

# Quantitative Robust Uncertainty Principles and Optimally Sparse Decompositions

Emmanuel J. Candès and Justin Romberg

Applied and Computational Mathematics, Caltech, Pasadena, CA 91125

June 3, 2005

## Abstract

In this paper, we develop a robust *uncertainty principle* for finite signals in  $\mathbb{C}^N$  which states that for nearly all choices  $T, \Omega \subset \{0, \dots, N-1\}$  such that

$$|T| + |\Omega| \asymp (\log N)^{-1/2} \cdot N,$$

there is no signal  $f$  supported on  $T$  whose discrete Fourier transform  $\hat{f}$  is supported on  $\Omega$ . In fact, we can make the above uncertainty principle *quantitative* in the sense that if  $f$  is supported on  $T$ , then only a small percentage of the energy (less than half, say) of  $\hat{f}$  is concentrated on  $\Omega$ .

As an application of this robust uncertainty principle (QRUP), we consider the problem of decomposing a signal into a sparse superposition of spikes and complex sinusoids

$$f(s) = \sum_{t \in T} \alpha_1(t) \delta(s-t) + \sum_{\omega \in \Omega} \alpha_2(\omega) e^{i2\pi\omega s/N} / \sqrt{N}.$$

We show that if a generic signal  $f$  has a decomposition  $(\alpha_1, \alpha_2)$  using spike and frequency locations in  $T$  and  $\Omega$  respectively, and obeying

$$|T| + |\Omega| \leq \text{Const} \cdot (\log N)^{-1/2} \cdot N,$$

then  $(\alpha_1, \alpha_2)$  is the *unique sparsest* possible decomposition (all other decompositions have more non-zero terms). In addition, if

$$|T| + |\Omega| \leq \text{Const} \cdot (\log N)^{-1} \cdot N,$$

then the sparsest  $(\alpha_1, \alpha_2)$  can be found by solving a convex optimization problem.

Underlying our results is a new probabilistic approach which insists on finding the correct uncertainty relation or the optimally sparse solution for nearly all subsets but not necessarily all of them, and allows to considerably sharpen previously known results [9, 10]. In fact, we show that the fraction of sets  $(T, \Omega)$  for which the above properties do not hold can be upper bounded by quantities like  $N^{-\alpha}$  for large values of  $\alpha$ .

The QRUP (and the application to finding sparse representations) can be extended to general pairs of orthogonal bases  $\Phi_1, \Phi_2$  of  $\mathbb{C}^N$ . For nearly all choices  $\Gamma_1, \Gamma_2 \subset \{0, \dots, N-1\}$  obeying

$$|\Gamma_1| + |\Gamma_2| \asymp \mu(\Phi_1, \Phi_2)^{-2} \cdot (\log N)^{-m},$$

where  $m \leq 6$ , there is no signal  $f$  such that  $\Phi_1 f$  is supported on  $\Gamma_1$  and  $\Phi_2 f$  is supported on  $\Gamma_2$  where  $\mu(\Phi_1, \Phi_2)$  is the *mutual coherence* between  $\Phi_1$  and  $\Phi_2$ .

**Keywords.** Uncertainty principle, applications of uncertainty principles, random matrices, eigenvalues of random matrices, sparsity, trigonometric expansion, convex optimization, duality in optimization, basis pursuit, wavelets, linear programming.

**Acknowledgments.** E. C. is partially supported by National Science Foundation grants DMS 01-40698 (FRG) and ACI-0204932 (ITR), and by an Alfred P. Sloan Fellowship. J. R. is supported by those same National Science Foundation grants. E. C. would like to thank Terence Tao for helpful conversations related to this project. These results were presented at the International Conference on Computational Harmonic Analysis, Nashville, Tennessee, May 2004.

## 1 Introduction

### 1.1 Uncertainty principles

The classical Weyl-Heisenberg uncertainty principle states that a continuous-time signal cannot be simultaneously well-localized in both time and frequency. Loosely speaking, this principle says that if most of the energy of a signal  $f$  is concentrated near a time-interval of length  $\Delta t$  and most of its energy in the frequency domain is concentrated near an interval of length  $\Delta\omega$ , then

$$\Delta t \cdot \Delta\omega \geq 1.$$

This principle is one of the major intellectual achievements of the 20th century, and since then much work has been concerned with extending such uncertainty relations to other contexts, namely by investigating to what extent it is possible to concentrate a function  $f$  and its Fourier transform  $\hat{f}$  when relaxing the assumption that  $f$  and  $\hat{f}$  be concentrated near intervals as in the work of Landau, Pollack and Slepian [17, 18, 21], or by considering signals supported on a discrete set [9, 22].

Because our paper is concerned with finite signals, we now turn our attention to “discrete uncertainty relations” and begin by recalling the definition of the discrete Fourier transform

$$\hat{f}(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} f(t) e^{-i2\pi\omega t/N}, \quad (1.1)$$

where the frequency index  $\omega$  ranges over the set  $\{0, 1, \dots, N-1\}$ . For signals of length  $N$ , [9] introduced a sharp uncertainty principle which simply states that the supports of a signal  $f$  in the time and frequency domains must obey

$$|\text{supp } f| + |\text{supp } \hat{f}| \geq 2\sqrt{N}. \quad (1.2)$$

We emphasize that there are no other restriction on the organization of the supports of  $f$  and  $\hat{f}$  other than the size constraint (1.2). It was also observed in [9] that the uncertainty relation (1.2) is tight in the sense that equality is achieved for certain special signals. For example, consider as in [9, 10] the *Dirac comb* signal: suppose that the sample size  $N$  is a perfect square and let  $f$  be equal to 1 at multiples of  $\sqrt{N}$  and 0 everywhere else

$$f(t) = \begin{cases} 1, & t = m\sqrt{N}, \quad m = 0, 1, \dots, \sqrt{N} - 1 \\ 0, & \text{elsewhere.} \end{cases} \quad (1.3)$$

Remarkably, the Dirac comb is invariant through the Fourier transform, i.e.  $\hat{f} = f$ , and therefore,  $|\text{supp } f| + |\text{supp } \hat{f}| = 2\sqrt{N}$ . In other words, (1.2) holds with equality.

In recent years, uncertainty relations have become very popular, in part because they help explain some miraculous properties of  $\ell_1$ -minimization procedures, as we will see below. Researchers have naturally developed similar uncertainty relations between pairs of bases other than the canonical basis and its conjugate. We single out the work of Elad and Bruckstein [12] which introduces a generalized uncertainty principle for pairs  $\Phi_1, \Phi_2$  of orthonormal bases. Define the *mutual coherence* [10, 12, 15] between  $\Phi_1$  and  $\Phi_2$  as

$$\mu(\Phi_1, \Phi_2) = \max_{\phi \in \Phi_1, \psi \in \Phi_2} |\langle \phi, \psi \rangle|; \quad (1.4)$$

then if  $\alpha_1$  is the (unique) representation of  $f$  in basis  $\Phi_1$  with  $\Gamma_1 = \text{supp } \alpha_1$ , and  $\alpha_2$  is the representation in  $\Phi_2$ , the supports must obey

$$|\Gamma_1| + |\Gamma_2| \geq \frac{2}{\mu(\Phi_1, \Phi_2)}. \quad (1.5)$$

Note that the mutual coherence  $\mu$  always obeys  $1/\sqrt{N} \leq \mu \leq 1$  and is a rough measure of the similarity of the two bases. The smaller the coherence, the stronger the uncertainty relation. To see how this generalizes the discrete uncertainty principle, observe that in the case where  $\Phi_1$  is the canonical or spike basis and  $\Phi_2$  is the Fourier basis,  $\mu = 1/\sqrt{N}$  (maximal incoherence) and (1.5) is, of course, (1.2).

## 1.2 The tightness of the uncertainty relation is fragile

It is true that there exist signals that saturate the uncertainty relations. But such signals are very special and are hardly representative of “generic” or “most” signals. Consider, for instance, the Dirac comb: the locations and heights of the  $\sqrt{N}$  spikes in the time domain carefully conspire to create an inordinate number of cancellations in the frequency domain. This will not be the case for sparsely supported signals in general. Simple numerical experiments confirm that signals with the same support as the Dirac comb but with different spike amplitudes almost always have Fourier transforms that are nonzero everywhere. Indeed, constructing pathological examples other than the Dirac comb requires mathematical wizardry.

Moreover, if the signal length  $N$  is prime (making signals like the Dirac comb impossible to construct), the discrete uncertainty principle is sharpened to [26]

$$|\text{supp } f| + |\text{supp } \hat{f}| > N, \quad (1.6)$$

which validates our intuition about the exceptionality of signals such as the Dirac comb.

## 1.3 Robust uncertainty principles

Excluding these exceedingly rare and exceptional pairs  $T := \text{supp } f, \Omega := \text{supp } \hat{f}$ , how tight is the uncertainty relation? That is, given two sets  $T$  and  $\Omega$ , how large need  $|T| + |\Omega|$  be so that it is possible to construct a signal whose time and frequency supports are  $T$  and  $\Omega$  respectively? In this paper, we introduce a *robust* uncertainty principle (for general  $N$ )

which illustrates that for “most” sets  $T, \Omega$ , (1.6) is closer to the truth than (1.2). Suppose that we choose  $(T, \Omega)$  at random from all pairs obeying

$$|T| + |\Omega| \leq \frac{N}{\sqrt{(\beta + 1) \log N}},$$

where  $\beta$  is a pre-determined constant. Then with overwhelming high probability—in fact, exceeding  $1 - O(N^{-\beta})$  (we shall give explicit values)—we will be unable to find a signal in  $\mathbb{C}^N$  supported on  $T$  in the time domain and  $\Omega$  in the frequency domain. In other words, remove a negligible fraction of sets and

$$|\text{supp } f| + |\text{supp } \hat{f}| > \frac{N}{\sqrt{(\beta + 1) \log N}}, \quad (1.7)$$

holds, which is very different than (1.2).

Our uncertainty principle is not only robust in the sense that it holds for most sets, it is also *quantitative*. Consider a random pair  $(T, \Omega)$  as before and put  $1_\Omega$  to be the indicator function of the set  $\Omega$ . Then with essentially the same probability as above, we have

$$\|\hat{f} \cdot 1_\Omega\|^2 \leq \|\hat{f}\|^2/2, \quad (1.8)$$

say, for all functions  $f$  supported on  $T$ . By symmetry, the same inequality holds by exchanging the role of  $T$  and  $\Omega$ ,

$$\|f \cdot 1_T\|^2 \leq \|f\|^2/2,$$

for all functions  $\hat{f}$  supported on  $\Omega$ . Moreover, as with the discrete uncertainty principle, the QRUP can be extended to arbitrary pairs of bases.

