JUST RELAX: CONVEX PROGRAMMING METHODS FOR SUBSET SELECTION AND SPARSE APPROXIMATION

JOEL A. TROPP

Abstract. Subset selection and sparse approximation problems request a good approximation of an input signal using a linear combination of elementary signals, yet they stipulate that the approximation may only involve a few of the elementary signals. This class of problems arises throughout electrical engineering, applied mathematics and statistics, but small theoretical progress has been made over the last fifty years. Subset selection and sparse approximation both admit natural convex relaxations, but the literature contains few results on the behavior of these relaxations for general input signals. This report demonstrates that the solution of the convex program frequently coincides with the solution of the original approximation problem. The proofs depend essentially on geometric properties of the ensemble of elementary signals. The results are powerful because sparse approximation problems are combinatorial, while convex programs can be solved in polynomial time with standard software. Comparable new results for a greedy algorithm, Orthogonal Matching Pursuit, are also stated. This report should have a major practical impact because the theory applies immediately to many real-world signal processing problems.

1. Introduction

Lately, there has been a lot of fuss about sparse approximation. This class of problems has two defining characteristics:

(1) An input signal is approximated with a linear combination of elementary signals. In many modern applications, the elementary signals are drawn from a large, linearly dependent collection.
(2) A preference for "sparse" linear combinations is imposed by penalizing nonzero coefficients. The most common penalty is the number of elementary signals that participate in the approximation.

The problem domain must justify the linear model, the choice of elementary signals and the sparsity criterion. This report discusses two abstract problems of this form, and it demonstrates that a well-known numerical method often yields nearly optimal solutions to both problems.

⋆ ⋆ ⋆ ⋆ ⋆

Sparse approximation has been studied for nearly a century, and it has numerous applications. Temlyakov [Tem02] locates the first example in a 1907 paper of Schmidt [Sch07]. In the 1950s, statisticians launched an extensive investigation of another sparse approximation problem called subset selection in regression [Mil02]. Later, approximation theorists began a systematic study of m-term approximation with respect to orthonormal bases and redundant systems [DeV98, Tem02].

Date: 14 February 2004. Revised 10 April 2004.

2000 *Mathematics Subject Classification.* Primary: 41A46, 90C25, 90C27; Secondary: 51N15, 52C17.

Key words and phrases. Highly nonlinear approximation, sparse approximation, subset selection, overcomplete representation, combinatorial optimization, relaxation, convex programming, approximation algorithm, Basis Pursuit, Orthogonal Matching Pursuit, lasso, projective spaces, packing, covering.

Citation information: ICES Report 0404, The University of Texas at Austin, February 2004. Available from `http://www.ices.utexas.edu/reports/2004.html`.

Over the last decade, the signal processing community—spurred by the work of Coifman et al. [CM89, CW92] and Mallat et al. [MZ93, DMZ94, DMA97]—has become interested in sparse representations for compression and analysis of audio [GB03], images [FViVK04] and video [NZ03]. Sparsity criteria also arise in deconvolution [TBM79], signal modeling [Ris79], pre-conditioning [GH97], machine learning [Gir98], de-noising [CDS99] and regularization [DDM03].

Most sparse approximation problems employ a linear model in which the collection of elementary signals is both linearly dependent and large. These models are often called redundant or overcomplete. Recent research suggests that overcomplete models offer a genuine increase in approximation power [RB98, FV03]. Unfortunately, they also raise a serious challenge. How do we find a good representation of the input signal among the plethora of possibilities? One method is to select a parsimonious or sparse representation. The exact rationale for invoking sparsity may range from engineering to economics to philosophy. Three common justifications:

(1) It is sometimes known *a priori* that the input signal can be expressed as a short linear combination of elementary signals that has been contaminated with noise.
(2) The approximation may have an associated cost that must be controlled. For example, the computational cost of evaluating the approximation depends on the number of elementary signals that participate. In compression, the goal is to minimize the number of bits required to store the approximation.
(3) Some researchers cite Occam's Razor, *"Pluralitas non est ponenda sine necessitate."* Causes must not be multiplied beyond necessity.[1]

Sparse approximation problems are computationally challenging because most reasonable sparsity measures are not convex. A formal hardness proof for one important class of problems independently appeared in [Nat95, DMA97]. Researchers have published a vast array of heuristic methods for producing sparse approximations, but the literature contains few guarantees on the performance of these approaches. The numerical techniques fall into five basic categories:

(1) The convex relaxation approach replaces the nonconvex sparsity measure with a related convex function to obtain a convex programming problem. The convex program can be solved in polynomial time with standard software [BV04], and one expects that it will yield a good sparse approximation. Much more will be said in the sequel.
(2) Greedy methods make a sequence of locally optimal choices in an effort to produce a good global solution to the approximation problem. This category includes forward selection procedures (such as matching pursuits), backward selection and others. Although these approaches sometimes succeed [CB00, GMS03, Tro03a, TGMS03], they can also fail spectacularly [DT96, CDS99]. The monographs of Miller [Mil02] and Temlyakov [Tem02] taste the many flavors of greedy heuristic.
(3) The Bayesian approach assumes that coefficients in the linear combination are random variables with a "sparse" prior distribution, such as Dirac + uniform or Laplace + Gaussian. Sometimes, the elementary signals are also treated as random variables with known prior distributions. Input signals are then used to estimate posterior probabilities, and the most likely models are selected for further attention. There do not seem to be any theoretical results for this paradigm [LS00, SO02, Mil02].
(4) Some researchers have developed specialized nonlinear programming software that attempts to solve sparse approximation problems directly using, for example, interior point methods [RKD99]. These techniques are only guaranteed to discover a locally optimal solution.
(5) Brute force methods sift through all potential approximations to find the global optimum. Exhaustive searches quickly become intractable as the problem size grows, and more sophisticated techniques, such as branch-and-bound, do not accelerate the hunt significantly enough to be practical [Mil02].

[1]Beware! The antiquity of Occam's Razor guarantees neither its accuracy nor its applicability [Dom99].

This paper focuses on the convex relaxation approach. Convex programming has been applied for over three decades to recover sparse representations of signals, and there is extensive empirical evidence that it often succeeds [TBM79, LF81, OSL83, SS86, Che95, CDS99, SDC03]. Although early theoretical results are beautiful, most lack practical value because they assume that the input signal admits a sparse approximation with zero error [DH01, EB02, GN03b, Fuc02, DE03, Tro03a, GN03a]. Other results are valid only in very proscribed settings [SS86, DS89]. Unhappily, the theory of sparse approximation has been marred by the lack of a proof that convex relaxation can recover the optimal sparse approximation of a *general* input signal with respect to a *general* ensemble of elementary signals. I am pleased therefore to report that this problem has finally been resolved.

This paper provides conditions which guarantee that the solution of a convex relaxation will identify the elementary signals that participate in the optimal sparse approximation of an input signal. The proofs rely on geometric properties of the ensemble of elementary signals, and a plausible argument is made that these properties are also necessary for convex relaxation to behave itself. The results here should have a significant practical impact because they cover many real applications in signal processing. Therefore, the next time one confronts a sparse approximation problem, just try to relax.

After the first draft of this report had been completed, it came to my attention that Donoho, Elad and Temlyakov had been studying some of the same problems [DET04]. Their methods and results have a rather different flavor, and their report will certainly generate significant interest on its own. I would like to ensure that the reader is aware of their investigation, and I wish to emphasize that my own work is entirely independent.

An outline of the paper would not be awry. After some preliminaries, Section 2 provides formal statements of two abstract sparse approximation problems, and it exhibits their convex relaxations. Partial statements of our major results are given without proof. Section 3 develops the algebraic, analytic and geometric concepts that will arise during the demonstrations. The fundamental lemmata appear in Section 4. Sections 5 and 6 prove that convex relaxation succeeds for the two sparse approximation problems we are studying. This theory is compared against results for a greedy algorithm, Orthogonal Matching Pursuit. Section 7 mentions some directions for further research. The first appendix provides numerical formulations of our convex relaxations, and the second appendix describes Orthogonal Matching Pursuit.

2. Sparse Approximation and Convex Relaxation

This section begins with the mathematical *mise en scène*. It then explains how sparsity is measured, which motivates the method of convex relaxation. We continue with a description of the subset selection problem and some of its properties. Then we give the convex relaxation of the subset selection problem and quote a basic result about when the relaxation succeeds. Afterward, we offer a similar treatment of the error-constrained sparse approximation problem. The section concludes with some historical remarks about convex relaxation.

2.1. The Dictionary.

We shall work in the finite-dimensional, complex inner-product space $\mathbb{C}^d$, which will be called the *signal space*.[2] The usual Hermitian inner product for $\mathbb{C}^d$ will be written as $\langle \cdot, \cdot \rangle$, and we shall denote the corresponding norm by $\|\cdot\|_2$. A *dictionary* for the signal space is a finite collection $\mathscr{D}$ of unit-norm elementary signals. The elementary signals are called *atoms*, and each is denoted by $\boldsymbol{\varphi}_\omega$, where the parameter ω is drawn from an index set Ω. The indices may have an interpretation, such as the time-frequency or time-scale localization of an atom, or they may

[2]Modifications for real signal spaces should be transparent, but the apotheosis to infinite dimensions may take additional effort.

simply be labels without an underlying metaphysics. The whole dictionary structure is written as

$$\mathscr{D} \quad = \quad \{\boldsymbol{\varphi}_\omega \ : \ \omega \in \Omega\}\,.$$

The letter N will denote the number of atoms in the dictionary.

A key parameter of the dictionary is the *coherence*, which is defined as the maximum absolute inner product between two distinct atoms:

$$\mu \quad \stackrel{\text{def}}{=} \quad \max_{\lambda \neq \omega} |\langle \boldsymbol{\varphi}_\omega, \boldsymbol{\varphi}_\lambda \rangle|\,. \tag{2.1}$$

When the coherence is large, the atoms look very similar to each other, which makes them difficult to distinguish. In the extreme case where $\mu = 1$, the dictionary contains two identical atoms. In the other direction, an orthonormal basis has zero coherence, so no atom will ever be confused for another. We say informally that a dictionary is *incoherent* when we judge that the coherence parameter is small. An incoherent dictionary may contain far more atoms than an orthonormal basis (i.e., $N \gg d$). Section 3.2 discusses this point in detail.

2.2. **Coefficient Space.** Every linear combination of atoms is parameterized by a list of coefficients. We may collect them into a *coefficient vector*, which formally belongs to $\mathbb{C}^\Omega$. In case this notation is unfamiliar, $\mathbb{C}^\Omega$ is the linear space that contains vectors whose complex components are labeled with indices from Ω. The canonical basis for this space is given by the vectors whose coordinate projections are identically zero, except for one unit coordinate. We shall denote the ω-th canonical basis vector by $\mathbf{e}_\omega$.

The *support* of a coefficient vector is the set of indices at which it is nonzero:

$$\operatorname{supp}(\boldsymbol{c}) \quad \stackrel{\text{def}}{=} \quad \{\omega \in \Omega \ : \ c_\omega \neq 0\}. \tag{2.2}$$

Suppose that $\Lambda \subset \Omega$. Without notice, we may embed "short" coefficient vectors from $\mathbb{C}^\Lambda$ into $\mathbb{C}^\Omega$ by extending them with zeros. Likewise, we may restrict "long" coefficient vectors from $\mathbb{C}^\Omega$ to their support. Both transubstantiations will be natural in context.

2.3. **Sparsity and Diversity.** The *sparsity* of a coefficient vector is the number of places where it equals zero. The complementary notion, *diversity*, counts the number of places where the coefficient vector does not equal zero. Diversity is calculated with the ℓ_0 *quasi-norm* $\|\cdot\|_0$.

$$\|\boldsymbol{c}\|_0 \quad \stackrel{\text{def}}{=} \quad |\operatorname{supp}(\boldsymbol{c})|\,. \tag{2.3}$$

We use $|\cdot|$ to indicate the cardinality of a finite set. For any positive number p, define

$$\|\boldsymbol{c}\|_p \quad \stackrel{\text{def}}{=} \quad \Big[\sum\nolimits_{\omega \in \Omega} |c_\omega|^p\Big]^{1/p} \tag{2.4}$$

with the convention that $\|\boldsymbol{c}\|_\infty \stackrel{\text{def}}{=} \max_{\omega \in \Omega} |c_\omega|$. As one might expect, there is an intimate connection between the definitions (2.3) and (2.4). Indeed, $\|\boldsymbol{c}\|_0 = \lim_{p \to 0} \|\boldsymbol{c}\|_p^p$. It is well known that the function (2.4) is convex if and only if $p \geq 1$, in which case it describes the ℓ_p norm.

Figure 1 shows how $\|\cdot\|_p$ penalizes coefficients for three different choices of p. One can see that the ℓ_1 norm is the "smallest" convex function that places a unit penalty on unit coefficients and gives away zero coefficients for free. The ℓ_2 norm satisfies the same normalization, but it charges much more than ℓ_1 for large coefficients and much less for small coefficients. From this perspective, it becomes clear that the ℓ_1 norm provides the natural convex relaxation of the ℓ_0 quasi-norm. In contrast, the ℓ_2 norm is completely inappropriate for promoting sparsity.

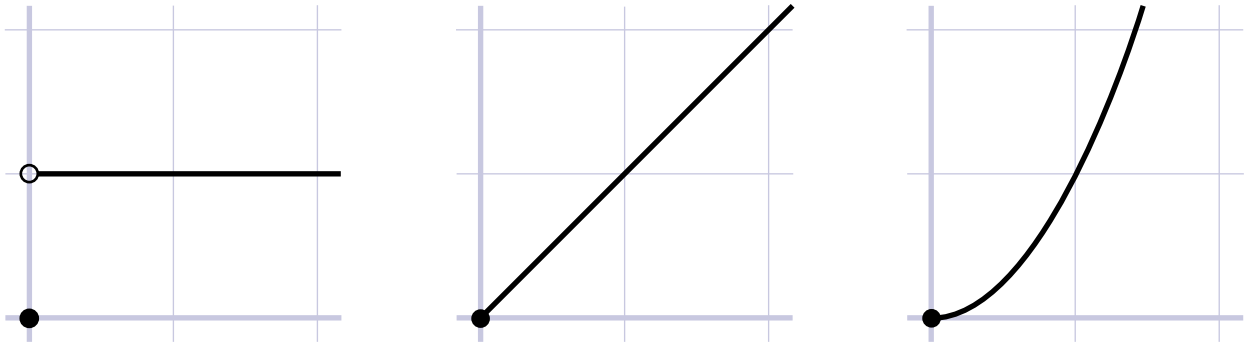

FIGURE 1. At left, the ℓ_0 quasi-norm charges one unit for each nonzero coefficient. At center, the ℓ_1 norm places a linear penalty on the magnitude of each coefficient. At right, the ℓ_2 norm imposes a quadratic cost.

2.4. **Dictionary Matrices.** Although one could use summations to express linear combinations of atoms, that notation muddies the water. Instead, let us define a matrix $\boldsymbol{\Phi}$, called the *dictionary synthesis matrix*, that maps coefficient vectors to signals. Formally,

$$\boldsymbol{\Phi} : \mathbb{C}^{\Omega} \longrightarrow \mathbb{C}^{d} \qquad \text{by the rule} \qquad \boldsymbol{\Phi} : \boldsymbol{c} \longmapsto \sum\nolimits_{\omega \in \Omega} c_{\omega}\, \boldsymbol{\varphi}_{\omega}.$$

The matrix $\boldsymbol{\Phi}$ describes the action of this linear transformation in the canonical bases of the underlying vector spaces. Therefore, the columns of $\boldsymbol{\Phi}$ are the atoms. The conjugate transpose of $\boldsymbol{\Phi}$ is called the *dictionary analysis matrix*, and it maps each signal to a coefficient vector that lists the inner products between signal and atoms.

$$\boldsymbol{\Phi}^{*} : \mathbb{C}^{d} \longrightarrow \mathbb{C}^{\Omega} \qquad \text{by the rule} \qquad \boldsymbol{\Phi}^{*} : \boldsymbol{s} \longmapsto \left[\langle \boldsymbol{s}, \boldsymbol{\varphi}_{\omega} \rangle\right]_{\omega \in \Omega}.$$

The rows of $\boldsymbol{\Phi}^{*}$ are atoms, conjugate-transposed.

2.5. **The Subset Selection Problem.** Statisticians often wish to predict the value of one random variable using a linear combination of other random variables. At the same time, they must negotiate a compromise between the number of variables involved and the mean squared prediction error to avoid overfitting. The problem of determining the correct variables is called *subset selection*, and it was probably the first type of sparse approximation to be studied in depth. As Miller laments, statisticians have made limited theoretical progress due to numerous complications that arise in the stochastic setting [Mil02].

We shall consider a deterministic version of subset selection that manages a simple tradeoff between the squared approximation error and the number of atoms that participate. Let $\boldsymbol{s}$ be an arbitrary input signal; we shall not assume that it has any particular structure nor that is is drawn from a probability distribution. Suppose that τ is a threshold that quantifies how much improvement in the approximation error is necessary before we admit an additional term into the approximation. We may state the formal problem

$$\min_{\boldsymbol{c} \in \mathbb{C}^{\Omega}} \quad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2^2 \;+\; \tau^2\, \|\boldsymbol{c}\|_0 \,. \tag{2.5}$$

Were the support of $\boldsymbol{c}$ fixed, the minimization (2.5) would be a least-squares problem. Selecting the optimal support, however, is a combinatorial nightmare; the naïve strategy involves sifting through all 2^N possibilities. In fact, if the dictionary is unrestricted, it must be NP-hard to solve (2.5) in consequence of results from [Nat95, DMA97]. Nevertheless, it may not be capricious to seek algorithms for a *particular* dictionary.

We shall attempt to solve the subset selection problem (2.5) using the convex relaxation

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \quad \tfrac{1}{2} \, \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}\|_2^2 \; + \; \gamma \, \|\boldsymbol{b}\|_1 \, . \tag{2.6}$$

Since the objective of (2.6) is an unconstrained convex function, we can use standard mathematical programming software to determine a minimizer. See Appendix A for the correct numerical formulation. The major change from (2.5) to (2.6) is obviously the substitution of the convex ℓ_1 norm for the discontinuous ℓ_0 quasi-norm. It will later become clear how γ and τ are related, and we shall understand why the two programs have slightly different structures.

The reader should be aware that convex programs of the form (2.6) have been proposed for many different applications. Geophysicists have long used them for deconvolution [TBM79, SS86]. An important technique for machine learning can sometimes be reduced to this form [Gir98]. Chen, Donoho, and Saunders have applied (2.6) to de-noise signals [CDS99], and Fuchs has put it forth for several other signal processing problems, e.g. in [Fuc97, Fuc98]. Daubechies, Defrise, and De Mol have proposed a related convex program for regularizing linear inverse problems [DDM03]. Most intriguing, perhaps, Olshausen and Field have argued that the mammalian visual cortex may solve similar minimization problems to produce sparse representations of images [OF96].

This paper asks what relationship, if any, a solution to the convex relaxation (2.6) might share with a solution to the subset selection problem (2.5). We shall see that, if the dictionary is incoherent and the threshold parameters are chosen correctly, then the relaxation identifies every significant atom from a solution to the subset selection problem and no others. Although the complete result will require some energy to understand, we can state a preliminary version now.

Theorem A. *Fix an input signal, and choose a threshold τ. Suppose that*

- *the vector $\boldsymbol{c}_{\text{opt}}$ is a solution to the subset selection problem* (2.5) *with threshold τ;*
- *the support of $\boldsymbol{c}_{\text{opt}}$ contains no more than $\frac{1}{3}\,\mu^{-1}$ indices; and*
- *the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (2.6) *with threshold $\gamma = 2\,\tau$.*

Then it follows that

- *the relaxation never selects a non-optimal atom since $\operatorname{supp}(\boldsymbol{b}_\star) \subset \operatorname{supp}(\boldsymbol{c}_{\text{opt}})$;*
- *the relaxation yields a nearly optimal coefficient vector since $\|\boldsymbol{b}_\star - \boldsymbol{c}_{\text{opt}}\|_\infty \leq 3\,\tau$;*
- *in particular, the relaxation identifies every significant atom from the optimal solution because $\lambda \in \operatorname{supp}(\boldsymbol{b}_\star)$ whenever $|\boldsymbol{c}_{\text{opt}}(\lambda)| > 3\,\tau$; and*
- *the relaxation has a unique solution.*

For orthonormal dictionaries, the full theorem yields somewhat stronger conclusions.

