Denoising and Decomposition of Moment Sequences using Convex Optimization

by

Badri Narayan Bhaskar

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Electrical and Computer Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2013

Date of final oral examination: 04/24/2013

The dissertation is approved by the following members of the Final Oral Committee:
John A. Gubner, Professor, Electrical and Computer Engineering
Robert D. Nowak, Professor, Electrical and Computer Engineering
Benjamin H. Recht, Professor, Computer Sciences Dept.
Barry D. Van Veen, Professor, Electrical and Computer Engineering
Stephen J. Wright, Professor, Computer Sciences Dept.

To my loving wife.

Acknowledgments

It is difficult to properly convey my deep sense of gratitude to the many wonderful people who have helped and enriched my life during my stay at UW. Suffice it to say I must be very blessed.

I consider myself very lucky to have Prof. Ben Recht as my advisor. He has a boundless enthusiasm for research and is a constant source of inspiration and energy. I would like to thank him for his excellent academic mentoring and his readiness to help with any problem. Without his pragmatism and brilliance, I doubt I would have spent my time fruitfully and I am grateful to him for setting a great example. He has fostered a unique collaborative atmosphere and I will cherish my stay at the wonderful Wisconsin Institute of Discovery (WID) for a long time to come.

The work I present in the thesis arose from a collaboration with Prof. Ben Recht and postdoctoral students Dr. Gongguo Tang and Dr. Parikshit Shah. I will definitely remember the many wonderful discussions I have had with them and their comradeship, and I wish them a wonderful future. My peers Gautam, Matt, Sumeet, Kevin, Aniruddha, Okan, Krishna, Victor, Charles, Nikhil and Zach made my stay at WID very enjoyable. I would also like to thank Laura, Aarti, Raman and Waheed for their friendship and their invaluable help during difficult times.

I would like to thank my committee for their support and guidance throughout. Prof. Barry Van Veen taught a very entertaining course which piqued my interest in Estimation Theory and I would like to thank him for his unique linear algebra perspective of it. Thanks to Prof. Nowak for engaging me with thoughtful discussions, offering excellent research courses, and especially also for building a wonderful team of students I've had the fortune of interacting with. I always felt very comfortable stopping by Prof. Steve Wright's office when I have a

question and I did so numerous times knowing that I can get a precise and quick answer. Thank you for sharing your expertise.

I took numerous courses taught by Prof. John Gubner and have always thoroughly enjoyed his wonderfully precise lectures. I admire his patience and readiness to provide a thorough answer to questions and would like to especially thank him for teaching me to be rigorous and meticulous. I owe Prof. Gubner many thanks for providing excellent advice, support, encouragement and giving me the gift of his time. I also feel especially indebted to Prof. Amir Assadi, Prof. Brian Gould and Prof. Ross Barmish for their encouragement, support and advice.

I could not have done anything without the unwavering faith my family has had in me. Thank you Amma, Appa and Karoo for believing in me and supporting me throughout. Last, but not the least, I have much to thank my wife for, who had to endure the several ups and downs of my life. Thank you for your unconditional love, your many selfless sacrifices, advice, encouragement and trust.

Contents

List of Figures

Abstract

Many high dimensional phenomena observed in applications are simple and can be approximated by a small combination of a potentially infinite number of building blocks or atoms. It is possible to estimate such simple objects robustly from a limited number of noisy measurements. Atomic norm regularization proposed in this thesis is a convex optimization problem that can be used for deriving efficient estimators of such high dimensional structures in a large number of cases.

This thesis provides a general approach to regularization using an atomic norm penalty which unifies previous literature on high dimensional statistics. We will revisit two fundamental problems in signal processing and systems theory – line spectral estimation and system identification, which are classically treated as nonlinear parameter estimation problems. We will see that the convex approach proposed in this thesis can provide a principled way of tackling these problems and provide optimal theoretical guarantees in the presence of noise. In contrast, parametric approaches often need to estimate the number of atoms or the model order and need heuristics to robustify nonlinear estimation.

The approach in this thesis can be thought of as a generalization of the Lasso estimator for handling continuous infinite dimensional sparse recovery problems. For the problem of line spectral estimation, I will provide efficient algorithms based on an exact semidefinite characterization of the proposed estimator and also more generally show that discretization provides a scalable alternative to approximate the solution for a number of problems.

1 Introduction

We live in a world of data abundance. Due to advancements in data collection, and massive storage capabilities, we now have a dizzying amount of high dimensional data to analyze. The task of the data scientist is to infer a simple model to describe the data. Fortunately, many naturally occurring phenomena often have a simple structure. This allows us to efficiently represent them using a few building blocks or features, even if they are observed in a high ambient dimension. The challenge is to exploit the simplicity of these objects and recover them robustly from limited measurements. This can be a formidable task even for simple instances of the problem.

For example, imagine a data series composed of n measurements that can be described by a linear model, in terms of a large number $p \gg n$ of potential predictor variables or features. Finding the linear model from these limited measurements is a hopelessly underdetermined problem. However, when the linear model is *simple* and it is known a priori that only a small subset of the features are actually active, the problem of determining the linear model is no longer ill posed. A simple description of the data series expresses the data series in terms of a few active features and corresponding feature weights. Our task of choosing the "right" model for the data can be cast as a combinatorial *feature selection* problem with the objective of finding the smallest subset of features that describe the data well. This is called the "small n, large p" problem in statistics and such datasets are ubiquitous - For example, a biologist might want to identify the active genes by looking at only a few samples of a microarray experiment which simultaneously measures thousands of genes. Machine learning practitioners might want to infer the active features from a large feature space with a relatively small number of measurements.

A naïve algorithm would explore all the possible subsets of features in increasing order

of size till we find a solution. There does not appear to be an efficient algorithm for this as there are an exponential number of subsets to choose from. In fact, this selection problem is provably NP-HARD[1] [75]. This does not however preclude the possibility of designing efficient algorithms that work on *most* instances. In fact, there is now a large body of literature on the theoretical understanding of the surprising success of convex relaxation methods in efficiently solving the sparse recovery problem most of the time.

Instead of the hard combinatorial problem of directly minimizing sparsity of the feature weights, the idea of using the ℓ_1 norm of the feature weights as a convex proxy was known to several early practitioners in Geophysics and seismology [35, 98, 68, 86]. The authors were also aware of the robustness of ℓ_1 regularization to outliers and noise. This was introduced in statistics as a sparsity inducing regularization in [100] and in Signal Processing as a means of exact and robust decomposition by [31] and subsequently studied by Donoho and his coworkers [42, 41]. Since the seminal publications of [17, 74], which rigorously established the exact recovery and model selection properties, the study of convex penalties for feature selection has been a subject of intense research. This theory was extended to other high dimensional structures including group sparse vectors[107] and low rank matrices [84].

The unifying theme in different structures in high dimension is some notion of simplicity like sparsity for vectors or rank for matrices. This thesis builds on a recent publication [30] which discusses the notion of simple objects as a sparse combination of atoms from a possibly infinite dictionary. This unifies several related problems in sparse recovery and approximation. Simple objects may be expressed as a sparse linear combination of a few basic atoms drawn from a possibly infinite dictionary of atoms. For instance, a sparse vector in high dimension is a combination of a few canonical unit vectors. A low rank matrix is a combination of a few

[1]Informally, this means that it is at least as hard as a large number of well known problems in complexity theory, widely believed to have no algorithm that can solve them in a time proportional to any polynomial function of the number of observations.

rank-1 matrices. A signal with a finite discrete spectrum is a combination of a few frequencies. The authors propose using "atomic norm" as a convex heuristic to recover simple models from linear measurements. The atomic norm may be thought of as a convex proxy to the combinatorial objective function that arises naturally in sparse recovery problems.

Furthermore, the atomic norm framework allows us to naturally extend convex relaxations to work on infinite dimensional objects which cannot be handled satisfactorily using standard finite dimensional Lasso. We will see how to denoise with atomic norms and then apply these ideas to revisit two fundamental signal processing problems – line spectral estimation and system identification, from the perspective of convex methods. I will also discuss efficient algorithms for these problems.

Line Spectral Estimation involves estimating frequencies and amplitudes from a limited number of noisy measurements. This is a fundamental problem in signal processing and there are a number of classical algorithms dating back to Prony. Surprisingly, using convex methods in this thesis, we can empirically outperform these classical techniques and also theoretically show near minimax optimal performance.

Another foundational problem in Signals and Systems theory is that of inferring a linear system from observed measurements. I will show that we can robustly recover a low order system from a limited set of noisy measurements, using the convex methods discussed in the thesis. We will see that it compares favorably in terms of prediction errors to the popular subspace ID method.

Contributions and Organization

In **Chapter 2**, I will present the atomic norm framework in depth and show how it unifies many linear inverse problems in high dimensional statistics. Extending this framework, I will

introduce a general regularized estimator using the atomic norm penalty which we will call Atomic Norm Soft Thresholding (AST). This may be thought of as an infinite dimensional version of the Lasso [100]. I will first establish universal properties of the estimator and indicate when accelerated convergence rates are possible. The choice of the regularization parameter for AST depends upon extremal properties of the dual atomic norm of noise, and we will examine general techniques for estimating the regularization parameter.

In **Chapter 3**, I show how to apply atomic norm soft thresholding estimator to denoise line spectral signals. I will show that this produces a consistent estimate for all signals which holds universally with very little assumptions on the measurement model. By exploiting properties of a dual certificate constructed in [20, 19], we will see that we can achieve accelerated convergence rates when the frequencies in the line spectral signal are well separated. As is the case for coherent designs using Lasso, this is nearly minimax optimal. This result may be thought of as a local version of coherence – Although our dictionary is highly coherent, as long as the signal we wish to recover is composed of relatively incoherent frequencies, it is possible to recover it robustly.

I will also show that the frequencies localized by AST tend to be near the true frequencies. With extensive experiments, I will demonstrate that the proposal in this thesis outperforms several classical line spectral estimation algorithms.

In **Chapter 4**, I will show how these techniques can be adapted for the System Identification problem. I will describe a general regularization problem for different kinds of linear measurements of the system. For the special case of frequency samples, I will show how to derive finite sample guarantees on the $\mathcal{H}^2$ prediction error of the transfer function of the linear system using AST.

Finally, in **Chapter 5**, I will discuss algorithms for AST. It turns out that we can efficiently solve AST as long as we have a reasonably efficient algorithm to test membership in the unit

ball of the atomic norm. For the case of Fourier measurements, the atomic norm balls can be characterized by a semidefinite program (SDP), which can be readily solved using any of the various SDP solvers. While the positive case is classical and is well known, the general case is a non-trivial extension. I will describe a fast parallelizable algorithm for the SDP using Alternating Directions Method of Multipliers (ADMM), and also provide an alternative efficient discretized version of AST which can be used in the absence of semidefinite characterizations, and in general for large problem sizes. This boils down to solving a Lasso problem on a grid. I will show the convergence of the Lasso solution which provides justification for discretization and Lasso on a grid as a general computational strategy.

2 Simple Models and Atomic Norms

The foundation that underlies the techniques discussed in this chapter is the work on *atomic norms* for linear inverse problems in [30]. In this work, the authors describe how to reconstruct models that can be expressed as sparse linear combinations of *atoms* from some basic set $\mathcal{A}$. The set $\mathcal{A}$ can be very general and is not assumed to be finite. For example, if the signal is known to be a low rank matrix, $\mathcal{A}$ could be the set of all unit norm rank-1 matrices, since a low rank matrix can indeed be written as a sparse combination of such rank-1 atoms.

I will first review the notion of simple models defined in [30] and describe how this generalizes various notions of sparsity and structure. Then, we will see how to use an atomic norm penalty to denoise a signal known to be a sparse nonnegative combination of atoms from a set $\mathcal{A}$. Atomic norms provide a natural convex penalty function for discouraging specialized notions of complexity. These norms generalize the ℓ_1 norm for sparse vector estimation [18] and the nuclear norm for low-rank matrix reconstruction [84, 21].

The first contribution described in this chapter, is an abstract theory of denoising with atomic norms. I show a unified approach to denoising with the atomic norm that provides a standard approach to computing low mean-squared-error (MSE) estimates. We will see how certain Gaussian statistics and geometrical quantities of particular atomic norms are sufficient to bound estimation rates with these penalty functions. This approach is essentially a generalization of the Lasso [100, 31] to infinite dictionaries.

Organization of this chapter

The denoising problem is obtaining an estimate $\hat{x}$ of the signal $x^\star$ from $y = x^\star + w$, where w is additive noise. Let us make the structural assumption that $x^\star$ is a sparse nonnegative combination of points from an arbitrary, possibly infinite set $\mathcal{A} \subset \mathbb{C}^n$. This assumption is very

expressive and generalizes many notions of sparsity [30]. The atomic norm $\|\cdot\|_{\mathcal{A}}$, introduced in [30], is a penalty function specially catered to the structure of $\mathcal{A}$ as we shall examine in depth in Section 2.1, and is defined as:

$$\|x\|_{\mathcal{A}} = \inf\left\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\right\}. \tag{2.1}$$

where $\mathrm{conv}(\mathcal{A})$ is the convex hull of points in $\mathcal{A}$. Then, we will look at the denoising performance of an estimate that uses the atomic norm to encourage sparsity in $\mathcal{A}$.

Decomposition. Section 2.2 considers vectors $x^\star$ that may be written as a sparse nonnegative combination of elements from the atomic set $\mathcal{A}$ and asks how one might certify that $x^\star$ has a unique sparsest decomposition in terms of the atoms. We will see how the existence of certain vectors in the dual space can reveal the composing atoms and certify uniqueness of the sparsest decomposition under some mild technical conditions.

Denoising. Section 2.3 characterizes the performance of the estimate $\hat{x}$ that solves

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\|x-y\|_2^2 + \tau\|x\|_{\mathcal{A}}. \tag{2.2}$$

where τ is an appropriately chosen regularization parameter, and $y = x^\star + w$ is a a vector of noisy measurements. I will show an upper bound on the MSE of the estimate when the noise statistics are known. Before stating the theorem, note that the dual norm $\|\cdot\|_{\mathcal{A}}^*$, corresponding to the atomic norm, is given by

$$\|z\|_{\mathcal{A}}^* = \sup_{a\in\mathcal{A}} \langle z, a\rangle,$$

where $\langle x, z\rangle = \mathfrak{Re}(z^* x)$ denotes the real inner product.

Theorem 2.1 (Universal Denoising Guarantee)**.** *Suppose we observe the signal* $y = x^\star + w$ *where* $x^\star \in \mathbb{C}^n$ *is a sparse nonnegative combination of points in* $\mathcal{A}$*. The estimate* $\hat{x}$ *of* $x^\star$ *given by the solution of the atomic soft thresholding problem* (2.2) *with* $\tau \geq \mathbb{E}\|w\|_{\mathcal{A}}^*$ *has the expected (per-element) MSE*

$$\frac{1}{n}\,\mathbb{E}\|\hat{x} - x^\star\|_2^2 \leq \frac{\tau}{n}\|x^\star\|_{\mathcal{A}}.$$

This theorem implies that when $\mathbb{E}\|w\|_{\mathcal{A}}^*$ is $o(n)$, the estimate $\hat{x}$ is consistent.

Choosing the regularization parameter. In Section 2.4, I will discuss the choice of regularization parameter τ. The lower bound on τ is in terms of the expected dual norm of the noise process w, equal to

$$\mathbb{E}\|w\|_{\mathcal{A}}^* = \mathbb{E}[\sup_{a \in \mathcal{A}} \langle w, a \rangle].$$

That is, the optimal τ and achievable MSE can be estimated by studying the extremal values of the stochastic process indexed by the atomic set $\mathcal{A}$.

Accelerated Convergence Rates The MSE rates given by Theorem 2.1 are not the best possible. Section 2.5, we will see a condition on the atomic sets that enables accelerate convergence rates. This is a generalization of the weak compatibility criterion [53] and unifies the condition for fast rates for several atomic sets. In particular, under this assumption, I will show that we can recover fast convergence rates published in literature for sparse vectors and low rank matrices.

2.1 Preliminaries

The notion of atomic sets and simple models introduced in [30] subsumes several structures in high dimensional geometry, including sparse vectors, low rank matrices, discrete measures,

linear systems of small order. The authors also define a notion of an *atomic norm* as a natural convex penalty for encouraging simplicity with respect to these structures. In this section, we shall review these preliminaries.

Let F be $\mathbb{C}$ or $\mathbb{R}$. The *atomic set* $\mathcal{A}$ is a possibly infinite subset of F^n and every signal we shall consider is composed of a combination of *atoms* from this dictionary. A target signal $x^\star \in F^n$ is *simple*, if it can be written as a nonnegative combination of small subset $T \subset \mathcal{A}$ of atoms. Call such a T, a *support* of $x^\star$. If the number of elements $k = |T|$ in the support is small relative to the dimension n of the signal, call such an $x^\star$ a (k) *simple* combination of atoms.

Allowing the dictionary or atomic set to be infinite offers tremendous flexibility in the kind of high dimensional structures that can be modeled. To see this, we consider specific instances of the setup, that will serve as motivation for this framework. In subsequent chapters, we will revisit some of these examples.

2.1.1 Examples

Estimating Nearly Black Objects Consider the estimation of the *nonnegative* vector $x^\star \in \mathbb{R}^m$ from the noisy observations $y = x^\star + w$ with the assumption that $x^\star$ is nearly black — i.e., most of its entries are zero. This problem was considered in [46, 61], in order to shed light on the superresolution properties of the maximum entropy estimator for nearly black MRI images. In this framework, such an $x^\star$ is a sparse combination of $\mathcal{A} = \{\pm e_1, \ldots, \pm e_n\}$ where e_i is the ith canonical unit vector.

Compressed Sensing and Sparse Recovery Many phenomena observed in high dimensions are sparse in a suitable dictionary. For instance, natural images can be well approximated by a small combination of basis functions of a suitable wavelet transform. This means the actual

data is highly redundant and it is actually sparse when suitably transformed. Compression techniques can exploit this sparse representation and represent the same high dimensional vector with fewer bits with no loss of accuracy and thus reduce storage costs. The idea of compressed sensing also reduces acquisition costs by only taking a few random linear measurements of the high dimensional vector. When the representation is sparse, a high p dimensional vector can be recovered from a small number $n \ll p$ of linear measurements.

Let $\theta^\star \in \mathbb{R}^p$ be sparse in a known orthogonal basis $\Phi \in \mathbb{R}^{p\times p}$. Suppose we make n observations $x_i^\star = \langle \psi_i, \theta^\star \rangle$ where $\{\psi_i\}_{i=1}^n$ are randomly chosen Gaussian vectors. Then, $x^\star$ is a simple combination of atoms from the atomic set given by $\left\{ \pm \Psi^T \Phi_j \;\middle|\; j = 1, \ldots, p \right\}$ where $\Psi \in \mathbb{R}^{p\times n}$ is the matrix with ψ_i for its columns.

Low Rank Matrix Estimation and Matrix Completion The problem of finding a minimum rank matrix from incomplete data arises in a number of applications including collaborative filtering, system identification and document classification. The data available to us could be a fraction of the entries of the matrix, or in general any set of linear measurements of the matrix. Let $\mathcal{A}$ be the one dimensional manifold composed of all unit norm rank-1 matrices. Then a low rank matrix $x^\star$ can be expressed as a sparse combination of rank-1 matrices, and consequently the observations of this matrix can be expressed as a sparse combination of observations due to these rank-1 matrices. In this example, note that it is not sufficient to consider a finite atomic set, since the set of all low rank matrices are not sparse in any finite dictionary.

Line Spectral Estimation Line Spectral Estimation concerns with the recovery of signals whose spectrum consists only of a finite number of frequencies. In other words, line spectral signals are simply a finite mixture of complex exponentials. An important signal processing

task is to recover the frequencies and amplitudes of the complex exponentials from a limited number of time samples. For each frequency f in $[-W, W]$, let $a(f)$ denote the vector of observed time samples of a complex exponential with a frequency f. Then the set $\mathcal{A}$ of different observations $a(f)$ due to each frequency f comprises the atomic set for samples of line spectral signals.

System Identification An important task in modeling is the identification of a linear system with the minimum number of states from its input and output. Let $\mathcal{A}$ denote the set of transfer functions corresponding to single pole systems. Then an LTI system with small order has a transfer function which is a sparse combination of atoms from $\mathcal{A}$. Therefore the task of system identification reduces to finding the comprising atoms given linear measurements.

The last two examples can be seen as an instance of continuous sparse recovery problem, since there are continuously many atoms. In most previous literature, sparse recovery problems are analyzed only in discrete settings. In fact, in previous literature, authors have advocated a discretization approach which amounts to working with a finite set $\tilde{\mathcal{A}}$ instead of the infinite atomic set $\mathcal{A}$. However, a finer gridding of the atomic set results in a more coherent dictionary and the usual compressed sensing theoretical guarantees degrade with the grid size. We will see that we can bypass this by directly analyzing the continuous case using a different approach. It turns out that the coherence of the dictionary is not a fundamental limitation and that we can get increasingly accurate results by finer and finer discretizations.

In succeeding chapters, I will concentrate on the application of the atomic norm framework to these examples. It is not the goal of this section to provide an exhaustive catalogue of examples of the framework and I would refer the interested reader to [30], which contains several more interesting applications.

2.1.2 Atomic Norm

Definition 2.2 (Atomic Norm)**.** *The atomic norm $\|\cdot\|_{\mathcal{A}}$ corresponding to $\mathcal{A} \subset F^n$ is the Minkowski functional (also called the gauge function) associated with $\mathrm{conv}(\mathcal{A})$ (the convex hull of $\mathcal{A}$):*

$$\|x\|_{\mathcal{A}} = \inf \{t > 0 \mid x \in t \,\mathrm{conv}(\mathcal{A})\} . \tag{2.3}$$

The gauge function is a norm in F^n if $\mathrm{conv}(\mathcal{A})$ is compact, centrally symmetric, and contains a ball of radius ϵ around the origin for some $\epsilon > 0$. Nevertheless, we shall call it atomic norm even if it is not as a norm, as the authors in [30] do. When $\mathcal{A}$ is the set of unit norm 1-sparse elements in $\mathbb{C}^n$, the atomic norm $\|\cdot\|_{\mathcal{A}}$ is the ℓ_1 norm [18]. Similarly, when $\mathcal{A}$ is the set of unit norm rank-1 matrices, the atomic norm is the nuclear norm [84]. In [30], the authors showed that minimizing the atomic norm subject to equality constraints provided exact solutions of a variety of linear inverse problems with nearly optimal bounds on the number of measurements required.

While not necessary for the definition of the atomic norm, we will assume some weak regularity conditions in the rest of the thesis. These are satisfied for a number of examples and they allow us to phrase our theorems without restating assumptions explicitly each time. We will assume that the atomic set $\mathcal{A}$ satisfies the following properties:

1. No atom can be written as a conical combination of other atoms in $\mathcal{A}$. In other words, we assume that $a \notin \mathrm{conv}(\mathcal{A} \backslash \{a\})$ for every $a \in \mathcal{A}$. This guarantees that elements in $\mathcal{A}$ are the extreme points of the set $\mathrm{conv}(\mathcal{A})$.

2. The set $\mathcal{A}$ is a closed subset of F^n. This assumption is always true for finite sets.

3. The spark of a set of vectors $\mathcal{A} \subset F^n$ is defined as the smallest number σ such that there exists a subcollection of σ elements of $\mathcal{A}$ which are linearly dependent. We will assume

that the spark of $\mathcal{A}$ is n. As a consequence, any decomposition of a vector $x^\star$ into $< n/2$ atoms is necessarily unique [41].

4. We will assume that $\mathrm{conv}(\mathcal{A})$ has a non-empty interior to avoid degenerate cases.

2.1.3 Dual Atomic Norm

Corresponding to the atomic norm, we can define the dual atomic norm, which is given by

$$\|z\|_{\mathcal{A}}^* = \sup_{\|x\|_{\mathcal{A}} \leq 1} \langle x, z \rangle, \tag{2.4}$$

implying

$$\langle x, z \rangle \leq \|x\|_{\mathcal{A}} \|z\|_{\mathcal{A}}^*. \tag{2.5}$$

We shall assume throughout the thesis that $\langle x, y \rangle$ always stands for the real inner product $\mathfrak{Re}\, x^* y$ even if $x, y \in \mathbb{C}^n$, unless otherwise noted. The supremum in (2.4) is achievable, namely, for any x, there is a z that achieves equality. Since $\mathcal{A}$ contains all extremal points of $\{x : \|x\|_{\mathcal{A}} \leq 1\}$, we are guaranteed that the optimal solution will actually lie in the set $\mathcal{A}$ (see [9] for a proof). So, we can write

$$\|z\|_{\mathcal{A}}^* = \sup_{a \in \mathcal{A}} \langle a, z \rangle. \tag{2.6}$$

The dual norm will play a critical role throughout, as our asymptotic error rates will be in terms of the dual atomic norm of noise processes. The dual atomic norm also appears in the dual problem of (2.2).

2.2 Decomposition

A natural question is whether it is always possible to recover the composing atoms of a simple $x^\star$. As stated, the problem is ill-posed, for there could many decompositions. However, under our assumptions, there is a unique sparsest decomposition of $x^\star$ provided there is a decomposition that is at most $n/2$ sparse. So, the question is well posed as long as we start with such an $x^\star$.

The goal of atomic norm decomposition is to write $x^\star$ in terms of composing atoms such that the sum of the coefficients is the atomic norm. We will call such a decomposition an atomic norm achieving decomposition. While not always true, the decomposition that achieves the atomic norm is often the sparsest and this section will provide some geometric insight to when this is true. This section also describes a procedure to find such a decomposition using the dual.

2.2.1 Dual Certificate

A useful device is that of a dual certificate, which is one of the subgradients of the atomic norm. For a convex differentiable function f, the gradient defines a global linear under approximator for the function. Subgradients generalize this property to nonsmooth convex functions by characterizing all linear under approximators of the function. A vector q is said to be in the subgradient of f at x if for every y,

$$f(y) \geq f(x) + \langle q, y - x \rangle.$$

When f is differentiable at x, the set of subgradients at x is simply the singleton set containing the gradient at x. It is easy to verify that q is a subgradient of $\|\cdot\|_{\mathcal{A}}$ at $x^\star$ if and only if

$\langle q, x^\star \rangle = \|x^\star\|_{\mathcal{A}}$ and $\|q\|_{\mathcal{A}}^* \leq 1$. A decomposition of a vector $x^\star$ gives an upper bound for its atomic norm. Subgradients in the dual space can provide a dual certificate of the optimality of a decomposition if it produces a matching lower bound.

Definition 2.3 (Dual Certificate)**.** *A vector q is a dual certificate for the support $T \subset \mathcal{A}$ if $\langle q, a \rangle = 1$ for every $a \in T$ and $\langle q, a \rangle \leq 1$ whenever $a \notin T$. Furthermore, the vector q is called a strict dual certificate if $\langle q, a \rangle < 1$ for every $a \notin T$.*

This has an intuitive geometric interpretation, when all the vectors are real. By definition of the dual certificate, for every non-empty set T, $\|q\|_{\mathcal{A}}^* = \sup_{a \in \mathcal{A}} \langle q, a \rangle = 1$, which guarantees that all the atoms (and hence $\mathrm{conv}(\mathcal{A})$) lie on one side of a half plane determined by q, and also that atoms in T (and hence $\mathrm{conv}(T)$) intersect this hyperplane. Geometrically, this says that q is a supporting hyperplane for the exposed face $\mathrm{conv}(T)$ of the convex set $\mathrm{conv}(\mathcal{A})$. The following proposition indicates why this is important.

Proposition 2.4. *Suppose $x^\star$ can be written in terms of atoms in T, i.e., $x^\star = \sum_{a \in T} c_a a$ for some $\{c_a\} > 0$. If q is a dual certificate for T, then, $\sum_{a \in T} c_a a$ is an atomic norm achieving decomposition of $x^\star$ and q is a subgradient of the atomic norm at $x^\star$.*

Proof. Then, by the definition of atomic norm $\|x^\star\|_{\mathcal{A}} \leq \sum_{a \in T} c_a$. Now,

$$\langle q, x^\star \rangle = \sum_{a \in T} c_a \langle q, a \rangle = \sum_{a \in T} c_a$$

But

$$\langle q, x^\star \rangle \leq \|q\|_{\mathcal{A}}^* \|x^\star\|_{\mathcal{A}} = \|x^\star\|_{\mathcal{A}}.$$

Combining these two equations, we get $\sum_{a \in T} c_a = \|x^\star\|_{\mathcal{A}} = \langle q, x \rangle$, which completes the proof. □

Conversely, the presence of a *strict* dual certificate which is also a subgradient of $\|\cdot\|_{\mathcal{A}}$ at $x^\star$ guarantees that $x^\star$ may be written as a combination of atoms in T.

Proposition 2.5. *Suppose q is a subgradient of $\|\cdot\|_{\mathcal{A}}$ at $x^\star \in \operatorname{cone}(\mathcal{A})$ and is a strict dual certificate for $T \subset \mathcal{A}$, then $x^\star$ has an atomic norm achieving decomposition in terms of atoms from T, i.e., there exists $c_a > 0$ such that $x^\star = \sum_{a\in T} c_a a$ with $\|x^\star\|_{\mathcal{A}} = \sum_{a\in T} c_a$.*

Under our assumptions (that $\mathcal{A}$ has full spark), we are guaranteed that this is the unique sparsest decomposition of $x^\star$ in terms of the atoms provided $|T| < n/2$. This is especially useful as it allows us to determine a support of $x^\star$ purely in terms of the properties of the dual. For a generic set, there is no guarantee that the atomic norm achieving decomposition is $n/2$ sparse and therefore that it recovers the *correct* support. However, we have a rich theory now that this is indeed true for many interesting cases[40, 18, 25].

Returning to our geometric interpretation, the existence of a dual certificate guarantees that $x^\star/\|x^\star\|_{\mathcal{A}}$ is in an exposed face of $\operatorname{conv}(\mathcal{A})$. Recall that a face in convex geometry is considered simplicial if every element in the face can be written as a unique convex combination. Due to our assumptions on $\mathcal{A}$, the faces are indeed simplicial and there is a unique way of writing $x^\star$ as a combination of vertices in the face. Furthermore, When the face is also low dimensional, the decomposition of $x^\star$ is unique. So, we can always recover the support of a sparse vector provided all sparse combinations lie on low dimensional exposed simplicial faces. Again, while this may not always be true, it holds with large probability for random constructions of the atomic set and we refer the interested reader to [44] for this geometric interpretation of sparse recovery.

In practice, it may be hard to find a *strict* dual certificate by optimization. However, if $\mathcal{A}$ has full spark like we assumed, any dual certificate will suffice to find a set of atoms for a decomposition that achieves the atomic norm. Given a simple $x^\star$, our recipe involves solving

the semi-infinite program

$$\begin{aligned} &\underset{q}{\text{maximize}} \ \langle q, x^\star \rangle \\ &\text{subject to } \langle q, a \rangle \leq 1, \text{ for every } a \in \mathcal{A}. \end{aligned} \tag{2.7}$$

to obtain a solution $\hat{q}$ which is one of the dual certificates of the support and a subgradient of $\|\cdot\|_{\mathcal{A}}$ at $x^\star$. However, $\langle q, a \rangle$ is 1 only for at most n atoms under our assumptions. In fact, if $\langle q, a \rangle$ has the same value for $n+1$ atoms $a_1, \ldots, a_{n+1}$, we have $\langle q, a_{i+1} - a_i \rangle = 0$ for $i = 1, \ldots, n$ and using the full spark assumption, we can conclude that q is identically zero. This means q is in fact a strict dual certificate for a support comprising at most n atoms which certifies that the atomic norm achieving decomposition can be composed in terms of the atoms where $\langle q, a \rangle = 1$. The coefficients of the decomposition can be determined by solving a linear system.