## 1.4 Significance of uncertainty principles

Over the past three years, there has been a series of papers starting with [10] establishing a link between discrete uncertainty principles and sparse approximation [10, 11, 15, 27]. In this field, the goal is to separate a signal  $f \in \mathbb{C}^N$  into two (or more) components, each representing contributions from different phenomena. The idea is as follows: suppose we have two (or possibly many more) orthonormal bases  $\Phi_1, \Phi_2$ ; we search among all the decompositions  $(\alpha_1, \alpha_2)$  of the signal  $f$

$$f = (\Phi_1 \quad \Phi_2) \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} := \Phi \alpha$$

for the shortest one

$$(P_0) \quad \min_{\alpha} \|\alpha\|_{\ell_0}, \quad \Phi \alpha = f, \quad (1.9)$$

where  $\|\alpha\|_{\ell_0}$  is simply the size of the support of  $\alpha$ ,  $\|\alpha\|_{\ell_0} := |\{\gamma, \alpha(\gamma) \neq 0\}|$ .

The discrete uncertainty principles (1.2) and (1.5) tell us when  $(P_0)$  has a unique solution. For example, suppose  $\Phi$  is the time-frequency dictionary constructed by concatenating the Dirac and Fourier orthobases:

$$\Phi = (I \quad F^*)$$

where  $F$  is the discrete Fourier matrix  $(F)_{t,\omega} = (1/\sqrt{N})e^{i2\pi\omega t/N}$ . It is possible to show that if a signal  $f$  has a decomposition  $f = \Phi\alpha$  consisting of spikes on subdomain  $T$  and frequencies on  $\Omega$ , and

$$|T| + |\Omega| < \sqrt{N}, \quad (1.10)$$

then  $\alpha$  is the unique minimizer of  $(P_0)$  [10]. In a nutshell, the reason is that if  $\Phi(\alpha_0 + \delta_0)$  were another decomposition,  $\delta_0$  would obey  $\Phi\delta_0 = 0$ , which dictates that  $\delta_0$  has the form  $\delta_0 = (\delta, -\hat{\delta})$ . Now (1.2) implies that  $\delta_0$  would have at least  $2\sqrt{N}$  nonzero entries which in turn would give  $\|\alpha_0 + \delta_0\|_{\ell_0} \geq \sqrt{N}$  for all  $\alpha_0$  obeying  $\|\alpha_0\|_{\ell_0} < \sqrt{N}$ , thereby proving the claim. Note that again the condition (1.10) is sharp because of the extremal signal (1.3). Indeed, the Dirac comb may be expressed as a superposition of  $\sqrt{N}$  terms in the time or in the frequency domain; for this special signal,  $(P_0)$  does not have a unique solution.

In [12], the same line of reasoning is followed for general pairs of orthogonal bases, and  $\ell_0$ -uniqueness is guaranteed when

$$|\Gamma_1| + |\Gamma_2| < \frac{1}{\mu(\Phi_1, \Phi_2)}. \quad (1.11)$$

Unfortunately, solving  $(P_0)$  directly is computationally infeasible because of the highly non-convex nature of the  $\|\cdot\|_{\ell_0}$  norm. To the best of our knowledge, finding the minimizer obeying the constraints would require searching over all possible *subsets* of columns of  $\Phi$ , an algorithm that is combinatorial in nature and has exponential complexity. Instead of solving  $(P_0)$ , we consider a similar program in the  $\ell_1$  norm which goes by the name of *Basis Pursuit* [7]:

$$(P_1) \quad \min_{\alpha} \|\alpha\|_{\ell_1}, \quad \Phi\alpha = f. \quad (1.12)$$

Unlike the  $\ell_0$  norm, the  $\ell_1$  norm is convex. As a result,  $(P_1)$  can be solved efficiently using standard “off the shelf” optimization algorithms. The  $\ell_1$ -norm itself can also be viewed as a measure of sparsity. Among the vectors that meet the constraints, it will favor those with a few large coefficients and many small coefficients over those where the coefficient magnitudes are approximately equal [7].

A beautiful result in [10] actually shows that if  $f$  has a sparse decomposition  $\alpha$  supported on  $\Gamma$  with

$$|\Gamma| < \frac{1}{2}(1 + \mu^{-1}), \quad (1.13)$$

then the minimizer of  $(P_1)$  is unique and is equal to the minimizer of  $(P_0)$  ([12] improves the constant in (1.13) from  $1/2$  to  $\approx .9142$ ). In these situations, we can replace the highly non-convex program  $(P_0)$  with the much tamer (and convex)  $(P_1)$ .

We now review a few applications of these types of ideas.

- *Geometric Separation.* Suppose we have a dataset and one wishes to separate point-like structures, from filamentary (edge-like) structures, from sheet-like structures. In 2 dimensions, for example, we might imagine synthesizing a signal as a superposition of wavelets and curvelets which are ideally adapted to represent point-like and curve-like structures respectively. Delicate space/orientation uncertainty principles show that the minimum  $\ell_1$ -norm decomposition in this combined dictionary automatically separates point and curve-singularities; the wavelet component in the decomposition (1.12) accurately captures all the pointwise singularities, while the curvelet component

captures all the edge curves. We refer to [8] for theoretical developments and to [23] for numerical experiments.

- *Texture-edges separation* Suppose now that we have an image we wish to decompose as a sum of a cartoon-like geometric part plus a texture part. The idea again is to use curvelets to represent the geometric part of the image and local cosines to represent the texture part. These ideas have recently been tested in practical settings, with spectacular success [24] (see also [20] for earlier and related ideas).

In a different direction, the QRUP is also implicit in some of our own work on the exact reconstruction of sparse signals from vastly undersampled frequency information [3]. Here, we wish to reconstruct a signal  $f \in \mathbb{C}^N$  from the data of only  $|\Omega|$  random frequency samples. The surprising result is that although most of the information is missing, one can still reconstruct  $f$  *exactly* provided that  $f$  is sparse. Suppose  $|\Omega|$  obeys the oversampling relation

$$|\Omega| \asymp |T| \cdot \log N$$

with  $T := \text{supp } f$ . Then with overwhelming probability, the object  $f$  (digital image, signal, and so on) is the exact and unique solution of the convex program that searches, among all signals that are consistent with the data, for that with minimum  $\ell_1$ -norm. We will draw on the tools developed in the earlier work, making the QRUP *explicit* and applying it to the problem of searching for sparse decompositions.

## 1.5 Innovations

Nearly all the existing literature on uncertainty relations and their consequences focuses on worst case scenarios, compare (1.2) and (1.13). What is new here is the development of probabilistic models which show that the performance of Basis Pursuit in an overwhelmingly large majority of situations is actually very different than that predicted by the overly “pessimistic” bounds (1.13). For the time-frequency dictionary, we will see that if a representation  $\alpha$  (with spike locations  $T$  and sinusoidal frequencies  $\Omega$ ) of a signal  $f$  exists with

$$|T| + |\Omega| \asymp N/\sqrt{\log N},$$

then  $\alpha$  is the sparsest representation of  $f$  nearly all of the time. If in addition,  $T$  and  $\Omega$  satisfy

$$|T| + |\Omega| \asymp N/\log N, \tag{1.14}$$

then  $\alpha$  can be recovered by solving the convex program  $(P_1)$ . In fact, numerical simulations reported in Section 6 suggest that (1.14) is far closer to the empirical behavior than (1.13), see also [9]. We show that similar results also hold for general pairs of bases  $\Phi_1, \Phi_2$ .

As discussed earlier, there now exists well-established machinery for turning uncertainty relations into statements about our ability to find sparse decompositions. We would like to point out that our results (1.14) are not an automatic consequence of the uncertainty relation (1.7) together with these existing ideas. Instead, our analysis relies on the study of eigenvalues of random matrices, which is completely new.

## 1.6 Organization of the paper

In Section 2 we develop a probability model that will be used throughout the paper to formulate our results. In Section 3, we will establish uncertainty relations such as (1.8). Sections 4 and 5 will prove uniqueness and equality of the  $(P_0)$  and  $(P_1)$  programs. In the case where the basis pair  $(\Phi_1, \Phi_2)$  is the time-frequency dictionary (Section 4), we will be very careful in calculating the constants appearing in the bounds. We will be somewhat less precise in the general case (Section 5), and will forgo explicit calculation of constants. We report on numerical experiments in Section 6 and close the paper with a short discussion (Section 7).

## 2 A Probability Model for $\Gamma_1, \Gamma_2$

To state our results precisely, we first need to specify a probabilistic model. Our statements will be centered on subsets of the time and/or frequency domains (and more generally the  $\Phi_1$  and  $\Phi_2$  domains) that have been sampled *uniformly at random*. For example, to randomly generate a subset of the frequency domain of size  $N_\Omega$ , we simply choose one of the  $\binom{N}{N_\Omega}$  subsets of  $\{0, \dots, N-1\}$  using a uniform distribution, i.e. each subset  $\Omega$  with  $|\Omega| = N_\Omega$  has an equal chance of being chosen.

As we will see in the next section, the robust uncertainty principle holds—with overwhelming probability—over sets  $T$  and  $\Omega$  randomly sampled as above. Our estimates are quantitative and introduce sufficient conditions so that the probability of “failure” be arbitrarily small, i.e. less than  $O(N^{-\beta})$  for some arbitrary  $\beta > 0$ . Deterministically, we can say that the estimates are true for the vast majority of sets  $T$  and  $\Omega$  of a certain size.

Further, to establish sparse approximation bounds (Section 4), we will also introduce a probability model on the “active” coefficients. Given a pair  $(T, \Omega)$ , we sample the coefficient vector  $\{\alpha(\gamma), \gamma \in \Gamma := T \cup \Omega\}$  from a distribution with identically and independently distributed coordinates; we also impose that each  $\alpha(\gamma)$  be drawn from a continuous probability distribution that is *circularly symmetric* in the complex plane; that is, the phase of  $\alpha(\gamma)$  is uniformly distributed on  $[0, 2\pi)$ .

## 3 Quantitative Robust Uncertainty Principles

Equipped with the probability model of Section 2, we now introduce our uncertainty relations. To state our result, we make use of the standard notation  $o(1)$  to indicate a numerical term tending to 0 as  $N$  goes to infinity. We will assume throughout the paper that the signal length  $N \geq 512$ , and that the parameter  $\beta$  obeys  $1 \leq \beta \leq (3/8) \cdot \log N$  (since even taking  $\beta \sim \log N$  distorts the meaning of  $O(N^{-\beta})$ ).

**Theorem 3.1** *Choose  $N_T$  and  $N_\Omega$  such that*

$$N_T + N_\Omega \leq \frac{N}{\sqrt{(\beta + 1) \log N}} (\rho_0/2 + o(1)), \quad \rho_0 = \sqrt{2/3} > 0.8164. \quad (3.1)$$

*Fix a subset  $T$  of the time domain with  $|T| = N_T$ . Let  $\Omega$  be a subset of size  $N_\Omega$  of the frequency domain generated uniformly at random. Then with probability at least*

$1 - O((\log N)^{1/2} \cdot N^{-\beta})$ , every signal  $f$  supported on  $T$  in the time domain has most of its energy in the frequency domain outside of  $\Omega$ :

$$\|\hat{f} \cdot 1_\Omega\|^2 \leq \frac{\|\hat{f}\|^2}{2};$$

and likewise, every signal  $f$  supported on  $\Omega$  in the frequency domain has most of its energy in the time domain outside of  $T$

$$\|f \cdot 1_T\|^2 \leq \frac{\|f\|^2}{2}.$$

As a result, it is impossible to find a signal  $f$  supported on  $T$  whose discrete Fourier transform  $\hat{f}$  is supported on  $\Omega$ . For all finite sample sizes  $N \geq 512$ , we can select  $T, \Omega$  as

$$|T| + |\Omega| \leq \frac{.2791 N}{\sqrt{(\beta + 1) \log N}}.$$

To establish this result, we will show that the singular values of the *partial Fourier matrix*  $F_{\Omega T} = R_\Omega F R_T^*$  are well controlled, where  $R_T, R_\Omega$  are the time- and frequency-domain restriction operator to sets  $T$  and  $\Omega$ . If  $f \in \mathbb{C}^N$  is supported on  $T$  (such that  $R_T^* R_T f = f$ ), then its Fourier coefficients on  $\Omega$  are given by  $F_{\Omega T} R_T f$ , and

$$\|\hat{f} \cdot 1_\Omega\|^2 = \|F_{\Omega T} R_T f\|^2 \leq \|F_{\Omega T}^* F_{\Omega T}\| \cdot \|f\|^2$$

where  $\|A\|$  is the operator norm (largest eigenvalue) of the matrix  $A$ . Therefore, it will suffice to show that  $\|F_{\Omega T}^* F_{\Omega T}\| \leq 1/2$  with high probability.