Theorem B. *Assume that the dictionary is orthonormal. Fix an input signal, and choose a threshold τ. Suppose that*

- *the vector $\boldsymbol{c}_{\text{opt}}$ is a solution to the subset selection problem* (2.5) *with threshold τ; and*
- *the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (2.6) *with threshold $\gamma = \tau$.*

It follows that $\operatorname{supp}(\boldsymbol{b}_\star) \subset \operatorname{supp}(\boldsymbol{c}_{\text{opt}})$ and that $\|\boldsymbol{b}_\star - \boldsymbol{c}_{\text{opt}}\|_\infty \leq \tau$. In particular, $\operatorname{supp}(\boldsymbol{b}_\star)$ contains the index λ whenever $|\boldsymbol{c}_{\text{opt}}(\lambda)| > \tau$.

This result should not come as a surprise to those who have studied the process of thresholding empirical wavelet coefficients to de-noise signals [DJ92, Mal99].

2.6. **Sparse Approximation with an Error Constraint.** In numerical analysis, a common problem is to approximate or interpolate a complicated function using a short linear combination of more elementary functions. The approximation must not commit too great an error. At the same time, one pays for each additional term in the linear combination whenever the approximation is evaluated—perhaps trillions of times. Therefore, one may wish to maximize the sparsity of the approximation subject to an error constraint [Nat95].

Suppose that $\boldsymbol{s}$ is an arbitrary input signal, and fix an error level ε. The sparse approximation problem we have described may be stated as

$$\min_{\boldsymbol{c} \in \mathbb{C}^\Omega} \|\boldsymbol{c}\|_0 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2 \;\leq\; \varepsilon. \tag{2.7}$$

Observe that this mathematical program will generally have many solutions that use the same number of atoms but yield different approximation errors. The solutions will always become sparser as the error tolerance increases.

An interesting special case of (2.7) occurs when the error tolerance ε is set to zero. This problem may be interpreted as finding the sparsest exact representation of the input signal. It is somewhat academic because the signals that can be represented with fewer than d atoms form a set of Lebesgue measure zero in $\mathbb{C}^d$ [Tro03a].

The obvious convex relaxation of (2.7) is

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2 \;\leq\; \delta. \tag{2.8}$$

Since (2.6) requests the minimizer of a convex function over a convex set, we may apply standard mathematical programming software to solve it. Appendix A provides a numerical formulation. For the case $\delta = 0$, the resulting convex program has been called *Basis Pursuit* [CDS99], and it has been studied extensively. See Section 6.1 for a discussion and a list of references.

Once again, we may ask how the solutions of (2.7) and (2.8) correspond. Our basic result is that, if the dictionary is incoherent, then the solution to the convex relaxation (2.8) for a given value of δ is at least as sparse as a solution to the sparse approximation problem (2.7) with an error tolerance somewhat smaller than δ. A preliminary statement of our result is the following.

Theorem C. *Fix an input signal, and let $m \leq \frac{1}{3}\mu^{-1}$. Suppose that*

- *the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (2.8) *with tolerance δ;*
- *the coefficient vector $\boldsymbol{c}_{\text{opt}}$ solves the sparse approximation problem* (2.7) *with error tolerance $\varepsilon = \delta/\sqrt{1 + 6\,m}$; and*
- *the support of $\boldsymbol{c}_{\text{opt}}$ contains no more than m atoms.*

Then it follows that

- *the coefficient vector $\boldsymbol{b}_\star$ is at least as sparse as $\boldsymbol{c}_{\text{opt}}$ since $\operatorname{supp}(\boldsymbol{b}_\star) \subset \operatorname{supp}(\boldsymbol{c}_{\text{opt}})$;*
- *yet $\boldsymbol{b}_\star$ is no sparser than a solution of the sparse approximation problem with tolerance δ;*
- *the vector $\boldsymbol{b}_\star$ is nearly optimal since $\|\boldsymbol{b}_\star - \boldsymbol{c}_{\text{opt}}\|_2 \leq \delta\,\sqrt{3/2}$; and*
- *the relaxation has a unique solution.*

The diminution of δ may seem unacceptably large, but one should be aware that this worst-case version of the result assumes that the residual left over after approximation is parallel with an atom. The full theorem depends strongly on how much the residual is correlated with the dictionary.

In other news, if the input signal happens to have an exact representation that requires fewer than $\frac{1}{2}\,(\mu^{-1} + 1)$ atoms, it has been proven that the convex relaxation with $\delta = 0$ will recover all the optimal atoms and their correct coefficients [Fuc02, DE03, Tro03a].

2.7. **Approximation with a Sparsity Constraint.** Approximation theorists prefer a third flavor of the sparse approximation problem, called *m-term approximation*. In this case, one is asked to provide the best approximation of a signal using a linear combination of m atoms or fewer from the dictionary. Formally,

$$\min_{\boldsymbol{c} \in \mathbb{C}^\Omega} \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2 \qquad \text{subject to} \qquad \|\boldsymbol{c}\|_0 \;\leq\; m. \tag{2.9}$$

Convex relaxation does not seem to be an appropriate method for solving this problem directly because it provides no control on the number of terms involved in the approximation. Greedy heuristics, on the other hand, can be halted when the approximation contains a specified number of

atoms. It is known that greedy algorithms succeed in some circumstances [CB00, GMS03, Tro03a, TGMS03], although they may derail catastrophically [DT96, CDS99]. Approximation theorists have also studied the asymptotic error rate of greedy approximations. Most of these results are not immediately relevant to the problems we have stated here. See the superb monographs of DeVore [DeV98] and Temlyakov [Tem02] for an introduction to this literature.

2.8. **A Brief History.** I believe that the ascendance of convex relaxation for sparse approximation was propelled by two theoretical-technological developments of the last half century. First, the philosophy and methodology of robust statistics—which derive from work of von Neumann, Tukey and Huber—show that ℓ_1 loss criteria can be applied to defend statistical estimators against outlying data points. Robust estimators qualitatively prefer a few large errors and many tiny errors to the armada of moderate deviations introduced by mean-squared-error criteria. Second, the elevation during the 1950s of linear programming to the level of *technology* and the interior-point revolution of the 1980s have made it both tractable and commonplace to solve the large-scale optimization problems that arise from convex relaxation.

It appears that a 1973 paper of Claerbout and Muir is the crucible in which these reagents were first combined for the express goal of yielding a sparse representation [CM73]. They write,

> In deconvolving any observed seismic trace, it is rather disappointing to discover that there is a nonzero spike at every point in time regardless of the data sampling rate. One might hope to find spikes only where real geologic discontinuities take place. Perhaps the L_1 norm can be utilized to give a [sparse] output trace....

This idea was subsequently developed in the geophysics literature by [TBM79, LF81, OSL83]. In 1986, Santosa and Symes proposed the convex relaxation (2.5) as a method for recovering sparse spike trains, and they proved that the method succeeds under moderate restrictions [SS86].

Around 1990, the work on ℓ_1 criteria in signal processing recycled to the statistics community. Donoho and Johnstone wrote a pathbreaking paper which proved that one could determine a nearly optimal minimax estimate of a smooth function contaminated with noise by solving the convex relaxation (2.5) where $\boldsymbol{\Phi}$ is an appropriate wavelet basis and γ is related to the variance of the noise. Slightly later, Tibshirani proposed that (2.5), which he calls *the lasso*, could be used to solve subset selection problems in the stochastic setting [Tib96]. From here, it is only a short step to Basis Pursuit and Basis Pursuit de-noising [CDS99].

This history could not be complete without mention of parallel developments in the theoretical computer sciences. It has long been known that some combinatorial problems are intimately bound up with continuous convex programming problems. In particular, the problem of determining the maximum value that an affine function attains at some vertex of a polytope can be solved using a linear program [PS98]. A major theme in modern computer science is that many other combinatorial problems can be solved approximately by means of a convex relaxation. For example, a celebrated paper of Goemans and Williamson proves that a certain convex program can be used to produce a graph cut whose weight exceeds 87% of the maximum cut [GW95]. The present work draws deeply on the fundamental idea that a combinatorial problem and its convex relaxation often have closely related solutions.

3. Dictionaries, Matrices and Geometry

The study of convex relaxation methods involves a lot of unusual matrix calculations. This section will prepare us for the proofs to come by describing the quantities that will arise. Its leitmotif is that even the most complicated expressions admit beautiful geometric explanations. In particular, one may interpret an incoherent dictionary as a configuration of lines that (nearly) satisfies several extremal properties. Most of the background material in this section may be traced to the usual suspects [HJ85, Kre89, GVL96], which will not generally be cited.

3.1. **Sub-dictionaries.** A linearly independent collection of atoms is called a *sub-dictionary*. If the atoms in a sub-dictionary are indexed by the set Λ, then we define a synthesis matrix $\boldsymbol{\Phi}_\Lambda : \mathbb{C}^\Lambda \to \mathbb{C}^d$ and an analysis matrix $\boldsymbol{\Phi}_\Lambda{}^* : \mathbb{C}^d \to \mathbb{C}^\Lambda$. These matrices are entirely analogous with the dictionary synthesis and analysis matrices. We shall frequently use the fact that the synthesis matrix $\boldsymbol{\Phi}_\Lambda$ has full column rank.

The *Gram matrix* of the sub-dictionary is given by $\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda$. Observe that the (λ, ω) entry of this matrix is the inner product $\langle \boldsymbol{\varphi}_\omega, \boldsymbol{\varphi}_\lambda \rangle$. Therefore, the Gram matrix is Hermitian, and it has a unit diagonal (since all atoms have unit Euclidean norm). One should interpret the Gram matrix as a table of the correlations between atoms listed by Λ. Note that the Gram matrix of a sub-dictionary is always invertible.

We shall encounter two other matrices frequently enough to single them out. The *Moore–Penrose generalized inverse* of the synthesis matrix is denoted by $\boldsymbol{\Phi}_\Lambda{}^+$, and it may be calculated using the formula $\boldsymbol{\Phi}_\Lambda{}^+ = (\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1} \boldsymbol{\Phi}_\Lambda{}^*$. For any signal $\boldsymbol{s}$, the coefficient vector $\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{s}$ synthesizes the best approximation of $\boldsymbol{s}$ using the atoms in Λ. The orthogonal projector that produces this best approximation will be denoted as $\boldsymbol{P}_\Lambda$. This projector may be expressed using the generalized inverse: $\boldsymbol{P}_\Lambda = \boldsymbol{\Phi}_\Lambda \boldsymbol{\Phi}_\Lambda{}^+$. Recall that $\boldsymbol{P}_\Lambda \boldsymbol{s}$ is always orthogonal to $(\boldsymbol{s} - \boldsymbol{P}_\Lambda \boldsymbol{s})$.

3.2. **Cumulative Coherence.** We have introduced the coherence parameter μ because it is easy to calculate, and yet it can be used to bound much more complicated quantities associated with the dictionary. A refinement of the coherence parameter is the *cumulative coherence*, which this author formerly called the *quasi-coherence* or *Babel function* [Tro03a, TGMS03]. It measures how much a collection of atoms can resemble a fixed, distinct atom. Formally,

$$\mu_1(m) \quad \stackrel{\text{def}}{=} \quad \max_{|\Lambda| = m} \max_{\omega \notin \Lambda} \sum\nolimits_{\lambda \in \Lambda} |\langle \boldsymbol{\varphi}_\omega, \boldsymbol{\varphi}_\lambda \rangle| \,.$$

We place the convention that $\mu_1(0) = 0$. The subscript on μ_1 serves as a mnemonic that the cumulative coherence is an absolute sum, and it distinguishes μ_1 from the coherence parameter μ. When the cumulative coherence grows slowly, we say informally that the dictionary is *incoherent* or *quasi-incoherent*.

The cumulative coherence function has an important interpretation in terms of sub-dictionaries. Suppose that Λ indexes m atoms. Then the number $\mu_1(m-1)$ gives an upper bound on the sum of the absolute off-diagonal entries in each row (or column) of the Gram matrix $\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda$. Several other facts about μ_1 follow immediately from the definition.

Proposition 3.1. *The cumulative coherence has the following properties:*

1. *It generalizes the coherence: $\mu_1(1) = \mu$ and $\mu_1(m) \leq m\,\mu$.*
2. *Its first differences are non-negative, and its second differences are non-positive.*
3. *For an orthonormal basis, $\mu_1(m) = 0$ for each m.*

It is possible to develop a lower bound on the coherence parameter in terms of the dimension d and the number of atoms N [SH03]:

$$\mu \quad \geq \quad \begin{cases} 0 & N \leq d \\ \sqrt{\frac{N-d}{d\,(N-1)}} & N > d. \end{cases}$$

If $N > d$ and the dictionary contains an orthonormal basis, the lower bound increases to $\mu \geq 1/\sqrt{d}$ [GN03b]. Although the coherence exhibits a quantum jump as soon as the number of atoms exceeds the dimension, it is possible to construct very large dictionaries with low coherence. When $d = 2^k$, Calderbank et al. have produced a striking example of a dictionary that contains $(d+1)$ orthonormal bases yet retains coherence $1/\sqrt{d}$ [CCKS97]. Gilbert, Muthukrishnan and Strauss have exhibited a method for constructing even larger dictionaries with slightly higher coherence [GMS03].

The coherence parameter of a dictionary was mentioned as a quantity of heuristic interest in [DMA97], but the first formal treatment appears in [DH01]. It is also related to an eponymous concept from the geometry of numbers [Yap00]. The concept of cumulative coherence was developed independently in [DE03, Tro03a]. In Section 3.10, we shall suggest geometric interpretations of both quantities.

3.3. **Operator Norms.** One of the most useful tools in our satchel is the operator norm. Let us treat the matrix A as a map between two finite-dimensional vector spaces equipped respectively with the ℓ_p and ℓ_q norms. The (p, q) *operator norm* of A measures the factor by which the matrix can increase the length of a vector. It may be calculated with any of the following expressions:

$$\|\mathsf{A}\|_{p,q} \stackrel{\text{def}}{=} \max_{\boldsymbol{z} \neq \mathbf{0}} \frac{\|\mathsf{A}\,\boldsymbol{z}\|_q}{\|\boldsymbol{z}\|_p} = \max_{\|\boldsymbol{z}\|_p = 1} \|\mathsf{A}\,\boldsymbol{z}\|_q = \max_{\|\boldsymbol{z}\|_p \leq 1} \|\mathsf{A}\,\boldsymbol{z}\|_q .$$

In words, the operator norm equals the maximum ℓ_q norm of any point in the image of the ℓ_p unit ball under A.

The topological dual of the finite-dimensional normed linear space $(\mathbb{C}^m, \ell_p)$ is the space $(\mathbb{C}^m, \ell_{p'})$ where $1/p + 1/p' = 1$. In particular, ℓ_1 and ℓ_∞ are dual to each other, while ℓ_2 is self-dual. If the matrix A maps from ℓ_p to ℓ_q, its conjugate transpose A^* should be viewed as a map between the dual spaces $\ell_{q'}$ and $\ell_{p'}$. Under this regime, the operator norm of a matrix always equals the operator norm of its conjugate transpose:

$$\|\mathsf{A}\|_{p,q} = \|\mathsf{A}^*\|_{q',p'} . \tag{3.1}$$

Therefore, any procedure for calculating the norm of a matrix can also be used to calculate the norm of its conjugate transpose.

A quantity related to the operator norm is the restricted minimum

$$\min_{\substack{\boldsymbol{z} \in \mathscr{R}(\mathsf{A}^*) \\ \boldsymbol{z} \neq \mathbf{0}}} \frac{\|\mathsf{A}\,\boldsymbol{z}\|_q}{\|\boldsymbol{z}\|_p} \tag{3.2}$$

where $\mathscr{R}(\cdot)$ denotes the range (i.e. column span) of its argument. The expression (3.2) measures the factor by which the non-singular part of A can decrease the length of a vector. If the matrix is surjective, we can express the minimum in terms of a generalized inverse.

Proposition 3.2. *The following bound holds for every matrix* A*.*

$$\min_{\substack{\boldsymbol{z} \in \mathscr{R}(\mathsf{A}^*) \\ \boldsymbol{z} \neq \mathbf{0}}} \frac{\|\mathsf{A}\,\boldsymbol{z}\|_q}{\|\boldsymbol{z}\|_p} \geq \left\|\mathsf{A}^+\right\|_{q,p}^{-1} . \tag{3.3}$$

If A *is surjective, equality holds in* (3.3)*. When* A *is bijective, this result implies*

$$\min_{\boldsymbol{z} \neq \mathbf{0}} \frac{\|\mathsf{A}\,\boldsymbol{z}\|_q}{\|\boldsymbol{z}\|_p} = \left\|\mathsf{A}^{-1}\right\|_{q,p}^{-1} .$$

Proof. Calculate that

$$\left[\min_{\substack{\boldsymbol{z} \in \mathscr{R}(\mathsf{A}^*) \\ \boldsymbol{z} \neq \mathbf{0}}} \frac{\|\mathsf{A}\,\boldsymbol{z}\|_q}{\|\boldsymbol{z}\|_p}\right]^{-1} = \max_{\substack{\boldsymbol{z} \in \mathscr{R}(\mathsf{A}^*) \\ \boldsymbol{z} \neq \mathbf{0}}} \frac{\|\boldsymbol{z}\|_p}{\|\mathsf{A}\,\boldsymbol{z}\|_q} = \max_{\substack{\boldsymbol{w} \in \mathscr{R}(\mathsf{A}) \\ \boldsymbol{w} \neq \mathbf{0}}} \frac{\|\mathsf{A}^+\boldsymbol{w}\|_p}{\|\boldsymbol{w}\|_q} \leq \max_{\boldsymbol{w} \neq \mathbf{0}} \frac{\|\mathsf{A}^+\boldsymbol{w}\|_p}{\|\boldsymbol{w}\|_q} .$$

The second equality holds because $\mathsf{A}^+\mathsf{A}$ is a projector onto the range of A^*. When A is surjective, the last two maxima are taken over the same set, which converts the inequality to an equality. Complete the proof by identifying the last maximum as $\|\mathsf{A}^+\|_{q,p}$. □

3.4. **Calculating Operator Norms.** Some basic operator norms can be determined with ease, while others are quite stubborn. The following table describes how to compute the most important ones.

		Co-domain		
		ℓ_1	ℓ_2	ℓ_∞
Domain	ℓ_1	Maximum ℓ_1 norm of a column	Maximum ℓ_2 norm of a column	Maximum absolute entry of matrix
	ℓ_2	NP-hard	Maximum singular value	Maximum ℓ_2 norm of a row
	ℓ_∞	NP-hard	NP-hard	Maximum ℓ_1 norm of a row

The computational complexity of the $(\infty, 1)$ norm is due to Rohn [Roh00]. Using his methods, one can prove that it is also NP-hard to calculate the $(\infty, 2)$ norm. The result for the $(2, 1)$ norm follows from equation (3.1).

3.5. **Singular Values.** Under any linear map A, the image of the Euclidean unit ball is an ellipsoid. The Euclidean norms of the axes of this ellipsoid are called the *singular values* of the map. The maximum singular value of A coincides with the $(2, 2)$ operator norm of the matrix. If A has more rows than columns, its singular values may also be defined algebraically as the square roots of the eigenvalues of $\mathsf{A}^*\mathsf{A}$.

Suppose that Λ indexes a collection of m atoms. If m is small enough, then we may develop good bounds on the singular values of $\mathbf{\Phi}_\Lambda$ using the cumulative coherence.

Proposition 3.3. *Suppose that $|\Lambda| = m$. Each singular value σ of the matrix $\mathbf{\Phi}_\Lambda$ satisfies*

$$1 \;-\; \mu_1(m-1) \quad\leq\quad \sigma^2 \quad\leq\quad 1 \;+\; \mu_1(m-1).$$

A version of this proposition was used implicitly by Gilbert, Muthukrishnan, and Strauss in [GMS03]. The present result was first published by Donoho and Elad [DE03].

Proof. Consider the Gram matrix $\mathsf{G} = \mathbf{\Phi}_\Lambda{}^*\mathbf{\Phi}_\Lambda$. The Geršgorin Disc Theorem [HJ85] states that every eigenvalue of G lies in one of the m discs

$$\Delta_\lambda \quad\stackrel{\text{def}}{=}\quad \left\{ z \;:\; |\mathsf{G}_{\lambda\lambda} - z| \;\leq\; \sum\nolimits_{\omega\neq\lambda} |\mathsf{G}_{\lambda\omega}| \right\} \qquad \text{for each } \lambda \text{ in } \Lambda.$$

The normalization of the atoms implies that $\mathsf{G}_{\lambda\lambda} \equiv 1$. Meanwhile, the sum is bounded above by $\mu_1(m-1)$. The result follows since the eigenvalues of G equal the squared singular values of $\mathbf{\Phi}_\Lambda$. □

If the m singular values of $\mathbf{\Phi}_\Lambda$ are all nonzero, then the columns of $\mathbf{\Phi}_\Lambda$ form a linearly independent set. It follows that every sufficiently small collection of atoms forms a sub-dictionary. This insight has an important consequence for the uniqueness of sparse representations.