2.3 Denoising

Now, let us look at a scheme that is robust to measurement noise. To set up the atomic norm denoising problem, suppose we observe a signal $y = x^\star + w$ where w is a noise vector and that we know *a priori* that $x^\star$ can be written as a linear combination of a few atoms from $\mathcal{A}$. One way to estimate $x^\star$ from these observations would be to search over all short linear combinations from $\mathcal{A}$ to select the one which minimizes $\|y - x\|_2$. However, this could be formidable: even if the set of atoms is a finite collection of vectors, this problem is the NP-hard SPARSEST VECTOR problem [75].

On the other hand, the problem (2.2) is convex, and reduces to many familiar denoising strategies for particular $\mathcal{A}$. The mapping from y to the optimal solution of (2.2) is called the proximal operator of the atomic norm applied to y, and can be thought of as a soft thresholded

version of y. Indeed, when $\mathcal{A}$ is the set of 1-sparse atoms, the atomic norm is the ℓ_1-norm, and the proximal operator corresponds to *soft-thresholding* y by element-wise shrinking towards zero [45]. Similarly, when $\mathcal{A}$ is the set of rank-1 matrices, the atomic norm is the nuclear norm and the proximal operator shrinks the singular values of the input matrix towards zero.

We now establish some universal properties about the problem (2.2). First, we collect a simple consequence of the optimality conditions in a lemma:

Lemma 2.6 (Optimality Conditions)**.** *$\hat{x}$ is the solution of* (2.2) *if and only if*
(i) $\|y - \hat{x}\|_{\mathcal{A}}^* \leq \tau$, (ii) $\langle y - \hat{x}, \hat{x} \rangle = \tau \|\hat{x}\|_{\mathcal{A}}$.

Proof. The function $f(x) = \frac{1}{2}\|y - x\|_2^2 + \tau\|x\|_{\mathcal{A}}$ is minimized at $\hat{x}$, if for all $\alpha \in (0, 1)$ and all x,

$$f(\hat{x} + \alpha(x - \hat{x})) \geq f(\hat{x})$$

or equivalently,

$$\alpha^{-1}\tau \left(\|\hat{x} + \alpha(x - \hat{x})\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}}\right) \geq \langle y - \hat{x}, x - \hat{x} \rangle - \frac{1}{2}\alpha\|x - \hat{x}\|_2^2 \tag{2.8}$$

Since $\|\cdot\|_{\mathcal{A}}$ is convex, we have

$$\|x\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}} \geq \alpha^{-1} \left(\|\hat{x} + \alpha(x - \hat{x})\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}}\right),$$

for all x and for all $\alpha \in (0, 1)$. Thus, by letting $\alpha \to 0$ in (2.8), we note that $\hat{x}$ minimizes $f(x)$ only if, for all x,

$$\tau \left(\|x\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}}\right) \geq \langle y - \hat{x}, x - \hat{x} \rangle. \tag{2.9}$$

However if (2.9) holds, then, for all x

$$\frac{1}{2}\|y-x\|_2^2+\tau\|x\|_{\mathcal{A}} \geq \frac{1}{2}\|y-\hat{x}+(\hat{x}-x)\|_2^2+\langle y-\hat{x}, x-\hat{x}\rangle+\tau\|\hat{x}\|_{\mathcal{A}}$$

implying $f(x) \geq f(\hat{x})$. Thus, (2.9) is necessary and sufficient for $\hat{x}$ to minimize $f(x)$.

Note. *The condition (2.9) simply says that* $\tau^{-1}(y-\hat{x})$ *is in the subgradient of* $\|\cdot\|_{\mathcal{A}}$ *at* $\hat{x}$ *or equivalently that* $0 \in \partial f(\hat{x})$.

We can rewrite (2.9) as

$$\tau\|\hat{x}\|_{\mathcal{A}}-\langle y-\hat{x}, \hat{x}\rangle \leq \inf_x \{\tau\|x\|_{\mathcal{A}}-\langle y-\hat{x}, x\rangle\} \tag{2.10}$$

But by definition of the dual atomic norm,

$$\sup_x \{\langle z, x\rangle-\|x\|_{\mathcal{A}}\} = I_{\{w:\|w\|_{\mathcal{A}}^* \leq 1\}}(z) = \begin{cases} 0 & \|z\|_{\mathcal{A}}^* \leq 1 \\ \infty & \text{otherwise.} \end{cases} \tag{2.11}$$

where $I_A(\cdot)$ is the convex indicator function. Using this in (2.10), we find that $\hat{x}$ is a minimizer if and only if $\|y-\hat{x}\|_{\mathcal{A}}^* \leq \tau$ and $\langle y-\hat{x}, \hat{x}\rangle \geq \tau\|\hat{x}\|_{\mathcal{A}}$. This proves the theorem. □

Lemma 2.7 (Dual Problem)**.** *The dual problem of* (2.2) *is*

$$\begin{aligned} &\underset{z}{\text{maximize}}\ \frac{1}{2}\left(\|y\|_2^2-\|y-z\|_2^2\right) \\ &\text{subject to } \|z\|_{\mathcal{A}}^* \leq \tau. \end{aligned}$$

The dual problem admits a unique solution $\hat{z}$ *due to strong concavity of the objective function. The primal solution* $\hat{x}$ *and the dual solution* $\hat{z}$ *are specified by the optimality conditions and there*

is no duality gap:

(i) $y = \hat{x} + \hat{z}$, (ii) $\|\hat{z}\|_{\mathcal{A}}^* \leq \tau$, (iii) $\langle \hat{z}, \hat{x} \rangle = \tau \|\hat{x}\|_{\mathcal{A}}$.

Proof. We can rewrite the primal problem (2.2) as a constrained optimization problem:

$$\begin{aligned} &\underset{x,u}{\text{minimize}} \, \frac{1}{2}\|y - x\|_2^2 + \|u\|_{\mathcal{A}} \\ &\text{subject to } u = x. \end{aligned}$$

Now, we can introduce the Lagrangian function

$$L(x, u, z) = \frac{1}{2}\|y - x\|_2^2 + \|u\|_{\mathcal{A}} + \langle z, x - u \rangle.$$

so that the dual function is given by

$$\begin{aligned} g(z) = \inf_{x,u} L(x, u, z) &= \inf_{x} \left(\frac{1}{2}\|y - x\|_2^2 + \langle z, x \rangle \right) + \inf_{u} \left(\tau \|u\|_{\mathcal{A}} - \langle z, u \rangle \right) \\ &= \frac{1}{2} \left(\|y\|_2^2 - \|y - z\|_2^2 \right) - I_{\left\{ w : \|w\|_{\mathcal{A}}^* \leq \tau \right\}}(z). \end{aligned}$$

where the first infimum follows by completing the squares and the second infimum follows from (2.11). Thus the dual problem of maximizing $g(z)$ can be written as in (2.7).

The solution to the dual problem is the unique projection $\hat{z}$ of y on to the closed convex set $C = \{z : \|z\|_{\mathcal{A}}^* \leq \tau\}$. By projection theorem for closed convex sets, $\hat{z}$ is a projection of y onto C if and only if $\hat{z} \in C$ and $\langle z - \hat{z}, y - \hat{z} \rangle \leq 0$ for all $z \in C$, or equivalently if $\langle \hat{z}, y - \hat{z} \rangle \geq \sup_z \langle z, y - \hat{z} \rangle = \tau \|y - \hat{z}\|_{\mathcal{A}}$. These conditions are satisfied for $\hat{z} = y - \hat{x}$ where $\hat{x}$ minimizes $f(x)$ by Lemma 2.6. Now the proof follows by the substitution $\hat{z} = y - \hat{x}$ in the previous lemma. The absence of duality gap can be obtained by noting that the primal

objective function at $\hat{x}$,

$$f(\hat{x}) = \frac{1}{2}\|y - \hat{x}\|_2^2 + \langle \hat{z}, \hat{x} \rangle = \frac{1}{2}\|\hat{z}\|_2^2 + \langle \hat{z}, \hat{x} \rangle = g(\hat{z}).$$

□

Conclusions (ii) and (iii) of the lemma say that $\tau^{-1}\hat{z}$ is a subgradient of the atomic norm at $x^\star$, where $\hat{z}$ is the solution to the dual problem (2.7). So, a straightforward corollary of Proposition 2.5 is a certificate of the support of the solution to (2.2):

Corollary 2.8 (Dual Certificate of Support)**.** *Suppose for some $S \subset \mathcal{A}$, $\hat{z}$ is a solution to the dual problem (2.7) satisfying*

1. $\langle \hat{z}, a \rangle = \tau$ *whenever* $a \in S$,
2. $|\langle \hat{z}, a \rangle| < \tau$ *if* $a \notin S$.

Then, any solution $\hat{x}$ of (2.2) *admits a decomposition $\hat{x} = \sum_{a \in S} c_a a$ with $\|\hat{x}\|_{\mathcal{A}} = \sum_{a \in S} c_a$.*

Thus the dual solution $\hat{z}$ provides a way to determine a decomposition of $\hat{x}$ into a set of elementary atoms that achieves the atomic norm of $\hat{x}$. In fact, one could evaluate the inner product $\langle \hat{z}, a \rangle$ and identify the atoms where the absolute value of the inner product is τ. When the signal-to-noise-ratio (SNR) is high, we expect that the decomposition identified in this manner should be close to the original decomposition of $x^\star$ under certain assumptions.

We are now ready to state a proposition which gives an upper bound on the MSE with the optimal choice of the regularization parameter.

Proposition 2.9. *If the regularization parameter $\tau > \|w\|_{\mathcal{A}}^*$, the optimal solution $\hat{x}$ of* (2.2) *has the MSE*

$$\frac{1}{n}\|\hat{x} - x^\star\|_2^2 \leq \frac{1}{n}\left(\tau\|x^\star\|_{\mathcal{A}} - \langle x^\star, w \rangle\right) \leq \frac{2\tau}{n}\|x^\star\|_{\mathcal{A}}. \tag{2.12}$$

Proof.

$$\|\hat{x} - x^\star\|_2^2 = \langle \hat{x} - x^\star, w - (y - \hat{x}) \rangle \tag{2.13}$$

$$= \langle x^\star, y - \hat{x} \rangle - \langle x^\star, w \rangle + \langle \hat{x}, w \rangle - \langle \hat{x}, y - \hat{x} \rangle$$

$$\leq \tau \|x^\star\|_{\mathcal{A}} - \langle x^\star, w \rangle + (\|w\|_{\mathcal{A}}^* - \tau)\|\hat{x}\|_{\mathcal{A}} \tag{2.14}$$

$$\leq (\tau + \|w\|_{\mathcal{A}}^*)\|x^\star\|_{\mathcal{A}} + (\|w\|_{\mathcal{A}}^* - \tau)\|\hat{x}\|_{\mathcal{A}} \tag{2.15}$$

where for (2.14) we have used Lemma 2.6 and (2.5). The theorem now follows from (2.14) and (2.15) since $\tau > \|w\|_{\mathcal{A}}^*$. The value of the regularization parameter τ to ensure the MSE is upper bounded thus, is $\|w\|_{\mathcal{A}}^*$. □

Example: Sparse Model Selection We can specialize our stability guarantee to Lasso [100] and recover known results. Let $\Phi \in \mathbb{R}^{n \times p}$ be a matrix with unit norm columns, and suppose we observe $y = x^\star + w$, where w is additive noise, and $x^\star = \Phi c^\star$ is an unknown k sparse combination of columns of Φ. In this case, the atomic set is the collection of columns of Φ and $-\Phi$, and the atomic norm is $\|x\|_{\mathcal{A}} = \min\{\|c\|_1 : x = \Phi c\}$. Therefore, the proposed optimization problem (2.2) coincides with the Lasso estimator [100]. This method is also known as Basis Pursuit Denoising [31]. If we assume that w is a gaussian vector with variance σ^2 for its entries, the expected dual atomic norm of the noise term, $\|w\|_{\mathcal{A}}^* = \|\Phi^* w\|_\infty$ is simply the expected maximum of p gaussian random variables. Using the well known result on the maximum of gaussian random variables [65], we have $\mathbb{E}\|w\|_{\mathcal{A}}^* \leq \sigma\sqrt{2\log(p)}$. If $\hat{x}$ is the denoised signal, we have from Theorem 2.1 that if $\tau = \mathbb{E}\|w\|_{\mathcal{A}}^* = \sigma\sqrt{2\log(p)}$,

$$\frac{1}{n}\,\mathbb{E}\|\hat{x} - x^\star\|_2^2 \leq \sigma \frac{\sqrt{2\log(p)}}{n}\|c^\star\|_1,$$

which is the stability result for Lasso reported in [56] assuming no conditions on Φ.

2.4 Dual Atomic Norm Bounds

As noted in Theorem 2.1, the optimal choice of the regularization parameter is dictated by the dual atomic norm of the noise process. To see why this is the case, let us consider the dual problem to Atomic Soft Thresholding, given by Lemma 2.7:

$$\begin{aligned} &\underset{z}{\text{minimize}} \ \|y - z\|_2 \\ &\text{subject to } \|z\|_{\mathcal{A}}^* \leq \tau. \end{aligned}$$

Using the optimality conditions in Lemma 2.7, we see that the primal solution $\hat{x}$ is a good estimate of the target $x^\star$ if the dual solution is a good estimate of the noise vector w. Thus τ should be proportional to noise and may be interpreted as a parameter controlling the amount of shrinkage towards origin. In fact, when $\tau > \|y\|_{\mathcal{A}}^*$, we have $\hat{z} = y$ and $\hat{x} = 0$. This suggests a choice of $\tau = \mathbb{E}\|w\|_{\mathcal{A}}^*$, which corresponds to the strongest mean-squared-error guarantee in Theorem 2.1. Specializing this to the case of sparse vectors in noise, we see that the recommendation $\tau = \mathbb{E}\|w\|_\infty \approx \sqrt{2\log(n)}$ coincides with the optimal tuning parameter in [45].

The quantity

$$\mathbb{E}\|w\|_{\mathcal{A}}^* = \mathbb{E}\sup_{a \in \mathcal{A}} \langle w, a \rangle$$

where $w \in \mathcal{N}(0, I_n)$ is called the Gaussian width of the atomic set $\mathcal{A}$. In some cases, the parameterization of the atoms in $\mathcal{A}$ allow the Gaussian width to be viewed as an extremum of a Gaussian process and thus this can be estimated using standard tools such as Dudley's inequality [66], or Talagrand's method of generic chaining [94].

2.5 Accelerated Convergence Rates

In this section, we provide conditions under which a faster convergence rate can be obtained for AST.

Proposition 2.10 (Fast Rates)**.** *Suppose the set of atoms $\mathcal{A}$ is centrosymmetric and $\|w\|_{\mathcal{A}}^*$ concentrates about its expectation so that $P(\|w\|_{\mathcal{A}}^* \geq \mathbb{E}\,\|w\|_{\mathcal{A}}^* + t) < \delta(t)$. For $\gamma \in [0,1]$, define the cone*

$$C_\gamma(x^\star, \mathcal{A}) = \mathrm{cone}(\{z : \|x^\star + z\|_{\mathcal{A}} \leq \|x^\star\|_{\mathcal{A}} + \gamma\|z\|_{\mathcal{A}}\}).$$

Suppose

$$\phi_\gamma(x^\star, \mathcal{A}) := \inf\left\{\frac{\|z\|_2}{\|z\|_{\mathcal{A}}} : z \in C_\gamma(x^\star, \mathcal{A})\right\} \tag{2.16}$$

is strictly positive for some $\gamma > \mathbb{E}\|w\|_{\mathcal{A}}^*/\tau$. Then

$$\|\hat{x} - x^\star\|_2^2 \leq \frac{(1+\gamma)^2\tau^2}{\gamma^2\phi_\gamma(x^\star, \mathcal{A})^2} \tag{2.17}$$

with probability at least $1 - \delta(\gamma\tau - \mathbb{E}\|w\|_{\mathcal{A}}^*)$.

Having the ratio of norms bounded below is a generalization of the Weak Compatibility criterion used to quantify when fast rates are achievable for the Lasso [53]. As shown in [53], this is a weak condition for fast MSE rates and generalizes the argument for Restricted Isometry[18], Restricted Eigenvalue [7] or Coherence conditions [24] which are often assumed in literature for deriving fast rates for the Lasso problem. One difference is that we define the corresponding cone C_γ where ϕ_γ must be controlled in parallel with the *tangent cones* studied in [30]. There, the authors showed that the mean width of the cone $C_0(x^\star, \mathcal{A})$ determined the number of random linear measurements required to recover $x^\star$ using atomic

norm minimization. In our case, γ is greater than zero, and represents a "widening" of the tangent cone. When $\gamma = 1$, the cone is all of $\mathbb{R}^n$ or $\mathbb{C}^n$ (via the triangle inequality), hence τ must be larger than the expectation to enable our proposition to hold.

Proof. Since $\hat{x}$ is optimal, we have,

$$\tfrac{1}{2}\|y - \hat{x}\|_2^2 + \tau\|\hat{x}\|_{\mathcal{A}} \leq \tfrac{1}{2}\|y - x^\star\|_2^2 + \tau\|x^\star\|_{\mathcal{A}}$$

Rearranging and using (2.5) gives

$$\tau\|\hat{x}\|_{\mathcal{A}} \leq \tau\|x^\star\|_{\mathcal{A}} + \langle w, \hat{x} - x^\star\rangle \tag{2.18}$$

$$\implies \tau\|\hat{x}\|_{\mathcal{A}} \leq \tau\|x^\star\|_{\mathcal{A}} + \|w\|_{\mathcal{A}}^*\|\hat{x} - x^\star\|_{\mathcal{A}}. \tag{2.19}$$

Since $\|w\|_{\mathcal{A}}^*$ concentrates about its expectation, with probability at least $1 - \delta(\gamma\tau - \mathbb{E}\|w\|_{\mathcal{A}}^*)$, we have $\|w\|_{\mathcal{A}}^* \leq \gamma\tau$ and hence $\hat{x} - x^\star \in C_\gamma$. Using (2.13), if $\tau > \|w\|_{\mathcal{A}}^*$,

$$\|\hat{x} - x^\star\|_2^2 \leq (\tau + \|w\|_{\mathcal{A}}^*)\|\hat{x} - x^\star\|_{\mathcal{A}} \leq \frac{(1+\gamma)\tau}{\gamma\phi_\gamma(x^\star, \mathcal{A})}\|\hat{x} - x^\star\|_2$$

So, with probability at least $1 - \delta(\gamma\tau - \mathbb{E}\|w\|_{\mathcal{A}}^*)$:

$$\|\hat{x} - x^\star\|_2^2 \leq \frac{(1+\gamma)^2\tau^2}{\gamma^2\phi_\gamma(x^\star, \mathcal{A})^2}$$

□

The main difference between (2.17) and (2.12) is that the MSE is controlled by τ^2 rather than $\tau\|x^*\|_{\mathcal{A}}$. As we will now see (2.17) provides minimax optimal rates for the examples of sparse vectors and low-rank matrices.

Example: Sparse Vectors in Noise Let $\mathcal{A}$ be the set of signed canonical basis vectors in $\mathbb{R}^n$. In this case, $\mathrm{conv}(\mathcal{A})$ is the unit cross polytope and the atomic norm $\|\cdot\|_{\mathcal{A}}$, coincides with the

ℓ_1 norm, and the dual atomic norm is the ℓ_∞ norm. Suppose $x^\star \in \mathbb{R}^n$ and $T := \mathrm{supp}(x^\star)$ has cardinality k. Consider the problem of estimating $x^\star$ from $y = x^\star + w$ where $w \sim \mathcal{N}(0, \sigma^2 I_n)$.

Proposition 2.11. *Let $\mathcal{A} = \{\pm e_1, \ldots, \pm e_n\}$, be the set of signed canonical unit vectors in $\mathbb{R}^n$. Suppose $x^\star \in \mathbb{R}^n$ has k nonzeros. Then $\phi_\gamma(x^\star, \mathcal{A}) \geq \frac{(1-\gamma)}{2\sqrt{k}}$.*

Proof. Let $z \in C_\gamma(x^\star, \mathcal{A})$. For some $\alpha > 0$ we have,

$$\|x^\star + \alpha z\|_1 \leq \|x^\star\|_1 + \gamma\|\alpha z\|_1$$

In the above inequality, set $z = z_T + z_{T^c}$ where z_T are the components on the support of T and z_{T^c} are the components on the complement of T. Since $x^\star + z_T$ and z_{T^c} have disjoint supports, we have,

$$\|x^\star + \alpha z_T\|_1 + \alpha\|z_{T^c}\|_1 \leq \|x^\star\|_1 + \gamma\|\alpha z_T\|_1 + \gamma\|\alpha z_{T^c}\|_1 \,.$$

This inequality implies

$$\|z_{T^c}\|_1 \leq \frac{1+\gamma}{1-\gamma}\|z_T\|_1$$

that is, z satisfies the null space property with a constant of $\frac{1+\gamma}{1-\gamma}$. Thus,

$$\|z\|_1 \leq \frac{2}{1-\gamma}\|z_T\|_1 \leq \frac{2\sqrt{k}}{1-\gamma}\|z\|_2$$

This gives the desired lower bound. We have therefore shown that in this case $\phi_\gamma(x^\star, \mathcal{A}) > \frac{(1-\gamma)}{2\sqrt{k}}$. We also have $\tau_0 = \mathbb{E}\|w\|_\infty \geq \sigma\sqrt{2\log(n)}$. Pick $\tau > \gamma^{-1}\tau_0$ for some $\gamma < 1$. Then, using

our lower bound for ϕ_γ in (2.17), we get a rate of

$$\frac{1}{n}\|\hat{x} - x^\star\|_2^2 = O\left(\frac{\sigma^2 k \log(n)}{n}\right) \tag{2.20}$$

for the AST estimate with high probability. This bound coincides with the minimax optimal rate derived by Donoho and Johnstone [43]. Note that if we had used (2.12) instead, our MSE would have instead been $O\left(\sqrt{\sigma^2 k \log n}\|x^\star\|_2/n\right)$, which depends on the norm of the input signal $x^\star$. □

Example: Low Rank Matrix in Noise Let $\mathcal{A}$ be the manifold of unit norm rank-1 matrices in $\mathbb{C}^{n\times n}$. In this case, the atomic norm $\|\cdot\|_{\mathcal{A}}$, coincides with the nuclear norm $\|\cdot\|_*$, and the corresponding dual atomic norm is the spectral norm of the matrix. Suppose $X^\star \in \mathbb{C}^{n\times n}$ has rank r, so it can be constructed as a combination of r atoms, and we are interested in estimating $X^\star$ from $Y = X^\star + W$ where W has independent $\mathcal{N}(0, \sigma^2)$ entries.

Proposition 2.12. *Let $\mathcal{A}$ be the manifold of unit norm rank-1 matrices in $\mathbb{C}^{n\times n}$. Suppose $X^\star \in \mathbb{C}^{n\times n}$ has rank r. Then $\phi_\gamma(X^\star, \mathcal{A}) \geq \frac{1-\gamma}{2\sqrt{2r}}$.*

Proof. Let $U\Sigma V^H$ be a singular value decomposition of $X^\star$ with $U \in \mathbb{C}^{n\times r}$, $V \in \mathbb{C}^{n\times r}$ and $\Sigma \in \mathbb{C}^{r\times r}$. Define the subspaces

$$T = \{UX + YV^H \ : \ X, Y \in \mathbb{C}^{n\times r}\}$$
$$T_0 = \{UMV^H \ : \ M \in \mathbb{C}^{r\times r}\}$$

and let $\mathcal{P}_{T_0}$, $\mathcal{P}_T$, and $\mathcal{P}_{T^\perp}$ be projection operators that respectively map onto the subspaces T_0, T, and the orthogonal complement of T. Now, if $Z \in C_\gamma(X^\star, \mathcal{A})$, then for some $\alpha > 0$, we

have

$$\|X^\star + \alpha Z\|_* \leq \|X^\star\|_* + \gamma\alpha\|Z\|_* \leq \|X^\star\|_* + \gamma\alpha\|\mathcal{P}_T(Z)\|_* + \gamma\alpha\|\mathcal{P}_{T^\perp}(Z)\|_*. \tag{2.21}$$

Now note that we have

$$\|X^\star + \alpha Z\|_* \geq \|X^\star + \alpha\mathcal{P}_{T_0}(Z)\|_* + \alpha\|\mathcal{P}_{T^\perp}(Z)\|_*$$

Substituting this in (2.21), we have,

$$\|X^\star + \alpha\mathcal{P}_{T_0}(Z)\|_* + \alpha\|\mathcal{P}_{T^\perp}(Z)\|_* \leq \|X^\star\|_* + \gamma\alpha\|\mathcal{P}_T(Z)\|_* + \gamma\alpha\|\mathcal{P}_{T^\perp}(Z)\|_*.$$

Since $\|\mathcal{P}_{T_0}(Z)\|_* \leq \|\mathcal{P}_T(Z)\|_*$, we have

$$\|\mathcal{P}_{T^\perp}(Z)\|_* \leq \frac{1+\gamma}{1-\gamma}\|\mathcal{P}_T(Z)\|_*.$$

Putting these computations together gives the estimate

$$\|Z\|_* \leq \|\mathcal{P}_T(Z)\|_* + \|\mathcal{P}_{T^\perp}(Z)\|_* \leq \frac{2}{1-\gamma}\|\mathcal{P}_T(Z)\| \leq \frac{2\sqrt{2r}}{1-\gamma}\|\mathcal{P}_T(Z)\|_F \leq \frac{2\sqrt{2r}}{1-\gamma}\|Z\|_F\,.$$

That is, we have $\phi_\gamma(X^\star, \mathcal{A}) \geq \frac{1-\gamma}{2\sqrt{2r}}$ as desired. □

Using this proposition, $\phi_\gamma(X^\star, \mathcal{A}) \geq \frac{1-\gamma}{2\sqrt{2r}}$. To obtain an estimate for τ, we note that the spectral norm of the noise matrix satisfies $\|W\| \leq 2\sqrt{n}$ with high probability [38]. Substituting these estimates for τ and ϕ_γ in (2.17), we get the minimax optimal MSE

$$\frac{1}{n^2}\|X - \hat{X}\|_F^2 = O\left(\frac{\sigma^2 r}{n}\right).$$

2.6 Conclusion

In this chapter, we defined atomic norms and the use of atomic norm penalty to denoise using AST (2.2) which may be thought of as a infinite dimensional version of Lasso. The chapter showed how the regularization parameter may be chosen for AST, and also a procedure to determine the composing atoms in the solution to AST. We also saw a universal convergence rate which holds for all atomic sets and all signals and saw general conditions under which a fast rate is possible. In the following chapters, we will use this framework and apply to the problems of line spectral estimation and system identification.

3 Line Spectrum Estimation

3.1 Introduction

Extracting the frequencies and relative phases of a superposition of complex exponentials from a small number of noisy time samples is a foundational problem in statistical signal processing. These *line spectral estimation* problems arise in a variety of applications, including the direction of arrival estimation in radar target identification [28], sensor array signal processing [64] and imaging systems [11] and also underlies techniques in ultra wideband channel estimation [73], spectroscopy [105], molecular dynamics [1], and power electronics [67]

Despite of hundreds of years of research on the fundamental problem of line spectrum estimation, there still remain several open questions in this area. This chapter addresses a central one of these problems: how well can we determine the locations and magnitudes of spectral lines from noisy temporal samples? We establish lower bounds on how well we can recover such signals and demonstrate that these worst case bounds can be nearly saturated by solving a convex programming problem. Moreover, we prove that the estimator approximately localizes the frequencies of the true spectral lines.

While polynomial interpolation using Prony's technique can estimate the frequency content of a signal *exactly* from as few as $2k$ samples if there are k frequencies, Prony's method is inherently unstable due to sensitivity of polynomial root finding. Several methods have been proposed to provide more robust polynomial interpolation [89, 85, 60] (for an extensive bibliography on the subject, see [91]), and these techniques achieve excellent noise performance in moderate noise. However, the denoising performance is often sensitive to the model order estimated, and theoretical guarantees for these methods are all asymptotic with no finite sample error bounds. Motivated by recent work on atomic norms [30], we will see how a

convex relaxation approach can denoise a mixture of complex exponentials, with theoretical guarantees of noise robustness and a better empirical performance than previous subspace based approaches.

Specializing the denoising results of the previous chapter to the line spectral estimation, I will provide mean-squared-error estimates for denoising line spectra with the atomic norm. The denoising algorithm amounts to soft thresholding the noise corrupted measurements in the atomic norm and so we may refer to the problem as *Atomic norm Soft Thresholding* (AST). Furthermore, it can be shown that AST achieves near minimax rates for estimating line spectral signals when the frequencies are reasonably well separated. We can give bounds on how well we can localize the frequencies using this technique.

3.1.1 Outline and summary of results

Denoising line spectral signals.

Let us specialize the results of the abstract denoising problem in the previous chapter to line spectral estimation in Section 3.2. Consider the continuous time signal $x^\star(t), t \in \mathbb{R}$ with a line spectrum composed of k unknown frequencies $\omega_1^\star, \ldots, \omega_k^\star$ bandlimited to $[-W, W]$. Then the Nyquist samples of the signal are given by

$$x_m^\star := x^\star\left(\tfrac{m}{2W}\right) = \sum_{l=1}^{k} c_l^\star e^{i2\pi m f_l^\star}, m = 0, \ldots, n-1 \tag{3.1}$$

where $c_1^\star, \ldots, c_k^\star$ are unknown *complex* coefficients and $f_l^\star = \frac{\omega_l^\star}{2W}$ for $l = 1, \ldots, k$ are the normalized frequencies. By swapping the roles of frequency and time or space, the signal model (3.1) also serves as a proper model for superresolution imaging where we aim to localize temporal events or spatial targets from noisy, low-frequency measurements [20, 19].

So, the vector $x^\star = [x_0^\star \ \cdots \ x_{n-1}^\star]^T \in \mathbb{C}^n$ can be written as a nonnegative linear combination of k points from the set of atoms

$$\mathcal{A} = \left\{ e^{i2\pi\phi}[1 \ e^{i2\pi f} \ \cdots \ e^{i2\pi(n-1)f}]^T, \ f \in [0,1], \phi \in [0,1] \right\}.$$

The set $\mathcal{A}$ can be viewed as an infinite dictionary indexed by the continuously varying parameters f and ϕ. When the number of observations, n, is much greater than k, $x^\star$ is k-sparse and thus line spectral estimation in the presence of noise can be thought of as a sparse approximation problem.

The first result is a global error rate that holds for line spectral signals by specializing the results in the previous chapter. In particular, we can apply AST and choose the regularization parameter for the strongest guarantee in Theorem 2.1 in terms of the expected dual norm of the noise. This can be explicitly computed for many noise models. For example, when the noise is Gaussian, we have the following theorem for the MSE:

Theorem 3.1. *Assume $x^\star \in \mathbb{C}^n$ is given by $x_m^\star = \sum_{l=1}^k c_l^\star e^{i2\pi m f_l^\star}$ for some unknown complex numbers $c_1^\star, \ldots, c_k^\star$, unknown normalized frequencies $f_1^\star, \ldots, f_k^\star \in [0,1]$ and $w \in \mathcal{N}(0, \sigma^2 I_n)$. Then the estimate $\hat{x}$ of $x^\star$ obtained from $y = x^\star + w$ given by the solution of atomic soft thresholding problem* (3.7) *with $\tau = \sigma\sqrt{n \log(n)}$ has the asymptotic MSE*

$$\frac{1}{n}\,\mathbb{E}\|\hat{x} - x^\star\|_2^2 \lesssim \sigma \sqrt{\frac{\log(n)}{n}} \sum_{l=1}^k |c_l^\star|.$$

It is instructive to compare this to the trivial estimator $\hat{x} = y$ which has a per-element MSE of σ^2. In contrast, Theorem 3.1 guarantees that AST produces a consistent estimate when $k = o\left(\sqrt{n/\log(n)}\right)$. While this rate holds for any line spectral signal, AST can perform considerably better when the frequencies are well separated.

