ALGORITHMS FOR SIMULTANEOUS SPARSE APPROXIMATION PART II: CONVEX RELAXATION

JOEL A. TROPP

Abstract. A simultaneous sparse approximation problem requests a good approximation of several input signals at once using different linear combinations of the same elementary signals. At the same time, the problem balances the error in approximation against the total number of elementary signals that participate. These elementary signals typically model coherent structures in the input signals, and they are chosen from a large, linearly dependent collection.

The first part of this paper presents theoretical and numerical results for a greedy pursuit algorithm, called Simultaneous Orthogonal Matching Pursuit.

The second part of the paper develops another algorithmic approach called convex relaxation. This method replaces the combinatorial simultaneous sparse approximation problem with a closely related convex program that can be solved efficiently with standard mathematical programming software. The paper develops conditions under which convex relaxation computes good solutions to simultaneous sparse approximation problems.

Date: Typeset on March 6, 2005.

Key words and phrases. Combinatorial optimization, convex relaxation, multiple measurement vectors, simultaneous sparse approximation.

The author may be reached by e-mail at jtropp@umich.edu or by post at the Department of Mathematics, 2074 East Hall, The University of Michigan, Ann Arbor, MI 48109-1109.

1. Introduction

A *simultaneous sparse approximation problem* has the following shape.

> *Given several input signals, approximate all these signals at once using different linear combinations of the same elementary signals, while balancing the error in approximating the data against the total number of elementary signals that are used.*

A natural example of a simultaneous sparse approximation problem occurs if we have multiple observations of a sparse signal that are contaminated with additive noise. This situation might occur in a communications system where a sparse signal is generated artificially and then transmitted through a noisy channel. One imagines that multiple looks could be used to provide a better estimate of the underlying sparse signal.

The first part of this paper [TGS04] contains some numerical experiments which compare simultaneous sparse approximation against simple sparse approximation. It also proposes a greedy algorithm, Simultaneous Orthogonal Matching Pursuit, for solving simultaneous sparse approximation problems. At each iteration, the greedy algorithm identifies an elementary signal that most improves the current approximation, and then it computes the best approximation of the input signals using all the elementary signals that have been chosen up to that stage. An analysis of the algorithm's theoretical performance shows that it succeeds in solving simultaneous sparse approximation problems when the elementary signals are weakly correlated with each other.

This second part of the paper presents an entirely different algorithmic approach, called *convex relaxation*. Here is one way to motivate this method. Suppose that $\mathbf{\Phi}$ is a matrix whose columns are elementary signals. Imagine that we measure several signals of the form

$$\boldsymbol{s}_k \quad = \quad \mathbf{\Phi}\, \boldsymbol{c}_{\text{opt}} \; + \; \boldsymbol{\nu}_k,$$

where $\boldsymbol{c}_{\text{opt}}$ is a sparse coefficient vector (that is, it has few nonzero entries) and $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_K$ are unknown noise vectors. Form a matrix $\boldsymbol{S}$ whose columns are the measurement vectors, and suppose that we also have a bound on the total energy of the noise, say $\sum \|\boldsymbol{\nu}_k\|_2^2 \leq \varepsilon^2$. One way to recover the ideal coefficient vector $\boldsymbol{c}_{\text{opt}}$ is to solve the mathematical program

$$\min_{\boldsymbol{C}} \, (\text{\# of nonzero rows in } \boldsymbol{C}) \qquad \text{subject to} \qquad \|\boldsymbol{S} - \mathbf{\Phi}\, \boldsymbol{C}\|_{\mathrm{F}} \; \leq \; \varepsilon.$$

By controlling the number of nonzero rows in $\boldsymbol{C}$, we limit the total number of elementary signals that may participate in the approximation of the signal matrix $\boldsymbol{S}$. If the nonzero rows of $\boldsymbol{C}$ correctly identify the nonzero entries of $\boldsymbol{c}_{\text{opt}}$, then we may use linear methods to find an estimate of $\boldsymbol{c}_{\text{opt}}$.

Although this plan is intuitively appealing, it suffers from a tragic flaw. The number of nonzero rows in the coefficient matrix $\boldsymbol{C}$ is a discrete-valued function, and the optimization problem is completely intractable in its current form. Fortunately, there is still hope. We can replace the objective function with a closely related convex function. This step converts the original combinatorial problem into a convex optimization problem that can be solved efficiently.

This paper develops the convex relaxation approach in more detail, and it provides theoretical justification that it actually works. The notational burden is somewhat heavy, so we must postpone a precise description of our results until later.

1.1. **Outline.** In Section 2, we give an overview of our notation. Section 3 develops the idea of convex relaxation in more detail, and it shows how two different formulations of simultaneous sparse approximation lead to two different convex relaxations. The major technical work appears in Section 4. Sections 5 and 6 provide results on the behavior of the two convex relaxations, and they give some examples to demonstrate how the theory can be applied. The final Section 7 shows how this work compares with the literature.

2. Notation

We begin with a very brief introduction to our notation. For more details and interpretation, please refer to Section 2 of [TGS04].

A *signal* is a vector from $\mathbb{C}^d$, and a *signal matrix* is a $d \times K$ complex matrix whose K columns are signals. We use $\|\cdot\|_2$ to denote the Euclidean norm on signals and $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm (aka the Hilbert–Schmidt norm or Schatten 2-norm) on signal matrices. The usual Hermitian inner product on vectors and matrices is written $\langle \cdot, \cdot \rangle$. All the results and derivations hold without modification for real signals. A reader more comfortable with the real setting may prefer to ignore the complications that arise in the complex setting.

A *dictionary* is a collection of unit-Euclidean-norm signals. These signals are called *atoms*, and each atom is written $\boldsymbol{\varphi}_\omega$, where the index is drawn from a set Ω. We form a matrix $\boldsymbol{\Phi}$ in $\mathbb{C}^{d \times \Omega}$ whose columns are atoms. The *coherence parameter* of the dictionary [DMA97, CDS99] is defined as

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

If the coherence parameter is small (e.g., $\mu = O(\sqrt{d})$), we say that the dictionary is *incoherent*. An incoherent dictionary may contain far more atoms than an orthonormal basis. See [Tro04a] for specific examples of incoherent dictionaries.

For simplicity, we will present our results in terms of the coherence parameter instead of the more general cumulative coherence function [DE03, Tro04a]. Note that the coherence parameter is not fundamental to sparse approximation but rather provides a simple way to confirm the hypotheses of our theorems.

Let Λ be a subset of Ω. A *coefficient matrix* is an element of $\mathbb{C}^{\Lambda \times K}$. If $\boldsymbol{C}$ is a coefficient matrix, the product $\boldsymbol{\Phi}\, \boldsymbol{C}$ is a signal matrix. The *row-support* of a coefficient matrix is defined as

$$\operatorname{rowsupp}(\boldsymbol{C}) \quad \stackrel{\text{def}}{=} \quad \{\omega \in \Omega : c_{\omega k} \neq 0 \text{ for some } k\},$$

and the row-ℓ_0 quasi-norm of a coefficient matrix is given by

$$\|\boldsymbol{C}\|_{\text{row-0}} \quad \stackrel{\text{def}}{=} \quad |\operatorname{supp}(\boldsymbol{C})| \,.$$

When the matrix $\boldsymbol{C}$ is a (column) vector, the row-support reduces to the support of the vector, and the row-ℓ_0 quasi-norm reduces to the usual ℓ_0 quasi-norm.

For $p \in [1, \infty]$, we use $\|\cdot\|_p$ to denote the ℓ_p vector norm. The operator norm between ℓ_p and ℓ_q will be written $\|\cdot\|_{p,q}$.

Suppose that Λ indexes a linearly independent collection of atoms. Denote by $\boldsymbol{\Phi}_\Lambda$ the matrix in $\mathbb{C}^{d \times \Lambda}$ whose λ-th column is the atom $\boldsymbol{\varphi}_\lambda$. Give a signal matrix $\boldsymbol{S}$, we denote by $\boldsymbol{A}_\Lambda$ the (unique) best Frobenius-norm approximation of $\boldsymbol{S}$ using the atoms listed in Λ. Let $\boldsymbol{C}_\Lambda$ be the (unique) coefficient matrix in $\mathbb{C}^{\Lambda \times K}$ that synthesizes $\boldsymbol{A}_\Lambda$. That is, $\boldsymbol{A}_\Lambda = \boldsymbol{\Phi}_\Lambda\, \boldsymbol{C}_\Lambda$.

3. Convex Relaxation

3.1. **Combinatorial Optimization.**

It is immediately clear that simultaneous sparse approximation is at least as hard as simple sparse approximation, which is just a special case. It was already established a decade ago that simple sparse approximation is NP-hard in the general case [Nat95, DMA97].