Proposition 3.4 (Donoho–Elad [DE03], Gribonval–Nielsen [GN03b])**.** *Suppose that $\mu_1(m-1) < 1$. If a signal can be written as a linear combination of k atoms, then any other exact representation of the signal requires at least $(m - k + 1)$ atoms.*

When $N \leq d$, it is possible to develop alternate bounds on the singular values of $\mathbf{\Phi}_\Lambda$ using interlacing theorems.

Proposition 3.5. *Suppose that $N \leq d$. Then each singular value σ of the matrix $\mathbf{\Phi}_\Lambda$ satisfies*

$$\sigma_{\min}(\mathbf{\Phi}) \quad\leq\quad \sigma \quad\leq\quad \sigma_{\max}(\mathbf{\Phi}),$$

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values of a matrix.

This result follows instantly from Theorem 7.3.9 of [HJ85].

3.6. **The Inverse Gram Matrix.** Suppose that Λ indexes a sub-dictionary, and let G denote the Gram matrix $\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda$. The (∞,∞) operator norm of G^{-1} will arise in our calculations, so we need to develop a bound on it. Afterward, we shall make a connection between this inverse and the dual system of the sub-dictionary.

Proposition 3.6. *Suppose that* $\mu_1(m-1) < 1$. *Then*

$$\left\|\mathsf{G}^{-1}\right\|_{\infty,\infty} \quad = \quad \left\|\mathsf{G}^{-1}\right\|_{1,1} \quad \leq \quad \frac{1}{1-\mu_1(m-1)}. \tag{3.4}$$

This proposition was established independently in [Fuc02, Tro03a]. For comparison, observe that $\|\mathsf{G}\|_{\infty,\infty} \leq 1 + \mu_1(m-1)$.

Proof. First, note that the two operator norms in (3.4) are equal because the inverse Gram matrix is Hermitian. Since the atoms are normalized, the Gram matrix has a unit diagonal. Therefore, we may split it as the sum of its diagonal and off-diagonal parts: $\mathsf{G} = \mathsf{I}_m + \mathsf{A}$. Each row of the matrix A lists the inner products between a fixed atom and $(m-1)$ other atoms. Therefore, $\|\mathsf{A}\|_{\infty,\infty} \leq \mu_1(m-1)$. Now invert G using a Neumann series:

$$\left\|\mathsf{G}^{-1}\right\|_{\infty,\infty} \quad = \quad \left\|\sum\nolimits_{k=0}^{\infty}(-\mathsf{A})^k\right\|_{\infty,\infty} \quad \leq \quad \sum\nolimits_{k=0}^{\infty}\|\mathsf{A}\|_{\infty,\infty}^k \quad = \quad \frac{1}{1-\|\mathsf{A}\|_{\infty,\infty}}.$$

Introduce the estimate for $\|\mathsf{A}\|_{\infty,\infty}$ to complete the proof. □

The inverse of the Gram matrix has a useful interpretation. The atoms in Λ form a linearly independent set, so there is a unique collection of *dual vectors* $\{\boldsymbol{\psi}_\lambda\}_{\lambda\in\Lambda}$ that has the same linear span as $\{\boldsymbol{\varphi}_\lambda\}_{\lambda\in\Lambda}$ and that satisfies the bi-orthogonal property

$$\langle\boldsymbol{\psi}_\lambda, \boldsymbol{\varphi}_\lambda\rangle = 1 \qquad \text{and} \qquad \langle\boldsymbol{\psi}_\lambda, \boldsymbol{\varphi}_\omega\rangle = 0 \qquad \text{for } \lambda,\omega \text{ in } \Lambda \text{ and } \omega \neq \lambda.$$

That is, each dual vector $\boldsymbol{\psi}_\lambda$ is orthogonal to the atoms with different indices, and it is extruded in this (unique) direction until its inner product with $\boldsymbol{\varphi}_\lambda$ equals one. The definition of the dual system suggests that it somehow inverts the sub-dictionary. Indeed, the dual vectors form the columns of $(\boldsymbol{\Phi}_\Lambda{}^+)^*$. We may calculate that

$$\mathsf{G}^{-1} \quad = \quad (\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda)^{-1} \quad = \quad (\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda)^{-1}(\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda)(\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda)^{-1} \quad = \quad \boldsymbol{\Phi}_\Lambda{}^+(\boldsymbol{\Phi}_\Lambda{}^+)^*.$$

Therefore, the inverse Gram matrix lists the inner products between the dual vectors.

3.7. **The Exact Recovery Coefficient.** Now we shall develop a measure of the similarity between a sub-dictionary and the remaining atoms from the dictionary. Let Λ index a sub-dictionary, and define the quantity

$$\mathrm{ERC}(\Lambda;\mathscr{D}) \quad \stackrel{\text{def}}{=} \quad 1 \;-\; \max_{\omega\notin\Lambda} \left\|\boldsymbol{\Phi}_\Lambda{}^+\,\boldsymbol{\varphi}_\omega\right\|_1.$$

The letters "ERC" abbreviate the term *Exact Recovery Coefficient*, so called because $\mathrm{ERC}(\Lambda;\mathscr{D}) > 0$ is a sufficient condition for several different algorithms to recover exact superpositions of atoms from Λ. *Nota bene* that every atom in the dictionary makes a critical difference in the value of $\mathrm{ERC}(\Lambda;\mathscr{D})$. Nevertheless, we shall usually omit the dictionary from the notation.

Proposition 3.7. *Suppose that* $|\Lambda| \leq m$. *A lower bound on the Exact Recovery Coefficient is*

$$\mathrm{ERC}(\Lambda) \quad \geq \quad \frac{1 \;-\; \mu_1(m-1) \;-\; \mu_1(m)}{1 \;-\; \mu_1(m-1)}.$$

It follows that $\mathrm{ERC}(\Lambda) > 0$ *whenever*

$$\mu_1(m-1) \;+\; \mu_1(m) \quad < \quad 1.$$

This argument independently appeared in [Fuc02, Tro03a] along with better estimates for incoherent dictionaries that have additional structure. For every sub-dictionary of an orthonormal basis, the Exact Recovery Coefficient equals one.

Proof. Begin the calculation by expanding the generalized inverse and applying a norm estimate.

$$\begin{aligned}\max_{\omega \notin \Lambda} \left\| \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega \right\|_1 &= \max_{\omega \notin \Lambda} \left\| (\boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{\Phi}_\Lambda)^{-1} \, \boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\varphi}_\omega \right\|_1 \\ &\leq \left\| (\boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{\Phi}_\Lambda)^{-1} \right\|_{1,1} \max_{\omega \notin \Lambda} \left\| \boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\varphi}_\omega \right\|_1 .\end{aligned}$$

For the first term, Proposition 3.6 provides an upper bound of $[1 - \mu_1(m-1)]^{-1}$. An estimate of the second term is

$$\max_{\omega \notin \Lambda} \left\| \boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\varphi}_\omega \right\|_1 = \max_{\omega \notin \Lambda} \sum\nolimits_{\lambda \in \Lambda} |\langle \boldsymbol{\varphi}_\omega, \boldsymbol{\varphi}_\lambda \rangle| \leq \mu_1(m).$$

Combine the inequalities to prove the result. □

Now let us turn to the geometric interpretation of the Exact Recovery Coefficient. Form the collection of signals

$$\mathscr{A}_1(\Lambda; \mathscr{D}) \stackrel{\text{def}}{=} \left\{ \boldsymbol{\Phi}_\Lambda \boldsymbol{b} \ : \ \boldsymbol{b} \in \mathbb{C}^\Lambda \text{ and } \|\boldsymbol{b}\|_1 \leq 1 \right\}.$$

This definition is adapted from the approximation theory literature [DeV98, Tem02]. The set $\mathscr{A}_1(\Lambda)$ might be called the *antipodal convex hull* of the sub-dictionary because it is the smallest convex set that contains every unit multiple of every atom. See Figure 2 for an illustration.

Recall that $\boldsymbol{\Phi}_\Lambda \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega$ gives the orthogonal projection of the atom $\boldsymbol{\varphi}_\omega$ onto the span of the atoms indexed by Λ. Therefore, the coefficient vector $\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega$ can be used to synthesize this projection. We conclude that the quantity $1 - \left\| \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega \right\|_1$ measures how far the projected atom $\mathsf{P}_\Lambda \boldsymbol{\varphi}_\omega$ lies from the boundary of $\mathscr{A}_1(\Lambda)$. If every projected atom lies well within the antipodal convex hull, then it is possible to recover superpositions of atoms from Λ. The intuition is that the coefficient associated with an atom outside Λ must be quite large to represent anything in the span of the sub-dictionary. Figure 2 exhibits the geometry.

The Exact Recovery Coefficient first arose during the analysis of a greedy algorithm, Orthogonal Matching Pursuit (OMP). It was proved that $\mathrm{ERC}(\Lambda) > 0$ is both necessary and sufficient for OMP to recover every exact superposition of atoms from Λ. The same article demonstrated that this condition is sufficient for a convex relaxation method, Basis Pursuit (BP), to accomplish the same feat [Tro03a]. Independently, Fuchs exhibited a slightly weaker sufficient condition for BP in the real setting [Fuc02]. His work was extended to the complex setting in [Tro03b]. Gribonval and Nielsen have established a necessary and sufficient condition for BP to recover every signal over Λ, but they have provided no means for checking that it holds [GN03b, GN03a].

3.8. **Projective Spaces.** Projective spaces provide the correct setting for understanding many geometric properties of a dictionary. To develop this concept, we begin with an equivalence relation on the collection of d-dimensional complex vectors:

$$\boldsymbol{w} \equiv \boldsymbol{z} \qquad \Longleftrightarrow \qquad \boldsymbol{w} = \zeta \, \boldsymbol{z} \quad \text{for a nonzero complex number } \zeta \text{ in } \mathbb{C}^\times.$$

Under this equivalence relation, every nonzero vector is identified with the one-dimensional subspace spanned by that vector. The zero vector lies in a class by itself. For our purposes, the $(d-1)$-dimensional *complex projective space* will be defined as the collection of nonzero, d-dimensional complex vectors, modulo this equivalence relation:

$$\mathbb{P}^{d-1}(\mathbb{C}) \stackrel{\text{def}}{=} \frac{\mathbb{C}^d \setminus \{\mathbf{0}\}}{\mathbb{C}^\times}.$$

In words, $\mathbb{P}^{d-1}(\mathbb{C})$ is the set of one-dimensional subspaces of $\mathbb{C}^d$. The real projective space $\mathbb{P}^{d-1}(\mathbb{R})$ is defined in much the same way, and it may be viewed as the collection of all lines through the origin of $\mathbb{R}^d$. On analogy, we shall refer to the elements of a complex projective space as *lines*.

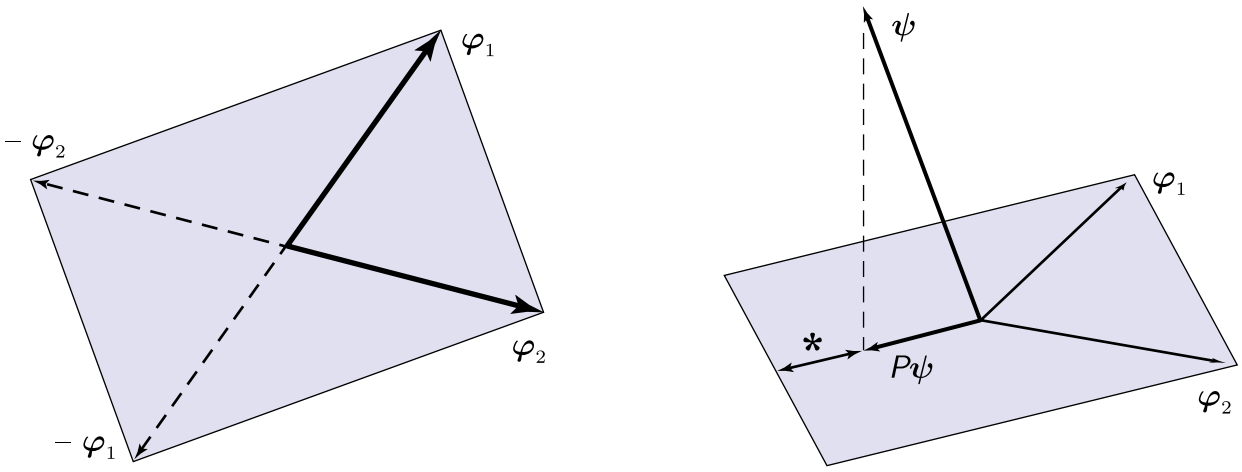

FIGURE 2. At left, we have shaded the antipodal convex hull of the two atoms $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$. At right, the asterisk ($*$) indicates the distance that the projection of the atom $\boldsymbol{\psi}$ lies from the edge of the antipodal convex hull (in signal space). The Exact Recovery Coefficient bounds the corresponding distance in coefficient space.

The natural metric for $\mathbb{P}^{d-1}(\mathbb{C})$ is the acute angle between two lines—or what is equivalent—the sine of the acute angle. Therefore, the projective distance between two d-dimensional vectors $\boldsymbol{z}$ and $\boldsymbol{w}$ will be calculated as

$$\operatorname{dist}(\boldsymbol{z}, \boldsymbol{w}) \quad \stackrel{\text{def}}{=} \quad \left[1 \; - \; \left(\frac{|\langle \boldsymbol{z}, \boldsymbol{w} \rangle|}{\|\boldsymbol{z}\|_2 \, \|\boldsymbol{w}\|_2}\right)^2\right]^{1/2}. \tag{3.5}$$

In particular, if both vectors have unit norm,

$$\operatorname{dist}(\boldsymbol{z}, \boldsymbol{w}) \quad = \quad \sqrt{1 \; - \; |\langle \boldsymbol{z}, \boldsymbol{w} \rangle|^2}.$$

Evidently, the distance between two lines ranges between zero and one. Equipped with this metric, $\mathbb{P}^{d-1}(\mathbb{C})$ forms a smooth, compact, Riemannian manifold [CHS96].

3.9. **Minimum Distance, Maximum Correlation.** We view the dictionary $\mathscr{D}$ as a finite set of lines in the projective space $\mathbb{P}^{d-1}(\mathbb{C})$. Given an arbitary nonzero signal $\boldsymbol{s}$, we shall calculate the *minimum distance* from the signal to the dictionary as

$$\min_{\omega \in \Omega} \operatorname{dist}(\boldsymbol{s}, \boldsymbol{\varphi}_\omega).$$

A complementary notion is the *maximum correlation* of the signal with the dictionary.

$$\operatorname{maxcor}(\boldsymbol{s}) \quad \stackrel{\text{def}}{=} \quad \max_{\omega \in \Omega} \frac{|\langle \boldsymbol{s}, \boldsymbol{\varphi}_\omega \rangle|}{\|\boldsymbol{s}\|_2} \quad = \quad \frac{\|\boldsymbol{\Phi}^* \boldsymbol{s}\|_\infty}{\|\boldsymbol{s}\|_2}.$$

Since the atoms are normalized, $0 \le \operatorname{maxcor}(\boldsymbol{s}) \le 1$. The relationship between the minimum distance and the maximum correlation is the following.

$$\min_{\omega \in \Omega} \operatorname{dist}(\boldsymbol{s}, \boldsymbol{\varphi}_\omega) \quad = \quad \sqrt{1 - \operatorname{maxcor}(\boldsymbol{s})^2}. \tag{3.6}$$

3.10. **Packing Radii.** We shall be interested in several extremal properties of the dictionary that are easiest to understand in a general setting. Let $\mathbb{X}$ be a compact metric space with metric $\mathrm{dist}_{\mathbb{X}}$, and choose $Y = \{y_k\}$ to be a discrete set of points in $\mathbb{X}$. The *packing radius* of the set Y is defined as

$$\mathrm{pack}_{\mathbb{X}}(Y) \quad \stackrel{\mathrm{def}}{=} \quad \min_{j \neq k} \mathrm{dist}_{\mathbb{X}}(y_j, y_k).$$

In words, the packing radius is the size of the largest open ball that can be centered at any point of Y without encompassing any other point of Y. An *optimal packing* of N points in $\mathbb{X}$ is a set Y_{opt} of cardinality N that has maximal packing radius, i.e., Y_{opt} solves the mathematical program

$$\max_{|Y| = N} \mathrm{pack}_{\mathbb{X}}(Y).$$

It is generally quite difficult to produce an optimal packing or even to check whether a collection of points gives an optimal packing. Figure 3 illustrates packing in the unit square, and Figure 4 examines the situation in a projective space. The standard reference on packing is the *magnum opus* of Conway and Sloane [CS98].

The packing radius of the dictionary in the projective space $\mathbb{P}^{d-1}(\mathbb{C})$ is given by

$$\mathrm{pack}(\mathscr{D}) \quad \stackrel{\mathrm{def}}{=} \quad \min_{\lambda \neq \omega} \mathrm{dist}(\boldsymbol{\varphi}_\lambda, \boldsymbol{\varphi}_\omega).$$

That is, the packing radius measures the minimum distance from the dictionary to itself. The coherence parameter is intimately related to this packing radius. Indeed,

$$\mu \quad = \quad \sqrt{1 - \mathrm{pack}(\mathscr{D})^2}.$$

Therefore, the dictionary provides a good packing if and only if the coherence is small. It is easily seen that the orthonormal bases for $\mathbb{C}^d$ give the only optimal packings of d points in $\mathbb{P}^{d-1}(\mathbb{C})$. For general d and N, it is quite difficult to construct minimally coherent dictionaries. The papers [CHS96, SH03, TDHS04] discuss the problem in detail.

The geometric interpretation of the cumulative coherence μ_1 is not as straightforward. One may imagine centering a collection of m open balls of non-decreasing radii at a fixed atom. The k-th ball is chosen so that it contains no more than $(k-1)$ other atoms. Roughly, the cumulative coherence $\mu_1(m)$ is complementary to the maximum total radius of a nested collection of m balls that can be centered at any atom and still retain this property.

3.11. **Covering Radii.** Let us return to our compact metric space $\mathbb{X}$, and let us select a finite set of points $Y = \{y_k\}$ from $\mathbb{X}$. The *covering radius* of the set Y is defined as

$$\mathrm{cover}_{\mathbb{X}}(Y) \quad \stackrel{\mathrm{def}}{=} \quad \max_{x \in \mathbb{X}} \min_{k} \mathrm{dist}_{\mathbb{X}}(x, y_k).$$

In words, the covering radius is the size of the largest open ball that can be centered at some point of $\mathbb{X}$ without encompassing a point of Y. An *optimal covering* with N points is a set Y_{opt} of cardinality N that has minimal covering radius. That is, Y_{opt} solves the mathematical program

$$\min_{|Y| = N} \mathrm{cover}_{\mathbb{X}}(Y).$$

N. J. A. Sloane has advanced the "meta-theorem" that optimal coverings are more regular than optimal packings [Slo02]. It is often extremely difficult to compute the covering radius of an ensemble of points, let alone to produce an optimal covering. See Figure 5 for an example of covering in the Euclidean unit square; Figure 6 demonstrates covering of a projective space. Conway and Sloane's book is also an important reference on covering [CS98].

In our projective space $\mathbb{P}^{d-1}(\mathbb{C})$, the covering radius of the dictionary is given by the formula

$$\mathrm{cover}(\mathscr{D}) \quad \stackrel{\mathrm{def}}{=} \quad \max_{\boldsymbol{s} \neq \mathbf{0}} \min_{\omega \in \Omega} \mathrm{dist}(\boldsymbol{s}, \boldsymbol{\varphi}_\omega).$$

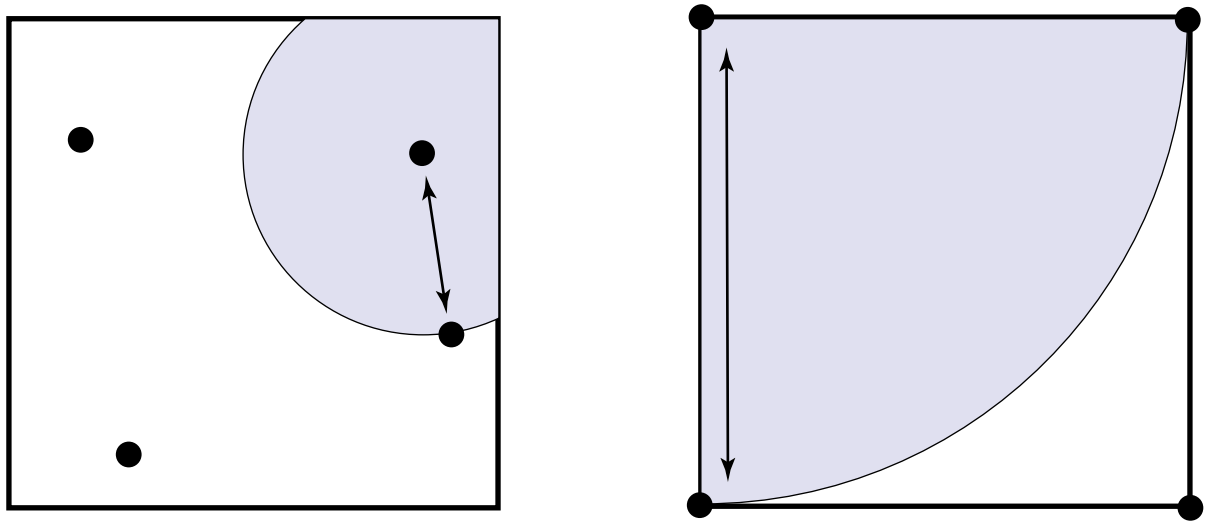

FIGURE 3. At left, the arrow indicates the packing radius of four points in the Euclidean unit square. At right, an *optimal* packing of four points in the unit square.

