Quantitative Robust Uncertainty Principles and Optimally Sparse Decompositions

Emmanuel J. Candès and Justin Romberg
Applied and Computational Mathematics, Caltech, Pasadena, CA 91125

November 11, 2004

Abstract

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

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 (α_1, α_2) using spike and frequency locations in T and Ω respectively, and obeying

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

then (α_1, α_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 (α_1, α_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, Ω) for which the above properties do not hold can be upper bounded by quantities like $N^{-\alpha}$ for large values of α.

The QRUP (and the application to sparse approximation) can be extended to general pairs of orthogonal bases Φ_1, Φ_2 of $\mathbb{C}^N$. For almost all choices $\Gamma_1, \Gamma_2 \subset \{0, \ldots, N-1\}$ obeying

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

there is no signal f such that $\Phi_1 f$ is supported on Γ_1 and $\Phi_2 f$ is supported on Γ_2 where $\mu(\Phi_1, \Phi_2)$ is the *mutual incoherence* between Φ_1 and Φ_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 Δt 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 setups, namely, by investigating to what extent it is possible to concentrate a function f and its Fourier transform $\hat{f}$, relaxing the assumption that f and $\hat{f}$ be concentrated near intervals as in the work of Landau, Pollack and Slepian [16, 17, 20], or by considering signals supported on a discrete set [9, 21].

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

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

where the frequency index ω ranges over the set $\{0, 1, \ldots, N-1\}$. 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

$$|\operatorname{supp} f| + |\operatorname{supp} \hat{f}| \geq 2\sqrt{N}. \tag{1.2}$$

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). [9] also observed 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: we 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

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

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

In recent years, uncertainty relations have become very popular, in part because they help explaining some miraculous properties of ℓ_1-minimization procedures as we will see below, and 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 Φ_1, Φ_2 of orthonormal bases. Define the *mutual incoherence* [10, 12, 15] between Φ_1 and Φ_2 as

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

then if α_1 is the (unique) representation of f in basis Φ_1 with $\Gamma_1 = \operatorname{supp} \alpha_1$, and α_2 is the representation in Φ_2, the supports must obey

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

Note that the mutual incoherence μ always obeys $1/\sqrt{N} \leq \mu \leq 1$ and measures how the two bases look alike. The smaller the incoherence, the stronger the uncertainty relation. To see how this generalizes the discrete uncertainty principle, observe that in the case where Φ_1 is the canonical or spike basis and Φ_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 the Dirac comb for instance; here 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 [25]

$$|\operatorname{supp} f| + |\operatorname{supp} \hat{f}| > N, \tag{1.6}$$

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

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

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

Then with overwhelming high probability—in fact, exceeding $1 - O(N^{-\beta\rho})$ for some positive constant ρ (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 Ω in the frequency domain. In other words, remove a negligible fraction of sets and

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

holds, not (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, Ω) as before and put 1_Ω to be the indicator function of the set Ω. Then with essentially the same probability as above, we have

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

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

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

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

1.4 Significance of uncertainty principles

In the last three years or so, there has been a series of papers starting with [10] establishing a link between discrete uncertainty principles and sparse approximation [10, 11, 15, 26]. 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 Φ_1, Φ_2; we search among all the decompositions (α_1, α_2) of the signal f

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

for the shortest one

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

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

The discrete uncertainty principles (1.2) and (1.5) are useful in the sense that they tell us when (P_0) has a unique solution. When Φ is the time-frequency dictionary, 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 Ω, and

$$|T| + |\Omega| < \sqrt{N}, \tag{1.10}$$

then α is the unique minimizer of (P_0) [10]. In a nutshell, the reason is that if $\Phi(\alpha_0 + \delta_0)$ were another decomposition, δ_0 would obey $\Phi\delta_0 = 0$ which says that δ_0 would be of the

form $\delta_0 = (\delta, -\hat{\delta})$. Now (1.2) implies that δ_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 α_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 ℓ_0-uniqueness is guaranteed when

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

Unfortunately, as far as finding the sparsest decomposition, 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 Φ, an algorithm that is combinatorial in nature and has exponential complexity. Instead of solving (P_0), we consider a similar program in the ℓ_1 norm which goes by the name of *Basis Pursuit* [7]:

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

Unlike the ℓ_0 norm, the ℓ_1 norm is convex. As a result, (P_1) can be solved efficiently using standard "off the shelf" optimization algorithms. The ℓ_1-norm can also be viewed as a "sparsity norm" which among the vectors that meet the constraints, 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 α supported on Γ with

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

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 ℓ_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 [22] 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 [23] (see also [19] 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 $T := \operatorname{supp} f$. 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 ℓ_1-norm. We will draw on the 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 its 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 α (with spike locations T and sinusoidal frequencies Ω) of a signal f exists with

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

then α is the sparsest representation of f almost all of the time. If in addition, T and Ω satisfy

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

then α 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 Φ_1, Φ_2.

As discussed earlier, there is by now a well-established machinery that allows turning uncertainty relations into statements about the 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, of course, is completely new.