Here is the intuition behind the computationally complexity result. Sparse approximation problems attempt to limit the number of elementary signals that participate in the approximation. For the worst instances of sparse approximation, it is necessary to search through all the possible subcollections of elementary signals to identify the best one. Since the number of different subcollections is exponential in the size of dictionary, the search is intractable for any problem of realistic size.

3.2. **Convex Relaxation.** A beautiful approach to simultaneous sparse approximation is to replace the row-ℓ_0 quasi-norm by a closely related convex function. One hopes that the solution of the convex relaxation will be very close to the solution of the difficult sparse approximation problem. But the relaxation is a convex optimization problem, and so it can be solved in polynomial time by standard mathematical programming software [BV04].

For motivation, consider the case of simple sparse approximation. These problems involve the ℓ_0 quasi-norm of a coefficient *vector*. In this case, convex relaxation replaces the ℓ_0 quasi-norm with the ℓ_1 vector norm [CDS99, Tro04b, DET04]. One rationale is that the ℓ_1 norm is the "smallest" convex function that satisfies the same normalization as the ℓ_0 quasi-norm. See [Tro04d, Sec. 2.1.5] for some discussion of this idea. Another intuition is that minimizing the ℓ_1 norm promotes sparsity whereas minimizing the ℓ_∞ norm promotes nonsparsity.

Let us return to simultaneous sparse approximation. Our intuition is that we want few atoms to participate in the approximation, but we want each atom to contribute to as many columns of the signal matrix as possible. In other words, most rows of the coefficient matrix should be zero, but the nonzero rows should have many nonzero entries. This leads us to consider the following relaxation of the row-ℓ_0 quasi-norm:

$$\|\boldsymbol{B}\|_{\mathrm{rx}} \quad \stackrel{\mathrm{def}}{=} \quad \sum\nolimits_{\omega \in \Omega} \max\nolimits_k \ |b_{\omega k}| \,. \tag{3.1}$$

In words, the norm identifies the maximum absolute entry in each row of $\boldsymbol{B}$ and adds them up. This is equivalent to applying the ℓ_∞ norm to rows (to promote nonsparsity) and then applying the ℓ_1 norm to the resulting vector (to promote sparsity).

Although (3.1) may appear mysterious, it is not far removed from more familiar norms. It is very easy to prove that $\|\cdot\|_{\mathrm{rx}}$ is the dual of the (∞, ∞) matrix norm. Note that, if the matrix $\boldsymbol{B}$ has a single column, the relaxed norm reduces to the ℓ_1 norm. Therefore, the relaxation of a simultaneous sparse approximation problem involving one signal will be the same as the relaxation of the corresponding simple sparse approximation problem. Apparently, the norm $\|\cdot\|_{\mathrm{rx}}$ also appears in the literature on interpolation spaces [BL76].

Other relaxations of the row-ℓ_0 quasi-norm are certainly possible, as we will discuss in Section 7. Unfortunately, there are few theoretical results available for these other relaxations. It would be very interesting to perform numerical experiments to compare the behavior of these different approaches.

3.3. **Relaxed Simultaneous Sparse Approximation.** Suppose that $\boldsymbol{S}$ is a signal matrix whose columns admit a good simultaneous approximation over the dictionary $\boldsymbol{\Phi}$. The most natural way to formulate the search for this simultaneous sparse approximation of $\boldsymbol{S}$ is

$$\min_{\boldsymbol{C} \in \mathbb{C}^{\Omega \times K}} \quad \|\boldsymbol{S} - \boldsymbol{\Phi}\,\boldsymbol{C}\|_{\mathrm{F}}^2 \ + \ \frac{\tau}{K}\ \|\boldsymbol{C}\|_{\text{row-0}} \,.$$

The parameter τ balances the approximation error against the number of nonzero rows in the coefficient matrix $\boldsymbol{C}$. As we have discussed, this problem is not tractable in its current form. Relaxing this combinatorial problem leads to the convex program

$$\min_{\boldsymbol{B} \in \mathbb{C}^{\Omega \times K}} \quad \tfrac{1}{2}\ \|\boldsymbol{S} - \boldsymbol{\Phi}\,\boldsymbol{B}\|_{\mathrm{F}}^2 \ + \ \gamma\ \|\boldsymbol{B}\|_{\mathrm{rx}} \,. \tag{RX-PENALTY}$$

Here the parameter γ negotiates a compromise between approximation error and row-sparsity. (The change of normalization in the convex program is inessential but very convenient.) In Section 5, we will prove that the (RX-PENALTY) offers an effective approach to simultaneous sparse approximation. Indeed, if $\boldsymbol{S}$ has a good approximation using the atoms in Λ, then the solution to (RX-PENALTY) will identify all the significant atoms in the approximation, and it will never identify an atom outside Λ.

A second formulation of simultaneous sparse approximation is the mathematical program described in the Introduction:

$$\min_{\boldsymbol{C} \in \mathbb{C}^{\Omega \times K}} \|\boldsymbol{C}\|_{\text{row-0}} \qquad \text{subject to} \qquad \|\boldsymbol{S} - \boldsymbol{\Phi}\,\boldsymbol{C}\|_{\text{F}} \;\leq\; \varepsilon.$$

In words, this problem requests the sparsest approximation of $\boldsymbol{S}$ that achieves an error no greater than ε. The convex relaxation of this combinatorial problem is

$$\min_{\boldsymbol{B} \in \mathbb{C}^{\Omega \times K}} \|\boldsymbol{B}\|_{\text{rx}} \qquad \text{subject to} \qquad \|\boldsymbol{S} - \boldsymbol{\Phi}\,\boldsymbol{B}\|_{\text{F}} \;\leq\; \delta. \qquad (\text{RX-ERROR})$$

In Section 6, we will prove that (RX-ERROR) also yields an effective approach to simultaneous sparse approximation. But our results suggest that the first convex relaxation is more powerful.

Although it is possible to develop interesting relationships between the solutions to the combinatorial problems and the solutions to the relaxations, we will not follow this route. Instead, we will demonstrate that the convex relaxations themselves can be used to solve simultaneous sparse approximation problems.

4. Lemmata

In this section, we will be studying minimizers of the convex function

$$L(\boldsymbol{B}) \quad = \quad \tfrac{1}{2}\,\|\boldsymbol{S} - \boldsymbol{\Phi}\,\boldsymbol{B}\|_{\text{F}}^2 \;+\; \gamma\,\|\boldsymbol{B}\|_{\text{rx}}\,. \qquad (4.1)$$

This is the objective function of (RX-PENALTY), and it is essentially the Lagrangian function of (RX-ERROR). If we understand the minimizers of this function, we will understand the performance of our convex relaxations. The casual reader can skip this section without any loss of continuity.

We will develop a sufficient condition for the minimizer of the function (4.1) to be supported on a given index set Λ. The proof of this condition is based primarily on the work in [Tro04b]. First, we characterize the unique minimizer of the objective function, when it is restricted to matrices supported on Λ. We use this characterization to show that any perturbation away from this restricted minimizer must increase the objective function. The idea of using a perturbation argument has appeared in Section 4.3 of [Tro04c], and it is also an element of the argument in [DET04]. The present approach is cleaner and more powerful than either of its precedents, and it provides a route toward studying several important extensions of simple and simultaneous sparse approximation.

4.1. **Convex Analysis.**

The proof relies on standard results from convex analysis. As it is usually presented, this subject addresses the properties of real-valued convex functions defined on real vector spaces. Nevertheless, it is possible to transport these results to the complex setting by defining an appropriate real-linear structure on the complex vector space. In this section, therefore, we use the bilinear inner product $\operatorname{Re}\langle \boldsymbol{X},\ \boldsymbol{Y}\rangle = \operatorname{Re}\operatorname{trace}(\boldsymbol{Y}^*\,\boldsymbol{X})$ instead of the usual sesquilinear (i.e., Hermitian) inner product. Note that both inner products generate the Frobenius norm.