Theorem 3.2. *Suppose the line spectral signal $x^\star$ is given by* (3.1) *and we observe n noisy consecutive samples $y_j = x_j^\star + w_j$ where w_j is i.i.d. complex Gaussian with variance σ^2. If the frequencies $\{f_l\}_{l=1}^k$ in $x^\star$ satisfy a minimum separation condition*

$$\min_{p \neq q} d(f_p, f_q) > 4/n \tag{3.2}$$

with $d(\cdot, \cdot)$ the distance metric on the torus, then we can determine an estimator $\hat{x}$ satisfying

$$\frac{1}{n}\|\hat{x} - x^\star\|_2^2 = O\left(\sigma^2 \frac{k \log(n)}{n}\right) \tag{3.3}$$

with high probability by solving a semidefinite programming problem.

Note that if we exactly knew the frequencies f_j, the best rate of estimation we could achieve would be $O(\sigma^2 k/n)$ [14]. This upper bound is merely a logarithmic factor larger than this rate. On the other hand, minimax theory can demonstrate that a logarithmic factor is unavoidable when the support is unknown. Hence, the proposed estimator is nearly minimax optimal.

It is instructive to compare this stability rate to the optimal rate achievable for estimating a sparse signal from a finite, discrete dictionary [22]. In the case that there are p incoherent dictionary elements, no method can estimate a k-sparse signal from n measurements corrupted by Gaussian noise at a rate less than $O(\sigma^2 \frac{k \log(p/k)}{n})$. In this problem, there are an infinite number of candidate dictionary elements and it is surprising that we can still achieve such a fast rate of convergence with our highly coherent dictionary. None of the standard techniques from sparse approximation can be immediately generalized to this case. Not only is the dictionary infinite, but also it does not satisfy the usual assumptions such as restricted eigenvalue conditions [7] or coherence conditions [24] that are used to derive stability results

in sparse approximation. Nonetheless, in terms of mean-square error performance, I will show results match those obtained when the frequencies are restricted to lie on a discrete grid.

In the absence of noise, polynomial interpolation can exactly recover a line spectral signal of k *arbitrary* frequencies with as few as $2k$ equispaced measurements. In the light of our minimum frequency separation requirement (3.2), why should one favor convex techniques for line spectral estimation? The stability result in this chapter coupled with minimax optimality establish that no method can perform better than convex methods when the frequencies are well-separated. And, while polynomial interpolation and subspace methods do not impose any resolution limiting assumptions on the constituent frequencies, these methods are empirically highly sensitive to noise. To the best of my knowledge, there is no result similar to Theorem 3.2 that provides finite sample guarantees about the noise robustness of polynomial interpolation techniques.

Localizing the frequencies using the Dual

The atomic formulation not only offers a way to denoise the line spectral signal, but also provides an efficient frequency localization method. After we obtain the signal estimate $\hat{x}$ by solving (3.7), we can also obtain the solution $\hat{z}$ to the dual problem as $\hat{z} = y - \hat{x}$. As we shall see in Corollary 1, the dual solution $\hat{z}$ both certifies the optimality of $\hat{x}$ and reveals the composing atoms of $\hat{x}$. For line spectral estimation, this provides an alternative to polynomial interpolation for localizing the constituent frequencies.

Indeed, when there is no noise, Candés and Fernandez-Granda showed the dual solution recovers these frequencies exactly under mild technical conditions [20]. This frequency localization technique is later extended in [95] to the random undersampling case to yield a compressive sensing scheme that is robust to basis mismatch. When there is noise, numerical simulations show that the atomic norm minimization problem (3.7) gives approximate

frequency localization.

I will theoretically characterize how well spectral lines can be localized from noisy observations. The frequencies estimated by any method will never exactly coincide with the true frequencies in the signal in the presence of noise. However, we can characterize the localization performance of our convex programming approach, and summarize this performance in Theorem 3.3.

Before stating the theorem, let us introduce a bit of notation. Define neighborhoods N_j around each frequency f_j in $x^\star$ by $N_j := \{f \in \mathbb{T} : d(f, f_j) \leq 0.16/n\}$. Also define $F = \mathbb{T} \backslash \cup_{j=1}^{k} N_j$ as the set of frequencies in $\mathbb{T}$ which are not near any true frequency. The letters N and F denote the regions that are *near* to and *far* from the true supporting frequencies. The following theorem summarizes the localization guarantees.

Theorem 3.3. *Let $\hat{x}$ be the solution to the same semidefinite programming (SDP) problem as referenced in Theorem 3.2 and $n > 256$. Let $\hat{c}_l$ and $\hat{f}_l$ form the decomposition of $\hat{x}$ into coefficients and frequencies, as revealed by the SDP. Then, there exist fixed numerical constants C_1, C_2 and C_3 such that with high probability*

i.) $\sum_{l:\hat{f}_l \in F} |\hat{c}_l| \leq C_1 \sigma \sqrt{\frac{k^2 \log(n)}{n}}$

ii.) $\sum_{l:\hat{f}_l \in N_j} |\hat{c}_l| \left\{ \min_{f_j \in T} d(f_j, \hat{f}_l) \right\}^2 \leq C_2 \sigma \sqrt{\frac{k^2 \log(n)}{n}}$

iii.) $\left| c_j - \sum_{l:\hat{f}_l \in N_j} \hat{c}_l \right| \leq C_3 \sigma \sqrt{\frac{k^2 \log(n)}{n}}$.

iv.) If for any frequency f_j, the corresponding amplitude $|c_j| > C_1 \sigma \sqrt{\frac{k^2 \log(n)}{n}}$, then with high probability there exists a corresponding frequency $\hat{f}_j$ in the recovered signal such that,

$$\left| f_j - \hat{f}_j \right| \leq \frac{\sqrt{C_2/C_1}}{n} \left(\frac{|c_j|}{C_1 \sigma \sqrt{\frac{k^2 \log(n)}{n}}} - 1 \right)^{-\frac{1}{2}}$$

Part (i) of Theorem 3.3 shows that the estimated amplitudes corresponding to frequencies far from the support are small. We rarely ever find any spurious frequencies in the far region, suggesting that our bound (i) is conservative. Parts (ii) and (iii) of the theorem show that in a neighborhood of each true frequency, the recovered signal has amplitude close to the true signal. Part (iv) shows that the larger a particular coefficient is, the better our method is able to estimate the corresponding frequency. In particular, note that if $|c_j| > 2C_1\sigma\sqrt{\frac{k^2\log(n)}{n}}$, then $\left|f_j - \hat{f}_j\right| \leq \frac{\sqrt{C_2/C_1}}{n}$. In all four parts, note that the localization error goes to zero as the number of samples grows.

Organization of this chapter

Section 3.2 describes how we can approach line spectral estimation using the framework of atomic norms. Section 3.3 describes how we can localize the frequencies using the dual problem. We will see a choice of the regularization parameter in 3.4 and to this end, derive nonasymptotic upper and lower bounds for the Gaussian width of the atomic set for line spectral estimation. Specializing the results of the previous chapter, we can derive a mean squared error rate that holds for all signals in Section 3.5. I present minimax lower bounds which show the best rate that can be achieved for well separated signals in 3.6. We will then see the proofs of Theorem 3.2 showing near minimax MSE and Theorem 3.3 showing frequency localization guarantees in Section 3.7. Section 3.8 contextualizes the results of this chapter in the canon of line spectral estimation and emphasize the advantages and shortcomings of prior art.

3.2 Denoising Line Spectral Signals

Suppose we wish to estimate the amplitudes and frequencies of a signal $x(t), t \in \mathbb{R}$ given as a mixture of k complex sinusoids:

$$x(t) = \sum_{l=1}^{k} c_l \exp(i2\pi f_l t)$$

where $\{c_l\}_{l=1}^k$ are unknown complex amplitudes corresponding to the k unknown frequencies $\{f_l\}_{l=1}^k$ assumed to be in the torus $\mathbb{T} = [0, 1]$. Such a signal may be thought of as a normalized band limited signal and has a Fourier transform given by a line spectrum:

$$\mu(f) = \sum_{l=1}^{k} c_l \delta(f - f_l) \tag{3.4}$$

Denote by $x^\star$ the $n = 2m + 1$ dimensional vector composed of equispaced Nyquist samples $\{x(j)\}_{j=-m}^m$ for $j = -m, \ldots, m$.

The goal of line spectral estimation is to estimate the frequencies and amplitudes of the signal $x(t)$ from the finite, noisy samples $y \in \mathbb{C}^n$ given by

$$y_j = x_j^\star + w_j$$

for $-m \leq j \leq m$, where $w_j \sim \mathcal{CN}(0, \sigma^2)$ is i.i.d. circularly symmetric complex Gaussian noise.

We can model the line spectral observations $x^\star = [x_{-m}^\star, \ldots, x_m^\star]^T \in \mathbb{C}^n$ as a sparse combination of atoms $a(f)$ which correspond to observations due to single frequencies. The atomic set in this case consists of samples of individual sinusoids, $a_{f,\phi} \in \mathbb{C}^n$, given by

$$a_{f,\phi} = e^{i2\pi\phi} \begin{bmatrix} 1 & e^{i2\pi f} & \cdots & e^{i2\pi(n-1)f} \end{bmatrix}^T . \tag{3.5}$$

The infinite set $\mathcal{A} = \{a_{f,\phi} : f \in [0,1], \phi \in [0,1]\}$ forms an appropriate collection of atoms for $x^\star$, since $x^\star$ in (3.1) can be written as a sparse nonnegative combination of atoms in $\mathcal{A}$. In fact, $x^\star = \sum_{l=1}^k c_l^\star a_{f_l^\star,0} = \sum_{l=1}^k |c_l^\star| a_{f_l^\star,\phi_l}$, where $c_l^\star = |c_l^\star| e^{i2\pi\phi_l}$.

The corresponding dual norm takes an intuitive form:

$$\|v\|_{\mathcal{A}}^* = \sup_{f,\phi} \langle v, a_{f,\phi} \rangle = \sup_{f\in[0,1]} \sup_{\phi\in[0,1]} e^{i2\pi\phi} \sum_{l=0}^{n-1} v_l e^{-2\pi i l f} = \sup_{|z|\leq 1} \left| \sum_{l=0}^{n-1} v_l z^l \right|. \tag{3.6}$$

In other words, $\|v\|_{\mathcal{A}}^*$ is the maximum absolute value attained on the unit circle by the polynomial $\zeta \mapsto \sum_{l=0}^{n-1} v_l \zeta^l$. Thus, in what follows, we will frequently refer to the *dual polynomial* as the polynomial whose coefficients are given by the dual optimal solution of the AST problem defined in (2.2), reproduced here for convenience:

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\|x - y\|_2^2 + \tau \|x\|_{\mathcal{A}}. \tag{3.7}$$

Chapter 5 studies algorithms for solving AST and we will see that we can solve it exactly using a semidefinite program and approximate it with a Lasso estimate. In this chapter, let us restrict ourselves to theoretically analyzing the performance of AST.

3.3 Determining the frequencies

As shown in Corollary 2.8, the dual solution can be used to identify the frequencies of the primal solution. For line spectra, a frequency $f \in [0,1]$ is in the support of the solution $\hat{x}$ of (3.7) if and only if

$$|\langle \hat{z}, a_{f,\phi} \rangle| = \left| \sum_{l=0}^{n-1} \hat{z}_l e^{-i2\pi l f} \right| = \tau$$

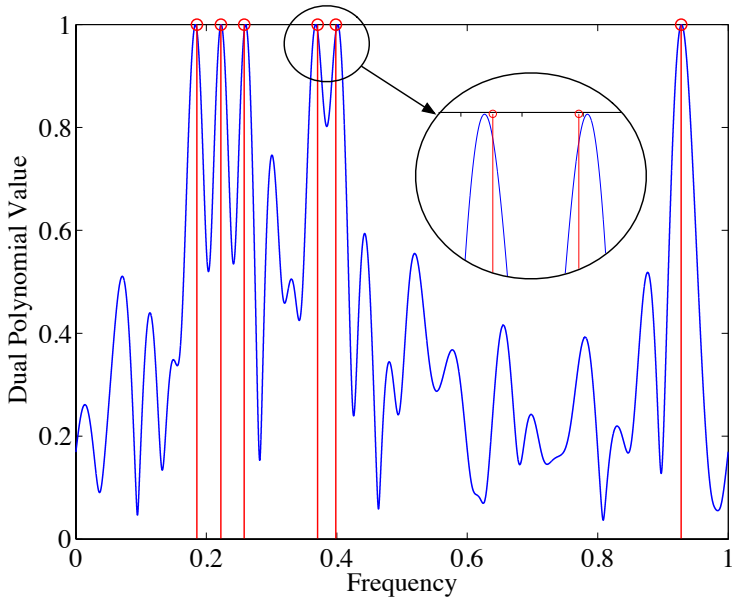


Figure 3.1: **Frequency Localization using Dual Polynomial**: The actual location of the frequencies in the line spectral signal $x^\star \in \mathbb{C}^{64}$ is shown in red. The blue curve is the dual polynomial obtained by solving (2.7) with $y = x^\star + w$ where w is noise of SNR 10 dB.

That is, f is in the support of $\hat{x}$ if and only if it is a point of maximum modulus for the dual polynomial. Thus, the support may be determined by finding frequencies f where the dual polynomial attains magnitude τ.

Figure 3.1 shows the dual polynomial for (3.7) with $n = 64$ samples and $k = 6$ randomly chosen frequencies. In the next section, let us consider the choice of the regularization parameter τ.

3.4 Choosing the regularization parameter

The choice of the regularization parameter is dictated by the noise model and we derive the optimal choice for white gaussian noise samples in our analysis. As noted in Theorem 2.1,

the optimal choice of the regularization parameter depends on the dual norm of the noise. A simple lower bound on the expected dual norm occurs when we consider the maximum value of n uniformly spaced points in the unit circle. Using the result of [65], the lower bound whenever $n \geq 5$ is

$$\sigma\sqrt{n\log(n) - \tfrac{n}{2}\log(4\pi\log(n))}\,.$$

Using standard results on the extreme value statistics of Gaussian distribution, we can also obtain a non-asymptotic upper bound on the expected dual norm of noise for $n > 3$:

$$\sigma\left(1 + \frac{1}{\log(n)}\right)\sqrt{n\log(n) + n\log(16\pi^3/2\log(n))}$$

We will examine these computations in detail in the following section.

3.4.1 Estimation of Gaussian Width

This section derives non-asymptotic upper and lower bounds on the expected dual norm of gaussian noise vectors, which are asymptotically tight upto $\log\log$ factors. Recall that the dual atomic norm of w is given by $\sqrt{n}\sup_{f\in[0,1]}|W_f|$ where

$$W_f = \frac{1}{\sqrt{n}}\sum_{m=0}^{n-1} w_m e^{-i2\pi mf}.$$

Here, the noise variables $w_1, w_2, \ldots$ are circularly symmetric independent sequence of standard complex normal variables.

If we define two independent $i.i.d$ sequences of standard normal numbers $\{g_k\}_1^\infty$ and $\{h_k\}_1^\infty$, note that we can write

$$W_f = \frac{1}{\sqrt{2n}}\sum_{k=0}^{n-1}\left[g_k\cos(2\pi kf) - h_k\sin(2\pi kf)\right]. \tag{3.8}$$

Note that W_f is a normal random variable with zero mean and a variance of $1/2$.

3.4.2 Upper Bound

Let us use a $1/N$-net of the torus $\mathbb{T}$ to estimate the expectation of $\sup_{f\in\mathbb{T}} W_f$. Define

$$\mathbb{T} = \{t \in \mathbb{T} \mid \ |t - k/N| \leq 1/N\}\,.$$

We have

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in T}|W_f|\right] &\leq \mathbb{E}\left[\sup_{1\leq k\leq N}|W_{k/N}|\right] + \mathbb{E}\left[\sup_{f\in\mathbb{T}_k}\left|W_f - W_{k/N}\right|\right] \\ &\leq \sqrt{\log(N)} + \frac{2\pi n}{\sqrt{3}N}\,\mathbb{E}\left[\sup_{f\in\mathbb{T}_k}|Y_f|\right] \end{aligned} \tag{3.9}$$

where

$$Y_f = \frac{\sqrt{3}N\left(W_f - W_{k/N}\right)}{4\pi n}. \tag{3.10}$$

We can use Dudley's integral inequality (See, for example [66]) to bound $\mathbb{E}\left[\sup_{f\in\mathbb{T}_k}|Y_f|\right]$. To proceed, let us compute the pseudometric ρ of the Gaussian process $\{W_f\}_f$ induced on the index set. For indices t and s in $\mathbb{T}_k$,

$$\begin{aligned}\rho^2(t,s) &:= \mathbb{E}\,|Y_t - Y_s|^2 \\ &= \frac{3N^2}{8\pi^2n^2}\,\mathbb{E}\,|X_t - X_s|^2 \\ &= \frac{3N^2}{2\pi^2n^3}\sum_{k=0}^{n-1}\sin^2(\pi k(t-s)) \\ &\leq N^2(t-s)^2.\end{aligned}$$

Thus, the diameter of the index set $\mathbb{T}_k$ with respect to ρ

$$\operatorname{diam}_\rho(\mathbb{T}_k) := \sup_{t,s\in\mathbb{T}_k} \rho(t,s) = 1.$$

Consequently the number $N(\mathbb{T}_k, \rho, \epsilon)$ of ϵ balls needed to cover T_k under this metric ρ is $1/\epsilon$. Now, the application Dudley's integral inequality yields

$$\mathbb{E}\left[\sup_{t\in\mathbb{T}_k} |Y_f|\right] \leq 24 \int_0^{\operatorname{diam}_\rho(T_k)} \sqrt{N(\mathbb{T}_k, \rho, \epsilon)} d\epsilon \tag{3.11}$$

$$\leq 24 \int_0^1 \sqrt{\log(1/\epsilon)} d\epsilon = 12\sqrt{\pi}. \tag{3.12}$$

Thus, from (3.12) and (3.9),

$$\mathbb{E}\left[\sup_{f\in\mathbb{T}} |W_f|\right] \leq \sqrt{\log(N)} + \frac{16\pi^{3/2} n}{N}$$

Substituting $N = 16n\sqrt{\pi^3 \log(n)}$, we get

$$\mathbb{E}\left[\sup_{f\in\mathbb{T}} W_f\right] \leq \left(1 + \frac{1}{\log(n)}\right) \sqrt{\log(n) + \log(16\pi^{3/2}\log(n))}$$

By the concentration to mean of the Gaussian process,

$$\sup_{f\in\mathbb{T}} |\langle w, a(f)\rangle| \leq 2\,\mathbb{E}\left[\sup_{f\in\mathbb{T}} |W_f|\right] \tag{3.13}$$

with high probability.

3.4.3 Lower Bound

The covariance function of W_f is

$$\mathbb{E}\left[W_{f_1} W_{f_2}^*\right] = \frac{1}{n}\sum_{m=0}^{n-1} \exp(2\pi m(f_1 - f_2)) = e^{\pi(n-1)(f_1-f_2)} \frac{\sin(n\pi(f_1 - f_2))}{n\sin(\pi(f_2 - f_2))}.$$

Thus, the n samples $\left\{W_{m/n}\right\}_{m=0}^{n-1}$ are uncorrelated and thus independent because of their joint gaussianity. This gives a simple non-asymptotic lower bound using the known result for maximum value of n independent gaussian random variables [65] whenever $n > 5$:

$$\mathbb{E}\left[\sup_{t\in T} |W_t|\right] \geq \mathbb{E}\left[\max_{m=0,\ldots,n-1} \mathfrak{Re}\left(W_{m/n}\right)\right] = \sqrt{\log(n) - \tfrac{\log\log(n)+\log(4\pi)}{2}}.$$

Combining this result with the upper bound, we can see that the lower bound is asymptotically tight neglecting $\log\log$ terms.

3.5 Universal Mean Squared Error Guarantee

We can set the regularization parameter τ greater than the upper bound on the expected dual atomic norm, i.e., we pick $\eta \in [1, \infty]$ and let

$$\tau = \sigma\eta\left(1 + \frac{1}{\log(n)}\right)\sqrt{n\log(n) + n\log(16\pi^{3/2}\log(n))}. \tag{3.14}$$

The application of Theorem 2.1 with the choice $\eta = 1$ guarantees Mean-Squared Error consistency of AST for Line spectral Estimation. This choice of τ then yields the asymptotic result in Theorem 3.1. However, as noted in Section 2.5, faster convergence rates may be possible under some conditions, whenever $\eta > 1$. Due to concentration to mean (3.13),

whenever $\eta > 1$, with overwhelming probability,

$$\sup_{f \in \mathbb{T}} |\langle w, a(f) \rangle| \leq 2\eta^{-1}\tau. \tag{3.15}$$

A recent result by Candes and Fernandez-Granda [20] establishes that in the noiseless case, the frequencies localized by the dual polynomial are exact provided the minimum separation between the frequencies is at least $4/n$ where n is the number of samples in the line spectral signal. Under similar separation condition, numerical simulations suggest that (3.7) achieves approximate frequency location in the noisy case.

In fact, we can also theoretically show that signals with well separated frequencies are well behaved and achieve faster convergence rates. Unlike previous work on fast rates for Lasso, the condition for fast rates is on the signal instead of the measurement operator. In fact, as frequencies can be arbitrarily close, the measurement operator which samples line spectral signals is highly coherent and it may be impossible to achieve robust recovery if frequencies can be close to each other.

3.6 What is the best rate we can expect?

Using results about minimax achievable rates for linear models [22, 83], we can deduce that the convergence rate stated in (3.3) is near optimal. Define the set of k well separated frequencies as

$$\mathcal{S}_k = \left\{ (f_1, \ldots, f_k) \in \mathbb{T}^k \;\middle|\; d(f_p, f_q) \geq 4/n, p \neq q \right\}$$

The expected minimax denoising error M_k for a line spectral signal with frequencies from $\mathcal{S}_k$ is defined as the lowest expected denoising error rate for any estimate $\hat{x}(y)$ for the worst

case signal $x^\star$ with support $T(x^\star) \in \mathcal{S}_k$. Note that we can lower bound M_k by restricting the set of candidate frequencies to smaller set. To that end, suppose we restrict the signal $x^\star$ to have frequencies only drawn from an equispaced grid on the torus $T_n := \{4j/n\}_{j=1}^{n/4}$. Note that any set of k frequencies from T_n are pairwise separated by at least $4/n$. If we denote by F_n a $n \times (n/4)$ partial DFT matrix with (unnormalized) columns corresponding to frequencies from T_n, we can write $x^\star = F_n c^\star$ for some $c^\star$ with $\|c^\star\|_0 = k$. Thus,

$$\begin{aligned} M_k &:= \inf_{\hat{x}} \sup_{T(x^\star)\in\mathcal{S}_k} \frac{1}{n}\mathbb{E}\|\hat{x} - x^\star\|_2^2 \\ &\geq \inf_{\hat{x}} \sup_{\|c^\star\|_0\leq k} \frac{1}{n}\mathbb{E}\|\hat{x} - F_n c^\star\|_2^2 \\ &\geq \inf_{\hat{c}} \sup_{\|c^\star\|_0\leq k} \frac{1}{n}\mathbb{E}\|F_n(\hat{c} - c^\star)\|_2^2 \\ &\geq \frac{n}{4}\left\{\inf_{\hat{c}} \sup_{\|c^\star\|_0\leq k} \frac{4}{n}\mathbb{E}\|\hat{c} - c^\star\|_2^2\right\}. \end{aligned}$$

Here, the first inequality is the restriction of $T(x^\star)$. The second inequality follows because we have projected out all components of $\hat{x}$ that do not lie in the span of F_n. Such projections can only reduce the Euclidean norm. The third inequality uses the fact that the minimum singular value of F_n is n since $F_n^* F_n = nI_{n/4}$. Now we may directly apply the lower bound for estimation error for linear models derived by Candés and Davenport. Namely, Theorem 1 of [22] states that

$$\inf_{\hat{c}} \sup_{\|c^\star\|_0\leq k} \frac{4}{n}\mathbb{E}\|\hat{c} - c^\star\|_2^2 \geq C\sigma^2 \frac{k\log\left(\frac{n}{4k}\right)}{\|F_n\|_{\mathrm{F}}^2}.$$

With the preceding analysis and the fact that $\|F_n\|_{\mathrm{F}}^2 = n^2/4$, we can thus deduce the following

theorem:

Theorem 3.4. *Let $x^\star$ be a line spectral signal as described by* (3.1) *with the support $T(x^\star) = \{f_1, \ldots, f_k\} \in \mathcal{S}_k$ and $y = x^\star + w$, where $w \in \mathbb{C}^n$ is circularly symmetric Gaussian noise with variance $\sigma^2 I_n$. Let $\hat{x}$ be any estimate of $x^\star$ using y. Then,*

$$M_k = \inf_{\hat{x}} \sup_{T(x^\star)\in\mathcal{S}_k} \frac{1}{n}\mathbb{E}\|\hat{x} - x^\star\|_2^2 \geq C\sigma^2 \frac{k \log\left(\frac{n}{4k}\right)}{n}$$

for some constant C that is independent of k, n, and σ.

This theorem and Theorem 3.2 certify that AST is nearly minimax optimal for spectral estimation of well separated frequencies.

3.7 Proofs for well separated frequencies

In this section, there are many numerical constants. Unless otherwise specified, C will denote a numerical constant whose value may change from equation to equation. Specific constants will be highlighted by accents or subscripts.

Before sketching the proof of Theorems 3.2 and 3.3, we will need some the preliminaries and notations. We will also need to recall some recent results that are relevant for the problem.

3.7.1 Preliminaries

The sample $x_j^\star$ may be regarded as the jth trigonometric moment of the discrete measure μ given by (3.4):

$$x_j^\star \quad = \quad \int_0^1 e^{i2\pi jf} \mu(df)$$

for $-m \leq j \leq m$. Thus, the problem of extracting the frequencies and amplitudes from noisy observations may be regarded as the inverse problem of estimating a measure from noisy trigonometric moments.

We can write the vector $x^\star$ of observations $[x^\star_{-m}, \ldots, x^\star_m]^T$ in terms of an *atomic decomposition*

$$x^\star = \sum_{l=1}^{k} c_l a(f_l)$$

or equivalently in terms of a corresponding *representing measure* μ given by (3.4) satisfying

$$x^\star = \int_0^1 a(f)\mu(df)$$

There is a one-one correspondence between atomic decompositions and representing measures. Note that there are infinite atomic decompositions of $x^\star$ and also infinite corresponding representing measures. However, since every collection of n atoms is linearly independent, $\mathcal{A}$ forms a full spark frame [41] and therefore the problem of finding the sparsest decomposition of $x^\star$ is well-posed if there is a decomposition which is at least $n/2$ sparse.

The atomic norm of a vector z defined in (2.1) is the minimum total variation norm [97, 29] $\|\mu\|_{\mathrm{TV}}$ of all representing measures μ of z. So, minimizing the total variation norm is the same as finding a decomposition that achieves the atomic norm.

3.7.2 Dual Certificate and Exact Recovery

Atomic norm minimization attempts to recover the sparsest decomposition by finding a decomposition that achieves the atomic norm, i.e., find c_l, f_l such that $x^\star = \sum_l c_l a(f_l)$ and $\|x^\star\|_{\mathcal{A}} = \sum_l |c_l|$ or equivalently, finding a representing measure μ of the form (3.4) that minimizes the total variation norm $\|\mu\|_{\mathrm{TV}}$. The authors of [20] showed that when

$n > 256$, the decomposition that achieves the atomic norm is the sparsest decomposition by explicitly constructing a dual certificate [23] of optimality, whenever the composing frequencies $f_1, \ldots, f_k$ satisfy a minimum separation condition (3.2). In the rest of the chapter, let us always make the technical assumption that $n > 256$. The following is just a restatement of Definition 2.3 for trigonometric moments:

Definition 3.5 (Dual Certificate)**.** *A vector $q \in \mathbb{C}^n$ is called a dual certificate for the decomposition*

$$x^\star = \sum_{l=1}^{k} c_l a(f_l)$$

if for the corresponding trigonometric polynomial $Q(f) := \langle q, a(f) \rangle$, we have

$$Q(f_l) = \operatorname{sign}(c_l), l = 1, \ldots, k$$

and

$$|Q(f)| < 1$$

whenever $f \notin \{f_1, \ldots, f_k\}$.

The authors of [20] not only explicitly constructed such a certificate characterized by the dual polynomial Q, but also showed that their construction satisfies some stability conditions, which is crucial for showing that denoising using the atomic norm provides stable recovery in the presence of noise.

Theorem 3.6 (Dual Polynomial Stability, Lemma 2.4 and 2.5 in [19])**.** *For any $f_1, \ldots, f_k$ satisfying the separation condition* (3.2) *and any sign vector $v \in \mathbb{C}^k$ with $|v_j| = 1$, there exists a trigonometric polynomial $Q = \langle q, a(f) \rangle$ for some $q \in \mathbb{C}^n$ with the following properties:*

1. *For each $j = 1, \ldots, k$, Q interpolates the sign vector v so that $Q(f_j) = v_j$*

2. *In each neighborhood N_j corresponding to f_j defined by $N_j = \{f : d(f, f_j) < 0.16/n\}$, the polynomial $Q(f)$ behaves like a quadratic and there exist constants C_a, C'_a so that*

$$|Q(f)| \leq 1 - \frac{C_a}{2} n^2 (f - f_j)^2 \tag{3.16}$$

$$|Q(f) - v_j| \leq \frac{C'_a}{2} n^2 (f - f_j)^2 \tag{3.17}$$

3. *When $f \in F = [0, 1] \backslash \cup_{j=1}^{k} N_j$, there is a numerical constant $C_b > 0$ such that*

$$|Q(f)| \leq 1 - C_b$$

This chapter uses results in [19] and [6] and borrows several ideas from the proofs in [19], with nontrivial modifications to establish the error rate of atomic norm regularization.