1.6 Organization of the paper

In Section 2 we develop a probability model that shall 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 (Φ_1, Φ_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 Γ_1, Γ_2

To state our results precisely, we first need to specify a probabilistic model. We let I_1 and I_2 be two independent Bernoulli sequences with parameters p_T and p_Ω respectively

$$\begin{aligned} I_1(t) = 1 \qquad & \text{with probability } \ p_T \\ I_2(\omega) = 1 \qquad & \text{with probability } \ p_\Omega \end{aligned}$$

where $t, \omega = 0, \ldots, N-1$, and define the support sets for the spikes and sinusoids (and in general for the bases Φ_1 and Φ_2) as

$$T = \{t \ \text{ s.t. } \ I_1(t) = 1\}, \qquad \Omega = \{\omega \ \text{ s.t. } \ I_2(\omega) = 1\}. \tag{2.1}$$

If both p_T and p_Ω are not too small, an application of the standard large deviations inequality shows us that our model is approximately equivalent to sampling $\mathbf{E}|T| = p_T \cdot N$ spike locations and $\mathbf{E}|\Omega| = p_\Omega \cdot N$ frequency locations uniformly at random.

As we will see in the next section, the robust uncertainty principle holds—with overwhelming probability—over sets T and Ω 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 $\beta > 0$. As a consequence, we will always assume that

$$\min(\mathbf{E}|T|, \mathbf{E}|\Omega|) \geq 4(\beta + 1) \cdot \log N \tag{2.2}$$

as otherwise, one would have to consider situations in which T or Ω (or both) are empty sets—a situation of rather limited interest. We also note that for p_T and p_Ω as above, we have

$$\mathbf{P}(|T| > 2p_T \cdot N) \leq N^{-\beta}, \tag{2.3}$$

as this follows from the well-known large deviation bound [1]

$$\mathbf{P}(|T| > \mathbf{E}|T| + t) \leq \exp\left(-\frac{t^2}{2\mathbf{E}|T| + 2t/3}\right).$$

Further, to establish sparse approximation bounds (section 4), we will also introduce a probability model on the "active" coefficients. Given a pair (T, Ω), we sample the coefficient vector $\{\alpha(\gamma), \ \gamma \in \Gamma\}$ 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 (2.1), 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.

Theorem 3.1 *Assume the parameters in the model* (2.1) *obey*

$$2\sqrt{\mathbf{E}|T| \cdot \mathbf{E}|\Omega|} \le \mathbf{E}|T| + \mathbf{E}|\Omega| \le \frac{N}{\sqrt{(\beta+1)\log N}}\,(\rho_0/2 + o(1)), \qquad \rho_0 = .7614 \tag{3.1}$$

(we will assume throughout the paper that $\beta \ge 1$ and $N \ge 512$) and let (T, Ω) be a randomly sampled support pair. Then with probability at least $1 - O(\log N \cdot N^{-\beta})$; every *signal f supported on T in the time domain has most of its energy in the frequency domain outside of Ω*

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

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

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

As a result, it is impossible to find a signal f supported on T whose discrete Fourier transform $\hat{f}$ is supported on Ω. For finite sample sizes N, we can select the parameters in (3.1) *as*

$$\mathbf{E}|T| + \mathbf{E}|\Omega| \le \frac{.2660\,N}{\sqrt{(\beta+1)\log N}}.$$

To establish this result, we introduce (as in [3]) the $|T| \times |T|$ auxiliary matrix $\mathcal{H}_T$

$$\mathcal{H}_T(t,t') = \begin{cases} 0 & t = t' \\ \sum_{\omega\in\Omega} e^{i\omega(t-t')} & t \ne t' \end{cases}. \tag{3.2}$$

The following lemma effectively says that the eigenvalues of $\mathcal{H}_T$ are small compared to N.

Lemma 3.1 *Fix q in $(0,1)$ and suppose that*

$$p_T + p_\Omega \le \rho_0 \cdot \frac{q}{\sqrt{(\beta+1)\log N}}, \qquad \rho_0 = .7614.$$

Then the the matrix $\mathcal{H}_T$ obeys

$$\mathbf{P}\left(\|\mathcal{H}_T\| \ge qN\right) \le (\beta+1)\log N \cdot N^{-\beta}.$$

Proof The Markov inequality gives

$$\mathbf{P}(\|\mathcal{H}_T\| \ge qN) \le \frac{\mathbf{E}\|\mathcal{H}_T^n\|^2}{q^{2n}N^{2n}}, \quad \text{for all } n \ge 1. \tag{3.3}$$

Recall next that the Frobenius norm $\|\cdot\|_F$ dominates the operator norm $\|\mathcal{H}_T\| \le \|\mathcal{H}_T\|_F$. This fact allows to leverage results from [3] which derives bounds for the conditional expectation $\mathbf{E}[\|\mathcal{H}_T^n\|_F^2 \,|\, T]$ (where the expectation is over Ω for a *fixed* T):

$$\mathbf{E}[\|\mathcal{H}_T^n\|_F^2 \,|\, T] \le (2n) \left(\frac{(1+\sqrt{5})^2}{2e(1-p_\Omega)} \right)^n n^n |T|^{n+1} p_\Omega^n N^n.$$

Our assumption about the size of $p_T + p_\Omega$ assures that $p_\Omega < .12$ so that $(1+\sqrt{5})^2/2(1-p_\Omega) \le 6$, whence

$$\mathbf{E}[\|\mathcal{H}_T^n\|_F^2 \,|\, T] \le 2n\,(6/e)^n n^n |T|^{n+1} p_\Omega^n N^n. \tag{3.4}$$

We will argue below that for $n \le (\beta+1)\log N$ and p_T obeying (2.2)

$$\mathbf{E}[|T|^{n+1}] \le 1.15^{n+1}\,[\mathbf{E}|T|]^{n+1} = 1.15^{n+1}\,(p_T\,N)^{n+1}. \tag{3.5}$$

Since $p_T \le .25$, we established

$$\mathbf{E}\|\mathcal{H}_T^n\|_F^2 \le (6 \times 1.15/e)^n\, n^{n+1} \cdot p_T^n\, p_\Omega^n\, N^{2n+1}.$$