If f is a convex function from a complex matrix space $\mathbb{M}$ to $\mathbb{R}$, then we define the *gradient* $\nabla f(\boldsymbol{X})$ as the usual (Fréchet) derivative of f at $\boldsymbol{X}$, computed with respect to the real inner product. The *subdifferential* of f at a matrix $\boldsymbol{X}$ is defined as

$$\partial f(\boldsymbol{X}) \quad \stackrel{\text{def}}{=} \quad \{\boldsymbol{G} \in \mathbb{M} : f(\boldsymbol{Y}) \geq f(\boldsymbol{X}) + \operatorname{Re}\langle \boldsymbol{Y} - \boldsymbol{X},\ \boldsymbol{G}\rangle \quad \text{for every } \boldsymbol{Y} \in \mathbb{M}\}.$$

The elements of the subdifferential are called *subgradients*. If f has a well-defined gradient at $\boldsymbol{X}$, the unique subgradient is the gradient. That is,

$$\partial f(\boldsymbol{X}) \quad = \quad \{\nabla f(\boldsymbol{X})\}.$$

The subdifferential of a sum is the (Minkowski) sum of the subdifferentials. Finally, if f is a (closed, proper) convex function then $\boldsymbol{X}$ is a minimizer of f if and only if $\mathbf{0} \in \partial f(\boldsymbol{X})$. The standard reference on this subject is [Roc70].

Remark 4.1. The subdifferential of a convex function provides a dual description of the function in terms of its supporting hyperplanes. In consequence, the appearance of the subdifferential in our proof is analogous with the familiar technique of studying the dual of a convex program.

4.2. **Restricted Minimizers.** Our first goal is to characterize the minimizers of the objective function (4.1). The following result generalizes Lemma 3 from [Tro04b] to the matrix case. The proof is identical.

Lemma 4.2. *Suppose that the matrix $\boldsymbol{B}_\star$ minimizes the objective function* (4.1) *over all coefficient matrices supported on Λ. A necessary and sufficient condition on such a minimizer is that*

$$\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star \quad = \quad \gamma\,({\boldsymbol{\Phi}_\Lambda}^*\,\boldsymbol{\Phi}_\Lambda)^{-1}\,\boldsymbol{G} \tag{4.2}$$

where $\boldsymbol{G}$ is drawn from the subdifferential $\partial\,\|\boldsymbol{B}_\star\|_{\mathrm{rx}}$. Moreover, the minimizer is unique.

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

$$F(\boldsymbol{B}) \quad \stackrel{\text{def}}{=} \quad \tfrac{1}{2}\,\|\boldsymbol{A}_\Lambda - \boldsymbol{\Phi}_\Lambda\,\boldsymbol{C}\|_{\mathrm{F}}^2 \;+\; \gamma\,\|\boldsymbol{B}\|_{\mathrm{rx}} \tag{4.3}$$

over matrices from $\mathbb{C}^{\Lambda\times K}$. 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 (4.3) 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 $\boldsymbol{0} \in \partial F(\boldsymbol{B}_\star)$ is a necessary and sufficient condition for $\boldsymbol{B}_\star$ to minimize F. The 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

$$({\boldsymbol{\Phi}_\Lambda}^*\,\boldsymbol{\Phi}_\Lambda)\,\boldsymbol{B}_\star - {\boldsymbol{\Phi}_\Lambda}^*\,\boldsymbol{A}_\Lambda + \gamma\,\boldsymbol{G} \quad = \quad \boldsymbol{0}$$

for some vector $\boldsymbol{G}$ drawn from the subdifferential $\partial\,\|\boldsymbol{B}_\star\|_{\mathrm{rx}}$. 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

$${\boldsymbol{\Phi}_\Lambda}^\dagger\,\boldsymbol{A}_\Lambda - \boldsymbol{B}_\star \quad = \quad \gamma\,({\boldsymbol{\Phi}_\Lambda}^*\,\boldsymbol{\Phi}_\Lambda)^{-1}\,\boldsymbol{G}.$$

Apply the fact that $\boldsymbol{C}_\Lambda = {\boldsymbol{\Phi}_\Lambda}^\dagger\,\boldsymbol{A}_\Lambda$ to reach the conclusion. ☐

Next, we identify the subdifferential of the relaxed norm $\|\cdot\|_{\mathrm{rx}}$. To that end, define the signum function as

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

The notation $\operatorname{conv}(S)$ indicates the convex hull of a set S.

Lemma 4.3. *A matrix $\boldsymbol{G}$ is a subgradient of $\|\cdot\|_{\mathrm{rx}}$ at the matrix $\boldsymbol{B}$ if and only if the ω-th row of $\boldsymbol{G}$ satisfies*

$${\mathbf{e}_\omega}^*\,\boldsymbol{G} \quad \in \quad \begin{cases} \{\boldsymbol{g}^* : \|\boldsymbol{g}\|_1 \le 1\} & \textit{if } b_{\omega k} = 0 \textit{ for each } k, \textit{ and} \\ \operatorname{conv}\{(\operatorname{sgn} b_{\omega k})\,{\mathbf{e}_k}^* : |b_{\omega k}| = \max_j |b_{\omega j}|\} & \textit{otherwise.} \end{cases}$$

In particular, $\|\boldsymbol{G}\|_{\infty,\infty} \le 1$ for any subgradient $\boldsymbol{G}$ of $\|\cdot\|_{\mathrm{rx}}$.

In words, we determine the rows of a subgradient of $\boldsymbol{B}$ in the following manner. In each nonzero row of $\boldsymbol{B}$, identify the component(s) that have the largest absolute value. The corresponding row of the subgradient must fall in the convex hull of the (signed) canonical basis vectors corresponding with these components. If a row of $\boldsymbol{B}$ equals zero, the corresponding row of the subgradient must lie in the ℓ_1 unit ball.

Proof. For a matrix $\boldsymbol{B}$ in $\mathbb{C}^{\Lambda \times K}$, the relaxed norm is computed as

$$\|\boldsymbol{B}\|_{\mathrm{rx}} \quad = \quad \sum\nolimits_{\lambda \in \Lambda} \max_k \; |b_{\lambda k}| \,.$$

By definition, a matrix $\boldsymbol{G}$ lies in $\partial \|\boldsymbol{B}\|_{\mathrm{rx}}$ if and only if

$$\|\boldsymbol{Z}\|_{\mathrm{rx}} \quad \geq \quad \|\boldsymbol{B}\|_{\mathrm{rx}} \;+\; \operatorname{Re} \langle \boldsymbol{Z} - \boldsymbol{B},\; \boldsymbol{G} \rangle$$

for every matrix $\boldsymbol{Z}$ in $\mathbb{C}^{\Lambda \times K}$. Now, we use the fact that the relaxed norm and the bilinear inner product are both row-separable to see that each row of $\boldsymbol{G}$ must belong to the subdifferential of the ℓ_∞ norm at the corresponding row of $\boldsymbol{B}$. It is well known that

$$\partial \|\boldsymbol{b}\|_\infty \quad = \quad \begin{cases} \{\boldsymbol{g} : \|\boldsymbol{g}\|_1 \leq 1\} & \text{if } \boldsymbol{b} = \boldsymbol{0}, \text{ and} \\ \operatorname{conv}\{(\operatorname{sgn} b_k)\, \mathbf{e}_k : |b_k| = \max_j |b_j|\} & \text{otherwise.} \end{cases}$$

The formula for $\partial \|\boldsymbol{B}\|_{\mathrm{rx}}$ just applies the ℓ_∞ subdifferential to each row of $\boldsymbol{B}$. □

In consequence, we may develop bounds on how much a solution to the restricted problem varies from the desired solution $\boldsymbol{C}_\Lambda$. The proof is essentially the same as that of Corollary 5 from [Tro04b].

Corollary 4.4 (Upper Bounds)**.** *Suppose that the matrix* $\boldsymbol{B}_\star$ *minimizes the function* (4.1) *over all coefficient matrices supported on* Λ. *The following bounds are in force:*

$$\|\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star\|_{\infty,\infty} \quad \leq \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^* \boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty} \tag{4.4}$$

$$\|\boldsymbol{\Phi}_\Lambda\, (\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star)\|_{\mathrm{F}} \quad \leq \quad \gamma \left\|\boldsymbol{\Phi}_\Lambda{}^\dagger\right\|_{2,1}. \tag{4.5}$$

Proof. We begin with the necessary and sufficient condition

$$\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star \quad = \quad \gamma\, (\boldsymbol{\Phi}_\Lambda{}^*\, \boldsymbol{\Phi}_\Lambda)^{-1}\, \boldsymbol{G} \tag{4.6}$$

where $\boldsymbol{G} \in \partial \|\boldsymbol{B}_\star\|_{\mathrm{rx}}$. To obtain (4.4), we take the (∞, ∞) norm of (4.6) and apply the usual estimate:

$$\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\infty,\infty} \quad = \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^*\, \boldsymbol{\Phi}_\Lambda)^{-1}\, \boldsymbol{G}\right\|_{\infty,\infty} \quad \leq \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^*\, \boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty} \|\boldsymbol{G}\|_{\infty,\infty}.$$

Lemma 4.3 shows that $\|\boldsymbol{G}\|_{\infty,\infty} \leq 1$, which proves the result.