With  $T$  fixed,  $F_{\Omega T}$  is a random matrix that depends on the random variable  $\Omega$ , where  $\Omega$  is drawn uniformly at random with a specified size  $|\Omega| = N_\Omega$ . We will write  $A(\Omega) = F_{\Omega T}$  to make this dependence explicit. To bound  $A(\Omega)$ , we will first derive a bound for  $A(\Omega')$  where  $\Omega'$  is sampled using independent coin flips, and show that in turn, it can be used to bound the uniform case.

Consider then the random set  $\Omega'$  sampled as follows: let  $I(\omega)$ ,  $\omega = 0, \dots, N - 1$  be a sequence of independent Bernoulli random variables with

$$\mathbf{P}(I(\omega) = 1) = \tau, \quad \mathbf{P}(I(\omega) = 0) = 1 - \tau,$$

and set

$$\Omega' = \{\omega \text{ s.t. } I(\omega) = 1\}.$$

Note that unlike  $\Omega$ , the size of the set  $\Omega'$  is random as well;

$$|\Omega'| = \sum_{\omega=0}^{N-1} I(\omega)$$

is a binomial random variable with mean  $\mathbf{E}|\Omega'| = \tau N$ . To relate the two models, we use

$$\begin{aligned} \mathbf{P}(\|A(\Omega')\|^2 > 1/2) &= \sum_{k=0}^N \mathbf{P}(\|A(\Omega')\|^2 > 1/2 \mid |\Omega'| = k) \cdot \mathbf{P}(|\Omega'| = k) \\ &= \sum_{k=0}^N \mathbf{P}(\|A(\Omega_k)\|^2 > 1/2) \cdot \mathbf{P}(|\Omega'| = k) \end{aligned} \quad (3.2)$$

where  $\Omega_k$  is a set of size  $|\Omega_k| = k$  selected uniformly at random. We will exploit two facts:



1.  $\mathbf{P}(\|A(\Omega_k)\|^2 > 1/2)$  is a non-decreasing function of  $k$ . For it is true that

$$\Omega_1 \subset \Omega_2 \quad \Rightarrow \quad \|A(\Omega_1)\|^2 \leq \|A(\Omega_2)\|^2.$$

Note that an equivalent method for sampling  $\Omega_{k+1}$  uniformly at random is to sample a set of size  $k$  uniformly at random, and then choose a single index from  $\{0, \dots, N-1\} \setminus \Omega_k$  again using a uniform distribution. From this, we can see

$$\mathbf{P}(\|A(\Omega_{k+1})\|^2 > 1/2) \geq \mathbf{P}(\|A(\Omega_k)\|^2 > 1/2).$$

2. If  $\mathbf{E}|\Omega'| = \tau N$  is an integer, then the median of  $|\Omega'|$  is also  $\tau N$ . That is,

$$\mathbf{P}(|\Omega'| \leq \tau N - 1) < 1/2 < \mathbf{P}(|\Omega'| \leq \tau N).$$

Proof of this fact can be found in [16, Theorem 3.2].

Drawing  $\Omega'$  with parameter  $\tau = N_\Omega/N$ , we can apply these facts (in order) to bound (3.2) as

$$\begin{aligned} \mathbf{P}(\|A(\Omega')\|^2 > 1/2) &\geq \sum_{k=N_\Omega}^N \mathbf{P}(\|A(\Omega_k)\|^2 > 1/2) \mathbf{P}(|\Omega'| = k) \\ &\geq \mathbf{P}(\|A(\Omega)\|^2 > 1/2) \sum_{k=N_\Omega}^N \mathbf{P}(|\Omega'| = k) \\ &\geq \frac{1}{2} \mathbf{P}(\|A(\Omega)\|^2 > 1/2). \end{aligned} \quad (3.3)$$

The failure rate for the uniform model is thus no more than twice the failure rate for the independent Bernoulli model.

It remains to show that  $\|A(\Omega')\|^2 \leq 1/2$  with high probability. To do this, we introduce (as in [3]) the  $|T| \times |T|$  auxiliary matrix  $\mathcal{H}(\Omega')$

$$\mathcal{H}(\Omega')_{t,t'} = \begin{cases} 0 & t = t' \\ \sum_{\omega \in \Omega'} e^{i2\pi\omega(t-t')/N} & t \neq t' \end{cases} \quad t, t' \in T. \quad (3.4)$$

We then have the identity

$$A(\Omega')^* A(\Omega') = \frac{|\Omega'|}{N} I + \frac{1}{N} \mathcal{H}$$

and

$$\|A(\Omega')\|^2 \leq \frac{|\Omega'|}{N} + \frac{1}{N} \|\mathcal{H}\|. \quad (3.5)$$

The following lemmas establish that both terms in (3.5) will be small.

**Lemma 3.1** *Fix  $q \in (0, 1/2]$  and let  $\Omega'$  be randomly generated with the Bernoulli model with*

$$\frac{(\log N)^2}{N} \leq \tau \leq \frac{\rho_0 q}{\sqrt{(\beta + 1) \log N}}. \quad (3.6)$$

Then

$$\frac{|\Omega'|}{N} \leq \frac{2\rho_0 q}{\sqrt{(\beta + 1) \log N}}$$

with probability greater than  $1 - N^{-\beta}$ .

**Proof** We apply the standard Bernstein inequality [1] for a Binomial random variable

$$\mathbf{P}(|\Omega'| > \mathbf{E}|\Omega'| + \lambda) \leq \exp\left(\frac{-\lambda^2}{2\mathbf{E}|\Omega'| + 2\lambda/3}\right), \quad \forall \lambda > 0,$$

and choose  $\lambda = \mathbf{E}|\Omega'| = \tau N$ . Then  $\mathbf{P}(|\Omega'| > 2\mathbf{E}|\Omega'|) \leq N^{-\beta}$ , since  $\tau \geq (\log N)^2/N \geq (8/3)\beta \log N/N$ , and

$$\frac{|\Omega'|}{N} \leq \frac{2\rho_0 q}{\sqrt{(\beta+1)\log N}}$$

with probability greater than  $1 - N^{-\beta}$ . ■

**Lemma 3.2** *Fix a set  $T$  in the time-domain and  $q \in (0, 1/2]$ . Let  $\Omega'$  be randomly generated under the Bernoulli model with parameter  $\tau$  such that*

$$|T| + \mathbf{E}|\Omega'| \leq \rho_0 \cdot \frac{qN}{\sqrt{(\beta+1)\log N}}, \quad \rho_0 = \sqrt{2/3} > 0.8164,$$

and let  $\mathcal{H}$  be defined as in (3.4). Then

$$\mathbf{P}(\|\mathcal{H}\| > qN) \leq 5 \cdot ((\beta+1)\log N)^{1/2} \cdot N^{-\beta}.$$

**Proof** To establish the lemma, we will leverage moment bounds from [3]. We have

$$\mathbf{P}(\|\mathcal{H}\| \geq qN) = \mathbf{P}(\|\mathcal{H}\|^{2n} \geq q^{2n} N^{2n}) = \mathbf{P}(\|\mathcal{H}^n\|^2 \geq q^{2n} N^{2n})$$

for all  $n \geq 1$ , where the last step follows from the fact that  $\mathcal{H}$  is self-adjoint. The Markov inequality gives

$$\mathbf{P}(\|\mathcal{H}\| \geq qN) \leq \frac{\mathbf{E}\|\mathcal{H}^n\|^2}{q^{2n} N^{2n}}. \quad (3.7)$$

Recall next that the Frobenius norm  $\|\cdot\|_F$  dominates the operator norm  $\|\mathcal{H}\| \leq \|\mathcal{H}\|_F$ , and so  $\|\mathcal{H}^n\|^2 \leq \|\mathcal{H}^n\|_F^2 = \text{Tr}(\mathcal{H}^{2n})$ , and

$$\mathbf{P}(\|\mathcal{H}\| \geq qN) \leq \frac{\mathbf{E}\|\mathcal{H}^n\|_F^2}{q^{2n} N^{2n}} = \frac{\mathbf{E}[\text{Tr}(\mathcal{H}^{2n})]}{q^{2n} N^{2n}}$$

In [3], the following bound for  $\mathbf{E}[\text{Tr}(\mathcal{H}^{2n}(\Omega'))]$  is derived:

$$\mathbf{E}[\text{Tr}(\mathcal{H}^{2n})] \leq (2n) \left( \frac{(1 + \sqrt{5})^2}{2e(1 - \tau)} \right)^n n^n |T|^{n+1} \tau^n N^n$$

for  $n \geq 4$  and  $\tau < 0.44$ . Our assumption about the size of  $|T| + \tau N$  assures us that  $\tau$  will be even smaller,  $\tau < 0.12$  so that  $(1 + \sqrt{5})^2/2(1 - \tau) \leq 6$ , whence

$$\mathbf{E}[\text{Tr}(\mathcal{H}^{2n})] \leq 2n (6/e)^n n^n |T|^{n+1} \tau^n N^n. \quad (3.8)$$

With  $\sqrt{|T| \cdot \tau N} \leq (|T| + \tau N)/2$ , this gives

$$\mathbf{E}[\text{Tr}(\mathcal{H}^{2n})] \leq 2n^{n+1} |T| e^{-n} \left( \frac{|T| + \tau N}{\rho_0} \right), \quad \rho_0 = \sqrt{2/3} > 0.8164. \quad (3.9)$$

We now specialize (3.9), and take  $n = \lfloor (\beta + 1) \log N \rfloor$  (which is greater than 4), where  $\lfloor x \rfloor$  is the largest integer less than or equal to  $x$ . Under our conditions for  $|T| + \tau N$ , we have

$$\begin{aligned} \mathbf{P}(\|\mathcal{H}\| > qN) &\leq 2e\rho_0q((\beta + 1) \log N)^{1/2} \cdot N^{-\beta} \\ &\leq 5 \cdot ((\beta + 1) \log N)^{1/2} \cdot N^{-\beta}. \end{aligned}$$

■

Combining the two lemmas, we have that in the case  $\tau \geq (\log N)^2/N$

$$\|A(\Omega')\|^2 \leq q \cdot \left( 1 + \frac{2\rho_0}{\sqrt{(\beta + 1) \log N}} \right)$$

with probability  $1 - O((\log N)^{1/2} \cdot N^{-\beta})$ . Theorem 3.1 is established by taking

$$q = \frac{1}{2} \cdot \left( 1 + \frac{2\rho_0}{\sqrt{(\beta + 1) \log N}} \right)^{-1} = 1/2 + o(1)$$

and applying (3.3). For the statement about finite sample sizes, we observe that for  $N \geq 512$  and  $\beta \geq 1$ ,  $q$  will be no lower than 0.3419, and the constant  $\rho_0q$  will be at least 0.2791.