Observe now that together with $\sqrt{p_T\, p_\Omega} \le (p_T + p_\Omega)/2$, this gives

$$\mathbf{P}(\|\mathcal{H}_T\| \ge qN) \le \left(\frac{p_T + p_\Omega}{\rho_0\, q} \right)^{2n} e^{-n}\, n^{n+1} N, \qquad \rho_0 = 1/\sqrt{6 \times 1.15} = .7614. \tag{3.6}$$

We now specialize (3.6) and take $n = \lceil (\beta+1)\log N \rceil$ where $\lceil x \rceil$ is the smallest integer greater or equal to x. Then if $p_T + p_\Omega$ obeys (3.1),

$$\mathbf{P}(\|\mathcal{H}_T\| \ge qN) \le [(\beta+1)\log N + 1] \cdot N^{-\beta}, \tag{3.7}$$

as claimed. ∎

We now return to (3.5) and write $|T|$ as

$$|T| = \mathbf{E}|T| \cdot (1+Y), \qquad Y = \frac{|T| - \mathbf{E}|T|}{\mathbf{E}|T|}.$$

Then

$$\mathbf{E}|T|^{n+1} = (\mathbf{E}|T|)^{n+1} \cdot \mathbf{E}(1+Y)^{n+1} \le (\mathbf{E}|T|)^{n+1} \cdot \mathbf{E}[\exp((n+1)Y)].$$

Observe that Y is a an affine function of a sum of independent Bernoulli random variables. Standard calculations then give

$$\mathbf{E}[\exp(nY)] = e^{-n} \cdot \left(1 + \frac{n}{N} \frac{e^\lambda - 1}{\lambda} \right)^N, \qquad \lambda = n/(Np_T).$$

Recall the assumption (2.2) which implies $\lambda \le 1/4$ which in turn gives $\lambda^{-1}(e^\lambda - 1) - 1 \le \log 1.15$. The claim follows.

We would like to remark that (3.5) might be considerably improved when $\mathbf{E}|T| = p_T \cdot N$ is much larger than n since in that case, the binomial will have enhanced concentration around its mean. For example,

$$\mathbf{E}|T|^{n+1} \le 2 \cdot [\mathbf{E}|T|]^{n+1}$$

in the event where $n \leq \rho \cdot \sqrt{p_T\, N}$ for some positive constant ρ that the above method allows to calculate explicitely. This would of course lead to improved constants in (3.1) and in the statement of Lemma 3.1. In this paper, we shall not pursue all these refinements as not to clutter the exposition.

Proof of Theorem 3.1 Let $f \in \mathbb{C}^N$ be supported on T; as such, $R_T^* R_T f = f$, where R_T is the restriction operator to T. Put $F_{\Omega T} = R_\Omega F R_T^*$. We have

$$\|\hat{f} \cdot 1_\Omega\|_2 = \|F_{\Omega T} f\|_2 \leq \|F_{\Omega T}\| \cdot \|f\|_2,$$

and since, $\|F_{\Omega T}\|^2 = \|F_{\Omega T}^* F_{\Omega T}\|$, it will suffice to show that that with high probability, the largest eigenvalue of $F_{\Omega T}^* F_{\Omega T}$ is less than $1/2$.

Using the definition of the auxiliary matrix in (3.2), it is not hard to verify the identity $F_{\Omega T}^* F_{\Omega T} = \frac{|\Omega|}{N} I + \frac{1}{N} \mathcal{H}_T$. Suppose that $p_T + p_\Omega$ obeys the condition in Lemma 3.1; then except for a set of probability less than $O(\log N \cdot N^{-\beta})$,

$$\frac{|\Omega|}{N} \leq 2p_\Omega \leq 2\rho_0 \cdot \frac{q}{\sqrt{(\beta+1)\log N}}, \quad \text{and} \quad \frac{1}{N}\|\mathcal{H}_T\| \leq q,$$

and, therefore,

$$\|F_{\Omega T}^* F_{\Omega T}\| \leq q \cdot \left(1 + \frac{2\rho_0}{\sqrt{(\beta+1)\log N}}\right) = q(1 + o(1)). \tag{3.8}$$

The theorem follows from taking $q = 1/2 + o(1)$. For the statement about finite sample sizes, we observe that for $\beta \geq 1$ and $N \geq 512$, $2/\sqrt{(\beta+1)\log N} \leq .567$ and, therefore, $\|F_{\Omega T}^* F_{\Omega T}\| \leq 1/2$ provided that $q \leq [2(1 + .567\rho_0)]^{-1}$. This establishes the first part of the theorem.

By symmetry of the discrete Fourier transform, the claim about the size of $\|f \cdot 1_T\|$ for $\hat{f}$ supported on a random set Ω is proven exactly in the same way. This concludes the proof of the theorem. ∎

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. ℓ_0-uniqueness: If $f \in \mathbb{C}^N$ has a decomposition α supported on $T \cup \Omega$ with $|T| + |\Omega| \asymp (\log N)^{-1/2} N$, then with high probability, α 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 α with overwhelmingly large probability.

4.1 ℓ_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 [25] derived the following exact, sharp discrete uncertainty principle.

Lemma 4.1 *[25] Suppose that the sample size N is a prime integer. Then*

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

Using Lemma 4.1, a strong ℓ_0-uniqueness result immediately follows:

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

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

Then the solution to (P_0) is unique and is equal to α. Conversely, there exist distinct vectors α_0, α_1 obeying $|\operatorname{supp} \alpha_0|, |\operatorname{supp} \alpha_1| \leq N/2 + 1$ and $\Phi\alpha_0 = \Phi\alpha_1$.

