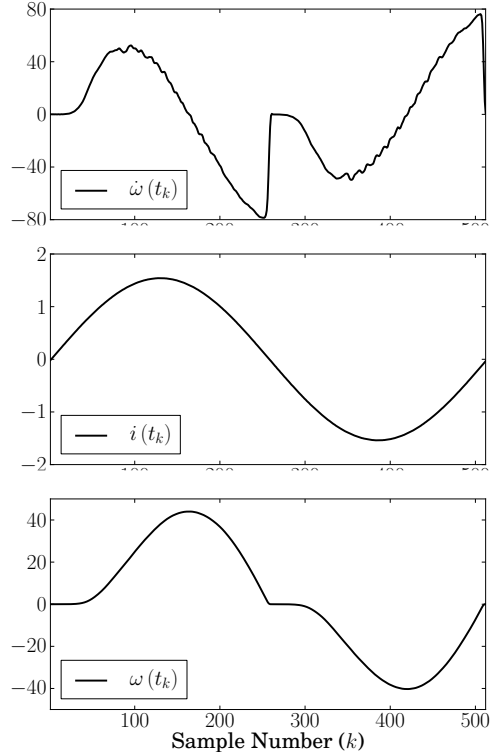


Figure 5.8: The experimental input-output data of a harmonic drive system. A sinusoidal electric current $i(t)$ with 5.12[s] period is applied to the system, and the rotation angle is sampled with constant sampling interval 0.01[s] (i.e., $t_k = 0.01(k - 1)$). The angular velocity ($\triangleq \omega(t)$) and its derivative $\dot{\omega}(t)$ are calculated by backward and central difference, respectively.

$$\dot{\omega}(t) = K(t)i(t) + D(t)\omega(t) + F(t), \quad (5.22)$$

where $D(t)$ corresponds to the factor of viscous friction, $K(t)$ corresponds to the torque factor, and $F(t)$ corresponds to some additional forces. In many applications in mechanical friction, and in order to simplify the varying behavior of (5.22), models with LTI parts (constant $K(t)$ and $D(t)$) and an additional piecewise constant part (such as Coulomb friction) are proposed [83]. This therefore indicates that a good LTV model with piecewise constant parameters exists for such systems. In this section, we show that such a time-varying ARX model can be obtained by the proposed method.

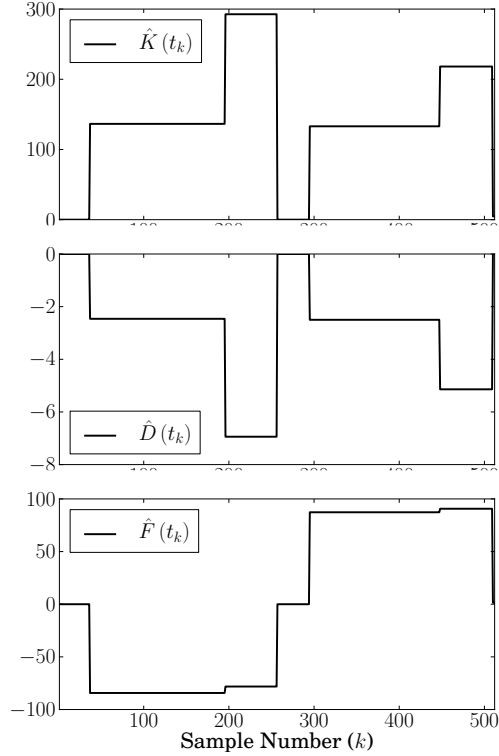


Figure 5.9: The estimated ARX model parameters corresponding to the DC motor system. As can be seen, the identified parameters have a piecewise-constant behavior with only a few changes happening over the course of experiment.

The sampled version of (5.22) can be represented as an ARX model where the parameter vector is defined as

$$\boldsymbol{\theta}(k) \triangleq [K(t_k), D(t_k), F(t_k)]^T, \quad (5.23)$$

where $t_k \triangleq 0.01(k-1)$ for $k = 1, 2, \dots, 512$, and the regression vector is defined as

$$\boldsymbol{\phi}(k) \triangleq [i(t_k), \omega(t_k), 1]^T. \quad (5.24)$$

Thus, the sampled output $y(k) \triangleq \dot{\omega}(t_k)$ can be written as

$$y(k) = \boldsymbol{\phi}^T(k) \boldsymbol{\theta}(k). \quad (5.25)$$

5.5.3 Identifiability Issue

The BOMP algorithm given in Algorithm 1 is initially designed for block-sparse recovery of signals that are measured by a matrix which is populated by i.i.d. random entries. This

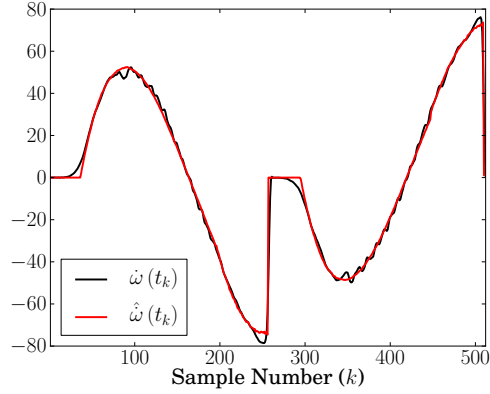


Figure 5.10: A comparison between the measured experimental output $\hat{\omega}(t_k)$ and the estimated output $\hat{\omega}(t_k)$ based on the identified piecewise-constant ARX model parameters.

assumption makes the columns of the matrix with almost equal norm and well conditioned. However, in our block-sparse formulation of the LTV-ARX identification given in (5.20), the regression matrix Ω_δ has the following structure:

$$\Omega_\delta = \begin{bmatrix} -\phi^T(1) & \mathbf{0} & \mathbf{0} & \dots \\ -\phi^T(2) & -\phi^T(2) & \mathbf{0} & \dots \\ -\phi^T(3) & -\phi^T(3) & -\phi^T(3) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Clearly, this matrix does not have random entries. Instead, it is populated by input-output data driven from the experiment. Furthermore, because of its particular structure the ℓ_2 norm of the last columns are much smaller compared to the ℓ_2 norm of the first columns. Moreover the experimental data is noisy. In the following subsections we explain how we treat these issues.

5.5.4 BOMP for Noisy Data

In a noisy case, we have

$$\mathbf{y} = \Omega_\delta \boldsymbol{\vartheta}_\delta + \mathbf{z},$$

where \mathbf{z} is a noise component with $\|\mathbf{z}\|_2 \leq \nu$. When noisy data are given, we modify the BOMP algorithm such that it incorporates our knowledge of noise level. We simply change the stopping criteria of the algorithm to

$$\|\mathbf{y} - \Omega_\delta \widehat{\boldsymbol{\vartheta}}_\delta\|_2 \leq \epsilon,$$

where $\epsilon \geq \nu$. The choice of ϵ in the BOMP algorithm tunes the trade-off between output matching ($\|\mathbf{y} - \widehat{\mathbf{y}}\|_2$) and parameter matching ($\|\boldsymbol{\vartheta}_\delta - \widehat{\boldsymbol{\vartheta}}_\delta\|_2$).

Observe that the chosen value for ϵ can affect the noisy recovery performance. By choosing small value for ϵ we are forcing the BOMP algorithm to over-fit the parameters, resulting in a solution $\widehat{\boldsymbol{\vartheta}}_\delta$ which is not the sparsest (in a block-sparse sense) possible solution. However, by forfeiting a small amount in the output matching error, BOMP can recover the true block-sparse solution. Increasing ϵ to much higher values will allow the BOMP algorithm to recover a solution which neither is sparse (increased parameter matching error) nor corresponds to acceptable output matching.

5.5.5 Scaling and Column Normalization

As mentioned earlier, the regression matrix Ω_δ have columns with different ℓ_2 norms. The first column of Ω_δ contains the electric current measurements ($-i(t_k)$), the second column of Ω_δ contains the angular velocity measurements ($-\omega(t_k)$), and the third column is all -1 . The remaining columns of Ω_δ are zero-padded versions of these 3 columns. Because of this particular structure, Ω_δ has columns with different ℓ_2 norms. We should take this issue into account before applying the BOMP algorithm. First observe that (as shown in Figure 5.8) there exists a substantial order difference between the current value $i(t_k)$ and the angular velocity $\omega(t_k)$. Therefore, we first scale our observations from these variables. This scaling makes the first 3 columns of Ω_δ have almost equal norms. However, even after scaling the variables, the ℓ_2 norm of the columns of Ω_δ decreases significantly as the column number increases. This is due to the particular structure of Ω_δ which results in poor recovery of parameters. In particular, we observe that this leads to undesirable effects in the second halves of the identified parameters. Observe that the second half of the parameters corresponds to the second half of the columns of Ω_δ . In order to overcome this issue, we take advantage of the fact that the data is periodic (we applied a sinusoidal periodic electric

current $i(t)$ to the system) and perform the parameter identification in two steps. We estimate the first half of the parameters by applying the algorithm on the data as shown in Figure 5.8. We then divide the input-output data into two halves and then wrap the signal around by changing the location of these two halves. We recover the second half of the parameters by applying the algorithm on this data set. The final estimate of the parameters is a concatenation of the two recovered halves.

5.5.6 Results

Figure 5.9 illustrates the identified ARX model parameters using the proposed approach. As can be seen, the identified parameters have a piecewise-constant behavior with only a few changes happening over the course of experiment. Figure 5.10 shows the estimated output $\hat{\omega}(t_k)$ corresponding to the identified piecewise-constant parameters versus the measured experimental output $\dot{\omega}(t_k)$. As can be seen, the output of the identified time-varying ARX model reasonably approximates the true data. The identified parameters also plausibly represent the DC motor system. For example, the identified $\hat{F}(t_k)$ represents the static friction. Moreover, different modes are assigned to the acceleration and deceleration phases. These seem quite likely appropriate for the model of DC motor system with a harmonic drive speed reducer.

CHAPTER 6

OBSERVABILITY WITH RANDOM OBSERVATIONS

Recovering or estimating the initial state of a high-dimensional system can require a large number of measurements. In this chapter⁷ we explain how this burden can be significantly reduced for certain linear systems when randomized measurement operators are employed. Our work builds upon recent results from Compressive Sensing (CS) and sparse signal recovery. In particular, we make the connection to CS analysis for block-diagonal measurement matrices [50, 84–87].

We first show that the observability matrix satisfies the Restricted Isometry Property (RIP) (a sufficient condition on the measurement matrix for stable recovery of sparse vectors) under certain conditions on the state transition matrix. For example, we show that if the state transition matrix is unitary, and if independent, randomly-populated measurement matrices are employed, then it is possible to uniquely recover a sparse high-dimensional initial state when the total number of measurements scales *linearly* in the sparsity level of the initial state. We support our analysis with a case study of a diffusion system.

We then derive Concentration of Measure (CoM) inequalities for the observability matrix and explain how the interaction between the state transition matrix and the initial state affect the concentration bounds. The concentration results cover a larger class of systems (not necessarily unitary) and initial states (not necessarily sparse). Aside from guaranteeing recovery of sparse initial states, the CoM results have potential applications in solving inference problems such as detection and classification of more general initial states and systems.

⁷This work is in collaboration with Tyrone L. Vincent and Michael B. Wakin [11, 12].

6.1 Introduction

In this chapter, we consider the problem of recovering the initial state of a high-dimensional system from compressive measurements (i.e., we take fewer measurements than the system dimension).

6.1.1 Measurement Burdens in Observability Theory

Consider an N -dimensional discrete-time linear dynamical system described by the state equation⁸

$$\begin{aligned}\mathbf{x}_k &= A\mathbf{x}_{k-1} \\ \mathbf{y}_k &= C_k\mathbf{x}_k\end{aligned}\tag{6.1}$$

where $\mathbf{x}_k \in \mathbb{R}^N$ represents the state vector at time $k \in \{0, 1, 2, \dots\}$, $A \in \mathbb{R}^{N \times N}$ represents the state transition matrix, $\mathbf{y}_k \in \mathbb{R}^M$ represents a set of measurements (or “observations”) of the state at time k , and $C_k \in \mathbb{R}^{M \times N}$ represents the measurement matrix at time k . Thus, the number of measurements at each time index is M . For any finite set $\Omega \subset \{0, 1, 2, 3, \dots\}$, define the *generalized observability matrix*

$$\mathcal{O}_\Omega := \begin{bmatrix} C_{k_0}A^{k_0} \\ C_{k_1}A^{k_1} \\ \vdots \\ C_{k_{K-1}}A^{k_{K-1}} \end{bmatrix} \in \mathbb{R}^{MK \times N},\tag{6.2}$$

where $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$ contains K observation times. In order to have general results, in this chapter we generalize the traditional definition of the observability matrix by considering arbitrary time samples in (6.2). This definition matches the traditional definition when $\Omega = \{0, 1, \dots, K-1\}$. The primary use of observability is in ensuring that a state (say, an initial state \mathbf{x}_0) can be recovered from a collection of measurements $\{\mathbf{y}_{k_0}, \mathbf{y}_{k_1}, \dots, \mathbf{y}_{k_{K-1}}\}$.

⁸The results of this chapter directly apply to systems described by a state equation of the form

$$\begin{aligned}\mathbf{x}_k &= A\mathbf{x}_{k-1} + B\mathbf{u}_k \\ \mathbf{y}_k &= C_k\mathbf{x}_k + D\mathbf{u}_k,\end{aligned}$$

where $\mathbf{u}_k \in \mathbb{R}^P$ is the input vector at sample time k and $B \in \mathbb{R}^{N \times P}$ and $D \in \mathbb{R}^{M \times P}$ are constant matrices. Indeed, initial state recovery is independent of B and D when it is assumed that the input vector \mathbf{u}_k is known for all sample times k .

In particular, defining

$$\mathbf{y}_\Omega := \begin{bmatrix} \mathbf{y}_{k_0}^T & \mathbf{y}_{k_1}^T & \cdots & \mathbf{y}_{k_{K-1}}^T \end{bmatrix}^T,$$

we have

$$\mathbf{y}_\Omega = \mathcal{O}_\Omega \mathbf{x}_0. \tag{6.3}$$

Although we will consider situations where C_k changes with each k , we first discuss the classical case where $C_k = C$ (C is assumed to have full row rank) for all k and $\Omega = \{0, 1, \dots, K - 1\}$ (K consecutive measurements). In this setting, an important and classical result [88] states that a system described by the state equation (6.1) is observable if and only if \mathcal{O}_Ω has rank N (full column rank), where $\Omega = \{0, 1, \dots, N - 1\}$. One challenge in exploiting this fact is that for some systems, N can be quite large. For example, distributed systems evolving on a spatial domain can have a large state space even after taking a spatially-discretized approximation. In such settings, we might therefore require a very large total number of measurements (MN) to identify an initial state, and moreover, inverting the matrix \mathcal{O}_Ω could be very computationally demanding.

This raises an interesting question: under what circumstances might we be able to infer the initial state of a system when $K < N$? We might imagine, for example, that the measurement burden could be alleviated in cases when there is a model for the state \mathbf{x}_0 that we wish to recover. Alternatively, we may have cases where, rather than needing to recover \mathbf{x}_0 from \mathbf{y}_Ω , we desire only to solve a much simpler inference problem such as a binary detection or a classification problem. In this work, inspired by CS [14, 26], we explain how such assumptions can indeed reduce the measurement burden and, in some cases, even allow recovery of the initial state when $MK < N$ and the system of equations (6.3) is guaranteed to be underdetermined.

6.1.2 Compressive Sensing and Randomized Measurements

The CS theory states that it is possible to solve certain rank-deficient sets of linear equations by imposing some model assumption on the signal to be recovered. In particular,

suppose $\mathbf{y} = \Phi \mathbf{x}$ where Φ is an $\widetilde{M} \times N$ matrix with $\widetilde{M} < N$. Suppose also that $\mathbf{x} \in \mathbb{R}^N$ is S -sparse, meaning that only S out of its N entries are non-zero.⁹ If Φ satisfies the RIP of order $2S$ for a sufficiently small isometry constant δ_{2S} , then it is possible to uniquely recover any S -sparse signal \mathbf{x} from the measurements $\mathbf{y} = \Phi \mathbf{x}$ using a tractable convex optimization program known as ℓ_1 -minimization [14, 22, 26]. In order to keep the chapter self-contained, we provide the definition of the RIP in the following.

Definition 6.4 *A matrix Φ is said to satisfy the RIP of order S and isometry constant $\delta_S := \delta_S(\Phi) \in (0, 1)$ if*

$$(1 - \delta_S) \|\mathbf{x}\|_2^2 \leq \|\Phi \mathbf{x}\|_2^2 \leq (1 + \delta_S) \|\mathbf{x}\|_2^2 \quad (6.5)$$

holds for all S -sparse vectors $\mathbf{x} \in \mathbb{R}^N$.

A simple way [32] of proving the RIP for a randomized construction of Φ involves first showing that the matrix satisfies a CoM inequality akin to the following.

Definition 6.6 [32, 89] *A random matrix (a matrix whose entries are drawn from a particular probability distribution) $\Phi \in \mathbb{R}^{\widetilde{M} \times N}$ is said to satisfy the Concentration of Measure (CoM) inequality if for any fixed signal $\mathbf{x} \in \mathbb{R}^N$ (not necessarily sparse) and any $\epsilon \in (0, \bar{\epsilon})$,*

$$\mathbf{P} \left\{ \left| \|\Phi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| > \epsilon \|\mathbf{x}\|_2^2 \right\} \leq 2 \exp \left\{ -\widetilde{M} f(\epsilon) \right\}, \quad (6.7)$$

where $f(\epsilon)$ is a positive constant that depends on the isometry constant ϵ , and $\bar{\epsilon} \leq 1$ is some maximum value of the isometry constant for which the CoM inequality holds.

Note that in the above definition, the failure probability decays exponentially fast in the number of measurements \widetilde{M} times some constant $f(\epsilon)$ that depends on the isometry constant ϵ . Baraniuk et al. [32] and Mendelson et al. [48] showed that a CoM inequality of the form (6.7) can be used to prove the RIP for random compressive matrices. This result is rephrased by Davenport [90] as follows.

⁹This is easily extended to the case where \mathbf{x} is sparse in some transform basis.

Lemma 6.8 [90] *Let \mathcal{X} denote an S -dimensional subspace in \mathbb{R}^N . Let $\delta_S \in (0, 1)$ denote a distortion factor and $\nu \in (0, 1)$ denote a failure probability, and suppose Φ is an $\widetilde{M} \times N$ random matrix that satisfies the CoM inequality (6.7) with*

$$\widetilde{M} \geq \frac{S \log(\frac{42}{\delta_S}) + \log(\frac{2}{\nu})}{f(\frac{\delta_S}{\sqrt{2}})}.$$

Then with probability at least $1 - \nu$,

$$(1 - \delta_S) \|\mathbf{x}\|_2^2 \leq \|\Phi \mathbf{x}\|_2^2 \leq (1 + \delta_S) \|\mathbf{x}\|_2^2$$

for all $\mathbf{x} \in \mathcal{X}$.

Through a union bound argument (see, for example, Theorem 5.2 in [32]) and by applying Lemma 6.8 for all $\binom{N}{S}$ S -dimensional subspaces that define the space of S -sparse signals in \mathbb{R}^N , one can show that Φ satisfies the RIP (of order S and with isometry constant δ_S) with high probability when \widetilde{M} scales *linearly* in S and *logarithmically* in N .

Aside from connections to the RIP, concentration inequalities such as the above can also be useful when solving other types of inference problems from compressive measurements. For example, rather than recovering a signal \mathbf{x} , one may wish only to solve a binary detection problem and determine whether a set of measurements \mathbf{y} correspond only to noise (the null hypothesis $\mathbf{y} = \Phi(\text{noise})$) or to signal plus noise ($\mathbf{y} = \Phi(\mathbf{x} + \text{noise})$). When Φ is random, the performance of a compressive detector (and of other multi-signal classifiers) can be studied using concentration inequalities [3, 4, 61], and in these settings it is not necessary to assume that \mathbf{x} is sparse.

6.1.3 Observability from Random, Compressive Measurements

In order to exploit CS concepts in observability analysis, we consider in this work scenarios where the measurement matrices C_k are populated with random entries. Physically, such randomized measurements may be taken using the types of CS protocols and hardware mentioned above. Our analysis is therefore appropriate in cases where one has some control over the sensing process.