3.7.3 Near optimal MSE

In this section, we will see a proof of Theorem 3.2. Let $\hat{\mu}$ be the representing measure for the solution $\hat{x}$ of (2.2) with minimum total variation norm, that is,

$$\hat{x} = \int_0^1 a(f) \hat{\mu}(df)$$

and $\|\hat{x}\|_{\mathcal{A}} = \|\hat{\mu}\|_{\mathrm{TV}}$. Denote the error vector by $e = x^\star - \hat{x}$. Then, the difference measure $\nu = \mu - \hat{\mu}$ is a representing measure for e. Express the denoising error $\|e\|_2^2$ as the integral of

the error function $E(f) = \langle e, a(f) \rangle$, against the difference measure ν:

$$\begin{aligned} \|e\|_2^2 &= \langle e, e \rangle \\ &= \left\langle e, \int_0^1 a(f)\nu(df) \right\rangle \\ &= \int_0^1 \langle e, a(f) \rangle \, \nu(df) \\ &= \int_0^1 E(f)\nu(df). \end{aligned}$$

Using a Taylor series approximation in each of the near regions N_j, we will see that the denoising error (or in general any integral of a trigonometric polynomial against the difference measure) can be controlled in terms of an integral in the far region F and the zeroth, first, and second moments of the difference measure in the near regions. The precise result is presented in the following lemma:

Lemma 3.7. *Define*

$$\begin{aligned} I_0^j &:= \left| \int_{N_j} \nu(df) \right| \\ I_1^j &:= n \left| \int_{N_j} (f - f_j)\nu(df) \right| \\ I_2^j &:= \frac{n^2}{2} \int_{N_j} (f - f_j)^2 |\nu|(df) \\ I_l &:= \sum_{j=1}^{k} I_l^j, \ \textit{ for } l = 0, 1, 2\,. \end{aligned}$$

Then for any mth order trigonometric polynomial X, we have

$$\int_0^1 X(f)\nu(df) \leq \|X(f)\|_\infty \left(\int_F |\nu|(df) + I_0 + I_1 + I_2 \right)$$

Proof. Split the domain of integration into the near and far regions.

$$\left|\int_0^1 X(f)\nu(df)\right| \leq \left|\int_F X(f)\nu(f)\right| + \sum_{j=1}^{k}\left|\int_{N_j} X(f)\nu(df)\right|$$
$$\leq \|X(f)\|_\infty \int_F |\nu|(df) + \sum_{j=1}^{k}\left|\int_{N_j} X(f)\nu(df)\right|. \tag{3.18}$$

by using Hölder's inequality for the last inequality. Using Taylor's theorem, we may expand the integrand $X(f)$ around f_j as

$$X(f) = X(f_j) + (f - f_j)X'(f_j) + \frac{1}{2}X''(\xi_j)(f - f_j)^2$$

for some $\xi_j \in N_j$. Thus,

$$\begin{aligned}
&|X(f) - X(f_j) - X'(f_j)(f - f_j)| \\
&\leq \sup_{\xi \in N_j} \frac{1}{2}|X''(\xi)|(f - f_j)^2 \\
&\leq \frac{1}{2}n^2\|X(f)\|_\infty (f - f_j)^2,
\end{aligned}$$

where for the last inequality follows from a theorem of Bernstein for trigonometric polynomials (see, for example [87]):

$$|X'(f_j)| \leq n\|X(f)\|_\infty$$
$$|X''(f_j)| \leq n^2\|X(f)\|_\infty.$$

As a consequence, we have

$$\begin{aligned}\left|\int_{N_j} X(f)\nu(df)\right| &\leq |X(f_j)|\left|\int_{N_j}\nu(df)\right| + |X'(f_j)|\left|\int_{N_j}(f-f_j)\nu(df)\right| \\ &\quad + \frac{1}{2}n^2\|X(f)\|_\infty \int_{N_j}(f-f_j)^2|\nu|(df) \\ &\leq \|X(f)\|_\infty\left(I_0^j + I_1^j + I_2^j\right).\end{aligned}$$

Substituting back into (3.18) yields the desired result. □

Applying Lemma 3.7 to the error function, we get

$$\|e\|_2^2 \leq \|E(f)\|_\infty\left(\int_F |\nu|(df) + I_0 + I_1 + I_2\right) \tag{3.19}$$

As a consequence of the choice of τ given by (3.14), we can show that $\|E(f)\|_\infty \leq (1+2\eta^{-1})\tau$ with high probability. In fact, we have

$$\begin{aligned}\|E(f)\|_\infty &= \sup_{f\in[0,1]}|\langle e, a(f)\rangle| \\ &= \sup_{f\in[0,1]}|\langle x^\star - \hat{x}, a(f)\rangle| \\ &\leq \sup_{f\in[0,1]}|\langle w, a(f)\rangle| + \sup_{f\in[0,1]}|\langle y-\hat{x}, a(f)\rangle| \\ &\leq \sup_{f\in[0,1]}|\langle w, a(f)\rangle| + \tau \\ &\leq (1+2\eta^{-1})\tau \leq 3\tau, \text{with high probability.}\end{aligned} \tag{3.20}$$

The second inequality follows from the optimality conditions for (2.2) and the penultimate inequality is from (3.15).

Therefore, to complete the proof, it suffices to show that the other terms on the right hand

side of (3.19) are $O(\frac{k\tau}{n})$. While there is no exact frequency recovery in the presence of noise, we can hope to get the frequencies approximately right. Hence, we can expect that the integral in the far region can be well controlled and the local integrals of the difference measure in the near regions are also small due to cancellations. Next, we could utilize the properties of the dual polynomial in Theorems 3.6 and another polynomial given in Theorem A.1 in Appendix A to show that the zeroth and first moments of ν may be controlled in terms of the other two quantities in (3.19) to upper bound the error rate. The following lemma is similar to Lemmas 2.2 and 2.3 in [19], but we have made several modifications to adapt it to our signal and noise model.

Lemma 3.8. *There exists numeric constants C_0 and C_1 such that*

$$I_0 \leq C_0 \left(\frac{k\tau}{n} + I_2 + \int_F |\nu|(df) \right)$$
$$I_1 \leq C_1 \left(\frac{k\tau}{n} + I_2 + \int_F |\nu|(df) \right).$$

Proof. Consider the polar form

$$\int_{N_j} \nu(df) = \left| \int_{N_j} \nu(df) \right| e^{i\theta_j}.$$

Set $v_j = e^{-i\theta_j}$ and let $Q(f)$ be the dual polynomial promised by Theorem 3.6 for this v. Then, we have

$$\begin{aligned} \left| \int_{N_j} \nu(df) \right| &= \int_{N_j} e^{-i\theta_j} \nu(df) \\ &= \int_{N_j} Q(f)\nu(df) + \int_{N_j} (e^{-i\theta_j} - Q(f))\nu(df) \end{aligned}$$

Summing over $j = 1, \ldots, k$ yields

$$
\begin{aligned}
I_0 &= \sum_{j=1}^{k} \left| \int_{N_j} \nu(df) \right| \\
&= \sum_{j=1}^{k} \int_{N_j} Q(f)\nu(df) + \sum_{j=1}^{k} \int_{N_j} (v_j - Q(f))\nu(df) \\
&\leq \left| \int_0^1 Q(f)\nu(df) \right| + \int_F |\nu|(df) + C_a' I_2, \text{ using triangle inequality and (3.17)} \\
&\leq \frac{Ck\tau}{n} + \int_F |\nu|(df) + C_a' I_2, \text{ using (A.13).} \qquad (3.21)
\end{aligned}
$$

We can use a similar argument for bounding I_1 but this time use the dual polynomial $Q_1(f)$ guaranteed by Theorem A.1. Again, start with the polar form

$$
\int_{N_j} (f - f_j)\nu(df) = \left| \int_{N_j} (f - f_j)\nu(df) \right| e^{i\theta_j} = I_1^j e^{i\theta_j}/n
$$

Set $v_j = e^{-i\theta_j}$ in Theorem A.1 to obtain

$$
\begin{aligned}
I_1^j &= n \int_{N_j} e^{-i\theta_j}(f - f_j)\nu(df) \\
&= n \int_{N_j} (v_j(f - f_j) - Q_1(f))\nu(df) + n \int_{N_j} Q_1(f)\nu(df)
\end{aligned}
$$

Summing over $j = 1, \ldots, k$ yields

$$\begin{aligned} I_1 &= \sum_{j=1}^{k} I_1^j \\ &= n \sum_{j=1}^{k} \int_{N_j} (v_j(f - f_j) - Q_1(f))\nu(df) + n \sum_{j=1}^{k} \int_{N_j} Q_1(f)\nu(df) \\ &\leq C_a^1 I_2 + n \left| \int_0^1 Q_1(f)\nu(df) \right| + n \left| \int_F Q_1(f)\nu(df) \right| \\ &\leq C_a^1 I_2 + \frac{Ck\tau}{n} + C_b^1 \int_F |\nu|(df) \end{aligned} \tag{3.22}$$

The first inequality uses (A.1) and triangle inequality, and the last inequality uses (A.14) and (A.2). Equations (3.21) and (3.22) complete the proof. □

All that remains to complete the proof is an upper bound on I_2 and $\int_F |\nu|(df)$. The key idea in establishing such a bound is deriving upper and lower bounds on the difference $\|P_{T^c}(\nu)\|_{\mathrm{TV}} - \|P_T(\nu)\|_{\mathrm{TV}}$ between the total variation norms of ν on and off the support. The upper bound can be derived using optimality conditions. We lower bound $\|P_{T^c}(\nu)\|_{\mathrm{TV}} - \|P_T(\nu)\|_{\mathrm{TV}}$ using the fact that a constructed dual certificate Q has unit magnitude for every element in the support T of $P_T(\nu)$ whence we have $\|P_T(\nu)\|_{\mathrm{TV}} = \int_{\mathbb{T}} Q(f)\nu(df)$. A critical element in deriving both the lower and upper bounds is that the dual polynomial Q has quadratic drop in each near regions N_j and is bounded away from one in the far region F. Finally, by combing these bounds and carefully controlling the regularization parameter, we get the desired result summarized in the following lemma.

Lemma 3.9. *Let $\tau = \eta\sigma\sqrt{n\log(n)}$. If $\eta > 1$ is large enough, then there exists a numerical constant C such that, with high probability*

$$\int_F |\nu|(df) + I_2 \leq \frac{Ck\tau}{n}.$$

Proof. Denote by $P_T(\nu)$ the projection of the difference measure ν on the support set $T = \{f_1, \ldots, f_k\}$ of $x^\star$ so that $P_T(\nu)$ is supported on T. Then, setting $Q(f)$ the polynomial in Theorem 3.6 that interpolates the sign of $P_T(\nu)$, we have

$$\begin{aligned}
\|P_T(\nu)\|_{\mathrm{TV}} &= \int_0^1 Q(f) P_T(\nu)(df) \\
&\leq \left| \int_0^1 Q(f)\nu(df) \right| + \left| \int_{T^c} Q(f)\nu(df) \right| \\
&\leq \frac{Ck\tau}{n} + \sum_{f_j \in T} \left| \int_{N_j/\{f_j\}} Q(f)\nu(df) \right| + \left| \int_F Q(f)\nu(df) \right|,
\end{aligned}$$

where the first inequality follows from triangle inequality and for the last inequality is given by (A.13). The integration over F is can be bounded using Hölder's inequality

$$\left| \int_F Q(f)\nu(df) \right| \leq (1 - C_b) \int_F |\nu|(df)$$

Continue with

$$\begin{aligned}
\left| \int_{N_j/\{f_j\}} Q(f)\nu(df) \right| &\leq \left| \int_{N_j/\{f_j\}} |Q(f)||\nu|(df) \right| \\
&\leq \int_{N_j/\{f_j\}} (1 - \tfrac{1}{2} n^2 C_a (f - f_j)^2) |\nu|(df) \\
&\leq \int_{N_j/\{f_j\}} |\nu|(df) - C_a I_2^j.
\end{aligned}$$

As a consequence, we have

$$\begin{aligned}\|P_T(\nu)\|_{\mathrm{TV}} &\leq \frac{Ck\tau}{n} + \sum_{f_j \in T} \int_{N_j/\{f_j\}} |\nu|(df) - C_a I_2 + (1 - C_b) \int_F |\nu|(df) \\ &\leq \frac{Ck\tau}{n} + \underbrace{\sum_{f_j \in T} \int_{N_j/\{f_j\}} |\nu|(df) + \int_F |\nu|(df)}_{\|P_{T^c}\|_{\mathrm{TV}}} - C_a I_2 - C_b \int_F |\nu|(df)\end{aligned}$$

or equivalently,

$$\|P_{T^c}(\nu)\|_{\mathrm{TV}} - \|P_T(\nu)\|_{\mathrm{TV}} \geq C_a I_2 + C_b \int_F |\nu|(df) - \frac{Ck\tau}{n}. \tag{3.23}$$

Now, appeal to the optimality conditions (2.18) of AST to obtain

$$\|\hat{x}\|_{\mathcal{A}} \leq \|x^\star\|_{\mathcal{A}} - \langle w, e \rangle / \tau$$

and thus

$$\|\hat{\mu}\|_{\mathrm{TV}} \leq \|\mu\|_{\mathrm{TV}} + |\langle w, e \rangle| / \tau. \tag{3.24}$$

Using Lemma 3.7,

$$\begin{aligned}|\langle w, e \rangle| &= |\langle w, \int_0^1 a(f) \nu(df) | \rangle \\ &= \left| \int_0^1 \langle w, a(f) \rangle \, \nu(df) \right| \qquad (3.25) \\ &\leq \|\langle w, a(f) \rangle\|_\infty \left(\frac{Ck\tau}{n} + I_0 + I_1 + I_2 \right) \\ &\leq 2\eta^{-1}\tau \left(\frac{Ck\tau}{n} + I_0 + I_1 + I_2 \right) \\ &\leq C\eta^{-1}\tau \left(\frac{k\tau}{n} + I_2 + \int_F |\nu|(df) \right) \qquad (3.26)\end{aligned}$$

with high probability, where for the penultimate inequality uses the choice of τ in (3.14) and thus $\|\langle w, a(f)\rangle\|_\infty \leq 2\eta^{-1}\tau$ with high probability from (3.15).

Substituting (3.26) in (3.24), we get

$$\begin{aligned}
&\|\mu\|_{\mathrm{TV}} + C\eta^{-1}\tau\left(\frac{k\tau}{n} + I_2 + \int_F |\nu|(df)\right) \\
&\geq \|\hat{\mu}\|_{\mathrm{TV}} \\
&= \|\mu + \nu\|_{\mathrm{TV}} \\
&\geq \|\mu\|_{\mathrm{TV}} - \|P_T(\nu)\|_{\mathrm{TV}} + \|P_{T^c}(\nu)\|_{\mathrm{TV}}
\end{aligned}$$

Canceling $\|\mu\|_{\mathrm{TV}}$ yields

$$\|P_{T^c}(\nu)\|_{\mathrm{TV}} - \|P_T(\nu)\|_{\mathrm{TV}} \leq C\eta^{-1}\tau\left(\frac{k\tau}{n} + I_2 + \int_F |\nu|(df)\right) \tag{3.27}$$

As a consequence of (3.23) and (3.27), we get,

$$C(1+\eta^{-1})\frac{k\tau}{n} \geq (C_b - \eta^{-1}C)\int_F |\nu|(df) + (C_a - \eta^{-1}C)I_2$$

whence the result follows for large enough η. □

Putting together Lemmas 3.7, 3.8 and 3.9, we can finally prove our main theorem:

$$\begin{aligned}
\frac{1}{n}\|e\|_2^2 &\leq \frac{\|E(f)\|_\infty}{n}\left(\int_F |\nu|(df) + I_0 + I_1 + I_2\right) \\
&\leq \frac{\|E(f)\|_\infty}{n}\left(\frac{C_1 k\tau}{n} + C_2\int_F |\nu|(df) + C_3 I_2\right) \\
&\leq \frac{\|E(f)\|_\infty}{n}\frac{Ck\tau}{n} \\
&\leq \frac{Ck\tau^2}{n^2} \\
&= O\left(\sigma^2\frac{k\log(n)}{n}\right).
\end{aligned}$$

The first three inequalities come from successive applications of Lemmas 1, 2 and 3 respectively. The fourth inequality follows from (3.20) and the fifth by the choice of τ according to Eq. (3.14). This completes the proof of Theorem 3.2.

3.7.4 Approximate Frequency Localization

In this section, we will see a proof of Theorem 3.3. The first two statements in Theorem 3.3 are direct consequences of Lemma 3.9. For (iii.), we will follow [52] and use the dual polynomial $Q_j^\star(f) = \langle q_j^\star, a(f)\rangle$ constructed in Lemma 2.2 of [52] which satisfies

$$\begin{aligned}
Q_j^\star(f_j) &= 1 \\
|1 - Q_j^\star(f)| &\leq n^2 C_1'(f - f_j)^2, f \in N_j \\
|Q_j^\star(f)| &\leq n^2 C_1'(f - f_{j'})^2, f \in N_{j'}, j' \neq j \\
|Q_j^\star(f)| &\leq C_2', f \in F.
\end{aligned}$$

Note that $c_j - \sum_{\hat{f}_l \in N_j} \hat{c}_l = \int_{N_j} \nu(df)$. Then, by applying triangle inequality several times,

$$\begin{aligned}
\left|\int_{N_j} \nu(df)\right| &\leq \left|\int_{N_j} Q_j^\star(f)\nu(df)\right| + \left|\int_{N_j} (1 - Q_j^\star(f))\nu(df)\right| \\
&\leq \left|\int_0^1 Q_j^\star(f)\nu(df)\right| + \left|\int_{N_j^c} Q_j^\star(f)\nu(df)\right| + \left|\int_{N_j} (1 - Q_j^\star(f))\nu(df)\right| \\
&\leq \left|\int_0^1 Q_j^\star(f)\nu(df)\right| + \left|\int_F Q_j^\star(f)\nu(df)\right| \\
&\qquad + \sum_{\substack{j'\neq j \\ j'=1}}^{k} \int_{N_{j'}} \left|Q_j^\star(f)\right| |\nu|(df) + \int_{N_j} \left|1 - Q_j^\star(f)\right| |\nu(df)| \,.
\end{aligned}$$

We will upper bound the first term using Lemma A.3 in Appendix A which yields

$$\left|\int_1^0 Q_j^\star(f)\nu(df)\right| \leq \frac{Ck\tau}{n}$$

The other terms can be controlled using the properties of $Q_j^\star$:

$$\left|\int_F Q_j^\star(f)\nu(df)\right| \leq C_2' \int_F |\nu|(df)$$

$$\sum_{\substack{j'\neq j \\ j'=1}}^{k} \int_{N_{j'}} \left|Q_j^\star(f)\right| |\nu|(df) + \int_{N_j} \left|1 - Q_j^\star(f)\right| |\nu|(df) \leq C_1' \sum_{j'=1}^{k} \int_{N_{j'}} n^2 (f - f_{j'})^2 |\nu|(df) = C_1 I_2$$

Using Lemma 3.9, both of the above are upper bounded by $\frac{Ck\tau}{n}$. Now, by combining these upper bounds, we finally have

$$\left| c_j - \sum_{l:\hat{f}_l \in N_j} \hat{c}_l \right| \leq \frac{C_3 k\tau}{n}$$

This shows part (iii) of the theorem. Part (iv) can be obtained by combining parts (ii) and

(iii).

3.8 Related Work

The classical methods of line spectral estimation, often called linear prediction methods, are built upon the seminal interpolation method of Prony [81]. In the noiseless case, with as little as $n = 2k$ measurements, Prony's technique can identify the frequencies exactly, no matter how close the frequencies are. However, Prony's technique is known to be sensitive to noise due to instability of polynomial rooting [62]. Following Prony, several methods have been employed to robustify polynomial rooting method including the Matrix Pencil algorithm [60], which recasts the polynomial rooting as a generalized eigenvalue problem and cleverly uses extra observations to guard against noise. The MUSIC [89] and ESPRIT [85] algorithms exploit the low rank structure of the autocorrelation matrix.

Cadzow [15] proposed a heuristic that improves over MUSIC by exploiting the Toeplitz structure of the matric of moments by alternately projecting between the linear space of Toeplitz matrices and the space of rank k matrices where k is the desired model order. Cadzow's technique is very similar [108] to a popular technique in time series literature [63, 54] called Singular Spectrum Analysis [103], which uses autocorrelation matrix instead of the matrix of moments for projection. Both these techniques may be viewed as instances of structured low rank approximation [34] which exploit additional structure beyond low rank structure used in subspace based methods such as MUSIC and ESPRIT. Cadzow's method has been identified as a fruitful preprocessing step for linear prediction methods [8]. A survey of classical linear prediction methods can be found in [8, 92] and an extensive list of references is given in [91].

Most, if not all of the linear prediction methods need to estimate the model order by

employing some heuristic and the performance of the algorithm is sensitive to the model order. In contrast, AST and the Lasso based approximation discussed in Chapter 5 only needs a rough estimate of the noise variance. The experiments in Chapter 5 provides the true model order to Matrix Pencil, MUSIC and Cadzow methods, but only an estimate of noise variance for AST. However, AST still compares favorably to the classical line spectral methods.

In contrast to linear prediction methods, a number of authors [32, 72, 12] have suggested using compressive sensing and viewing the frequency estimation as a sparse approximation problem. For instance, [72] notes that the Lasso based method has better empirical localization performance than the popular MUSIC algorithm. However, the theoretical analysis of this phenomenon is complicated because of the need to replace the continuous frequency space by an oversampled frequency grid. Compressive sensing based results (see, for instance, [47]) need to carefully control the incoherence of their linear maps to apply off-the-shelf tools from compressed sensing. It is important to note that the performance of our algorithm improves as the grid size increases. But this seems to contradict conventional wisdom in compressed sensing because our design matrix Φ becomes more and more coherent, and limits how fine we can grid for the theoretical guarantees to hold.

Directly working in the continuous parameter space circumvents the problems in the conventional compresssive sensing analysis, and allows us step away from such notions as coherence, focusing on the geometry of the atomic set as the critical feature. By showing that the continuous approach is the limiting case of the Lasso based methods using the convergence of the corresponding atomic norms, we could justify denoising line spectral signals using Lasso on a large grid. Furthemore, Candès and Fernandez-Granda [20] showed that our SDP formulation exactly recovers the correct frequencies in the noiseless case.

More recently, approaches based on convex optimization have gained favor and have been demonstrated to perform well on a variety of spectrum estimation tasks [72, 12, 3,

111]. These convex programming methods restrict the frequencies to lie on a finite grid of points and view line spectral signals as a sparse combination of single frequencies. While these methods are reported to have significantly better localization properties than subspace methods (see for example, [72]) and admit fast and robust algorithms, they have two significant drawbacks. First, while finer gridding may lead to better performance, very fine grids are often numerically unstable. Furthermore, traditional compressed sensing theory does not adequately characterize the performance of fine gridding in these algorithms as the dictionary becomes highly coherent.

Some very recent work [6, 20, 19] bridges the gap between the performant discretized algorithms and continuous subspace approaches by developing a new theory of convex relaxations for infinite continuous dictionary of frequencies. Chapter 5 demonstrates empirically that the algorithm proposed in the chapter compares favorably with both the classical and recent convex approaches which assume the frequencies are on an oversampled DFT grid. We saw that we can derive a weak but asymptotically consistent convergence rate with no assumption about the separation between frequencies. When the frequencies are well separated, we saw that much faster convergence rates can be achieved.

This work is closely related to recent results established by Candès and Fernandez-Granda [20] on exact recovery using convex methods and their recent work [19] on exploiting the robustness of their dual polynomial construction to show super-resolution properties of convex methods. The total variation norm formulation used in [19] is equivalent to the atomic norm specialized to the line spectral estimation problem.

Robustness bounds were established both in earlier work [6] and in the work of Candès and Fernandez-Granda [19]. In [6], a slow convergence rate was established with no assumptions about the separation of frequencies in the true signal. In [19], the authors provide guarantees on the L_1 energy of error in the frequency domain in the case that the frequencies are well

separated. The noise is assumed to be adversarial with a small L_1 spectral energy. In contrast, we saw near minimax denoising error under Gaussian noise in this chapter. It is also not clear that there is a computable formulation for the optimization problem analyzed in [19]. While the guarantees the authors derive in [19] are not comparable with our results, several of their mathematical constructions are used in the proofs here.

Additional recent work derives conditions for approximate support recovery under the Gaussian noise model using the Beurling-Lasso [2]. There, the authors show that there is a true frequency in the neighborhood of every estimated frequency with large enough amplitude. The Beurling-Lasso is equivalent to the atomic norm algorithm in this chapter. A more recent paper by Fernandez-Granda[52] improves this result by giving conditions on recoverability in terms of the true signal instead of the estimated signal and prove a theorem similar to Theorem 3.3, but use a worst case L_2 bound on the noise samples. This chapter improves these results in Theorem 3.3, providing tighter guarantees under the Gaussian noise model.

3.9 Conclusion

The Atomic norm formulation of line spectral estimation provides several advantages over prior approaches. Performing the analysis in the continuous domain permits deriving simple closed form rates using fairly straightforward techniques. This approach allowed us to circumvent some of the more complicated theoretical arguments that arise when using concepts from compressed sensing or random matrix theory.

This chapter demonstrated stability of atomic norm regularization by analysis of specific properties of the atomic set of moments and the associated dual space of trigonometric polynomials. The key to this analysis is the existence and properties of various trigonometric polynomials associated with signals with well separated frequencies.

4 System Identification

Identifying dynamical systems from noisy observation of their input-output behavior is of fundamental importance in systems and control theory. Often times models derived from physical first principles are not available to the control engineering, and computing a surrogate model from data is essential to the design of a control system. System identification from data is thus ubiquitous in problem domains ranging from process engineering, dynamic modeling of mechanical and aerospace systems, and systems biology. Though there are a myriad of approaches and excellent texts on the subject (see, for example [70]), there is still no universally agreed upon approach for this problem. One reason is that quantifying the interplay between system parameters, measurement noise, and model mismatch tends to be challenging.

This chapter draws novel connections between contemporary high-dimensional statistics, operator theory, and linear systems theory to prove consistent estimators of linear systems from small measurement sets. In particular, building on recent studies of *atomic norms* in estimation theory [30, 6], I propose a penalty function which encourages estimated models to have small McMillan degree.

A related family of system identification techniques use finite sample Hankel matrices to estimate dynamical system models, using either singular value decompositions (e.g, [104, 76]) or semidefinite programming [51, 69, 90, 84]. In all of these techniques, no statistical guarantees were given about the quality of estimation with finite noisy data, and it was difficult to determine how sensitive these methods were to the hidden system parameters or measurement noise. Moreover, since these problems were dealing with finite, truncated Hankel matrices, it is never certain if the size of the Hankel matrix is sufficient to reveal the true McMillan degree. Moreover, the techniques based on semidefinite programming are

challenging to scale to very large problems, as their complexity grows superlinearly with the number of measurements.

In contrast, the atomic norm regularizer proposed in this chapter is not only equivalent to the sum of the Hankel singular values (the Hankel nuclear norm), but is also well approximated by a finite dimensional, ℓ_1 minimization problem. I show that solving least-squares problems regularized by atomic norm is consistent, and scales gracefully with the stability radius, the McMillan degree of the system to be identified, and the number of measurements. Numerical experiments validate these theoretical underpinnings, and show that this method has great promise to provide concrete estimates on the hard limits of estimating linear systems.

Notation

We will adopt standard notation; $\mathbb{D}$ and $\mathbb{S}$ will denote respectively the open unit ball and the unit circle in the complex plane $\mathbb{C}$. $\mathcal{H}_2$ and $\mathcal{H}_\infty$ will denote the Hardy spaces of functions analytic outside $\mathbb{D}$, with the norms

$$\|f\|_{\mathcal{H}_2} = \tfrac{1}{2\pi}\int_0^{2\pi} |f(e^{i\theta})|^2 d\theta \quad \text{and} \quad \|f\|_{\mathcal{H}_\infty} = \sup_{z\in\mathbb{S}} |f(z)|$$

respectively. $\ell_2([a,b])$ will denote the set of square summable sequences on the integers in $[a,b]$.

4.1 Atomic Decompositions of Transfer Functions

In this chapter, we will concern ourselves with Single Input Single Output (SISO) systems. Suppose we wish to estimate a SISO, LTI system with transfer function $G_\star(z)$ from a finite collection of measurements $y = \Phi(G_\star)$. The set of all transfer functions is an infinite dimensional space, so reconstructing $G_\star$ from this data is ill-posed. In order to make it well

posed, a common regularization approach constructs a penalty function $\mathrm{pen}(\cdot)$ that encourages "low-complexity" models and solves the optimization problem

$$\underset{G}{\text{minimize}} \ \|\Phi(G) - y\|_2^2 + \mu \,\mathrm{pen}(G)\,. \tag{4.1}$$

This formulation uses the parameter μ to balance between model complexity and fidelity to the data. The least-squares cost can be modified to other convex loss functions if knowledge about measurement noise is available (as in [90, 77]), though in general it is less clear how to design a good penalty function.

In many applications, we know that the true model can be decomposed as a linear combination of very simple building blocks. For instance, as described in Chapter 2, sparse vectors can be written as short linear combinations of vectors from some discrete dictionary and low-rank matrices can be written as a sum of a few rank-one factors. Here, let us recall some of the preliminaries from Chapter 2. In [30], Chandrasekaran et al. proposed a universal heuristic for constructing regularizers based on such prior information. If we assumed that

$$G_\star = \sum_{i=1}^{r} c_i a_i\,, \text{ for some } a_i \in \mathcal{A}, c_i \in \mathbb{C}\,,$$

where $\mathcal{A}$ is an origin-symmetric set of atoms normalized to have unit norm and r is relatively small, then the appropriate penalty function is the atomic norm induced by the atomic set $\mathcal{A}$:

$$\|G\|_{\mathcal{A}} := \inf\left\{t \ : \ G \in t\,\mathrm{conv}(\mathcal{A})\right\} = \inf\left\{\sum_{a\in\mathcal{A}} |c_a| \ : \ G = \sum_{a\in\mathcal{A}} c_a a\right\}. \tag{4.2}$$

In [30], it is shown that minimizing the atomic norm subject to compressed measurements yielded the tightest known bounds for recovering many classes of models from linear measurements. Moreover, in Chapter 2, we saw the atomic norm regularizer in the context of

denoising problems and saw that it produces consistent estimates at nearly optimal estimation error rates for many classes of atoms.

Corresponding to the atomic norm, there is a dual atomic norm. For an atomic set $\mathcal{A}$, the dual norm is given by

$$\|z\|_{\mathcal{A}}^* = \sup_{a \in \mathcal{A}} \langle a, z \rangle \,.$$

Note that for this norm, we have the generalization of Hölder's inequality $\langle x, z \rangle \leq \|x\|_{\mathcal{A}} \|z\|_{\mathcal{A}}^*$. Moreover, note that

$$\alpha \|x\|_{\mathcal{A}'} \leq \|x\|_{\mathcal{A}} \leq \beta \|x\|_{\mathcal{A}'}$$

for some $\alpha \leq 1$ and $\beta \geq 1$ for all x if and only if

$$\beta^{-1} \|z\|_{\mathcal{A}'}^* \leq \|z\|_{\mathcal{A}}^* \leq \alpha^{-1} \|z\|_{\mathcal{A}'}^*$$

for all z. For more details, see Chapter 2

To apply these atomic norm techniques to system identification, we must first determine the appropriate set of atoms. For discrete time LTI systems with small McMillan degree, we can always decompose any finite dimensional, strictly proper system $G(z)$ as:

$$G(z) = \sum_{i=1}^{s} \frac{c_i}{z - a_i} \,.$$

via a partial fraction expansion. Hence, it makes sense that our set of atoms should be single-pole transfer functions. So, define the atomic set for linear systems by

$$\mathcal{A} = \left\{ \varphi_w(z) = \frac{1 - |w|^2}{z - w} \;:\; w \in \mathbb{D} \right\} .$$

The numerator is normalized so that the Hankel norm of each atom is 1. See the discussion in

Section 4.2 for precisely why this normalization is desirable.

The atomic norm penalty function associated with these atoms is

$$\|G(z)\|_{\mathcal{A}} = \inf \left\{ \sum_{w \in \mathbb{D}} |c_w| \ : \ G(z) = \sum_{w \in \mathbb{D}} \frac{c_w(1-|w|^2)}{z-w} \right\}, \tag{4.3}$$

where the summation implies that only a countable number of terms have nonzero coefficients c_w. This expression finds the decomposition of $G(z)$ into a linear combination of single pole systems such that the ℓ_1 norm, weighted by the norms of the single poles, is as small as possible.

With this penalty function in hand, we can now turn to analyzing its utility. In Section 4.2, I will first show that for most systems of interest $\|G\|_{\mathcal{A}}$ is a well-defined, bounded quantity. Moreover, we will see that the atomic norm is equivalent to the nuclear norm of the Hankel operator associated with G. Hence, the models that are preferred by our penalty function will have low-rank Hankel operators, and thus low McMillan degrees.

Section 4.2.3 describes practical algorithms for approximating atomic norm regularization problems for several classes of measurements. We will show that with finite data, our atomic norm minimization problem is well-approximated by a finite-dimensional ℓ_1 norm regularization problem. In particular, using specialized algorithms adapted to the solution of the LASSO [106], we can solve atomic norm regularization problems in time competitive with respect to techniques that regularize with the nuclear norm and SVD-based subspace identification methods. We will leave the actual analysis of the approximation by discretization to Chapter 5 on algorithms.

Finally, Section 4.3 analyzes the statistical performance of atomic norm minimization and shows that the proposed algorithm is asymptotically consistent over several measurement ensembles of interest. We will focus on sampling the transfer function on the unit circle and

present $\mathcal{H}_2$ error bounds in terms of the stability radius, Hankel singular values, $\mathcal{H}_\infty$ norm, and McMillan degree of the system to be estimated.

4.2 The Hankel Nuclear Norm and Atomic Norm Minimization

Let us first show that most LTI systems of interest do indeed have finite atomic norm, and, moreover, that the atomic norm is closely connected with the sum of the Hankel singular values.