To develop the second bound (4.5), left-multiply (4.6) by the matrix $\boldsymbol{\Phi}_\Lambda$ and compute the Frobenius norm:

$$\|\boldsymbol{\Phi}_\Lambda\, (\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda)\|_{\mathrm{F}} \quad = \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^\dagger)^*\, \boldsymbol{G}\right\|_{\mathrm{F}} \quad \leq \quad \gamma \left\|(\boldsymbol{\Phi}_\Lambda{}^\dagger)^*\right\|_{\infty,2} \|\boldsymbol{G}\|_{\infty,\infty}.$$

(Note that the norm bound here is not trivial to establish. One approach is to apply Proposition 2.1 from [TGS04] use this characterization to show that $\|\cdot\|_{(\infty,\infty)\to\mathrm{F}} = \|\cdot\|_{\infty,2}$.) To complete the argument, recall that $\|\boldsymbol{G}\|_{\infty,\infty} \leq 1$. Finally, we use adjointness to switch from the $(\infty, 2)$ operator norm to the $(2, 1)$ operator norm. □

4.3. **Perturbation Theory.** Now, we develop a sufficient condition for the minimizer of (4.1) to be supported inside the index set Λ. This result is the key to our analysis.

Lemma 4.5 (Correlation Condition)**.** *A sufficient condition for the minimizer* $\boldsymbol{B}_\star$ *of the objective function* (4.1) *to be supported on* Λ *is that*

$$\|\boldsymbol{\Phi}^*\, (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \quad \leq \quad \gamma \left[1 \;-\; \max_{\omega \notin \Lambda} \left\|\boldsymbol{G}^*\, \boldsymbol{\Phi}_\Lambda{}^\dagger\, \boldsymbol{\varphi}_\omega\right\|_1\right]$$

where $\boldsymbol{G} \in \partial \|\boldsymbol{B}_\star\|_{\mathrm{rx}}$. *In particular, it sufficient that*

$$\|\boldsymbol{\Phi}^*\, (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \quad \leq \quad \gamma \left[1 \;-\; \max_{\omega \notin \Lambda} \left\|\boldsymbol{\Phi}_\Lambda{}^\dagger\, \boldsymbol{\varphi}_\omega\right\|_1\right].$$

We will typically abbreviate the bracket in the second sufficient condition:

$$\mathrm{ERC}(\Lambda) \quad \stackrel{\text{def}}{=} \quad 1 \; - \; \max_{\omega \notin \Lambda} \left\| \mathbf{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega \right\|_1. \tag{4.7}$$

For a geometric description of this quantity, see Section 3.6 of [Tro04c].

Proof. Let $\boldsymbol{B}_\star$ be the minimizer of the objective function (4.1) over coefficient matrices supported on Λ. We will develop a condition which guarantees that the objective function increases when $\boldsymbol{B}_\star$ is perturbed so that its support extends outside Λ. Since the objective function is convex and unconstrained, this fact implies that the minimizer must be supported on Λ. Therefore, $\boldsymbol{B}_\star$ is the global minimizer of the objective function.

Choose an index ω that is not contained in Λ, and let $\boldsymbol{\delta}$ be a nonzero K-dimensional column vector. We must develop a condition which ensures that

$$L(\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*) \; - \; L(\boldsymbol{B}_\star) \quad > \quad 0.$$

To that end, we expand the left-hand side of this equation to obtain

$$L(\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*) \; - \; L(\boldsymbol{B}_\star) \quad = \quad \tfrac{1}{2} \left[\|(\boldsymbol{S} - \mathbf{\Phi} \, \boldsymbol{B}_\star) - \boldsymbol{\varphi}_\omega \, \boldsymbol{\delta}^*\|_{\mathrm{F}}^2 \; - \; \|\boldsymbol{S} - \mathbf{\Phi} \, \boldsymbol{B}_\star\|_{\mathrm{F}}^2 \right] \; + \; \gamma \left[\|\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*\|_{\mathrm{rx}} \; - \; \|\boldsymbol{B}_\star\|_{\mathrm{rx}} \right].$$

We denote by $\mathbf{e}_\omega$ the ω-th canonical basis vector in $\mathbb{C}^\Omega$. Next, simplify the first bracket by expanding the Frobenius norms and canceling like terms. Simplify the second bracket by applying the definition of the relaxed norm and recognizing that the row-support of $\boldsymbol{B}_\star$ is disjoint from the row-support of $\mathbf{e}_\omega \, \boldsymbol{\delta}^*$.

$$L(\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*) \; - \; L(\boldsymbol{B}_\star) \quad = \quad \tfrac{1}{2} \, \|\boldsymbol{\varphi}_\omega \, \boldsymbol{\delta}^*\|_{\mathrm{F}}^2 \; - \; \mathrm{Re} \, \langle \boldsymbol{S} - \mathbf{\Phi} \, \boldsymbol{B}_\star, \; \boldsymbol{\varphi}_\omega \, \boldsymbol{\delta}^* \rangle \; + \; \gamma \, \|\boldsymbol{\delta}\|_\infty .$$

Next, transfer the atom $\boldsymbol{\varphi}_\omega$ to the other side of the inner product via adjointness. Then add and subtract $\mathbf{\Phi}_\Lambda \, \boldsymbol{C}_\Lambda$ on the left-hand side of the inner product, and use linearity to split the inner product into two pieces. We reach

$$L(\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*) \; - \; L(\boldsymbol{B}_\star) \quad = \quad \tfrac{1}{2} \, \|\boldsymbol{\varphi}_\omega \, \boldsymbol{\delta}^*\|_{\mathrm{F}}^2 \; - \; \mathrm{Re} \, \langle \boldsymbol{\varphi}_\omega{}^* \, (\boldsymbol{S} - \mathbf{\Phi}_\Lambda \, \boldsymbol{C}_\Lambda), \; \boldsymbol{\delta}^* \rangle \; - \; \mathrm{Re} \, \langle \boldsymbol{\varphi}_\omega{}^* \, \mathbf{\Phi}_\Lambda \, (\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star), \; \boldsymbol{\delta}^* \rangle \; + \; \gamma \, \|\boldsymbol{\delta}\|_\infty .$$

We will bound the right-hand side below. To that end, observe that the first term is strictly positive, and invoke the lower triangle inequality.

$$L(\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*) \; - \; L(\boldsymbol{B}_\star) \quad > \quad \gamma \, \|\boldsymbol{\delta}\|_\infty \; - \; |\langle \boldsymbol{\varphi}_\omega{}^* \, (\boldsymbol{S} - \mathbf{\Phi}_\Lambda \, \boldsymbol{C}_\Lambda), \; \boldsymbol{\delta}^* \rangle| \; - \; |\langle \boldsymbol{\varphi}_\omega{}^* \, \mathbf{\Phi}_\Lambda \, (\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star), \; \boldsymbol{\delta}^* \rangle| . \tag{4.8}$$

It remains to inspect the right-hand side of (4.8).

Let us examine the second of the three terms on the right-hand side of (4.8). Identify $\boldsymbol{A}_\Lambda = \mathbf{\Phi}_\Lambda \, \boldsymbol{C}_\Lambda$, and then apply Hölder's Inequality to obtain

$$|\langle \boldsymbol{\varphi}_\omega{}^* \, (\boldsymbol{S} - \mathbf{\Phi}_\Lambda \, \boldsymbol{C}_\Lambda), \; \boldsymbol{\delta}^* \rangle| \quad \leq \quad \|(\boldsymbol{S} - \boldsymbol{A}_\Lambda)^* \, \boldsymbol{\varphi}_\omega\|_1 \, \|\boldsymbol{\delta}\|_\infty . \tag{4.9}$$

We continue with the third term on the right-hand side of (4.8). Lemma 4.2 characterizes the difference $(\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star)$. Introduce this characterization, and identify the pseudo-inverse of $\mathbf{\Phi}_\Lambda$ to discover that

$$|\langle \boldsymbol{\varphi}_\omega{}^* \, \mathbf{\Phi}_\Lambda \, (\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star), \; \boldsymbol{\delta}^* \rangle| \quad = \quad \gamma \left| \left\langle \boldsymbol{\varphi}_\omega{}^* \, (\mathbf{\Phi}_\Lambda{}^\dagger)^* \, \boldsymbol{G}, \; \boldsymbol{\delta}^* \right\rangle \right|$$

where $\boldsymbol{G} \in \partial \, \|\boldsymbol{B}_\star\|_{\mathrm{rx}}$. Apply Hölder's Inequality again.

$$|\langle \boldsymbol{\varphi}_\omega{}^* \, \mathbf{\Phi}_\Lambda \, (\boldsymbol{C}_\Lambda - \boldsymbol{B}_\star), \; \boldsymbol{\delta}^* \rangle| \quad \leq \quad \gamma \left\| \boldsymbol{G}^* \, \mathbf{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega \right\|_1 \|\boldsymbol{\delta}\|_\infty . \tag{4.10}$$

Now we introduce the bounds (4.9) and (4.10) into the bound (4.8).

$$L(\boldsymbol{B}_\star + \mathbf{e}_\omega \, \boldsymbol{\delta}^*) \; - \; L(\boldsymbol{B}_\star) \quad > \quad \left[\gamma \; - \; \|(\boldsymbol{S} - \boldsymbol{A}_\Lambda)^* \, \boldsymbol{\varphi}_\omega\|_1 \; - \; \gamma \left\| \boldsymbol{G}^* \, \mathbf{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega \right\|_1 \right] \|\boldsymbol{\delta}\|_\infty . \tag{4.11}$$

Our final goal is to ensure that the left-hand side of this relation is strictly positive.