As is apparent from (6.2), even with randomness in the matrices C_k , the observability matrices \mathcal{O}_Ω will contain some structure and cannot simply modeled as being populated with independent and identically distributed (i.i.d.) Gaussian random variables and thus, existing results can not be directly applied. In this chapter, we show that, under certain conditions on A , the observability matrix \mathcal{O}_Ω will satisfy the RIP with high probability, when the total number of measurements MK scales linearly in S and logarithmically in N . This work builds on two recent papers by Yap et al. [86] and Eftekhari et al. [87] which establishes the RIP for block-diagonal matrices via establishing a tail probability bound on the isometry constant. We also derive a CoM bound for \mathcal{O}_Ω . As we demonstrate, the concentration performance of such a matrix depends on properties of both the state transition matrix A and the initial state \mathbf{x}_0 . This work builds on few recent papers in which CoM bounds are derived for random, block-diagonal measurement matrices [50, 84, 85]. Apart from recovery, other inference problems concerning \mathbf{x}_0 (such as detection or classification) can also be solved from the random, compressive measurements, and the performance of such techniques can be studied using the CoM bound that we provide.

6.1.4 Related Work

Questions involving observability in compressive measurement settings have also been raised in a recent paper [91] concerned with tracking the state of a system from nonlinear observations. Due to the intrinsic nature of the problems in that paper, however, the observability issues raised are quite different. For example, one argument appears to assume that $M \geq S$, a requirement that we do not have. In a recent technical report [92], Dai et al. have also considered a similar sparse initial state recovery problem. However, their approach is quite different and the results are only applicable in noiseless and perfectly sparse initial state recovery problems. In this chapter, we establish the RIP for the observability matrix, which implies not only that perfectly sparse initial states can be recovered exactly when the measurements are noiseless but also that the recovery process is robust with respect to noise and that nearly-sparse initial states can be recovered with high accuracy [27]. Finally, we

note that a matrix vaguely similar to the observability matrix has been studied by Yap et al. in the context of quantifying the memory capacity of echo state networks [93].

6.1.5 Chapter Organization

In Section 6.2, we establish the RIP for the observability matrix when the state transition matrix A is a scaled unitary matrix (i.e., $A = aU, \forall a$ and unitary matrix U) and in the case when the measurement matrices C_k are independent of each other, populated with i.i.d. random variables. To this end, in Section 6.2.1, we first show that the expected value of the isometry constant associated with the observability matrix \mathcal{O}_Ω is small. In Section 6.2.2, we then derive a tail probability bound that shows the isometry constant does not deviate from its expected value with high probability. Based on these two steps, we show that \mathcal{O}_Ω satisfies the RIP with high probability if the total number of measurements scales linearly in the sparsity level of the initial state.

In Section 6.3, we derive CoM inequalities for the observability matrix. Assuming the measurement matrices C_k are populated with i.i.d. random variables, we derive a CoM bound for \mathcal{O}_Ω and discuss the implications of the properties of A and \mathbf{x}_0 . Whilst our CoM results are general and cover a broad class of systems and initial states, they have important implications for establishing the RIP in the case of a sparse initial state \mathbf{x}_0 . Such RIP results provide a sufficient number of measurements for exact initial state recovery when the initial state is known to be sparse a priori. We show that under certain conditions on A (e.g., for unitary and certain symmetric matrices A), the observability matrix \mathcal{O}_Ω will satisfy the RIP with high probability when the total number of measurements MK scales linearly in S and logarithmically in N .

Finally, we support our results with a case study involving a diffusion system. As an example of a diffusion process with sparse initial state, one may imagine the sparse contaminants introduced into particular (i.e., sparse) locations in a water supply or in the air. From the available measurements, we would like to find the source of the contamination.

6.2 The RIP and the Observability Matrix

In this section, we establish the RIP for the observability matrix. In all of the results in this section, we assume that all of the measurement matrices $C_k \in \mathbb{R}^{M \times N}$ are generated independently of each other, populated with i.i.d. sub-Gaussian random variables with zero mean, $\frac{1}{M}$ variance, and $\frac{\tau}{\sqrt{M}}$ sub-Gaussian norm. A sub-Gaussian random variable is defined as follows.¹⁰

Definition 6.9 [94, Definition 5.7] *A random variable x is called a sub-Gaussian random variable if it satisfies*

$$(\mathbf{E}[|x|^p])^{\frac{1}{p}} \leq C\sqrt{p}, \quad \text{for all } p \geq 1, \quad (6.10)$$

where $C > 0$ is an absolute constant. The sub-Gaussian norm of x , denoted $\|x\|_{\psi_2}$, is defined as the smallest C in (6.10). Formally,

$$\|x\|_{\psi_2} := \sup_{p \geq 1} p^{-\frac{1}{2}} (\mathbf{E}[|x|^p])^{\frac{1}{p}}.$$

The following theorems state the main results on the RIP for the observability matrix.

Theorem 6.11 *Assume $\Omega = \{0, 1, \dots, K-1\}$. Suppose that $A \in \mathbb{R}^{N \times N}$ can be represented as $A = aU$ where $a \in \mathbb{R}$ and $U \in \mathbb{R}^{N \times N}$ is unitary. Assume $|a| < 1$ and define $b := 1 + a^2 + a^4 + \dots + a^{2(K-1)}$. Let c_1, c_2, c_3, c_4, c_7 , and c_8 be absolute constants such that $c_2^2 c_4^2 c_7 \geq 3$. Define $c_5 := c_2 \sqrt{c_3} c_4$, $c_6^2(a, K) := (1 - a^2)K + a^2$, $c_9(a, K) := 2c_5 c_6(a, K) c_8 (c_1 + c_5 c_6(a, K))$, and $c_{10}(a, K) := c_9(a, K) + 2(1 + \sqrt{5})c_1 c_5 c_6(a, K)$. If for $\nu \in (0, 1)$,*

$$\widetilde{M} \geq \nu^{-2} c_{10}^2(a, K) \tau^2 \log^6(N) S, \quad (6.12)$$

then the matrix $\frac{1}{\sqrt{b}} \mathcal{O}_\Omega$ satisfies the RIP of order S and with the isometry constant $\delta_S \left(\frac{1}{\sqrt{b}} \mathcal{O}_\Omega \right) \leq \nu$, except with a probability of at most N^{-1} .

A similar result to Theorem 6.11 can be achieved when $|a| > 1$.

¹⁰The class of sub-Gaussian random variables contains the standard normal and all bounded (including Bernoulli) random variables.

Theorem 6.13 *Assume the same notation as in Theorem 6.11. Assume $|a| > 1$. If for $\nu \in (0, 1)$,*

$$\widetilde{M} \geq \nu^{-2} c_{10}^2 (a^{-1}, K) \tau^2 \log^6(N) S, \quad (6.14)$$

then the matrix $\frac{1}{\sqrt{b}} \mathcal{O}_\Omega$ satisfies the RIP of order S and with the isometry constant $\delta_S \left(\frac{1}{\sqrt{b}} \mathcal{O}_\Omega \right) \leq \nu$, except with a probability of at most N^{-1} .

When the state transition matrix A is unitary, we have the following results when K arbitrary-chosen samples are taken, i.e., $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$.

Theorem 6.15 *Assume $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$. Suppose that $A \in \mathbb{R}^{N \times N}$ is unitary. Let c_1, c_2, c_3, c_4, c_7 , and c_8 be absolute constants such that $c_2^2 c_4^2 c_7 \geq 3$. Define $c_5 := c_2 \sqrt{c_3} c_4$, $c_9 := 2c_5 c_8 (c_1 + c_5)$ and $c_{10} := c_9 + 2(1 + \sqrt{5})c_1 c_5$. If for $\nu \in (0, 1)$,*

$$\widetilde{M} \geq \nu^{-2} c_{10}^2 \tau^2 \log^6(N) S, \quad (6.16)$$

then the matrix $\frac{1}{\sqrt{b}} \mathcal{O}_\Omega$ satisfies the RIP of order S and with the isometry constant $\delta_S \left(\frac{1}{\sqrt{b}} \mathcal{O}_\Omega \right) \leq \nu$, except with a probability of at most N^{-1} .

These theorems state that under the assumed conditions, $\frac{1}{\sqrt{b}} \mathcal{O}_\Omega$ satisfies the RIP of order S with high probability when the total number of measurements MK scale linearly in the sparsity level S and logarithmically in the state ambient dimension N . Consequently under these assumptions, *unique* recovery of any S -sparse initial state \mathbf{x}_0 is possible from $\mathbf{y}_\Omega = \mathcal{O}_\Omega \mathbf{x}_0$ by solving the ℓ_1 -minimization problem whenever $MK \sim \mathcal{O}(S \log^6(N))$. This is in fact a significant reduction in the sufficient total number of measurement for correct initial state recovery as compared to traditional observability theory. In the rest of this section, we present the proof of Theorem 6.11. Proofs of Theorem 6.13 and Theorem 6.15 involve essentially the same steps as of Theorem 6.11.

We start the analysis by noting that if $\Omega = \{0, 1, \dots, K-1\}$ the observability matrix can be decomposed as

$$\mathcal{O}_\Omega = \sqrt{b} \begin{array}{c} \overbrace{\left[\begin{array}{ccc} C_0 & & \\ & C_1 & \\ & & \ddots \\ & & & C_{K-1} \end{array} \right]}^{\widetilde{\mathcal{O}}_\Omega \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}} \underbrace{\left[\begin{array}{c} \frac{1}{\sqrt{b}} I_N \\ \frac{1}{\sqrt{b}} A \\ \vdots \\ \frac{1}{\sqrt{b}} A^{K-1} \end{array} \right]}_{\mathcal{A}_\Omega \in \mathbb{R}^{\widetilde{N} \times N}} \end{array}, \quad (6.17)$$

where $\widetilde{N} := NK$, $\widetilde{M} := MK$, and $b := 1 + a^2 + a^4 + \dots + a^{2(K-1)}$. In other words, $\frac{1}{\sqrt{b}}\mathcal{O}_\Omega$ can be written as a product of a block-diagonal matrix \mathcal{C}_Ω and a matrix \mathcal{A}_Ω with normalized columns (Observe that $\mathcal{A}_\Omega^T \mathcal{A}_\Omega = I_N$). Thanks to this connection between $\frac{1}{\sqrt{b}}\mathcal{O}_\Omega$ and block-diagonal matrices, we adapt a recent RIP analysis for block-diagonal matrices [86, 87] and establish the RIP for $\widetilde{\mathcal{O}}_\Omega$. To this end, we derive a bound on the isometry constant defined in (6.5) associated with $\widetilde{\mathcal{O}}_\Omega$.

Let Σ_S be the set of all signals $\mathbf{x}_0 \in \mathbb{R}^N$ with $\|\mathbf{x}_0\|_2 \leq 1$ and $\|\mathbf{x}_0\|_0 \leq S$ where $\|\cdot\|_0$ is the ℓ_0 norm and simply counts the non-zero entries of a vector. Observe that $\|\cdot\|_0$ is not a proper norm. Formally,

$$\Sigma_S := \{ \mathbf{x}_0 \in \mathbb{R}^N : \|\mathbf{x}_0\|_0 \leq S \text{ and } \|\mathbf{x}_0\|_2 \leq 1 \}.$$

The isometry constant δ_S can be equivalently written as

$$\begin{aligned} \delta_S := \delta_S(\widetilde{\mathcal{O}}_\Omega) &= \sup_{\|\mathbf{x}_0\|_0 \leq S} \left| \frac{\|\widetilde{\mathcal{O}}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} - 1 \right| \stackrel{2}{=} \sup_{\mathbf{x}_0 \in \Sigma_S} \left| \frac{\|\widetilde{\mathcal{O}}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} - 1 \right| \\ &\stackrel{3}{=} \sup_{\mathbf{x}_0 \in \Sigma_S} \left| \mathbf{x}_0^T \left(\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega - I_N \right) \mathbf{x}_0 \right| =: \|\|\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega - I_N\|\|. \end{aligned} \quad (6.18)$$

It can be easily shown that $\|\|\cdot\|\|$ is a valid norm. In (6.18), the second and third equalities are due to invariance of the ratio $\frac{\|\widetilde{\mathcal{O}}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2}$ to the scaling of \mathbf{x}_0 [33, 86, 87]. Following the steps stated in [33, 86, 87] we first show that $\mathbf{E}[\delta_S]$ is small. We then show that δ_S does not deviate from $\mathbf{E}[\delta_S]$ with high probability.

6.2.1 Bounding $\mathbf{E}[\delta_S]$

In this section, we show that the expected value of δ_S is bounded and small. Observe that if $A = aU$,

$$\mathbf{E} \left[\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega \right] = \mathbf{E} \left[\mathcal{A}_\Omega^T \mathcal{C}_\Omega^T \mathcal{C}_\Omega \mathcal{A}_\Omega \right] = \mathcal{A}_\Omega^T \mathbf{E} \left[\mathcal{C}_\Omega^T \mathcal{C}_\Omega \right] \mathcal{A}_\Omega = \mathcal{A}_\Omega^T I_{\widetilde{N}} \mathcal{A}_\Omega = I_N.$$

Note that $\mathbf{E} \left[\mathcal{C}_\Omega^T \mathcal{C}_\Omega \right] = I_{\widetilde{N}}$ as the entries of \mathcal{C}_Ω are i.i.d. sub-Gaussian random variables with zero mean and $\frac{1}{M}$ variance. Adapting the results of [86] and following similar steps, we have the following Lemma.

Lemma 6.19 *Let c_1 be an absolute constant and let $\widetilde{\mathcal{O}}_\Omega(\cdot) \in \mathbb{R}^{\widetilde{M}N}$ be a vector containing the entries of $\widetilde{\mathcal{O}}_\Omega$. Then*

$$\mathbf{E}[\delta_S] \leq 2c_1 \sqrt{S} \log^2 N \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} \sqrt{\mathbf{E}[\delta_S] + 1}, \quad (6.20)$$

where $\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty = \max_{\substack{\widetilde{m} \in [\widetilde{M}] \\ n \in [N]}} |\widetilde{\mathcal{O}}_\Omega(\widetilde{m}, n)|$. The proof of Lemma 6.19 depends on the following Lemmas.

Lemma 6.21 *Let $\widetilde{\mathcal{O}}_{\widetilde{m}}^T$ ($\widetilde{\mathcal{O}}_{\widetilde{m}} \in \mathbb{R}^N$) be the \widetilde{m} -th row of $\widetilde{\mathcal{O}}_\Omega$. Then $\left\{ \widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T \right\}_{\widetilde{m}=1}^{\widetilde{M}}$ is a sequence of \widetilde{M} independent random (rank-one) matrices in $\mathbb{R}^{N \times N}$. Let $\boldsymbol{\xi} := \{\xi_{\widetilde{m}}\}_{\widetilde{m}=1}^{\widetilde{M}}$ be a Rademacher sequence¹¹ independent of $\left\{ \widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T \right\}_{\widetilde{m}=1}^{\widetilde{M}}$. Then,*

$$\mathbf{E}[\delta_S] \leq 2\mathbf{E}_{\mathcal{C}_\Omega, \boldsymbol{\xi}} \left[\left\| \sum_{\widetilde{m} \in [\widetilde{M}]} \xi_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T \right\| \right]. \quad (6.22)$$

Proof of Lemma 6.21 The proof is based on a symmetrization argument [55, Lemma 6.7].

Since $\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega = \sum_{\widetilde{m} \in [\widetilde{M}]} \widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T$,

¹¹A sequence $\boldsymbol{\xi}$ of independent Rademacher variables is called a Rademacher sequence. A Rademacher variable is a random variable which takes values $+1$ or -1 , each with probability $\frac{1}{2}$ [55].

$$\begin{aligned}
\mathbf{E} [\delta_S] &= \mathbf{E} \left[\left\| \widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega - I_N \right\| \right] \\
&= \mathbf{E} \left[\left\| \sum_{\tilde{m} \in [\widetilde{M}]} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T - I_N \right\| \right] \\
&= \mathbf{E} \left[\left\| \sum_{\tilde{m} \in [\widetilde{M}]} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T - \mathbf{E} \left[\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega \right] \right\| \right] \\
&= \mathbf{E} \left[\left\| \sum_{\tilde{m} \in [\widetilde{M}]} \left(\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T - \mathbf{E} \left[\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right] \right) \right\| \right] \\
&\leq 2\mathbf{E} \left[\left\| \sum_{\tilde{m} \in [\widetilde{M}]} \xi_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right\| \right],
\end{aligned}$$

where $\sum_{\tilde{m} \in [\widetilde{M}]} \xi_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T$ is the associated Rademacher sum¹² of the sequence of independent random matrices $\left\{ \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right\}_{\tilde{m}=1}^{\widetilde{M}}$. \blacksquare

The following Lemma is due to Rudelson and Vershynin [95] which is rephrased by Tropp et al. [96, Lemma 17].

Lemma 6.23 [96, Lemma 17] *Suppose that $\left\{ \widetilde{\mathcal{O}}_{\tilde{m}} \right\}_{\tilde{m}=1}^{\widetilde{M}}$ is a sequence of \widetilde{M} vectors in \mathbb{R}^N where $\widetilde{M} \leq N$, and assume that each vector satisfies the bound $\|\widetilde{\mathcal{O}}_{\tilde{m}}\|_\infty \leq B$. Let $\boldsymbol{\xi} := \{\xi_{\tilde{m}}\}_{\tilde{m}=1}^{\widetilde{M}}$ be a Rademacher sequence independent of $\left\{ \widetilde{\mathcal{O}}_{\tilde{m}} \right\}_{\tilde{m}=1}^{\widetilde{M}}$. Then*

$$\mathbf{E}_{\boldsymbol{\xi}} \left[\left\| \sum_{\tilde{m} \in [\widetilde{M}]} \xi_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right\| \right] \leq \beta \left\| \sum_{\tilde{m} \in [\widetilde{M}]} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right\|^{1/2},$$

where $\beta \leq c_1 B \sqrt{S} \log^2 N$, and c_1 is an absolute constant.

Proof of Lemma 6.19 Using Lemma 6.23,

$$\mathbf{E}_{\boldsymbol{\xi}} \left[\left\| \sum_{\tilde{m} \in [\widetilde{M}]} \xi_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right\| \right] \leq c_1 \sqrt{S} \log^2 N \|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty \|\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega\|^{1/2}.$$

Therefore, using Lemma 6.21,

¹²Let $\boldsymbol{\xi} := \{\xi_{\tilde{m}}\}_{\tilde{m}=1}^{\widetilde{M}}$ be a Rademacher sequence. The associated Rademacher sum is defined as $\sum_{\tilde{m}=1}^{\widetilde{M}} \xi_{\tilde{m}} x_{\tilde{m}}$, where the $\{x_{\tilde{m}}\}_{\tilde{m}=1}^{\widetilde{M}}$ are scalars, vectors, or matrices[55].

$$\begin{aligned}
\mathbf{E} [\delta_S] &\leq 2\mathbf{E}_{\mathcal{C}_\Omega} \mathbf{E}_\xi \left[\left\| \sum_{\tilde{m} \in [\tilde{M}]} \xi_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \right\| \right] \\
&\leq 2c_1 \sqrt{S} \log^2 N \mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty \|\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega\|^{1/2} \right] \\
&\leq 2c_1 \sqrt{S} \log^2 N \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega\| \right]} \\
&\leq 2c_1 \sqrt{S} \log^2 N \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega - I_N\| + \|I_N\| \right]} \\
&= 2c_1 \sqrt{S} \log^2 N \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} \sqrt{\mathbf{E} [\delta_S] + 1},
\end{aligned}$$

where the third inequality is due to the Cauchy-Schwarz inequality, and the fourth inequality is due to the triangle inequality for the norm $\|\cdot\|$. This completes the proof of Lemma 6.19.

■

Lemma 6.19 indicates that bounding $\mathbf{E} [\delta_S]$ can be achieved by bounding $\sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]}$. This bound is achieved in Lemma 6.26. Before stating Lemma 6.26 and its proof, however, we state the following Lemma which is an adaptation of [55, Lemma 6.6] and is rephrased by Eftekhari et al. [87].