4.2.1 Preliminaries: the Hankel operator

Recall that the *Hankel operator*, Γ_G, of the transfer function G is defined as the mapping from the past to the future under the transfer function G. Given a signal u supported on $(-\infty, -1]$, the output under G is given by $g * u$ where "$*$" denotes convolution and g is the impulse response of G:

$$G(z) = \sum_{k=1}^{\infty} g_k z^{-k} \,.$$

Γ_G is then simply the projection of $g * u$ onto $[0, \infty)$. An introduction to Hankel operators in control theory can be found in [48, Chapter 4] or [109, Chapter 7].

The *Hankel norm* of G is the operator norm of Γ_G considered as an operator mapping $\ell_2(-\infty, -1]$ to $\ell_2[0, \infty)$. The *Hankel nuclear norm* of G is the nuclear norm (aka the trace norm or Schatten 1-norm) of Γ_G. To be precise, an operator T is in the *trace class* S_1 if the trace of $(T^*T)^{1/2}$ is finite. This implies first that T is a compact operator and admits a singular value decomposition

$$T(f) = \sum_{i=1}^{\infty} \sigma_i \langle v_i, f \rangle u_i \,.$$

The sequence σ_i are called the *Hankel singular values* of T. Moreover, the Schatten 1-norm of

T is given by

$$\|T\|_1 = \operatorname{trace}\left((T^*T)^{1/2}\right) = \sum_{i=1}^{\infty} \sigma_i \,.$$

4.2.2 The atomic norm is equivalent to the Hankel nuclear norm

The rank of the Hankel operator determines the McMillan degree of the linear system defined by G. Rank minimization is notoriously computationally challenging (see [84] for a discussion), and we don't expect to be able to directly penalize the norm of the Hankel operator in implementations. Thus, as is common, a reasonable heuristic for minimizing the rank of the Hankel operator would be to minimize the sum of the Hankel singular values, i.e., to minimize the Schatten 1-norm of the Hankel operator. For rational transfer functions, we can compute the Hankel nuclear norm via a balanced realization [109]. On the other hand, while the maximal Hankel singular value can be written variationally as an LMI, we are not aware of any such semidefinite programming formulations for the Hankel nuclear norm.

The following theorem provides a path towards minimizing the Hankel nuclear norm, minimizing the atomic norm $\|G(z)\|_{\mathcal{A}}$ as a proxy. Indeed, from the view of Banach space theory, the atomic norm is *equivalent* to the Hankel nuclear norm.

Theorem 4.1. *Let $G \in \mathcal{H}_2$. Then Γ_G is trace class if and only if there exists a sequence $\{\lambda_k\} \in \ell_1$ and a sequence $\{w_k\}$ with $w_k \in \mathbb{D}$ such that*

$$g(z) = \sum_{i=1}^{\infty} \lambda_k \frac{1 - |w_k|^2}{z - w_k} \,. \tag{4.4}$$

Moreover, we have the following chain of inequalities

$$\tfrac{\pi}{8}\|G\|_{\mathcal{A}} \leq \|\Gamma_G\|_1 \leq \|G\|_{\mathcal{A}} \tag{4.5}$$

where $\|G\|_{\mathcal{A}}$ is given by (4.2).

Proof Outline Theorem 4.1 follows by carefully combining several different results from operator theory. Peller first showed that transfer functions with trace class Hankel operators formed a *Besov space* [80]. Peller's argument can be found in his book [79]. The atomic decomposition of such operators is due to Coifman and Rochberg [36]. The norm bounds (4.5) were proven by Bonsall and Walsh [10]. There they show that the $\frac{\pi}{8}$ is the best possible lower bound. They also show that if $\|\Gamma_g\|_1 \leq C\|g\|_{\mathcal{A}}$ for all g, then C must be at least $\frac{1}{2}$, so the chain of inequalities is nearly optimal. A concise presentation of the full argument can be found in [78]. A modern perspective using the theory of reproducing kernels can be found in [110].

Theorem 4.1 asserts that a transfer function has a finite atomic norm if and only if the sum of its Hankel singular values is finite. In particular, this means that every rational transfer function has a finite atomic norm. More importantly, the atomic norm is equivalent to the Hankel nuclear norm. Thus if we can approximately solve atomic norm-minimization, we can approximately solve Hankel nuclear norm minimization and vice-versa.

4.2.3 System Identification using Atomic Norms

From here on, let us assume that the $G_\star$ that we seek to estimate has all of its poles of magnitude at most ρ (we will call ρ the *stability radius*, and treat it as a known parameter). Let $\mathbb{D}_\rho$ denote the set of all complex numbers with norm at most ρ. Note that if $G_\star$ has stability radius ρ then

$$\|G\|_{\mathcal{A}} := \inf \left\{ \sum_{w \in \mathbb{D}_\rho} |c_w| \ : \ G(z) = \sum_{w \in \mathbb{D}_\rho} \frac{c_w(1-|w|^2)}{z-w} \right\}.$$

That is, we can restrict the set of atoms to only be those single pole systems with stability radius equal to ρ. For the remainder of this manuscript, assume that $\mathcal{A}$ only consists of such

single pole systems.

In what follows, let us focus on linear measurement maps. Let $\mathcal{L}_i : \mathcal{H} \mapsto \mathbb{C}$ be a linear functional that serves as a measurement operator for the system $G(z)$. Many maps of interest can be phrased as linear functionals of the transfer function,

1. Samples of the frequency response $\mathcal{L}_k(G) := G(e^{i\theta_k})$ for $k = 1, \ldots, n$. From a control theoretic perspective, this measurement operator corresponds to measuring the gain and phase of the linear system at different frequencies.

2. Samples of the impulse response, $\mathcal{L}_k(G) := g_{i_k}$ for $k = 1, \ldots, n$ and $i_k \in [1, \infty)$.

3. Convolutions of the impulse response with a pseudorandom signal u_k: $\mathcal{L}_k(G) := \sum_{j=1}^{\infty} g_j u_{k-j}$.

In all of these cases, consider the problem

$$\underset{G}{\text{minimize}} \; \tfrac{1}{2} \sum_{i=1}^{n} |\mathcal{L}_i(G) - y_i|^2 + \mu \|G\|_{\mathcal{A}} \,. \tag{4.6}$$

This problem is equivalent to the constrained, semi-infinite programming problem

$$\begin{array}{ll} \text{minimize}_{x,G} & \tfrac{1}{2} \sum_{k=1}^{n} |x_k - y_k|^2 + \mu \sum_{w \in \mathbb{D}_\rho} |c_w| \\ \text{subject to} & x_k = \mathcal{L}_k(G) \;\; \text{for } i = 1, \ldots, n \\ & G = \sum_{w \in \mathbb{D}_\rho} \frac{c_w (1 - |w|^2)}{z - w} \end{array}$$

Eliminating the equality constraint gives yet another equivalent formulation

$$\begin{array}{ll} \text{minimize}_{x} & \tfrac{1}{2} \sum_{k=1}^{n} |x_k - y_k|^2 + \mu \sum_{w \in \mathbb{D}_\rho} |c_w| \\ \text{subject to} & x_k = \sum_{w \in \mathbb{D}_\rho} c_w \mathcal{L}_k \left(\frac{1 - |w|^2}{z - w} \right) \;\; \text{for } i = 1, \ldots, n \,. \end{array} \tag{4.7}$$

Note that in this final formulation, our decision variable is x, a finite dimensional vector, and c_w, the coefficients of the atomic decomposition. The infinite dimensional variable G has been eliminated. Let us define a norm on $\mathbb{R}^n$ based on the formulation (4.7)

$$\|x\|_{\mathcal{L}(\mathcal{A})} = \inf \left\{ \sum_{w \in \mathbb{D}_\rho} |c_w| \ : \ x_i = \sum_{w \in \mathbb{D}_\rho} c_w \mathcal{L}_i \left(\frac{1-|w|^2}{z-w} \right) \right\} .$$

Then we see that problem (4.6) is equivalent to the denoising problem

$$\underset{x}{\text{minimize}} \, \tfrac{1}{2}\|x-y\|_2^2 + \mu \|x\|_{\mathcal{L}(\mathcal{A})} . \tag{4.8}$$

Note that the first term is simply the squared Euclidean distance between y and x in $\mathbb{R}^n$. The second term is an atomic norm on $\mathbb{R}^n$ induced by the linear map of the set of transfer functions via the measurement operator $\mathcal{L}$.

In order to tractably solve (4.6), we will need computational schemes for approximating $\|x\|_{\mathcal{L}(\mathcal{A})}$. Such computational considerations are treated in Chapter 5. In that chapter, I show that a sufficiently fine discretization of the unit disk can yield a good approximation for the solution. In particular, we will need the following proposition proved in Chapter 5.

Proposition 4.2. *Let $\mathbb{D}_\rho^{(\epsilon)}$ be a finite subset of the unit disc such that for any $w \in \mathbb{D}_\rho$ there exists a $v \in \mathbb{D}_\rho^{(\epsilon)}$ satisfying $|w-v| \leq \epsilon$. Define*

$$\|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)} = \inf \left\{ \sum_{w \in \mathbb{D}_\rho^{(\epsilon)}} |c_w| \ : \ x_i = \sum_{w \in \mathbb{D}_\rho^{(\epsilon)}} c_w \mathcal{L}_i \left(\frac{1-|w|^2}{z-w} \right) \right\} .$$

Then there exists a constant $C_\epsilon \in [0,1]$ such that

$$C_\epsilon \|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)} \leq \|x\|_{\mathcal{L}(\mathcal{A})} \leq \|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)} .$$

The set $\mathbb{D}_\rho^{(\epsilon)}$ is called an ϵ-net for the set $\mathbb{D}_\rho$. We show in chapter 5 that when $\mathcal{L}_k(G) = G(e^{i\theta_k})$, C_ϵ is at least $(1 - \frac{16\rho\epsilon}{\pi(1-\rho)})$. Other measurement ensembles can be treated similarly. Chapter 5 gives a way to solve (4.8) approximately.

4.3 Statistical Bounds

Let $\mathcal{L}_i : \mathcal{H} \mapsto \mathbb{C}$ be a linear functional that serves as a measurement operator for the system $H(z)$. In this section, let us suppose that we obtain noisy measurements of the form

$$y_i = \mathcal{L}_i\left(H(z)\right) + \omega_i \qquad i = 1, \ldots, n.$$

where ω_i is a noise sequence consisting of independent, identically distributed random variables. In this section, we will specialize our results to the case where $\mathcal{L}$ returns samples from the frequency response at uniformly spaced frequencies:

$$\mathcal{L}_k(H(z)) = H(z_k), \qquad z_k = e^{\frac{2\pi i k}{m}},\ k = 1, \ldots, n.$$

Our goal in this section is to prove that solving the DAST optimization problem yields a good approximation to the transfer function we are probing. The following theorem provides a precise statistical guarantee on the performance of our algorithm.

Before going to the proof, let us recall the optimality conditions of atomic soft thresholding proved in Chapter 2, reproduced here for convenience:

Theorem 4.3. *Let $\mathcal{Q} \subset \mathbb{R}^n$ be an arbitrary set of atoms. Suppose that we observe $y = x_\star + \omega$ where $\omega \sim \mathcal{N}(0, \sigma^2 I)$. Let $\hat{x}$ denote the optimal solution of*

$$\underset{x}{\text{minimize}}\ \tfrac{1}{2}\|x - y\|_2^2 + \mu\|x\|_{\mathcal{Q}}$$

with $\mu \geq \|\omega\|_{\mathcal{Q}}^*$. *Then we have*

$$\|\hat{x} - x_\star\|_2^2 \leq 2\mu \|x_\star\|_{\mathcal{Q}} \tag{4.9}$$

$$\|\hat{x}\|_{\mathcal{Q}} \leq \|x_\star\|_{\mathcal{Q}} + \mu^{-1} \langle \omega, \hat{x} - x_\star \rangle \,. \tag{4.10}$$

We will use these inequalities in the proof of the main theorem.

Theorem 4.4. *Let* $G_\star$ *be a strictly proper transfer function with bounded Hankel nuclear norm. Suppose the noise sequence* ω_i *is i.i.d. Gaussian with mean zero and variance* σ^2. *Choose* $\delta \in (0,1)$ *and set* $\epsilon = \frac{\pi(1-\rho)\delta}{16\rho}$. *Let* $\mathbb{D}_\rho^{(\epsilon)}$ *be as in Proposition 4.2 and let* $\hat{c}$ *be the optimal solution of (5.29) with*

$$\mu = 2\sigma \sqrt{n \log\left(\frac{11\rho^2}{\delta(1-\rho)}\right)}\,.$$

Set $\hat{G}(z) = \sum_{w \in \mathbb{D}_\rho^{(\epsilon)}} \hat{c}_w \frac{1-|w|^2}{z-w}$. *Then if the set of vectors* $\{\mathcal{L}(\varphi_a) \in \mathbb{R}^n \ : \ a \in \mathbb{D}_\rho^{(\epsilon)}\}$ *spans* $\mathbb{R}^n$, *we have*

$$\|\hat{G}(z) - G_\star(z)\|_{\mathcal{H}_2}^2 \leq 186 \frac{1+\rho}{1-\rho} \left(\sqrt{\sigma^2 \log\left(\frac{11\rho^2}{(1-\rho)\epsilon}\right)} \sqrt{\frac{\|\Gamma_{G_\star}\|_1^2}{n(1-\delta)^2}} + \frac{4\|\Gamma_{G_\star}\|_1^2}{\pi n (1-\delta)^2} \right)$$

with probability $1 - e^{-o(n)}$.

Proof. To upper bound the $\mathcal{H}_2$ norm, let us use some properties of functions which admit atomic decompositions. Note that for any function, $H(z) = \sum_{w \in \mathbb{D}} c_w \varphi_w(z)$ with $\|H\|_{\mathcal{A}} =$

$\sum_{w\in\mathbb{D}}|c_w|$, we have, for any $z_1, z_2 \in \mathbb{S}$,

$$\begin{aligned}|H(z_1)|^2 - |H(z_2)|^2 &= (|H(z_1)| - |H(z_2)|)(|H(z_1)| + |H(z_2)|) \\ &\leq 2|H(z_1) - H(z_2)|\|H\|_{\mathcal{H}_\infty} \\ &\leq 2\left(\sum_{w\in\mathbb{D}}|c_w||\varphi_w(z_1) - \varphi_w(z_2)|\right)\left(\sum_{w\in\mathbb{D}}|c_w|\|\varphi_w\|_{\mathcal{H}\infty}\right), \end{aligned} \tag{4.11}$$

where the last inequality is just the triangle inequality. Now, to upper bound the first term in (4.11), we could connect the distance between the poles to the distance between the atoms. Let $z_j = e^{i\theta_j}, j = 1, 2$. Then,

$$\begin{aligned}|\varphi_a(z_1) - \varphi_a(z_2)| &= \left|\frac{1-|a|^2}{z_1 - a} - \frac{1-|a|^2}{z_2 - a}\right| \\ &\leq (1-|a|^2)\left|\frac{z_1 - z_2}{(z_1 - a)(z_2 - a)}\right| \\ &\leq \frac{(1-|a|^2)}{(1-|a|)^2}|z_1 - z_2| \leq \frac{1+\rho}{1-\rho}|\theta_1 - \theta_2|\,.\end{aligned}$$

Next, we could upper the $\mathcal{H}_\infty$ norm (4.11) globally. Note that for any $a \in D$, $\|\varphi_a(z)\|_{\mathcal{H}_\infty} \leq 2$. In fact, for $z = \exp(i\theta)$, $|\frac{1-|a|^2}{z-a}| \leq |\frac{1-|a|^2}{1-|a|}| \leq 1 + |a| \leq 2$. Using these computations, we finally have

$$\begin{aligned}|H(z_1)|^2 - |H(z_2)|^2 &\leq 4\frac{1+\rho}{1-\rho}\left(\sum_{w\in\mathbb{D}}|c_w|\right)^2|\theta_1 - \theta_2| \\ &= 4\frac{1+\rho}{1-\rho}\|H\|_{\mathcal{A}}^2|\theta_1 - \theta_2|\,.\end{aligned}$$

Let $\Delta = G_\star - \hat{G}$ and $\theta_k = \frac{2\pi k}{n}$. Then we can bound the norm of Δ as

$$\begin{aligned}
\|\Delta\|_{\mathcal{H}_2}^2 &= \frac{1}{2\pi}\int_0^{2\pi} |\Delta(e^{i\theta})|^2 d\theta \\
&= \frac{1}{2\pi}\sum_{k=0}^{n-1}\int_{\theta_k}^{\theta_{k+1}} |\Delta(e^{i\theta})|^2 d\theta \\
&\leq \frac{1}{2\pi}\sum_{k=0}^{n-1}\int_{\theta_k}^{\theta_{k+1}} \left(|\Delta(e^{i\theta_k})|^2 + 4\frac{1+\rho}{1-\rho}\|\Delta\|_{\mathcal{A}}^2|\theta-\theta_k|\right) d\theta \\
&= \frac{1}{n}\sum_{k=0}^{n-1} |\Delta(e^{i\theta_k})|^2 + \frac{4\pi}{n}\frac{1+\rho}{1-\rho}\|\Delta\|_{\mathcal{A}}^2 .
\end{aligned}$$

The inequality here follows from our preceding argument.

Now, in this expression, we need to both upper bound the size of Δ on the measured frequencies and its atomic norm. The following chain of inequalities bounds the latter in terms of the former:

$$\begin{aligned}
\|\Delta\|_{\mathcal{A}} &\leq \|G_\star\|_{\mathcal{A}} + \|\hat{G}\|_{\mathcal{A}} && (4.12)\\
&= \|G_\star\|_{\mathcal{A}} + \|\mathcal{L}(\hat{G})\|_{\mathcal{L}(\mathcal{A}_\epsilon)} && (4.13)\\
&\leq \|G_\star\|_{\mathcal{A}} + \|\mathcal{L}(G_\star)\|_{\mathcal{L}(\mathcal{A}_\epsilon)} + \mu^{-1}\langle\omega, \mathcal{L}(\hat{G}-G_\star)\rangle && (4.14)\\
&\leq \|G_\star\|_{\mathcal{A}} + (1-\delta)^{-1}\|\mathcal{L}(G_\star)\|_{\mathcal{L}(\mathcal{A})} + \mu^{-1}\langle\omega, \mathcal{L}(\hat{G}-G_\star)\rangle && (4.15)\\
&\leq \frac{2-\delta}{1-\delta}\|G_\star\|_{\mathcal{A}} + \mu^{-1}\langle\omega, \mathcal{L}(\hat{G}-G_\star)\rangle && (4.16)\\
&\leq \frac{2-\delta}{1-\delta}\|G_\star\|_{\mathcal{A}} + \mu^{-1}\|\omega\|_2 \left(\sum_{k=1}^{n-1} |\Delta(e^{i\theta_k})|^2\right)^{1/2} . && (4.17)
\end{aligned}$$

(4.12) is the triangle inequality. (4.13) follows from how we defined $\hat{G}$. (4.14) follows from Theorem 4.3. (4.15) follows from Proposition 5.14. (4.16) follows because $\|\mathcal{L}(H)\|_{\mathcal{L}(\mathcal{A})} \leq \|H\|_{\mathcal{A}}$ for all linear maps $\mathcal{L}$ and transfer functions H. (4.17) is Hölder's inequality. Note that

the quantity $\|\mathcal{L}(G_\star)\|_{\mathcal{L}(\mathcal{A}_\epsilon)}$ could be infinite if the set $\{\mathcal{L}(\phi_a) \ : \ a \in \mathbb{D}_\rho^{(\epsilon)}\}$ does not span $\mathbb{R}^n$. This is precisely what leads to us including this assumption in the theorem statement.

To bound the size of Δ on the frequency grid, use Theorem 4.3:

$$\frac{1}{n}\sum_{k=1}^{n-1}|\Delta(e^{i\theta_k})|^2 \leq \frac{2\mu}{n}\|\mathcal{L}(G_\star)\|_{\mathcal{L}(\mathcal{A}_\epsilon)} \leq \frac{2\mu}{n}(1-\delta)^{-1}\|G_\star\|_{\mathcal{A}}\,.$$

Let $\omega \sim \mathcal{N}(0,\sigma^2 I_n)$. Using the well known upper bound for maximum of Gaussian variables (see, for example [65]), we have

$$\mathbb{E}[\|\omega\|^*_{\mathcal{L}(\mathcal{A}_\epsilon)}] = \mathbb{E}\left[\sup_{a\in\mathbb{D}_\rho^{(\epsilon)}}\langle\mathcal{L}(\varphi_a),\omega\rangle\right] \leq \sigma\left(\sup_{a\in\mathbb{D}_\rho}\|\mathcal{L}(\varphi_a)\|_2\right)\sqrt{2\log|\mathbb{D}_\rho^{(\epsilon)}|}\,.$$

Now, $\|\mathcal{L}(\varphi_a)\|_2 \leq \sqrt{n}\|\varphi_a\|_{\mathcal{H}_\infty} \leq 2\sqrt{n}$. Moreover, by a simple volume argument, $|\mathbb{D}_\rho^{(\epsilon)}| \leq \frac{1024\rho^4}{\pi^2(1-\rho)^2\delta^2}$. To see this, suppose $\mathcal{S}$ is a maximal set of points on $\mathbb{D}_\rho$ which are separated by at least τ. The maximal size of such a set is at most $\frac{4\rho^2}{\tau^2}$. Moreover, $|\mathcal{S}| \geq |\mathbb{D}_\rho^{(\delta)}|$. Now set $\tau = \epsilon = \frac{\pi(1-\rho)\delta}{16\rho}$. In particular, note that we have $\mathbb{E}[\|\omega\|^*_{\mathcal{L}(\mathcal{A}_\epsilon)}] = \frac{1}{2}\mu$.

Now we can put all of the ingredients together.

$$\begin{aligned}
\|\Delta\|^2_{\mathcal{H}_2} &\leq \left(1+\frac{8\pi\|\omega\|_2^2}{\mu^2}\frac{1+\rho}{1-\rho}\right)\frac{1}{n}\sum_{k=1}^{n-1}|\Delta(e^{i\theta_k})|^2 \\
&\qquad + \frac{16\pi}{n(1-\delta)^2}\frac{1+\rho}{1-\rho}\|G_\star\|^2_{\mathcal{A}} \\
&\leq \left(1+2\pi\frac{1+\rho}{1-\rho}\right)\frac{8\sigma}{1-\delta}\sqrt{\frac{2\log(\frac{1024\rho^4}{\pi^2(1-\rho)^2\delta^2})}{n}}\|G_\star\|_{\mathcal{A}} \\
&\qquad + \frac{16\pi}{n(1-\delta)^2}\frac{1+\rho}{1-\rho}\|G_\star\|^2_{\mathcal{A}} \\
&\leq 59\frac{1+\rho}{1-\rho}\left(\sqrt{4\sigma^2\log\left(\frac{11\rho^2}{(1-\rho)\delta}\right)}\sqrt{\frac{\|G_\star\|^2_{\mathcal{A}}}{n(1-\delta)^2}} + \frac{\|G_\star\|^2_{\mathcal{A}}}{n(1-\delta)^2}\right)
\end{aligned}$$

as desired.

Applying the inequality $\|G_\star\|_{\mathcal{A}} \leq \frac{8}{\pi}\|\Gamma_{G_\star}\|_1$ completes the proof. □

Corollary 4.5. *There is a quantity C depending on ρ and σ such that for sufficiently large n*

$$\|\hat{G}(z) - G_\star(z)\|_{\mathcal{H}_2}^2 \leq C\|\Gamma_{G_\star}\|_1 n^{-\frac{1}{2}}$$

with probability exceeding $1 - e^{-o(n)}$.

Now, let us look the consequence of the theorem. First of all, the right hand side is a parameter of the number of samples, the Hankel nuclear norm of the true system, and the stability radius of the true system. Also, if the McMillan degree of $G_\star(z)$ is d, then we can upper bound the Hankel nuclear norm by the product of the McMillan degree and the Hankel norm of $G_\star$: $\|\Gamma_{G_\star}\|_1 \leq d\|\Gamma_{G_\star}\|$. Second, note that as n tends to infinity, the right hand side tends to zero. In particular, this means that the discretized algorithm is consistent, and we can quantify the worst case convergence rate.

4.4 Conclusion

By using the atomic norm framework of [30], this chapter posits a reasonable regularizer for linear systems. We looked at a computational method to handle such a regularizer, and analyzed its statistical performance. Since it is closely connected to the Hankel nuclear norm but is computationally more practical, the atomic norm could be useful in a variety of practical implementations and also in theoretical analysis.

5 Algorithms

In the preceding chapters, we concentrated on the theoretical performance of regularization using an atomic norm penalty, but did not show how one may compute it. The problems of atomic norm decomposition and regularization are convex optimization problems with linear and quadratic objectives, and can be efficiently solved if we can test membership in the constraint sets efficiently. The constraint sets of the primal and dual problems are the sublevel sets of the atomic and the dual atomic norm. Therefore, it is sufficient to develop efficient characterizations of the the unit ball of the atomic norm.

For the special case of Fourier measurements, the atomic norm ball has an exact semidefinite characterization which I will derive in this chapter. The positive case is classical and comes from moment theory and the dual theory of positive polynomials. I will review the results for the positive case and provide the proofs for completeness. We will also see a semidefinite characterization for the more general complex case using these results. I then provide a reasonably fast method for solving this SDP via the Alternating Direction Method of Multipliers (ADMM) [4, 13]. The ADMM implementation given in this chapter can solve instances with a thousand observations in a few minutes.

Discretization essentially reduces to solving a Lasso problem on an overcomplete grid. The proofs for discretized atomic soft thresholding (DAST) demonstrate why Lasso is often successful even for off-grid data. In fact, we will see that solving the Lasso problem on an oversampled grid of frequencies approximates the solution of the atomic norm minimization problem to a resolution sufficiently high to guarantee excellent mean-squared error (MSE). For line spectral estimation, the gridded problem reduces to the Lasso, and by leveraging the Fast Fourier Transform (FFT), can be rapidly solved with freely available software such as SpaRSA [106]. A Lasso problem with thousands of observations can be solved in under a

second using Matlab on a laptop. The prediction error and the localization accuracy for line spectral estimation both increase as the oversampling factor increases, even if the actual set of frequencies in the line spectral signal are off the Lasso grid.

In this chapter, we will also see the experimental results using these algorithms on Line Spectral Signals. Section 5.6 compares and contrasts AST and Lasso, with classical line spectral algorithms including MUSIC [89],and Cadzow's [15] and Matrix Pencil [60] methods. The experiments indicate that both AST and the Lasso approximation outperform classical methods in low SNR even when we provide the exact model order to the classical approaches. Moreover, AST has the same complexity as Cadzow's method, alternating between a least-squares step and an eigenvalue thresholding step. The discretized Lasso-based algorithm has even lower computational complexity, consisting of iterations based upon the FFT and simple linear time soft-thresholding.

Summary and Organization of this chapter

The goal in this chapter is to develop an algorithm to solve the Atomic Soft Thresholding (AST) problem, defined in Chapter 2:

$$\underset{x}{\text{minimize}} \frac{1}{2}\|y - x\|_2^2 + \tau\|x\|_{\mathcal{A}}, \tag{5.1}$$

and its corresponding dual problem:

$$\begin{aligned} &\underset{q}{\text{maximize}} \frac{1}{2}\left(\|y\|_2^2 - \|y - \tau q\|_2^2\right) \\ &\text{subject to } \|q\|_{\mathcal{A}}^* \leq 1. \end{aligned} \tag{5.2}$$

For certain atomic sets, we can compute the atomic norms efficiently. We will develop computable characterizations of the atomic norm and its dual for line spectral atomic sets

defined in Chapter 3. First, let us recall the definition of the trigonometric monomial $a(f)$, which form the basic atoms for Line Spectral Estimation:

$$a(f) = \begin{pmatrix} 1 \\ e^{i2\pi f} \\ \vdots \\ e^{i2\pi f(n-1)} \end{pmatrix}.$$

The first part of this chapter gives a computable algebraic characterization of the atomic norm balls of the following atomic sets derived from these atoms:

$$\mathcal{A}_+ = \{a(f) \mid f \in \mathbb{T}\} \tag{5.3}$$

$$\mathcal{A} = \{a(f)\exp(i2\pi\phi) \mid f, \phi \in \mathbb{T}\}. \tag{5.4}$$

The first set $\mathcal{A}_+$ is a one dimensional manifold called the trigonometric moment curve and the second set $\mathcal{A}$ is the two dimensional manifold corresponding to a phase-symmetric trigonometric moment curve. The unit norm ball of the atomic sets are precisely the convex hulls of the atomic sets. Section 5.1 examines these atomic sets and describes the classical characterization of the conical hull of $\mathcal{A}_+$ which is called the moment cone. This corresponds to allowable observations for the case of line spectral estimation with nonnegative amplitudes, or in other words, valid trigonometric moments of a positive measure on the torus $\mathbb{T}$.

Based on these classical results, I derive semidefinite characterizations of the atomic sets in Section 5.2. As a straightforward consequence, we can write down a semidefinite characterization of $\mathrm{conv}(\mathcal{A}_+)$. Before stating the result, we will need a bit of notation. Define

the map $T_n : \mathbb{C}^n \to \mathbb{C}^{n\times n}$ which creates a Hermitian Toeplitz matrix out of its input, that is

$$T_n(x) = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ x_2^* & x_1 & \dots & x_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^* & x_{n-1}^* & \dots & x_1 \end{bmatrix}$$

In the following, the subscript in T_n may be supressed unless the idea is to explicitly characterize the dependence on n. So we will write it as $T(x)$ or Tx. Also, denote the corresponding adjoint map by T^*.

Theorem 5.1 (SDP for Positive Trigonometric Moments)**.** *Suppose* $x = (x_0 \cdots x_{n-1})^T \in \mathbb{C}^n$. *Then,*

$$\|x\|_{\mathcal{A}_+} = \begin{cases} x_0, & Tx \succeq 0, \\ +\infty, & \textit{otherwise.} \end{cases}$$

Note that this is an example when atomic norm is not really a norm on F^n, but is nevertheless useful. In contrast to $\mathrm{conv}(\mathcal{A}_+)$, the semidefinite characterization of $\mathrm{conv}(\mathcal{A})$ requires some calculation. The characterization of the moment cone allows us to describe the convex hull of the general atomic moments $\mathrm{conv}(\mathcal{A})$ and we will derive the following semidefinite characterization:

Theorem 5.2 (SDP for General Trigonometric Moments)**.** *For* $x \in \mathbb{C}^n$*,*

$$\|x\|_{\mathcal{A}} = \inf \left\{ \tfrac{1}{2n}\operatorname{tr}(T_n(u)) + \tfrac{1}{2}t \;\middle|\; \begin{bmatrix} T_n(u) & x \\ x^* & t \end{bmatrix} \succeq 0 \right\}. \tag{5.5}$$

The characterization of atomic norm ball of $\mathcal{A}$ can also be derived by working on the dual

problem, as we shall see in Section 5.3. In fact, for any x, the optimization problem

$$\begin{aligned} &\underset{q}{\text{minimize}} \ \langle q, x \rangle \\ &\text{such that } \|q\|_{\mathcal{A}}^* \leq 1. \end{aligned} \tag{5.6}$$

is the dual characterization which has an optimum value $\|x\|_{\mathcal{A}}$. For the atomic set of general trigonometric moments, the constraint set in Equation (5.6)

$$\{q \in \mathbb{C}^n \mid \|q\|_{\mathcal{A}}^* \leq 1\} = \left\{ q \in \mathbb{C}^n \;\middle|\; \sup_{a \in \mathcal{A}} \langle q, a \rangle \leq 1 \right\} \tag{5.7}$$

$$= \left\{ q \in \mathbb{C}^n \;\middle|\; \left| \sum_{j=0}^{n-1} q_j \exp(i 2\pi j f) \right| \leq 1, \text{ for all } f \in \mathbb{T} \right\} \tag{5.8}$$

corresponds precisely to the set of complex trigonometric polynomials with a maximum modulus of 1 and is characterized by the Bounded Real Lemma. We provide details of applying Alternating Directions Method of Multipliers (ADMM) for solving the SDP for the AST problem in Section 5.4.