The left-hand side of (4.11) is positive whenever the bracket is nonnegative. Therefore, we need

$$\|(\boldsymbol{S} - \boldsymbol{A}_\Lambda)^* \boldsymbol{\varphi}_\omega\|_1 \quad \leq \quad \gamma \left[1 \; - \; \left\|\boldsymbol{G}^* \, \boldsymbol{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega\right\|_1\right].$$

This expression must hold for every index ω that does not belong to Λ. Optimizing both sides over ω in $\Omega \setminus \Lambda$, we reach the stronger condition

$$\max_{\omega \notin \Lambda} \|(\boldsymbol{S} - \boldsymbol{A}_\Lambda)^* \boldsymbol{\varphi}_\omega\|_1 \quad \leq \quad \gamma \left[1 \; - \; \max_{\omega \notin \Lambda} \left\|\boldsymbol{G}^* \, \boldsymbol{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega\right\|_1\right].$$

Since $(\boldsymbol{S} - \boldsymbol{A}_\Lambda)$ is orthogonal to the atoms listed in Λ, the left-hand side does not change if we maximize over all ω from Ω. We rewrite the left-hand side to conclude that the relation

$$\|\boldsymbol{\Phi}^* \, (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \quad \leq \quad \gamma \left[1 \; - \; \max_{\omega \notin \Lambda} \left\|\boldsymbol{G}^* \, \boldsymbol{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega\right\|_1\right]$$

is a sufficient condition for every perturbation away from $\boldsymbol{B}_\star$ to increase the objective function L. In particular, since $\|\boldsymbol{G}\|_{\infty,\infty} \leq 1$, it is also sufficient that

$$\|\boldsymbol{\Phi}^* \, (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \quad \leq \quad \gamma \left[1 \; - \; \max_{\omega \notin \Lambda} \left\|\boldsymbol{\Phi}_\Lambda{}^\dagger \, \boldsymbol{\varphi}_\omega\right\|_1\right].$$

This completes the argument. □

5. The Penalty Method

Suppose that $\boldsymbol{S}$ is an observed signal matrix, and we wish to approximate the columns of $\boldsymbol{S}$ simultaneously using linear combinations of the same atoms from $\boldsymbol{\Phi}$. In this section, we approach this problem by means of the convex program

$$\min_{\boldsymbol{B} \in \mathbb{C}^{\Omega \times K}} \quad \tfrac{1}{2} \, \|\boldsymbol{S} - \boldsymbol{\Phi} \, \boldsymbol{B}\|_{\mathrm{F}}^2 \; + \; \gamma \, \|\boldsymbol{B}\|_{\mathrm{rx}} \, . \qquad (\text{RX-PENALTY})$$

Intuitively, the parameter γ negotiates a compromise between the approximation error and the level of sparsity. If the signal matrix contains a single column $\boldsymbol{s}$, observe that (RX-PENALTY) reduces to

$$\min_{\boldsymbol{b} \in \mathbb{C}^{\Omega}} \quad \tfrac{1}{2} \, \|\boldsymbol{s} - \boldsymbol{\Phi} \, \boldsymbol{b}\|_2^2 \; + \; \gamma \, \|\boldsymbol{b}\|_1 \, . \qquad (\ell_1\text{-PENALTY})$$

This method for simple sparse approximation has been studied in [Fuc04c, Fuc04b, Tro04b].

In this section, we will present a theorem that describes the behavior of the minimizer of (RX-PENALTY). Afterward, we will show how to apply this theorem to the problem of recovering a sparse signal from multiple observations that have been contaminated with additive noise. Other related examples for the case of simple sparse approximation are available in [Tro04b].

5.1. **Performance of the Relaxation.** Our major theorem on the behavior of (RX-PENALTY) follows immediately from the lemmata in the last section. The definition of $\mathrm{ERC}(\Lambda)$ appears in equation (4.7).

Theorem 5.1. *Let Λ index a linearly independent collection of atoms for which* $\mathrm{ERC}(\Lambda) \geq 0$. *Suppose that $\boldsymbol{S}$ is a signal matrix whose best approximation $\boldsymbol{A}_\Lambda$ over Λ satisfies the condition*

$$\|\boldsymbol{\Phi}^* \, (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \quad \leq \quad \gamma \, \mathrm{ERC}(\Lambda)$$

Define $\boldsymbol{B}_\star$ to be the unique solution to (RX-PENALTY) *with parameter γ. We may conclude that*

- *the row-support of $\boldsymbol{B}_\star$ is contained in Λ, and*
- *the distance between $\boldsymbol{B}_\star$ and the optimal coefficient matrix $\boldsymbol{C}_\Lambda$ satisfies*

$$\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\infty,\infty} \quad \leq \quad \gamma \, \left\|(\boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{\Phi}_\Lambda)^{-1}\right\|_{\infty,\infty} .$$

- *In particular,* rowsupp$(\boldsymbol{B}_\star)$ *contains every index* λ *from* Λ *for which*

$$\sum_{k=1}^{K} |\boldsymbol{C}_\Lambda(\lambda, k)| \quad > \quad \gamma \left\| (\boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{\Phi}_\Lambda)^{-1} \right\|_{\infty,\infty}.$$

That is, we approximate the signal matrix over Λ, and we assume that the remaining atoms are weakly correlated with the residual. If so, the unique solution to (rx-penalty) identifies every atom that participates significantly in $\boldsymbol{A}_\Lambda$, and it never selects an atom outside Λ. Note that the uniqueness of the minimizer is one of the conclusions of the theorem.

To use this theorem, it is necessary to leverage information about the problem domain to determine an appropriate choice for the set Λ. In the sequel, we will show how to apply the theorem to a couple of specific cases. One can also use the coherence parameter to simplify the statement of the theorem. This formulation has the advantage that the index set Λ plays a smaller role.

Corollary 5.2. *Suppose that* $m\,\mu \leq \frac{1}{2}$ *and that* Λ *contains no more than* m *indices. Suppose that* $\boldsymbol{S}$ *is a signal matrix whose best approximation over* Λ *satisfies*

$$\left\| \boldsymbol{\Phi}^* \left(\boldsymbol{S} - \boldsymbol{A}_\Lambda \right) \right\|_{\infty,\infty} \quad \leq \quad \gamma \, \frac{1 \; - \; (2m - 1)\, \mu}{1 \; - \; (m - 1)\, \mu}.$$

Define $\boldsymbol{B}_\star$ *to be the unique solution to* (rx-penalty) *with parameter* γ. *We may conclude that*

- *the row-support of* $\boldsymbol{B}_\star$ *is contained in* Λ, *and*
- *the distance between* $\boldsymbol{B}_\star$ *and* $\boldsymbol{C}_\Lambda$ *satisfies*

$$\left\| \boldsymbol{B}_\star - \boldsymbol{C}_\Lambda \right\|_{\infty,\infty} \quad \leq \quad \frac{\gamma}{1 \; - \; (m-1)\, \mu}.$$

- *In particular,* rowsupp$(\boldsymbol{B}_\star)$ *contains every index* λ *from* Λ *for which*

$$\sum_{k=1}^{K} |\boldsymbol{C}_\Lambda(\lambda, k)| \quad > \quad \frac{\gamma}{1 \; - \; (m-1)\, \mu}.$$

Proof. The result follows immediately from the coherence estimates of [Tro04c, Sec. 3]. ☐

The corollary takes a particularly attractive form if we assume that $m\,\mu \leq \frac{1}{3}$. Then the sufficient condition becomes $\left\| \boldsymbol{\Phi}^* \left(\boldsymbol{S} - \boldsymbol{A}_\Lambda \right) \right\|_{\infty,\infty} \leq \gamma/2$, and the distance between the coefficient vectors satisfies $\left\| \boldsymbol{B}_\star - \boldsymbol{C}_\Lambda \right\|_{\infty,\infty} \leq 3\,\gamma/2$.