For the case  $\tau < (\log N)^2/N$ , we have

$$\mathbf{P}(|\Omega'| > 2(\log N)^2) \leq N^{-\beta}$$

and

$$\|A(\Omega')\|^2 \leq q + \frac{2 \cdot (\log N)^2}{N}$$

with probability  $1 - O((\log N)^{1/2} \cdot N^{-\beta})$ . Taking  $q = 1/2 - 2(\log N)^2/N$  (which is also greater than 0.3419 for  $N \geq 512$ ) establishes the theorem.

For a function with Fourier transform  $\hat{f}$  supported on  $\Omega$ , we have also established

$$\|f \cdot 1_T\|^2 \leq \|F_{\Omega T}^*\|^2 \cdot \|f\|^2 \leq (1/2) \cdot \|f\|^2$$

since  $F_{\Omega T}$  and  $F_{\Omega T}^*$  have the same maximum singular value.

**Remarks.** Note that by the symmetry of the discrete Fourier transform, we could just as easily fix a subset in the frequency domain and then choose a subset of the time domain at random. Also, the choice of 1/2 can be generalized; we can show that

$$\|F_{\Omega T}^* F_{\Omega T}\| \leq c \quad \Rightarrow \quad \|\hat{f} \cdot 1_\Omega\|^2 \leq c \cdot \|f\|^2$$

for any  $c > 0$  if we make the constants in Theorem 3.1 small enough.

## 4 Robust UPs and Basis Pursuit: Spikes and Sinusoids

As in [10, 12, 15], our uncertainty principles are directly applicable to finding sparse approximations in redundant dictionaries. In this section, we look exclusively at the case of spikes and sinusoids. We will leverage Theorem 3.1 in two different ways:

1.  $\ell_0$ -uniqueness: If  $f \in \mathbb{C}^N$  has a decomposition  $\alpha$  supported on  $T \cup \Omega$  with  $|T| + |\Omega| \asymp (\log N)^{-1/2}N$ , then with high probability,  $\alpha$  is the sparsest representation of  $f$ .
2. Equivalence of  $(P_0)$  and  $(P_1)$ : If  $|T| + |\Omega| \asymp (\log N)^{-1}N$ , then  $(P_1)$  recovers  $\alpha$  with overwhelmingly large probability.

## 4.1 $\ell_0$ -uniqueness

To illustrate that it is possible to do much better than (1.10), we first consider the case in which  $N$  is a prime integer. Tao [26] derived the following exact, sharp discrete uncertainty principle.

**Lemma 4.1** [26] *Suppose that the signal length  $N$  is a prime integer. Then*

$$|\text{supp } f| + |\text{supp } \hat{f}| > N, \quad \forall f \in \mathbb{C}^N.$$

Using Lemma 4.1, a strong  $\ell_0$ -uniqueness result immediately follows:

**Corollary 4.1** *Let  $T$  and  $\Omega$  be subsets of  $\{0, \dots, N-1\}$  for  $N$  prime, and let  $\alpha$  (with  $\Phi\alpha = f$ ) be a vector supported on  $\Gamma = T \cup \Omega$  such that*

$$|T| + |\Omega| < N/2.$$

*Then the solution to  $(P_0)$  is unique and is equal to  $\alpha$ . Conversely, there exist distinct vectors  $\alpha_0, \alpha_1$  obeying  $|\text{supp } \alpha_0|, |\text{supp } \alpha_1| < N/2 + 1$  and  $\Phi\alpha_0 = \Phi\alpha_1$ .*

**Proof** As we have seen in the introduction, one direction is trivial. If  $\alpha + \delta_0$  is another decomposition, then  $\delta_0$  is of the form  $\delta_0 := (\delta, -\hat{\delta})$ . Lemma 4.1 gives  $\|\delta_0\|_{\ell_0} > N$  and thus  $\|\alpha + \delta_0\|_{\ell_0} \geq \|\delta\|_{\ell_0} - \|\alpha\|_{\ell_0} > N/2$ . Therefore,  $\|\alpha + \delta_0\|_{\ell_0} > \|\alpha\|_{\ell_0}$ .

For the converse, we know that since  $\Phi$  has rank at most  $N$ , we can find  $\delta \neq 0$  with  $|\text{supp } \delta| = N + 1$  such that  $\Phi\delta = 0$ . (Note that it is of course possible to construct such  $\delta$ 's for any support of size greater than  $N$ ). Consider a partition of  $\text{supp } \delta = \Gamma_0 \cup \Gamma_1$  where  $\Gamma_0$  and  $\Gamma_1$  are two disjoint sets with  $|\Gamma_0| = (N + 1)/2$  and  $|\Gamma_1| = (N + 1)/2$ , say. The claim follows by taking  $\alpha_0 = \delta|_{\Gamma_0}$  and  $\alpha_1 = -\delta|_{\Gamma_1}$ . ■

A slightly weaker statement addresses arbitrary sample sizes.

**Theorem 4.1** *Let  $f = \Phi\alpha$  be a signal of length  $N \geq 512$  with support set  $\Gamma = T \cup \Omega$  sampled uniformly at random with  $N_\Gamma := N_T + N_\Omega = |T| + |\Omega|$  obeying (3.1), and with coefficients  $\alpha$  sampled as in Section 2. Then with probability at least  $1 - O((\log N)^{1/2} \cdot N^{-\beta})$ , the solution to  $(P_0)$  is unique and equal to  $\alpha$ .*

To prove Theorem 4.1, we shall need the following lemma:

**Lemma 4.2** *Suppose  $T$  and  $\Omega$  are fixed subsets of  $\{0, \dots, N-1\}$  with associated restriction operators  $R_T$  and  $R_\Omega$ . Put  $\Gamma = T \cup \Omega$ , and let  $\Phi_\Gamma := \Phi R_\Gamma^*$  be the  $N \times (|T| + |\Omega|)$  matrix  $\Phi_\Gamma = (R_T^* \quad F^* R_\Omega^*)$ , where  $F$  is the discrete Fourier transform matrix. Then*

$$|\Gamma| < 2N \quad \Rightarrow \quad \dim(\text{Null}(\Phi_\Gamma)) < \frac{|\Gamma|}{2}.$$

**Proof** Obviously,

$$\dim(\text{Null}(\Phi_\Gamma)) = \dim(\text{Null}(\Phi_\Gamma^* \Phi_\Gamma)),$$

and we then write the  $|\Gamma| \times |\Gamma|$  matrix  $\Phi_\Gamma^* \Phi_\Gamma$  as

$$\Phi_\Gamma^* \Phi_\Gamma = \begin{pmatrix} I & F_{\Omega T}^* \\ F_{\Omega T} & I \end{pmatrix}$$

with  $F_{\Omega T}$  the partial Fourier transform from  $T$  to  $\Omega$   $F_{\Omega T} := R_\Omega F R_T^*$ . The dimension of the nullspace of  $\Phi_\Gamma^* \Phi_\Gamma$  is simply the number of eigenvalues of  $\Phi_\Gamma^* \Phi_\Gamma$  that are zero. Put

$$G := I - \Phi_\Gamma^* \Phi_\Gamma = \begin{pmatrix} 0 & -F_{\Omega T}^* \\ -F_{\Omega T} & 0 \end{pmatrix}, \quad \text{so that} \quad G^* G = \begin{pmatrix} F_{\Omega T}^* F_{\Omega T} & 0 \\ 0 & F_{\Omega T} F_{\Omega T}^* \end{pmatrix}.$$

Since  $G$  is symmetric,

$$\text{Tr}(G^* G) = \lambda_1^2(G) + \lambda_2^2(G) + \cdots + \lambda_{|T|+|\Omega|}^2(G), \quad (4.1)$$

where  $\lambda_1(G), \dots, \lambda_{|T|+|\Omega|}(G)$  are the eigenvalues of  $G$ . We also have that  $\text{Tr}(G^* G) = \text{Tr}(F_{\Omega T}^* F_{\Omega T}) + \text{Tr}(F_{\Omega T} F_{\Omega T}^*)$ , so the eigenvalues in (4.1) will appear in duplicate,

$$\lambda_1^2(G) + \lambda_2^2(G) + \cdots + \lambda_{|T|+|\Omega|}^2(G) = 2 \cdot (\lambda_1^2(F_{\Omega T}^* F_{\Omega T}) + \cdots + \lambda_{|T|}^2(F_{\Omega T}^* F_{\Omega T})). \quad (4.2)$$

We calculate

$$(F_{\Omega T}^* F_{\Omega T})_{t,t'} = \frac{1}{N} \sum_{\omega \in \Omega} e^{i2\pi\omega(t-t')/N} \quad (F_{\Omega T} F_{\Omega T}^*)_{\omega,\omega'} = \frac{1}{N} \sum_{t \in T} e^{i2\pi t(\omega-\omega')/N}.$$

and thus

$$\text{Tr}(G^* G) = \frac{2(|T| \cdot |\Omega|)}{N}. \quad (4.3)$$

Since the eigenvalues of  $\Phi_\Gamma^* \Phi_\Gamma$  are  $1 - \lambda_1(G), \dots, 1 - \lambda_{|T|+|\Omega|}(G)$ , at least  $K$  of the eigenvalues in (4.2) must have magnitude greater than or equal to 1 for the dimension of  $\text{Null}(\Phi_\Gamma^* \Phi_\Gamma)$  to have dimension  $K$ . As a result

$$\text{Tr}(G^* G) < 2K \Rightarrow \dim(\text{Null}(\Phi_\Gamma^* \Phi_\Gamma)) < K.$$

Using the fact that  $(a+b) \geq 4ab/(a+b)$  (arithmetic mean dominates geometric mean), we see that if  $|T| + |\Omega| < 2N$ , then  $2|T| \cdot |\Omega|/N < |T| + |\Omega|$ , yielding

$$\text{Tr}(G^* G) < |T| + |\Omega| \Rightarrow \dim(\text{Null}(\Phi_\Gamma^* \Phi_\Gamma)) < \frac{|T| + |\Omega|}{2}$$

as claimed. ■

**Proof of Theorem 4.1** We assume  $\Gamma$  is selected such that  $\Phi_\Gamma$  has full rank, meaning that  $\alpha$  is the unique representation of  $f$  on  $\Gamma$ . This happens if  $\|F_{\Omega T}^* F_{\Omega T}\| < 1$  and Theorem 3.1 states that this occurs with probability at least  $1 - O((\log N)^{1/2} \cdot N^{-\beta})$ .