Lemma 6.24 *Let c_2 be an absolute constant and let $\|\cdot\|_{\psi_2}$ denote the sub-Gaussian norm of a random variable. Assume the entries of \mathcal{C}_Ω are sub-Gaussian random variables with zero mean, $\frac{1}{M}$ variance, and $\frac{\tau}{\sqrt{M}}$ sub-Gaussian norm. Since the entries of $\widetilde{\mathcal{O}}_\Omega$ are sub-Gaussian random variables,¹³*

$$\sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} = \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\max_{\substack{\tilde{m} \in [\tilde{M}] \\ n \in [N]}} \widetilde{\mathcal{O}}_\Omega^2(\tilde{m}, n) \right]} \leq c_2 \max_{\substack{\tilde{m} \in [\tilde{M}] \\ n \in [N]}} \|\widetilde{\mathcal{O}}_\Omega(\tilde{m}, n)\|_{\psi_2} \sqrt{\log(\tilde{M}N)}. \tag{6.25}$$

Lemma 6.26 *Let c_3 and c_4 be absolute constants and let $c_5 := c_2 \sqrt{c_3} c_4$. Then*

¹³A linear combination of sub-Gaussian random variables is a sub-Gaussian random variable [94, 97].

$$\begin{aligned}\sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} &\leq c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{KM}}, \quad |a| < 1, \\ \sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} &\leq c_5 c_6(a^{-1}, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{KM}}, \quad |a| > 1,\end{aligned}$$

where $c_6^2(a, K) := (1 - a^2)K + a^2, \forall a \in \mathbb{R}$.

Proof of Lemma 6.26 From (6.25), an upper bound on $\sqrt{\mathbf{E}_{\mathcal{C}_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]}$ follows from an upper bound on $\max_{\substack{\tilde{m} \in [\tilde{M}] \\ n \in [N]}} \|\widetilde{\mathcal{O}}_\Omega(\tilde{m}, n)\|_{\psi_2}$. First, observe that for $k \in \{0, 1, \dots, K-1\}$

$$\widetilde{\mathcal{O}}_\Omega(kM + m, n) = \frac{1}{\sqrt{b}} \langle C_k^{m \rightarrow}, A_{n \downarrow}^k \rangle = \frac{1}{\sqrt{b}} \sum_{q \in [N]} C_k^{m \rightarrow}(q) A_{n \downarrow}^k(q),$$

where $C_k^{m \rightarrow}$ is the m -th row of $C_k \in \mathbb{R}^{M \times N}$ and $A_{n \downarrow}^k$ is the n -th column of $A^k \in \mathbb{R}^{N \times N}$.

Therefore,

$$\begin{aligned}\|\widetilde{\mathcal{O}}_\Omega(kM + m, n)\|_{\psi_2}^2 &= \frac{1}{b} \left\| \sum_{q \in [N]} C_k^{m \rightarrow}(q) A_{n \downarrow}^k(q) \right\|_{\psi_2}^2 \\ &\leq \frac{c_3}{b} \sum_{q \in [N]} \|C_k^{m \rightarrow}(q) A_{n \downarrow}^k(q)\|_{\psi_2}^2 \\ &= \frac{c_3}{b} \sum_{q \in [N]} \|C_k^{m \rightarrow}(q)\|_{\psi_2}^2 |A_{n \downarrow}^k(q)|^2 \\ &= \frac{c_3}{b} \left(\frac{\tau}{\sqrt{M}} \right)^2 \sum_{q \in [N]} |A_{n \downarrow}^k(q)|^2 \\ &= \frac{c_3 \tau^2}{bM} \|A_{n \downarrow}^k\|_2^2,\end{aligned}\tag{6.27}$$

where c_3 is an absolute constant and the rotation invariance Lemma [94, Lemma 5.9] implies the inequality. Observe that by assumption $\|C_k^{m \rightarrow}(q)\|_{\psi_2}^2 = \left(\frac{\tau}{\sqrt{M}}\right)^2$ for all k, m , and q .

From (6.25) and (6.27),

$$\begin{aligned}
\sqrt{\mathbf{E}_{c_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} &\leq c_2 \sqrt{c_3} \tau \frac{\sqrt{\log(\widetilde{M}N)}}{\sqrt{b\widetilde{M}}} \max_{\substack{k \in [K] \\ n \in [N]}} \|A_{n\downarrow}^k\|_2 \\
&\leq c_2 \sqrt{c_3} c_4 \tau \frac{\sqrt{\log(N)}}{\sqrt{b\widetilde{M}}} \max_{\substack{k \in [K] \\ n \in [N]}} \|A_{n\downarrow}^k\|_2 \\
&=: c_5 \tau \frac{\sqrt{\log(N)}}{\sqrt{b\widetilde{M}}} \max_{\substack{k \in [K] \\ n \in [N]}} \|A_{n\downarrow}^k\|_2, \tag{6.28}
\end{aligned}$$

where in the second inequality c_4 is an absolute constant and it is assumed that $\widetilde{M} \leq N$. It is trivial to see that if $|a| < 1$ then $\|A_{n\downarrow}^k\|_2 \leq 1$ and if $|a| > 1$ then $\|A_{n\downarrow}^k\|_2 \leq a^{(K-1)}$, for all k, n . Also observe that when $|a| < 1$,¹⁴

$$\frac{1 - a^2}{1 - a^{2K}} \leq (1 - a^2) + \frac{a^2}{K}. \tag{6.29}$$

Similarly, when $|a| > 1$,

$$\frac{1 - a^{-2}}{1 - a^{-2K}} \leq (1 - a^{-2}) + \frac{a^{-2}}{K}. \tag{6.30}$$

Consequently, when $|a| < 1$

$$\frac{1}{b} = \frac{1}{\sum_{k=1}^K a^{2(k-1)}} = \frac{1 - a^2}{1 - a^{2K}} \leq (1 - a^2) + \frac{a^2}{K} = \frac{(1 - a^2)K + a^2}{K}. \tag{6.31}$$

Therefore, from (6.28) and (6.31),

$$\sqrt{\mathbf{E}_{c_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} \leq c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{KM}}, \tag{6.32}$$

where $c_6^2(a, K) := (1 - a^2)K + a^2$. When $|a| > 1$,

$$\frac{a^{2(K-1)}}{b} = \frac{a^{2(K-1)}}{\sum_{k=1}^K a^{2(k-1)}} = \frac{1 - a^{-2}}{1 - a^{-2K}} \leq (1 - a^{-2}) + \frac{a^{-2}}{K} = \frac{(1 - a^{-2})K + a^{-2}}{K} = \frac{c_6^2(a^{-1}, K)}{K}. \tag{6.33}$$

¹⁴In order to prove (6.29), assume that for a given $|a| < 1$, there exists a constant $C(a)$ such that for all K , $\frac{1}{1 - a^{2K}} \leq 1 + \frac{C(a)}{K}$. By this assumption, $C(a) \geq \frac{Ka^{2K}}{1 - a^{2K}} =: g(a, K)$. Observe that for a given $|a| < 1$, $g(a, K)$ is a decreasing function of K (K only takes positive integer values) and its maximum is achieved when $K = 1$. Choosing $C(a) = g(a, 1) = \frac{a^2}{1 - a^2}$ completes the proof of (6.29). A similar approach can be taken to prove (6.30) when $|a| > 1$.

Thus, when $|a| > 1$ from (6.28) and (6.33),

$$\sqrt{\mathbf{E}_{c_\Omega} \left[\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 \right]} \leq c_5 c_6 (a^{-1}, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{KM}}. \quad (6.34)$$

■

We now provide a bound on $\mathbf{E}[\delta_S]$. Consider the case when $|a| < 1$. Then, from (6.20) and (6.32),

$$\mathbf{E}[\delta_S] \leq \frac{2c_1 c_5 c_6 (a, K) \tau \log^{2.5}(N) \sqrt{S}}{\sqrt{\widetilde{M}}} \sqrt{\mathbf{E}[\delta_S] + 1} =: Q \sqrt{\mathbf{E}[\delta_S] + 1}.$$

It is trivial to see that if $Q \leq 1$, then $\mathbf{E}[\delta_S] \leq \frac{1+\sqrt{5}}{2}Q$. Therefore,

$$\mathbf{E}[\delta_S] \leq \left(1 + \sqrt{5}\right) c_1 c_5 c_6 (a, K) \tau \log^{2.5}(N) \sqrt{\frac{S}{\widetilde{M}}} \quad (6.35)$$

if

$$\widetilde{M} \geq (2c_1 c_5 c_6 (a, K))^2 \tau^2 \log^5(N) S. \quad (6.36)$$

Similar results can be obtained if $|a| > 1$ just by replacing $c_6(a, K)$ with $c_6(a^{-1}, K)$ in (6.35) and (6.36). So far we have shown in (6.35) and (6.36) that when a sufficient total number of measurements are taken, $\mathbf{E}[\delta_S]$ is upper bounded by a small value. In the next section we show that when a sufficient total number of measurements are taken, δ_S does not deviate from $\mathbf{E}[\delta_S]$ with high probability.

6.2.2 Tail Bound for δ_S

In this section, we derive a tail probability bound for δ_S . We follow the similar steps as explained by Eftekhari et al. [87]. The following Lemma states a tail probability bound for every sub-Gaussian random variable.

Lemma 6.37 [94, Definition 5.7] *For every sub-Gaussian random variable x and for all $t \geq 0$*

$$\mathbf{P}\{|x| > t\} \leq \exp\left(1 - c_7 \frac{t^2}{\|x\|_{\psi_2}^2}\right),$$

where c_7 is an absolute constant.

Using Lemma 6.37, we derive a tail probability bound for $\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty$. Without loss of generality, we assume $|a| < 1$. Similar steps can be taken when $|a| > 1$.

Lemma 6.38 *Assume $|a| < 1$. Let c_2, c_3, c_4 , and c_7 be absolute constants such that $c_2^2 c_4^2 c_7 \geq 3$. Define $c_5 := c_2 \sqrt{c_3} c_4$ and $c_6^2(a, K) := (1 - a^2) K + a^2$. Then,*

$$\mathbf{P} \left\{ \|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty \geq c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{\widetilde{M}}} \right\} \leq eN^{-1}. \quad (6.39)$$

Proof of Lemma 6.38 Since the entries of $\widetilde{\mathcal{O}}_\Omega$ are sub-Gaussian, we can use Lemma 6.37.

We have

$$\begin{aligned} & \mathbf{P} \left\{ \|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty \geq c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{\widetilde{M}}} \right\} \\ &= \mathbf{P} \left\{ \max_{\substack{\widetilde{m} \in [\widetilde{M}] \\ n \in [N]}} |\widetilde{\mathcal{O}}_\Omega(\widetilde{m}, n)| \geq c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{\widetilde{M}}} \right\} \\ &\leq \widetilde{M} N \max_{\substack{\widetilde{m} \in [\widetilde{M}] \\ n \in [N]}} \mathbf{P} \left\{ |\widetilde{\mathcal{O}}_\Omega(\widetilde{m}, n)| \geq c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{\widetilde{M}}} \right\} \\ &\leq \widetilde{M} N \max_{\substack{\widetilde{m} \in [\widetilde{M}] \\ n \in [N]}} \exp \left(1 - \frac{c_5^2 c_6^2(a, K) c_7 \tau^2 \log(N)}{\widetilde{M} \|\widetilde{\mathcal{O}}_\Omega(\widetilde{m}, n)\|_{\psi_2}^2} \right) \\ &\leq e \widetilde{M} N \exp(-c_2^2 c_4^2 c_7 \log(N)) \\ &\leq e \widetilde{M} N \exp(-3 \log(N)) \leq e \frac{\widetilde{M}}{N^2} \leq eN^{-1}, \end{aligned}$$

where we assumed $c_2^2 c_4^2 c_7 \geq 3$ and $\widetilde{M} \leq N$ in the last inequality. The second inequality is due to Lemma 6.37 and the third inequality is due to (6.27), (6.31), and because $\|\widetilde{\mathcal{O}}_\Omega(\widetilde{m}, n)\|_{\psi_2}^2 \leq \frac{c_3 c_6^2(a, K) \tau^2}{\widetilde{M}}$. \blacksquare

After deriving a tail probability bound for $\|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty$, we now state the proof of Theorem 6.11 which derives a tail probability bound for δ_S . Proofs of Theorem 6.13 and Theorem 6.15 follow similar steps as compared to Theorem 6.11.

Proof of Theorem 6.11 In order to derive a tail bound for δ_S , we use a symmetrization argument [55, Lemma 6.7]. Recall that the isometry constant can be written as $\delta_S := \delta_S(\widetilde{\mathcal{O}}_\Omega) = \|\|\|\widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega - I_N\|\|\|$. Let ϑ be the symmetrized version of δ_S defined as

$$\vartheta := \|\|\| \sum_{\widetilde{m} \in [\widetilde{M}]} \left(\widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T - \widetilde{\mathcal{O}}'_{\widetilde{m}} \widetilde{\mathcal{O}}'_{\widetilde{m}}{}^T \right) \|\|\|,$$

where $\{\widetilde{\mathcal{O}}'_{\widetilde{m}}\}$ is an independent copy of $\{\widetilde{\mathcal{O}}_{\widetilde{m}}\}$ and forms the matrix $\widetilde{\mathcal{O}}'_\Omega$. As explained in [96, Proposition 4], the tail of the symmetrized variable ϑ is closely related to the tail of δ_S . In fact, we have [96, Proposition 4] for $u > 0$

$$\mathbf{P} \{ \delta_S > 2\mathbf{E}[\delta_S] + u \} \leq \mathbf{P} \{ \vartheta > u \}.$$

Also observe that the symmetrized random variable ϑ and its Rademacher series has the same distribution. The Rademacher series of ϑ is defined as

$$\vartheta' := \|\|\| \sum_{\widetilde{m} \in [\widetilde{M}]} \xi_{\widetilde{m}} \left(\widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T - \widetilde{\mathcal{O}}'_{\widetilde{m}} \widetilde{\mathcal{O}}'_{\widetilde{m}}{}^T \right) \|\|\|,$$

where $\{\xi_{\widetilde{m}}\}$ is an independent Rademacher sequence. As ϑ' and ϑ have the same distribution, the tail for ϑ' also bounds the tail for δ_S , so that

$$\mathbf{P} \{ \delta_S > 2\mathbf{E}[\delta_S] + u \} \leq \mathbf{P} \{ \vartheta > u \} = \mathbf{P} \{ \vartheta' > u \}. \quad (6.40)$$

Observe that conditioned on $\widetilde{\mathcal{O}}_\Omega$ and $\widetilde{\mathcal{O}}'_\Omega$, we have

$$\begin{aligned} \mathbf{E}_\xi [\vartheta'] &\leq \mathbf{E}_\xi \left[\|\|\| \sum_{\widetilde{m} \in [\widetilde{M}]} \xi_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}} \widetilde{\mathcal{O}}_{\widetilde{m}}^T \|\|\| \right] + \mathbf{E}_\xi \left[\|\|\| \sum_{\widetilde{m} \in [\widetilde{M}]} \xi_{\widetilde{m}} \widetilde{\mathcal{O}}'_{\widetilde{m}} \widetilde{\mathcal{O}}'_{\widetilde{m}}{}^T \|\|\| \right] \\ &\leq c_1 \sqrt{S} \log^2 N \|\|\| \widetilde{\mathcal{O}}_\Omega(\cdot) \|\|\|_\infty \|\|\| \widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega \|\|\|^{1/2} + c_1 \sqrt{S} \log^2 N \|\|\| \widetilde{\mathcal{O}}'_\Omega(\cdot) \|\|\|_\infty \|\|\| \widetilde{\mathcal{O}}'^T_\Omega \widetilde{\mathcal{O}}'_\Omega \|\|\|^{1/2} \\ &\leq c_1 \sqrt{S} \log^2 N \|\|\| \widetilde{\mathcal{O}}_\Omega(\cdot) \|\|\|_\infty \left(\|\|\| \widetilde{\mathcal{O}}_\Omega^T \widetilde{\mathcal{O}}_\Omega - I_N \|\|\| + 1 \right)^{1/2} \\ &\quad + c_1 \sqrt{S} \log^2 N \|\|\| \widetilde{\mathcal{O}}'_\Omega(\cdot) \|\|\|_\infty \left(\|\|\| \widetilde{\mathcal{O}}'^T_\Omega \widetilde{\mathcal{O}}'_\Omega - I_N \|\|\| + 1 \right)^{1/2} \\ &= c_1 \sqrt{S} \log^2 N \|\|\| \widetilde{\mathcal{O}}_\Omega(\cdot) \|\|\|_\infty \sqrt{\delta_S(\mathcal{C}_\Omega, \mathcal{A}_\Omega) + 1} + c_1 \sqrt{S} \log^2 N \|\|\| \widetilde{\mathcal{O}}'_\Omega(\cdot) \|\|\|_\infty \sqrt{\delta_S(\mathcal{C}'_\Omega, \mathcal{A}_\Omega) + 1}. \end{aligned}$$

Using the triangle inequality for the norm $\|\|\| \cdot \|\|\|$, conditioned on $\widetilde{\mathcal{O}}_\Omega$ and $\widetilde{\mathcal{O}}'_\Omega$, for every \widetilde{m} ,

$$\|\|\xi_{\tilde{m}} \left(\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T - \widetilde{\mathcal{O}}'_{\tilde{m}} \widetilde{\mathcal{O}}'_{\tilde{m}}{}^T \right)\|\| \leq \|\|\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T\|\| + \|\|\widetilde{\mathcal{O}}'_{\tilde{m}} \widetilde{\mathcal{O}}'_{\tilde{m}}{}^T\|\|.$$

Also note that for every \tilde{m} ,

$$\|\|\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T\|\| = \sup_{\mathbf{x}_0 \in \Sigma_S} \left| \mathbf{x}_0^T \widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T \mathbf{x}_0 \right| = \sup_{\mathbf{x}_0 \in \Sigma_S} \left| \langle \widetilde{\mathcal{O}}_{\tilde{m}}, \mathbf{x}_0 \rangle \right|^2 \leq S \|\widetilde{\mathcal{O}}_{\tilde{m}}\|_\infty^2 \leq S \|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2.$$

Similarly, $\|\|\widetilde{\mathcal{O}}'_{\tilde{m}} \widetilde{\mathcal{O}}'_{\tilde{m}}{}^T\|\| \leq S \|\widetilde{\mathcal{O}}'_\Omega(\cdot)\|_\infty^2$. Consequently, we get

$$\|\|\xi_{\tilde{m}} \left(\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T - \widetilde{\mathcal{O}}'_{\tilde{m}} \widetilde{\mathcal{O}}'_{\tilde{m}}{}^T \right)\|\| \leq S \|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty^2 + S \|\widetilde{\mathcal{O}}'_\Omega(\cdot)\|_\infty^2.$$

Define the following events.

$$\begin{aligned} \mathcal{E}_1 &:= \left\{ \|\widetilde{\mathcal{O}}_\Omega(\cdot)\|_\infty < c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{M}} \right\}, \text{ and} \\ \mathcal{E}_2 &:= \left\{ \|\widetilde{\mathcal{O}}'_\Omega(\cdot)\|_\infty < c_5 c_6(a, K) \tau \frac{\sqrt{\log(N)}}{\sqrt{M}} \right\}. \end{aligned}$$

Note that Lemma 6.38 implies $\mathbf{P}\{\mathcal{E}_1^c\} \leq eN^{-1}$ and $\mathbf{P}\{\mathcal{E}_2^c\} \leq eN^{-1}$. Also define

$$\begin{aligned} \mathcal{E}_3(\alpha) &:= \{\delta_S(\mathcal{C}_\Omega \mathcal{A}_\Omega) \leq \alpha\}, \text{ and} \\ \mathcal{E}_4(\alpha) &:= \{\delta_S(\mathcal{C}'_\Omega \mathcal{A}_\Omega) \leq \alpha\}, \end{aligned}$$

for $\alpha \geq 0$. Conditioned on $\widetilde{\mathcal{O}}_\Omega$, $\widetilde{\mathcal{O}}'_\Omega$, and events \mathcal{E}_1 , \mathcal{E}_2 , $\mathcal{E}_3(\alpha)$, and $\mathcal{E}_4(\alpha)$, we have

$$\begin{aligned} \mathbf{E}_\xi[\vartheta'] &\leq c_1 c_5 c_6(a, K) \tau \log^{2.5} N \sqrt{\frac{S}{M}} \sqrt{\alpha + 1} + c_1 c_5 c_6(a, K) \tau \log^{2.5} N \sqrt{\frac{S}{M}} \sqrt{\alpha + 1} \\ &= 2c_1 c_5 c_6(a, K) \tau \log^{2.5} N \sqrt{\frac{S}{M}} \sqrt{\alpha + 1}, \end{aligned}$$

for $\alpha \geq 0$. The following proposition is useful in deriving a tail probability bound for ϑ' .