5.2. **Comparison with Simple Sparse Approximation.** It may be illuminating to compare Theorem 5.1 against the analogous result for simple sparse approximation [Tro04b, Thm. 8]. First, we note that Theorem 5.1 reduces to the original result for simple sparse approximation if the signal matrix $\boldsymbol{S}$ has a single column. Therefore, nothing has been lost in the analysis.

Still, there are some qualitative differences. In the case of simultaneous sparse approximation, we expect that the left-hand side of the condition

$$\left\| \boldsymbol{\Phi}^* \left(\boldsymbol{S} - \boldsymbol{A}_\Lambda \right) \right\|_{\infty,\infty} \quad \leq \quad \gamma \; \mathrm{ERC}(\Lambda)$$

will be somewhat larger than it is in the case of simple sparse approximation. As a result, it may be necessary to increase the value of the parameter γ to ensure that the condition is in force. In compensation, we also expect that

$$\sum_{k=1}^{K} |\boldsymbol{C}_\Lambda(\lambda, k)|$$

will be somewhat larger than it is in the simple case. As a result, it should be easier for an index λ to lie in the row-support of $\boldsymbol{B}_\star$. To determine whether simultaneous sparse approximation is valuable for a given problem domain, one must check which of the two effects dominates.

5.3. **Sparse Representation of Signal Matrices.** We obtain an important corollary when the signal matrix $\boldsymbol{S}$ can be represented exactly using the atoms in Λ. This result provides a sufficient condition for convex relaxation to identify the sparsest exact representation of a signal matrix. The analogous result for simple sparse approximation is due to Fuchs [Fuc04a]. See also [Tro04b, Cor. 10].

Corollary 5.3. *Assume that* Λ *lists a linearly independent collection of atoms for which* $\mathrm{ERC}(\Lambda) > 0$. *Choose an arbitrary coefficient matrix* $\boldsymbol{C}_\Lambda$ *who row-support equals* Λ, *and fix a signal matrix* $\boldsymbol{S} = \boldsymbol{\Phi}\, \boldsymbol{C}_\Lambda$. *Let* $\boldsymbol{B}_\star(\gamma)$ *denote the unique minimizer of* (RX-PENALTY) *with parameter* γ. *We may conclude that*

- *there is a positive number* γ_0 *for which* $\gamma < \gamma_0$ *implies that* $\mathrm{rowsupp}(\boldsymbol{B}_\star(\gamma)) = \Lambda$.
- *As* $\gamma \downarrow 0$, *we have* $\boldsymbol{B}_\star(\gamma) \to \boldsymbol{C}_\Lambda$.

Proof. Since $\boldsymbol{S}$ can be expressed using the atoms in Λ, its best approximation $\boldsymbol{A}_\Lambda = \boldsymbol{S}$. Therefore, $\|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} = 0$. It follows from Theorem 5.1 that the minimizer $\boldsymbol{B}_\star(\gamma)$ is unique for every positive number γ. Moreover,

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

It follows immediately that $\boldsymbol{B}_\star(\gamma) \to \boldsymbol{C}_\Lambda$ as the parameter $\gamma \downarrow 0$. Finally, observe that $\mathrm{rowsupp}(\boldsymbol{B}_\star(\gamma))$ contains every index Λ provided that

$$\frac{\min_{\lambda \in \Lambda} \ \sum_{k=1}^{K} |\boldsymbol{C}_\Lambda(\lambda, k)|}{\|(\boldsymbol{\Phi}_\Lambda{}^* \, \boldsymbol{\Phi}_\Lambda)^{-1}\|_{\infty,\infty}} \quad > \quad \gamma.$$

The left-hand side of this equation furnishes an explicit value for γ_0. □

It is also important to recognize that the limit of the solutions to (4.1) as the parameter γ declines to zero will solve the problem

$$\min_{\boldsymbol{B} \in \mathbb{C}^{\Omega \times K}} \ \|\boldsymbol{B}\|_{\mathrm{rx}} \qquad \text{subject to} \qquad \boldsymbol{S} = \boldsymbol{\Phi}\, \boldsymbol{B}.$$

The converse, however, is not necessary true. One may wish to compare this result with [CH04b, Thm. 3.8].

5.4. **Example: Identifying Noisy Sparse Signals.** The most natural application of Theorem 5.1 is to the case where the columns of the signal matrix $\boldsymbol{S}$ are multiple observations of an ideal signal that have been corrupted with additive noise. In this subsection, we will present a basic example to demonstrate how the theorem applies.

Let us begin with a model for our ideal signals. Suppose that $\boldsymbol{\Phi}$ is a dictionary with coherence μ, and suppose that $m\,\mu \leq \frac{1}{3}$. We will form ideal signals by taking linear combinations of m atoms from $\boldsymbol{\Phi}$ with all coefficients equal to one. More formally, suppose that $\boldsymbol{c}_{\mathrm{opt}}$ is a vector whose m nonzero entries equal one. Then each ideal signal has the form $\boldsymbol{\Phi}\, \boldsymbol{c}_{\mathrm{opt}}$.

For simplicity, we will not model the noise statistically. Let $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_K$ be arbitrary noise vectors whose ℓ_2 norms do not exceed ε. Suppose that we measure K signals

$$\boldsymbol{s}_k \quad = \quad \boldsymbol{\Phi}\, \boldsymbol{c}_{\mathrm{opt}} \ + \ \boldsymbol{\nu}_k.$$

We form a signal matrix $\boldsymbol{S}$ whose columns are $\boldsymbol{s}_1, \dots, \boldsymbol{s}_K$. Given this signal matrix, the goal is to determine which atoms participated in the ideal signal.

First, we need to choose γ so that the hypotheses of Corollary 5.2 hold. Let $\Lambda = \mathrm{supp}(\boldsymbol{c}_{\mathrm{opt}})$, and let $\boldsymbol{P}_\Lambda$ denote the orthogonal projector onto the span of the atoms listed in Λ. Recall that

$\boldsymbol{A}_\Lambda = \boldsymbol{P}_\Lambda \boldsymbol{S}$. According to the Pythagorean Theorem,

$$\begin{aligned} \|\boldsymbol{s}_k - \boldsymbol{P}_\Lambda \boldsymbol{s}_k\|_2^2 &= \|\boldsymbol{s}_k - \boldsymbol{\Phi}\, \boldsymbol{c}_{\text{opt}}\|_2^2 - \|\boldsymbol{\Phi}\, \boldsymbol{c}_{\text{opt}} - \boldsymbol{P}_\Lambda \boldsymbol{s}_k\|_2^2 \\ &= \|\boldsymbol{\nu}_k\|_2^2 - \|\boldsymbol{\Phi}\, \boldsymbol{c}_{\text{opt}} - \boldsymbol{P}_\Lambda \boldsymbol{s}_k\|_2^2 \\ &\le \varepsilon^2. \end{aligned}$$

In other words, each column of $(\boldsymbol{S} - \boldsymbol{A}_\Lambda)$ has ℓ_2 norm no greater than ε. Since the columns of $\boldsymbol{\Phi}$ have unit ℓ_2 norm, it follows directly that

$$\|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \le \varepsilon K.$$

Referring to the remarks after Corollary 5.2, we discover that it suffices to choose $\gamma = 2\,\varepsilon K$. Invoking the corollary, we see that $\boldsymbol{B}_\star$ has row-support inside Λ and also that

$$\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\infty,\infty} \le 3\,\gamma/2 = 3\,\varepsilon K.$$

We wish to ensure that the λ-th row of $\boldsymbol{B}_\star$ is nonzero for each $\lambda \in \Lambda$. From this fact, it will follow that $\operatorname{rowsupp}(\boldsymbol{B}_\star) = \Lambda$.