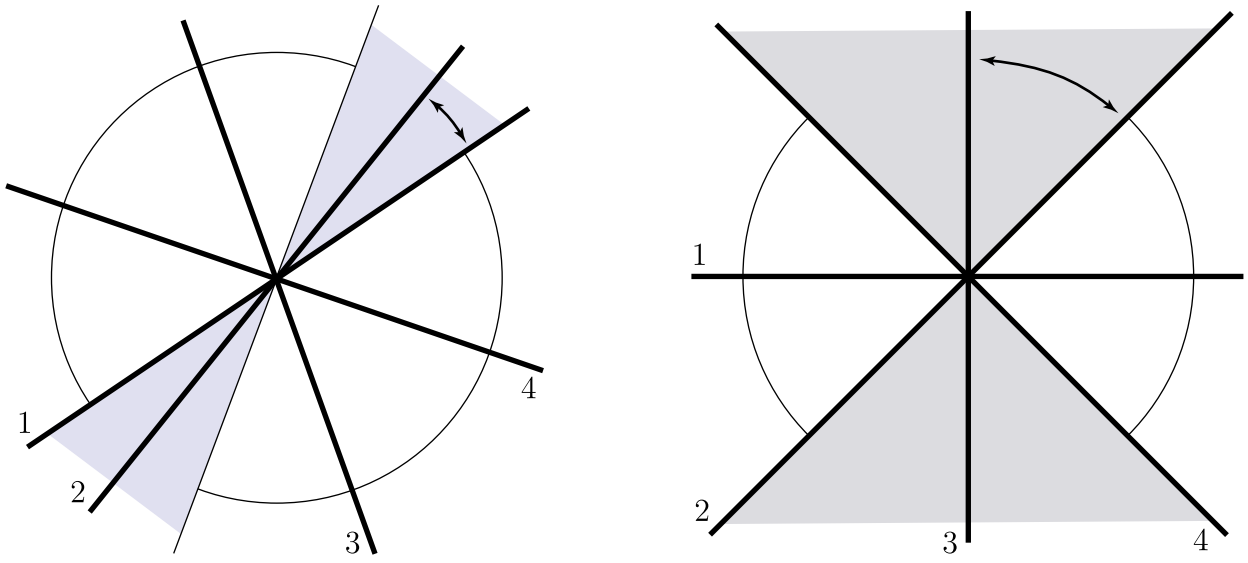

FIGURE 4. At left, the arrow indicates the packing radius of a collection of numbered lines, considered as elements of $\mathbb{P}^1(\mathbb{R})$. Note that, in this projective space, a "ball" becomes a doubly-infinite cone. At right, an *optimal* packing of four lines in $\mathbb{P}^1(\mathbb{R})$.

It follows from the relation (3.6) that the covering radius is attained at a signal (called a *deep hole*) whose maximum correlation with the dictionary is smallest:

$$\operatorname{cover}(\mathscr{D}) \quad = \quad \max_{\boldsymbol{s} \neq \mathbf{0}} \sqrt{1 - \operatorname{maxcor}(\boldsymbol{s})^2}.$$

We shall be most interested in how well a sub-dictionary covers its span. To that end, define

$$\operatorname{cover}(\Lambda; \mathscr{D}) \quad \stackrel{\text{def}}{=} \quad \max_{\substack{\boldsymbol{s} \in \mathscr{R}(\boldsymbol{\Phi}_\Lambda) \\ \boldsymbol{s} \neq \mathbf{0}}} \min_{\lambda \in \Lambda} \operatorname{dist}(\boldsymbol{s}, \boldsymbol{\varphi}_\lambda). \tag{3.7}$$

Without the range restriction on $\boldsymbol{s}$, the covering radius of the sub-dictionary would be one unless the atoms in Λ spanned the entire signal space.

Proposition 3.8. *The covering radius of a sub-dictionary satisfies the identity*

$$\operatorname{cover}(\Lambda)^2 \quad = \quad 1 \; - \; \left\| \boldsymbol{\Phi}_\Lambda{}^+ \right\|_{2,1}^{-2}.$$

Proof. Begin with (3.7), and apply the definition (3.5) of projective distance to see that

$$\operatorname{cover}(\Lambda)^2 \quad = \quad 1 \; - \; \min_{\substack{\boldsymbol{s} \in \mathscr{R}(\boldsymbol{\Phi}_\Lambda) \\ \boldsymbol{s} \neq \mathbf{0}}} \max_{\lambda \in \Lambda} \frac{|\langle \boldsymbol{s}, \boldsymbol{\varphi}_\lambda \rangle|^2}{\|\boldsymbol{s}\|_2^2} \quad = \quad 1 \; - \; \min_{\substack{\boldsymbol{s} \in \mathscr{R}(\boldsymbol{\Phi}_\Lambda) \\ \boldsymbol{s} \neq \mathbf{0}}} \frac{\|\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{s}\|_\infty^2}{\|\boldsymbol{s}\|_2^2}.$$

Since the atoms indexed by Λ form a linearly independent set, $\boldsymbol{\Phi}_\Lambda{}^*$ is surjective. It follows from Proposition 3.2 that the minimum equals $\left\| (\boldsymbol{\Phi}_\Lambda{}^+)^* \right\|_{\infty,2}^{-2}$. Apply the identity (3.1) to switch from the $(\infty, 2)$ norm to the $(2, 1)$ norm. □

One interpretation of this proposition is that $\left\| \boldsymbol{\Phi}_\Lambda{}^+ \right\|_{2,1}$ gives the secant of the largest acute angle between a vector in the span of the sub-dictionary and the closest atom from the sub-dictionary. We also learn that calculating $\operatorname{cover}(\Lambda)$ is likely to be NP-hard.

Using the cumulative coherence function, we can develop reasonable estimates for $\operatorname{cover}(\Lambda)$. This result will show that sub-dictionaries of incoherent dictionaries form good coverings.

Proposition 3.9. *Suppose that Λ lists m linearly independent atoms and that $\mu_1(m-1) < 1$. Then*

$$\begin{aligned} \left\| \boldsymbol{\Phi}_\Lambda{}^+ \right\|_{2,1} \quad &\leq \quad \left[\frac{m}{1 \; - \; \mu_1(m-1)} \right]^{1/2}, \qquad \textit{and therefore} \\ \operatorname{cover}(\Lambda) \quad &\geq \quad \left[1 \; - \; \frac{1 \; - \; \mu_1(m-1)}{m} \right]^{1/2}. \end{aligned}$$

Proof. Write the definition of the operator norm, and estimate the ℓ_1 norm with the ℓ_2 norm:

$$\left\| \boldsymbol{\Phi}_\Lambda{}^+ \right\|_{2,1} \quad = \quad \max_{\|\boldsymbol{s}\|_2 = 1} \left\| \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{s} \right\|_1 \quad \leq \quad \sqrt{m} \max_{\|\boldsymbol{s}\|_2 = 1} \left\| \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{s} \right\|_2 \quad = \quad \sqrt{m} \left\| \boldsymbol{\Phi}_\Lambda{}^+ \right\|_{2,2}.$$

The $(2, 2)$ operator norm of $\boldsymbol{\Phi}_\Lambda{}^+$ is the reciprocal of the minimum singular value of $\boldsymbol{\Phi}_\Lambda$. To complete the proof, apply the lower bound on this singular value given by Proposition 3.3. □

A second consequence of the argument is that the covering radius of m vectors strictly exceeds $\sqrt{1 - 1/m}$ unless the vectors are orthonormal. It follows that orthonormal bases give the only optimal coverings of $\mathbb{P}^{d-1}(\mathbb{C})$ using d points. The result also provides an intuition why balls in infinite-dimensional Hilbert spaces cannot be compact: a collection of m vectors must cover its span worse and worse as m increases. We can develop a second version of Proposition 3.9 by estimating the minimum singular value with Proposition 3.5.

Proposition 3.10. *Assume that $N \leq d$, and suppose that Λ lists m atoms. Then*

$$\left\| \boldsymbol{\Phi}_\Lambda{}^+ \right\|_{2,1} \quad \leq \quad \sqrt{m} \; / \; \sigma_{\min}(\boldsymbol{\Phi}).$$

The material here on covering projective spaces seems to be novel.

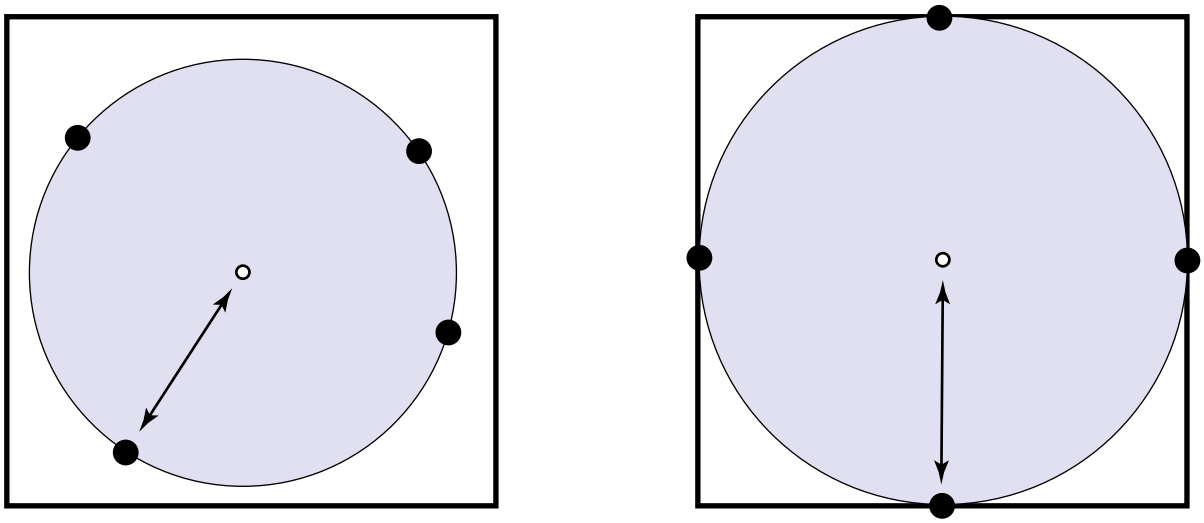

FIGURE 5. At left, the arrow indicates the covering radius of four points in the Euclidean unit square, and the open circle marks the point at which the covering radius is attained. At right, an *optimal* covering of the unit square with four points.

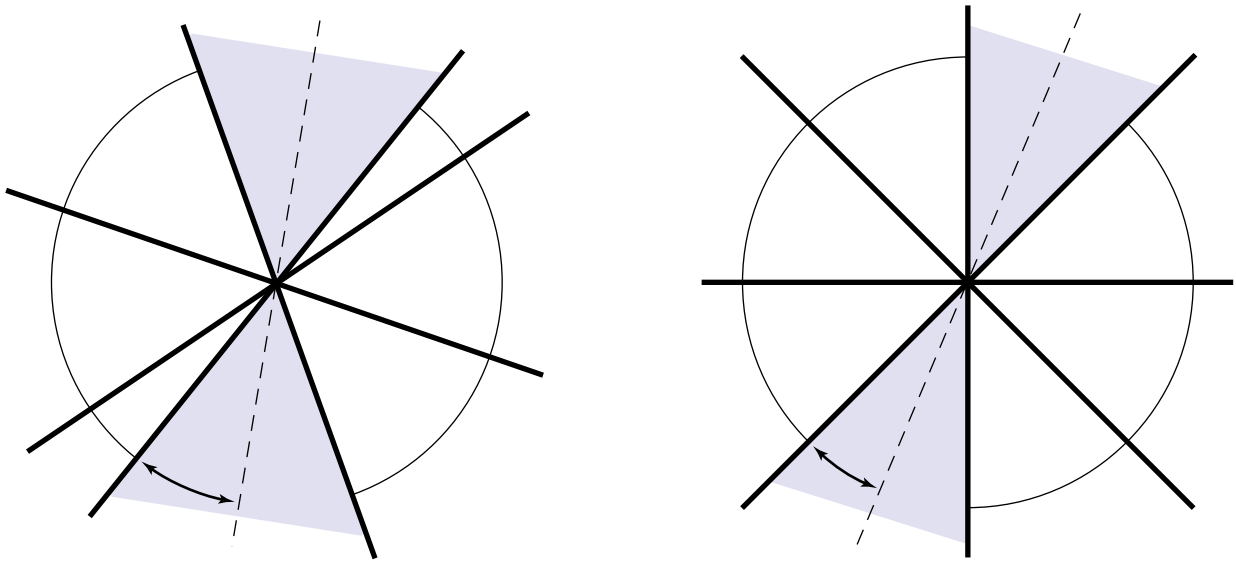

FIGURE 6. At left, the covering radius of four lines in the real projective space $\mathbb{P}^1(\mathbb{R})$. Dashes mark the line at which the covering radius is attained. At right, an *optimal* covering of $\mathbb{P}^1(\mathbb{R})$ with four lines.

3.12. **Quantization.** Now let us consider a probability space $(\mathbb{X}, \Sigma, \mathrm{d}x)$ in which metric balls are Σ-measurable, and choose another finite set of points $Y = \{y_k\}$. In this setting, $\min_k \operatorname{dist}(x, y_k)$ is called the *quantization error* for the point x. The expected quantization error is the usual measure of how well Y represents the distribution $\mathrm{d}x$. It is computed with the integral

$$\operatorname{quant}(Y) \quad \stackrel{\text{def}}{=} \quad \int \min_k \operatorname{dist}_{\mathbb{X}}(x, y_k) \, \mathrm{d}x.$$

The relationship $\operatorname{quant}(Y) \leq \operatorname{cover}(Y)$ is always in force. An N-point *optimal codebook* for quantizing $\mathrm{d}x$ is a set that solves the mathematical program

$$\min_{|Y| = N} \operatorname{quant}(Y).$$

Optimal quantization is a difficult problem, and it has been studied extensively [GG92]. Heuristic methods are available for constructing good codebooks [Llo57, Csi84, Ros98].

Suppose that we define a probability measure $\mathrm{d}\boldsymbol{\nu}$ on the projective space $\mathbb{P}^{d-1}(\mathbb{C})$. The expected error in quantizing $\mathrm{d}\boldsymbol{\nu}$ with the dictionary is defined as

$$\operatorname{quant}(\mathscr{D}) \quad \stackrel{\text{def}}{=} \quad \int \min_{\omega} \operatorname{dist}(\boldsymbol{\nu}, \boldsymbol{\varphi}_\omega) \, \mathrm{d}\boldsymbol{\nu} \quad = \quad \int \sqrt{1 - \operatorname{maxcor}(\boldsymbol{\nu})^2} \, \mathrm{d}\boldsymbol{\nu}.$$

Now imagine that we are trying to recover a short linear combination of atoms that has been contaminated with additive noise whose *direction* is distributed according to $\mathrm{d}\boldsymbol{\nu}$. In this situation, the best dictionaries for sparse approximation do a *horrible* job quantizing the direction of the noise. As a result, it is highly likely that the signal can be recovered, even if the noise has significant magnitude.

4. Fundamental Lemmata

When solving the convex relaxation of a sparse approximation problem, we immediately encounter the objective function

$$L(\boldsymbol{b}; \gamma, \boldsymbol{s}) \quad \stackrel{\text{def}}{=} \quad \tfrac{1}{2} \, \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}\|_2^2 \, + \, \gamma \, \|\boldsymbol{b}\|_1 . \tag{4.1}$$

In this section, we shall reach a detailed understanding of the minimizers of this function. With these results at hand, the major theorems of this paper will follow easily.

Throughout this section, we shall use the following notations. Fix the input signal $\boldsymbol{s}$ and the parameter γ so that the function L depends only on the coefficient vector $\boldsymbol{b}$. Let Λ index a subdictionary, which will be used to approximate the input signal. The best approximation of the input signal over the atoms in Λ can be written as $\boldsymbol{a}_\Lambda = P_\Lambda \, \boldsymbol{s}$. The coefficient vector $\boldsymbol{c}_\Lambda = {\boldsymbol{\Phi}_\Lambda}^+ \boldsymbol{s}$ may be used to synthesize $\boldsymbol{a}_\Lambda$. Although $\boldsymbol{c}_\Lambda$ lies in $\mathbb{C}^\Lambda$, we often extend it to $\mathbb{C}^\Omega$ by padding it with zeros.

4.1. **The Correlation Condition Lemma.** Suppose that the atoms outside Λ have small inner products (i.e. are weakly correlated) with the residual signal $(\boldsymbol{s} - \boldsymbol{a}_\Lambda)$. The following lemma shows that any coefficient vector which minimizes the objective function (4.1) must be supported inside Λ. This result is the soul of the paper.

Lemma 4.1 (Correlation Condition)**.** *Suppose that the maximum inner product between the residual signal and any atom fulfills the condition*

$$\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty \quad \leq \quad \gamma \, \operatorname{ERC}(\Lambda). \tag{4.2}$$

Then any coefficient vector $\boldsymbol{b}_\star$ that minimizes the function (4.1) *must satisfy* $\operatorname{supp}(\boldsymbol{b}_\star) \subset \Lambda$.

Some remarks and corollaries are in order. First, observe that the lemma is worthless unless $\mathrm{ERC}(\Lambda)$ is positive. Return to Section 3.7 for a geometric description of this condition. Next, if the index set does satisfy $\mathrm{ERC}(\Lambda) > 0$, the sufficient condition (4.2) will always hold if γ is large enough. But if γ is too large, then $\boldsymbol{b}_\star = \mathbf{0}$. Third, the lemma is indifferent to the choice of Λ, so long as the inequality (4.2) holds for the residual $(\boldsymbol{s} - \boldsymbol{a}_\Lambda)$. Therefore, the support of $\boldsymbol{b}_\star$ is actually contained in the intersection of all such index sets. Fourth, we write the left-hand side of (4.2) as

$$\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty \quad = \quad \mathrm{maxcor}(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\ \|\boldsymbol{s} - \boldsymbol{a}_\Lambda\|_2 \,.$$

It follows that the result is strongest when the magnitude of the residual and its maximum correlation with the dictionary are both small. Since the maximum correlation never exceeds one, we obtain a (much weaker) result that depends only on the magnitude of the residual.

Corollary 4.2. *Suppose that the approximation error satisfies*

$$\|\boldsymbol{s} - \boldsymbol{a}_\Lambda\|_2 \quad \leq \quad \gamma\ \mathrm{ERC}(\Lambda).$$

Then any coefficient vector $\boldsymbol{b}_\star$ that minimizes the function (4.1) *must be supported inside Λ.*

By normalizing the input signal, we may also obtain a result that depends only on the maximum correlation.

Corollary 4.3. *Suppose that the maximum correlation between the residual signal and the dictionary satisfies the bound*

$$\mathrm{maxcor}(\boldsymbol{s} - \boldsymbol{a}_\Lambda) \quad \leq \quad \gamma\ \mathrm{ERC}(\Lambda).$$

If we scale the input signal to have unit norm and then compute a minimizer $\boldsymbol{b}_\star$ of the function (4.1)*, it follows that $\boldsymbol{b}_\star$ must be supported inside Λ.*

If the input signal can be expressed as an exact superposition of the atoms in the sub-dictionary, we reach another interesting corollary.

Corollary 4.4. *Suppose that the input signal has an exact representation using the atoms in Λ and that $\mathrm{ERC}(\Lambda) > 0$. For every positive γ, every coefficient vector that minimizes* (4.1) *must be supported inside Λ.*

One might wonder whether the condition $\mathrm{ERC}(\Lambda) > 0$ is really necessary to prove these results. The answer is a qualified affirmative. Section 4.4 offers a partial converse of Corollary 4.4.

4.2. **Proof of Correlation Condition Lemma.** The proof requires a theorem that has appeared in recent literature.