When there are no efficient characterizations, we may resort to a discretized approach which may be thought of as an approximation to the atomic norm defined by equation (5.6) by relaxing the infinite linear constraints in the semi-infinite program to a finite set of inequalities. In Section 5.5, we will look at convergence rates of this approximation for the line spectral estimation problem and the system identification problem. In this case, the atomic soft thresholding problem can be approximated by solving a Lasso problem on an overcomplete grid.

A number of Prony-like techniques have been devised that are able to achieve excellent denoising and frequency localization even in the presence of noise. In Section 5.6, we will turn to comparing our ADMM implementation of SDP and DAST by Lasso with these classical

algorithms. The experiments in this section demonstrate that the proposed estimation algorithms outperform Matrix Pencil, MUSIC and Cadzow's methods. Both AST and the discretized Lasso algorithms obtain lower MSE compared to previous approaches, and the discretized algorithm is much faster on large problems. Furthermore, the semidefinite programming approach outperforms MUSIC [89] and Cadzow's technique [16], in terms of the localization metrics defined by parts (i), (ii) and (iii) of Theorem 3.3.

5.1 Preliminaries

As described in section 3.7.1, the observations in a spectral estimation problem may be regarded as trigonometric moments of a measure on the torus $\mathbb{T} = [0, 1]$.

Definition 5.3 (Trigonometric Moments)**.** *Given a measure μ defined on the torus $\mathbb{T}$, the mth trigonometric moment is defined by*

$$x_m = \int_{\mathbb{T}} e^{i2\pi mf} \mu(df) \tag{5.9}$$

By a simple reparametrization, the mth trigonometric moment may also be described as complex moments $\int_{\mathbb{T}} z^m \mu(dz)$ when we regard μ as a measure supported on the unit circle in the complex plane. This section will review two classical results for trigonometric moments due to *positive measures*.

The first theorem, due to Herglotz gives a complete characterization of the infinite sequence of Trigonometric moments for positive measures. Since the proof is not too long, I will provide it here.

Theorem 5.4 (Herglotz Theorem)**.** *The sequence of complex numbers $\{x_m\}_{m=-\infty}^{\infty}$ are trigonometric moments of a positive Borel measure μ on $\mathbb{T}$ if and only if the sequence $\{x_m\}_{m=-\infty}^{\infty}$ is*

positive definite.

Proof. A sequence $\{x_m\}$ is positive definite if for every $n \in \mathbb{N}$, and every sequence $(c_1, \ldots, c_n)$ of n complex numbers, $\sum_{k,l=1}^{n} x_{k-l} c_k c_l^* \geq 0$. Alternatively, we may write this condition as $T_n x \succeq 0$ for every $n \in \mathbb{N}$.

Suppose indeed the set $\{x_m\}$ is a sequence of trigonometric moments corresponding to some measure $\mu > 0$ on $\mathbb{T}$. Then, for any $n \in \mathbb{N}$,

$$\begin{aligned}\sum_{k,l=1}^{n} x_{k-l} c_k c_l^* &= \sum_{k,l=1}^{n} c_k c_l^* \int \exp(i2\pi(k-l)t)\mu(dt) \\ &= \int \left(\sum_{k=1}^{n} c_k \exp(i2\pi kt)\right)\left(\sum_{k=1}^{n} c_l \exp(i2\pi lt)\right)^* \mu(dt) \\ &= \int \left|\sum_{k=1}^{n} c_k \exp(i2\pi kt)\right|^2 \mu(dt) \geq 0.\end{aligned}$$

Conversely, if the sequence $\{x_m\}$ is positive semidefinite, then for any $t \in \mathbb{R}$, we have

$$X_n(t) = \frac{1}{n} \sum_{k,l=1}^{n} x_{k-l} e^{i2\pi(k-l)t} \geq 0$$

But, we have

$$X_n(t) = \sum_{k=-(n-1)}^{(n-1)} \left(1 - \frac{|k|}{n}\right) x_k \exp(i2\pi kt)$$

By Fourier Series theory, we have that for every $|k| \leq n$,

$$\left(1 - \frac{|k|}{n}\right) x_k = \int_0^1 X_n(t) \exp(i2\pi kt) dt \tag{5.10}$$

Define a sequence of measures μ_n given by

$$\mu_n(B) = \int_B X_n(t)dt$$

for every Borel subset $B \subset [0,1]$. The sequence of measures $\{\mu_n\}$ are tight and therefore by the Helly selection theorem, there exists a subsequence μ_{n_k} which converges weakly to a measure μ on $\mathbb{T}$. Finally, using 5.10 we conclude that $\{x_k\}$ is the sequence of moments for the measure μ. □

The Herglotz theorem characterizes arbitrary positive measures on the torus. However, for line spectral estimation, we are interested in *discrete* positive measures, i.e., measures composed of a finite number of atoms at $f_1, \ldots, f_k$, so we can write

$$\mu(f) = \sum_{l=1}^{k} c_l \xi_{f_l}$$

where ξ_f denotes the point measure at f and $\{c_l\} > 0$ are the amplitudes. The first n moments of such a measure concentrated on k atoms corresponds to a k simple combination of atomic moment sequences $a(f)$. In fact,

$$\begin{aligned} x_m &= \int_{\mathbb{T}} e^{i2\pi mf} \sum_{l=1}^{k} c_l \xi_{f_l}(df) \\ &= \sum_{l=1}^{k} c_l e^{i2\pi f_l m}. \end{aligned}$$

So, the corresponding moment vector $x = (x_0 \ \cdots \ x_{n-1}) \in \mathbb{C}^n$ given by

$$x = \sum_{l=1}^{k} c_l a(f_l)$$

is a simple combination of line spectral atoms. To see when such simple moment vectors actually correspond to a partially observed sequence of moments of a positive discrete measure, we will now review a second classical result due to Caratheodory and Toeplitz.

Theorem 5.5 (Caratheodory-Toeplitz theorem, [101, 27, 26])**.** *$x \in \mathbb{C}^n$ corresponds to the first n trigonometric moments of a measure μ (i.e., $x \in \mathrm{cone}(\mathcal{A}_+)$) if and only if $Tx \succeq 0$. Furthermore, if $k = \mathrm{rank}(Tx)$, there exists positive numbers $c_1, \ldots, c_k$ and $f_1, \ldots, f_k \in \mathbb{T}$ such that*

$$\mu = \sum_{l=1}^{k} c_l \delta(f - f_l)$$

so that

$$\begin{aligned} x &= \int_{\mathbb{T}} a(f)\mu(df) \\ &= \sum_{l=1}^{k} c_l a(f_l) \end{aligned}$$

Finally, when $k < n$, there is a unique extension of x to an infinite positive definite sequence and only a unique measure μ with x for its moments.

This theorem can be proved using Herglotz theorem [59] and the theorems on flat extensions of moment sequences studied in [37]. See [57] for an algebraic proof of this theorem. This theorem guarantees the existence of a unique discrete measure under easily verifiable conditions. Due to this guarantee, it is possible to write a set of recurrence relations connecting the moments and algebraically solve for the atoms of this measure, as demonstrated by Prony [81]. A straightforward corollary of Caratheodory's theorem is the following Vandermonde Decomposition for positive definite Toeplitz matrices.

Corollary 5.6 (Vandermonde Decomposition)**.** *Any positive semidefinite Toeplitz matrix* $P \in \mathbb{C}^{n\times n}$ *can be represented as follows*

$$P = VDV^*,$$

where

$$V = [a(f_1) \cdots a(f_r)] ,$$

$$D = \operatorname{diag}([d_1 \cdots d_r]) ,$$

d_k *are real positive numbers, and* $r = \operatorname{rank}(P)$.

Proof. Write P as $T_n(x)$ for some $x \in \mathbb{C}^n$. By assumption, $T_n(x) \succeq 0$ and $\operatorname{rank}(T_n(x)) = r$. Therefore by Caratheodory-Toeplitz theorem, x can be written as $\sum_{l=1}^{r} d_l a(f_l)$. Thus,

$$\begin{aligned} P = T_n(x) &= \sum_{l=1}^{r} d_l T_n(a(f_l)) \\ &= \sum_{l=1}^{r} d_l a(f_l) a(f_l)^* \\ &= VDV^*, \end{aligned}$$

where V and D are defined as in the statement of the theorem. □

5.2 SDP for Trigonometric Moments

5.2.1 Positive Trigonometric Moments

As a consequence of the Caratheodory's characterization of the moment cone, we can prove the characterization of the the convex hull of $\mathcal{A}_+$, as given by Theorem 5.1.

If $Tx \succeq 0$, then there exists an atomic decomposition $x = \sum_l c_l a(f_l)$ by Caratheodory-Toeplitz theorem since it is in the conical hull of the atomic set $\mathcal{A}_+$. Since we have an invariant

$x_0 = \sum_l c_l$ for any such decomposition, we have that

$$\|x\|_{\mathcal{A}_+} = \inf \left\{ \sum_l c_l \;\middle|\; x = \sum_l c_l a(f_l) \right\} = x_0.$$

On the other hand, if $Tx \not\succeq 0$, $x \notin \mathrm{cone}(\mathcal{A}_+)$ and thus $\|x\|_{\mathcal{A}_+} = +\infty$. In other words, whenever x is a valid (partial) sequence of trigonometric moments, the atomic norm is x_0. Otherwise, it is unbounded by definition.

5.2.2 General Trigonometric Moments

For the general trigonometric atoms, let us denote the atoms by two indices for frequency and phase,

$$a(f, \phi) = a(f) \exp(i2\pi\phi)$$

so that $\mathcal{A} = \{a(f, \phi) \mid f, \phi \in \mathbb{T}\}$.

Proof of Theorem 5.2. Denote the value of the right hand side of (5.5) by $\mathrm{SDP}(x)$. Suppose $x = \sum_k c_k a(f_k, \phi_k)$ with $c_k > 0$. Defining $u = \sum_k c_k a(f_k, 0)$ and $t = \sum_k c_k$, we note that

$$T(u) = \sum_k c_k a(f_k, 0) a(f_k, 0)^* = \sum_k c_k a(f_k, \phi_k) a(f_k, \phi_k)^*.$$

Therefore,

$$\begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} = \sum_k c_k \begin{bmatrix} a(f_k, \phi_k) \\ 1 \end{bmatrix} \begin{bmatrix} a(f_k, \phi_k) \\ 1 \end{bmatrix}^* \succeq 0 \tag{5.11}$$

Now, $\frac{1}{n} \mathrm{trace}(T(u)) = t = \sum_k c_k$ so that $SDP(x) \leq \sum_k c_k$. Since this holds for any decomposition of x, we conclude that $\|x\|_{\mathcal{A}} \geq \mathrm{SDP}(x)$.

Conversely, suppose for some u and x,

$$\begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \succeq 0\,. \tag{5.12}$$

In particular, $T(u) \succeq 0$. Form a Vandermonde decomposition

$$T(u) = VDV^*$$

as promised by Lemma 5.6. Since $VDV^* = \sum_k d_k a(f_k, 0)a(f_k, 0)^*$ and $\|a(f_k, 0)\|_2 = \sqrt{n}$, we have $\frac{1}{n}\operatorname{tr}(T(u)) = \operatorname{tr}(D)$.

Using this Vandermonde decomposition and the matrix inequality (5.12), it follows that x is in the range of V, and hence

$$x = \sum_k w_k a(f_k, 0) = Vw$$

for some complex coefficient vector $w = [\cdots, w_k, \cdots]^T$. Finally, by the Schur Complement Lemma, we have

$$VDV^* \succeq t^{-1} Vww^*V^*$$

Let q be any vector such that $V^*q = \operatorname{sign}(w)$. Such a vector exists because V is full rank. Then

$$\operatorname{tr}(D) = q^*VDV^*q \succeq t^{-1}q^*Vww^*V^*q = t^{-1}\left(\sum_k |w_k|\right)^2.$$

implying that $\operatorname{tr}(D)t \geq (\sum_k |w_k|)^2$. By the arithmetic geometric mean inequality,

$$\tfrac{1}{2n}\operatorname{tr}(T(u)) + \tfrac{1}{2}t = \tfrac{1}{2}\operatorname{tr}(D) + \tfrac{1}{2}t \geq \sqrt{\operatorname{tr}(D)t} \geq \sum_k |w_k| \geq \|x\|_{\mathcal{A}}$$

implying that $\mathrm{SDP}(x) \geq \|x\|_{\mathcal{A}}$ since the previous chain of inequalities hold for any choice of u, t that are feasible. □

5.3 SDP for Trigonometric Polynomials

The dual problem to AST involves trigonometric polynomials instead of moments.

5.3.1 Positive Trigonometric Polynomials

Definition 5.7. *A vector $q \in \mathbb{C}^n$ is a positive trigonometric polynomial if for every $f \in \mathbb{T}$,*

$$\langle q, a(f)\rangle = \mathfrak{Re}\left(\sum_{j=0}^{n-1} q_j \exp(i2\pi j f)\right) \geq 0.$$

Such polynomials have a simple characterization due to spectral factorization theorem.

Theorem 5.8. *$q \in \mathbb{C}^n$ is a positive polynomial if and only if there exists $Q \succeq 0$ such that $T^*Q = q$.*

Proof. Suppose there exists $Q \succeq 0$ such that $T^*Q = q$. Then, for any $f \in \mathbb{T}$,

$$\langle q, a(f)\rangle = \langle T^*Q, a(f)\rangle = \langle Q, Ta(f)\rangle.$$

Since $a(f)$ is a valid moment sequence $Ta(f) \succeq 0$. Thus, $\langle q, a(f)\rangle \geq 0$ since the inner product between two positive nonnegative matrices is nonnegative.

On the other hand, we can show that Q exists by using the spectral factorization theorem. Define $P(f) = \langle q, a(f)\rangle$. Since $P(f)$ is positive on the unit circle, there exists $\tilde{P}(f) = \sum_{j=0}^{n-1} \tilde{q}_j e^{i2\pi fj}$ such that $P(f) = \tilde{P}(f)\tilde{P}(f)^*$. Introduce the notation $\tilde{q} = \begin{pmatrix} \tilde{q}_0 & \dots & \tilde{q}_{n-1} \end{pmatrix}^T \in \mathbb{C}^n$ so that $\tilde{P}(f) = a(f)^*\tilde{q}$. Define $Q = \tilde{q}\tilde{q}^* \succeq 0$. Now, for any $f \in \mathbb{T}$,

$$\begin{aligned} P(f) &= \tilde{P}(f)\tilde{P}(f)^* \\ &= a(f)^*\tilde{q}\tilde{q} * a(f) \\ &= \mathrm{tr}(a(f)^*Qa(f)) \\ &= \mathrm{tr}(Qa(f)a(f)^*) \\ &= \langle Q, T\left(a(f)\right)\rangle \\ &= \langle T^*Q, a(f)\rangle. \end{aligned}$$

Comparing coefficients, we conclude that indeed $q = T^*Q$. □

5.3.2 General Trigonometric Polynomials

Recall from (3.6) that the dual atomic norm of a vector $v \in \mathbb{C}^n$ is the maximum absolute value of a complex trigonometric polynomial $V(f) = \sum_{l=0}^{n-1} v_l e^{-2\pi ilf}$. As a consequence, a constraint on the dual atomic norm is equivalent to a bound on the magnitude of $V(f)$:

$$\|v\|_{\mathcal{A}}^* \leq \tau \Leftrightarrow |V(f)|^2 \leq \tau^2, \forall f \in [0,1].$$

Equivalently, function $q(f) = \tau^2 - |V(f)|^2$ is thus a positive trigonometric polynomial. Using the characterization given by Theorem 5.8, one can derive the following result called the Bounded Real Lemma. (See for example, Theorem 4.24 in [49].)

Lemma 5.9 (Bounded Real Lemma)**.** *For any given causal trigonometric polynomial* $V(f) = \sum_{l=0}^{n-1} v_l e^{-2\pi i l f}$*,* $|V(f)| \leq \tau$ *if and only if there exists complex Hermitian matrix* Q *such that*

$$T^*(Q) = \tau^2 e_1 \;\; \textit{and} \;\; \begin{bmatrix} Q & v \\ v^* & 1 \end{bmatrix} \succeq 0.$$

Here, e_1 *is the first canonical basis vector with a one at the first component and zeros elsewhere and* v^* *denotes the Hermitian adjoint (conjugate transpose) of* v*.*

Proof. The polynomial $|V(f)|^2$ can be written as

$$\begin{aligned} |V(f)|^2 &= |\langle v, a(f)\rangle|^2 \\ &= a(f)^* v v^* a(f) \\ &= \operatorname{tr}\left(a(f)^* v v^* a(f)\right) \\ &= \langle v v^*, T a(f)\rangle \\ &= \langle T^*(v v^*), a(f)\rangle. \end{aligned}$$

Thus, $|q(f)|^2 = \tau^2 - |V(f)|^2 = \langle \tau^2 e_1 - T^*(vv^*), a(f)\rangle$. Now,Theorem 5.8 promises the existence of $P \succeq 0$ such that $\tau^2 e_1 - T^*(vv^*) = T^* P$. Equivalently, there exists $Q \succeq vv^*$ (equal to $P + vv^*$) such that $\tau^2 e_1 = T^* Q$. Using Schur complements, we can rewrite $Q \succeq vv^*$ as $\begin{bmatrix} Q & v \\ v^* & 1 \end{bmatrix} \succeq 0$, and this completes the proof. □

5.3.3 Deriving the primal characterization from dual

The dual characterization in the previous section via the theory of positive polynomials provides an alternative route to rederive the SDP characterization of the atomic norm derived

in section 5.2.2.

Using Lemma 5.9, we rewrite the atomic norm $\|x\|_{\mathcal{A}} = \sup_{\|v\|_{\mathcal{A}}^* \leq 1} \langle x, v \rangle$ as the following semidefinite program:

$$\begin{array}{ll} \text{maximize}_{v,\, Q} & \langle x, v \rangle \\ \text{subject to} & \begin{bmatrix} Q & v \\ v^* & 1 \end{bmatrix} \succeq 0, \qquad T^*(Q) = e_1. \end{array} \tag{5.13}$$

The dual problem of (5.13) (after a trivial rescaling) is then equal to the atomic norm of x:

$$\begin{array}{lll} \|x\|_{\mathcal{A}} = & \min_{t,u} & \frac{1}{2}(t + u_1) \\ & \text{subject to} & \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \succeq 0. \end{array}$$

Thus, we have another proof of the characterization in Theorem 5.2.

5.4 AST using Alternating Direction Method of Multipliers

Using the characterization of the atomic norm given by Lemma 5.9, we can see that the atomic denoising problem (2.2) for the set of trigonometric atoms is equivalent to

$$\begin{array}{ll} \text{minimize}_{t,u,x} & \frac{1}{2}\|x - y\|_2^2 + \frac{\tau}{2}(t + u_1) \\ \text{subject to} & \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \succeq 0. \end{array} \tag{5.14}$$

This semidefinite program (5.14) can be solved by off-the-shelf solvers such as SeDuMi [93] and SDPT3 [102]. However, these solvers tend to be slow for large problems. So, we will look at a reasonably efficient algorithm based upon the Alternating Direction Method of Multipliers.

A thorough survey of the ADMM algorithm is given in [13]. I only present the details essential to the implementation of atomic norm soft thresholding. To put the problem in an appropriate form for ADMM, rewrite (5.14) as

$$\begin{aligned} &\text{minimize}_{t,u,x,Z} && \tfrac{1}{2}\|x-y\|_2^2 + \tfrac{\tau}{2}(t+u_1) \\ &\text{subject to} && Z = \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \\ & && Z \succeq 0. \end{aligned}$$

and dualize the equality constraint via an Augmented Lagrangian:

$$\mathcal{L}_\rho(t,u,x,Z,\Lambda) = \frac{1}{2}\|x-y\|_2^2 + \frac{\tau}{2}(t+u_1) + \left\langle \Lambda, Z - \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \right\rangle + \frac{\rho}{2}\left\| Z - \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \right\|_F^2$$

ADMM then consists of the update steps:

$$\begin{aligned} (t^{l+1}, u^{l+1}, x^{l+1}) &\leftarrow \arg\min_{t,u,x} \mathcal{L}_\rho(t,u,x,Z^l,\Lambda^l) \\ Z^{l+1} &\leftarrow \arg\min_{Z\succeq 0} \mathcal{L}_\rho(t^{l+1},u^{l+1},x^{l+1},Z,\Lambda^l) \\ \Lambda^{l+1} &\leftarrow \Lambda^l + \rho\left(Z^{l+1} - \begin{bmatrix} T(u^{l+1}) & x^{l+1} \\ x^{l+1^*} & t^{l+1} \end{bmatrix}\right). \end{aligned}$$

The updates with respect to t, x, and u can be computed in closed form:

$$
\begin{aligned}
t^{l+1} &= Z^l_{n+1,n+1} + \left(\Lambda^l_{n+1,n+1} - \frac{\tau}{2}\right)/\rho \\
x^{l+1} &= \frac{1}{2\rho+1}(y + 2\rho z^l_1 + 2\lambda^l_1) \\
u^{l+1} &= W\left(T^*(Z^l_0 + \Lambda^l_0/\rho) - \frac{\tau}{2\rho}e_1\right)
\end{aligned}
$$

Here W is the diagonal matrix with entries

$$
W_{ii} = \begin{cases} \frac{1}{n} & i = 1 \\ \frac{1}{2(n-i+1)} & i > 1 \end{cases}
$$

and we introduced the partitions:

$$
Z^l = \begin{bmatrix} Z^l_0 & z^l_1 \\ z^{l\,*}_1 & Z^l_{n+1,n+1} \end{bmatrix} \quad \text{and} \quad \Lambda^l = \begin{bmatrix} \Lambda^l_0 & \lambda^l_1 \\ \lambda^{l\,*}_1 & \Lambda^l_{n+1,n+1} \end{bmatrix}.
$$

The Z update is simply the projection onto the positive definite cone

$$
Z^{l+1} := \arg\min_{Z \succeq 0} \left\| Z - \begin{bmatrix} T(u^{l+1}) & x^{l+1} \\ x^{l+1\,*} & t^{l+1} \end{bmatrix} + \Lambda^l/\rho \right\|^2_F. \tag{5.15}
$$

Projecting a matrix Q onto the positive definite cone is accomplished by forming an eigenvalue decomposition of Q and setting all negative eigenvalues to zero.

To summarize, the update for (t, u, x) requires averaging the diagonals of a matrix (which is equivalent to projecting a matrix onto the space of Toeplitz matrices), and hence operations that are $O(n)$. The update for Z requires projecting onto the positive definite cone and

requires $O(n^3)$ operations. The update for Λ is simply addition of symmetric matrices.

Note that the dual solution $\hat{z}$ can be obtained as $\hat{z} = y - \hat{x}$ from the primal solution $\hat{x}$ obtained from ADMM by using Lemma 2.7.

5.5 Discretization

In general the atomic sets may not have a computable characterization. Even if there is an SDP characterization and an efficient implementation using techniques like ADMM, when the number of samples is larger than a few hundred, the running time of the ADMM method is dominated by the cost of computing eigenvalues and is usually expensive. This is not avoidable as there is no cheap way to find projections on the positive definite cone. For very large problems, we will see that we can use discretization and hence Lasso as an alternative to the semidefinite program (5.14).

5.5.1 Discretized Atomic Soft Thresholding

Suppose we solve AST (2.2) on a different set $\widetilde{\mathcal{A}}$ (say, an ϵ-net of $\mathcal{A}$) instead of $\mathcal{A}$. If for some $M > 0$,

$$M^{-1}\|x\|_{\widetilde{\mathcal{A}}} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\widetilde{\mathcal{A}}}$$

holds for every x, then Theorem 2.1 still applies with a constant factor M. Our justification for using the finite dimensional Lasso as an alternative to the general infinite atomic norm soft thresholding problem relies on approximation guarantees for epsilon-nets of the atomic sets. To be precise, in the next section we see that M approaches unity as $\epsilon \to 0$. Thus, the solution to the discretized atomic soft thresholding (DAST) problem approaches the AST solution.

The following proposition demonstrates that the universal guarantee in Theorem 2.1 continues to hold with only a penalty of a small multiplicative constant when DAST is used in

place of AST.

Proposition 5.10. *Suppose*

$$\|z\|_{\widetilde{\mathcal{A}}}^* \leq \|z\|_{\mathcal{A}}^* \leq M\|z\|_{\widetilde{\mathcal{A}}}^* \textit{ for every } z, \tag{5.16}$$

or equivalently

$$M^{-1}\|x\|_{\widetilde{\mathcal{A}}} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\widetilde{\mathcal{A}}} \textit{ for every } x, \tag{5.17}$$

then under the same conditions as in Theorem 2.1,

$$\frac{1}{n}\,\mathbb{E}\|\tilde{x} - x^\star\|_2^2 \leq \frac{M\tau}{n}\|x^\star\|_{\mathcal{A}}$$

where $\tilde{x}$ is the optimal solution for (2.2) *with $\mathcal{A}$ replaced by $\widetilde{\mathcal{A}}$.*

Proof. By assumption, $\mathbb{E}\left(\|w\|_{\mathcal{A}}^*\right) \leq \tau$. Now, (5.16) implies $\mathbb{E}\left(\|w\|_{\widetilde{\mathcal{A}}}^*\right) \leq \tau$. Applying Theorem 2.1, and using (5.17), we get

$$\frac{1}{n}\,\mathbb{E}\|\tilde{x} - x^\star\|_2^2 \leq \frac{\tau}{n}\|x^\star\|_{\widetilde{\mathcal{A}}} \leq \frac{M\tau}{n}\|x^\star\|_{\mathcal{A}}. \qquad \square$$

5.5.2 Approximated Atomic Norms

In this section, we will see that atomic norms can be approximated by choosing a large ϵ-net of atoms instead of the original set of atoms. The proof of approximation and the characterization of the convergence depends upon using a Euclidean ϵ cover of the atomic set $\mathcal{A}$, and the equivalence between Euclidean and atomic norms in finite dimensions.

If $\mathcal{A} \in F^n$, there exists C, possibly depending on n such that $\|x\|_{\mathcal{A}} \leq C\|x\|_2$ for any choice of atomic set. The minimum volume ellipsoid that circumscribes $\mathrm{conv}(\mathcal{A})$ is called the

Löwner-John ellipsoid after Charles Löwner who discovered the uniqueness of the minimum volume circumscribing ellipsoid for any convex sets and Fritz John who proved the existence. This is often a scaled version of the Euclidean ball when the atomic set is symmetric and it provides an optimal choice of C.

Using the characterization of John, we have the following condition for Euclidean ball to be the minimum volume circumscribing ellipsoid of $\mathrm{conv}(\mathcal{A})$.

Theorem 5.11. *If there exists positive numbers $\lambda_1, \ldots, \lambda_m > 0$ and atoms $a_1, \ldots, a_m$ for $m \geq n$ such that*

$$\sum_{i=1}^{m} \lambda_i a_i = 0 \text{ and } I_n = \sum_{i=1}^{m} \lambda_i (a_i a_i^*),$$

the Euclidean ball $B = \{x \mid \|x\|_2 \leq 1\}$ is the unique minimum volume ellipsoid that circumscribes $\mathrm{conv}(\mathcal{A})$. If in addition, $\mathcal{A}$ is centrosymmetric, for any x,

$$\|x\|_2 \leq \|x\|_{\mathcal{A}} \leq \sqrt{n}\|x\|_2$$

The conditions in the previous theorem hold for a number of interesting atomic set including Fourier measurements discussed in Chapter 3, and one may argue that the minimum volume ellipsoid is often a Euclidean sphere. When the condition in Theorem 5.11 is satisfied, defining $\mathcal{A}_\epsilon$ as an ϵ-net of $\mathcal{A}$, we can write

$$\begin{aligned}
\|z\|_{\mathcal{A}}^* &= \sup_{x \in \mathcal{A}} \langle z, x \rangle \\
&\leq \sup_{\hat{x} \in \mathcal{A}_\epsilon} \langle z, \hat{x} \rangle + \inf_{\hat{x} \in \mathcal{A}_\epsilon} \|z\|_{\mathcal{A}}^* \|x - \hat{x}\|_{\mathcal{A}} \\
&\leq \sup_{\hat{x} \in \mathcal{A}_\epsilon} \langle z, \hat{x} \rangle + \|z\|_{\mathcal{A}}^* \sup_{\|w\|_2 \leq \epsilon} w_{\mathcal{A}} \\
&\leq \|z\|_{\mathcal{A}_\epsilon}^* + \sqrt{n}\epsilon \|z\|_{\mathcal{A}}^*,
\end{aligned}$$

whence we get $\|z\|_{\mathcal{A}}^* \leq (1-\sqrt{n}\epsilon)^{-1}\|z\|_{\mathcal{A}_\epsilon}^*$. Also, since $\mathcal{A}_\epsilon \subset \mathcal{A}$, by definition $\|z\|_{\mathcal{A}_\epsilon}^* \leq \|z\|_{\mathcal{A}}^*$. Combining these two, we have the following approximation

$$\|z\|_{\mathcal{A}_\epsilon}^* \leq \|z\|_{\mathcal{A}}^* \leq (1-\sqrt{n}\epsilon)^{-1}\|z\|_{\mathcal{A}_\epsilon}^*. \tag{5.18}$$

We will need the following lemma to restate the ϵ approximation in terms of the atomic norm, instead of dual.

Lemma 5.12. *$\|z\|_{\mathcal{A}}^* \leq M\|z\|_{\widetilde{\mathcal{A}}}^*$ for every z iff $M^{-1}\|x\|_{\widetilde{\mathcal{A}}} \leq \|x\|_{\mathcal{A}}$ for every z.*

Proof. Let us show the forward implication – the converse will follow since the dual of the dual norm is again the primal norm. By (2.5), for any x, there exists a z with $\|z\|_{\widetilde{\mathcal{A}}}^* \leq 1$ and $\langle x, z\rangle = \|x\|_{\widetilde{\mathcal{A}}}$. So,

$$\begin{aligned} M^{-1}\|x\|_{\widetilde{\mathcal{A}}} &= M^{-1}\langle x, z\rangle \\ &\leq M^{-1}\|z\|_{\mathcal{A}}^*\|x\|_{\mathcal{A}} && \text{by (2.5)} \\ &\leq \|x\|_{\mathcal{A}} && \text{by assumption.} \end{aligned}$$

□

Now, we can write the approximation for atomic norm. For any x,

$$(1-\sqrt{n}\epsilon)\|x\|_{\mathcal{A}_\epsilon} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}_\epsilon}. \tag{5.19}$$

While this method is generic, simpler and sometimes stronger arguments may be possible for specific atomic sets. In the succeeding sections, we use techniques similar to what we just outlined and characterize the atomic norm approximation for the atomic sets for line spectral estimation and system identification.

5.5.3 DAST for Line Spectral Signals

To proceed, pick a uniform grid of N frequencies and form

$$\mathcal{A}_N = \left\{ a_{m/N,\phi} \;\middle|\; 0 \leq m \leq N-1 \right\} \subset \mathcal{A}$$

and solve (2.2) on this grid. i.e., we solve the problem

$$\text{minimize } \frac{1}{2}\|x-y\|_2^2 + \tau \|x\|_{\mathcal{A}_N}. \tag{5.20}$$

To see why this is to our advantage, define Φ be the $n \times N$ Fourier matrix with mth column $a_{m/N,0}$. Then for any $x \in \mathbb{C}^n$ we have $\|x\|_{\mathcal{A}_N} = \min\{\|c\|_1 : x = \Phi c\}$. So, we solve

$$\text{minimize } \frac{1}{2}\|\Phi c - y\|_2^2 + \tau \|c\|_1. \tag{5.21}$$

for the optimal point $\hat{c}$ and set $\hat{x}_N = \Phi \hat{c}$ or the first n terms of the N term discrete Fourier transform (DFT) of $\hat{c}$. Furthermore, $\Phi^* z$ is simply the N term inverse DFT of $z \in \mathbb{C}^n$. This observation coupled with Fast Fourier Transform (FFT) algorithm for efficiently computing DFTs gives a fast method to solve (5.20), using standard compressed sensing software for $\ell_2 - \ell_1$ minimization, for example, SparSA [106].