Proposition 6.41 [96, Proposition 19] *Let Y_1, \dots, Y_R be independent, symmetric random variables in a Banach space X , and assume each random variable satisfies $\|Y_r\|_X \leq B$ almost surely. Let $Y = \|\sum_r Y_r\|_X$. Let c_8 be an absolute constant. Then*

$$\mathbf{P} \{Y > c_8 [u\mathbf{E}[Y] + tB]\} \leq \exp(-u^2) + \exp(-t)$$

for all $u, t \geq 1$.

We apply Proposition 6.41 to $\vartheta' = \left\| \sum_{\tilde{m} \in [\tilde{M}]} \xi_{\tilde{m}} \left(\widetilde{\mathcal{O}}_{\tilde{m}} \widetilde{\mathcal{O}}_{\tilde{m}}^T - \widetilde{\mathcal{O}}_{\tilde{m}}' \widetilde{\mathcal{O}}_{\tilde{m}}'^T \right) \right\|$ with $t = \log(N)$ and $u = \log^{0.5}(N)$. Conditioned on $\widetilde{\mathcal{O}}_{\Omega}, \widetilde{\mathcal{O}}_{\Omega}'$, and events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3(\alpha)$, and $\mathcal{E}_4(\alpha)$, we have

$$\begin{aligned} & \mathbf{P} \left\{ \vartheta' > c_9(a, K) \tau \log^3(N) \sqrt{\alpha + 1} \sqrt{\frac{S}{\widetilde{M}}} \middle| \widetilde{\mathcal{O}}_{\Omega}, \widetilde{\mathcal{O}}_{\Omega}' \right\} \\ & \leq \mathbf{P} \left\{ \vartheta' > c_8 \left[\log^{0.5}(N) \mathbf{E}_{\xi}[\vartheta'] + \log(N) \left(S \|\widetilde{\mathcal{O}}_{\Omega}(\cdot)\|_{\infty}^2 + S \|\widetilde{\mathcal{O}}_{\Omega}'(\cdot)\|_{\infty}^2 \right) \right] \middle| \widetilde{\mathcal{O}}_{\Omega}, \widetilde{\mathcal{O}}_{\Omega}' \right\} \\ & \leq 2N^{-1}, \end{aligned} \tag{6.42}$$

where $c_9(a, K) := 2c_5c_6(a, K)c_8(c_1 + c_5c_6(a, K))$ and it is assumed that $\tau \sqrt{\frac{S}{\widetilde{M}}} \leq 1$. The rest of the proof follows from [87] where first the conditioning in (6.42) is removed and then a tail probability bound for δ_S is achieved via (6.40) and the bound on $\mathbf{E}[\delta_S]$ given in (6.35).

Consequently, we show that

$$\mathbf{P} \left\{ \delta_S \geq c_{10}(a, K) \tau \log^3(N) \sqrt{\frac{S}{\widetilde{M}}} \right\} \leq c_{11}N^{-1},$$

where $c_{10}(a, K) := c_9(a, K) + 2(1 + \sqrt{5})c_1c_5c_6(a, K)$ and c_{11} is an absolute constant. Consequently, one can conclude that for $\nu \in (0, 1)$,

$$\mathbf{P} \{\delta_S \geq \nu\} \leq c_{11}N^{-1}$$

if

$$\widetilde{M} \geq \nu^{-2}c_{10}^2(a, K) \tau^2 \log^6(N) S.$$

This complete the proof of Theorem 6.11. ■

Remark 6.43 *One should note that when $A = aU$ ($a \neq 1$), the results have a dependency on K (the number of sampling times). This dependency is not desired in general. With simple modifications of the proofs, one can show that when $a = 1$ (i.e., A is unitary), a result can be obtained in which the total number of measurements \widetilde{M} scales linearly in S and with no*

dependency on K . One way to see this is by observing that $c_6(a, K) = 1$ when $a = 1$. Our general results for $A = aU$ also indicate that when $|a|$ is close to the origin (i.e., $|a| \ll 1$), and by symmetry when $|a| \gg 1$, worse recovery performance is expected as compared to the case when $a = 1$. When $|a| \ll 1$, as an example, the effect of the initial state will be highly attenuated as we take measurements at later times. A similar intuition can be made when $|a| \gg 1$. When A is unitary (i.e., $a = 1$), we can further relax the K consecutive sample times requirement and instead take K arbitrary-chosen samples, i.e., $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$. The proof when A is unitary follows the really similar steps as explained.

6.3 CoM Inequalities and the Observability Matrix

In this section, we derive CoM inequalities for the observability matrix when there is no restrictions on the state transition matrix A and the initial state \mathbf{x}_0 . Whilst our CoM results are general and cover a broad class of systems and initial states, they have implications in the case of unitary A and sparse initial state \mathbf{x}_0 as will be discussed in this section. Our derivation is based on the CoM results for block-diagonal matrices. As in Section 6.2, we exploit the fact that the observability matrix can be decomposed as a product of two matrices, one of which has a block-diagonal structure, where diagonals of this matrix are the measurement matrices C_k . We derive CoM inequalities for two cases. We first consider the case where all matrices C_k are generated independently of each other. We then consider the case where all matrices C_k are the same. In either case, we assume that each matrix C_k is populated with i.i.d. Gaussian random variables having zero mean and $\sigma^2 = \frac{1}{M}$ variance. To begin, note that we can write

$$\mathcal{O}_\Omega = \sqrt{K} \begin{array}{c} \overbrace{\left[\begin{array}{cccc} C_{k_0} & & & \\ & C_{k_1} & & \\ & & \ddots & \\ & & & C_{k_{K-1}} \end{array} \right]}^{\widetilde{\mathcal{O}}_\Omega \in \mathbb{R}^{\widetilde{M} \times N}} \underbrace{\left[\begin{array}{c} \frac{1}{\sqrt{K}} A^{k_0} \\ \frac{1}{\sqrt{K}} A^{k_1} \\ \vdots \\ \frac{1}{\sqrt{K}} A^{k_{K-1}} \end{array} \right]}_{\mathcal{A}_\Omega \in \mathbb{R}^{\widetilde{N} \times N}} \end{array}, \quad (6.44)$$

$\underbrace{\hspace{15em}}_{C_\Omega \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}}$

where $\tilde{N} := NK$ and $\tilde{M} := MK$. Note that the matrix \mathcal{A}_Ω is not defined the same as in (6.17) as in this section we allow arbitrary observation times in our analysis.

6.3.1 Independent Random Measurement Matrices

In this section, we assume all matrices C_k are generated independently of each other. The matrix \mathcal{C}_Ω in (6.44) is block diagonal, and focusing just on this matrix for the moment, we have the following bound on its concentration behavior.¹⁵

Theorem 6.45 [84] *Let $\mathbf{v}_{k_0}, \mathbf{v}_{k_1}, \dots, \mathbf{v}_{k_{K-1}} \in \mathbb{R}^N$ and define*

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_{k_0}^T & \mathbf{v}_{k_1}^T & \cdots & \mathbf{v}_{k_{K-1}}^T \end{bmatrix}^T \in \mathbb{R}^{KN}.$$

Then

$$\mathbf{P} \left\{ \left| \|\mathcal{C}_\Omega \mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2 \right| > \epsilon \|\mathbf{v}\|_2^2 \right\} \leq \begin{cases} 2 \exp\left\{-\frac{M\epsilon^2 \|\boldsymbol{\gamma}\|_1^2}{256 \|\boldsymbol{\gamma}\|_2^2}\right\}, & 0 \leq \epsilon \leq \frac{16 \|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_\infty \|\boldsymbol{\gamma}\|_1} \\ 2 \exp\left\{-\frac{M\epsilon \|\boldsymbol{\gamma}\|_1}{16 \|\boldsymbol{\gamma}\|_\infty}\right\}, & \epsilon \geq \frac{16 \|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_\infty \|\boldsymbol{\gamma}\|_1}, \end{cases}$$

where

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}(\mathbf{v}) := \begin{bmatrix} \|\mathbf{v}_{k_0}\|_2^2 \\ \|\mathbf{v}_{k_1}\|_2^2 \\ \vdots \\ \|\mathbf{v}_{k_{K-1}}\|_2^2 \end{bmatrix} \in \mathbb{R}^K.$$

As we will be frequently concerned with applications where ϵ is small, let us consider the first of the cases given in the right-hand side of the above bound. (It can be shown [84] that this case always permits any value of ϵ between 0 and $\frac{16}{\sqrt{K}}$.) We define

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\mathbf{v}) := \frac{\|\boldsymbol{\gamma}(\mathbf{v})\|_1^2}{\|\boldsymbol{\gamma}(\mathbf{v})\|_2^2} = \frac{(\|\mathbf{v}_{k_0}\|_2^2 + \|\mathbf{v}_{k_1}\|_2^2 + \cdots + \|\mathbf{v}_{k_{K-1}}\|_2^2)^2}{\|\mathbf{v}_{k_0}\|_2^4 + \|\mathbf{v}_{k_1}\|_2^4 + \cdots + \|\mathbf{v}_{k_{K-1}}\|_2^4} \quad (6.46)$$

¹⁵All results in Section 6.3.1 may be extended to the case where the matrices C_k are populated with sub-Gaussian random variables, as in [84].

and note that for any $\mathbf{v} \in \mathbb{R}^{KN}$, $1 \leq \Gamma(\mathbf{v}) \leq K$. This simply follows from the standard relation that $\|\mathbf{z}\|_2 \leq \|\mathbf{z}\|_1 \leq \sqrt{K}\|\mathbf{z}\|_2$ for all $\mathbf{z} \in \mathbb{R}^K$. The case $\Gamma(\mathbf{v}) = K$ is quite favorable because the failure probability will decay exponentially fast in the total number of measurements MK . In this case, we get the same degree of concentration from the $MK \times NK$ block-diagonal matrix \mathcal{C}_Ω as we would get from a *dense* $MK \times NK$ matrix populated with i.i.d. Gaussian random variables. A CoM result for a dense Gaussian matrix is stated in Definition 6.6. This event happens if and only if the components \mathbf{v}_{k_i} have equal energy, i.e., if and only if

$$\|\mathbf{v}_{k_0}\|_2 = \|\mathbf{v}_{k_1}\|_2 = \cdots = \|\mathbf{v}_{k_{K-1}}\|_2.$$

On the other hand, the case $\Gamma(\mathbf{v}) = 1$ is quite unfavorable and implies that we get the same degree of concentration from the $MK \times NK$ block-diagonal matrix \mathcal{C}_Ω as we would get from a dense Gaussian matrix having size only $M \times NK$. This event happens if and only if $\|\mathbf{v}_{k_i}\|_2 = 0$ for all $i \in \{0, 1, \dots, K-1\}$ but one i . Thus, more uniformity in the values of the $\|\mathbf{v}_{k_i}\|_2$ ensures a higher probability of concentration.

We now note that, when applying the observability matrix to an initial state, we will have

$$\mathcal{O}_\Omega \mathbf{x}_0 = \sqrt{K} \mathcal{C}_\Omega \mathcal{A}_\Omega \mathbf{x}_0.$$

This leads us to the following corollary of Theorem 6.45.

Corollary 6.47 *Fix any state $\mathbf{x}_0 \in \mathbb{R}^N$. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,*

$$\mathbf{P} \left\{ \left| \|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2 - K \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \right| > \epsilon K \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \right\} \leq 2 \exp \left\{ -\frac{M\Gamma(\mathcal{A}_\Omega \mathbf{x}_0) \epsilon^2}{256} \right\}. \quad (6.48)$$

There are two important phenomena to consider in this result, and both are impacted by the interaction of A with \mathbf{x}_0 . First, on the left-hand side of (6.48), we see that the point

of concentration of $\|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2$ is around $K\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2$, where

$$K\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 = \|A^{k_0} \mathbf{x}_0\|_2^2 + \|A^{k_1} \mathbf{x}_0\|_2^2 + \cdots + \|A^{k_{K-1}} \mathbf{x}_0\|_2^2. \quad (6.49)$$

For a concentration bound of the same form as Definition 6.6, however, we might like to ensure that $\|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2$ concentrates around some constant multiple of $\|\mathbf{x}_0\|_2^2$. In general, for different initial states \mathbf{x}_0 and transition matrices A , we may see widely varying ratios $K \frac{\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2}$. However, in below we discuss one scenario where this ratio is predictable and fixed. Second, on the right-hand side of (6.48), we see that the exponent of the concentration failure probability scales with

$$\mathbf{\Gamma}(\mathcal{A}_\Omega \mathbf{x}_0) = \frac{(\|A^{k_0} \mathbf{x}_0\|_2^2 + \|A^{k_1} \mathbf{x}_0\|_2^2 + \cdots + \|A^{k_{K-1}} \mathbf{x}_0\|_2^2)^2}{\|A^{k_0} \mathbf{x}_0\|_2^4 + \|A^{k_1} \mathbf{x}_0\|_2^4 + \cdots + \|A^{k_{K-1}} \mathbf{x}_0\|_2^4}. \quad (6.50)$$

As mentioned earlier, $1 \leq \mathbf{\Gamma}(\mathcal{A}_\Omega \mathbf{x}_0) \leq K$. The case $\mathbf{\Gamma}(\mathcal{A}_\Omega \mathbf{x}_0) = K$ is quite favorable and happens when $\|A^{k_0} \mathbf{x}_0\|_2 = \|A^{k_1} \mathbf{x}_0\|_2 = \cdots = \|A^{k_{K-1}} \mathbf{x}_0\|_2$; in Section 6.3.2, we discuss one scenario where this is guaranteed to occur. The case $\mathbf{\Gamma}(\mathcal{A}_\Omega \mathbf{x}_0) = 1$ is quite unfavorable and happens if and only if $\|A^{k_i} \mathbf{x}_0\|_2^2 = 0$ for all $i \in \{0, 1, 2, \dots, K-1\}$ except for one i . This happens when $k_0 = 0$ and $\mathbf{x}_0 \in \text{null}(A)$ for $\mathbf{x}_0 \neq 0$.

6.3.2 Unitary and Symmetric System Matrices

In the special case where A is unitary (i.e., $\|A^{k_i} \mathbf{u}\|_2^2 = \|\mathbf{u}\|_2^2$ for all $\mathbf{u} \in \mathbb{R}^N$ and for any power k_i), we can draw a particularly strong conclusion. Because a unitary A guarantees both that $K\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 = \|\mathbf{x}_0\|_2^2$ and that $\mathbf{\Gamma}(\mathcal{A}_\Omega \mathbf{x}_0) = K$, we have the following result.

Corollary 6.51 *Fix any state $\mathbf{x}_0 \in \mathbb{R}^N$ and suppose that A is a unitary operator. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,*

$$\mathbf{P} \left\{ \left| \frac{1}{\sqrt{K}} \|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2 - \|\mathbf{x}_0\|_2^2 \right| > \epsilon \|\mathbf{x}_0\|_2^2 \right\} \leq 2 \exp \left\{ -\frac{MK\epsilon^2}{256} \right\}. \quad (6.52)$$

What this means is that we get the same degree of concentration from the $MK \times N$ observability matrix \mathcal{O}_Ω as we would get from a fully random dense $KM \times N$ matrix popu-

lated with i.i.d. Gaussian random variables. Observe that this concentration result is valid for any $\mathbf{x} \in \mathbb{R}^N$ (not necessarily sparse) and can be used, for example, to prove that finite point clouds [98] and low-dimensional manifolds [99] in \mathbb{R}^N can have stable, approximate distance-preserving embeddings under the matrix \mathcal{O}_Ω . In each of these cases we may be able to solve very powerful signal inference and recovery problems with $MK \ll N$.

We further extend our analysis and establish the RIP for certain symmetric matrices A . We believe this analysis has important consequences in analyzing problems of practical interest such as diffusion (see, for example, Section 6.4). Motivated by such an application, in particular, suppose that $A \in \mathbb{R}^{N \times N}$ is a positive semidefinite matrix with the eigendecomposition

$$A = U\Lambda U^T = [U_1|U_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} [U_1|U_2]^T, \quad (6.53)$$

where $U \in \mathbb{R}^{N \times N}$ is unitary, $\Lambda \in \mathbb{R}^{N \times N}$ is a diagonal matrix with non-negative entries, $U_1 \in \mathbb{R}^{N \times L}$, $U_2 \in \mathbb{R}^{N \times (N-L)}$, $\Lambda_1 \in \mathbb{R}^{L \times L}$, and $\Lambda_2 \in \mathbb{R}^{(N-L) \times (N-L)}$. The submatrix Λ_1 contains the L largest eigenvalues of A . The value for L can be chosen as desired; our results below give the strongest bounds when all eigenvalues in Λ_1 are large compared to all eigenvalues in Λ_2 . Let $\lambda_{1,\min}$ denote the smallest entry of Λ_1 , $\lambda_{1,\max}$ denote the largest entry of Λ_1 , and $\lambda_{2,\max}$ denote the largest entry of Λ_2 .

In the following, we show that in the special case where the matrix $U_1^T \in \mathbb{R}^{L \times N}$ ($L < N$) happens to itself satisfy the RIP (up to a scaling), then \mathcal{O}_Ω satisfies the RIP (up to a scaling). Although there are many state transition matrices A that do not have a collection of eigenvectors U_1 with this special property, we do note that if A is a circulant matrix, its eigenvectors will be the Discrete Fourier Transform (DFT) basis vectors, and it is known that a randomly selected set of DFT basis vectors will satisfy the RIP with high probability [100]. Other cases where U_1 could be modeled as being randomly generated could also fit into this scenario, though such cases may primarily be of academic interest.

Theorem 6.54 Assume $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$. Assume A has the eigendecomposition given in (6.53) and $U_1^T \in \mathbb{R}^{L \times N}$ ($L < N$) satisfies a scaled version¹⁶ of the RIP of order S with isometry constant δ_S . Formally, assume for $\delta_S \in (0, 1)$ that

$$(1 - \delta_S) \frac{L}{N} \|\mathbf{x}_0\|_2^2 \leq \|U_1^T \mathbf{x}_0\|_2^2 \leq (1 + \delta_S) \frac{L}{N} \|\mathbf{x}_0\|_2^2 \quad (6.55)$$

holds for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. Assume each of the measurement matrices $C_k \in \mathbb{R}^{M \times N}$ is populated with i.i.d. Gaussian random entries with mean zero and variance $\frac{1}{M}$. Assume all matrices C_k are generated independently of each other. Let $\nu \in (0, 1)$ denote a failure probability and $\delta \in (0, \frac{16}{\sqrt{K}})$ denote a distortion factor. Then with probability exceeding $1 - \nu$,

$$(1 - \delta) \left((1 - \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1, \min}^{2k_i} \right) \leq \frac{\|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} \leq (1 + \delta) \left((1 + \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1, \max}^{2k_i} + \sum_{i=0}^{K-1} \lambda_{2, \max}^{2k_i} \right) \quad (6.56)$$

for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$ whenever

$$MK \geq \frac{512K \left(S \left(\log\left(\frac{42}{\delta}\right) + 1 + \log\left(\frac{N}{S}\right) \right) + \log\left(\frac{2}{\nu}\right) \right)}{\rho \delta^2}, \quad (6.57)$$

where

$$\rho := \inf_{S\text{-sparse } \mathbf{x}_0 \in \mathbb{R}^N} \Gamma(\mathcal{A}_\Omega \mathbf{x}_0) \quad (6.58)$$

and

$$\Gamma(\mathcal{A}_\Omega \mathbf{x}_0) := \frac{\left(\|A^{k_0} \mathbf{x}_0\|_2^2 + \|A^{k_1} \mathbf{x}_0\|_2^2 + \dots + \|A^{k_{K-1}} \mathbf{x}_0\|_2^2 \right)^2}{\|A^{k_0} \mathbf{x}_0\|_2^4 + \|A^{k_1} \mathbf{x}_0\|_2^4 + \dots + \|A^{k_{K-1}} \mathbf{x}_0\|_2^4}.$$

Proof of Theorem 6.54 We start the analysis by showing that $\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2$ lies within a small neighborhood around $\|\mathbf{x}_0\|_2^2$ for any S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. To this end, we derive the following lemma.