Theorem 4.5 (Fuchs [Fuc02], Tropp [Tro03a])**.** *Suppose that $\mathrm{ERC}(\Lambda) > 0$ and that the signal $\boldsymbol{s}$ is a linear combination of the atoms indexed by Λ. It follows that there is a unique coefficient vector supported on Λ that synthesizes the signal. Moreover, this vector is the unique solution of the convex program*

$$\min_{\boldsymbol{b}} \quad \|\boldsymbol{b}\|_1 \qquad \textit{subject to} \qquad \boldsymbol{\Phi}\,\boldsymbol{b} = \boldsymbol{s}.$$

A simple but powerful geometric idea underlies the proof of the lemma. The expense of using indices outside Λ is not compensated by the improvement in the approximation error. Therefore, any approximation involving other atoms can be projected onto the atoms in Λ to reduce the value of the objective function.

Proof of Lemma 4.1. We shall be studying minimizers of the objective function

$$L(\boldsymbol{b}) \quad \stackrel{\mathrm{def}}{=} \quad \tfrac{1}{2}\ \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2^2 \ +\ \gamma\ \|\boldsymbol{b}\|_1 \,. \tag{4.3}$$

Assume that there is a coefficient vector $\boldsymbol{b}_\star$ which minimizes (4.3) even though $\boldsymbol{b}_\star$ uses an index outside of Λ. That is, $\mathrm{supp}(\boldsymbol{b}_\star) \setminus \Lambda$ is non-empty. We shall compare $\boldsymbol{b}_\star$ against the projected

coefficient vector $\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star$, which is supported on Λ. Since $\boldsymbol{b}_\star$ minimizes the objective function, the inequality $L(\boldsymbol{b}_\star) \le L(\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star)$ must hold. Rearranging the terms of this relation gives

$$2\,\gamma\,\left[\,\|\boldsymbol{b}_\star\|_1 \;-\; \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star\right\|_1\,\right] \quad\le\quad \|\boldsymbol{s} - \boldsymbol{P}_\Lambda \boldsymbol{\Phi}\, \boldsymbol{b}_\star\|_2^2 \;-\; \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}_\star\|_2^2 . \tag{4.4}$$

We shall provide a lower bound on the left-hand side of (4.4) and an upper bound on the right-hand side. The correlation condition (4.2) will reverse the consequent inequality.

First, we claim that $\boldsymbol{\Phi}\, \boldsymbol{b}_\star$ does not lie in the range of $\boldsymbol{\Phi}_\Lambda$. Suppose that it did. Then $\boldsymbol{b}_\star$ and $\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star$ synthesize the same signal, and so the right-hand side of (4.4) equals zero. At the same time, Theorem 4.5 shows that the coefficient vector $\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star$ is the unique solution of the convex program

$$\min_{\boldsymbol{b}} \quad \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \boldsymbol{\Phi}\, \boldsymbol{b} = \boldsymbol{\Phi}\, \boldsymbol{b}_\star .$$

In consequence, the left-hand side of (4.4) is positive. This combination of events is impossible.

Now we shall develop a lower bound on the left-hand side of (4.4). To accomplish this, let us split the coefficient vector into two parts: $\boldsymbol{b}_\star = \boldsymbol{b}_\Lambda + \boldsymbol{b}_{\text{bad}}$. The first vector $\boldsymbol{b}_\Lambda$ contains the components with indices in Λ. The second vector contains the (undesirable) remaining components—those from $\Omega \setminus \Lambda$. This splitting yields the identity

$$\|\boldsymbol{b}_\star\|_1 \;-\; \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star\right\|_1 \quad=\quad \|\boldsymbol{b}_\Lambda\|_1 \;+\; \|\boldsymbol{b}_{\text{bad}}\|_1 \;-\; \left\|\boldsymbol{b}_\Lambda + \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_{\text{bad}}\right\|_1 .$$

The upper triangle inequality allows us to cancel the terms involving $\boldsymbol{b}_\Lambda$:

$$\|\boldsymbol{b}_\star\|_1 \;-\; \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star\right\|_1 \quad\ge\quad \|\boldsymbol{b}_{\text{bad}}\|_1 \;-\; \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_{\text{bad}}\right\|_1 . \tag{4.5}$$

Let us focus on the second term from the right-hand side of (4.5). To be rigorous, we temporarily define a diagonal matrix $\boldsymbol{R}_{\text{bad}}$ that zeros the components of a coefficient vector indexed by Λ and ignores the rest. Then

$$\begin{aligned} \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_{\text{bad}}\right\|_1 &= \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi} \boldsymbol{R}_{\text{bad}}\, \boldsymbol{b}_{\text{bad}}\right\|_1 \\ &\le \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi} \boldsymbol{R}_{\text{bad}}\right\|_{1,1} \|\boldsymbol{b}_{\text{bad}}\|_1 \\ &= \max_{\omega \notin \Lambda} \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega\right\|_1 \|\boldsymbol{b}_{\text{bad}}\|_1 . \end{aligned}$$

Re-introducing this expression into (4.5) gives

$$\|\boldsymbol{b}_\star\|_1 \;-\; \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star\right\|_1 \quad\ge\quad \left[1 \;-\; \max_{\omega \notin \Lambda} \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega\right\|_1\right] \|\boldsymbol{b}_{\text{bad}}\|_1 .$$

We identify the bracketed quantity as the Exact Recovery Coefficient of Λ to reach the lower bound

$$\|\boldsymbol{b}_\star\|_1 \;-\; \left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\Phi}\, \boldsymbol{b}_\star\right\|_1 \quad\ge\quad \text{ERC}(\Lambda)\, \|\boldsymbol{b}_{\text{bad}}\|_1 . \tag{4.6}$$

Next, we need to provide an upper bound on the right-hand side of (4.4). Apply the Law of Cosines to the triangle formed by the signals $\boldsymbol{s}$ and $\boldsymbol{\Phi}\, \boldsymbol{b}_\star$ and $\boldsymbol{P}_\Lambda \boldsymbol{\Phi}\, \boldsymbol{b}_\star$ to obtain the identity

$$\begin{aligned} \|\boldsymbol{s} - \boldsymbol{P}_\Lambda \boldsymbol{\Phi}\, \boldsymbol{b}_\star\|_2^2 \;-\; \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}_\star\|_2^2 \quad= \\ 2\,\text{Re}\,\langle (\mathsf{I}_d - \boldsymbol{P}_\Lambda) \boldsymbol{\Phi}\, \boldsymbol{b}_\star, \boldsymbol{s} - \boldsymbol{P}_\Lambda \boldsymbol{\Phi}\, \boldsymbol{b}_\star \rangle \;-\; \|(\mathsf{I}_d - \boldsymbol{P}_\Lambda) \boldsymbol{\Phi}\, \boldsymbol{b}_\star\|_2^2 . \end{aligned} \tag{4.7}$$

Alternately, one might expand the squared norms on the right-hand side of (4.4) and simplify the consequent monstrosity. Since the matrix $(\mathsf{I}_d - \boldsymbol{P}_\Lambda)$ annihilates the atoms listed by Λ, it follows that

$$(\mathsf{I}_d - \boldsymbol{P}_\Lambda) \boldsymbol{\Phi}\, \boldsymbol{b}_\star \quad=\quad (\mathsf{I}_d - \boldsymbol{P}_\Lambda) \boldsymbol{\Phi}\, \boldsymbol{b}_{\text{bad}} . \tag{4.8}$$

Using (4.8) and the fact that $(\mathsf{I}_d - P_\Lambda)$ is an orthogonal projector, we may manipulate the inner product in (4.7):

$$\begin{aligned}\langle (\mathsf{I}_d - P_\Lambda)\boldsymbol{\Phi}\,\boldsymbol{b}_\star, \boldsymbol{s} - P_\Lambda \boldsymbol{\Phi}\,\boldsymbol{b}_\star\rangle &= \langle (\mathsf{I}_d - P_\Lambda)\boldsymbol{\Phi}\,\boldsymbol{b}_{\mathrm{bad}}, \boldsymbol{s} - P_\Lambda \boldsymbol{\Phi}\,\boldsymbol{b}_\star\rangle \\ &= \langle \boldsymbol{b}_{\mathrm{bad}}, \boldsymbol{\Phi}^*(\mathsf{I}_d - P_\Lambda)(\boldsymbol{s} - P_\Lambda \boldsymbol{\Phi}\,\boldsymbol{b}_\star)\rangle \\ &= \langle \boldsymbol{b}_{\mathrm{bad}}, \boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\rangle\,.\end{aligned}$$

Substitute (4.8) and the expression for the inner product into equation (4.7) to obtain the identity

$$\|\boldsymbol{s} - P_\Lambda \boldsymbol{\Phi}\,\boldsymbol{b}_\star\|_2^2 \;-\; \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}_\star\|_2^2 \quad = \quad 2\,\mathrm{Re}\,\langle \boldsymbol{b}_{\mathrm{bad}}, \boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\rangle \;-\; \|(\mathsf{I}_d - P_\Lambda)\boldsymbol{\Phi}\,\boldsymbol{b}_{\mathrm{bad}}\|_2^2\,.$$

Since $\boldsymbol{\Phi}\,\boldsymbol{b}_\star$ does not lie in the range of $\boldsymbol{\Phi}_\Lambda$, neither does $\boldsymbol{\Phi}\,\boldsymbol{b}_{\mathrm{bad}}$. So the second term on the right-hand side is strictly positive. Nevertheless, this term is negligible in comparison with the first term whenever $\boldsymbol{b}_{\mathrm{bad}}$ is small. Therefore, we simply discard the second term, and we apply Hölder's Inequality to reach the upper bound

$$\|\boldsymbol{s} - P_\Lambda \boldsymbol{\Phi}\,\boldsymbol{b}_\star\|_2^2 \;-\; \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}_\star\|_2^2 \quad < \quad 2\,\|\boldsymbol{b}_{\mathrm{bad}}\|_1\,\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty\,. \tag{4.9}$$

To proceed, we combine the bounds (4.6) and (4.9) into the inequality (4.4) to discover that

$$\gamma\;\mathrm{ERC}(\Lambda)\;\|\boldsymbol{b}_{\mathrm{bad}}\|_1 \quad < \quad \|\boldsymbol{b}_{\mathrm{bad}}\|_1\;\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty\,.$$

We have assumed that the support of the coefficient vector $\boldsymbol{b}_\star$ contains at least one index outside Λ, so the vector $\boldsymbol{b}_{\mathrm{bad}}$ cannot be null. Therefore, we may divide the most recent inequality by $\|\boldsymbol{b}_{\mathrm{bad}}\|_1$ to conclude that

$$\gamma\;\mathrm{ERC}(\Lambda) \quad < \quad \|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty\,.$$

If this inequality fails, then we must discard our hypothesis that the minimizer $\boldsymbol{b}_\star$ involves an index outside Λ. □

4.3. **Restricted Minimizers.** The sufficient condition of Lemma 4.1 guarantees that the minimizer of (4.1) is supported on the index set Λ. Unfortunately, this does not even rule out the zero vector as a possible minimizer. Therefore, we must develop bounds on how much the minimizing coefficient vector can vary from the desired coefficient vector $\boldsymbol{c}_\Lambda$.

The argument involves standard techniques from convex analysis, but the complex setting demands an artifice. In this subsection only, we shall decompose complex vectors into independent real and imaginary parts, i.e. $\boldsymbol{z} = \boldsymbol{x} + \mathrm{i}\,\boldsymbol{y}$. Then we may define the *complex gradient* of a real-valued function $f : \mathbb{C}^m \to \mathbb{R}$ as the vector

$$\nabla f \quad \stackrel{\mathrm{def}}{=} \quad \nabla_{\boldsymbol{x}} f \;+\; \mathrm{i}\,\nabla_{\boldsymbol{y}} f.$$

Here, $\nabla_{\boldsymbol{x}}$ indicates the (real) derivative of f taken with respect to the real variables while fixing the imaginary variables; the definition of $\nabla_{\boldsymbol{y}}$ is similar. If f does not depend on the imaginary variables, then ∇f reduces to the usual real gradient. One may wish to read the article [vdB94] for a more elegant treatment of complex gradients.

In the same spirit, the *complex subdifferential* of a convex function $f : \mathbb{C}^m \to \mathbb{R}$ at a complex vector $\boldsymbol{z}$ may be defined as

$$\partial f(\boldsymbol{z}) \quad \stackrel{\mathrm{def}}{=} \quad \{\; \boldsymbol{g} \in \mathbb{C}^m \;:\; f(\boldsymbol{w}) \geq f(\boldsymbol{z}) + \mathrm{Re}\,\langle \boldsymbol{g}, \boldsymbol{w} - \boldsymbol{z}\rangle \quad \text{for all } \boldsymbol{w} \in \mathbb{C}^m \;\}.$$

The vectors contained in the subdifferential are called *subgradients*, and they provide affine lower bounds on the function. If the function has a complex gradient at a point, the complex gradient gives the unique subgradient there. In addition, the complex subdifferential is additive, viz. $\partial(f_1 + f_2)(\boldsymbol{z}) = \partial f_1(\boldsymbol{z}) + \partial f_2(\boldsymbol{z})$. It is straightforward (but tedious) to verify that the complex subdifferential satisfies all the properties of real subdifferentials [Roc70].

Lemma 4.6. *Suppose that the vector $\boldsymbol{b}_\star$ minimizes the objective function* (4.1) *over all coefficient vectors supported on Λ. A necessary and sufficient condition on such a minimizer is that*

$$\boldsymbol{c}_\Lambda - \boldsymbol{b}_\star \quad = \quad \gamma \, (\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1} \boldsymbol{g} \tag{4.10}$$

where the vector $\boldsymbol{g}$ is drawn from $\partial \|\boldsymbol{b}_\star\|_1$. Moreover, the minimizer is unique.

Fuchs developed this necessary and sufficient condition in the real setting using essentially the same method [Fuc02].

Proof. Apply the Pythagorean Theorem to (4.1) to see that minimizing L over coefficient vectors supported on Λ is equivalent to minimizing the function

$$F(\boldsymbol{b}) \quad \stackrel{\text{def}}{=} \quad \tfrac{1}{2} \, \|\boldsymbol{a}_\Lambda - \boldsymbol{\Phi}_\Lambda \, \boldsymbol{b}\|_2^2 \, + \, \gamma \, \|\boldsymbol{b}\|_1 \tag{4.11}$$

over coefficient vectors from $\mathbb{C}^\Lambda$. Recall that the atoms indexed by Λ form a linearly independent collection, so $\boldsymbol{\Phi}_\Lambda$ has full column rank. It follows that the quadratic term in (4.11) is strictly convex, and so the whole function F must also be strictly convex. Therefore, its minimizer is unique.

The function F is convex and unconstrained, so $\mathbf{0} \in \partial F(\boldsymbol{b}_\star)$ is a necessary and sufficient condition for the coefficient vector $\boldsymbol{b}_\star$ to minimize F. The complex gradient of the first term of F equals $(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda) \, \boldsymbol{b}_\star - \boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{a}_\Lambda$. From the additivity of subdifferentials, it follows that

$$(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda) \, \boldsymbol{b}_\star \, - \, \boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{a}_\Lambda \, + \, \gamma \, \boldsymbol{g} \quad = \quad \mathbf{0}$$

for some vector $\boldsymbol{g}$ drawn from the subdifferential $\partial \|\boldsymbol{b}_\star\|_1$. Since the atoms indexed by Λ are linearly independent, we may pre-multiply this relation by $(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1}$ to reach

$$\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{a}_\Lambda - \boldsymbol{b}_\star \quad = \quad \gamma \, (\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1} \boldsymbol{g}.$$

Apply the fact that $\boldsymbol{c}_\Lambda = \boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{a}_\Lambda$ to reach the conclusion. □

Now we identify the subdifferential of the ℓ_1 norm. To that end, define the signum function as

$$\operatorname{sgn}(r \, \mathrm{e}^{\mathrm{i}\,\theta}) \quad \stackrel{\text{def}}{=} \quad \begin{cases} \mathrm{e}^{\mathrm{i}\,\theta} & \text{for } r > 0 \\ 0 & \text{for } r = 0. \end{cases}$$

One may extend the signum function to vectors by applying it to each component.

Proposition 4.7. *Let $\boldsymbol{z}$ be a complex vector. The complex vector $\boldsymbol{g}$ lies in the complex subdifferential $\partial \|\boldsymbol{z}\|_1$ if and only if*

- $|g_k| \le 1$ *whenever* $z_k = 0$, *and*
- $g_k = \operatorname{sgn} z_k$ *whenever* $z_k \neq 0$.

In particular, $\|\boldsymbol{g}\|_\infty = 1$ unless $\boldsymbol{z} = \mathbf{0}$, in which case $\|\boldsymbol{g}\|_\infty \le 1$.

We omit the proof. At last, we may develop bounds on how much a solution to the restricted problem varies from the desired solution $\boldsymbol{c}_\Lambda$.

Corollary 4.8 (Upper Bounds)**.** *Suppose that the vector $\boldsymbol{b}_\star$ minimizes the function* (4.1) *over all coefficient vectors supported on Λ. The following bounds are in force:*

$$\|\boldsymbol{c}_\Lambda - \boldsymbol{b}_\star\|_\infty \quad \le \quad \gamma \, \left\|(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty} \tag{4.12}$$

$$\|\boldsymbol{\Phi}_\Lambda \, (\boldsymbol{c}_\Lambda - \boldsymbol{b}_\star)\|_2 \quad \le \quad \gamma \, \left\|\boldsymbol{\Phi}_\Lambda{}^+\right\|_{2,1}. \tag{4.13}$$

Proof. We begin with the necessary and sufficient condition

$$\boldsymbol{c}_\Lambda - \boldsymbol{b}_\star \quad = \quad \gamma \, (\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1} \boldsymbol{g} \tag{4.14}$$

where $\boldsymbol{g} \in \partial \|\boldsymbol{b}_\star\|_1$. To obtain (4.12), we take the ℓ_∞ norm of (4.14) and apply the usual estimate:

$$\|\boldsymbol{b}_\star - \boldsymbol{c}_\Lambda\|_\infty \quad = \quad \gamma \, \left\|(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1} \boldsymbol{g}\right\|_\infty \quad \le \quad \gamma \, \left\|(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty} \|\boldsymbol{g}\|_\infty.$$

Proposition 4.7 shows that $\|\boldsymbol{g}\|_\infty \le 1$, which proves the result.

To develop the second bound (4.13), we pre-multiply (4.14) by the matrix $\boldsymbol{\Phi}_\Lambda$ and compute the Euclidean norm:

$$\|\boldsymbol{\Phi}_\Lambda(\boldsymbol{b}_\star - \boldsymbol{c}_\Lambda)\|_2 \quad = \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^+)^*\boldsymbol{g}\right\|_2 \quad \leq \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^+)^*\right\|_{\infty,2} \|\boldsymbol{g}\|_\infty .$$

As before, $\|\boldsymbol{g}\|_\infty \leq 1$. Finally, we apply the identity (3.1) to switch from the $(\infty, 2)$ operator norm to the $(2, 1)$ operator norm. □

For the record, we present a lower bound.

Corollary 4.9 (Lower Bounds)**.** *Suppose that the vector $\boldsymbol{b}_\star$ minimizes the function* (4.1) *over all coefficient vectors supported on Λ. For every index λ in* $\operatorname{supp}(\boldsymbol{b}_\star)$*,*

$$|\boldsymbol{c}_{\text{opt}}(\lambda) \ - \ \boldsymbol{b}_\star(\lambda)| \quad \geq \quad \gamma \left[2 \ - \ \left\|(\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty}\right].$$

We shall not use this result, so we leave its proof as an exercise for the reader.

4.4. **Is the ERC Necessary?** Let Λ index a sub-dictionary for which $\text{ERC}(\Lambda) > 0$, and suppose that the input signal can be written as an exact superposition of atoms from Λ. It follows from the Correlation Condition Lemma and from Corollary 4.8 that, for all sufficiently small γ, the minimizer of the function (4.1) has support equal to Λ. The following theorem shows that this type of result cannot hold if $\text{ERC}(\Lambda) < 0$.

Theorem 4.10. *Suppose that* $\text{ERC}(\Lambda) < 0$*. Then we may construct an input signal that has an exact representation using the atoms in Λ and yet the minimizer of the function* (4.1) *is* not *supported on Λ when γ is small.*

It is known that some dictionaries (e.g., Fourier + Dirac) possess a sub-dictionary Λ such that $\text{ERC}(\Lambda) < 0$ and yet, if a signal can be written as $\boldsymbol{\Phi}_\Lambda\, \boldsymbol{c}$, then the coefficient vector $\boldsymbol{c}$ provides the unique maximally sparse representation of the signal [FN03, Tro03a]. It follows that there is an input signal whose sparsest representation over the dictionary cannot be recovered by solving (4.1) with small γ.