Because of the relatively simple structure of the atomic set, the optimal solution $\hat{x}$ for (5.20) can be made arbitrarily close to (5.14) by picking N a constant factor larger than n. In fact, the following section furnishes the proof that the atomic norms induced by $\mathcal{A}$ and a discretized version $\mathcal{A}_N$ are equivalent.

Due to the efficiency of the FFT, the discretized approach has a much lower algorithmic complexity than either Cadzow's alternating projections method or the ADMM method de-

scribed in section 5.4, which each require computing an eigenvalue decomposition at each iteration. Indeed, fast solvers for (5.21) converge to an ϵ optimal solution in no more than $1/\sqrt{\epsilon}$ iterations. Each iteration requires a multiplication by Φ and a simple "shrinkage" step. Multiplication by Φ or Φ^* requires $O(N \log N)$ time and the shrinkage operation can be performed in time $O(N)$.

This fast form of basis pursuit has been proposed by several authors. However, analyzing this method with tools from compressed sensing has proven daunting because the matrix Φ is nowhere near a restricted isometry. Indeed, as N tends to infinity, the columns become more and more coherent. However, common sense says that a larger grid should give better performance, for both denoising and frequency localization! Indeed, by appealing to the atomic norm framework, makes it possible to show exactly this point: the larger one makes N, the closer one approximates the desired atomic norm soft thresholding problem. Moreover, we do not have to choose N to be too large in order to achieve nearly the same performance as the AST.

Approximation of the Dual Atomic Norm

Note that the dual atomic norm of w is given by

$$\|w\|_{\mathcal{A}}^* = \sqrt{n} \sup_{f \in [0,1]} \left| W_n(e^{i2\pi f}) \right|. \tag{5.22}$$

i.e., the maximum modulus of the polynomial W_n defined by

$$W_n(e^{i2\pi f}) = \frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} w_m e^{-i2\pi m f}. \tag{5.23}$$

Treating W_n as a function of f, with a slight abuse of notation, define

$$\|W_n\|_\infty := \sup_{f\in[0,1]} \left|W_n(e^{i2\pi f})\right|.$$

We will see that we can approximate the maximum modulus by evaluating W_n in a uniform grid of N points on the unit circle. To show that as N becomes large, the approximation is close to the true value, let us bound the derivative of W_n using Bernstein's inequality for polynomials.

Theorem 5.13 (Bernstein, See, for example [88])**.** *Let p_n be any polynomial of degree n with complex coefficients. Then,*

$$\sup_{|z|\le 1} |p'(z)| \le n \sup_{|z|\le 1} |p(z)|.$$

Note that for any $f_1, f_2 \in [0,1]$, we have

$$\begin{aligned}
\left|W_n(e^{i2\pi f_1})\right| - \left|W_n(e^{i2\pi f_2})\right| &\le \left|e^{i2\pi f_1} - e^{i2\pi f_2}\right| \|W_n'\|_\infty \\
&= 2|\sin(2\pi(f_1-f_2))|\|W_n'\|_\infty \\
&\le 4\pi(f_1-f_2)\|W_n'\|_\infty \\
&\le 4\pi n(f_1-f_2)\|W_n\|_\infty,
\end{aligned}$$

where the last inequality follows by Bernstein's theorem. Letting s take any of the N values $0, \frac{1}{N} \ldots, \frac{N-1}{N}$, we see that,

$$\|W_n\|_\infty \le \max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right| + \frac{2\pi n}{N}\|W_n\|_\infty.$$

Since the maximum on the grid is a lower bound for maximum modulus of W_n, we have

$$\max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right| \leq \|W_n\|_\infty \tag{5.24}$$

$$\leq \left(1-\frac{2\pi n}{N}\right)^{-1} \max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right|$$

$$\leq \left(1+\frac{4\pi n}{N}\right) \max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right|. \tag{5.25}$$

Thus, for every w,

$$\|w\|^*_{\mathcal{A}_N} \leq \|w\|^*_{\mathcal{A}} \leq \left(1-\frac{2\pi n}{N}\right)^{-1} \|w\|^*_{\mathcal{A}_N} \tag{5.26}$$

or equivalently, for every x,

$$\left(1-\frac{2\pi n}{N}\right) \|x\|_{\mathcal{A}_N} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}_N} \tag{5.27}$$

$$\left(1-\frac{2\pi n}{N}\right) \|x\|_{\mathcal{A}_N} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}_N}, \forall x \in \mathbb{C}^n \tag{5.28}$$

Using Proposition 5.10 and (3.14), we conclude

$$\frac{1}{n}\,\mathbb{E}\|\hat{x}_N - x^\star\|_2^2 \leq \frac{\sigma\left(\frac{\log(n)+1}{\log(n)}\right)\|x^\star\|_{\mathcal{A}}\sqrt{n\log(n)+n\log(4\pi\log(n))}}{n\left(1-\frac{2\pi n}{N}\right)} = O\left(\sigma\sqrt{\frac{\log(n)}{n}}\frac{\|x^\star\|_{\mathcal{A}}}{\left(1-\frac{2\pi n}{N}\right)}\right)$$

5.5.4 DAST for System Identification

The following proposition asserts that we can approximate this finite dimensional atomic norm defined the atomic set of measurements of single pole systems via a sufficiently fine discretization of the unit disk.

Proposition 5.14. *Let $\mathbb{D}_\rho^{(\epsilon)}$ be a finite subset of the unit disc such that for any $w \in \mathbb{D}_\rho$ there exists a $v \in \mathbb{D}_\rho^{(\epsilon)}$ satisfying $|w - v| \leq \epsilon$. Define*

$$\|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)} = \inf \left\{ \sum_{w \in \mathbb{D}_\rho^{(\epsilon)}} |c_w| \ : \ x_i = \sum_{w \in \mathbb{D}_\rho^{(\epsilon)}} c_w \mathcal{L}_i \left(\frac{1 - |w|^2}{z - w} \right) \right\} .$$

Then there exists a constant $C_\epsilon \in [0, 1]$ such that

$$C_\epsilon \|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)} \leq \|x\|_{\mathcal{L}(\mathcal{A})} \leq \|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)} .$$

The set $\mathbb{D}_\rho^{(\epsilon)}$ is called an ϵ-net for the set $\mathbb{D}_\rho$. We will see that when $\mathcal{L}_k(G) = G(e^{i\theta_k})$, C_ϵ is at least $(1 - \frac{16\rho\epsilon}{\pi(1-\rho)})$. Other measurement ensembles can be treated similarly.

When we replace $\|x\|_{\mathcal{L}(\mathcal{A})}$ with its discretized counterpart $\|x\|_{\mathcal{L}(\mathcal{A})}$ in (4.8),

$$\underset{x}{\text{minimize}} \ \tfrac{1}{2}\|x - y\|_2^2 + \mu \|x\|_{\mathcal{L}(\mathcal{A}_\epsilon)}$$

is equivalent to

$$\underset{c}{\text{minimize}} \ \tfrac{1}{2}\|Mc - y\|_2^2 + \mu \sum_{w \in \mathbb{D}_\rho^{(\epsilon)}} |c_w| \tag{5.29}$$

where

$$M_{ij} = \mathcal{L}_i \left(\tfrac{1 - |w_j|^2}{z - w_j} \right)$$

and j indexes the set $\mathbb{D}_\rho^{(\epsilon)}$. That is M is an $n \times |\mathbb{D}_\rho^{(\epsilon)}|$ matrix. Problem (5.29) is a weighted ℓ_1 regularization problem with real or complex data depending on specific problem.

This DAST problem can be solved very efficiently with a variety of off-the-shelf tools including SPARSA [106], FPC [58] or even more general purpose packages such as YALMIP [71] or CVX [55]. DAST yields an approximate solution to problem (4.6), and, as we will see, yields

a statistically consistent estimate provided the parameter ϵ is adjusted to meet the desired numerical accuracy.

As in section 5.5.2, the proof of proposition 5.14 depends upon connecting the Euclidean distance between the atoms indexed by the discretization of the unit disc to the distances induced by the atomic norm. We already know from theorem 4.1 in Chapter 4 that the Hankel nuclear norm and the atomic norm are equivalent. So, we will need the following lemma that connects the Hankel nuclear norm of the atomic functions $\{\varphi_a\}$ to the distances in the complex planes for poles $a \in \mathbb{C}$.

Lemma 5.15. *For any $a, b \in \mathbb{D}_\rho$,*

$$\|\Gamma_{\varphi_a} - \Gamma_{\varphi_b}\|_1 \leq \frac{2\rho}{1-\rho}|a-b|\,. \tag{5.30}$$

Proof. The Hankel operator for $\varphi_a(z)$ is given by the semi-infinite, rank one matrix

$$(1-|a|^2)\begin{bmatrix} 1 & a & a^2 & a^3 & \cdots \\ a & a^2 & a^3 & a^4 & \cdots \\ a^2 & a^3 & a^4 & a^5 & \cdots \\ \vdots & \ddots & & & \end{bmatrix} = (1-|a|^2)\begin{bmatrix} 1 \\ a \\ a^2 \\ a^3 \\ \vdots \end{bmatrix}\begin{bmatrix} 1 \\ a \\ a^2 \\ a^3 \\ \vdots \end{bmatrix}^T .$$

Let $\zeta_a = \sqrt{1-|a|^2}\begin{bmatrix} 1 & a & a^2 & a^3 & \cdots \end{bmatrix}^T$. Note that $\zeta_a \in \ell_2$ with norm equal to 1. Also note that we have

$$\langle \zeta_a, \zeta_b \rangle = \frac{\sqrt{1-|a|^2}\sqrt{1-|b|^2}}{1-\bar{a}b}\,. \tag{5.31}$$

Then we have

$$\begin{aligned}
\|\Gamma_{\varphi_a} - \Gamma_{\varphi_b}\|_1 &= \|\zeta_a\zeta_a^T - \zeta_b\zeta_b^T\|_1 \\
&= \|\zeta_a(\zeta_a - \zeta_b)^T + (\zeta_a - \zeta_b)\zeta_b^T\|_1 && \text{(5.32a)} \\
&\leq \|\zeta_a(\zeta_a - \zeta_b)^T\|_1 + \|(\zeta_a - \zeta_b)\zeta_b^T\|_1 && \text{(5.32b)} \\
&= 2\|\zeta_a - \zeta_b\|_{\ell_2} && \text{(5.32c)} \\
&= 2\sqrt{2}\sqrt{1 - \mathfrak{Re}\frac{\sqrt{1-|a|^2}\sqrt{1-|b|^2}}{1-\bar{a}b}} && \text{(5.32d)} \\
&\leq \frac{2\rho}{1-\rho}|a-b|
\end{aligned}$$

Here, (5.32b) is the triangle inequality. (5.32c) follows because the nuclear norm of a rank one operator is equal to the product of the ℓ_2 norm of the factors. (5.32d) follows from (5.31). The final inequality follows from analyzing the taylor series of the preceding expression. □

Now, we are ready to prove our main proposition.

Proof of Proposition 5.14

First note that for any atomic sets $\mathcal{A} \subset \mathcal{A}'$, $\|x\|_{\mathcal{A}'} \leq \|x\|_{\mathcal{A}}$. The harder part of this proposition is the lower bound. To proceed, let us use the dual norm.

Let $\mathbb{D}_\rho^{(\tau)}$ be a subset of $\mathbb{D}_\rho$ such that for every $a \in \mathbb{D}_\rho$, there exists an $\hat{a} \in \mathbb{D}_\rho^{(\tau)}$ satisfying $|a - \hat{a}| \leq \tau$. For each $a \in \mathbb{D}_\rho$, denote $\hat{a}$ as the closest point in $\mathbb{D}_\rho^{(\tau)}$ to a.

Now observe that

$$\|\mathcal{L}(\varphi_{\hat{a}} - \varphi_a)\|_{\mathcal{L}(\mathcal{A})} \leq \|\varphi_{\hat{a}} - \varphi_a\|_{\mathcal{A}} \leq \frac{8}{\pi}\|\Gamma_{\varphi_{\hat{a}}} - \Gamma_{\varphi_a}\|_1 \leq \frac{16\rho\tau}{\pi(1-\rho)}\,.$$

Here, the first inequality follows from the reasoning in Section 4.2.3. The second inequality is Theorem 4.1, and the final inequality is by (5.30).

We can then compute

$$\begin{aligned}
\|z\|^*_{\mathcal{L}(\mathcal{A})} &= \sup_{a\in\mathbb{D}_\rho} \langle \mathcal{L}(\varphi_a), z\rangle \\
&= \sup_{a\in\mathbb{D}_\rho} \langle \mathcal{L}(\varphi_{\hat{a}}), z\rangle + \langle \mathcal{L}(\varphi_a - \varphi_{\hat{a}}), z\rangle \\
&\leq \sup_{a\in\mathbb{D}_\rho^{(\tau)}} \langle \mathcal{L}(\varphi_{\hat{a}}), z\rangle + \sup_{a\in\mathbb{D}_\rho} \langle \mathcal{L}(\varphi_a - \varphi_{\hat{a}}), z\rangle \\
&= \|z\|^*_{\mathcal{L}(\mathcal{A}_\tau)} + \sup_{a\in\mathbb{D}_\rho} \langle \mathcal{L}(\varphi_a - \varphi_{\hat{a}}), z\rangle \\
&\leq \|z\|^*_{\mathcal{L}(\mathcal{A}_\tau)} + \sup_{a\in\mathbb{D}_\rho} \|\mathcal{L}(\varphi_{\hat{a}} - \varphi_a)\|_{\mathcal{L}(\mathcal{A})} \|z\|^*_{\mathcal{L}(\mathcal{A})} \\
&\leq \|z\|^*_{\mathcal{L}(\mathcal{A}_\tau)} + \frac{16\rho\tau}{\pi(1-\rho)} \|z\|^*_{\mathcal{L}(\mathcal{A})} \,.
\end{aligned}$$

Rearranging both sides of this inequality gives

$$\|z\|^*_{\mathcal{L}(\mathcal{A})} \leq C_\tau^{-1} \|z\|^*_{\mathcal{L}(\mathcal{A}_\tau)}$$

with $C_\tau = 1 - \frac{16\rho\tau}{\pi(1-\rho)}$, completing the proof.

5.6 Experiments for Line Spectral Estimation

I compared the performance of AST, the discretized Lasso approximation, the Matrix Pencil, MUSIC and Cadzow's method, both in terms of the mean squared estimation error as in Theorem 3.2 and frequency localization. For my experiments, I generated k normalized frequencies $f_1^\star, \ldots, f_k^\star$ uniformly randomly chosen from $[0, 1]$ such that every pair of frequencies

are separated by at least $1/2n$. The signal $x^\star \in \mathbb{C}^n$ is generated according to (3.1) with k random amplitudes independently chosen from $\chi^2(1)$ distribution (squared Gaussian). All of our sinusoids were then assigned a random phase (equivalent to multiplying $c_k^\star$ by a random unit norm complex number). Then, the observation y is produced by adding complex white gaussian noise w such that the input signal to noise ratio (SNR) is $-10, -5, 0, 5, 10, 15$ or 20 dB. We compare the average MSE of the various algorithms in 20 trials for various values of number of observations $(n = 64, 128, 256)$, and number of frequencies $(k = n/4, n/8, n/16)$.

AST needs an estimate of the noise variance σ^2 to pick the regularization parameter according to (3.14). In many situations, this variance is not known to us *a priori*. However, we can construct a reasonable estimate for σ when the phases are uniformly random. It is known that the autocorrelation matrix of a line spectral signal (see, for example Chapter 4 in [92]) can be written as a sum of a low rank matrix and $\sigma^2 I$ if we assume that the phases are uniformly random. Since the empirical autocorrelation matrix concentrates around the true expectation, we can estimate the noise variance by averaging a few smallest eigenvalues of the empirical autocorrelation matrix. For the following experiments, I formed the empirical autocorrelation matrix using the MATLAB routine `corrmtx` using a prediction order $m = n/3$ and averaging the lower 25% of the eigenvalues, and used this estimate in equation (3.14) to determine the regularization parameter for both the AST and Lasso experiments.

First, I implemented AST using the ADMM method and used the stopping criteria described in [13] and set $\rho = 2$ for all experiments, and used the dual solution $\hat{z}$ to determine the support of the optimal solution $\hat{x}$ using the procedure described in Section 3.3. Once the frequencies $\hat{f}_l$ are extracted, I ran the least squares problem $\text{minimize}_\alpha \|U\alpha - y\|^2$ where $U_{jl} = \exp(i2\pi j \hat{f}_l)$ to obtain a *debiased* solution. After computing the optimal solution α_{opt}, one can compute the prediction $\hat{x} = U\alpha_{\text{opt}}$.

After implementing Lasso, I obtained an estimate $\hat{x}$ of $x^\star$ from y by solving the optimization

problem (5.20) with debiasing. I used the algorithm described in Section 5.5.3 with grid of $N = 2^m$ points where $m = 10, 11, 12, 13, 14$ and 15. Because of the basis mismatch effect, the optimal c_{opt} has significantly more non-zero components than the true number of frequencies. However, we can observe that the frequencies corresponding to the non-zero components of c_{opt} cluster around the true ones. We therefore extract one frequency from each cluster of non-zero values by identifying the grid point with the maximum absolute c_{opt} value and zero everything else in that cluster. I then ran a debiasing step which solves the least squares problem minimize$_\beta \|\Phi_S \beta - y\|^2$ where Φ_S is the submatrix of Φ whose columns correspond to frequencies identified from c_{opt}. This gave the estimate $\hat{x} = \Phi_S \beta_{\text{opt}}$. I used the freely downloadable implementation of SpaRSA to solve the Lasso problem. A stopping parameter of 10^{-4}, but otherwise used the default parameters.

I implemented Cadzow's method as described by the pseudocode in [8], the Matrix Pencil as described in [60] and MUSIC [89] using the MATLAB routine `rootmusic`. All these algorithms need an estimate of the number of sinusoids. Rather than implementing a heuristic to estimate k, *I fed the true k to our solvers*. This provides a huge advantage to these algorithms. Neither AST or the Lasso based algorithm are provided the true value of k, and the noise variance σ^2 required in the regularization parameter is estimated from y.

Let $\{\hat{c}_l\}$ and $\{\hat{f}_l\}$ denote the amplitudes and frequencies estimated by any of the algorithms - AST, MUSIC or Cadzow. I used the following error metrics to characterize the frequency localization of various algorithms:

(i) Sum of the absolute value of amplitudes in the far region F, $m_1 = \sum_{l:\hat{f}_l \in F} |\hat{c}_l|$

(ii) The weighted frequency localization error, $m_2 = \sum_{l:\hat{f}_l \in N_j} |\hat{c}_l| \{\min_{f_j \in T} d(f_j, \hat{f}_l)\}^2$

(iii) Error in approximation of amplitudes in the near region, $m_3 = \left| c_j - \sum_{l:\hat{f}_l \in N_j} \hat{c}_l \right|$

These are precisely the quantities that we prove tend to zero in Theorem 3.3.

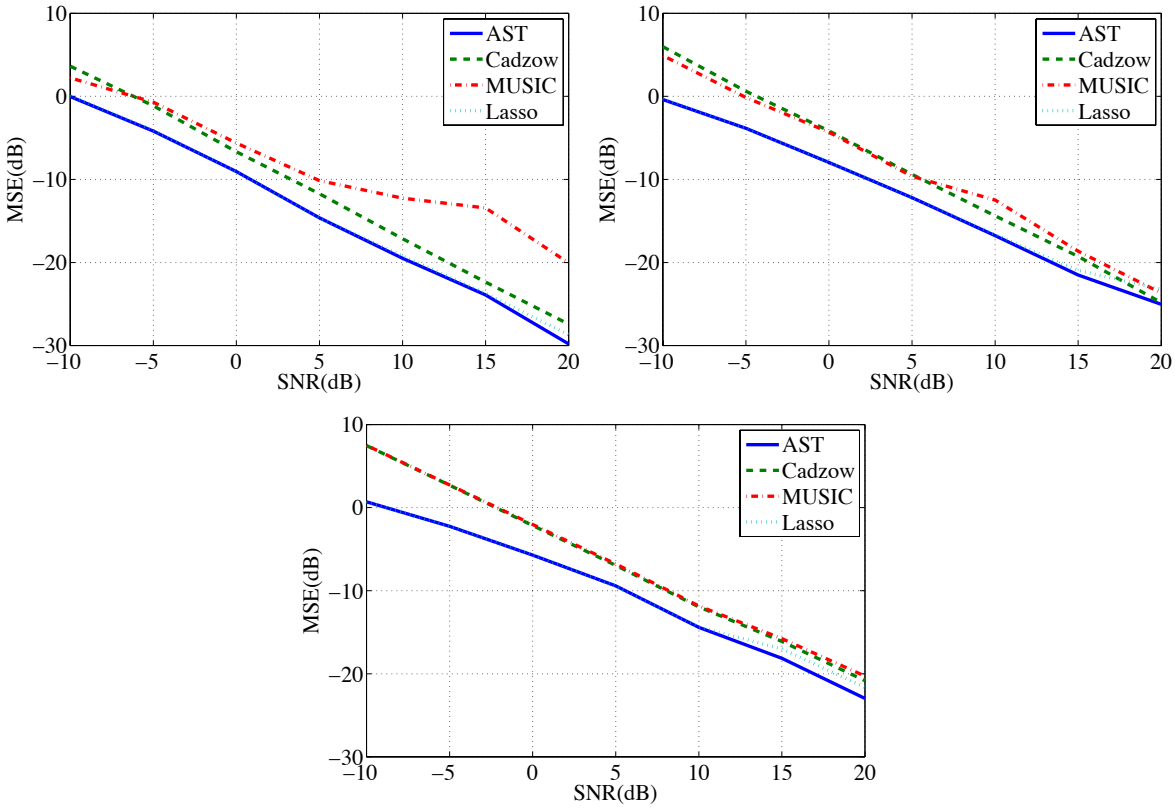
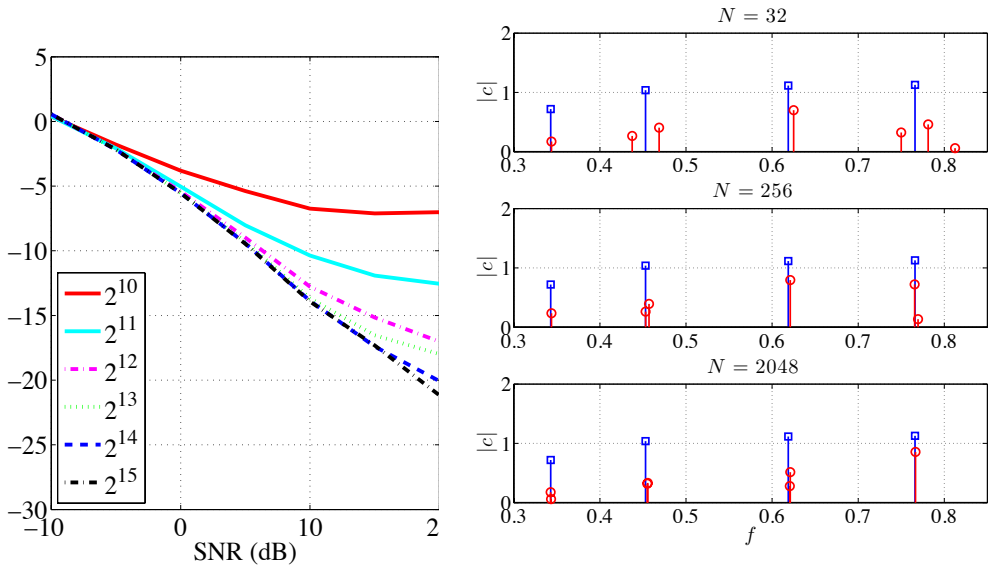


Figure 5.1: **MSE vs SNR plots:** This graph compares MSE vs SNR for a subset of experiments with $n = 128$ samples. From top left, clockwise, the plots are for combinations of 8, 16, and 32 sinusoids with amplitudes and frequencies sampled at random.

Figure 5.1 shows MSE vs SNR plots for a subset of experiments when $n = 128$ time samples are taken to take a closer look at the differences. It can be seen from these plots that the performance difference between classical algorithms such as MUSIC and Cadzow with respect to the convex optimization based AST and Lasso is most pronounced at lower sparsity levels. When the noise dominates the signal (SNR ≤ 0 dB), all the algorithms are comparable. However, AST and Lasso outperform the other algorithms in almost every regime.



(a) (b)

Figure 5.2: (a) Plot of MSE vs SNR for Lasso at different grid sizes for a subset of experiments with $n = 128, k = 16$
(b) Lasso Frequency localization with $n = 32, k = 4$, SNR = 10 dB. Blue represents the true frequencies, while red are given by Lasso. For better visualization, we threshold the Lasso solution by 10^{-6}.

Note that the denoising performance of Lasso improves with increased grid size as shown in the MSE vs SNR plot in Figure 5.2(a). The figure shows that the performance improvement for larger grid sizes is greater at high SNRs. This is because when the noise is small, the discretization error is more dominant and finer gridding helps to reduce this error. Figures 5.2(a) and (b) also indicate that the benefits of increasing discretization levels are diminishing with the grid sizes, at a higher rate in the low SNR regime, suggesting a tradeoff among grid size, accuracy, and computational complexity.

Finally, Figure 5.2(b) provides numerical evidence supporting the assertion that frequency localization improves with increasing grid size. Lasso identifies more frequencies than the

true ones due to basis mismatch. However, these frequencies cluster around the true ones, and more importantly, finer discretization improves clustering, suggesting over-discretization coupled with clustering and peak detection as a means for frequency localization for Lasso. This observation does not contradict the results of [33] where the authors look at the full Fourier basis ($N = n$) and the noise-free case. This is the situation where discretization effect is most prominent. We instead look at the scenario where $N \gg n$.

Figure 5.3 shows how the error metrics vary with increasing SNR for AST, MUSIC and Cadzow. The plots only show experiments with $n = 256$ samples. These plots demonstrate that AST localizes frequencies substantially better than MUSIC and Cadzow even for low signal to noise ratios as there is very little energy in the far region of the frequencies (m_1) and has the smallest weighted mean square frequency deviation (m_2). Although we have plotted the average value in these plots, we observed spikes in the plots for Cadzow's algorithm as the average is dominated by the worst performing instances. These large errors are due to the numerical instability of polynomial root finding.

I use *performance profiles* to summarize the behavior of the various algorithms across all of the parameter settings. Performance profiles provide a good visual indicator of the relative performance of many algorithms under a variety of experimental conditions[39]. Let $\mathcal{P}$ be the set of experiments and let $\mathrm{MSE}_s(p)$ be the MSE of experiment $p \in \mathcal{P}$ using the algorithm s. Then the ordinate $P_s(\beta)$ of the graph at β specifies the fraction of experiments where the ratio of the MSE of the algorithm s to the minimum MSE across all algorithms for the given experiment is less than β, i.e.,

$$P_s(\beta) = \frac{\#\left\{p \in \mathcal{P} \ : \ \mathrm{MSE}_s(p) \leq \beta \min_s \mathrm{MSE}_s(p)\right\}}{\#(\mathcal{P})}$$

From the performance profile in Figure 5.4(a), we see that AST is the best performing

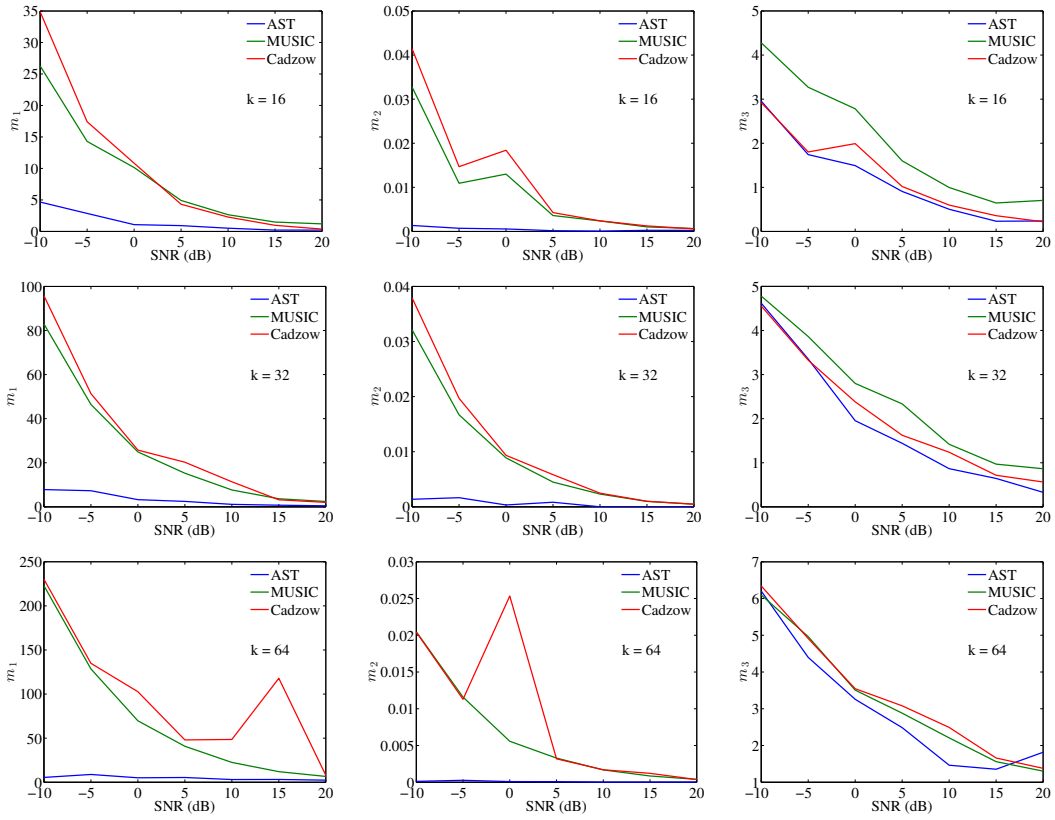
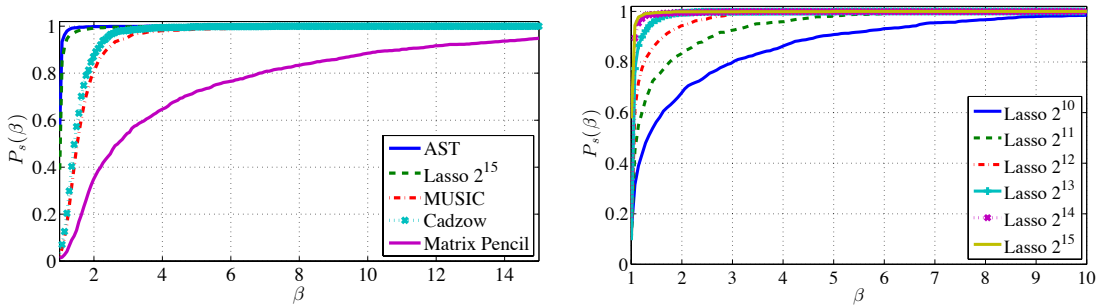


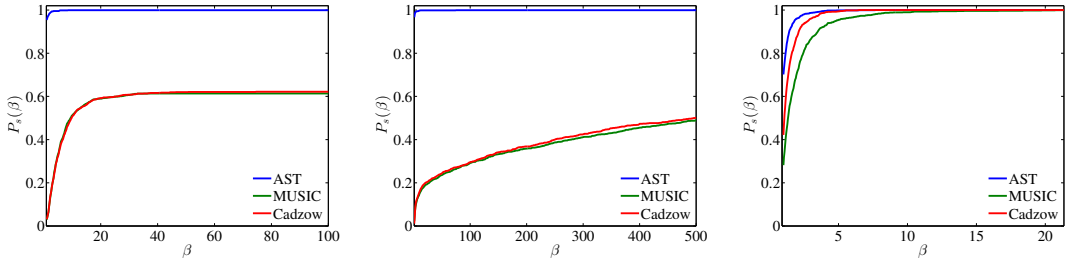
Figure 5.3: For $n = 256$ samples, the plots from left to right in order measure the average value over 20 random experiments for the error metrics m_1, m_2 and m_3 respectively. The top, middle and the bottom third of the plots respectively represent the subset of the experiments with the number of frequencies $k = 16, 32$ and 64.