Lemma 6.59 Assume $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$. Assume A has the eigendecomposition given in (6.53) and $U_1^T \in \mathbb{R}^{L \times N}$ ($L < N$) satisfies a scaled version of the RIP of order S with

¹⁶The $\frac{L}{N}$ scaling in (6.55) is to account for the unit-norm rows of U_1^T .

isometry constant δ_S as given in (6.55). Then, for $\delta_S \in (0, 1)$,

$$(1 - \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1,\min}^{2k_i} \leq \frac{\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} \leq (1 + \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1,\max}^{2k_i} + \sum_{i=0}^{K-1} \lambda_{2,\max}^{2k_i} \quad (6.60)$$

holds for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$.

Proof of Lemma 6.59 If A is of the form given in (6.53), we have $A\mathbf{x}_0 = U_1 \Lambda_1 U_1^T \mathbf{x}_0 + U_2 \Lambda_2 U_2^T \mathbf{x}_0$, and consequently,

$$\|A\mathbf{x}_0\|_2^2 = \mathbf{x}_0^T U_1 \Lambda_1^2 U_1^T \mathbf{x}_0 + \mathbf{x}_0^T U_2 \Lambda_2^2 U_2^T \mathbf{x}_0 \geq \|\Lambda_1 U_1^T \mathbf{x}_0\|_2^2 \geq \lambda_{1,\min}^2 \|U_1^T \mathbf{x}_0\|_2^2.$$

On the other hand,

$$\begin{aligned} \|A\mathbf{x}_0\|_2^2 &= \mathbf{x}_0^T U_1 \Lambda_1^2 U_1^T \mathbf{x}_0 + \mathbf{x}_0^T U_2 \Lambda_2^2 U_2^T \mathbf{x}_0 \leq \lambda_{1,\max}^2 \|U_1^T \mathbf{x}_0\|_2^2 + \lambda_{2,\max}^2 \|U_2^T \mathbf{x}_0\|_2^2 \\ &\leq \lambda_{1,\max}^2 \|U_1^T \mathbf{x}_0\|_2^2 + \lambda_{2,\max}^2 \|\mathbf{x}_0\|_2^2. \end{aligned}$$

Thus,

$$\lambda_{1,\min}^2 \|U_1^T \mathbf{x}_0\|_2^2 \leq \|A\mathbf{x}_0\|_2^2 \leq \lambda_{1,\max}^2 \|U_1^T \mathbf{x}_0\|_2^2 + \lambda_{2,\max}^2 \|\mathbf{x}_0\|_2^2. \quad (6.61)$$

If U_1^T satisfies the scaled RIP, then from (6.55) and (6.61) for $\delta_S \in (0, 1)$,

$$(1 - \delta_S) \frac{L}{N} \lambda_{1,\min}^2 \leq \frac{\|A\mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} \leq (1 + \delta_S) \frac{L}{N} \lambda_{1,\max}^2 + \lambda_{2,\max}^2 \quad (6.62)$$

holds for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. Similarly, one can show that for $i \in \{0, 1, \dots, K-1\}$,

$$\lambda_{1,\min}^{2k_i} \|U_1^T \mathbf{x}_0\|_2^2 \leq \|A^{k_i} \mathbf{x}_0\|_2^2 \leq \lambda_{1,\max}^{2k_i} \|U_1^T \mathbf{x}_0\|_2^2 + \lambda_{2,\max}^{2k_i} \|\mathbf{x}_0\|_2^2,$$

and consequently, for $\delta_S \in (0, 1)$,

$$(1 - \delta_S) \frac{L}{N} \lambda_{1,\min}^{2k_i} \leq \frac{\|A^{k_i} \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} \leq (1 + \delta_S) \frac{L}{N} \lambda_{1,\max}^{2k_i} + \lambda_{2,\max}^{2k_i} \quad (6.63)$$

holds for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. Consequently using (6.49), for $\delta_S \in (0, 1)$,

$$(1 - \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1,\min}^{2k_i} \leq \frac{\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} \leq (1 + \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1,\max}^{2k_i} + \sum_{i=0}^{K-1} \lambda_{2,\max}^{2k_i}$$

holds for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. ■

Lemma 6.59 provides deterministic bounds on the ratio $\frac{\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2}$ for all S -sparse \mathbf{x}_0 when U_1^T satisfies the scaled RIP. Using this deterministic result, we can now state the proof of Theorem 6.54 where we show that a scaled version of \mathcal{C}_Ω satisfies the RIP with high probability. First observe that when all matrices C_k are independent and populated with i.i.d. Gaussian random entries, from Corollary 6.47 we have the following CoM inequality for \mathcal{C}_Ω . For any fixed S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$, let $\mathbf{v} = \mathcal{A}_\Omega \mathbf{x}_0 \in \mathbb{R}^{NK}$. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,

$$\mathbf{P} \left\{ \left| \|\mathcal{C}_\Omega \mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2 \right| > \epsilon \|\mathbf{v}\|_2^2 \right\} \leq 2 \exp \left\{ -\frac{M\Gamma(\mathbf{v})\epsilon^2}{256} \right\}. \quad (6.64)$$

As can be seen, the right-hand side of (6.64) is signal dependent. However, we need a universal failure probability bound (that is independent of \mathbf{x}_0) in order to prove the RIP based a CoM inequality. Define

$$\rho := \inf_{S\text{-sparse } \mathbf{x}_0 \in \mathbb{R}^N} \Gamma(\mathcal{A}_\Omega \mathbf{x}_0). \quad (6.65)$$

Therefore from (6.64) and (6.65), for any fixed S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$ and for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,

$$\mathbf{P} \left\{ \left| \|\mathcal{C}_\Omega \mathcal{A}_\Omega \mathbf{x}_0\|_2^2 - \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \right| > \epsilon \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \right\} \leq 2 \exp \left\{ -\frac{M\rho\epsilon^2}{256} \right\} = 2 \exp \left\{ -\widetilde{M}f(\epsilon) \right\}, \quad (6.66)$$

where $f(\epsilon) := \frac{\rho\epsilon^2}{256K}$, $\widetilde{M} := MK$, and $\widetilde{N} := NK$. Let $\nu \in (0, 1)$ denote a failure probability and $\delta \in (0, \frac{16}{\sqrt{K}})$ denote a distortion factor. Through a union bound argument and by applying Lemma 6.8 for all $\binom{N}{S}$ S -dimensional subspaces in \mathbb{R}^N , whenever $\mathcal{C}_\Omega \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}$ satisfies the CoM inequality (6.66) with

$$MK \geq \frac{S \log(\frac{42}{\delta}) + \log(\frac{2}{\nu}) + \log(\binom{N}{S})}{f(\frac{\delta}{\sqrt{2}})}, \quad (6.67)$$

then with probability exceeding $1 - \nu$,

$$(1 - \delta) \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \leq \|\mathcal{C}_\Omega \mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \leq (1 + \delta) \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2,$$

for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. Consequently using the deterministic bound on $\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2$ derived in (6.60), with probability exceeding $1 - \nu$,

$$(1 - \delta) \left((1 - \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1,\min}^{2k_i} \right) \leq \frac{\|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2} \leq (1 + \delta) \left((1 + \delta_S) \frac{L}{N} \sum_{i=0}^{K-1} \lambda_{1,\max}^{2k_i} + \sum_{i=0}^{K-1} \lambda_{2,\max}^{2k_i} \right)$$

for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$. ■

The result of Theorem 6.54 is particularly interesting in applications where the largest eigenvalues of A all cluster around each other and the rest of the eigenvalues cluster around zero. Put formally, we are interested in applications where

$$0 \approx \lambda_{2,\max} \ll \frac{\lambda_{1,\min}}{\lambda_{1,\max}} \approx 1.$$

The following corollary of Theorem 6.54 considers an extreme case when $\lambda_{1,\max} = \lambda_{1,\min}$ and $\lambda_{2,\max} = 0$.

Corollary 6.68 *Assume $\Omega = \{k_0, k_1, \dots, k_{K-1}\}$. Assume each of the measurement matrices $C_k \in \mathbb{R}^{M \times N}$ is populated with i.i.d. Gaussian random entries with mean zero and variance $\frac{1}{M}$. Assume all matrices C_k are generated independently of each other. Suppose A has the eigendecomposition given in (6.53) and $U_1^T \in \mathbb{R}^{L \times N}$ ($L < N$) satisfies a scaled version of the RIP of order S with isometry constant δ_S as given in (6.55). Assume $\lambda_{1,\max} = \lambda_{1,\min} = \lambda$ ($\lambda \neq 0$) and $\lambda_{2,\max} = 0$. Let $\nu \in (0, 1)$ denote a failure probability and $\delta \in (0, 1)$ denote a distortion factor. Define $C := \sum_{i=0}^{K-1} \lambda^{2k_i}$ and $\delta'_S := \delta_S + \delta + \delta_S \delta$. Then with probability exceeding $1 - \nu$,*

$$(1 - \delta'_S) \|\mathbf{x}_0\|_2^2 \leq \left\| \sqrt{\frac{N}{LC}} \mathcal{O}_\Omega \mathbf{x}_0 \right\|_2^2 \leq (1 + \delta'_S) \|\mathbf{x}_0\|_2^2 \quad (6.69)$$

for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$ whenever

$$MK \geq \begin{cases} \frac{512(1 + \delta_S)^2 \lambda^{-4(k_{K-1} - k_0)} \left(S \left(\log\left(\frac{42}{\delta}\right) + 1 + \log\left(\frac{N}{S}\right) \right) + \log\left(\frac{2}{\nu}\right) \right)}{(1 - \delta_S)^2 \delta^2}, & \lambda < 1 \\ \frac{512(1 + \delta_S)^2 \lambda^{4(k_{K-1} - k_0)} \left(S \left(\log\left(\frac{42}{\delta}\right) + 1 + \log\left(\frac{N}{S}\right) \right) + \log\left(\frac{2}{\nu}\right) \right)}{(1 - \delta_S)^2 \delta^2}, & \lambda > 1. \end{cases}$$

While the result of Corollary 6.68 is generic and valid for any λ , an important RIP result can be obtained when $\lambda = 1$. The following corollary states the result.

Corollary 6.70 *Suppose the same notation and assumptions as in Corollary 6.68 and additionally assume $\lambda = 1$. Then with probability exceeding $1 - \nu$,*

$$(1 - \delta'_S)\|\mathbf{x}_0\|_2^2 \leq \left\| \sqrt{\frac{N}{LK}} \mathcal{O}_\Omega \mathbf{x}_0 \right\|_2^2 \leq (1 + \delta'_S)\|\mathbf{x}_0\|_2^2 \quad (6.71)$$

for all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$ whenever

$$MK \geq \frac{512(1 + \delta_S)^2 \left(S \left(\log\left(\frac{42}{\delta}\right) + 1 + \log\left(\frac{N}{S}\right) \right) + \log\left(\frac{2}{\nu}\right) \right)}{(1 - \delta_S)^2 \delta^2}. \quad (6.72)$$

Proof of Corollary 6.68 and Corollary 6.70

We simply need to derive a lower bound on $\Gamma(\mathcal{A}_\Omega \mathbf{x}_0)$ as an evaluation of ρ . Recall (6.50) and define

$$\mathbf{z}_0 := \left[\|A^{k_0} \mathbf{x}_0\|_2^2 \ \|A^{k_1} \mathbf{x}_0\|_2^2 \ \cdots \ \|A^{k_{K-1}} \mathbf{x}_0\|_2^2 \right]^T \in \mathbb{R}^K.$$

If all the entries of \mathbf{z}_0 lie within some bounds as $\ell_\ell \leq z_0(i) \leq \ell_h$ for all i , then one can show that

$$\Gamma(\mathcal{A}_\Omega \mathbf{x}_0) \geq K \left(\frac{\ell_\ell}{\ell_h} \right)^2. \quad (6.73)$$

Using the deterministic bound derived in (6.63) on $\|A^{k_i} \mathbf{x}_0\|_2^2$ for all $i \in \{0, 1, \dots, K-1\}$, one can show that when $\lambda = 1$ ($\lambda_{1,\max} = \lambda_{1,\min} = \lambda$ and $\lambda_{2,\max} = 0$), $\ell_\ell = (1 - \delta_S) \frac{L}{N} \|\mathbf{x}_0\|_2^2$ and $\ell_h = (1 + \delta_S) \frac{L}{N} \|\mathbf{x}_0\|_2^2$, and thus,

$$\rho \geq K \frac{(1 - \delta_S)^2}{(1 + \delta_S)^2}.$$

Similarly one can show that when $\lambda < 1$,

$$\rho \geq K \frac{(1 - \delta_S)^2}{(1 + \delta_S)^2} \lambda^{4(k_{K-1} - k_0)}, \quad (6.74)$$

and when $\lambda > 1$,

$$\rho \geq K \frac{(1 - \delta_S)^2}{(1 + \delta_S)^2} \lambda^{-4(k_{K-1} - k_0)}. \quad (6.75)$$

Using these lower bounds on ρ (recall that ρ is defined in (6.65) as the infimum of $\Gamma(\mathcal{A}_\Omega \mathbf{x}_0)$ over all S -sparse $\mathbf{x}_0 \in \mathbb{R}^N$) in the result of Theorem 6.54 completes the proof. We also note that when $\lambda_{1,\max} = \lambda_{1,\min} = \lambda$ and $\lambda_{2,\max} = 0$, the upper bound given in (6.66) can be used to bound the left-hand side failure probability even when $\epsilon \geq \frac{16}{\sqrt{K}}$. In fact, we can show that (6.66) holds for any $\epsilon \in (0, 1)$. The RIP results of Corollaries 6.68 and 6.70 follow based on this CoM inequality.

These results essentially indicate that the more λ deviates from one, the more total measurements MK are required to ensure unique recovery of any S -sparse initial state \mathbf{x}_0 . The bounds on ρ (which we state in Appendix B to derive Corollaries 6.68 and 6.70 from Theorem 6.54) also indicate that when $\lambda \neq 1$, the smallest number of measurements are required when the sample times are *consecutive* (i.e., when $k_{K-1} - k_0 = K$). Similar to what we mentioned earlier in our analysis for a scaled unitary A , when $\lambda \neq 1$ the effect of the initial state will be highly attenuated as we take measurements at later times (i.e., when $k_{K-1} - k_0 > K$) which results in a larger total number of measurements MK sufficient for exact recovery.

6.3.3 Identical Random Measurement Matrices

In this section, we consider the case where all matrices C_k are identical and equal to some $M \times N$ matrix C which is populated with i.i.d. Gaussian entries having zero mean and variance $\sigma^2 = \frac{1}{M}$. Once again note that we can write $\mathcal{O}_\Omega = \mathcal{C}_\Omega \mathcal{A}_\Omega$, where this time

$$\mathcal{C}_\Omega := \begin{bmatrix} C_{k_0} & & & \\ & C_{k_1} & & \\ & & \ddots & \\ & & & C_{k_{K-1}} \end{bmatrix} = \begin{bmatrix} C & & & \\ & C & & \\ & & \ddots & \\ & & & C \end{bmatrix}, \quad (6.76)$$

and \mathcal{A}_Ω is as defined in (6.44). The matrix \mathcal{C}_Ω is block diagonal with equal blocks on its main diagonal, and we have the following bound on its concentration behavior.

Theorem 6.77 [50] *Assume each of the measurement matrices $C_k \in \mathbb{R}^{M \times N}$ is populated with i.i.d. Gaussian random entries with mean zero and variance $\frac{1}{M}$. Assume all matrices C_k are the same (i.e., $C_k = C, \forall k$). Let $\mathbf{v}_{k_0}, \mathbf{v}_{k_1}, \dots, \mathbf{v}_{k_{K-1}} \in \mathbb{R}^N$ and define*

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_{k_0}^T & \mathbf{v}_{k_1}^T & \cdots & \mathbf{v}_{k_{K-1}}^T \end{bmatrix}^T \in \mathbb{R}^{KN}.$$

Then,

$$\mathbf{P} \left\{ \left| \|\mathcal{C}_\Omega \mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2 \right| > \epsilon \|\mathbf{v}\|_2^2 \right\} \leq \begin{cases} 2 \exp\left\{-\frac{M\epsilon^2 \|\boldsymbol{\lambda}\|_1^2}{256 \|\boldsymbol{\lambda}\|_2^2}\right\}, & 0 \leq \epsilon \leq \frac{16 \|\boldsymbol{\lambda}\|_2^2}{\|\boldsymbol{\lambda}\|_\infty \|\boldsymbol{\lambda}\|_1} \\ 2 \exp\left\{-\frac{M\epsilon \|\boldsymbol{\lambda}\|_1}{16 \|\boldsymbol{\lambda}\|_\infty}\right\}, & \epsilon \geq \frac{16 \|\boldsymbol{\lambda}\|_2^2}{\|\boldsymbol{\lambda}\|_\infty \|\boldsymbol{\lambda}\|_1}, \end{cases} \quad (6.78)$$

where

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}(\mathbf{v}) := \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{\min(K,N)} \end{bmatrix} \in \mathbb{R}^{\min(K,N)},$$

and $\{\lambda_1, \lambda_2, \dots, \lambda_{\min(K,N)}\}$ are the first (non-zero) eigenvalues of the $K \times K$ matrix $V^T V$, where

$$V = \begin{bmatrix} \mathbf{v}_{k_0} & \mathbf{v}_{k_1} & \cdots & \mathbf{v}_{k_{K-1}} \end{bmatrix} \in \mathbb{R}^{N \times K}.$$

Consider the first of the cases given in the right-hand side of the above bound. (This case permits any value of ϵ between 0 and $\frac{16}{\sqrt{\min(K,N)}}$.) Define

$$\Lambda(\mathbf{v}) := \frac{\|\boldsymbol{\lambda}(\mathbf{v})\|_1^2}{\|\boldsymbol{\lambda}(\mathbf{v})\|_2^2} \quad (6.79)$$

and note that for any $\mathbf{v} \in \mathbb{R}^{NK}$, $1 \leq \Lambda(\mathbf{v}) \leq \min(K, N)$. Moving forward, we will assume for simplicity that $K \leq N$, although this assumption can be removed. The case $\Lambda(\mathbf{v}) = K$ is quite favorable and implies that we get the same degree of concentration from the $MK \times NK$ block diagonal matrix \mathcal{C}_Ω as we would get from a dense $MK \times NK$ matrix populated with i.i.d. Gaussian random variables. This event happens if and only if $\lambda_1 = \lambda_2 = \cdots = \lambda_K$,

which happens if and only if

$$\|\mathbf{v}_{k_0}\|_2 = \|\mathbf{v}_{k_1}\|_2 = \cdots = \|\mathbf{v}_{k_{K-1}}\|_2$$

and $\langle \mathbf{v}_{k_i}, \mathbf{v}_{k_\ell} \rangle = 0$ for all $0 \leq i, \ell \leq K-1$ with $i \neq \ell$. On the other hand, the case $\Lambda(\mathbf{v}) = 1$ is quite unfavorable and implies that we get the same degree of concentration from the $MK \times NK$ block diagonal matrix \mathcal{C}_Ω as we would get from a dense Gaussian matrix having only M rows. This event happens if and only if the dimension of $\text{span}\{\mathbf{v}_{k_0}, \mathbf{v}_{k_1}, \dots, \mathbf{v}_{k_{K-1}}\}$ equals 1. Thus, comparing to Section 6.3.1, uniformity in the norms of the vectors \mathbf{v}_k is no longer sufficient for a high probability of concentration; in addition to this we must have diversity in the directions of the \mathbf{v}_{k_i} .

The following corollary of Theorem 6.77 derives a CoM inequality for the observability matrix. Recall that $\mathcal{O}_\Omega \mathbf{x}_0 = \mathcal{C}_\Omega \mathcal{A}_\Omega \mathbf{x}_0$ where \mathcal{C}_Ω is a block diagonal matrix whose diagonal blocks are repeated.