Proof. Since $\text{ERC}(\Lambda) < 0$, there must exist an atom $\boldsymbol{\varphi}_\omega$ for which $\left\|\boldsymbol{\Phi}_\Lambda{}^+ \boldsymbol{\varphi}_\omega\right\|_1 > 1$ even though $\omega \notin \Lambda$. Perversely, we select the input signal to be $\boldsymbol{s} = P_\Lambda\, \boldsymbol{\varphi}_\omega$. To synthesize $\boldsymbol{s}$ using the atoms in Λ, we use the coefficient vector $\boldsymbol{c}_\Lambda = \boldsymbol{\Phi}_\Lambda{}^+\boldsymbol{\varphi}_\omega$.

According to Corollary 4.8, the minimizer $\boldsymbol{b}_\star$ of the function

$$L(\boldsymbol{b}) \quad = \quad \tfrac{1}{2}\, \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}\|_2^2 \ + \ \gamma\, \|\boldsymbol{b}\|_1$$

over all coefficient vectors supported on Λ must satisfy $\|\boldsymbol{c}_\Lambda - \boldsymbol{b}_\star\|_\infty \leq \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^*\boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty}$. Since $\|\boldsymbol{c}_\Lambda\|_1 > 1$ by construction, we may choose γ small enough that the bound $\|\boldsymbol{b}_\star\|_1 > 1$ is also in force. Define the corresponding approximation $\boldsymbol{a}_\star = \boldsymbol{\Phi}\, \boldsymbol{b}_\star$.

Now we construct a parameterized coefficient vector

$$\boldsymbol{b}(t) \quad \stackrel{\text{def}}{=} \quad (1 - t)\, \boldsymbol{b}_\star \ + \ t\, \mathbf{e}_\omega \qquad \text{for } t \text{ in } [0, 1].$$

For positive t, it is clear that the support of $\boldsymbol{b}(t)$ is not contained in Λ. We shall prove that $L(\boldsymbol{b}(t)) < L(\boldsymbol{b}_\star)$ for small, positive t. Since $\boldsymbol{b}_\star$ minimizes L over all coefficient vectors supported on Λ, no global minimizer of L can be supported on Λ.

To proceed, calculate that

$$\begin{aligned} L(\boldsymbol{b}(t)) \quad &= \quad \tfrac{1}{2}\, \|(\boldsymbol{s} - \boldsymbol{a}_\star) \ + \ t\, (\boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega)\|_2^2 \ + \ \gamma\, \|(1 - t)\, \boldsymbol{b}_\star \ + \ t\, \mathbf{e}_\omega\|_1 \\ &= \quad \tfrac{1}{2}\, \|\boldsymbol{s} - \boldsymbol{a}_\star\|_2^2 \ + \ t\, \operatorname{Re} \langle \boldsymbol{s} - \boldsymbol{a}_\star, \boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega \rangle \ + \ \tfrac{1}{2} t^2\, \|\boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega\|_2^2 \ + \ \gamma\, (1 - t)\, \|\boldsymbol{b}_\star\|_1 \ - \ t\gamma. \end{aligned}$$

Differentiate this expression with respect to t and evaluate the derivative at $t = 0$:

$$\left.\frac{\mathrm{d} L(\boldsymbol{b}(t))}{\mathrm{d} t}\right|_{t=0} \quad = \quad \operatorname{Re} \langle \boldsymbol{s} - \boldsymbol{a}_\star, \boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega \rangle \ + \ \gamma\, (1 - \|\boldsymbol{b}_\star\|_1).$$

By construction of $\boldsymbol{b}_\star$, the second term is negative. The first term is non-positive because

$$\begin{aligned}\langle \boldsymbol{s} - \boldsymbol{a}_\star, \boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega \rangle &= \langle P_\Lambda(\boldsymbol{s} - \boldsymbol{a}_\star), \boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega \rangle = \langle \boldsymbol{s} - \boldsymbol{a}_\star, P_\Lambda(\boldsymbol{a}_\star - \boldsymbol{\varphi}_\omega) \rangle \\ &= \langle \boldsymbol{s} - \boldsymbol{a}_\star, \boldsymbol{a}_\star - \boldsymbol{s} \rangle = -\, \|\boldsymbol{s} - \boldsymbol{a}_\star\|_2^2 .\end{aligned}$$

Therefore, the derivative is negative, and $L(\boldsymbol{b}(t)) < L(\boldsymbol{b}(0))$ for small, positive t. Since $\boldsymbol{b}(0) = \boldsymbol{b}_\star$, the proof is complete. □

5. Subset Selection

For a fixed signal $\boldsymbol{s}$ and a threshold τ, recall that the subset selection problem is

$$\min_{\boldsymbol{c} \in \mathbb{C}^\Omega} \quad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2^2 \;+\; \tau^2 \;\|\boldsymbol{c}\|_0 \,. \tag{5.1}$$

One should observe that the atoms which participate in a solution must form a linearly independent set, or else some could be discarded to improve the objective function.

The parameter τ indicates how much an atom must improve the approximation error before it is allowed to participate. When τ reaches the norm of the input signal, the zero vector is the unique solution of (5.1). On the other hand, as τ approaches zero, solutions will involve as many atoms as it takes to represent the signal exactly. A solution will require fully d atoms except when the input signal belongs to a set with Lebesgue measure zero in $\mathbb{C}^d$ [Tro03a].

If the dictionary is orthonormal, one may solve the subset selection problem by applying a hard threshold operator (see Figure 7) with cutoff τ to each coefficient in the orthogonal expansion of the signal [Mal99]. In effect, one retains every atom whose inner product with the signal is larger than τ and discards the rest. This heuristic is nearly correct even if the dictionary is not orthonormal.

Proposition 5.1. *Fix an input signal $\boldsymbol{s}$, and choose a threshold τ. Suppose that the coefficient vector $\boldsymbol{c}_{\mathrm{opt}}$ solves the subset selection problem, and set $\boldsymbol{a}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$.*

- *For each $\lambda \in \mathrm{supp}(\boldsymbol{c}_{\mathrm{opt}})$, we have $|\boldsymbol{c}_{\mathrm{opt}}(\lambda)| \geq \tau$.*
- *For each $\omega \notin \mathrm{supp}(\boldsymbol{c}_{\mathrm{opt}})$, we have $|\langle \boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}, \boldsymbol{\varphi}_\omega \rangle| \leq \tau$.*

If we insist that $\boldsymbol{c}_{\mathrm{opt}}$ is a maximally diverse (i.e., minimally sparse) solution to the subset selection problem, the second inequality is strict.

For continuity, we shall postpone the proof until Section 5.2.

A convex relaxation of the subset selection problem is

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \quad \tfrac{1}{2}\, \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2^2 \;+\; \gamma\, \|\boldsymbol{b}\|_1 \,. \tag{5.2}$$

Our theory will supply the correct relationship between γ and τ. Suppose for a moment that the dictionary is orthonormal. Then one may solve the convex relaxation (5.2) by applying a soft threshold operator (see Figure 7) with cutoff γ to each coefficient in the orthogonal expansion of the signal [Mal99]. This amounts to retaining every atom whose inner product with the signal is strictly greater than γ and discarding the rest. We see that the form of the relaxation has been adapted so that the parameter still determines the location of the cutoff.

One should not expect to solve the subset selection problem directly by means of convex relaxation because the ℓ_1 penalty has the effect of shrinking the optimal coefficients. Statisticians have exploited this property to improve the behavior of their estimators [DJ92, Tib96]. In the present setting, it is a nuisance. Our hope is that the coefficient vector which solves the convex relaxation has the same *support* as the optimal coefficient vector. Then we may solve the original subset selection problem by projecting the signal onto the atoms indexed by the support.

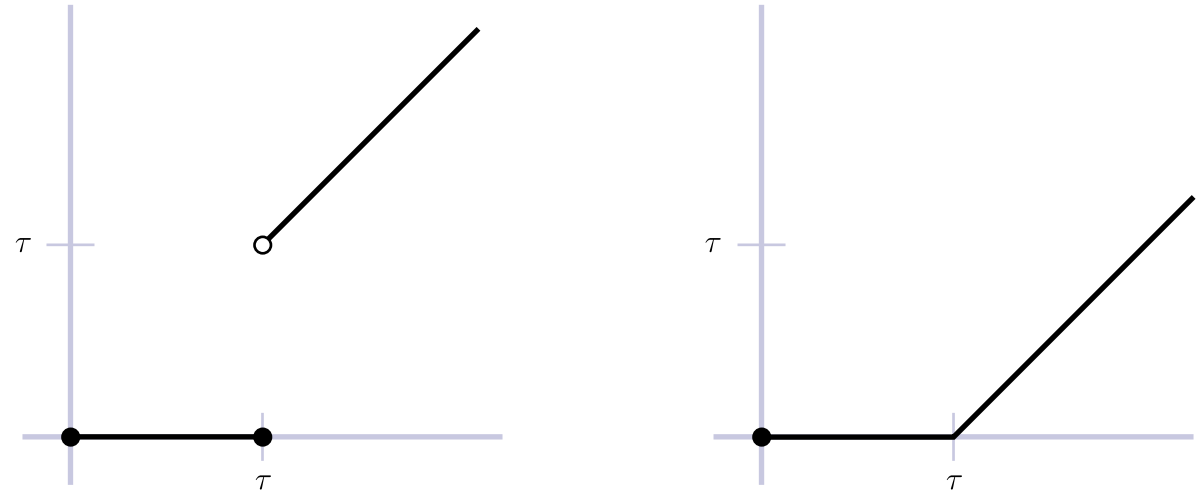

FIGURE 7. At left, the hard thresholding operator with cutoff τ. At right, the soft thresholding operator with cutoff τ.

5.1. **Results.** If the dictionary is incoherent and the threshold parameters are correctly chosen, then convex relaxation identifies every significant atom from the solution to the subset selection problem and no others.

To simplify the statement of the results, we shall extract some of the hypotheses. Fix an input signal $\boldsymbol{s}$, and choose a threshold parameter τ. Suppose that the coefficient vector $\boldsymbol{c}_{\mathrm{opt}}$ is a solution to the subset selection problem (5.1) with threshold τ and that $\boldsymbol{a}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$ is the corresponding approximation of the signal. Let $\Lambda_{\mathrm{opt}} = \operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and define the corresponding synthesis matrix $\boldsymbol{\Phi}_{\mathrm{opt}}$. Assume, moreover, that $\operatorname{ERC}(\Lambda_{\mathrm{opt}}) > 0$.

Theorem 5.2 (Relaxed Subset Selection)**.** *Suppose that the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (5.2) *with threshold $\gamma = \tau \,/\, \operatorname{ERC}(\Lambda_{\mathrm{opt}})$. Then it follows that*

- *the relaxation never selects a non-optimal atom since $\operatorname{supp}(\boldsymbol{b}_\star) \subset \operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$;*
- *the solution of the relaxation is nearly optimal since*

$$\|\boldsymbol{b}_\star - \boldsymbol{c}_{\mathrm{opt}}\|_\infty \quad \leq \quad \frac{\left\|(\boldsymbol{\Phi}_{\mathrm{opt}}{}^*\boldsymbol{\Phi}_{\mathrm{opt}})^{-1}\right\|_{\infty,\infty}}{\operatorname{ERC}(\Lambda_{\mathrm{opt}})}\,\tau;$$

- *in particular,* $\operatorname{supp}(\boldsymbol{b}_\star)$ *contains every index λ for which*

$$|\boldsymbol{c}_{\mathrm{opt}}(\lambda)| \quad > \quad \frac{\left\|(\boldsymbol{\Phi}_{\mathrm{opt}}{}^*\boldsymbol{\Phi}_{\mathrm{opt}})^{-1}\right\|_{\infty,\infty}}{\operatorname{ERC}(\Lambda_{\mathrm{opt}})}\,\tau;$$

- *and the solution of the convex relaxation is unique.*

We postpone the proof to Section 5.2 so that we may discuss the consequences of the theorem. On account of Proposition 5.1, every nonzero coefficient in $\boldsymbol{c}_{\mathrm{opt}}$ has a magnitude of at least τ. Therefore, convex relaxation will not miss a coefficient unless it barely reaches the threshold τ. Observe that the result depends on the Exact Recovery Coefficient of the optimal sub-dictionary Λ_{opt}, so the non-optimal atoms must not resemble the optimal atoms too strongly. The theorem also prefers that the dual system of the sub-dictionary exhibit small pairwise inner products. From the discussion in Section 3, we see that convex relaxation performs best when the dictionary is a good packing of lines in projective space. Furthermore, we note that the theorem holds for every

solution of the sparse approximation problem, and so the convex relaxation will only identify atoms that appear in *every* solution of the sparse approximation problem.

As a reality check, let us apply Theorem 5.2 to the case where dictionary is orthonormal. Every sub-dictionary has an Exact Recovery Coefficient of one. In addition, the inverse Gram matrix equals the identity, so its (∞, ∞) norm is one. Therefore, the theorem advises that we solve the convex relaxation with $\gamma = \tau$, and it states that the solution will involve just those atoms whose optimal coefficients are strictly larger than τ. In other words, the theorem describes the behavior of the soft-thresholding operator with cutoff τ.

Using the incoherence function, we may develop versions of the theorem that depend only on the size of the optimal index set.

Corollary 5.3. *Suppose the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (5.2) *with threshold*

$$\gamma \quad = \quad \frac{1 \;-\; \mu_1(m-1)}{1 \;-\; \mu_1(m-1) \;-\; \mu_1(m)} \; \tau.$$

Then $\operatorname{supp}(\boldsymbol{b}_\star)$ *is contained in* $\operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$*, and*

$$\|\boldsymbol{b}_\star - \boldsymbol{c}_{\mathrm{opt}}\|_\infty \quad \leq \quad \frac{\tau}{1 \;-\; \mu_1(m-1) \;-\; \mu_1(m)}.$$

This result follows immediately from the estimates in Propositions 3.6 and 3.7. By positing a specific bound for $\mu_1(m)$, we may develop a more quantitative result.

Corollary 5.4. *Assume that that the support of $\boldsymbol{c}_{\mathrm{opt}}$ indexes m atoms or fewer, where $\mu_1(m) \leq \frac{1}{3}$, and suppose that the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (2.6) *with threshold $\gamma = 2\,\tau$. It follows that* $\operatorname{supp}(\boldsymbol{b}_\star) \subset \operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$ *and that $\|\boldsymbol{b}_\star - \boldsymbol{c}_{\mathrm{opt}}\|_\infty \leq 3\,\tau$.*

Of course, smaller bounds will give better conclusions. We developed Theorem A follows from this corollary and the fact that $\mu_1(m) \leq m\,\mu$. Finally, note that, for exactly sparse signals, the theorem reduces to the known result that convex relaxation can recover all the atoms in the signal.

Corollary 5.5 (Fuchs [Fuc02])**.** *Suppose that* $\mathrm{ERC}(\Lambda_{\mathrm{opt}}) > 0$*, and assume that the input signal can be expressed exactly with the atoms in Λ_{opt}. For sufficiently small γ, the solution to the convex relaxation* (5.2) *has support equal to Λ_{opt}.*

5.2. **Proofs.** Let us begin with a proof of Proposition 5.1, which describes solutions of the subset selection problem.

Proof of Proposition 5.1. For a given threshold τ and input signal $\boldsymbol{s}$, suppose that the coefficient vector $\boldsymbol{c}_{\mathrm{opt}}$ is a solution of the subset selection problem (5.1). Let $\boldsymbol{a}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$.

Take an index ω outside $\operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and let $\boldsymbol{P}_{\mathrm{opt}}$ denote the orthogonal projector onto the atoms listed in $\operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$. Adding the atom $\boldsymbol{\varphi}_\omega$ to the approximation would diminish the squared error by exactly

$$\frac{|\langle \boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}, \boldsymbol{\varphi}_\omega \rangle|^2}{\|\boldsymbol{P}_{\mathrm{opt}}\,\boldsymbol{\varphi}_\omega\|_2^2}. \tag{5.3}$$

This quantity must be less than or equal to τ^2, or else we could immediately construct a solution to the subset selection problem that is strictly better than $\boldsymbol{c}_{\mathrm{opt}}$. Every atom has unit Euclidean norm, and projections can only attenuate the Euclidean norm. It follows that $\|\boldsymbol{P}_{\mathrm{opt}}\,\boldsymbol{\varphi}_\omega\|_2^2 \leq 1$, and so $|\langle \boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}, \boldsymbol{\varphi}_\omega \rangle| \leq \tau$.

Suppose now that $\boldsymbol{c}_{\mathrm{opt}}$ is a maximally diverse solution to the subset selection problem. Then the expression (5.3) must be strictly smaller τ^2, or else we would be able to construct a solution that involves an additional atom, which is impossible. Following the same train of reasoning as before, we conclude that $|\langle \boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}, \boldsymbol{\varphi}_\omega \rangle| < \tau$.

Choose an index λ inside $\operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and let P denote the orthogonal projector onto the span of the atoms listed by $\operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}}) \setminus \{\lambda\}$. Removing the atom $\boldsymbol{\varphi}_\lambda$ from the approximation would increase the squared error by exactly

$$|\boldsymbol{c}_{\mathrm{opt}}(\lambda)|^2 \, \|(\mathsf{I} - \mathsf{P})\boldsymbol{\varphi}_\lambda\|_2^2 .$$

This quantity must be at least τ^2. Since $(\mathsf{I} - \mathsf{P})$ is an orthogonal projector, $\|(\mathsf{I} - \mathsf{P})\boldsymbol{\varphi}_\lambda\|_2^2 \leq 1$. We conclude that $|\boldsymbol{c}_{\mathrm{opt}}(\lambda)| \geq \tau$. □

One could obviously prove much more about the solutions of the subset selection problem using similar techniques, but these results are too tangential to pursue here. Now we turn our attention to the proof of Theorem 5.2. This result involves a straightforward application of the fundamental lemmata.

Proof of Theorem 5.2. Suppose that the coefficient vector $\boldsymbol{c}_{\mathrm{opt}}$ is a solution to the subset selection problem (5.1) with threshold parameter τ. The associated approximation of the signal is $\boldsymbol{a}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$. Define $\Lambda_{\mathrm{opt}} = \operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and denote the corresponding synthesis matrix by $\boldsymbol{\Phi}_{\mathrm{opt}}$.

Let us develop an upper bound on the inner product between any atom and the residual vector $(\boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}})$. First, note that every atom indexed by Λ_{opt} has a zero inner product with the residual since $\boldsymbol{a}_{\mathrm{opt}}$ is the best approximation of $\boldsymbol{s}$ using the atoms in Λ_{opt}. Choose $\omega \notin \Lambda_{\mathrm{opt}}$. Then Proposition 5.1 shows that $|\langle \boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}, \boldsymbol{\varphi}_\omega \rangle| < \tau$. This relation holds for all $\omega \in \Omega$, and so

$$\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}})\|_\infty \quad \leq \quad \tau. \tag{5.4}$$

Note that the strict inequality relies on the fact that the dictionary is finite.

The Correlation Condition Lemma states that, for any threshold γ satisfying

$$\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}})\|_\infty \quad \leq \quad \gamma \operatorname{ERC}(\Lambda_{\mathrm{opt}}),$$

the solution $\boldsymbol{b}_\star$ to the convex relaxation (5.2) is supported on Λ_{opt}. Using the inequality (5.4), we determine that the choice

$$\gamma \quad = \quad \frac{\tau}{\operatorname{ERC}(\Lambda_{\mathrm{opt}})}$$

is sufficient to ensure that $\operatorname{supp}(\boldsymbol{b}_\star) \subset \Lambda_{\mathrm{opt}}$. The uniqueness of the minimizer $\boldsymbol{b}_\star$ follows from Lemma 4.6 since Λ_{opt} indexes a linearly independent collection of atoms. From Corollary 4.8, we obtain the upper bound

$$\|\boldsymbol{c}_{\mathrm{opt}} - \boldsymbol{b}_\star\|_\infty \quad \leq \quad \gamma \left\| (\boldsymbol{\Phi}_{\mathrm{opt}}{}^* \boldsymbol{\Phi}_{\mathrm{opt}})^{-1} \right\|_{\infty,\infty}.$$

For any index λ at which $|\boldsymbol{c}_{\mathrm{opt}}(\lambda)| > \gamma \left\| (\boldsymbol{\Phi}_{\mathrm{opt}}{}^* \boldsymbol{\Phi}_{\mathrm{opt}})^{-1} \right\|_{\infty,\infty}$, it follows that the corresponding coefficient $\boldsymbol{b}_\star(\lambda)$ must be nonzero. □

5.3. **Orthogonal Matching Pursuit.** One might wish to compare the theoretical performance of convex relaxation against a greedy algorithm. With some care, the arguments of [Tro03a] can be extended to prove a strong result about the behavior of Orthogonal Matching Pursuit (OMP) on the subset selection problem. A brief introduction to OMP appears in Appendix B.