Proof As we have seen in the introduction, one direction is trivial. If $\alpha_0 + \delta_0$ is another decomposition, then δ_0 is of the form $\delta_0 := (\delta, -\hat{\delta})$. Lemma 4.1 gives $\|\delta_0\|_{\ell_0} > N$ and thus $\|\alpha_0 + \delta_0\|_{\ell_0} \geq \|\delta\|_{\ell_0} - \|\alpha\|_{\ell_0} > N/2$. Therefore, $\|\alpha_0 + \delta_0\|_{\ell_0} > \|\alpha\|_{\ell_0}$.

For the converse, we know that since Φ has rank at most N, we can find $\delta \neq 0$ with $|\operatorname{supp} \delta| = N + 1$ such that $\Phi\delta = 0$. (Note that it is of course possible to construct such δ's for any support of size greater than N). Consider a partition of $\operatorname{supp} \delta = \Gamma_0 \cup \Gamma_1$ where Γ_0 and Γ_1 are two disjoint sets with $|\Gamma_0| = N/2 + 1$ and $|\Gamma_1| = N/2$, 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 with support set $\Gamma = T \cup \Omega$ and coefficients α sampled as in Section 2, and with parameters obeying (3.1). Then with probability at least $1 - O(\log N \cdot N^{-\beta})$, the solution to (P_0) is unique and equal to α.*

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

Lemma 4.2 *Suppose T and Ω are fixed subsets of $\{0, \ldots, N-1\}$, put $\Gamma = T \cup \Omega$, and let $\Phi_\Gamma := \Phi R_\Gamma^*$ be the $N \times (|T| + |\Omega|)$ matrix $\Phi_\Gamma = \begin{pmatrix} R_T^* & F^* R_\Omega^* \end{pmatrix}$. Then*

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

Proof Obviously,

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

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 Ω $F_{\Omega T} := R_\Omega F R_T^*$. 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

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

Letting $\lambda_j(\cdot)$ denotes the jth largest eigenvalue of a matrix, observe that $\lambda_j(\Phi_\Gamma^*\Phi_\Gamma) = 1 - \lambda_j(G)$, and since G is symmetric

$$\mathrm{Tr}(G^*G) = \lambda_1^2(G) + \lambda_2^2(G) + \cdots + \lambda_{|T|+|\Omega|}^2(G). \tag{4.1}$$

We also have that $\mathrm{Tr}(G^*G) = \mathrm{Tr}(F_{\Omega T}^* F_{\Omega T}) + \mathrm{Tr}(F_{\Omega T} F_{\Omega T}^*)$, so the eigenvalues in (4.1) will appear in duplicate,

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

We calculate

$$(F_{\Omega T}^* F_{\Omega T})_{t,t'} = \frac{1}{N} \sum_{\omega \in \Omega} e^{i\omega(t-t')} \qquad (F_{\Omega T} F_{\Omega T}^*)_{\omega,\omega'} = \frac{1}{N} \sum_{t \in T} e^{it(\omega-\omega')}.$$

and thus

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

Observe now that for the null space of $\Phi_\Gamma^*\Phi_\Gamma$ to have dimension K, at least K of the eigenvalues in (4.2) must have magnitude greater than or equal to 1. As a result

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

Using the fact that $(a+b) \geq 4ab/(a+b)$ (arithmetic mean dominates geometric mean), we see that if $|T| + |\Omega| < 2N$, then $2|T| \cdot |\Omega|/N < |T| + |\Omega|$ which implies (4.1) (and hence $\dim(\mathrm{Null}(\Phi_\Gamma^*\Phi_\Gamma))$) is less than $(|T| + |\Omega|)/2$. ∎

Proof of Theorem 4.1 We assume Γ is selected such that Φ_Γ has full rank. This happens if $\|F_{\Omega T}^* F_{\Omega T}\| < 1$ and Theorem 3.1 states that this occurs with probability at least $1 - O(\log N \cdot N^{-\beta})$.

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

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

As such, the set of signals in $\mathrm{Range}(\Phi_\Gamma)$ that have expansions on a $\Gamma' \neq \Gamma$ that are *at least* as sparse as their expansions on Γ is a finite union of subspaces of dimension strictly smaller than $|\Gamma|$. This set has measure zero as a subset of $\mathrm{Range}(\Phi_\Gamma)$, 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 $\dim(\mathrm{Range}(\Phi_{\Gamma'})) < |\Gamma'|$, then (4.4) is certainly true. For a set of coefficients α supported on Γ and α' supported on Γ' to have the same image under Φ, $\Phi\alpha = \Phi\alpha'$ (or equivalently $\Phi_\Gamma R_\Gamma \alpha = \Phi_{\Gamma'} R_{\Gamma'} \alpha'$), two things must be true:

1. α and α' must agree on $\Gamma \cap \Gamma'$. This is a direct consequence of $\Phi_{\Gamma'}$ being full rank (its columns are linearly independent).

2. There is a $\delta \in \mathrm{Null}(\Phi)$ such that $\alpha' = \alpha + \delta$. Of course,

$$\delta(\gamma) = 0, \quad \gamma \in (\Gamma \cup \Gamma')^c.$$

   By item 1 above, we will also have

$$\delta(\gamma) = 0, \quad \Gamma \cap \Gamma'.$$

   Thus, $\mathrm{supp}\, \delta \subset (\Gamma \backslash \Gamma') \cup (\Gamma' \backslash \Gamma)$.