Given this  $\Gamma$ , the (continuous) probability distribution on the  $\{\alpha(\gamma), \gamma \in \Gamma\}$  induces a continuous probability distribution on  $\text{Range}(\Phi_\Gamma)$ . We will show that for every  $\Gamma' \neq \Gamma$  with  $|\Gamma'| \leq |\Gamma|$

$$\dim(\text{Range}(\Phi_{\Gamma'}) \cap \text{Range}(\Phi_\Gamma)) < |\Gamma|. \quad (4.4)$$

As such, the set of signals in  $\text{Range}(\Phi_\Gamma)$  that have expansions on  $\Gamma'$  that are *at least* as sparse as their expansions on  $\Gamma$  is a finite union of subspaces of dimension strictly smaller

than  $|\Gamma|$ . This set has measure zero as a subset of  $\text{Range}(\Phi_\Gamma)$ , and hence the probability of observing such a signal is zero.

To show (4.4), we may assume that  $\Phi_{\Gamma'}$  also has full rank, since if  $\dim(\text{Range}(\Phi_{\Gamma'})) < |\Gamma'|$ , then (4.4) is certainly true. Under this premise, we will argue by contradiction. For  $\dim(\text{Range}(\Phi_{\Gamma'}) \cap \text{Range}(\Phi_\Gamma)) = |\Gamma|$  we must have  $\text{Range}(\Phi_{\Gamma'}) = \text{Range}(\Phi_\Gamma)$ . Since  $\Phi_{\Gamma \cap \Gamma'}$  is a submatrix of both  $\Phi_\Gamma$  and  $\Phi_{\Gamma'}$  (which are both full rank), it must follow that

$$\text{Range}(\Phi_{\Gamma \setminus \Gamma'}) = \text{Range}(\Phi_{\Gamma' \setminus \Gamma}).$$

As a result, to each  $\alpha$  supported on  $\Gamma \setminus \Gamma'$  there corresponds a unique vector in the nullspace of  $\Phi$  supported on  $(\Gamma \setminus \Gamma') \cup (\Gamma' \setminus \Gamma)$ :  $\delta \in \text{Null}(\Phi_{(\Gamma \setminus \Gamma') \cup (\Gamma' \setminus \Gamma)})$ . Therefore, we must have

$$\dim(\text{Null}(\Phi_{(\Gamma \setminus \Gamma') \cup (\Gamma' \setminus \Gamma)})) \geq |\Gamma \setminus \Gamma'| \geq \frac{|(\Gamma \setminus \Gamma') \cup (\Gamma' \setminus \Gamma)|}{2}.$$

But this is a direct contradiction to Lemma 4.2. Hence  $\dim(\text{Range}(\Phi_{\Gamma'}) \cap \text{Range}(\Phi_\Gamma)) < |\Gamma|$ , and the theorem follows.  $\blacksquare$

## 4.2 Recovery via $\ell_1$ -minimization

The problem  $(P_0)$  is combinatorial and solving it directly is infeasible even for modest-sized signals. This is the reason why we consider instead, the convex relaxation (1.12).

**Theorem 4.2** *Suppose  $f = \Phi\alpha$  is a signal with support set  $\Gamma = T \cup \Omega$  sampled uniformly at random with*

$$|\Gamma| = |T| + |\Omega| \leq \frac{N}{(\beta + 1) \log N} \cdot (1/4 + o(1)) \quad (4.5)$$

*and with coefficients  $\alpha$  sampled as in Section 2. Then with probability at least  $1 - O(N^{-\beta})$ , the solutions of  $(P_1)$  and  $(P_0)$  are identical and equal to  $\alpha$ .*

In addition to being computationally tractable, there are analytical advantages which come with  $(P_1)$ , as our arguments will essentially rely on a strong duality result [2]. In fact, the next section shows that  $\alpha$  is a unique minimizer of  $(P_1)$  if and only if there exists a “dual vector”  $S$  satisfying certain properties. The crucial part of the analysis relies on the fact that “partial” Fourier matrices  $F_{\Omega T} := R_\Omega F R_T^*$  have very well-behaved eigenvalues, hence the connection with robust uncertainty principles.

### 4.2.1 $\ell_1$ -duality

For a vector of coefficients  $\alpha \in \mathbb{C}^{2N}$  supported on  $\Gamma := \Gamma_1 \cup \Gamma_2$ , define the ‘sign’ vector  $\text{sgn } \alpha$  by  $(\text{sgn } \alpha)(\gamma) := \alpha(\gamma)/|\alpha(\gamma)|$  for  $\gamma \in \Gamma$  and  $(\text{sgn } \alpha)(\gamma) = 0$  otherwise. We say that  $S \in \mathbb{C}^N$  is a *dual vector* associated to  $\alpha$  if  $S$  obeys

$$(\Phi^* S)(\gamma) = (\text{sgn } \alpha)(\gamma) \quad \gamma \in \Gamma \quad (4.6)$$

$$|(\Phi^* S)(\gamma)| < 1 \quad \gamma \in \Gamma^c. \quad (4.7)$$

With this notion, we recall a standard strong duality result which is similar to that presented in [3], see also [14] and [28]. We include the proof to help keep the paper self-contained.

**Lemma 4.3** Consider a vector  $\alpha \in \mathbb{C}^{2N}$  with support  $\Gamma = \Gamma_1 \cup \Gamma_2$  and put  $f = \Phi\alpha$ .

- Suppose that there exists a dual vector and that  $\Phi_\Gamma$  has full rank. Then the minimizer  $\alpha^\sharp$  to the problem  $(P_1)$  is unique and equal to  $\alpha$ .
- Conversely, if  $\alpha$  is the unique minimizer of  $(P_1)$ , then there exists a dual vector.

**Proof** The program dual to  $(P_1)$  is

$$(D1) \quad \max_S \operatorname{Re}(S^*f) \quad \text{subject to} \quad \|\Phi^*S\|_{\ell_\infty} \leq 1. \quad (4.8)$$

It is a classical result in convex optimization that if  $\tilde{\alpha}$  is a minimizer of  $(P_1)$ , then  $\operatorname{Re}(S^*\Phi\tilde{\alpha}) \leq \|\tilde{\alpha}\|_{\ell_1}$  for all feasible  $S$ . Since the primal is a convex functional subject only to equality constraints, we will have  $\operatorname{Re}(\tilde{S}^*\Phi\tilde{\alpha}) = \|\tilde{\alpha}\|_{\ell_1}$  if and only if  $\tilde{S}$  is a maximizer of  $(D1)$  [2, Chap. 5].

First, suppose that  $\Phi R_\Gamma^*$  has full rank and that a dual vector  $S$  exists. Set  $P = \Phi^*S$ . Then

$$\begin{aligned} \operatorname{Re}\langle \Phi\alpha, S \rangle &= \operatorname{Re}\langle \alpha, \Phi^*S \rangle \\ &= \operatorname{Re} \sum_{\gamma=0}^{N-1} \overline{P(\gamma)}\alpha(\gamma) \\ &= \operatorname{Re} \sum_{\gamma \in \Gamma} \overline{\operatorname{sgn} \alpha(\gamma)}\alpha(\gamma) \\ &= \|\alpha\|_{\ell_1} \end{aligned}$$

and  $\alpha$  is a minimizer of  $(P_1)$ . Since  $|P(\gamma)| < 1$  for  $\gamma \in \Gamma^c$ , all minimizers of  $(P_1)$  must be supported on  $\Gamma$ . But  $\Phi R_\Gamma^*$  has full rank, so  $\alpha$  is the unique minimizer.

For the converse, suppose that  $\alpha$  is the unique minimizer of  $(P_1)$ . Then there exists at least one  $S$  such that with  $P = \Phi^*S$ ,  $\|P\|_{\ell_\infty} \leq 1$  and  $S^*f = \|\alpha\|_{\ell_1}$ . Then

$$\begin{aligned} \|\alpha\|_{\ell_1} &= \operatorname{Re}\langle \Phi\alpha, S \rangle \\ &= \operatorname{Re}\langle \alpha, \Phi^*S \rangle \\ &= \operatorname{Re} \sum_{\gamma \in \Gamma} \overline{P(\gamma)}\alpha(\gamma). \end{aligned}$$

Since  $|P(\gamma)| \leq 1$ , equality above can only hold if  $P(\gamma) = \operatorname{sgn} \alpha(\gamma)$  for  $\gamma \in \Gamma$ .

We will argue geometrically that for one of these  $S$ , we have  $|P(\gamma)| < 1$  for  $\gamma \in \Gamma^c$ . Let  $V$  be the hyperplane  $\{d \in \mathbb{C}^{2N} : \Phi d = f\}$ , and let  $B$  be the polytope  $B = \{d \in \mathbb{C}^{2N} : \|d\|_{\ell_1} \leq \|\alpha\|_{\ell_1}\}$ . Each of the  $S$  above corresponds to a hyperplane  $H_S = \{d : \operatorname{Re}\langle d, \Phi^*S \rangle = \|\alpha\|_{\ell_1}\}$  that contains  $V$  (since  $\operatorname{Re}\langle f, S \rangle = \|\alpha\|_{\ell_1}$ ) and which defines a halfspace  $\{d : \operatorname{Re}\langle d, \Phi^*S \rangle \leq 1\}$  that contains  $B$  (and for each such hyperplane, a  $S$  exists that describes it as such). Since  $\alpha$  is the unique minimizer, for one of these  $S'$ , the hyperplane  $H_{S'}$  intersects  $B$  only on the minimal facet  $\{d : \operatorname{supp} d \subset \Gamma\}$ , and we will have  $P(\gamma) < 1$ ,  $\gamma \in \Gamma^c$ . ■

Thus to show that  $(P_1)$  recovers a representation  $\alpha$  from a signal observation  $\Phi\alpha$ , it is enough to prove that a dual vector with properties (4.6)–(4.7) exists.

As a sufficient condition for the equivalence of  $(P_0)$  and  $(P_1)$ , we construct the *minimum energy* dual vector

$$\min \|P\|_2, \quad \text{subject to} \quad P \in \operatorname{Range}(\Phi^*) \text{ and } P(\gamma) = \operatorname{sgn}(\alpha)(\gamma), \forall \gamma \in \Gamma.$$

This minimum energy vector is somehow “small,” and we hope that it obeys the inequality constraints (4.7). Note that  $\|P\|_2 = 2\|S\|_2$ , and the problem is thus the same as finding that  $S \in \mathbb{C}^N$  with minimum norm and obeying the constraint above; the solution is classical and given by

$$S = \Phi_\Gamma(\Phi_\Gamma^* \Phi_\Gamma)^{-1} R_\Gamma \operatorname{sgn} \alpha$$

where again,  $R_\Gamma$  is the restriction operators to  $\Gamma$ . Setting

$$P = \Phi^* S = \Phi^* \Phi_\Gamma(\Phi_\Gamma^* \Phi_\Gamma)^{-1} R_\Gamma \operatorname{sgn} \alpha, \quad (4.9)$$

we need to establish that

1.  $\Phi_\Gamma^* \Phi_\Gamma$  is invertible (so that  $S$  exists), and if so
2.  $|P(\gamma)| < 1$  for  $\gamma \in \Gamma^c$ .