For reference, we restate our hypotheses. Fix an input signal $\boldsymbol{s}$ and a threshold τ. Suppose that $\boldsymbol{c}_{\mathrm{opt}}$ is a solution to the subset selection problem (5.1) with threshold τ. Let $\Lambda_{\mathrm{opt}} = \operatorname{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and define $m = |\Lambda_{\mathrm{opt}}|$. Finally, assume that $\operatorname{ERC}(\Lambda_{\mathrm{opt}}) > 0$.

Theorem 5.6. *Suppose that one halts Orthogonal Matching Pursuit as soon as the maximal inner product between an atom and the residual declines below $\tau / \operatorname{ERC}(\Lambda_{\mathrm{opt}})$. Up to this point, OMP has only selected atoms listed in Λ_{opt}, and OMP has chosen every index λ for which*

$$|\boldsymbol{c}_{\mathrm{opt}}(\lambda)| \quad > \quad \frac{1 - M}{1 - 2M} \, \frac{\tau}{\operatorname{ERC}(\Lambda_{\mathrm{opt}})},$$

where the number M is determined by the formula

$$M \quad \stackrel{\text{def}}{=} \quad \max_{k=0,\dots,m} \frac{\mu_1(m-k)}{1 \;-\; \mu_1(k)}.$$

Estimating M and the Exact Recovery Coefficient by means of the coherence parameter, one reaches a more spartan result.

Corollary 5.7. *Suppose that one halts Orthogonal Matching Pursuit as soon as the maximal inner product between an atom and the residual declines below $\tau / \operatorname{ERC}(\Lambda_{\text{opt}})$. Up to this point, OMP has only selected atoms listed in Λ_{opt}, and OMP has chosen every index λ for which*

$$|\boldsymbol{c}_{\text{opt}}(\lambda)| \quad > \quad \left[\frac{1 \;-\; m\,\mu}{1 \;-\; 2\,m\,\mu}\right]^2 \tau.$$

In comparison, convex relaxation discovers every coefficient that exceeds $\tau \,/\, (1 - 2\,m\,\mu)$.

Proofs of these results will appear in a later publication. Please be aware that the last draft of this report did misstate the stopping criterion for the greedy algorithm to perform as advertised.

6. Sparse Approximation with an Error Constraint

Let $\boldsymbol{s}$ be an input signal, and choose an error tolerance ε. In this section, we shall consider the error-constrained sparse approximation problem

$$\min_{\boldsymbol{c} \in \mathbb{C}^\Omega} \; \|\boldsymbol{c}\|_0 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2 \;\leq\; \varepsilon. \tag{6.1}$$

First, notice that the support of any solution must index a linearly independent collection of atoms, or else some could be discarded. A second important point is that the solutions of (6.1) will generally form a convex set or a union of convex sets. Each minimizer will have the same level of sparsity, but they will yield different approximation errors. One may remove some of this multiplicity by considering the more convoluted mathematical program

$$\min_{\boldsymbol{c} \in \mathbb{C}^\Omega} \; \|\boldsymbol{c}\|_0 \;+\; \tfrac{1}{2}\,\|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2\,\varepsilon^{-1} \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{c}\|_2 \;\leq\; \varepsilon. \tag{6.2}$$

Any minimizer of (6.2) also solves (6.1), but it produces the smallest approximation error possible at that level of sparsity. Observe that, when ε reaches the norm of the input signal, the unique solution to either (6.1) or (6.2) is the zero vector. On the other hand, as ε approaches zero, solutions will involve as many atoms as necessary to represent the signal exactly. Essentially every signal in $\mathbb{C}^d$ requires d atoms [Tro03a].

We shall try to produce an error-constrained sparse approximation by relaxing the basic problem (6.1). This results in the convex program

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \; \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2 \;\leq\; \delta. \tag{6.3}$$

Our theory will supply the correct relationship between δ and ε. Note that if $\delta \geq \|\boldsymbol{s}\|_2$ then the optimal solution to the convex program is the zero vector. When δ is smaller, the approximation error is always equal to δ. Therefore, the minimizer of the convex program will rarely solve the sparse approximation problem (6.2). To improve the approximation obtained by relaxation, one should project the signal onto the atoms indexed by the support of the minimal coefficient vector.

6.1. Basis Pursuit.

We begin with the case where $\varepsilon = 0$. This restriction translates (6.1) into the problem of recovering a superposition of a small number of atoms. This situation is qualitatively different from (and much easier to analyze than) the general problem, so we shall treat it separately. Formally, we wish to solve

$$\min_{\boldsymbol{c} \in \mathbb{C}^\Omega} \; \|\boldsymbol{c}\|_0 \qquad \text{subject to} \qquad \boldsymbol{s} \;=\; \boldsymbol{\Phi}\,\boldsymbol{c}. \tag{6.4}$$

There are several motivations for considering (6.4). Although natural signals are not perfectly sparse, one can imagine applications in which a sparse signal is constructed and transmitted without error. Second, analysis of the simpler problem can provide lower bounds on the computational complexity of the general case. When the easy problem is hard, so is the harder problem [Tro03a].

As early as 1994, Chen and Donoho proposed that the convex program

$$\min_{\boldsymbol{b} \in \mathbb{C}^{\Omega}} \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \boldsymbol{s} = \boldsymbol{\Phi}\,\boldsymbol{b}. \tag{6.5}$$

could be used to recover exactly sparse linear combinations [CD94]. They called this approach *Basis Pursuit*. Their paper [CDS99] gives copious numerical evidence that the method succeeds but provides no rigorous proof.

Later Donoho and Huo were able to show that Basis Pursuit can recover exactly sparse linear combinations from an incoherent union of two orthonormal bases [DH01]. These early results have been extended to cover the entire class of incoherent dictionaries in a series of recent papers [EB02, GN03b, Fuc02, DE03, Tro03a]. To indicate the tenor of this work, we quote one theorem.

Theorem 6.1 (Fuchs [Fuc02], Tropp [Tro03a])**.** *Suppose that* $\mathrm{ERC}(\Lambda_{\mathrm{opt}}) > 0$ *and that the input signal is a superposition of the atoms indexed by* Λ_{opt}*. Then the unique solution of the Basis Pursuit problem* (6.5) *is identical with the unique solution of the sparse approximation problem* (6.4)*.*

Note that the same theorem holds for the greedy algorithm Orthogonal Matching Pursuit [Tro03a]. Proposition 3.7 furnishes an immediate corollary.

Corollary 6.2 (Donoho–Elad [DE03], Tropp [Tro03a])**.** *Suppose that* $\mu_1(m-1)+\mu_1(m) < 1$*. Then Basis Pursuit recovers any exact superposition of* m *atoms or fewer.*

Gribonval and Nielsen have also established a necessary and sufficient condition for Basis Pursuit to succeed [GN03b, GN03a]. Their results show that the theorem and corollary can be improved, but they have not provided a simple method for determining when or how much.

6.2. **The General Case.** Our major theorem proves that, under appropriate conditions, the solution to the relaxation (6.3) for a given δ is at least as sparse as a solution to the sparse approximation problem (6.2) for a smaller value of ε.

To make the statement of the result more transparent, let us extract some of the hypotheses. Fix an input signal $\boldsymbol{s}$, and choose an error tolerance ε. Suppose that the coefficient vector $\boldsymbol{c}_{\mathrm{opt}}$ solves the sparse approximation problem (6.2) with tolerance ε, and let the corresponding approximation of the signal be $\boldsymbol{a}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$. Define the optimal index set $\Lambda_{\mathrm{opt}} = \mathrm{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and let $\boldsymbol{\Phi}_{\mathrm{opt}}$ be the associated synthesis matrix. Assume, moreover, that $\mathrm{ERC}(\Lambda_{\mathrm{opt}}) > 0$.

Theorem 6.3 (Relaxed Sparse Approximation)**.** *Suppose that the coefficient vector* $\boldsymbol{b}_\star$ *solves the convex relaxation* (6.3) *with an error tolerance*

$$\delta \quad \geq \quad \left[1 \; + \; \left(\frac{\mathrm{maxcor}(\boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}) \; \left\| \boldsymbol{\Phi}_{\mathrm{opt}}{}^{+} \right\|_{2,1}}{\mathrm{ERC}(\Lambda_{\mathrm{opt}})} \right)^2 \right]^{1/2} \varepsilon. \tag{6.6}$$

Then it follows that

- *this solution is at least as sparse as* $\boldsymbol{c}_{\mathrm{opt}}$ *since* $\mathrm{supp}(\boldsymbol{b}_\star) \subset \mathrm{supp}(\boldsymbol{c}_{\mathrm{opt}})$*;*
- *yet* $\boldsymbol{b}_\star$ *is no sparser than a solution of the sparse approximation problem with tolerance* δ*;*
- *the coefficient vector* $\boldsymbol{b}_\star$ *is nearly optimal since* $\|\boldsymbol{b}_\star - \boldsymbol{c}_{\mathrm{opt}}\|_2 \leq \delta \; \left\| \boldsymbol{\Phi}_{\mathrm{opt}}{}^{+} \right\|_{2,2}$*; and*
- *the relaxation has no other solution.*

As usual, we postpone the proof until we have completed the commentary. Notice that the result depends strongly on several geometric properties of the dictionary. First, the dependence on the Exact Recovery Coefficient shows that non-optimal atoms must not resemble the optimal atoms

too strongly. Second, the presence of the $(2, 1)$ operator norm shows that the optimal atoms should cover their span well. Third, the optimal solution is easiest to recover when the residual left over after approximation is badly correlated with the dictionary. Section 3 treats these factors in more detail.

A practical question is to estimate the size of the bracketed constant. Suppose that the signal has a good approximation over a small sub-dictionary Λ_{opt} with a moderate Exact Recovery Coefficient. If the dictionary is incoherent, one expects that the maximum correlation of the residual with the dictionary is on the order of $1/\sqrt{d}$ and that the $(2, 1)$ operator norm is on the order of $\sqrt{|\Lambda_{\text{opt}}|}$. Under these assumptions, the constant may well be less than $\sqrt{2}$. On the other hand, the maximum correlation between the residual and the dictionary may be as large as one, which yields a worst-case corollary.

Corollary 6.4. *Suppose that the coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation* (6.3) *with an error tolerance*

$$\delta \quad \geq \quad \left[1 \ + \ \left(\frac{\left\| \boldsymbol{\Phi}_{\text{opt}}{}^{+} \right\|_{2,1}}{\text{ERC}(\Lambda_{\text{opt}})} \right)^2 \right]^{1/2} \varepsilon.$$

Then the conclusions of Theorem 6.3 are in force.

Using the cumulative coherence function, we can develop results that do not depend on the specific index set Λ_{opt} at all.

Corollary 6.5. *Fix an input signal, and choose an error tolerance ε. Suppose that $\boldsymbol{c}_{\text{opt}}$ solves the sparse approximation problem* (6.2) *with tolerance ε and that* $\text{supp}(\boldsymbol{c}_{\text{opt}})$ *contains m indices. Assume that the incoherence condition $\mu_1(m-1) + \mu_1(m) < 1$ holds, and pick the parameter*

$$\delta \quad \geq \quad \left[1 \ + \ \frac{1 \ - \ \mu_1(m-1)}{[1 \ - \ \mu_1(m-1) \ - \ \mu_1(m)]^2} \ m \right]^{1/2} \varepsilon.$$

It follows that

- *the* unique *solution $\boldsymbol{b}_\star$ to the convex relaxation* (6.3) *with error tolerance δ is supported inside* $\text{supp}(\boldsymbol{c}_{\text{opt}})$*;*
- *the coefficient vector $\boldsymbol{b}_\star$ is nearly optimal since $\|\boldsymbol{b}_\star - \boldsymbol{c}_{\text{opt}}\|_2 \leq \delta \,/\, \sqrt{1 - \mu_1(m-1)}$; and*
- *yet $\boldsymbol{b}_\star$ is no sparser than a solution of the sparse approximation problem with tolerance δ.*

By positing a bound on the cumulative coherence, we may reach a more quantitative result.

Corollary 6.6. *Fix an input signal, and choose an error tolerance ε. Suppose that an optimal solution to the sparse approximation problem* (6.2) *with tolerance ε requires m atoms or fewer, where m satisfies the incoherence condition $\mu_1(m) \leq \frac{1}{3}$. Select $\delta \geq \varepsilon \sqrt{1 + 6\,m}$. It follows that the unique solution to the convex relaxation* (6.3) *with tolerance δ involves a subset of the optimal atoms. This solution diverges in Euclidean norm from the optimal coefficient vector by no more than $\delta \sqrt{3/2}$. Moreover, it is no sparser than a solution of the sparse approximation problem with tolerance δ.*

We have lost quite a lot by failing to account for the direction of the residual.

6.3. **Proof of the Relaxation Theorem.** Now we prove the result. The argument is rather more difficult than that of Theorem 5.2 because it involves Karush–Kuhn–Tucker conditions.

Proof of Theorem 6.3. Let $\boldsymbol{s}$ be an input signal, and suppose that $\boldsymbol{c}_{\text{opt}}$ solves the sparse approximation problem (6.2) with error tolerance ε. Denote the optimal approximation as $\boldsymbol{a}_{\text{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\text{opt}}$. Let $\Lambda_{\text{opt}} = \text{supp}(\boldsymbol{c}_{\text{opt}})$, and define $\boldsymbol{\Phi}_{\text{opt}}$ to be the associated synthesis matrix. Assume also that $\|\boldsymbol{s}\|_2 > \delta$, or else the zero vector is the unique solution of the relaxation (6.3). It is self-evident

that the solution of the relaxation can be no sparser than a solution of the sparse approximation problem with tolerance δ.

To prove the theorem, we shall find a coefficient vector supported on Λ_{opt} and a corresponding Lagrange multiplier that give a potential solution of the relaxation. We shall argue that the condition (6.6) guarantees that this coefficient vector actually minimizes the relaxation. Then we shall demonstrate that this coefficient vector gives the unique minimizer. As a coda, we shall estimate how much the solution of the relaxation varies from the optimal coefficient vector.

A coefficient vector $\boldsymbol{b}_\star$ solves the convex relaxation

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2 \;\leq\; \delta. \tag{6.7}$$

if and only if the Karush–Kuhn–Tucker conditions are satisfied [Roc70]. That is, there exists a Lagrange multiplier $\gamma_\star$ for which

$$\boldsymbol{b}_\star \quad \in \quad \arg\min_{\boldsymbol{b}} \quad \tfrac{1}{2}\,\|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2^2 \;+\; \gamma_\star\,\|\boldsymbol{b}\|_1 \tag{6.8}$$

$$\|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}_\star\|_2 \quad = \quad \delta \tag{6.9}$$

$$\gamma_\star \quad > \quad 0. \tag{6.10}$$

The KKT conditions are both necessary and sufficient because the objective function and constraint set are convex. Note that (6.9) and (6.10) hold because $\|\boldsymbol{s}\|_2 > \delta$ implies that the error constraint is strictly binding. Since the Lagrange multiplier $\gamma_\star$ is positive, we have transferred it to the ℓ_1 term in (6.8) to simplify the application of the formulae we have developed.

Let $(\boldsymbol{b}_\star, \gamma_\star)$ be a solution to the restricted problem

$$\min_{\text{supp}(\boldsymbol{b}) \subset \Lambda_{\text{opt}}} \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2 \;\leq\; \delta. \tag{6.11}$$

The hypothesis $\|\boldsymbol{s}\|_2 > \delta$ implies that the error constraint in (6.11) is strictly binding, so (6.9) and (6.10) are both in force. Applying the Pythagorean Theorem to (6.9), we obtain the identity

$$\|\boldsymbol{a}_{\text{opt}} - \boldsymbol{\Phi}_{\text{opt}}\,\boldsymbol{b}_\star\|_2 \quad = \quad \left[\delta^2 \;-\; \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2^2\right]^{1/2}. \tag{6.12}$$

Corollary 4.8 furnishes the estimate

$$\|\boldsymbol{a}_{\text{opt}} - \boldsymbol{\Phi}_{\text{opt}}\,\boldsymbol{b}_\star\|_2 \quad \leq \quad \gamma_\star\,\left\|\boldsymbol{\Phi}_{\text{opt}}{}^+\right\|_{2,1}.$$

Introducing (6.12) into this relation, we obtain a lower bound on the multiplier:

$$\gamma_\star \quad \geq \quad \left[\delta^2 \;-\; \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2^2\right]^{1/2}\,\left\|\boldsymbol{\Phi}_{\text{opt}}{}^+\right\|_{2,1}^{-1}. \tag{6.13}$$

Meanwhile, the Correlation Condition Lemma gives a sufficient condition,

$$\gamma_\star \quad \geq \quad \frac{\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})\|_\infty}{\text{ERC}(\Lambda_{\text{opt}})}, \tag{6.14}$$

which ensures that any coefficient vector satisfying (6.8) is supported on Λ_{opt}. Combine (6.13) into (6.14), and rearrange to obtain

$$\|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2^2 \;+\; \frac{\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})\|_\infty^2\,\left\|\boldsymbol{\Phi}_{\text{opt}}{}^+\right\|_{2,1}^2}{\text{ERC}(\Lambda_{\text{opt}})^2} \quad \leq \quad \delta^2.$$

Factor $\|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2^2$ out from the left-hand side, identify the maximum correlation of $(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})$ with the dictionary, and take square roots to reach

$$\left[1 \;+\; \left(\frac{\text{maxcor}(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})\,\left\|\boldsymbol{\Phi}_{\text{opt}}{}^+\right\|_{2,1}}{\text{ERC}(\Lambda_{\text{opt}})}\right)^2\right]^{1/2} \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2 \quad \leq \quad \delta.$$

Since $\boldsymbol{a}_{\mathrm{opt}}$ is an approximation of $\boldsymbol{s}$ with error less than or equal to ε, the hypothesis (6.6) is a sufficient condition for the pair $(\boldsymbol{b}_\star, \gamma_\star)$ to satisfy all three KKT conditions (6.8), (6.9) and (6.10). It follows that our coefficient vector $\boldsymbol{b}_\star$ gives a solution to the convex relaxation (6.7).

Now we shall demonstrate that the coefficient vector $\boldsymbol{b}_\star$ provides the *unique* minimizer of the convex relaxation. This requires some work because we have not proven that every solution of the convex program is necessarily supported on Λ_{opt}.

Suppose that $\boldsymbol{b}_{\mathrm{alt}}$ is another coefficient vector that solves (6.7). First, we argue that $\boldsymbol{\Phi}\,\boldsymbol{b}_{\mathrm{alt}} = \boldsymbol{\Phi}\,\boldsymbol{b}_\star$ by assuming the contrary. The condition (6.9) must hold at every solution, so the signals $\boldsymbol{\Phi}\,\boldsymbol{b}_{\mathrm{alt}}$ and $\boldsymbol{\Phi}\,\boldsymbol{b}_\star$ both lie on a Euclidean sphere of radius δ centered at the input signal $\boldsymbol{s}$. Since Euclidean balls are strictly convex, the signal $\frac{1}{2}\,\boldsymbol{\Phi}\,(\boldsymbol{b}_{\mathrm{alt}} + \boldsymbol{b}_\star)$ is strictly closer than δ to the input signal. Thus $\frac{1}{2}\,(\boldsymbol{b}_{\mathrm{alt}} + \boldsymbol{b}_\star)$ cannot be a solution of the convex relaxation. But the solutions to a convex program always form a convex set, which is a contradiction. In consequence, any alternate solution $\boldsymbol{b}_{\mathrm{alt}}$ synthesizes the same signal as $\boldsymbol{b}_\star$. Moreover, $\boldsymbol{b}_{\mathrm{alt}}$ and $\boldsymbol{b}_\star$ share the same ℓ_1 norm because they both solve (6.7). Under our hypothesis that $\mathrm{ERC}(\Lambda_{\mathrm{opt}}) > 0$, Theorem 4.5 shows that $\boldsymbol{b}_\star$ is the *unique* solution to the problem

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \quad \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \boldsymbol{\Phi}\,\boldsymbol{b} \;=\; \boldsymbol{\Phi}\,\boldsymbol{b}_\star.$$

Thus $\boldsymbol{b}_{\mathrm{alt}} = \boldsymbol{b}_\star$. We conclude that $\boldsymbol{b}_\star$ is the unique minimizer of the convex relaxation.