Corollary 6.80 *Suppose the same notation and assumptions as in Theorem 6.77 and suppose $K \leq N$. Then for any fixed initial state $\mathbf{x}_0 \in \mathbb{R}^N$ and for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,*

$$\mathbf{P} \left\{ \left| \|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2 - \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \right| > \epsilon \|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 \right\} \leq 2 \exp \left\{ -\frac{M \Lambda(\mathcal{A}_\Omega \mathbf{x}_0) \epsilon^2}{256} \right\}. \quad (6.81)$$

Once again, there are two important phenomena to consider in this result, and both are impacted by the interaction of A with \mathbf{x}_0 . First, on the left hand side of (6.81), we see that the point of concentration of $\|\mathcal{O}_\Omega \mathbf{x}_0\|_2^2$ is around $\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2$. Second, on the right-hand side of (6.81), we see that the exponent of the concentration failure probability scales with $\Lambda(\mathcal{A}_\Omega \mathbf{x}_0)$, which is determined by the eigenvalues of the $K \times K$ Gram matrix $V^T V$, where

$$V = [A^{k_0} \mathbf{x}_0 \quad A^{k_1} \mathbf{x}_0 \quad \cdots \quad A^{k_{K-1}} \mathbf{x}_0] \in \mathbb{R}^{N \times K}.$$

As mentioned earlier, $1 \leq \Lambda(\mathcal{A}_\Omega \mathbf{x}_0) \leq K$. The case $\Lambda(\mathcal{A}_\Omega \mathbf{x}_0) = K$ is quite favorable and happens when $\|A^{k_0} \mathbf{x}_0\|_2 = \|A^{k_1} \mathbf{x}_0\|_2 = \cdots = \|A^{k_{K-1}} \mathbf{x}_0\|_2$ and $\langle A^{k_i} \mathbf{x}_0, A^{k_\ell} \mathbf{x}_0 \rangle = 0$ for all

$0 \leq i, \ell \leq K - 1$ with $i \neq \ell$. The case $\Lambda(\mathcal{A}_\Omega \mathbf{x}_0) = 1$ is quite unfavorable and happens if the dimension of $\text{span}\{A^{k_0} \mathbf{x}_0, A^{k_1} \mathbf{x}_0, \dots, A^{k_{K-1}} \mathbf{x}_0\}$ equals 1.

In the special case where A is unitary, we know that $\|\mathcal{A}_\Omega \mathbf{x}_0\|_2^2 = K\|\mathbf{x}_0\|_2^2$. However, a unitary system matrix does not guarantee a favorable value for $\Lambda(\mathcal{A}_\Omega \mathbf{x}_0)$. Indeed, if $A = I_{N \times N}$ we obtain the worst case value $\Lambda(\mathcal{A}_\Omega \mathbf{x}_0) = 1$. If, on the other hand, A acts as a rotation that takes a state into an orthogonal subspace, we will have a stronger result.

Corollary 6.82 *Suppose the same notation and assumptions as in Theorem 6.77 and suppose $K \leq N$. Suppose that A is a unitary operator. Suppose also that $\langle A^{k_i} \mathbf{x}_0, A^{k_\ell} \mathbf{x}_0 \rangle = 0$ for all $0 \leq i, \ell \leq K - 1$ with $i \neq \ell$. Then for any fixed initial state $\mathbf{x}_0 \in \mathbb{R}^N$ and for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,*

$$\mathbf{P} \left\{ \left| \left\| \frac{1}{\sqrt{K}} \mathcal{O}_\Omega \mathbf{x}_0 \right\|_2^2 - \|\mathbf{x}_0\|_2^2 \right| > \epsilon \|\mathbf{x}_0\|_2^2 \right\} \leq 2 \exp \left\{ -\frac{MK\epsilon^2}{256} \right\}. \quad (6.83)$$

This result requires a particular relationship between A and \mathbf{x}_0 , namely that $\langle A^{k_i} \mathbf{x}_0, A^{k_\ell} \mathbf{x}_0 \rangle = 0$ for all $0 \leq i, \ell \leq K - 1$ with $i \neq \ell$. Thus, given a particular system matrix A , it is possible that it might hold for some \mathbf{x}_0 and not others. One must therefore be cautious in using this concentration result for CS applications (such as proving the RIP) that involve applying the concentration bound to a prescribed collection of vectors [32]; one must ensure that the “orthogonal rotation” property holds for each vector in the prescribed set.

Remark 6.84 *As mentioned in Remark 6.43, when the state transition matrix A is unitary and the measurement matrices C_k are independent randomly-populated matrices, our RIP result in Section 6.2 states that we can recover any S -sparse initial state \mathbf{x}_0 from K arbitrarily-chosen observations via solving an ℓ_1 -minimization problem when the total number of measurements KM scales linearly in S . It is worth mentioning that a similar linear RIP estimate can be achieved using the CoM results given in Section 6.3, namely Corollary 6.3.2. Baraniuk et al. [32] and Mendelson et al. [48] showed that CoM inequalities can be used*

to prove the RIP for random unstructured compressive matrices. For structured matrices, however, using the CoM inequalities usually does not lead to strong RIP bound [3, 4, 50]. This motivated us to directly prove the RIP for structured matrices [33, 56, 86]. However, in the case of the observability matrix and the special case of unitary A , the CoM inequalities actually result in a strong RIP result comparable to the result achieved from our direct analysis in Section 6.2.

6.4 Case Study: Estimating the Initial State in a Diffusion Process

So far we have provided theorems that provide a sufficient number of measurements for stable recovery of a sparse initial state under certain conditions on the state transition matrix and under the assumption that the measurement matrices are independent and populated with random entries. In this section, we use a case study to illustrate some of the phenomena raised in the previous sections.

6.4.1 System Model

We consider the problem of estimating the initial state of a system governed by the diffusion equation

$$\frac{\partial x}{\partial t} = \nabla \cdot (D(p)\nabla x(p, t)),$$

where $x(p, t)$ is the concentration, or density, at position p at time t , and $D(p)$ is the diffusion coefficient at position p . If D is independent of position, then this simplifies to

$$\frac{\partial x}{\partial t} = D\nabla^2 x(p, t).$$

The boundary conditions can vary according to the surroundings of the domain Π . If Π is bounded by an impermeable surface (e.g., a lake surrounded by the shore), then the boundary conditions are $n(p) \cdot \frac{\partial x}{\partial p} \Big|_{p \in \partial \Pi} = 0$, where $n(p)$ is the normal to $\partial \Pi$ at p . We will work with an approximate model discretized in time and in space. For simplicity, we explain a one-dimensional (one spatial dimension) diffusion process here but a similar approach of discretization can be taken for a diffusion process with two or three spatial dimension. Let

$\mathbf{p} := \begin{bmatrix} p(1) & p(2) & \cdots & p(N) \end{bmatrix}^T$ be a vector of equally-spaced locations with spacing Δ_s , and $\mathbf{x}(p, t) := \begin{bmatrix} x(p(1), t) & x(p(2), t) & \cdots & x(p(N), t) \end{bmatrix}^T$. Then a first difference approximation in space gives the model

$$\dot{\mathbf{x}}(p, t) = G\mathbf{x}(p, t), \quad (6.85)$$

where G represents the discrete Laplacian. We have

$$G = -L = \frac{D}{\Delta_s^2} \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & & & 1 & -1 \end{bmatrix},$$

where L is the Laplacian matrix associated with a path (one spatial dimension). This discrete Laplacian G has eigenvalues $\lambda_i = -\frac{2D}{\Delta_s^2} (1 - \cos(\frac{\pi}{N}(i-1)))$ for $i = 1, 2, \dots, N$.

To obtain a discrete-time model, we choose a sampling time T_s and let the vector $\mathbf{x}_k = \mathbf{x}(p, kT_s)$ be the concentration at positions $p(1), p(2), \dots, p(N)$ at sampling time k . Using a first difference approximation in time, we have

$$\mathbf{x}_k = A\mathbf{x}_{k-1},$$

where $A = I_N + GT_s$. For a diffusion process with two spatial dimensions, a similar analysis would follow, except one would use the Laplacian matrix of a grid (instead of the Laplacian matrix of a one-dimensional path) in $A = I_N + GT_s$. For all simulations in this section we take $D = 1$, $\Delta_s = 1$, $N = 100$, and $T_s = 0.1$. An example simulation of a one-dimensional diffusion is shown in Figure 6.1, where we have initialized the system with a sparse initial state \mathbf{x}_0 containing unit impulses at $S = 10$ randomly chosen locations.

In Section 6.4.3, we provide several simulations which demonstrate that recovery of a sparse initial state is possible from compressive measurements.

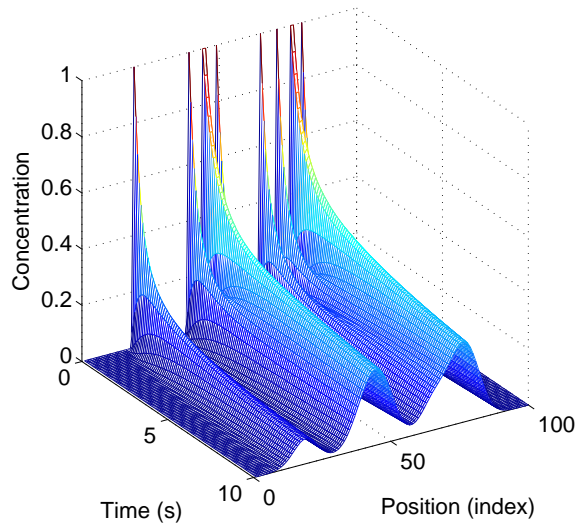


Figure 6.1: One-dimensional diffusion process. At time zero, the concentration (the state) is non-zero only at a few locations of the path graph of $N = 100$ nodes.

6.4.2 Diffusion and its Connections to Theorem 6.54

Before presenting the recovery results from compressive measurements, we would like to mention that our analysis in Theorem 6.54 gives some insight into (but is not precisely applicable to) the diffusion problem. In particular, the discrete Laplacian matrix G and the corresponding state transition matrix A (see below) are almost circulant, and so their eigenvectors will closely resemble the DFT basis vectors. The largest eigenvalues correspond to the lowest frequencies, and so the U_1 matrix corresponding to G or A will resemble a basis of the lowest frequency DFT vectors. While such a matrix does not technically satisfy the RIP, matrices formed from random sets of DFT vectors do satisfy the RIP with high probability [100]. Thus, even though we cannot apply Theorem 6.54 directly to the diffusion problem, it does provide some intuition that sparse recovery should be possible in the diffusion setting.

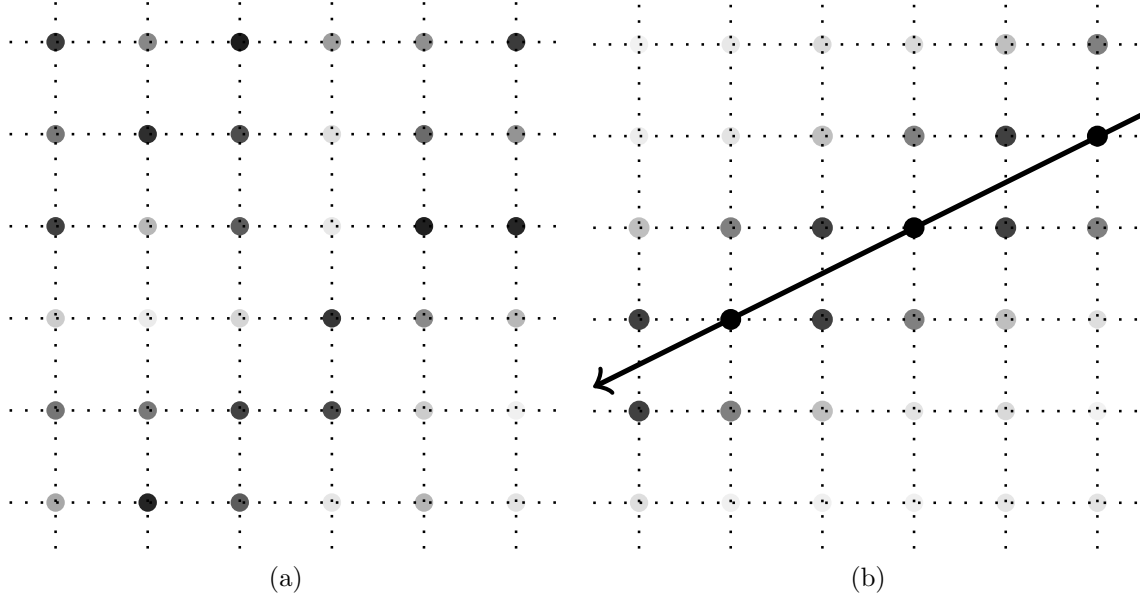


Figure 6.2: Dense Measurements versus Line Measurements. The color of a node indicates the corresponding weight of that node. The darker the node color, the higher the weight. These weights are the entries of each row of each C_k . (a) Dense Measurements. The weights are drawn from a Gaussian distribution with mean zero and variance $\frac{1}{M}$. These values are random and change for each measurement. (b) Line Measurements. The weights are generated as a function of the perpendicular distances of all nodes of the grid to the line. The slope and the intercept of the line are random and change for each measurement.

6.4.3 State Recovery from Compressive Measurements

In this section, we consider a two-dimensional diffusion process. As mentioned earlier, the state transition matrix A associated with this process is of the form $A = I_N + GT_s$, where T_s is the sampling time and G is the Laplacian matrix of a grid. In these simulations, we consider a grid of size 10×10 with $T_s = 0.1$.

We also consider two types of measuring processes. We first look at random measurement matrices $C_k \in \mathbb{R}^{M \times N}$ where the entries of each matrix are i.i.d. Gaussian random variables with mean zero and variance $\frac{1}{M}$. Note that this type of measurement matrix falls within the assumptions of our theorems in Sections 6.2 and 6.3.1. In this measuring scenario, all of the nodes of the grid (i.e., all of the states) will be measured at each sample time. Formally, at each observation time we record a random linear combination of all nodes. In the following,

we refer to such measurements as “Dense Measurements.” Figure Figure 6.2(a) illustrates an example of how the random weights are spread over the grid. The weights (the entries of each row of each C_k) are shown using grayscale. The darker the node color, the higher the corresponding weight. We also consider a more practical measuring process in which at each sample time the operator measures the nodes of the grid occurring along a line with random slope and random intercept. Formally, $C_k(i, j) = \exp\left(-\frac{d_k(i, j)}{c}\right)$ where $d_k(i, j)$ is the perpendicular distance of node j ($j = 1, \dots, N$) to the i th ($i = 1, \dots, M$) line with random slope and random intercept and c is an absolute constant that determines how fast the node weights decrease as their distances increase from the line. Figure Figure 6.2(b) illustrates an example of how the weights are spread over the grid in this scenario. Observe that the nodes that are closer to the line are darker, indicating higher weights for those nodes. We refer to such measurements as “Line Measurements.”

To address the problem of recovering the initial state \mathbf{x}_0 , let us first consider the situation where we collect measurements only of $\mathbf{x}_0 \in \mathbb{R}^{100}$ itself. We fix the sparsity level of \mathbf{x}_0 to $S = 9$. For various values of M , we construct measurement matrices C_0 according to the two models explained above. At each trial, we collect the measurements $\mathbf{y}_0 = C_0\mathbf{x}_0$ and attempt to recover \mathbf{x}_0 given \mathbf{y}_0 and C_0 using the canonical ℓ_1 -minimization problem from CS:

$$\hat{\mathbf{x}}_0 = \arg \min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{y}_k = C_k A^k \mathbf{x} \quad (6.86)$$

with $k = 0$. (In the next paragraph, we repeat this experiment for different k .) In order to imitate what might happen in reality (e.g., a drop of poison being introduced to a lake of water at $k = 0$), we assume the initial contaminant appears in a cluster of nodes on the associated diffusion grid. In our simulations, we assume the $S = 9$ non-zero entries of the initial state correspond to a 3×3 square-neighborhood of nodes on the grid. For each M , we repeat the recovery problem for 300 trials; in each trial we generate a random sparse initial state \mathbf{x}_0 (an initial state with a random location of the 3×3 square and random values of the 9 non-zero entries) and a measurement matrix C_0 as explained above.

Figure 6.3(a) depicts, as a function of M , the percent of trials (with \mathbf{x}_0 and C_0 randomly chosen in each trial) in which the initial state is recovered perfectly, i.e., $\hat{\mathbf{x}}_0 = \mathbf{x}_0$. Naturally, we see that as we take more measurements, the recovery rate increases. When Line Measurements are taken, with almost 35 measurements we recover every sparse initial state of dimension 100 with sparsity level 9. When Dense Measurements are employed, however, we observe a slightly weaker recovery performance at $k = 0$ as almost 45 measurements are required to see exact recovery. In order to see how the diffusion phenomenon affects the recovery, we repeat the same experiment at $k = 10$. In other words, we collect the measurements $\mathbf{y}_{10} = C_{10}\mathbf{x}_{10} = C_{10}A^{10}\mathbf{x}_0$ and attempt to recover \mathbf{x}_0 given \mathbf{y}_{10} and $C_{10}A^{10}$ using the canonical ℓ_1 -minimization problem (6.86). As shown in Fig. Figure 6.3(b), the recovery performance is improved when Line and Dense Measurements are employed (with almost 25 measurements exact recovery is possible). Qualitatively, this suggests that due to diffusion, at $k = 10$, the initial contaminant is now propagating and consequently a larger surface of the lake (corresponding to more nodes of the grid) is contaminated. In this situation, a higher number of contaminated nodes will be measured by Line Measurements which potentially can improve the recovery performance of the initial state.

In order to see how the recovery performance would change as we take measurements at different times, we repeat the previous example for $k = \{0, 1, 2, 8, 50, 100\}$. The results are shown in Fig. Figure 6.4(a) and Fig. Figure 6.4(b) for Dense and Line Measurements, respectively. In both cases, the recovery performance starts to improve as we take measurements at later times. However, in both measuring scenarios, the recovery performance tends to decrease if we wait too long to take measurements. For example, as shown in Fig. Figure 6.4(a), the recovery performance is significantly decreased at time $k = 100$ when Dense Measurements are employed. A more dramatic decrease in the recovery performance can be observed when Line Measurements are employed in Fig. Figure 6.4(b). Again this behavior is as expected and can be interpreted with the diffusion phenomenon. If we wait too long to take measurements from the field of study (e.g., the lake of water), the effect of the initial

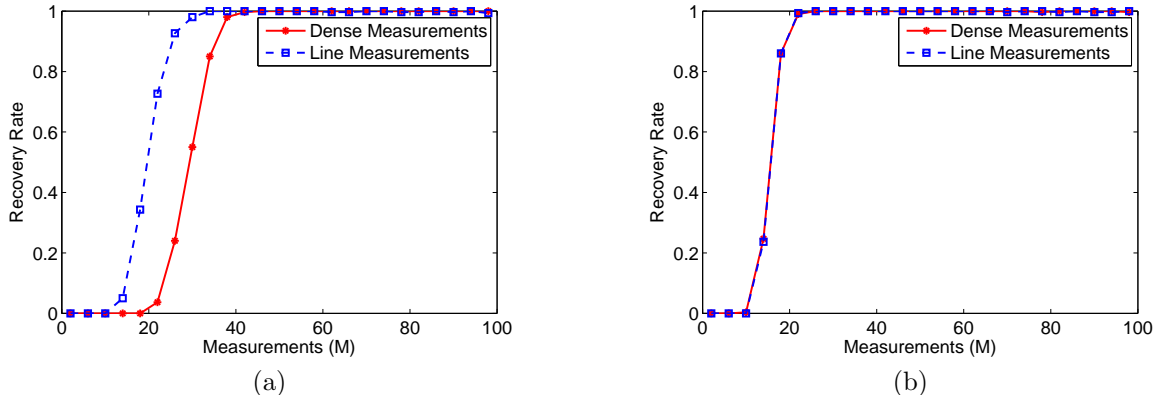


Figure 6.3: Signal recovery from compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. The plots show the percent of trials (out of 300 trials in total) with perfect recovery of the initial state \mathbf{x}_0 versus the number of measurements M . (a) Recovery from compressive measurements at time $k = 0$. (b) Recovery from compressive measurements at time $k = 10$.

contaminant starts to disappear in the field (due to diffusion) and consequently measurements at later times contain less information. In summary, one could conclude from these observations that taking compressive measurements of a diffusion process at times that are too early or too late might decrease the recovery performance.

In another example, we fix $M = 32$, consider the same model for the sparse initial states with $S = 9$ as in the previous examples, introduce white noise in the measurements with standard deviation 0.05, use a noise-aware version of the ℓ_1 recovery algorithm [27], and plot a histogram of the recovery errors $\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_2$. We perform this experiment at $k = 2$ and $k = 10$. As can be seen in Fig. Figure 6.5(a), at time $k = 2$ the Dense Measurements have lower recovery errors (almost half) compared to the Line Measurements. However, if we take measurements at time $k = 10$, the recovery error of both measurement processes tends to be similar, as depicted in Fig. Figure 6.5(b).