The next section shows that for  $|T| + |\Omega| \asymp N/\log N$ , not only is  $\Phi_\Gamma^* \Phi_\Gamma$  invertible with high probability but in addition, the eigenvalues of  $(\Phi_\Gamma^* \Phi_\Gamma)^{-1}$  are all less than two, say. These size estimates will also be very useful in showing that  $P$  is small componentwise.

#### 4.2.2 Invertibility

**Lemma 4.4** *Under the conditions of Theorem 4.2, the matrix  $\Phi_\Gamma^* \Phi_\Gamma$  is invertible and obeys*

$$\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| = 1 + o(1).$$

with probability exceeding  $1 - O(N^{-\beta})$ .

**Proof** We begin by recalling that with  $F_{\Omega T}$  as before,  $\Phi_\Gamma^* \Phi_\Gamma$  is given by

$$\Phi_\Gamma^* \Phi_\Gamma = I + \begin{pmatrix} 0 & F_{\Omega T}^* \\ F_{\Omega T} & 0 \end{pmatrix}.$$

Clearly,  $\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| = 1/\lambda_{\min}(\Phi_\Gamma^* \Phi_\Gamma)$  and since  $\lambda_{\min}(\Phi_\Gamma^* \Phi_\Gamma) \geq 1 - \sqrt{\|F_{\Omega T}^* F_{\Omega T}\|}$ , we have

$$\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \leq \frac{1}{1 - \sqrt{\|F_{\Omega T}^* F_{\Omega T}\|}}.$$

We then need to prove that  $\|F_{\Omega T}^* F_{\Omega T}\| = o(1)$  with the required probability. This follows from taking  $q = (4\rho_0((\beta+1)\log N)^{1/2})^{-1}$  in the proof of Theorem 3.1, which the condition (4.5) allows. Note that this gives more than what is claimed since

$$\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \leq 1 + \frac{1}{4\rho_0 \sqrt{(\beta+1)\log N}} + O(1/\log N).$$

Taking  $q \sim ((\beta+1)\log N)^{-1/2}$  also allows us to remove the  $\sqrt{\log N}$  term from the probability of error in Lemma 3.2. ■

**Remark.** Note that Theorem 3.1 assures us that it is sufficient to take  $|T| + |\Omega|$  of the order of  $N/\sqrt{\log N}$  (rather than of the order of  $N/\log N$  as Theorem 4.2 states) and still have invertibility with  $\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \leq 2$ , for example. The reason why we actually need the stronger condition will become apparent in the next subsection.



### 4.2.3 Proof of Theorem 4.2

To prove our theorem, it remains to show that  $|P(\gamma)| < 1$  on  $\Gamma^c$  with high probability.

**Lemma 4.5** *Under the hypotheses of Theorem 4.2, and for  $P$  as in (4.9), for each  $\gamma \in \Gamma^c$*

$$\mathbf{P}(|P(\gamma)| \geq 1) \leq 4N^{-(\beta+1)}.$$

As a result,

$$\mathbf{P}\left(\max_{\gamma \in \Gamma^c} |P(\gamma)| \geq 1\right) \leq 8N^{-\beta}.$$

**Proof** The image of the dual vector  $P$  is given by

$$P := \begin{pmatrix} P_1(t) \\ P_2(\omega) \end{pmatrix} = \Phi^* \Phi_\Gamma (\Phi_\Gamma^* \Phi_\Gamma)^{-1} R_\Gamma \operatorname{sgn} \alpha,$$

where the matrix  $\Phi^* \Phi_\Gamma$  may be expanded in the time and frequency subdomains as

$$\Phi^* \Phi_\Gamma = \begin{pmatrix} R_T^* & F^* R_\Omega^* \\ F R_T^* & R_\Omega^* \end{pmatrix}.$$

Consider first  $P_1(t)$  for  $t \in T^c$  and let  $V_t \in \mathbb{C}^{|\Gamma|}$  be the conjugate transpose of the row of the matrix  $\begin{pmatrix} R_T^* & F^* R_\Omega^* \end{pmatrix}$  corresponding to index  $t$ . For  $t \in T^c$ , the row of  $R_T^*$  with index  $t$  is zero, and  $V_t$  is then the  $(|T| + |\Omega|)$ -dimensional vector

$$V_t = \begin{pmatrix} 0 \\ \left\{ \frac{1}{\sqrt{N}} e^{-i2\pi\omega t/N}, \omega \in \Omega \right\} \end{pmatrix}.$$

These notations permit us to express  $P_1(t)$  as the inner product

$$\begin{aligned} P_1(t) &= \langle (\Phi_\Gamma^* \Phi_\Gamma)^{-1} R_\Gamma \operatorname{sgn} \alpha, V_t \rangle \\ &= \langle R_\Gamma \operatorname{sgn} \alpha, (\Phi_\Gamma^* \Phi_\Gamma)^{-1} V_t \rangle \\ &= \sum_{\gamma \in \Gamma} \overline{W(\gamma)} \operatorname{sgn} \alpha(\gamma) \end{aligned}$$

where  $W = (\Phi_\Gamma^* \Phi_\Gamma)^{-1} V_t$ . The signs of  $\alpha$  on  $\Gamma$  are statistically independent of  $\Gamma$  (and hence of  $W$ ), and therefore for a fixed support set  $\Gamma$ ,  $P_1(t)$  is a weighted sum of independent complex-valued random variables

$$P_1(t) = \sum_{\gamma \in \Gamma} X_\gamma$$

with  $\mathbf{E}X_\gamma = 0$  and  $|X_\gamma| \leq |W(\gamma)|$ . Applying the complex Hoeffding inequality (see the Appendix) gives a bound on the conditional distribution of  $P(t)$

$$\mathbf{P}(|P_1(t)| \geq 1 \mid \Gamma) \leq 4 \exp\left(-\frac{1}{4\|W\|_2^2}\right).$$

Thus it suffices to develop a bound on the magnitude of the vector  $W$ . Controlling the eigenvalues of  $\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\|$  is essential here, as

$$\|W\| \leq \|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \cdot \|V_t\|. \quad (4.10)$$

Using the fact that  $\|V_t\| = \sqrt{|\Omega|/N}$ , and Lemma 4.4 which states that  $\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \leq 1 + o(1)$  with the desired probability, we can control  $\|W\|$ :

$$\|W\|^2 \leq (1 + o(1)) \cdot |\Omega|/N.$$

This gives

$$\mathbf{P}(|P_1(t)| \geq 1) \leq 4 \exp\left(-\frac{N}{4|\Omega| \cdot (1 + o(1))}\right).$$

Our assumption that  $|\Omega|$  obeys (4.5) yields

$$\mathbf{P}(|P_1(t)| \geq 1) \leq 4 \exp(-(\beta + 1) \log N) \leq 4N^{-(\beta+1)}$$

and

$$\mathbf{P}\left(\max_{t \in T^c} |P_1(t)| \geq 1\right) \leq 4N^{-\beta}.$$

As we alluded to earlier, the bound about the size of each individual  $P(t)$  one would obtain assuming that  $|T| + |\Omega|$  be only of the order  $N/\sqrt{\log N}$  would not allow taking the supremum via the standard union bound. Our approach requires  $|T| + |\Omega|$  to be of the order  $N/\log N$ .

By the symmetry of the Fourier transform, the same is true for  $P_2(\omega)$ . This finishes the proof of Lemma 4.5 and of Theorem 4.2.  $\blacksquare$

## 5 Robust UPs and Basis Pursuit

The results of Sections 3 and 4 extend to the general situation where the dictionary  $\Phi$  is a union of two orthonormal bases  $\Phi_1, \Phi_2$ . In this section, we present results for pairs of orthogonal bases that parallel those for the time-frequency dictionary presented in previous sections. The bounds will depend critically on the degree of similarity of  $\Phi_1$  and  $\Phi_2$ , which we measure using the mutual coherence defined in (1.4),  $\mu := \mu(\Phi_1, \Phi_2)$ . As we will see, our generalization introduces additional “ $\log N$ ” factors. It is our conjecture that bounds that do not include these factors exist.

As before, the key result is the quantitative robust uncertainty principle. We use the same probabilistic model to sample the support sets  $\Gamma_1, \Gamma_2$  in the  $\Phi_1$  and  $\Phi_2$  domains respectively. The statement below is the analogue of Theorem 3.1.

**Theorem 5.1** *Let  $\Phi := (\Phi_1 \ \Phi_2)$  be a dictionary composed of a union of two orthonormal bases with mutual coherence  $\mu$ . Let  $\Gamma_2$  be a sampled uniformly at random of size  $|\Gamma_2| = N_{\Gamma_2}$ . Then with probability  $1 - O(N^{-\beta})$ , every signal  $f$  with  $\Phi f$  supported on a set  $\Gamma_1$  of size  $|\Gamma_1| = N_{\Gamma_1}$  with*

$$N_{\Gamma_1} + N_{\Gamma_2} \leq \frac{C_\beta}{\mu^2 \cdot (\log N)^6} \tag{5.1}$$

*has most of its energy in the  $\Phi_2$ -domain outside of  $\Gamma_2$ :*

$$\|\Phi_2 f \cdot 1_{\Gamma_2}\|^2 \leq \|f\|^2/2,$$

*and vice versa. As a result, for nearly all pairs  $(\Gamma_1, \Gamma_2)$  with sizes obeying (5.1), it is impossible to find a signal  $f$  supported on  $\Gamma_1$  in the  $\Phi_1$ -domain and  $\Gamma_2$  in the  $\Phi_2$ -domain.*

We would like to re-emphasize the significant difference between these results and (1.5). Namely, (5.1) effectively squares the size of the joint support since, ignoring log-like factors, the factor  $1/\mu$  is replaced by  $1/\mu^2$ . For example, in the case where the two bases are maximally incoherent, i.e.  $\mu = 1/\sqrt{N}$ , our condition says that it is nearly impossible to concentrate a function in both domains simultaneously unless (again, up to logarithmic factors)

$$|\Gamma_1| + |\Gamma_2| \sim N,$$

which needs to be compared with (1.5)

$$|\Gamma_1| + |\Gamma_2| \geq 2\sqrt{N}.$$

For mutual coherences scaling like a power-law  $\mu \sim N^{-\gamma}$ , our condition essentially reads  $|\Gamma_1| + |\Gamma_2| \sim N^{2\gamma}$  compared to  $|\Gamma_1| + |\Gamma_2| \sim N^\gamma$ .