In light of these observations, we see that for $\dim(\mathrm{Range}(\Phi_{\Gamma'}) \cap \mathrm{Range}(\Phi_\Gamma)) = |\Gamma|$, we need that for *every* α supported on Γ, there is a $\delta \in \mathrm{Null}(\Phi)$ that is supported on $(\Gamma \backslash \Gamma') \cup (\Gamma' \backslash \Gamma)$ such that

$$\delta(\gamma) = -\alpha(\gamma) \quad \gamma \in \Gamma \backslash \Gamma'.$$

In other words, we need

$$\dim(\mathrm{Null}(Q_{(\Gamma \backslash \Gamma') \cup (\Gamma' \backslash \Gamma)})) \geq \left|\Gamma \backslash \Gamma'\right|.$$

However, Lemma 4.2 tells us

$$\begin{aligned} \dim(\mathrm{Null}(Q_{(\Gamma \backslash \Gamma') \cup (\Gamma' \backslash \Gamma)})) &< \frac{|\Gamma \backslash \Gamma'| + |\Gamma' \backslash \Gamma|}{2} \\ &\leq \left|\Gamma \backslash \Gamma'\right|, \end{aligned}$$

since $|\Gamma'| \leq |\Gamma|$. Hence $\dim(\mathrm{Range}(\Phi_{\Gamma'}) \cap \mathrm{Range}(\Phi_\Gamma)) < |\Gamma|$, and the theorem follows. ∎

4.2 Recovery via ℓ_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 random signal sampled as in Section 2 and with parameters obeying*

$$\mathbf{E}|T| + \mathbf{E}|\Omega| \leq \frac{N}{(\beta + 1)\log N} \cdot (1/8 + o(1)). \tag{4.5}$$

Then with probability at least $1 - O((\log N) \cdot N^{-\beta})$, the solutions of (P_1) and (P_0) are identical and equal to α.

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 α is a unique minimizer of (P_1) if and only if there exists a "dual vector" S satisfying certain properties. Here, 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 ℓ_1-duality

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

$$(\Phi^* S)(\gamma) = (\operatorname{sgn}\alpha)(\gamma) \qquad \gamma \in \Gamma \tag{4.6}$$

$$|(\Phi^* S)(\gamma)| < 1 \qquad \gamma \in \Gamma^c. \tag{4.7}$$

With this notion, we introduce a strong duality result which is similar to that presented in [3], see also [14].

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

Proof The program dual to (P_1) is

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

It is a classical result in convex optimization that if $\tilde{\alpha}$ is a minimizer of (P_1), then $\operatorname{Re}(S^*\Phi\tilde{\alpha}) \le \|\tilde{\alpha}\|_{\ell_1}$ 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 ΦR_Γ^* 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 α is a minimizer of (P_1). Since $|P(\gamma)| < 1$ for $\gamma \in \Gamma^c$, all minimizers of (P_1) must be supported on Γ. But ΦR_Γ^* has full rank, so α is the unique minimizer.

For the converse, suppose that α is the unique minimizer of (P_1). Then there exists at least one S such that with $P = \Phi^* S$, $\|P\|_{\ell_\infty} \le 1$ 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 $|P(\gamma)| \le 1$, 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 $H_S = \{d : \mathrm{Re}\langle d, \Phi^* S\rangle = \|\alpha\|_{\ell_1}\}$ that contains V (since $\mathrm{Re}\langle f, S\rangle = \|\alpha\|_{\ell_1}$) and which defines a halfspace $\{d : \mathrm{Re}\langle d, \Phi^* S\rangle \leq 1\}$ that contains B (and for each such hyperplane, a S exists that describes it as such). Since α is the unique minimizer, for one of these S', the hyperplane $H_{S'}$ intersects B only on the minimal facet $\{d : \mathrm{supp}\, d \subset \Gamma\}$, and we will have $P(\gamma) < 1,\ \gamma \in \Gamma^c$. ∎

Thus to show that (P_1) recovers a representation α 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

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

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

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

where again, R_Γ is the restriction operators to Γ. Setting $P = \Phi^* S$, 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 be very useful to show that P is small componentwise.

4.2.2 Invertibility

Lemma 4.4 *Fix $\beta \geq 1$ and the parameters as in* (4.5)*. Then the matrix $\Phi_\Gamma^* \Phi_\Gamma$ is invertible and obeys*

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

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

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 the conclusion of Lemma 3.1 which (4.5) alows to specialize to the value $1/q = 8\rho_0\sqrt{(\beta+1)\log N}$. Note that this gives more than what is claimed since

$$\|(\Phi^*_\Gamma \Phi_\Gamma)^{-1}\| \leq 1 + \frac{1}{8\rho_0\sqrt{(\beta+1)\log N}} + O(1/\log N).$$

■

Remark. Note that Lemma 3.1 assures us that it is sufficient to take $\mathbf{E}|T| + \mathbf{E}|\Omega|$ of the order of $N/\sqrt{\log N}$ (rather than of the order of $N/\log N$ as the Theorem states) and still have invertibility with $\|(\Phi^*_\Gamma \Phi_\Gamma)^{-1}\| \leq 2$, say. 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, with high probability, $|P(\gamma)| < 1$ on Γ^c.