Of course, it is not necessary to take all of the measurements only at one observation time. What may not be obvious a priori is how spreading the measurements over time may impact the initial state recovery. To this end, we perform the signal recovery experiments

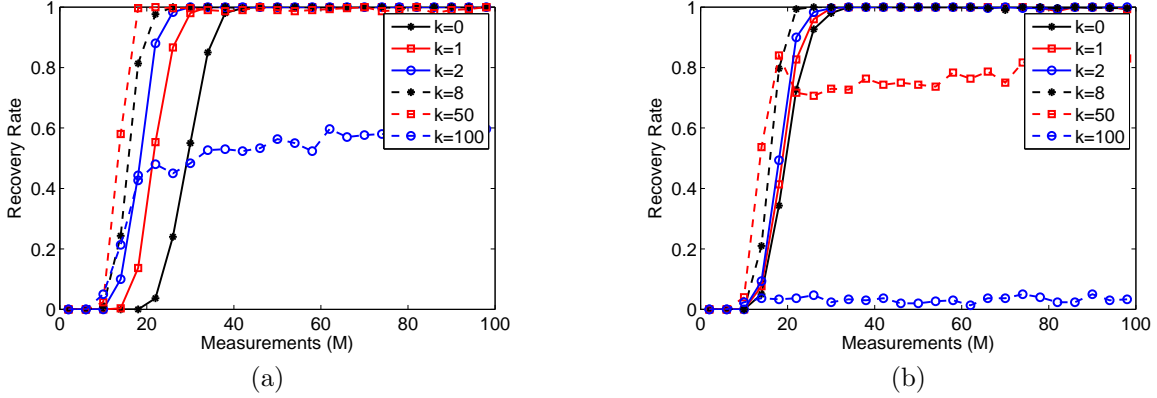


Figure 6.4: Signal recovery from compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. The plots show the percent of trials (out of 300 trials in total) with perfect recovery of the initial state \mathbf{x}_0 versus the number of measurements M taken at observation times $k = \{0, 1, 2, 8, 50, 100\}$. (a) Recovery from compressive Dense Measurements. (b) Recovery from compressive Line Measurements.

when a total of $MK = 32$ measurements are spread over $K = 4$ observation times (at each observation time we take $M = 8$ measurements). In order to see how different observation times affect the recovery performance, we repeat the experiment for different sample sets, Ω_i . We consider 10 sample sets as $\Omega_1 = \{0, 1, 2, 3\}$, $\Omega_2 = \{4, 5, 6, 7\}$, $\Omega_3 = \{8, 9, 10, 11\}$, $\Omega_4 = \{10, 20, 30, 40\}$, $\Omega_5 = \{20, 21, 22, 23\}$, $\Omega_6 = \{10, 30, 50, 70\}$, $\Omega_7 = \{51, 52, 53, 54\}$, $\Omega_8 = \{60, 70, 80, 90\}$, $\Omega_9 = \{91, 92, 93, 94\}$, and $\Omega_{10} = \{97, 98, 99, 100\}$. Figure Figure 6.6(a) illustrates the results. For both of the measuring scenarios, the overall recovery performance improves when we take measurements at later times. As mentioned earlier, however, if we wait too long to take measurements the recovery performance drops. For sample sets Ω_2 through Ω_6 , we have perfect recovery of the initial state only from $MK = 32$ total measurements, either using Dense or Line Measurements. The overall recovery performance is not much different compared to, say, taking $M = 32$ measurements at a single instant and so there is no significant penalty that one pays by slightly spreading out the measurement collection process in time, as long as a different random measurement matrix is used at each sample time. We repeat the same experiment when the measurements are noisy. We intro-

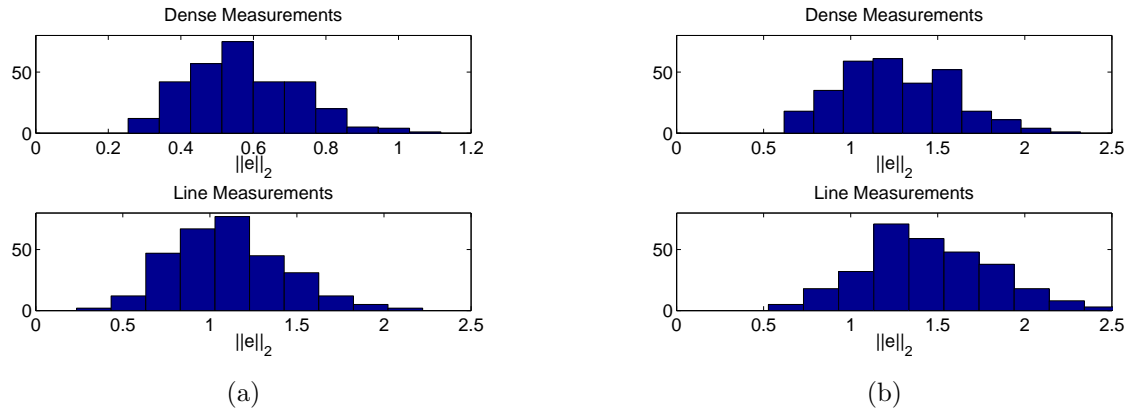


Figure 6.5: Signal recovery from $M = 32$ compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. The plots show the recovery error of the initial state $\|e\|_2 = \|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_2$ over 300 trials. (a) Recovery from compressive measurements at time $k = 2$. (b) Recovery from compressive measurements at time $k = 10$.

duce white noise in the measurements with standard deviation 0.05 and use a noise-aware version of the ℓ_1 -minimization problem to recover the true solution. Figure Figure 6.6(b) depicts a histogram of the recovery errors $\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_2$ when $MK = 32$ measurements are spread over $K = 4$ sample times $\Omega_4 = \{10, 20, 30, 40\}$.

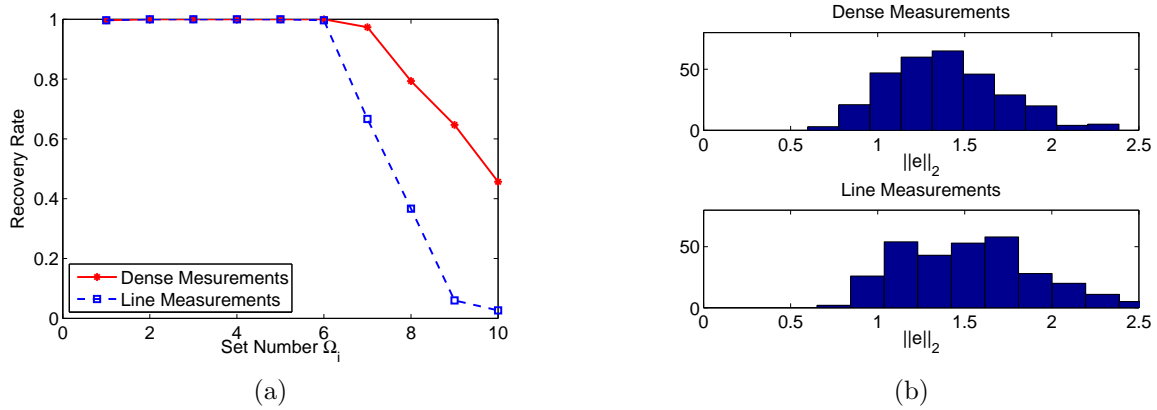


Figure 6.6: Signal recovery from compressive measurements of a diffusion process which has initiated from a sparse initial state of dimension $N = 100$ and sparsity level $S = 9$. A total of $KM = 32$ measurements are spread over $K = 4$ observation times while at each time, $M = 8$ measurements are taken. (a) Percent of trials (out of 300 trials in total) with perfect recovery of the initial state \mathbf{x}_0 are shown for different sample sets, Ω_i . (b) Recovery error of the initial state $\|e\|_2 = \|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_2$ over 300 trials for set Ω_4 .

CHAPTER 7

CONCLUSIONS

Many dynamical systems of practical interest often contain a notion of simplicity in their intrinsic structure or representation. In this thesis, we developed new theory and algorithms for exploiting this type of simplicity in the analysis of high-dimensional dynamical systems with a particular focus on system identification and estimation.

Our work was inspired by the field of Compressive Sensing (CS) which is a recent paradigm in signal processing for sparse signal recovery. The CS recovery problem states that a sparse signal can be recovered from a small number of random linear measurements. Compared to the standard CS setup, in most of the problems considered in this thesis we dealt with structured signals and measurement matrices where the structure was implied by the system under study.

As summarized in the next few sections of this chapter, our contributions in this thesis are: *new Concentration of Measure (CoM) inequalities* for Toeplitz matrices with applications; *a new Compressive Topology Identification (CTI) framework* for the analysis of large-scale interconnected dynamical systems; *a new algorithm* for recovering clustered-sparse signals with an application in CTI; *new analysis* on the recovery conditions of high-dimensional but sparse Linear Time-Invariant (LTI) Auto Regressive with eXternal input (ARX) systems; *a new pre-filtering scheme* for coherence reduction; *a new framework* for compressive identification of time-varying ARX models with an application in DC motor identification; *new theory, recovery guarantees, and analysis on the observability problem* of linear systems with high-dimensional but sparse initial states. As briefly discussed in Section 7.3, there are many possible future directions for this research.

7.1 Sparse Recovery and CSI

Throughout this thesis we showed that in many applications of practical interest it is possible to perform system identification from few observations if the system under study contains a notion of simplicity. We demonstrated in Chapter 3 that impulse response recovery is possible from few convolution-based measurements when it is known a priori that the unknown impulse response is high-dimensional but sparse. In Chapter 4 we considered topology identification of large-scale interconnected dynamical systems and showed that compressive topology identification is possible using block-sparse or clustered-sparse recovery algorithms. In Chapter 5 we considered LTI and Linear Time-Variant (LTV) ARX models and showed that recovery of the model parameters is possible from few measurements. In Chapter 6 we demonstrated that initial state recovery is possible from few observations when it is known a priori that the initial state is high-dimensional but sparse. Whilst for all of these applications, we provided simulation results, a significant amount of our work and this thesis was devoted to analyzing the recovery guarantees and conditions.

7.2 Recovery Guarantees

For the problems considered in this thesis, we analyzed and derived recovery guarantees using several different approaches. As we have mentioned, each approach has its own importance and implications.

7.2.1 Concentration of Measure Inequalities

In Chapter 3 we derived CoM inequalities for randomized compressive Toeplitz matrices. These inequalities showed that the norm of a high-dimensional signal mapped by a compressive Toeplitz matrix to a low-dimensional space concentrates around its mean with a tail probability bound that decays exponentially in the dimension of the range space divided by a quantity which is a function of the signal. This implies that the CoM inequalities for compressive Toeplitz matrices are non-uniform and signal-dependent. We further analyzed the behavior of the introduced quantity. In particular, we showed that for the class of *sparse*

signals, the introduced quantity is bounded by the sparsity level of the signal. However, we observed that this bound is highly pessimistic for most sparse signals and we showed that if a random distribution is imposed on the non-zero entries of the signal, the typical value of the quantity is bounded by a term that scales logarithmically in the ambient dimension. We also extended our analysis for signals that are sparse in a generic orthobasis. To this end, we introduced the notion of the *Fourier coherence* of an arbitrary orthobasis and stated our generic results based on this measure. We considered the time and frequency domains as specific cases.

In Chapter 6 we derived CoM inequalities for the observability matrix and explained how the interaction between the state transition matrix and the initial state affect the concentration bounds. Our derivation was based on the CoM results for block-diagonal matrices. We observed that the observability matrix can be decomposed as a product of two matrices, one of which has a block-diagonal structure, where the diagonal blocks of this matrix are the measurement matrices C_k . We derived CoM inequalities for two cases. We first considered the case where all matrices C_k are generated independently of each other. We then considered the case where all matrices C_k are the same. In either case, we assumed that each matrix C_k is populated with independent and identically distributed (i.i.d.) Gaussian random variables. The concentration results cover a larger class of systems (e.g., not necessarily unitary) and initial states (not necessarily sparse). Aside from guaranteeing recovery of sparse initial states, the CoM results have potential applications in solving inference problems such as detection and classification of more general initial states and systems.

7.2.2 Restricted Isometry Property

A sufficient condition for correct recovery of sparse signals is the Restricted Isometry Property (RIP). In Chapter 3 we showed that as an implication of the CoM inequalities, one can derive the RIP for compressive Toeplitz matrices. Based on this result, we showed that one can establish the RIP for Toeplitz matrices if the number of rows (the number of measurements) scales quadratically in the sparsity level.

In Chapter 6 we first established the RIP for the observability matrix. We showed that the observability matrix satisfies the RIP under certain conditions on the state transition matrix. For example, we showed that if the state transition matrix is unitary, and if independent, randomly-populated measurement matrices are employed, then it is possible to uniquely recover a sparse high-dimensional initial state when the total number of measurements scales *linearly* in the sparsity level of the initial state. We also showed that a similar linear RIP estimate can be achieved using the CoM results for the observability matrix.

7.2.3 Coherence

Establishing the RIP for structured matrices is usually a hard task. It usually involves complicated mathematical analyses and tools. As a simpler but coarser recovery condition, we also analyzed coherence-based recovery guarantees.

In Chapter 4 we showed that for the type of matrices that appear in CTI, the coherence-based metrics have an asymptotic behavior; as we take more node measurements, the coherence-based metrics associated with the measurements matrix approach a non-zero value which depends on the link impulse responses.

We performed a similar analysis in Chapter 5 for the matrices that appear in the analysis of the ARX models. We observed that while the associated coherence-based recovery guarantees indicate no or very limited recovery, exact (or stable) recovery of such models is indeed possible from few measurements. In fact, this observation is not a contradiction, as these guarantees are sufficient (and not necessary) recovery conditions and furthermore, they are typically blunt tools as they reflect the worst correlations in the matrix. As a first step towards investigating this gap, we suggested a pre-filtering scheme by which the coherence of such matrices can be reduced.

7.3 Future Research Directions

In this thesis, we took some first steps towards analyzing high-dimensional dynamical systems when it is known a priori that the system under study contains a notion of sim-

plicity. To this end, we adapted the tools, theorems, and algorithms from CS and sparse signal processing. There are many possible future directions for this research. On one hand, there are many possible directions in developing new analyses when dynamical systems are involved. In particular, our work will continue with analyzing and developing new recovery conditions. As mentioned earlier, there exists a big gap between the recovery performance and the existing recovery guarantees. Therefore, one open research direction is to reduce this gap by developing new recovery guarantees, applying new techniques such as the pre-filtering, or considering new frameworks such as input design. Another open front is to apply the methods and algorithms developed in this thesis on real-world applications and data such as the gene regulatory network.

REFERENCES CITED

- [1] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, 2nd edition, 1999.
- [2] J. Romberg. Compressive sensing by random convolution. *SIAM Journal on Imaging Sciences*, 2(4):1098–1128, 2009.
- [3] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Concentration of measure inequalities for compressive Toeplitz matrices with applications to detection and system identification. in *Proc. 49th IEEE Conference on Decision and Control*, pages 2922–2929, 2010.
- [4] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Concentration of measure inequalities for Toeplitz matrices with applications. *to appear in IEEE Transactions on Signal Processing*, 2012.
- [5] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Concentration of measure inequalities for Toeplitz matrices with applications. technical report, October 2012. URL <http://arxiv.org/abs/1112.1968>.
- [6] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, 2010.
- [7] B. M. Sanandaji, T. L. Vincent, K. Poolla, and M. B. Wakin. A tutorial on recovery conditions for compressive system identification of sparse channels. in *Proc. 51th IEEE Conference on Decision and Control*, 2012.
- [8] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Exact topology identification of large-scale interconnected dynamical systems from compressive observations. in *Proc. 2011 American Control Conference*, pages 649–656, 2011.
- [9] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Compressive topology identification of interconnected dynamic systems via clustered orthogonal matching pursuit. in *Proc. 50th IEEE Conference on Decision and Control and European Control Conference*, pages 174–180, 2011.
- [10] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. A review of sufficient conditions for structure identification in interconnected systems. in *Proc. 16th IFAC Symposium on System Identification*, 2012.

- [11] M. B. Wakin, B. M. Sanandaji, and T. L. Vincent. On the observability of linear systems from random, compressive measurements. *in Proc. 49th IEEE Conference on Decision and Control*, pages 4447–4454, 2010.
- [12] B. M. Sanandaji, M. B. Wakin, and T. L. Vincent. Observability with random observations. *ArXiv preprint arXiv:1211.4077*, 2012. URL <http://arxiv.org/abs/1211.4077>.
- [13] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- [14] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- [15] Y. C. Eldar and M. Mishali. Block-sparsity and sampling over a union of subspaces. *in Proc. 16th International Conference on Digital Signal Processing*, 2009.
- [16] Y. C. Eldar, P. Kuppinger, and H. Bölcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.
- [17] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.
- [18] T. K. Moon and W. C. Stirling. *Mathematical methods and algorithms for signal processing*, volume 1. Prentice hall New York, 2000.
- [19] M. B. Wakin. The geometry of low-dimensional signal models. *PhD Dissertation, Rice University*, 2006.
- [20] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.
- [21] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.
- [22] E. J. Candès and T. Tao. Decoding via linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- [23] E. J. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2):227–254, 2006.

- [24] S. Chen, D. L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
- [25] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.
- [26] E. J. Candès. Compressive sampling. in *Proc. International Congress of Mathematicians*, pages 1433–1452, 2006.
- [27] E. J. Candès. The restricted isometry property and its implications for compressed sensing. in *Compte Rendus de l’Academie des Sciences, Paris, Series I*, 346:589–592, 2008.
- [28] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.
- [29] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.
- [30] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.
- [31] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.
- [32] R. G. Baraniuk, M. A. Davenport, R. A. DeVore, and M. B. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.
- [33] H. Rauhut, J. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. *Applied and Computational Harmonic Analysis*, 32:242–254, 2012.
- [34] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.
- [35] D. Healy and D. J. Brady. Compression at the physical interface. *IEEE Signal Processing Magazine*, 25(2):67–71, 2008.

- [36] M. B. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, J. Yoo, J. K. Romberg, A. Emami-Neyestanak, and E. J. Candès. A non-uniform sampler for wide-band spectrally-sparse environments. *preprint*, 2012.
- [37] J. Yoo, C. Turnes, E. Nakamura, C. Le, S. Becker, E. Sovero, M. B. Wakin, M. Grant, J. K. Romberg, A. Emami-Neyestanak, , and E. J. Candès. A compressed sensing parameter extraction platform for radar pulse signal acquisition. *preprint*, 2012.
- [38] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.
- [39] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.
- [40] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.
- [41] D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete and Computational Geometry*, 35(4):617–652, 2006.
- [42] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *in Proc. National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.
- [43] D. L. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *American Mathematical Society*, 22(1):1–53, 2009.
- [44] M. Ledoux. *The concentration of measure phenomenon*. Amer Mathematical Society, 2001.
- [45] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.
- [46] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.
- [47] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- [48] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and sub-Gaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

- [49] H. Rauhut. Circulant and Toeplitz matrices in compressed sensing. *Arxiv preprint arXiv:0902.4394*, 2009. URL <http://arxiv.org/abs/0902.4394>.
- [50] J. Y. Park, H. L. Yap, C. J. Rozell, and M. B. Wakin. Concentration of measure for block diagonal matrices with applications to compressive signal processing. *IEEE Transactions on Signal Processing*, 59(12):5859–5875, 2011.
- [51] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak. Compressed channel sensing. in *Proc. 42nd Annual Conference on Information Sciences and Systems*, 2008.
- [52] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, R. Tóth, and K. Poolla. Compressive system identification of LTI and LTV ARX models. in *Proc. 50th IEEE Conference on Decision and Control and European Control Conference*, pages 791–798, 2011.
- [53] J. A. Tropp, M. B. Wakin, M. F. Duarte, D. Baron, and R. G. Baraniuk. Random filters for compressive sampling and reconstruction. in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:872–875, 2006.
- [54] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak. Toeplitz-structured compressed sensing matrices. in *Proc. 14th IEEE Workshop on Statistical Signal Processing*, pages 294–298, 2007.
- [55] H. Rauhut. Compressive sensing and structured random matrices. *Theoretical Foundations and Numerical Methods for Sparse Recovery*, 9:1–92, 2010.
- [56] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Arxiv preprint arXiv:1207.0235*, 2012. URL <http://arxiv.org/abs/1207.0235>.
- [57] H. L. Yap and C. J. Rozell. On the relation between block diagonal matrices and compressive Toeplitz matrices. technical report, October 2011. URL <http://users.ece.gatech.edu/~crozell/pubs/yapGTTR2011.pdf>.
- [58] S. G. Hwang. Cauchy’s interlace theorem for eigenvalues of Hermitian matrices. *The American Mathematical Monthly*, 111(2):157–159, 2004.
- [59] G. Lugosi. Concentration-of-measure inequalities. *Lecture Notes*, 2004.
- [60] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and processes*. Springer, 1991.
- [61] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk. Signal processing with compressive measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, 2010.