(a) (b)

Figure 5.4: (a) Performance Profile comparing various algorithms and AST. (b) Performance profiles for Lasso with different grid sizes.

algorithm, with Lasso coming in second. Cadzow does not perform as well as AST, even though it is fed the true number of sinusoids. When Cadzow is fed an incorrect k, even off by 1, the performance degrades drastically, and never provides adequate mean-squared error. Figure 5.4(b) shows that the denoising performance improves with grid size.



(a) m_1 (b) m_2 (c) m_3

Figure 5.5: Performance Profiles for AST, MUSIC and Cadzow. (a) Sum of the absolute value of amplitudes in the far region (m_1) (b) The weighted frequency localization error, m_2 (c) Error in approximation of amplitudes in the near region, m_3

The performance profiles in Figure 5.5 show that AST is the best performing algorithm for all the three metrics for frequency localization. AST in fact outperforms MUSIC and Cadzow by a substantial margin for metrics m_1 and m_2.

6 Conclusion and Future Work

The convex optimization perspective on problems in signal processing and systems holds a lot of promise. This work forms the rudiments of continuous sparse recovery which underlies a number of estimation problems in parametric signal processing, systems theory and machine learning. In fact, using the atomic norm framework introduced by [30], we were able to revisit classical problems and empirically and theoretically demonstrate performance favorable to and comparable to the state-of-art.

The primary contribution of this thesis is a novel way to analyze regularization with convex penalties for the case of continuous dictionaries that occur naturally in several problems. In fact, we saw that the recovery performance is more tied to the target we wish to estimate and not the arbitrary coherence of the atomic set. This turns the attention away from global properties of dictionaries which answer the question "Which dictionaries enable sparse recovery using relaxations?" to properties of signals. By employing a local coherence condition characterized by minimum separation between frequencies for line spectral signals, we instead ask question "Which target signals are recoverable using convex relaxations?".

However, we have only scratched the surface of what is possible by continuing this line of inquiry. In [97], the authors extended the work descrived in this thesis to show that using the atomic norm framework, we could achieve continuous compressed sensing using Fourier measurements. My theoretical analysis of the optimal performance for continuous sparse recovery is heavily tied to Fourier measurements and properties of trigonometric polynomials derived in [20, 19]. Future work should explore abstract conditions when we can guarantee similarly optimal rates.

Though chapter 3 makes significant progress at understanding the theoretical limits of line-spectral estimation and superresolution, the bounds could still be improved. For instance, it

remains open as to whether the logarithmic term in Theorem 3.2 can be improved to $\log(n/k)$. Deriving such an upper bound or improving our minimax lower bound would provide an interesting direction for future work. Additionally, it is not clear if our localization bounds in Theorem 3.3 have the optimal dependence on the number of sinusoids k. For instance, one expects that the condition on signal amplitudes for approximate support recovery should not depend on k, by comparison with similar guarantees that have been established for Lasso [24]. It is reasonable to conjecture that for a large enough regularization parameter, there should will be no spurious recovered frequencies in the solution. That is, there should be no non-zero coefficients in the "far region" F in Theorem 3.3. Future work should investigate whether better guarantees on frequency localization are possible.

The analysis in chapter 4 also focused on the particular case of sampling the frequency response at regular intervals. While this example contains the critical ingredients to computing convergence rate, it is not straightforward to extend to the other sampling models described in Section 4.2.3. A useful line of study would extend this analysis to estimating transfer functions from pairs of input-output time series. The rates derived demonstrate that the DAST algorithm is asymptotically consistent, but is quite crude. It may be possible to improve the rates by leveraging more of the geometry of the set of single-pole transfer functions. It would be interesting to find reasonable lower-bounds on the reconstruction error from limited measurements, and to see how close we can match these worst-case estimates via a new analysis.

Algorithmically, the success of the method depends on whether we can efficiently compute atomic norms. We saw that we can sometimes get exact semidefinite characterizations and provided scalable algorithms. We also saw that discretization approximates the solution although it does not yield as sparse a solution as the exact approach. An interesting line of work is showing how to bootstrap an approximate solution from a fast discretization algorithm

and incrementally increase the accuracy of the solution. While gridding enables us to quickly solve atomic norm problems, a main drawback is that we can never exactly localize the true composing atoms without an extremely fine grid. One recent proposal to enable such a localization uses a linearization technique to simultaneously fit a model on the grid points and at the derivatives of the transfer functions at these grid points [50]. It would be interesting to see whether this argument can give rigorous guarantees of localization for general atomic sets.

Analysis of the convergence of greedy algorithms provides yet another fruitful direction for scaling the performance of atomic norm regularization. While there is already some interesting work on employing greedy methods for atomic norms [99, 82], the proof of convergence depends upon global restricted smoothness property. It would be interesting to pursue improvements to this assumption.

A Appendix

In this section, I collect useful results about the dual polynomial for line spectral signals from previous work, which is used in the analysis in Chapter 3. For some of these theorems, I also derive some useful corollaries for some of these theorems which we used. In addition to Theorem 3.6, we will need another result in [19] where the authors show the existence of a trigonometric polynomial Q_1 that is linear in each N_j which is also an essential ingredient in the proofs.

Theorem A.1 (Lemma 2.7 in [19])**.** *For any $f_1, \ldots, f_k$ satisfying* (3.2) *and any sign vector $v \in \mathbb{C}^k$ with $|v_j| = 1$, there exists a polynomial $Q_1 = \langle q_1, a(f) \rangle$ for some $q_1 \in \mathbb{C}^n$ with the following properties:*

1. *For every $f \in N_j$, there exists a numerical constant C_a^1 such that*

$$|Q_1(f) - v_j(f - f_j)| \leq \frac{n}{2} C_a^1 (f - f_j)^2 \tag{A.1}$$

2. *For $f \in F$, there exists a numerical constant C_b^1 such that*

$$|Q_1(f)| \leq \frac{C_b^1}{n}. \tag{A.2}$$

We will also need the following straightforward consequence of the constructions of the polynomials in Theorem 3.6, Theorem A.1, and Section 3.7.4.

Lemma A.2. *There exists a numerical constant C such that the constructed $Q(f)$ in Theorem*

3.6, $Q_1(f)$ in Theorem A.1, and $Q_j^\star(f)$ in Section 3.7.4 satisfy respectively

$$\|Q(f)\|_1 := \int_0^1 |Q(f)|df \leq \frac{Ck}{n} \tag{A.3}$$

$$\|Q_1(f)\|_1 \leq \frac{Ck}{n^2} \tag{A.4}$$

$$\|Q_j^\star\|_1 \leq \frac{Ck}{n}. \tag{A.5}$$

Proof. I will give a detailed proof of (A.3), and list the necessary modifications for proving (A.4) and (A.5). The dual polynomial $Q(f)$ constructed in [20] is of the form

$$Q\left(f\right) \quad = \quad \sum_{f_j \in T} \alpha_j K\left(f - f_j\right) + \sum_{f_j \in T} \beta_j K'\left(f - f_j\right) \tag{A.6}$$

where $K\left(f\right)$ is the squared Fejér kernel (recall that $m = (n-1)/2$)

$$K\left(f\right) \quad = \quad \left(\frac{\sin\left(\left(\frac{m}{2}+1\right)\pi f\right)}{\left(\frac{m}{2}+1\right)\sin\left(\pi f\right)} \right)^4$$

and for $n \geq 257$, the coefficients $\alpha \in \mathbb{C}^k$ and $\beta \in \mathbb{C}^k$ satisfy [20, Lemma 2.2]

$$\begin{aligned} \|\alpha\|_\infty &\leq C_\alpha \\ \|\beta\|_\infty &\leq \frac{C_\beta}{n} \end{aligned}$$

for some numerical constants C_α and C_β. Using (A.6) and triangle inequality, we bound

$\|Q(f)\|_1$ as follows:

$$\begin{aligned}
\|Q(f)\|_1 &= \int_0^1 |Q(f)|\, df \\
&\leq k\|\alpha\|_\infty \int_0^1 |K(f)|\, df + k\|\beta\|_\infty \int_0^1 |K'(f)|\, df \qquad \text{(A.7)} \\
&\leq C_\alpha k \int_0^1 |K(f)|\, df + \frac{C_\beta}{n} k \int_0^1 |K'(f)|\, df, \qquad \text{(A.8)}
\end{aligned}$$

To continue, note that $\int_0^1 |K(f)|df = \int_0^1 |G(f)|^2 df =: \|G(f)\|_2^2$ where $G(f)$ is the Fejér kernel, since $K(f)$ is the squared Fejér kernel. We can write

$$G(f) = \left(\frac{\sin\left(\pi\left(\frac{m}{2}+1\right)f\right)}{\left(\frac{m}{2}+1\right)\sin(\pi f)} \right)^2 = \sum_{l=-m/2}^{m/2} g_l e^{-i2\pi f l} \qquad \text{(A.9)}$$

where $g_l = \left(\frac{m}{2}+1-|l|\right) / \left(\frac{m}{2}+1\right)^2$. Now, by using Parseval's identity, we obtain

$$\begin{aligned}
\int_0^1 |K(f)|df &= \int_0^1 |G(f)|^2 df = \sum_{l=-m/2}^{m/2} |g_l|^2 \\
&= \frac{1}{\left(\frac{m}{2}+1\right)^4} \left(\left(\frac{m}{2}+1\right)^2 + 2\sum_{l=1}^{m/2} \left(\frac{m}{2}+1-l\right)^2 \right) \\
&= \frac{1}{\left(\frac{m}{2}+1\right)^4} \left(\left(\frac{m}{2}+1\right)^2 + 2\sum_{l=1}^{m/2} l^2 \right) \\
&\leq \frac{C}{n} \qquad \text{(A.10)}
\end{aligned}$$

for some numerical constant C when $n = 2m+1 \geq 10$.

Now let us turn our attention to $\int_0^1 |K'(f)|df$. Since $K(f) = G(f)^2$, we have

$$\int_0^1 |K'(f)|df = 2\int_0^1 |G(f)G'(f)|df \leq 2\|G(f)\|_2 \|G'(f)\|_2 \qquad \text{(A.11)}$$

We have already established that $\|G(f)\|_2^2 \leq C/n$. Let us now show that $\|G'(f)\|_2^2 \leq C'n$. Differentiating the expression for $G(f)$ in (A.9), we get

$$G'(f) \;=\; -2\pi i \sum_{l=-m/2}^{m/2} l g_l e^{-i2\pi f l}$$

Therefore, by applying Parseval's identity again, we get

$$\begin{aligned} \|G'(f)\|_2^2 &= 4\pi^2 \sum_{l=-m/2}^{m/2} l^2 |g_l|^2 \\ &\leq \pi^2 m^2 \sum_{l=-m/2}^{m/2} |g_l|^2 \\ &\leq C'n \end{aligned}$$

Plugging back into (A.11) yields

$$\int_0^1 |K'(f)| df \leq C \tag{A.12}$$

for some constant C. Combining (A.12) and (A.10) with (A.8) gives the desired result in (A.3).

The dual polynomial $Q_1(f)$ is also of the form (A.6) with coefficient vectors α_1 and β_1, which satisfy [19, Proof of Lemma 2.7]

$$\begin{aligned} \|\alpha_1\|_\infty &\leq \frac{C_{\alpha_1}}{n}, \\ \|\beta_1\|_\infty &\leq \frac{C_{\beta_1}}{n^2}. \end{aligned}$$

Combining the above two bounds with (A.7), (A.12) and (A.10) gives the desired result in

(A.4).

The last polynomial $Q_j^\star$ also has the form (A.6) with coefficient vectors $\alpha^\star$ and $\beta^\star$. According to [52, Proof of Lemma 2.2], these coefficients satisfy

$$\|\alpha^\star\|_\infty \leq C_{\alpha_\star},$$
$$\|\beta_\star\|_\infty \leq \frac{C_{\beta_\star}}{n},$$

which yields (A.5) following the same argument leading to (A.3).

□

Using Lemma A.2, we can derive the estimates we need in the following lemma.

Lemma A.3. *Let $\nu = \hat{\mu} - \mu$ be the difference measure. Then, there exists numerical constant $C > 0$ such that*

$$\left| \int_0^1 Q(f)\nu(df) \right| \leq \frac{Ck\tau}{n} \tag{A.13}$$
$$\left| \int_0^1 Q_1(f)\nu(df) \right| \leq \frac{Ck\tau}{n^2} \tag{A.14}$$
$$\left| \int_0^1 Q_j^\star(f)\nu(df) \right| \leq \frac{Ck\tau}{n}. \tag{A.15}$$

Proof. Let $Q_0 = \langle q_0, a(f) \rangle$ be a general trigonometric polynomial associated with $q_0 \in \mathbb{C}^n$.

Then,

$$\begin{aligned}\left|\int_0^1 Q_0(f)\nu(df)\right| &= \left|\int_0^1 \langle q_0, a(f)\rangle \nu(df)\right| \\ &= \left|\langle q_0, \int_0^1 a(f)\nu(df)\rangle\right| \\ &= |\langle q_0, e\rangle| \\ &= |\langle Q_0(f), E(f)\rangle| \\ &\leq \|Q_0(f)\|_1 \|E(f)\|_\infty .\end{aligned}$$

The penultimate equality follows from Parseval's theorem and the last inequality is an application of Hölder's inequality. Then, the result follows by using Lemma A.2 and (3.20). □

Bibliography

[1] X. Andrade, J. Sanders, and A. Aspuru-Guzik. “Application of compressed sensing to the simulation of atomic systems”. *Proc. of the National Academy of Sciences* 109.35 (2012), pp. 13928–13933.

[2] J.-M. Azais, Y. De Castro, and F. Gamboa. “Spike detection from inaccurate samplings”. Preprint available at `http://arxiv.org/abs/1301.5873`. 2013.

[3] R. Baraniuk et al. “Model-based compressive sensing”. *IEEE Trans. on Information Theory* 56.4 (2010), pp. 1982–2001.

[4] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1997.

[5] B. N. Bhaskar and B. Recht. “Atomic Norm Denoising with Applications to Line Spectral Estimation”. *Proceedings of the 49th Annual Allerton Conference*. 2011.

[6] B. N. Bhaskar, G. Tang, and B. Recht. “Atomic norm denoising with applications to line spectral estimation”. *Submitted to IEEE Transactions on Signal Processsing* (2012). Preprint available at `http://arxiv.org/1204.0562`.

[7] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. “Simultaneous analysis of Lasso and Dantzig selector”. *The Annals of Statistics* 37.4 (2009), pp. 1705–1732.

[8] T. Blu et al. “Sparse sampling of signal innovations”. *IEEE Signal Processing Magazine* 25.2 (2008), pp. 31–40.

[9] F. F. Bonsall. “A general atomic decomposition theorem and Banach’s closed range theorem”. *The Quarterly Journal of Mathematics* 42.1 (1991), pp. 9–14.

[10] F. F. Bonsall and D. Walsh. “Symbols for trace class Hankel operators with good estimates for norms”. *Glasgow Mathematical Journal* 28 (1986), pp. 47–54.

[11] L. Borcea et al. “Imaging and time reversal in random media”. *Inverse Problems* 18 (2002), p. 1247.

[12] S. Bourguignon, H. Carfantan, and J. Idier. “A sparsity-based method for the estimation of spectral lines from irregularly sampled data”. *IEEE Journal of Selected Topics in Signal Processing* 1.4 (2007), pp. 575–585.

[13] S. Boyd et al. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. *Foundations and Trends in Machine Learning* 3.1 (2011), pp. 1–122.

[14] F. Bunea, A. Tsybakov, and M. Wegkamp. “Sparsity oracle inequalities for the Lasso”. *Electronic Journal of Statistics* 1 (2007), pp. 169–194.

[15] J. A. Cadzow. “Signal enhancement-a composite property mapping algorithm”. *IEEE Trans. on Acoustics, Speech and Signal Processing* 36.1 (1988), pp. 49–62.

[16] J. A. Cadzow. “Spectral estimation: An overdetermined rational model equation approach”. *Proc. of the IEEE* 70.9 (2005), pp. 907–939.

[17] E. J. Candès, J. Romberg, and T. Tao. “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”. *IEEE Trans. Inform. Theory* 52.2 (2006), pp. 489–509.

[18] E. Candès, J. Rombergh, and T. Tao. “Stable signal recovery from incomplete and inaccurate measurements”. *Comm. on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223.

[19] E. Candès and C. Fernandez-Granda. “Super-resolution from noisy data”. Preprint available at `http://arxiv.org/abs/1211.0290`. 2012.

[20] E. Candès and C. Fernandez-Granda. “Towards a mathematical theory of super-resolution”. *To appear in Communications on Pure and Applied Mathematics* (2012). Preprint available at `http://arxiv.org/abs/1203.5871`.

[21] E. Candès and B. Recht. “Exact Matrix Completion via Convex Optimization”. *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[22] E. J. Candès and M. A. Davenport. “How well can we estimate a sparse vector?” *Applied and Computational Harmonic Analysis* (2013).

[23] E. J. Candès and Y. Plan. “A probabilistic and RIPless theory of compressed sensing”. *IEEE Trans. on Information Theory* 57.11 (2011), pp. 7235–7254.

[24] E. J. Candès and Y. Plan. “Near-ideal model selection by ℓ_1 minimization”. *The Annals of Statistics* 37.5A (2009), pp. 2145–2177.

[25] E. J. Candès and T. Tao. “Decoding by linear programming”. *Information Theory, IEEE Transactions on* 51.12 (2005), pp. 4203–4215.

[26] C. Carathéodory. “Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen”. *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 32.1 (1911), pp. 193–217.

[27] C. Carathéodory and L. Fejér. “Über den Zusammenhang der extremen von harmonischen Funktionen mit ihren Koeffizienten und über den Picard-Landauschen Satz”. *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 32.1 (1911), pp. 218–239.

[28] R. Carriere and R. Moses. “High resolution radar target modeling using a modified Prony estimator”. *IEEE Trans. on Antennas and Propagation* 40.1 (1992), pp. 13–18.

[29] Y. de Castro and F. Gamboa. “Exact reconstruction using Beurling minimal extrapolation”. *Journal of Mathematical Analysis and Applications* 395.1 (2012), pp. 336 –354.

[30] V. Chandrasekaran et al. “The Convex Geometry of Linear Inverse Problems”. English. *Foundations of Computational Mathematics* 12 (6 2012), pp. 805–849.

[31] S. S. Chen, D. L. Donoho, and M. A. Saunders. “Atomic decomposition by basis pursuit”. *SIAM journal on scientific computing* 20.1 (1998), pp. 33–61.

[32] S. Chen and D. Donoho. “Application of basis pursuit in spectrum estimation”. *Proc. IEEE Intl. Conference on Acoustics, Speech and Signal Processing*. Vol. 3. IEEE. 1998, pp. 1865–1868.

[33] Y. Chi et al. “Sensitivity to basis mismatch in compressed sensing”. *Signal Processing, IEEE Transactions on* 59.5 (2011), pp. 2182–2195.

[34] M. Chu, R. Funderlic, and R. Plemmons. “Structured low rank approximation”. *Linear algebra and its applications* 366 (2003), pp. 157–172.

[35] J. F. Claerbout and F. Muir. “Robust modeling with erratic data”. *Geophysics* 38.5 (1973), pp. 826–844.

[36] R. R. Coifman and R. Rochberg. “Representation theorems for holomorphic and harmonic functions in L^p”. *Asterisque* 77 (1980), pp. 11–66.

[37] R. Curto. “An operator-theoretic approach to truncated moment problems”. *Banach Center Publications* 38 (1997), pp. 75–104.

[38] K. R. Davidson and S. J. Szarek. “Local Operator Theory, Random Matrices and Banach Spaces”. *Handbook on the Geometry of Banach spaces*. Ed. by W. B. Johnson and J. Lindenstrauss. Elsevier Scientific, 2001, pp. 317–366.

[39] E. Dolan and J. Moré. “Benchmarking optimization software with performance profiles”. *Mathematical Programming* 91.2 (2002), pp. 201–213.

[40] D. L. Donoho. “For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution”. *Communications on pure and applied mathematics* 59.6 (2006), pp. 797–829.

[41] D. L. Donoho and M. Elad. “Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization”. *Proceedings of the National Academy of Sciences* 100.5 (2003), pp. 2197–2202.

[42] D. L. Donoho and X. Huo. “Uncertainty principles and ideal atomic decomposition”. *Information Theory, IEEE Transactions on* 47.7 (2001), pp. 2845–2862.

[43] D. L. Donoho and I. M. Johnstone. “Ideal Spatial Adaptation by Wavelet Shrinkage”. *Biometrika* 81.3 (1994), pp. 425–455.

[44] D. L. Donoho and J. Tanner. “Neighborliness of randomly projected simplices in high dimensions”. *Proceedings of the National Academy of Sciences of the United States of America* 102.27 (2005), pp. 9452–9457.

[45] D. Donoho. “De-noising by soft-thresholding”. *IEEE Trans. on Information Theory* 41.3 (1995), pp. 613–627.

[46] D. Donoho et al. “Maximum entropy and the nearly black object”. *Journal of the Royal Statistical Society. Series B (Methodological)* (1992), pp. 41–81.

[47] M. Duarte and R. Baraniuk. “Spectral compressive sensing”. *Applied and Computational Harmonic Analysis* (2012).

[48] G. E. Dullerud and F. G. Paganini. *A course in robust control theory: A convex approach*. Springer, 2000.

[49] B. A. Dumitrescu. *Positive Trigonometric Polynomials and Signal Processing Applications*. Netherlands: Springer, 2007.

[50] C. Ekanadham, D. Tranchina, and E. P. Simoncelli. "Recovery of Sparse Translation-Invariant Signals with Continuous Basis Pursuit". *IEEE Transactions on Signal Processing* 59.10 (2011), pp. 4735–4744.

[51] M. Fazel, H. Hindi, and S. Boyd. "A rank minimization heuristic with application to minimum order system approximation". *Proceedings of the American Control Conference*. 2001.

[52] C. Fernandez-Granda. "Support detection in super-resolution". Preprint available at `http://arxiv.org/abs/1302.3921`. 2013.

[53] S. van de Geer and P. Bühlmann. "On the conditions used to prove oracle results for the lasso". *Electronic Journal of Statistics* 3 (2009), pp. 1360–1392.

[54] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of Time Series Structure: SSA and related techniques*. Vol. 90. Chapman & Hall/CRC, 2001.

[55] M. Grant and S. Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 1.21*. `http://cvxr.com/cvx`. May 2010.

[56] E. Greenshtein and Y. Ritov. "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization". *Bernoulli* 10.6 (2004), pp. 971–988.

[57] U. Grenander and G. Szegő. *Toeplitz forms and their applications*. Chelsea Pub Co, 2001.

[58] T Hale, W Yin, and Y Zhang. "A fixed-point continuation method for ℓ_1-regularized minimization: Methodology and convergence". *SIAM Joural on Optimization* 19 (2008), pp. 1107–1130.

[59] G Herglotz. “Über Potenzreihen mit positivem, reellem Teil im Einheitskreis.” *Ber. Verh. Sachs. Akad. Wiss. Leipzig* 63 (1911), pp. 501–511.

[60] Y. Hua and T. Sarkar. “Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise”. *IEEE Trans. on Acoustics, Speech and Signal Processing* 38.5 (1990), pp. 814–824.

[61] I. Johnstone. “On minimax estimation of a sparse normal mean vector”. *The Annals of Statistics* 22.1 (1994), pp. 271–289.

[62] M. Kahn et al. “On the consistency of Prony’s method and related algorithms”. *Journal of Computational and Graphical Statistics* 1.4 (1992), pp. 329–349.

[63] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Vol. 7. Cambridge University Press, 2003.

[64] H. Krim and M. Viberg. “Two decades of array signal processing research: the parametric approach”. *IEEE Signal Processing Magazine* 13.4 (1996), pp. 67–94.

[65] T. Lai and H. Robbins. “Maximally dependent random variables”. *Proc. of the National Academy of Sciences* 73.2 (1976), p. 286.

[66] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Vol. 23. Springer, 2011.

[67] Z. Leonowicz, T. Lobos, and J. Rezmer. “Advanced spectrum estimation methods for signal analysis in power electronics”. *IEEE Trans. on Industrial Electronics* 50.3 (2003), pp. 514 –519.

[68] S. Levy and P. K. Fullagar. “Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution”. *Geophysics* 46.9 (1981), pp. 1235–1243.

[69] Z. Liu and L. Vandenberghe. “Interior-point method for nuclear norm approximation with application to system identification”. *SIAM Journal on Matrix Analysis and Applications* 31.3 (2009), pp. 1235–1256.

[70] L. Ljung. *System Identification. Theory for the user*. 2nd. Upper Saddle River, NJ: Prentice Hall, 1998.

[71] J. Löfberg. “YALMIP: A Toolbox for Modeling and Optimization in MATLAB”. *Proceedings of the CACSD Conference*. Taipei, Taiwan, 2004.

[72] D. Malioutov, M. Çetin, and A. Willsky. “A sparse signal reconstruction perspective for source localization with sensor arrays”. *IEEE Trans. on Signal Processing* 53.8 (2005), pp. 3010–3022.

[73] I. Maravic, J. Kusuma, and M. Vetterli. “Low-sampling rate UWB channel characterization and synchronization”. *Journal of Communications and Networks* 5.4 (2003), pp. 319–327.

[74] N. Meinshausen and P. Bühlmann. “High-dimensional graphs and variable selection with the lasso”. *The Annals of Statistics* 34.3 (2006), pp. 1436–1462.

[75] B. K. Natarajan. “Sparse Approximate Solutions to Linear Systems”. *SIAM Journal of Computing* 24.2 (1995), pp. 227–234.

[76] P. V. Overschee and B. D. Moor. “N4SID: Subspace algorithms for the identification of combined deterministic– stochastic systems”. *Automatica* 30 (1994), pp. 75–93.

[77] F. G. Paganini. “A set-based approach for white noise modeling”. *IEEE Transactions on Automatic Control* 41.10 (1996), pp. 1453–1465.

[78] J. R. Partington. *An Introduction to Hankel Operators*. Cambridge University Press, 1989.

[79] V. V. Peller. *Hankel Operators and Their Applications*. Springer Monographs in Mathematics. Berlin: Springer, 2003.

[80] V. V. Peller. *Nuclearity of Hankel operators*. Tech. rep. E-I-79. Leningrad: LOMI Preprints, 1979.

[81] R. Prony. “Essai experimental et analytique”. *J. Ec. Polytech.(Paris)* 2 (1795), pp. 24–76.

[82] N. Rao et al. “A Greedy Forward-Backward Algorithm for Atomic Norm constrained minimization” ().

[83] G. Raskutti, M. J. Wainwright, and B. Yu. “Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q balls.” *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6976–6994.

[84] B. Recht, M. Fazel, and P. Parrilo. “Guaranteed Minimum Rank Solutions of Matrix Equations via Nuclear Norm Minimization”. *SIAM Review* 52.3 (2010), pp. 471–501.

[85] R. Roy and T. Kailath. “ESPRIT - estimation of signal parameters via rotational invariance techniques”. *IEEE Trans. on Acoustics, Speech and Signal Processing* 37.7 (1989), pp. 984–995.

[86] F. Santosa and W. W. Symes. “Linear inversion of band-limited reflection seismograms”. *SIAM Journal on Scientific and Statistical Computing* 7.4 (1986), pp. 1307–1330.

[87] A. Schaeffer. “Inequalities of A. Markoff and S. Bernstein for polynomials and related functions”. *Bull. Amer. Math. Soc* 47 (1941), pp. 565–579.

[88] A. Schaeffer. “Inequalities of A. Markoff and S. Bernstein for polynomials and related functions”. *Bull. Amer. Math. Soc* 47 (1941), pp. 565–579.

[89] R. Schmidt. “Multiple emitter location and signal parameter estimation”. *IEEE Trans. on Antennas and Propagation* 34.3 (1986), pp. 276–280.

[90] R. S. Smith. “Nuclear norm minimization methods for frequency domain subspace identification”. *Proceedings of the American Control Conference*. 2012.

[91] P. Stoica. “List of references on spectral line analysis”. *Signal Processing* 31.3 (1993), pp. 329–340.

[92] P. Stoica and R. Moses. *Spectral analysis of signals*. Pearson/Prentice Hall, 2005.

[93] J. F. Sturm. “Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones”. *Optimization Methods and Software* 11-12 (1999), pp. 625–653.

[94] M. Talagrand. *The generic chaining*. Springer, 2005.

[95] G. Tang et al. “Compressed Sensing off the Grid”. *arXiv Preprint 1207.6053* (2012).

[96] G. Tang, B. N. Bhaskar, and B. Recht. “Near Minimax Line Spectral Estimation”. *Submitted to IEEE Transactions on Information Theory* (2013). Preprint available at `http://arxiv.org/abs/1303.4348`.

[97] G. Tang et al. “Compressed Sensing off the Grid”. *Submitted to IEEE Transactions on Information Theory* (2012). Preprint available at `http://arxiv.org/abs/1207.6053`.

[98] H. L. Taylor, S. C. Banks, and J. F. McCoy. “Deconvolution with the 1 norm”. *Geophysics* 44.1 (1979), pp. 39–52.

[99] A. Tewari, P. Ravikumar, and I. S. Dhillon. “Greedy algorithms for structurally constrained high dimensional problems”. *Adv. NIPS* (2011).

[100] R. Tibshirani. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society. Series B* 58.1 (1996), pp. 267–288.

[101] O. Toeplitz. “Zur Theorie der quadratischen und bilinearen Formen von unendlichvielen Veränderlichen”. *Mathematische Annalen* 70.3 (1911), pp. 351–376.

[102] K.-C. Toh, M. J. Todd, and R. H. Tütüncü. “SDPT3: a MATLAB software package for semidefinite programming, version 1.3”. *Optimization Methods and Software* 11.1-4 (1999), pp. 545–581.

[103] R. Vautard, P. Yiou, and M. Ghil. “Singular-spectrum analysis: A toolkit for short, noisy chaotic signals”. *Physica D: Nonlinear Phenomena* 58.1 (1992), pp. 95–126.

[104] M. Verhaegen and P. Dewilde. “Subspace model identification”. *Int. J. of Control* 56.5 (1992), pp. 1187–1210.

[105] V. Viti, C. Petrucci, and P. Barone. “Prony methods in NMR spectroscopy”. *Intl. Journal of Imaging Systems and Technology* 8.6 (1997), pp. 565–571.

[106] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. “Sparse reconstruction by separable approximation”. *IEEE Trans. on Signal Processing* 57.7 (2009), pp. 2479–2493.

[107] M. Yuan and Y. Lin. “Model selection and estimation in regression with grouped variables”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

[108] A. Zhigljavsky. “Singular Spectrum Analysis for time series: Introduction to this special issue”. *Statistics and its Interface* 3.3 (2010), pp. 255–258.

[109] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. New Jersey: Prentice Hall, 1995.

[110] K. Zhu. *Operator Theory in Function Spaces*. Providence: American Mathematical Society, 2007.

[111] G. Zweig. “Super-resolution Fourier transforms by optimisation, and ISAR imaging”. *IEE Proc. on Radar, Sonar and Navigation*. Vol. 150. IET. 2003, pp. 247–52.