Just as with Theorem 3.1, the key to establishing Theorem 5.1 is showing that the eigenvalues of  $A := R_{\Gamma_2} \Phi_2^* \Phi_1 R_{\Gamma_1}^*$  (playing the role of  $F_{\Omega T}^* F_{\Omega T}$  in Theorem 3.1) can be controlled. To do this, we invoke results from [4, 5]. These works establish that if  $\Gamma_2$  is selected using the Bernoulli model of Section 3, and  $\Gamma_1$  is any set with

$$|\Gamma_1| \leq \frac{C_\beta \cdot \mathbf{E}|\Gamma_2|}{(N\mu^2) \cdot (\log N)^6}, \quad (5.2)$$

then

$$\|A^* A\| \leq (3/2) \cdot \mathbf{E}|\Gamma_2|/N$$

with probability  $1 - O(N^{-\beta})$  (here and in (5.1),  $C_\beta$  is some universal constant that depends only on  $\beta$ ). Using the same reasoning as in the proof of Theorem 3.1, we can translate this result for the case where  $\Gamma_2$  is sampled under the uniform probability model of Section 2. If  $\Gamma_2$  is selected uniformly at random of size  $|\Gamma_2| = N_{\Gamma_2}$ , and  $\Gamma_1$  is any set of size  $|\Gamma_1| = N_{\Gamma_1}$ , and  $N_{\Gamma_1}, N_{\Gamma_2}$  obey (5.1), then

$$\|A^* A\| \leq (3/2) \cdot |\Gamma_2|/N \quad (5.3)$$

with the required probability. (The asymmetry between the set sizes suggested in (5.2) can be removed since  $\|A^* A\|$  can only be reduced as rows of  $A$  are removed; we really just need that  $|\Gamma_2| \leq N/3$  and  $|\Gamma_1| \lesssim N/(\log N)^6$ .) As we have seen, the size estimate (5.3) is all that is required to establish the Theorem.

Note that Theorem 5.1 is slightly stronger than Theorem 3.1. In Theorem 5.1 we generate the random subset of the  $\Phi_2$ -domain once, and (5.3) holds for all subsets of the  $\Phi_1$ -domain of a certain size. In contrast, Theorem 3.1 fixes the subset  $\Gamma_1$  before selecting  $\Gamma_2$ , meaning that the sets  $\Gamma_2$  the RUP fails on could be different for each  $\Gamma_1$ . The price we pay for this stronger statement are additional  $\log N$  factors and unknown constants.

The generalized  $\ell_0$ -uniqueness result follows directly from Theorem 5.1:

**Theorem 5.2** *Let  $f = \Phi\alpha$  be an observed signal with support set  $\Gamma = \Gamma_1 \cup \Gamma_2$  sampled uniformly at random with*

$$|\Gamma| = |\Gamma_1| + |\Gamma_2| \leq \frac{C'_\beta}{\mu^2 \cdot (\log N)^6}.$$

*and coefficients sampled as in Section 2. Then with probability  $1 - O(N^{-\beta})$ , the solution to  $(P_0)$  is unique and equal to  $\alpha$ .*

The only change to the proof presented in Section 4.1 is in the equivalent of Lemma 4.2. When calculating the trace of  $G^*G$ , each term can be bounded by  $\mu^2$  (instead of  $1/N$ ), and so  $\text{Tr}(G^*G) \leq 2(|\Gamma_1| \cdot |\Gamma_2|)\mu^2$ . As a result, if  $|\Gamma| < 2/\mu^2$ , then  $\dim(\text{Null}(Q_\Gamma)) < \frac{|\Gamma|}{2}$ , and generalized  $\ell_0$ -uniqueness is established.

The conditions for the equivalence of  $(P_0)$  and  $(P_1)$  can also be generalized.

**Theorem 5.3** *Let  $f = \Phi\alpha$  be a random signal generated as in Theorem 5.2 with*

$$|\Gamma_1| + |\Gamma_2| \leq \frac{C''_\beta}{\mu^2 \cdot (\log N)^6}.$$

*Then with probability  $1 - O(N^{-\beta})$ , the solutions of  $(P_0)$  and  $(P_1)$  are identical and equal to  $\alpha$ .*

The proof of Theorem 5.3 is again almost exactly the same as what we have already seen. Using Theorem 5.1, the eigenvalues of  $(\Phi_\Gamma^* \Phi_\Gamma)^{-1}$  are controlled, allowing us to construct a dual vector meeting the conditions (4.6) and (4.7) of Section 4.2.1. Note that the  $(\log N)^6$  term in the denominator means that  $\mathbf{P}(|P(\gamma)| < 1)$ ,  $\gamma \in \Gamma^c$  goes to zero at a much faster speed than a negative power of  $N$ , it decays as  $\exp(-\rho(\log N)^6)$  for some positive constant  $\rho > 0$ .

## 6 Numerical Experiments

From a practical standpoint, the ability of  $(P_1)$  to recover sparse decompositions is nothing short of amazing. To illustrate this fact, we consider a 256 point signal composed of 60 spikes and 60 sinusoids;  $|T| + |\Omega| \approx N/2$ , see Figure 1. Solving  $(P_1)$  recovers the original decomposition *exactly*.

We then empirically validate the previous numerical result by repeating the experiment for various signals and sample sizes, see Figure 2. These experiments were designed as follows:

1. set  $N_\Gamma$  as a percentage of the signal length  $N$ ;
2. select a support set  $\Gamma = T \cup \Omega$  of size  $|\Gamma| = N_\Gamma$  uniformly at random;
3. sample a vector  $\alpha$  on  $\Gamma$  with independent and identically distributed Gaussian entries<sup>1</sup>;
4. make  $f = \Phi\alpha$ ;
5. solve  $(P_1)$  and obtain  $\hat{\alpha}$ ;
6. compare  $\alpha$  to  $\hat{\alpha}$ ;
7. repeat 100 times for each  $N_\Gamma$ ;
8. repeat for signal lengths  $N = 256, 512, 1024$ .

---

<sup>1</sup>The results presented here do not seem to depend on the actual distribution used to sample the coefficients.

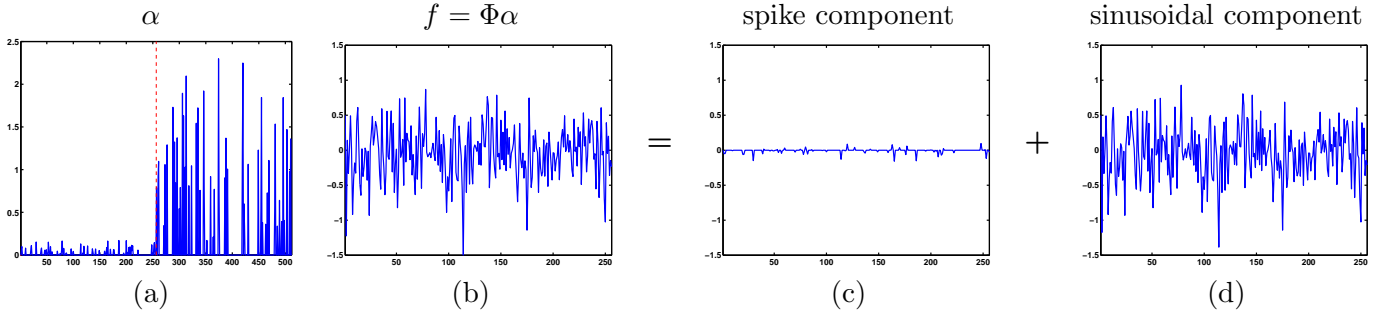


Figure 1: Recovery of a “sparse” decomposition. (a) Magnitudes of a randomly generated coefficient vector  $\alpha$  with 120 nonzero components. The spike components are on the left (indices 1–256) and the sinusoids are on the right (indices 257–512). The spike magnitudes are made small compared to the magnitudes of the sinusoids for effect; we cannot locate the spikes by inspection from the observed signal  $f$ , whose real part is shown in (b). Solving  $(P_1)$  separates  $f$  into its spike (c) and sinusoidal components (d) (the real parts are plotted).

Figure 2(a) shows that we are numerically able to recover “sparse” superpositions of spikes and sinusoids when  $|T| + |\Omega|$  is close to  $N/2$ , at least for this range of sample sizes  $N$  (we use the quotations since decompositions of this order can hardly be considered sparse). Figure 2(b) plots the success rate of the sufficient condition for the recovery of the sparsest  $\alpha$  developed in Section 4.2.1 (i.e. the minimum energy signal is a dual vector). Numerically, the sufficient condition holds when  $|T| + |\Omega|$  is close to  $N/5$ .

The time-frequency dictionary is special in that it is maximally incoherent ( $\mu = 1$ ). But as suggested in [10], incoherence between two bases is the rule, rather than the exception. To illustrate this, the above experiment was repeated for  $N = 256$  with a dictionary that is a union of the spike basis and of an orthobasis sampled uniformly at random (think about orthogonalizing  $N$  vectors sampled independently and uniformly on the unit-sphere of  $\mathbb{C}^N$ ). As shown in Figure 3, the results are very close to those obtained with time-frequency dictionaries; we recover “sparse” decompositions of size about  $|\Gamma_1| + |\Gamma_2| \leq 0.4 \cdot N$ .

## 7 Discussion

In this paper, we have demonstrated that except for a negligible fraction of pairs  $(T, \Omega)$ , the behavior of the discrete uncertainty relation is very different from what worst case scenarios—which have been the focus of the literature thus far—suggest. We introduced probability models and a robust uncertainty principle showing that for nearly all pairs  $(T, \Omega)$ , it is actually impossible to concentrate a discrete signal on  $T$  and  $\Omega$  simultaneously unless the size of the joint support  $|T| + |\Omega|$  be at least of the order of  $N/\sqrt{\log N}$ . We derived significant consequences of this new uncertainty principle, showing how one can recover sparse decompositions by solving simple convex programs.

Our sampling models were selected in perhaps the most natural way, giving each subset of the time and frequency domains with set size an equal chance of being selected. Now there is little doubt that conclusions similar to those derived in this paper would hold for other probability models. In fact, our analysis develops a machinery amenable to other setups. The centerpiece is the study of the singular values of partial Fourier transforms. For other sampling models such as models biased toward low or high frequencies for example,

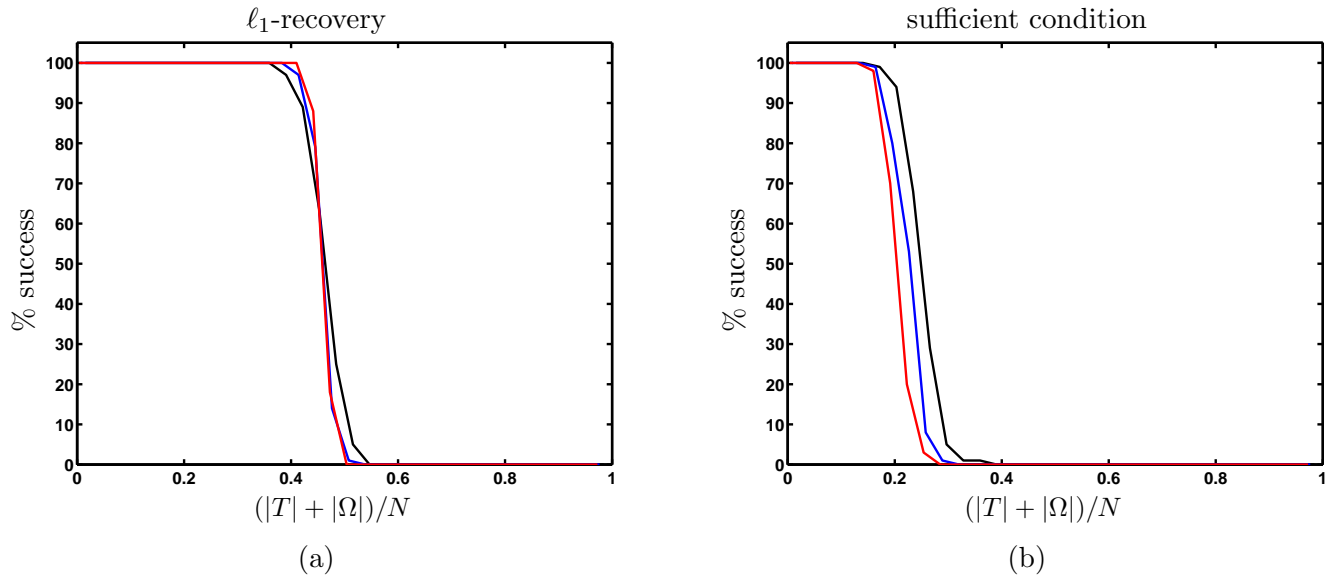


Figure 2:  $\ell_1$ -recovery for the time-frequency dictionary. (a) Success rate of  $(P_1)$  in recovering the sparsest decomposition versus the number of nonzero terms. (b) Success rate of the sufficient condition (the minimum energy signal is a dual vector).