Lemma 4.5 *Under the hypotheses of Theorem 4.2, for each* $\gamma \in \Gamma^c$

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

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

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

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

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

These notations permit 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 α on Γ are statistically independent of Γ (and hence of W) and, therefore, for a fixed support set Γ, $P_1(t)$ 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 $|X_\gamma| \le |W(\gamma)|$. Applying the complex Hoeffding inequality (see the Appendix) gives a bound on the conditional ditribution of $P(t)$

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

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

$$\|W\| \le \|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \cdot \|V_t\|. \tag{4.9}$$

On the one hand, $\|V_t\| = \sqrt{|\Omega|/N}$ and as we have seen, size estimates about $|\Omega|$ give $\|V_t\| \le \sqrt{2(p_T + p_\Omega)}$ with the desired probability. On the other hand, we have also seen that $\|(\Phi_\Gamma^* \Phi_\Gamma)^{-1}\| \le 1 + o(1)$— also with the desired probability—and, therefore,

$$\|W\|^2 \le 2 \cdot (1 + o(1))) \cdot (p_T + p_\Omega).$$

This gives

$$P\left(|P_1(t)| \ge 1\right) \le 4\exp\left(-\frac{1}{8(p_T + p_\Omega)(1 + o(1))}\right).$$

Select $p_T + p_\Omega$ as in (4.5). Then

$$\mathbf{P}\left(|P_1(t)| \ge 1\right) \le 4\exp(-(\beta+1)\log N) \le 4N^{-(\beta+1)}$$

and

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

As we alluded earlier, the bound about the size of each individual $P(t)$ one would obtain assuming that $\mathbf{E}|T| + \mathbf{E}|\Omega|$ be only of the order $N/\sqrt{\log N}$ would not allow taking the supremum via the standard union bound. Our approach requires $\mathbf{E}|T| + \mathbf{E}|\Omega|$ 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 and 4.5 and of Theorem 4.2. ∎

5 Robust UPs and Basis Pursuit

The results of Sections 3 and 4 extend to the general situation where the dictionary Φ is a union of two orthonormal bases Φ_1, Φ_2. In this section, we present results for pairs of orthogonal bases that parallel those for the time-frequency dictionary presented in Sections 3 and 4. The bounds will depend critically on the degree of similarity of Φ_1 and Φ_2, which we measure using the the mutual incoherence 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 setup to sample the support sets Γ_1, Γ_2 in the Φ_1 and Φ_2 domains respectively. The statement below is the analogue of Theorem 3.1.

Theorem 5.1 *Let $\Phi := \begin{pmatrix}\Phi_1 & \Phi_2\end{pmatrix}$ be a dictionary composed of a union of two orthonormal bases with mutual incoherence μ. Suppose the sampling parameters obey*

$$\mathbf{E}|\Gamma_1| + \mathbf{E}|\Gamma_2| \leq \frac{C_1}{\mu^2 \cdot ((\beta+1)\log N)^{5/2}} \tag{5.1}$$

for some positive constant $C_1 > 0$. Assume $\mu \leq 1/\sqrt{2(\beta+1)\log N}$. Then with probability at least $1 - O(\log N \cdot N^{-\beta})$, every signal f with $\Phi_1 f$ supported on Γ_1 has most of its energy in the Φ_2-domain outside of Γ_2:

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

and vice versa. As a result, for nearly all pairs (Φ_1, Φ_2) with sizes obeying (5.1), it is impossible to find a signal f supported on Γ_1 in the Φ_1-domain and Γ_2 in the Φ_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 incoherences scaling like a power-law $\mu \sim N^{-\gamma}$, our condition essentially reads $|\Gamma_1| + |\Gamma_2| \sim N^{2\gamma}$ compared to $|\Gamma_1| + |\Gamma_2| \sim N^{\gamma}$.

The proof of Theorem 5.1 directly parallels that of Theorem 3.1, with $A := R_{\Gamma_2}\Phi_2^*\Phi_1 R_{\Gamma_1}^*$ playing the role of the partial Fourier transform from T to Ω. Our argument calls for bounds on the eigenvalues of the random matrix A^*A which we write as the sum of two terms; a diagonal and an off-diagonal term

$$A^*A = D + \mathcal{H}_1.$$

We use large deviation theory to control the norm of D while bounds on the size of $\mathcal{H}_1$ are obtained by using moment estimates. This calculation involves estimates about the expected value of the Frobenius norm of large powers of A^*A and is very delicate. We do not reproduce all these arguments here (this is the scope of a whole separate article) and simply state a result which is proved in [4]

$$\mathbf{P}(\|A^*A\| \geq 1/2) \leq C \cdot \log N \cdot N^{-\beta} \tag{5.2}$$

for $\mathbf{E}|\Gamma_1|+\mathbf{E}|\Gamma_2|$ obeying (5.1) (here C is some universal positive constant). Now for (5.2) to hold, we also need that the incoherence be not too large and obeys $\mu > 1/\sqrt{2(\beta+1)\log N}$ which is the additional condition stated in the hypothesis. The idea that μ cannot be too large is somewhat natural as otherwise for $\mu = 1$, say, the two bases would share at least one element and we would have $\|A^*A\| = 1$ as soon as Γ_1 and Γ_2 would contain a common element. As we have seen in section 3, the size estimate (5.2) would then establish the theorem.

The generalized ℓ_0-uniqueness result follows directly from Theorem 5.1:

Theorem 5.2 *Let $f = \Phi\alpha$ be an observed signal sampled as in Section 2, and with parameters obeying*

$$\mathbf{E}|\Gamma_1| + \mathbf{E}|\Gamma_2| \leq \frac{C_2}{\mu^2 \cdot ((\beta+1)\log N)^{5/2}}.$$

Assume $\mu \leq 1/\sqrt{2(\beta+1)\log N}$. Then with probability $1 - O(\log N \cdot N^{-\beta})$, the solution to (P_0) is unique and equal to α.