Finally, let us estimate how far $\boldsymbol{b}_\star$ varies from $\boldsymbol{c}_{\mathrm{opt}}$. We begin with the equation (6.12), which can be written

$$\|\boldsymbol{\Phi}_{\mathrm{opt}}\,(\boldsymbol{c}_{\mathrm{opt}} - \boldsymbol{b}_\star)\|_2 \quad = \quad \left[\delta^2 \;-\; \|\boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}}\|_2^2\right]^{1/2}.$$

The right-hand side clearly does not exceed δ, while the left-hand side may be bounded below as

$$\left\|\boldsymbol{\Phi}_{\mathrm{opt}}{}^+\right\|_{2,2}^{-1} \|\boldsymbol{c}_{\mathrm{opt}} - \boldsymbol{b}_\star\|_2 \quad \leq \quad \|\boldsymbol{\Phi}_{\mathrm{opt}}\,(\boldsymbol{c}_{\mathrm{opt}} - \boldsymbol{b}_\star)\|_2\,.$$

Combine the two bounds and rearrange to complete the argument. □

6.4. **Results for Greedy Algorithms.** The literature contains a few results on the application of greedy algorithms to the error-constrained sparse approximation problem (6.2). For an introduction to one greedy algorithm, Orthogonal Matching Pursuit (OMP), turn to Appendix B.

The paper [Tro03a] offers an approximation guarantee for OMP. A slight modification of the argument there yields the following result. Fix an input signal $\boldsymbol{s}$, and choose an error tolerance ε. Suppose that the vector $\boldsymbol{c}_{\mathrm{opt}}$ solves the sparse approximation problem (6.2) with tolerance ε, and let $\boldsymbol{a}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$ be the corresponding approximation. Set $\Lambda_{\mathrm{opt}} = \mathrm{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and denote by $\boldsymbol{\Phi}_{\mathrm{opt}}$ the associated synthesis matrix. Assume, moreover, that $\mathrm{ERC}(\Lambda_{\mathrm{opt}}) > 0$.

Theorem 6.7. *Suppose that one halts Orthogonal Matching Pursuit as soon as the approximation error has declined past*

$$\left[1 \;+\; \left(\frac{\mathrm{maxcor}(\boldsymbol{s} - \boldsymbol{a}_{\mathrm{opt}})\;\left\|\boldsymbol{\Phi}_{\mathrm{opt}}{}^+\right\|_{2,1}}{\mathrm{ERC}(\Lambda_{\mathrm{opt}})}\right)^2\right]^{1/2} \varepsilon.$$

Then every atom that Orthogonal Matching Pursuit has chosen must be listed in Λ_{opt}.

It is remarkable that the analyses of convex relaxation and Orthogonal Matching Pursuit lead to nearly identical results. It is also surprising because empirical evidence suggests that convex relaxation is more powerful than OMP [CDS99]. Incidentally, one should be aware that Theorem 6.7 is pessimistic about the performance of OMP when the dictionary is an orthonormal basis.

Another worthwhile result concerns the behavior of the forward selection method [Nat95].

Theorem 6.8 (Natarajan). *Assume that $N \leq d$. Given an input signal $\boldsymbol{s}$, suppose that one uses the forward selection method to compute an approximation with error tolerance δ. The number of atoms chosen is no greater than*

$$\left\lceil 18\, m\, \left\|\boldsymbol{\Phi}^{+}\right\|_{2,2}^{2} \ln \frac{\|\boldsymbol{s}\|_2}{\delta} \right\rceil,$$

where m is the number of atoms that participate in a solution of the sparse approximation problem (6.2) *with error tolerance $\varepsilon = \frac{1}{2}\,\delta$.*

Note that Natarajan's paper is missing the hypothesis $N \leq d$. With a little effort, one may adapt his analysis to Orthogonal Matching Pursuit, and reduce the factor 18 down to 8.

Finally, we should mention that Couvreur and Bresler have obtained a qualitative result for the backward selection procedure [CB00].

Theorem 6.9 (Couvreur–Bresler). *Assume that $N \leq d$ and that the entire dictionary forms a linearly independent set. The backward selection procedure can solve the error-constrained sparse approximation problem for every input signal that is sufficiently close to an exact superposition of atoms.*

The definition of "sufficiently close" depends on which superposition of atoms we are trying to recover, and no quantitative estimates are presently available.

7. Horizons

We have seen conclusively that convex relaxation may be used to determine nearly optimal solutions of the subset selection problem and the error-constrained sparse approximation problem. Moreover, the efficacy of the convex relaxation is intimately related to the geometric properties of the dictionary that is being used for approximation. Even though convex relaxation has been applied to similar problems for over thirty years, the present results are unprecedented in the published literature because they do not require the input signal to have an exact, sparse representation. Nor do they assume that the dictionary has any kind of structure beyond incoherence. This report should have a significant impact because it proves how convex relaxation will behave in a practical setting.

Nevertheless, there are still many directions for further research. First of all, the ℓ_0 quasi-norm is not always the correct way to promote sparsity. For example, an accurate model of compression would measure diversity as the number of bits necessary to represent the coefficient vector with a certain precision. Gribonval and Nielsen have made some preliminary progress on this problem [GN03a]. Second, one may wish to consider sparse approximation problems with respect to error measures other than the usual Euclidean distance. For example, the uniform norm would promote sparse approximations that commit small maximum errors, while the absolute sum error is much less sensitive to wild samples in the signal. A third variation is to consider the simultaneous sparse approximation of a collection of signals [LT03]. This problem can be dispatched with the same techniques given here and in [Tro03a], although the details are more complicated [Tro04]. Fourth, one might study sparse approximation with respect to very specific dictionaries, such as a wavelet packet dictionary or a Gabor dictionary, in hope of determining sharper results than one can obtain in the general setting. A fifth direction is to allow the input signal and/or the dictionary vectors to be random variables. As mentioned, the statistics community has spent a lot of effort here [Mil02]. In each of these cases, one should study the structure of optimal solutions, and one should attempt to exhibit algorithms that provably deliver nearly optimal solutions.

The biggest question of all may be to develop necessary and sufficient conditions on the dictionary and input signal under which sparse approximation problems are computationally tractable. Today, we are still leagues away from a detailed understanding of the computational complexity of sparse

approximation. But as Vergil reminds us, *"Tantae molis erat Romanam condere gentem."* Such a burden it was to establish the Roman race.

Acknowledgments

I wish to thank my supervisors, I. S. Dhillon and A. C. Gilbert, as well as my colleagues, R. W. Heath, S. Muthukrishnan, M. J. Strauss and T. Strohmer. Thanks also to the Austinites, who have told me for years that I should just relax. During the research and development of this report, I have been supported by an NSF Graduate Fellowship.

Appendix A. Computational Formulations

The convex programs that we have given in the text are attractive conceptually, and they are amenable to analysis. Unfortunately, they require some modification before they can be solved numerically. The reason is simple: the ℓ_1 norm is not a differentiable function. Happily, one may convert our relaxations into smooth convex programs with some standard tricks [BV04].

Let us begin with the convex relaxation of the subset selection problem:

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \ \tfrac{1}{2} \ \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}\|_2^2 \ + \ \gamma \ \|\boldsymbol{b}\|_1 \,. \tag{A.1}$$

This convex program is equivalent to the smooth, constrained convex program

$$\min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi} \in \mathbb{R}^\Omega} \ \tfrac{1}{2} \ \|\boldsymbol{s} - \boldsymbol{\Phi}\,(\boldsymbol{x} + \mathrm{i}\,\boldsymbol{y})\|_2^2 \ + \ \gamma\, \mathbf{1}^T \boldsymbol{\xi} \qquad \text{subject to} \qquad \boldsymbol{\xi} \ \geq \ \mathbf{0} \qquad \text{and}$$

$$x_\omega{}^2 + y_\omega{}^2 \ \leq \ \xi_\omega{}^2 \qquad \text{for } \omega \text{ in } \Omega.$$

As one may suspect, $\mathbf{1}$ is a conformal vector of ones. The solution of the original convex relaxation may be recovered using the formula $\boldsymbol{b}_\star = \boldsymbol{x}_\star + \mathrm{i}\,\boldsymbol{y}_\star$. In the real setting, a different formulation of (A.1) is likely to be more efficient:

$$\min_{\boldsymbol{p},\boldsymbol{n} \in \mathbb{R}^\Omega} \ \tfrac{1}{2} \ \|\boldsymbol{s} - \boldsymbol{\Phi}\,(\boldsymbol{p} - \boldsymbol{n})\|_2^2 \ + \ \gamma\, \mathbf{1}^T (\boldsymbol{p} + \boldsymbol{n}) \qquad \text{subject to} \qquad \boldsymbol{p}, \boldsymbol{n} \ \geq \ \mathbf{0}.$$

In this case, we use $\boldsymbol{b}_\star = \boldsymbol{p}_\star - \boldsymbol{n}_\star$ to recover the original solution. Note that this formulation gives a semi-definite quadratic program with *linear* constraints.

The convex relaxation of the error-constrained sparse approximation problem is

$$\min_{\boldsymbol{b} \in \mathbb{C}^\Omega} \ \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}\|_2 \ \leq \ \delta. \tag{A.2}$$

The smooth formulation is

$$\min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi} \in \mathbb{R}^\Omega} \ \mathbf{1}^T \boldsymbol{\xi} \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,(\boldsymbol{x} + \mathrm{i}\,\boldsymbol{y})\|_2^2 \ \leq \ \delta^2,$$

$$\boldsymbol{\xi} \ \geq \ \mathbf{0} \qquad \text{and}$$

$$x_\omega{}^2 + y_\omega{}^2 \ \leq \ \xi_\omega{}^2 \qquad \text{for } \omega \text{ in } \Omega.$$

Of course, we have $\boldsymbol{b}_\star = \boldsymbol{x}_\star + \mathrm{i}\,\boldsymbol{y}_\star$. In the real setting, the correct formulation of (A.2) is

$$\min_{\boldsymbol{p},\boldsymbol{n} \in \mathbb{R}^\Omega} \ \mathbf{1}^T (\boldsymbol{p} + \boldsymbol{n}) \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\,(\boldsymbol{p} - \boldsymbol{n})\|_2^2 \ \leq \ \delta^2 \qquad \text{and}$$

$$\boldsymbol{p}, \boldsymbol{n} \ \geq \ \mathbf{0}.$$

The solution of the original program is $\boldsymbol{b}_\star = \boldsymbol{p}_\star - \boldsymbol{n}_\star$. Once again, we have reached a semi-definite quadratic program with *linear* constraints.

In principle, any of these problems may be solved in polynomial time with standard convex programming software [BV04]. As part of an ongoing campaign for reproducible research, Donoho offers a publicly available software package called Atomizer that solves sparse approximation problems using Basis Pursuit. It can take advantage of special structure in the dictionary synthesis matrix to accelerate computations [CDS99]. Sardy, Bruce, and Tseng have also written a paper on

using cyclic minimization to solve the Basis Pursuit problem when the dictionary has a block-basis structure [SBT00]. Starck, Donoho, and Candès have proposed an iterative method for solving (A.1) when the dictionary has block-basis structure [SDC03]. Most recently, Daubechies, Defrise, and De Mol have developed an iterative method for solving a class of ℓ_1-regularized inverse problems that are similar to (A.1) in form [DDM03].

Appendix B. Orthogonal Matching Pursuit

If the dictionary is an orthonormal basis, it is computationally easy to solve sparse approximation problems. One may build the approximation one term at a time by selecting at each step the atom which correlates most strongly with the residual signal. Greedy techniques for sparse approximation extend this idea to more general dictionaries.

The algorithm Orthogonal Matching Pursuit (OMP) begins by making a trivial initial approximation $\boldsymbol{a}_0 = \mathbf{0}$. At step k, OMP chooses an atom $\boldsymbol{\varphi}_{\lambda_k}$ that solves the easy optimization problem

$$\max_{\omega} \ |\langle \boldsymbol{s} - \boldsymbol{a}_{k-1}, \boldsymbol{\varphi}_\omega \rangle| \,.$$

In practice, this criterion will often be weakened by choosing a nearly maximal inner product. Suppose that $\Lambda_k = \{\lambda_1, \dots, \lambda_k\}$ lists the atoms that have been selected at step k. Then the k-th approximant $\boldsymbol{a}_k$ solves

$$\min_{\boldsymbol{a}} \ \|\boldsymbol{s} - \boldsymbol{a}\|_2 \qquad \text{subject to} \qquad \boldsymbol{a} \ \in \ \operatorname{span}\{\boldsymbol{\varphi}_\lambda : \lambda \in \Lambda_k\}.$$

This minimization can be performed incrementally with standard least-squares techniques.

One must also supply a stopping rule to decide when the procedure should halt. The following choices are common.

(1) Stop after the approximation contains a specified number of terms.
(2) Stop when the error has declined to a specified level.
(3) Stop as soon as the maximum inner product between an atom and the residual reaches a certain threshold.

Note that the residual always equals zero after d steps. If the dictionary is an orthonormal basis, then $\boldsymbol{a}_m$ is always an m-term approximation with minimal error.

Orthogonal Matching Pursuit was developed independently by many researchers. The earliest reference appears to be a 1989 paper of Chen, Billings and Luo [CBL89]. The first signal processing papers on OMP arrived in 1993 [PRK93, DMZ94]. See the monographs of Miller [Mil02] and Temlyakov [Tem02] for information about other greedy heuristics.

References

[BV04] S. Boyd and L. Vanderberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[CB00] C. Couvreur and Y. Bresler. On the optimality of the Backward Greedy Algorithm for the subset selection problem. *SIAM J. Matrix Anal. Appl.*, 21(3):797–808, 2000.

[CBL89] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *Intl. J. Control*, 50(5):1873–1896, 1989.

[CCKS97] A. R. Calderbank, P. J. Cameron, W. M. Kantor, and J. J. Seidel. $\mathbb{Z}_4$-Kerdock codes, orthogonal spreads and extremal Euclidean line sets. *Proc. London Math. Soc. (3)*, 75(2):436–480, 1997.

[CD94] S. Chen and D. L. Donoho. Basis Pursuit. Statistics Dept. Technical Report, Stanford Univ., 1994.

[CDS99] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, 1999.

[Che95] S. Chen. *Basis Pursuit*. Ph.d. dissertation, Statistics Dept., Stanford Univ., 1995.

[CHS96] J. H. Conway, R. H. Hardin, and N. J. A. Sloane. Packing lines, planes, etc.: Packings in Grassmannian spaces. *Experimental Math.*, 5(2):139–159, 1996.

[CM73] J. F. Claerbout and F. Muir. Robust modeling of erratic data. *Geophysics*, 38(5):826–844, October 1973.

[CM89] R. R. Coifman and Y. Meyer. Nouvelles bases orthonormées de $L^2(\mathbb{R})$ ayant la structure du système de Walsh. Preprint, Mathematics Dept., Yale Univ., 1989.

[CS98] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Number 290 in Grundlehren der mathematischen Wissenschaften. Springer Verlag, 3rd edition, 1998.

[Csi84] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Annals of Probability*, 12:768–793, 1984.

[CW92] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Trans. Inform. Theory*, 1992.

[DDM03] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Submitted for publication, April 2003.

[DE03] D. L. Donoho and M. Elad. Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100:2197–2202, March 2003.

[DET04] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. Working draft, February 2004.

[DeV98] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[DH01] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47:2845–2862, Nov. 2001.

[DJ92] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. Statistics Dept. Technical Report, Stanford Univ., 1992.

[DMA97] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *J. Constr. Approx.*, 13:57–98, 1997.

[DMZ94] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Optical Eng.*, July 1994.

[Dom99] P. Domingos. The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3:409–425, 1999.

[DS89] D. L. Donoho and P. B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49(3):906–931, June 1989.

[DT96] R. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5:173–187, 1996.

[EB02] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 48(9):2558–2567, 2002.

[FN03] A. Feuer and A. Nemirovsky. On sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 49(6):1579–1581, June 2003.

[Fuc97] J.-J. Fuchs. Extension of the Pisarenko Method to sparse linear arrays. *IEEE Trans. Signal Process.*, 45:2413–2421, Oct. 1997.

[Fuc98] J.-J. Fuchs. Estimation and detection of superimposed signals. In *Proc. of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998.

[Fuc02] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. IRISA Technical Report, Univ. de Rennes I, Dec. 2002. Submitted to *IEEE Trans. Inform. Theory*.

[FV03] P. Frossard and P. Vandergheynst. Redundant representations in image processing. In *Proc. of the 2003 IEEE International Conference on Image Processing*, 2003. Special session.

[FViVK04] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura, and M. Kunt. A posteriori quantization of progressive Matching Pursuit streams. *IEEE Trans. Signal Process.*, 52(2):525–535, Feb. 2004.

[GB03] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with Matching Pursuit. *IEEE Trans. Signal Process.*, 51(1):101–111, 2003.

[GG92] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[GH97] M. Grote and T. Huckle. Parallel preconditioning with sparse approximate inverses. *SIAM J. Sci. Comput.*, 18(3):838–853, 1997.

[Gir98] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. *Neural Comput.*, 10(6):1455–1480, 1998.

[GMS03] A. C. Gilbert, M. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, Jan. 2003.

[GN03a] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. Technical report, Aalborg University, Oct. 2003.

[GN03b] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325, Dec. 2003.

[GVL96] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Univ. Press, 3rd edition, 1996.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42:1115–1145, 1995.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, 1985.

[Kre89] E. Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1989.

[LF81] 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):1235–1243, 1981.

[Llo57] S. P. Lloyd. Least squares quantization in PCM. Technical note, Bell Laboratories, 1957.

[LS00] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Comput.*, 12:337–365, 2000.

[LT03] D. Leviatan and V. N. Temlyakov. Simultaneous approximation by greedy algorithms. IMI Report 2003:02, Univ. of South Carolina at Columbia, 2003.

[Mal99] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, London, 2nd edition, 1999.

[Mil02] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 2nd edition, 2002.

[MZ93] S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[Nat95] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1995.

[NZ03] T. Nguyen and A. Zakhor. Matching Pursuits based multiple description video coding for lossy environments. In *Proc. of the 2003 IEEE International Conference on Image Processing*, Barcelona, 2003.

[OF96] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[OSL83] D. W. Oldenburg, T. Scheuer, and S. Levy. Recovery of the acoustic impedance from reflection seismograms. *Geophysics*, 48:1318–1337, 1983.

[PRK93] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal Matching Pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proc. of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, Nov. 1993.

[PS98] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Press, 1998. Corrected republication.

[RB98] B. D. Rao and Y. Bresler. Signal processing with sparseness constraints. In *Proc. of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998. Special session.

[Ris79] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1979.

[RKD99] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Process.*, 47(1):187–200, 1999.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[Roh00] J. Rohn. Computing the norm $\|A\|_{\infty,1}$ is NP-hard. *Linear and Multilinear Algebra*, 47:195–204, 2000.

[Ros98] K. Rose. Deterministing annealing for clustering, compression, classification, regression and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, Nov. 1998.

[SBT00] S. Sardy, A. G. Bruce, and P. Tseng. Block Coordinate Relaxation methods for nonparametric wavelet denoising. *Comp. and Graph. Stat.*, 9(2), 2000.

[Sch07] E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen, I. *Math. Annalen*, 63:433–476, 1906–1907.

[SDC03] J. L. Starck, D. L. Donoho, and E. J. Candès. Astronomical image representation by the Curvelet Transform. *Astronomy and Astrophysics*, 398:785–800, 2003.

[SH03] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comp. Harmonic Anal.*, 14(3):257–275, May 2003.

[Slo02] N. J. A. Sloane. Packing planes in four dimensions and other mysteries, Oct. 2002. Invited talk, Conference on Applied Mathematics, Univ. of Oklahoma at Edmond.

[SO02] P. Sallee and B. A. Olshausen. Learning sparse, multi-scale image representations. In *Adv. Neural Inform. Process.*, 2002.

[SS86] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.*, 7(4):1307–1330, 1986.

[TBM79] H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

[TDHS04] J. A. Tropp, I. S. Dhillon, R. W. Heath, and T. Strohmer. Constructing Grassmannian packings via alternating projection. In preparation, 2004.

[Tem02] V. Temlyakov. Nonlinear methods of approximation. *Foundations of Comp. Math.*, July 2002.

[TGMS03] J. A. Tropp, A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Improved sparse approximation over quasi-incoherent dictionaries. In *Proc. of the 2003 IEEE International Conference on Image Processing*, Barcelona, 2003.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc.*, 58, 1996.

[Tro03a] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. ICES Report 0304, The Univ. of Texas at Austin, Feb. 2003. Accepted to *IEEE Trans. Inform. Theory*.

[Tro03b] J. A. Tropp. Recovery of short, complex linear combinations via ℓ_1 minimization. Unpublished, August 2003.

[Tro04] J. A. Tropp. *Topics in Sparse Approximation*. Ph.d. dissertation, Computational and Applied Mathematics, The Univ. of Texas at Austin, 2004. Expected completion, summer 2004.

[vdB94] A. van den Bos. Complex gradient and Hessian. *IEE Proc.-Vis. Image Signal Process.*, 141(6), Dec. 1994.

[Yap00] C. K. Yap. *Fundamental Problems of Algorithmic Algebra*. Oxford Univ. Press, 2000.