Let $\mathbf{e}$ denote a $K \times 1$ vector of ones. Clearly, the λ-th row of $\boldsymbol{c}_{\text{opt}}\, \mathbf{e}^*$ has absolute sum equal to K for each λ in Λ. Therefore, we can ensure that the λ-th row of $\boldsymbol{B}_\star$ is nonzero if $\|\boldsymbol{B}_\star - \boldsymbol{c}_{\text{opt}}\, \mathbf{e}^*\|_{\infty,\infty} < K$. Collecting the noise vectors into a $d \times K$ matrix $\boldsymbol{N}$, we may calculate

$$\begin{aligned} \|\boldsymbol{C}_\Lambda - \boldsymbol{c}_{\text{opt}}\, \mathbf{e}^*\|_{\infty,\infty} &= \left\|{\boldsymbol{\Phi}_\Lambda}^\dagger (\boldsymbol{S} - \boldsymbol{\Phi}_\Lambda\, \boldsymbol{c}_{\text{opt}}\, \mathbf{e}^T)\right\|_{\infty,\infty} \\ &= \left\|{\boldsymbol{\Phi}_\Lambda}^\dagger \boldsymbol{N}\right\|_{\infty,\infty} \\ &\le K \left\|{\boldsymbol{\Phi}_\Lambda}^\dagger\right\|_{2,\infty} \|\boldsymbol{N}\|_{1,2} \\ &\le \varepsilon K \sqrt{3/2}. \end{aligned}$$

Here, we have used the coherence bound $\left\|{\boldsymbol{\Phi}_\Lambda}^\dagger\right\|_{2,\infty} \le \sqrt{3/2}$ from Appendix III of [Tro04b]. The triangle inequality shows that

$$\|\boldsymbol{B}_\star - \boldsymbol{c}_{\text{opt}}\, \mathbf{e}^*\|_{\infty,\infty} \le \varepsilon K (3 + \sqrt{3/2})$$

Therefore, we require that

$$\varepsilon K (3 + \sqrt{3/2}) < K.$$

Rearranging this relation yields

$$\varepsilon < \frac{1}{3 + \sqrt{3/2}} \approx 0.2367.$$

If the noise level ε satisfies this condition, then the solution $\boldsymbol{B}_\star$ to the convex relaxation with parameter $\gamma = 2\,\varepsilon K$ identifies each and every atom in Λ, and it makes no mistakes.

This calculation does not show that we accrue any advantage from multiple observations of the signal. This fact is not surprising because we allowed the noise vectors to be completely arbitrary. Each observation of the signal could be contaminated by the same worst-case noise vector. To see the benefit of simultaneous sparse approximation, one must model the noise statistically. In that case, it would be highly unlikely for all the noise vectors to be directed unfavorably. Unfortunately, we have not made the complicated calculations that are necessary to support this point.

6. The Error-Constrained Method

Another approach to simultaneous sparse approximation is to seek the sparsest coefficient matrix that synthesizes a signal matrix within a fixed distance of the given signal matrix. We can attempt to solve this problem by way of the convex program

$$\min_{\boldsymbol{B} \in \mathbb{C}^{\Omega \times K}} \|\boldsymbol{B}\|_{\text{rx}} \quad \text{subject to} \quad \|\boldsymbol{S} - \boldsymbol{\Phi}\, \boldsymbol{B}\|_{\text{F}} \le \delta. \qquad (\text{RX-ERROR})$$

Minimizing the norm $\|\cdot\|_{\text{rx}}$ promotes row-sparsity in the coefficient matrix, while the parameter δ describes the amount of approximation error we are willing to tolerate. In the case where the signal matrix has a single column $\boldsymbol{s}$, the convex program reduces to

$$\min_{\boldsymbol{b} \in \mathbb{C}^{\Omega}} \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \|\boldsymbol{s} - \boldsymbol{\Phi}\, \boldsymbol{b}\|_2 \ \leq\ \delta. \tag{ℓ_1-ERROR}$$

This method for simple sparse approximation has been studied extensively in [DET04, Tro04b]. The evidence in [Tro04b] suggests that the penalty method for simple sparse approximation is much more powerful than the error-constrained method, and this paper will provide some evidence that the same phenomenon occurs in the simultaneous case.

In this section, we will present a theorem on the performance of (RX-ERROR), and we will show how to apply this theorem to a specific problem. Several related examples for simple sparse approximation appear in [Tro04b].

6.1. **Performance of the Relaxation.** Our major theorem provides conditions under which the solution to (RX-ERROR) has row-support contained inside a given index set Λ. The definition of $\text{ERC}(\Lambda)$ appears in equation (4.7).

Theorem 6.1. *Let Λ index a linearly independent collection of atoms for which $\text{ERC}(\Lambda) > 0$, and fix a signal matrix $\boldsymbol{S}$. Let $\boldsymbol{A}_\Lambda$ be the best approximation of $\boldsymbol{S}$ over Λ, and let $\boldsymbol{C}_\Lambda$ be the associated synthesis matrix. Select an error tolerance δ that satisfies*

$$\delta^2 \quad \geq \quad \|\boldsymbol{S} - \boldsymbol{A}_\Lambda\|_{\text{F}}^2 \ + \ \left[\frac{\left\|{\boldsymbol{\Phi}_\Lambda}^{\dagger}\right\|_{2,1} \|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty}}{\text{ERC}(\Lambda)} \right]^2 .$$

Then it follows that the unique solution $\boldsymbol{B}_\star$ to (RX-ERROR) *with tolerance δ has the following properties.*

- *The row-support of $\boldsymbol{B}_\star$ is contained in Λ.*
- *The distance between $\boldsymbol{B}_\star$ and $\boldsymbol{C}_\Lambda$ satisfies*

$$\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\text{F}} \quad \leq \quad \delta \left\|{\boldsymbol{\Phi}_\Lambda}^{\dagger}\right\|_{2,2}.$$

- *In particular,* $\text{rowsupp}(\boldsymbol{B}_\star)$ *contains every index λ from Λ for which*

$$\sum_{k=1}^{K} |\boldsymbol{C}_\Lambda(\lambda, k)| \quad > \quad \delta \left\|{\boldsymbol{\Phi}_\Lambda}^{\dagger}\right\|_{2,2}.$$

In words, we select the tolerance δ somewhat larger than the Frobenius norm of the residual $(\boldsymbol{S} - \boldsymbol{A}_\Lambda)$. This ensures that the solution to the convex relaxation identifies every atom that participates significantly in the signal, and the relaxation never identifies any atom outside Λ. The proof of this theorem is substantially identical with the proof of Theorem 14 from [Tro04b]. In consequence, we omit the details.

As with Theorem 5.1, one must leverage information about the problem domain to select the index set Λ and the tolerance δ. We can also present a version of the theorem that uses the coherence parameter to estimate some key quantities.

Corollary 6.2. *Let Λ index a linearly independent collection of atoms for which $\text{ERC}(\Lambda) > 0$, and fix a signal matrix $\boldsymbol{S}$. Let $\boldsymbol{A}_\Lambda$ be the best approximation of $\boldsymbol{S}$ over Λ, and let $\boldsymbol{C}_\Lambda$ be the associated synthesis matrix. Select an error tolerance δ that satisfies*

$$\delta^2 \quad \geq \quad \|\boldsymbol{S} - \boldsymbol{A}_\Lambda\|_{\text{F}}^2 \ + \ \frac{m\,(1 \ - \ (m-1)\,\mu}{(1 \ - \ (2m-1)\,\mu)^2} \ \|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty}^2 . \tag{6.1}$$

Then it follows that the unique solution $\boldsymbol{B}_\star$ to (RX-ERROR) *with tolerance δ has the following properties. The row-support of $\boldsymbol{B}_\star$ is contained in Λ, and the distance between the coefficient vectors satisfies $\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\text{F}} \leq \delta \sqrt{1 - (m-1)\,\mu}$.*

This result follows directly from the coherence estimates developed in Section 3 of [Tro04c]. The corollary takes an especially pretty form if $m\,\mu \le \frac{1}{3}$. Then the fraction in (6.1) reduces to $6\,m$, and the distance between the coefficient vectors satisfies $\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\mathrm{F}} \le \delta\,\sqrt{3/2}$.

6.2. **Comparison with Simple Sparse Approximation.** Theorem 6.1 contains the analogous result for simple sparse approximation [Tro04b, Thm. 14], so nothing has been lost in our generalization. Once again, there are some qualitative differences between the two results. We expect that it will be necessary to choose δ somewhat larger in the simultaneous case than in the simple case, and this will increase the distance between $\boldsymbol{B}_\star$ and $\boldsymbol{C}_\Lambda$. In compensation, we expect that the absolute row sums of $\boldsymbol{C}_\Lambda$ will be somewhat larger in the simultaneous case. The details of the application will determine which effect dominates.

6.3. **Example: Identifying Noisy Sparse Signals (Redux).** We will apply Theorem 6.1 to the same example that we described in Section 5.4. That is, we will show that the convex program (RX-ERROR) can be used to recover a synthetic sparse signal from multiple observations that have been contaminated with noise.