The only change to the proof presented in Section 4.1 is in the analogue to Lemma 4.2:

Lemma 5.1 *Let Γ_1, Γ_2 be fixed subsets of $\{0, \ldots, N-1\}$, let $\Gamma = \Gamma_1 \cup \Gamma_2$, and let Q_Γ be the $N \times |\Gamma|$ matrix*

$$Q_\Gamma = \begin{pmatrix} \Phi_1 R^*_{\Gamma_1} & \Phi_2 R^*_{\Gamma_2} \end{pmatrix}.$$

If $|\Gamma| < 2/\mu^2$, then

$$\dim(\mathrm{Null}(Q_\Gamma)) < \frac{|\Gamma|}{2}.$$

The proof of Lemma 5.1 has exactly the same structure as the proof of Lemma 4.2. The only modification comes in calculating the trace of G^*G; here each term can be bounded by μ^2, and we have $\mathrm{Tr}(G^*G) \leq 2(|\Gamma_1| \cdot |\Gamma_2|)\mu^2$. Lemma 5.1 follows.

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 Section 2 with*

$$\mathbf{E}|\Gamma_1| + \mathbf{E}|\Gamma_2| \leq \frac{C_3}{\mu^2 \cdot ((\beta+1)\log N)^{5/2}}.$$

Assume $\mu \leq 1/\sqrt{2(\beta+1)\log N}$. Then with probability $1 - O(\log N \cdot N^{-\beta})$, the solutions of (P_0) and (P_1) are identical and equal to α.

The proof of Theorem 5.3 is again almost exactly the same as that 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 $(\log N)^{5/2}$ term in the denominator means that 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 $\exp(-\rho(\log N)^5)$ 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_Γ 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 α on Γ with independent and identically distributed Gaussian entries[1];
4. make $f = \Phi\alpha$;
5. solve (P_1) and obtain $\hat{\alpha}$;
6. compare α to $\hat{\alpha}$;
7. repeat 100 times for each N_Γ;
8. repeat for signal lengths $N = 256, 512, 1024$.

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 α 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, Ω), 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 for nearly all pairs

[1] The results presented here do not seem to depend on the actual distribution used to sample the coefficients.

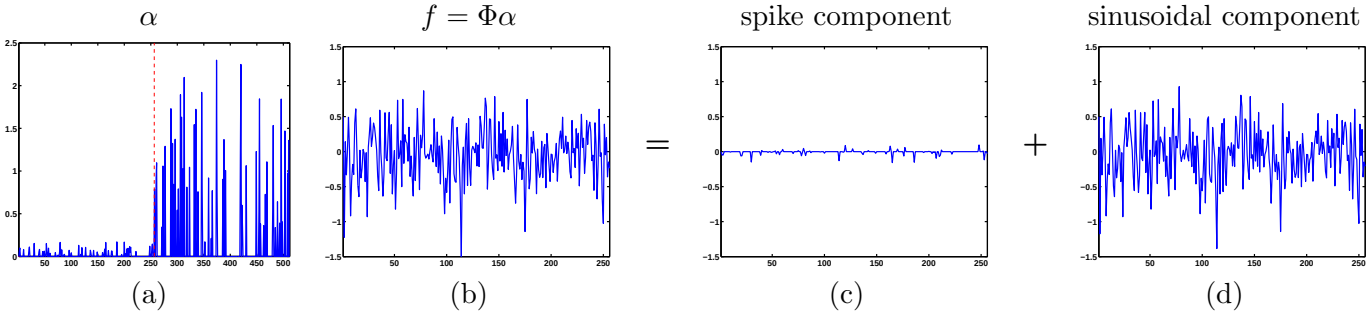

Figure 1: *Recovery of a "sparse" decomposition. (a) Magnitudes of a randomly generated coefficient vector α with 120 nonzero components. The spike components are on the left (indices 1–256) and the sinusoids are on the right (indices 257–512). The spike magnitudes are made small compared to the magnitudes 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).*

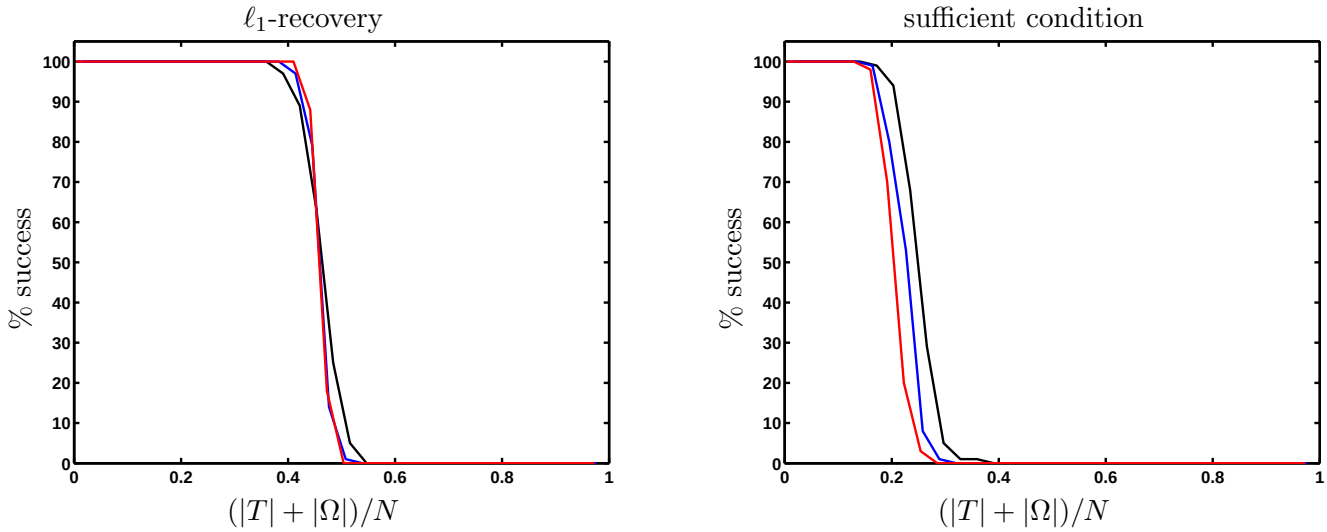

Figure 2: *ℓ_1-recovery for the time-frequency dictionary. (a) Success rate of (P_1) in recovering the sparsest decomposition versus the number of nonzero terms. (b) Success rate of the sufficient condition (the minimum energy signal is a dual vector).*