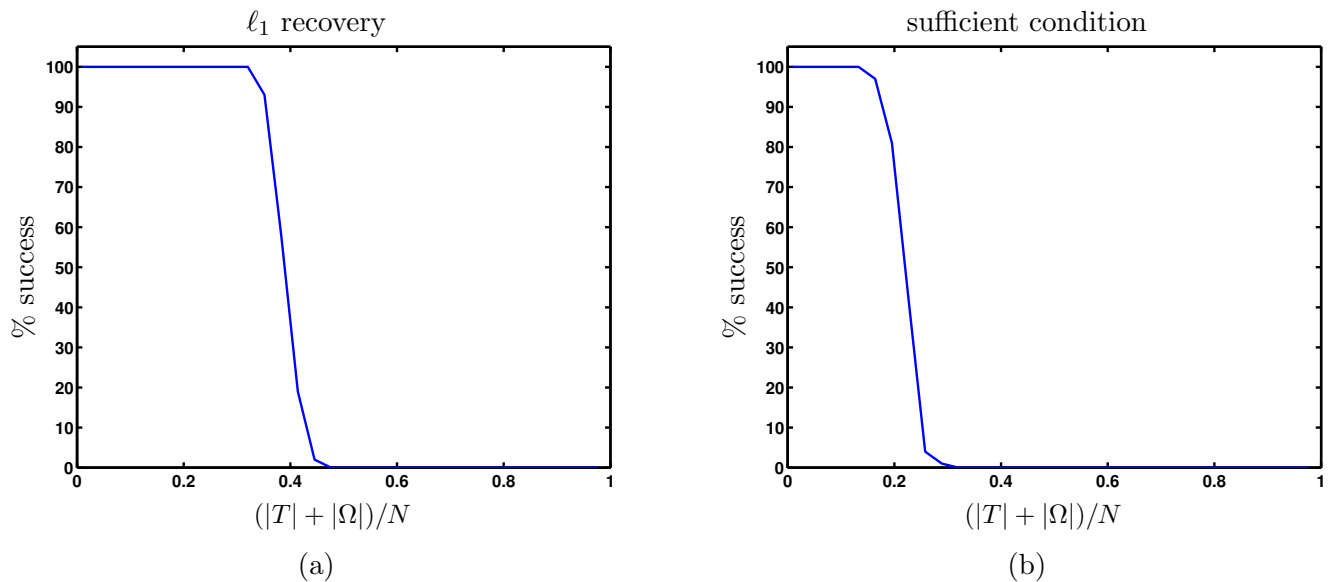


Figure 3:  $\ell_1$ -recovery for the spike-random dictionary. (a) Success rate of  $(P_1)$  in recovering the sparsest decomposition versus the number of nonzero terms. (b) Success rate of the sufficient condition.

one would need to develop analogues of Theorem 3.1. Our machinery would then nearly automatically transform these new estimates into corresponding claims.

In conclusion, we would like to mention areas for possible improvement and refinement. First, although we have made an effort to obtain explicit constants in all our statements (with the exception of section 5), there is little doubt that a much more sophisticated analysis would yield better estimates for the singular values of partial Fourier transforms, and thus provide better constants. Another important question we shall leave for future research, is whether the  $1/\sqrt{\log N}$  factor in the QRUP (Theorem 3.1) and the  $1/\log N$  for the exact  $\ell_1$ -reconstruction (Theorem 4.2) are necessary. Finally, we already argued that one really needs to randomly sample the support to derive our results but we wonder whether one needs to assume that the signs of the coefficients  $\alpha$  (in  $f = \Phi\alpha$ ) need to be randomized as well. Or would it be possible to show analogs of Theorem 4.2 ( $\ell_1$  recovers the sparsest decomposition) for all  $\alpha$ , provided that the support of  $\alpha$  may not be too large (and randomly selected)? Recent work [5,6] suggests that this might be possible—at the expense of additional logarithmic factors.

## 8 Appendix: Concentration-of-Measure Inequalities

The Hoeffding inequality is a well-known large deviation bound for sums of independent random variables. For a proof and interesting discussion, see [19].

**Lemma 8.1** (*Hoeffding inequality*) *Let  $X_0, \dots, X_{N-1}$  be independent real-valued random variables such that  $\mathbf{E}X_j = 0$  and  $|X_j| \leq a_j$  for some positive real numbers  $a_j$ . For  $\epsilon > 0$*

$$\mathbf{P} \left( \left| \sum_{j=0}^{N-1} X_j \right| \geq \epsilon \right) \leq 2 \exp \left( -\frac{\epsilon^2}{2\|\mathbf{a}\|_2^2} \right)$$

where  $\|\mathbf{a}\|_2^2 = \sum_j a_j^2$ .

**Lemma 8.2** (*complex Hoeffding inequality*) *Let  $X_0, \dots, X_{N-1}$  be independent complex-valued random variables such that  $\mathbf{E}X_j = 0$  and  $|X_j| \leq a_j$ . Then for  $\epsilon > 0$*

$$\mathbf{P} \left( \left| \sum_{j=0}^{N-1} X_j \right| \geq \epsilon \right) \leq 4 \exp \left( -\frac{\epsilon^2}{4\|\mathbf{a}\|_2^2} \right).$$

**Proof** Separate the  $X_j$  into their real and imaginary parts;  $X_j^r = \operatorname{Re} X_j$ ,  $X_j^i = \operatorname{Im} X_j$ . Clearly,  $|X_j^r| \leq a_j$  and  $|X_j^i| \leq a_j$ . The result follows immediately from Lemma 8.1 and the fact that

$$\mathbf{P} \left( \left| \sum_{j=0}^{N-1} X_j \right| \geq \epsilon \right) \leq \mathbf{P} \left( \left| \sum_{j=0}^{N-1} X_j^r \right| \geq \epsilon/\sqrt{2} \right) + \mathbf{P} \left( \left| \sum_{j=0}^{N-1} X_j^i \right| \geq \epsilon/\sqrt{2} \right).$$

■

## References

- [1] S. Boucheron, G. Lugosi, and P. Massart, A sharp concentration inequality with applications, *Random Structures Algorithms* **16** (2000), 277–292.
- [2] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [3] E. J. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, Technical Report, California Institute of Technology. Submitted to *IEEE Transactions on Information Theory*, June 2004. Available on the ArXiv preprint server: [math.GM/0409186](https://arxiv.org/abs/math.GM/0409186).
- [4] E. J. Candès, and J. Romberg, The Role of Sparsity and Incoherence for Exactly Reconstructing a Signal from Limited Measurements, Technical Report, California Institute of Technology.
- [5] E. J. Candès, and T. Tao, Near optimal signal recovery from random projections: universal encoding strategies? Submitted to *IEEE Transactions on Information Theory*, October 2004. Available on the ArXiv preprint server: [math.CA/0410542](https://arxiv.org/abs/math.CA/0410542).
- [6] E. J. Candès, and T. Tao, Decoding of random linear codes. Manuscript, October 2004.
- [7] S. S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM J. Scientific Computing* **20** (1999), 33–61.
- [8] D. L. Donoho, *Geometric separation using combined curvelet/wavelet representations*. Lecture at the International Conference on Computational Harmonic Analysis, Nashville, Tennessee, May 2004.
- [9] D. L. Donoho, P. B. Stark, Uncertainty principles and signal recovery, *SIAM J. Appl. Math.* **49** (1989), 906–931.
- [10] D. L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Transactions on Information Theory*, **47** (2001), 2845–2862.
- [11] D. L. Donoho and M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell_1$  minimization. *Proc. Natl. Acad. Sci. USA* **100** (2003), 2197–2202.
- [12] M. Elad and A. M. Bruckstein, A generalized uncertainty principle and sparse representation in pairs of  $\mathbb{R}^N$  bases, *IEEE Transactions on Information Theory*, **48** (2002), 2558–2567.
- [13] A. Feuer and A. Nemirovsky, On sparse representations in pairs of bases, Accepted to the *IEEE Transactions on Information Theory* in November 2002.
- [14] J. J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Transactions on Information Theory*, **50** (2004), 1341–1344.
- [15] R. Gribonval and M. Nielsen, Sparse representations in unions of bases. *IEEE Trans. Inform. Theory* **49** (2003), 3320–3325.



- [16] K. Jogdeo and S. M. Samuels, Monotone convergence of binomial probabilities and a generalization of Ramanujan’s equation. *Annals of Mathematical Statistics*, **39** (1968), 1191–1195.
- [17] H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty II, *Bell Systems Tech. Journal*, **40** (1961), pp. 65–84.
- [18] H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty III, *Bell Systems Tech. Journal*, **41** (1962), pp. 1295–1336.
- [19] G. Lugosi, Concentration-of-measure Inequalities, Lecture Notes.
- [20] F. G. Meyer, A. Z. Averbuch and R. R. Coifman, Multi-layered Image Representation: Application to Image Compression, *IEEE Transactions on Image Processing*, **11** (2002), 1072-1080.
- [21] D. Slepian and H. O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty I, *Bell Systems Tech. Journal*, **40** (1961), pp. 43–64.
- [22] D. Slepian, Prolate spheroidal wave functions, Fourier analysis and uncertainty V — the discrete case, *Bell Systems Tech. Journal*, **57** (1978), pp. 1371–1430.
- [23] J. L. Starck, E. J. Candès, and D. L. Donoho. Astronomical image representation by the curvelet transform *Astronomy & Astrophysics*, **398** 785–800.
- [24] J. L. Starck, M. Elad, and D. L. Donoho. Image Decomposition Via the Combination of Sparse Representations and a Variational Approach. Submitted to *IEEE Transactions on Image Processing*, 2004.
- [25] P. Stevenhagen, H. W. Lenstra Jr., Chebotarëv and his density theorem, *Math. Intelligencer* **18** (1996), no. 2, 26–37.
- [26] T. Tao, An uncertainty principle for cyclic groups of prime order, preprint. math.CA/0308286
- [27] J. A. Tropp, Greed is good: Algorithmic results for sparse approximation, Technical Report, The University of Texas at Austin, 2003.
- [28] J. A. Tropp, Recovery of short, complex linear combinations via  $\ell_1$  minimization. To appear in *IEEE Transactions on Information Theory*, 2005.