Let us repeat the setup. Suppose that $\boldsymbol{\Phi}$ is a dictionary with coherence μ, and suppose that $m\,\mu \le \frac{1}{3}$. Let $\boldsymbol{c}_{\mathrm{opt}}$ be a vector whose m nonzero entries equal one. Each ideal signal has the form $\boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}}$. Take $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_K$ to be arbitrary noise vectors whose ℓ_2 norms do not exceed ε. Suppose that we measure K signals

$$\boldsymbol{s}_k \quad = \quad \boldsymbol{\Phi}\,\boldsymbol{c}_{\mathrm{opt}} \; + \; \boldsymbol{\nu}_k.$$

We form a signal matrix $\boldsymbol{S}$ whose columns are $\boldsymbol{s}_1, \dots, \boldsymbol{s}_K$. Given this signal matrix, the goal is to determine which atoms participated in the ideal signal.

First, we must determine an appropriate value for the tolerance δ. Referring to the calculations in Section 5.4, we discover that $\|\boldsymbol{S} - \boldsymbol{A}_\Lambda\|_{\mathrm{F}}^2 \le \varepsilon^2 K$ and that $\|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_\Lambda)\|_{\infty,\infty} \le \varepsilon\,K$. The comments after Corollary 6.2 show that it is sufficient to choose

$$\delta \quad = \quad \varepsilon\,K\,\sqrt{1/K + \; 6\,m}.$$

Invoking the corollary, we discover that the row-support of the minimizer $\boldsymbol{B}_\star$ is contained in Λ and that

$$\|\boldsymbol{B}_\star - \boldsymbol{C}_\Lambda\|_{\infty,\infty} \quad \le \quad \delta\,\sqrt{3/2}.$$

We wish to ensure that the λ-th row of $\boldsymbol{B}_\star$ is nonzero for each λ in Λ. We follow the same approach as before.

Recall that the λ-th row of $\boldsymbol{c}_{\mathrm{opt}}\,\mathbf{e}^*$ has absolute sum equal to K for each λ in Λ. Section 5.4 shows that

$$\|\boldsymbol{C}_\Lambda - \boldsymbol{c}_{\mathrm{opt}}\,\mathbf{e}^*\|_{\infty,\infty} \quad \le \quad \varepsilon\,K\,\sqrt{3/2}.$$

It follows from the triangle inequality that

$$\|\boldsymbol{B}_\star - \boldsymbol{c}_{\mathrm{opt}}\,\mathbf{e}^*\|_{\infty,\infty} \quad \le \quad \varepsilon\,K\,\sqrt{3/2}\,(1 \; + \; \sqrt{1/K + 6\,m}).$$

Therefore, the λ-th row of $\boldsymbol{B}_\star$ is nonzero provided that

$$\varepsilon\,K\,\sqrt{3/2}\,(1 \; + \; \sqrt{1/K + 6\,m}) \quad < \quad K.$$

Solving this equation for ε yields

$$\varepsilon \quad < \quad \frac{\sqrt{2/3}}{1 \; + \; \sqrt{1/K + 6\,m}}.$$

This condition on ε is sufficient to ensure that the solution $\boldsymbol{B}_\star$ to the convex relaxation has row-support equal to Λ.

It is interesting to see that additional observations do improve the performance of the convex relaxation (RX-ERROR). For a fixed value of m, the upper bound on ε increases with K. Nevertheless, the bound on ε decreases as m grows. For every value of m, the theoretical evidence suggests that the penalty method will perform better than the error-constrained method.

7. Comparison with Previous Work

There are many different ways to relax the row-ℓ_0 quasi-norm. One may define an entire family of relaxations [CREKD04, Eqn. 13] of the form

$$J_{p,q}(\boldsymbol{B}) \quad \stackrel{\text{def}}{=} \quad \sum\nolimits_{\omega \in \Omega} \left[\sum\nolimits_{k=1}^{K} |b_{\omega k}|^q \right]^{p/q}. \tag{7.1}$$

Typically $p \leq 1$ and $q \geq 1$. That is, the ℓ_q norm is applied to rows of the coefficient matrix, and the ℓ_p quasi-norm is applied to the resulting vector of norms. Note that $J_{p,q}$ is convex whenever $p, q \geq 1$.

It seems that the family (7.1) encompasses all the relaxations of $\|\cdot\|_{\text{row-0}}$ that have appeared in the literature. The relaxation $\|\cdot\|_{\text{rx}}$ that we have considered in this paper corresponds with the case $p = 1$ and $q = \infty$ (with the usual convention). Chen and Huo [CH04a, CH04b] present theory for the case $p, q = 1$. Cotter et al. [CREKD04] and Malioutov et al. [MÇW03] focus on the case $p \leq 1$ and $q = 2$. At present, there is little empirical or theoretical evidence by which to prefer one relaxation over another.

The most detailed theoretical work so far is due to Chen and Huo [CH04a, CH04b], who have studied convex relaxations involving $J_{1,1}$. They prove that the convex program

$$\min \ J_{1,1}(\boldsymbol{B}) \qquad \text{subject to} \qquad \boldsymbol{S} = \boldsymbol{\Phi}\, \boldsymbol{B}$$

can be used to recover signals that have a sufficiently sparse representation. General intuitions about sparse approximation suggest that this method is most appropriate when the coefficient matrix is very sparse (in the usual sense) and not merely row-sparse. An interesting feature of this optimization problem is that it separates into independent simple sparse approximation problems [CH04b, Rem. 5.3]. This fact makes the analysis much cleaner. On comparison with the arguments in this paper, one may judge that simplicity is a worthwhile end.

Cotter et al. have developed an algorithm, M-FOCUSS, for solving

$$\min \ J_{p,2}(\boldsymbol{B}) \qquad \text{subject to} \qquad \boldsymbol{S} = \boldsymbol{\Phi}\, \boldsymbol{B}$$

where $p \leq 1$. A pragmatic reason for considering the relaxation $J_{p,2}$ is that it leads to efficient algorithms of the factored-gradient type. Cotter et al. prove that their algorithms converge to a local minimum, which in case of $J_{1,2}$ must also be a global minimum [CREKD04].

Malioutov et al. have considered convex programs that can handle sparse signal matrices contaminated with noise. In particular, they consider problems of the form

$$\min \ \|\boldsymbol{S} - \boldsymbol{\Phi}\, \boldsymbol{B}\|_{\text{F}}^2 \ + \ \gamma\, J_{1,2}(\boldsymbol{B}).$$

They provide empirical evidence that these convex programs offer an effective method for source-localization using linear arrays of sensors. See [MÇW03] and [Mal03, Sec. 5.1.6] for details.

The methods of this paper can certainly be adapted to develop results for convex relaxations involving other members of the family $\{J_{1,q}\}$. (Our methods will not work in case $p < 1$.) It would be interesting to see how these relaxations compare with each other theoretically. It would also be valuable to perform numerical experiments to determine which of the convex relaxations offers the best balance of accuracy and computational cost.

Acknowledgments

I sincerely thank the anonymous referees for their technical suggestions and their advice on decorum.

References

[BL76] J. Bergh and J. Löfström. *Interpolation Spaces: An Introduction*. Number 223 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 1976.

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

[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.

[CH04a] J. Chen and X. Huo. Sparse representations for multiple measurement vectors (mmv) in an overcomplete dictionary. Submitted to ICASSP 2005, 2004.

[CH04b] J. Chen and X. Huo. Theoretical results about finding the sparsest representations of multiple measurement vectors (mmv) in an overcomplete dictionary using ℓ_1 minimization and greedy algorithms. Manuscript, 2004.

[CREKD04] S. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions of linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 2004. Accepted.

[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.

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

[Fuc04a] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory*, 50(6):1341–1344, June 2004.

[Fuc04b] J.-J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. Unpublished manuscript, Apr. 2004.

[Fuc04c] J.-J. Fuchs. Recovery of exact sparse representations in the presence of noise. In *Proc. of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montréal, May 2004.

[Mal03] D. Malioutov. A sparse signal reconstruction perspective for source localization with sensor arrays. Ms.c. thesis, MIT, July 2003.

[MÇW03] D. Malioutov, M. Çetin, and A. Willsky. Source localization by enforcing sparsity through a Laplacian prior: an SVD-based approach. In *IEEE Statistical Signal Processing Workshop*, pages 553–556, Oct. 2003.

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

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

[TGS04] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Submitted for publication, Nov. 2004.

[Tro04a] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10), Oct. 2004.

[Tro04b] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals. Submitted for publication, June 2004.

[Tro04c] J. A. Tropp. Just relax: Convex programming methods for subset selection and sparse approximation. ICES Report 04-04, The University of Texas at Austin, 2004.

[Tro04d] J. A. Tropp. *Topics in Sparse Approximation*. Ph.D. dissertation, Computational and Applied Mathematics, The University of Texas at Austin, August 2004.