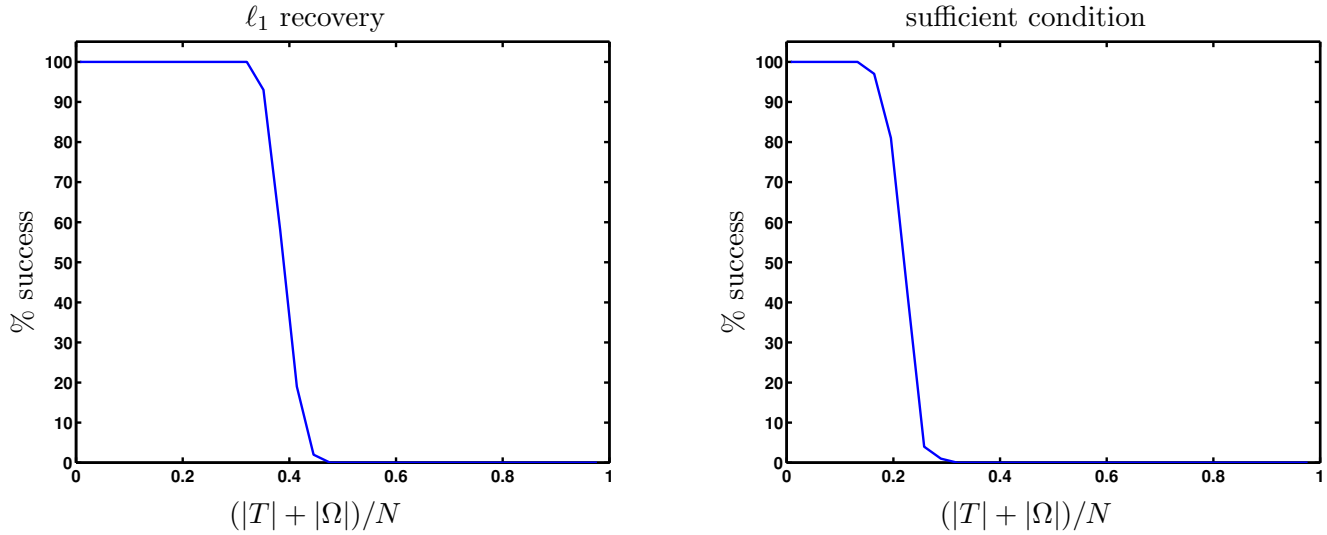

Figure 3: *ℓ_1-recovery for the spike-random dictionary. (a) Success rate of (P_1) in recovering the sparsest decomposition versus the number of nonzero terms. (b) Success rate of the sufficient condition.*

(T, Ω), it is actually impossible to concentrate a discrete signal on T and Ω 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 to each time point and to each frequency point the same chance of being sampled, independently of the others. 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 ferequencies for example, one would need to develop analogues of Lemma 3.1. Our machinery would then nearly automatically transforms 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 ℓ_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 α (in $f = \Phi\alpha$) need to be randomized as well. Or would it be possible to show analogs of Theorem 4.2 (ℓ_1 recovers the sparsest decomposition) for all α, provided that the support of α 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 [18].

Lemma 8.1 *(Hoeffding inequality) Let $X_0, \ldots, X_{N-1}$ 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 $X_0, \ldots, X_{N-1}$ 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; $X_j^{\mathsf{r}} = \operatorname{Re} X_j$, $X_j^{\mathsf{i}} = \operatorname{Im} X_j$. Clearly, $|X_j^{\mathsf{r}}| \leq a_j$ and $|X_j^{\mathsf{i}}| \leq a_j$. The result follows immediately from Lemma 8.1 and the fact that

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

■

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`.

[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`.

[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 ℓ_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] 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.

[17] 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.

[18] G. Lugosi, Concentration-of-measure Inequalities, Lecture Notes.

[19] 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.

[20] D. Slepian and H. O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty I, *Bell Systems Tech. Journal*, **40** (1961), pp. 43–64.

[21] D. Slepian, Prolate spheroidal wave functions, Fourier analysis and uncertainty V — the discete case, *Bell Systems Tech. Journal*, **57** (1978), pp. 1371–1430.

[22] J. L. Starck, E. J. Candès, and D. L. Donoho. Astronomical image representation by the curvelet transform *Astronomy & Astrophysics*, **398** 785–800.

[23] 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.

[24] P. Stevenhagen, H. W. Lenstra Jr., Chebotarëv and his density theorem, *Math. Intelligencer* **18** (1996), no. 2, 26–37.

[25] T. Tao, An uncertainty principle for cyclic groups of prime order, preprint. math.CA/0308286

[26] J. A. Tropp, Greed is good: Algorithmic results for sparse approximation, Technical Report, The University of Texas at Austin, 2003.