- [62] J. Pakanen and S. Karjalainen. Estimating static heat flows in buildings for energy allocation systems. *Energy and Buildings*, 38(9):1044–1052, 2006.
- [63] N. Chaturvedi and J. E. Braun. An inverse gray-box model for transient building load prediction. *International Journal of Heating, Ventilating, Air-Conditioning and Refrigerating Research*, 8(1):73–100, 2002.
- [64] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.
- [65] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999.
- [66] S. Hara, H. Shimizu, and T. H. Kim. Consensus in hierarchical multi-agent dynamical systems with low-rank interconnections: Analysis of stability and convergence rates. in *Proc. 2009 American Control Conference*, pages 5192–5197, 2009.
- [67] R. Olfati-Saber and R. M. Murray. Consensus and cooperation in networked multi-agent systems. in *Proc. IEEE*, 95(1):215–233, 2007.
- [68] A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt. Controllability of multi-agent systems from a graph-theoretic perspective. *SIAM Journal on Control and Optimization*, 48(1):162–186, 2009.
- [69] G. Innocenti and D. Materassi. A modeling approach to multivariate analysis and clusterization theory. *Journal of Physics A: Mathematical and Theoretical*, 41(1):205101, 2008.
- [70] D. Materassi, G. Innocenti, and L. Giarré. Reduced complexity models in the identification of dynamical networks: links with sparsification problems. in *Proc. 48th IEEE Conference on Decision and Control*, pages 4796–4801, 2009.
- [71] D. Materassi and G. Innocenti. Topological identification in networks of dynamical systems. *IEEE Transactions on Automatic Control*, 55(8):1860–1871, 2010.
- [72] A. Bolstad, B.D. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *IEEE Trans. Signal Processing*, 59(6):2628–2641, 2001.
- [73] V. Cevher, P. Indyk, C. Hegde, and R. G. Baraniuk. Recovery of clustered sparse signals from compressive measurements. in *Proc. International Conference on Sampling Theory and Applications*, 2009.
- [74] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

- [75] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent. Compressive system identification in the linear time-invariant framework. *in Proc. 50th IEEE Conference on Decision and Control and European Control Conference*, pages 783–790, 2011.
- [76] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [77] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.
- [78] C. Lyzell, J. Roll, and L. Ljung. The use of nonnegative garrote for order selection of ARX models. *in Proc. 47th IEEE Conference on Decision and Control*, pages 1974–1979, 2008.
- [79] H. Ohlsson, L. Ljung, and S. Boyd. Segmentation of ARX-models using sum-of-norms regularization. *Automatica*, 46(6):1107–1111, 2010.
- [80] I. Maruta and T. Sugie. A new approach for modeling hybrid systems based on the minimization of parameters transition in linear time-varying models. *in Proc. 49th IEEE Conference on Decision and Control*, pages 117–1182, 2010.
- [81] S. Bjorklund and L. Ljung. A review of time-delay estimation techniques. *in Proc. 42th IEEE Conference on Decision and Control*, pages 2502–2507, 2003.
- [82] H. D. Taghirad and P. R. Bélanger. Modeling and parameter identification of harmonic drive systems. *Dynamic Systems, Measurement, and Control*, 120(4):439–444, 1998.
- [83] B. Armstrong-Héouvry, P. Dupont, and C. Canudas De Wit. A survey of models, analysis tools and compensation methods for the control of machines with friction. *Automatica*, 30(7):1083–1138, 1994.
- [84] M. B. Wakin, J. Y. Park, H. L. Yap, and C. J. Rozell. Concentration of measure for block diagonal measurement matrices. *in Proc. International Conference on Acoustics, Speech, Signal Processing*, 2010.
- [85] C. J. Rozell, H. L. Yap, J. Y. Park, and M. B. Wakin. Concentration of measure for block diagonal matrices with repeated blocks. *in Proc. 44th Annual Conference on Information Sciences and Systems*, 2010.
- [86] H. L. Yap, A. Eftekhari, M. B. Wakin, and C. J. Rozell. The restricted isometry property for block diagonal matrices. *in Proc. 45th Annual Conference on Information Sciences and Systems*, 2011.

- [87] A. Eftekhari, H. L. Yap, M. B. Wakin, and C. J. Rozell. Restricted isometry property for random block-diagonal matrices. *ArXiv preprint arXiv:1210.3395*, 2012. URL <http://arxiv.org/abs/1210.3395>.
- [88] C. T. Chen. *Linear System Theory and Design*. Oxford University Press, 3rd edition, 1999.
- [89] R. A. DeVore, G. Petrova, and P. Wojtaszczyk. Instance-optimality in probability with an ℓ_1 -minimization decoder. *Applied and Computational Harmonic Analysis*, 27(3):275–288, 2009.
- [90] Mark A. Davenport. *Random Observations on Random Observations: Sparse Signal Acquisition and Processing*. PhD thesis, Rice University, 2010.
- [91] E. Wang, J. Silva, and L. Carin. Compressive particle filtering for target tracking. in *Proc. Statistical Signal Processing Workshop*, pages 233–236, 2009.
- [92] W. Dai and S. Yuksel. Technical report: Observability of a linear system under sparsity constraints. *ArXiv preprint arXiv:1204.3097*, 2012. URL <http://arxiv.org/abs/1204.3097>.
- [93] H. L. Yap, A. S. Charles, and C. J. Rozell. The restricted isometry property for echo state networks with applications to sequence memory capacity. in *Proc. 2012 IEEE Statistical Signal Processing Workshop*, pages 580–583, 2012.
- [94] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arxiv:1011.3027*, 2011. URL <http://arxiv.org/abs/1011.3027>.
- [95] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. in *Proc. 40th Annual Conference on Information Sciences and Systems*, pages 207–212, 2006.
- [96] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *IEEE Transactions on Information Theory*, 56(1):520–544, 2010.
- [97] Y. C. Eldar. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.
- [98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. in *Proc. ACM Symposium on Theory of Computing*, pages 604–613, 1998.

- [99] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.
- [100] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

APPENDIX - PROOFS

A.1 Proof of Lemma 3.48

We start by proving a more general version of Lemma 3.48 and (3.49). Let z_1, z_2, \dots, z_n be random variables.

Consider the event $\mathcal{E}_A \triangleq \{z_1 < c_1 U \text{ and } z_2 < c_2 U \text{ and } \dots z_n < c_n U\}$ where c_1, c_2, \dots, c_n are fixed numbers that sum to 1. It is trivial to see that if \mathcal{E}_A happens, then the event $\mathcal{E}_B \triangleq \{z_1 + z_2 + \dots + z_n < U\}$ must also occur. Consequently, $\mathbf{P}\{(\mathcal{E}_B)^c\} \leq \mathbf{P}\{(\mathcal{E}_A)^c\}$, where

$$(\mathcal{E}_A)^c = \{z_1 \geq c_1 U \text{ or } z_2 \geq c_2 U \text{ or } \dots z_n \geq c_n U\}.$$

Using the union bound, we have $\mathbf{P}\{(\mathcal{E}_A)^c\} \leq \mathbf{P}\{z_1 \geq c_1 U\} + \mathbf{P}\{z_2 \geq c_2 U\} + \dots + \mathbf{P}\{z_n \geq c_n U\}$ which completes the proof. The inequality (3.49) is a special case of this result with $c_1 = c_2 = 0.5$.

We follow a similar approach for proving (3.50) where z_1 and z_2 are positive random variables. Consider the event $\mathcal{E}_A \triangleq \{z_1 < U_1 \text{ and } z_2 > U_2\}$. If \mathcal{E}_A occurs, then $\mathcal{E}_B \triangleq \left\{\frac{z_1}{z_2} < \frac{U_1}{U_2}\right\}$ must also occur. Consequently, $\mathbf{P}\{(\mathcal{E}_B)^c\} \leq \mathbf{P}\{(\mathcal{E}_A)^c\}$, where $(\mathcal{E}_A)^c = \{z_1 \geq U_1 \text{ or } z_2 \leq U_2\}$. Using the union bound, we have $\mathbf{P}\{(\mathcal{E}_A)^c\} \leq \mathbf{P}\{z_1 \geq U_1\} + \mathbf{P}\{z_2 \leq U_2\}$.

A.2 Proof of Proposition 3.52

Before proving Proposition 3.52, we state the following lemma.

Lemma A.1 (Hoeffding's inequality for real-valued Gaussian sums) *Let $\mathbf{b} \in \mathbb{R}^N$ be fixed, and let $\boldsymbol{\epsilon} \in \mathbb{R}^N$ be a random vector whose N entries are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma^2)$. Then, for any $u > 0$,*

$$\mathbf{P}\left\{\left|\sum_{i=1}^N \epsilon_i b_i\right| \geq u\right\} \leq e^{-\frac{u^2}{2\sigma^2\|\mathbf{b}\|_2^2}}.$$

Proof First note that the random variable $\sum_{i=1}^N \epsilon_i b_i$ is also Gaussian with distribution $\mathcal{N}(0, \sigma^2 \|\mathbf{b}\|_2^2)$. Applying a Gaussian tail bound to this distribution yields the inequality [28].

■

Using the result of Lemma A.1, we can complete the proof of Proposition 3.52.

Let $b_i = r_i + q_i j \forall i$ where r_i is the real part of b_i and q_i is the imaginary part. Then we have

$$\begin{aligned}
\mathbf{P} \left\{ \left| \sum_{i=1}^N \epsilon_i b_i \right| \geq \|\mathbf{b}\|_2 u \right\} &= \mathbf{P} \left\{ \left| \left(\sum_{i=1}^N \epsilon_i r_i \right) + \left(\sum_{i=1}^N \epsilon_i q_i \right) j \right| \geq \|\mathbf{b}\|_2 u \right\} \\
&= \mathbf{P} \left\{ \left| \left(\sum_{i=1}^N \epsilon_i r_i \right) + \left(\sum_{i=1}^N \epsilon_i q_i \right) j \right|^2 \geq \|\mathbf{b}\|_2^2 u^2 \right\} \\
&\leq \mathbf{P} \left\{ \left(\sum_{i=1}^N \epsilon_i r_i \right)^2 \geq \frac{\|\mathbf{b}\|_2^2 u^2}{2} \right\} + \mathbf{P} \left\{ \left(\sum_{i=1}^N \epsilon_i q_i \right)^2 \geq \frac{\|\mathbf{b}\|_2^2 u^2}{2} \right\} \\
&= \mathbf{P} \left\{ \left| \sum_{i=1}^N \epsilon_i r_i \right| \geq \frac{\|\mathbf{b}\|_2 u}{\sqrt{2}} \right\} + \mathbf{P} \left\{ \left| \sum_{i=1}^N \epsilon_i q_i \right| \geq \frac{\|\mathbf{b}\|_2 u}{\sqrt{2}} \right\} \\
&\leq e^{-\frac{\|\mathbf{b}\|_2^2 u^2}{4\|\mathbf{r}\|_2^2 \sigma^2}} + e^{-\frac{\|\mathbf{b}\|_2^2 u^2}{4\|\mathbf{q}\|_2^2 \sigma^2}} \leq 2e^{-\frac{u^2}{4\sigma^2}},
\end{aligned}$$

where the first inequality uses Lemma 3.48 and the last inequality uses Lemma A.1 and the facts that $\frac{\|\mathbf{b}\|_2}{\|\mathbf{r}\|_2} \geq 1$ and $\frac{\|\mathbf{b}\|_2}{\|\mathbf{q}\|_2} \geq 1$.

A.3 Proof of Lemma 4.23

We can show that for $M > m$

$$\mathbf{E} \left[\hat{A}_1^T \hat{A}_2 \right] = \frac{1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}} \cdot \mathcal{T}(\mathbf{x}_2^1)_m^m. \tag{A.2}$$

Using a conservative lower bound for $\|\mathcal{T}(\mathbf{x}_2^1)_m^m\|_2 \geq \|\mathbf{x}_2^1\|_2$, from (A.2) we get

$$\|\mathbf{E} \left[\hat{A}_1^T \hat{A}_2 \right]\|_2 \geq \frac{\|\mathbf{x}_2^1\|_2}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}}. \tag{A.3}$$

We can derive an upper bound for $\|\mathbf{E} \left[\hat{A}_1^T \hat{A}_2 \right]\|_2$ using the Cauchy Interlacing Theorem [58]. Let $X_{1,2} =: \mathbf{E} \left[\hat{A}_1^T \hat{A}_2 \right]$. We have

$$\|X_{1,2}\|_2^2 = \max_i \lambda_i(X_{1,2}^T X_{1,2}) \leq \max_i \lambda_i(\tilde{X}_{1,2}^T \tilde{X}_{1,2}), \quad (\text{A.4})$$

where $\tilde{X}_{1,2}$ is $(2m-1) \times (2m-1)$ circulant matrix with

$$\tilde{\mathbf{x}}_2^1 = \frac{1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}} [0, \dots, 0, x_2^1(m-1), \dots, x_2^1(2), x_2^1(1)],$$

as its first row [3]. Since $\lambda_i(\tilde{X}_{1,2}^T \tilde{X}_{1,2}) = |\lambda_i(\tilde{X}_{1,2})|^2$, an upper bound for $\|X_{1,2}\|_2^2$ is provided by the maximum eigenvalue of $\tilde{X}_{1,2}$. Because $\tilde{X}_{1,2}$ is circulant, $\lambda_i(\tilde{X}_{1,2})$ simply equals the un-normalized $(2m-1)$ -length Discrete Fourier Transform (DFT) of the first row of $\tilde{X}_{1,2}$. As a result,

$$\lambda_i(\tilde{X}_{1,2}) = \frac{1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}} \sum_{k=1}^{m-1} x_2^1(k) e^{-j2\pi(i-1)k/(2m-1)}. \quad (\text{A.5})$$

From (A.5) and by applying the triangle inequality, we get

$$|\lambda_i(\tilde{X}_{1,2})| \leq \frac{1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}} \sum_{k=1}^{m-1} |x_2^1(k)|. \quad (\text{A.6})$$

Therefore, combining (A.4) and (A.6), we have

$$\|X_{1,2}\|_2 \leq \frac{1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}} \sum_{k=1}^{m-1} |x_2^1(k)| \leq \frac{\|\mathbf{x}_2^1\|_1}{\sqrt{1 + \|\mathbf{x}_2^1\|_2^2}}.$$

A.4 Proof of Theorem 5.8

Without loss of generality, assume $d = 0$ as the input delays do not affect the coherence of Φ . Using the definition of $\mu(\Phi)$, we can write $\mu(\Phi) = \|\boldsymbol{\mu}_\Phi\|_\infty$ where $\boldsymbol{\mu}_\Phi$ is a vector whose entries are all the normalized distinct inner products of the columns of Φ and $\|\cdot\|_\infty$ is the maximum absolute entry of a vector. From Jensen's inequality for convex functions ($\|\cdot\|_\infty$), we have

$$\mathbf{E}[\mu(\Phi)] = \mathbf{E}[\|\boldsymbol{\mu}_\Phi\|_\infty] \geq \|\mathbf{E}[\boldsymbol{\mu}_\Phi]\|_\infty.$$

First we look at the numerator of the entries of $\boldsymbol{\mu}_\Phi$. From the definition of Φ , $\forall \phi_i, \phi_{i+s} \in \Phi_y$, $s \neq 0$,

$$\phi_i^T \phi_{i+s} = \sum_{t=t_0}^{t_0+M} y(t)y(t-s). \quad (\text{A.7})$$

Combining (A.7) with (5.9) and reordering the sums we have

$$\begin{aligned} \phi_i^T \phi_{i+s} &= \sum_{t=t_0}^{t_0+M} y(t)y(t-s) = \sum_{t=t_0}^{t_0+M} \left(\sum_{k=-\infty}^{\infty} h(k)u(t-k) \right) \left(\sum_{\ell=-\infty}^{\infty} h(\ell)u(t-\ell-s) \right) \\ &= \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h(k)h(\ell) \sum_{t=t_0}^{t_0+M} u(t-k)u(t-\ell-s). \end{aligned} \quad (\text{A.8})$$

Taking the expected value of both sides of (A.8), we have

$$\mathbf{E} [\phi_i^T \phi_{i+s}] = M \sum_{\ell=-\infty}^{\infty} h(\ell)h(\ell+s) \quad (\text{A.9})$$

where we used the fact that $\mathbf{E}[u(t-k)u(t-\ell-s)] = 1$ for $k = \ell + s$ and 0 otherwise.

Similarly, $\forall \phi_i \in \Phi_y, \forall \phi_{i+s} \in \Phi_u,$

$$\begin{aligned} \phi_i^T \phi_{i+s} &= \sum_{t=t_0}^{t_0+M} y(t)u(t-s) = \sum_{t=t_0}^{t_0+M} \left(\sum_{\ell=-\infty}^{\infty} h(\ell)u(t-\ell) \right) u(t-s) \\ &= \sum_{\ell=-\infty}^{\infty} h(\ell) \sum_{t=t_0}^{t_0+M} u(t-\ell)u(t-s). \end{aligned} \quad (\text{A.10})$$

Taking the expected value of both sides of (A.10), we have

$$\mathbf{E} [\phi_i^T \phi_{i+s}] = Mh(s). \quad (\text{A.11})$$

It is trivial to see that $\forall \phi_i, \phi_{i+s} \in \Phi_u$ with $s \neq 0$, $\mathbf{E} [\phi_i^T \phi_{i+s}] = 0$. Using concentration of measure inequalities, it can be shown that as $M \rightarrow \infty$, the entries of the denominator of μ_Φ are highly concentrated around their expected value [3]. We have $\forall \phi_i \in \Phi_u, \mathbf{E} [\|\phi_i\|_2^2] = M$ and $\forall \phi_i \in \Phi_y, \mathbf{E} [\|\phi_i\|_2^2] = M\|h\|_2^2$. By putting together (A.9) and (A.11) and applying the required column normalizations the proof is complete.

A.5 Proof of Theorem 5.11

We follow a similar argument to the proof of Theorem 5.8. Define $u_g(t) = \sum_{k=-\infty}^{\infty} g(k)u(t-k)$ and $y_g(t) = \sum_{k=-\infty}^{\infty} g(k)y(t-k)$. Then we have

$$\sum_{t=t_0}^{t_0+M} y_g(t)u_g(t-s) = \sum_{t=t_0}^{t_0+M} \left(\sum_{k=-\infty}^{\infty} g(k)y(t-k) \right) \left(\sum_{\ell=-\infty}^{\infty} g(\ell)u(t-\ell-s) \right) \quad (\text{A.12})$$

and by taking the expected value of both sides of (A.12), we get

$$\mathbf{E} \left[\sum_{t=t_0}^{t_0+M} y_g(t)u_g(t-s) \right] = M \sum_{\ell=-\infty}^{\infty} g(\ell)f(\ell+s),$$

where $f = g * h$. In a similar way,

$$\begin{aligned} \mathbf{E} \left[\sum_{t=t_0}^{t_0+M} y_g(t)y_g(t-s) \right] &= M \sum_{k=-\infty}^{\infty} f(k)f(k+s), \\ \mathbf{E} \left[\sum_{t=t_0}^{t_0+M} u_g(t)y_g(t-s) \right] &= M \sum_{k=-\infty}^{\infty} f(k)g(k+s), \\ \mathbf{E} \left[\sum_{t=t_0}^{t_0+M} u_g(t)u_g(t-s) \right] &= M \sum_{k=-\infty}^{\infty} g(k)g(k+s). \end{aligned}$$