

Uncertainty principles and sparse approximation

In this lecture, we will consider the special case where the dictionary Ψ is composed of a pair of orthobases. We will see that our ability to find a sparse approximation to a signal (when one exists) hinges on the existence of an *uncertainty principle* between the two bases.

Essentially, this uncertainty principle ensures that it would take many elements from one of the bases to represent the space spanned by a small number of elements from the other basis — signals that are concentrated in one of the bases must be spread out in the other. We will see how this type of relationship ensures that sparse combinations of the dictionary elements can only be written one way with a small number of terms.

Spikes and Sinusoids

As an illustrative example, we will consider in detail the case where our dictionary is the union of two particular orthobases: the “spike” basis (identity) and the “sine” bases (Fourier):

$$\Psi = [I \quad F] .$$

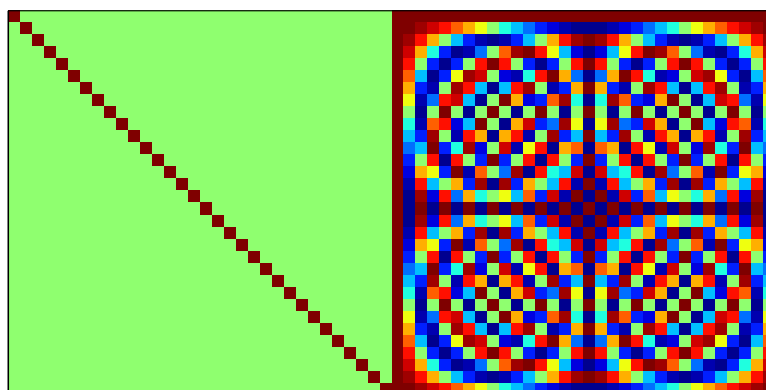
I is the $n \times n$ identity,

F is the $n \times n$ (normalized) discrete Fourier matrix:

$$F[m, \ell] = \frac{1}{\sqrt{n}} e^{j2\pi(m-1)(\ell-1)/n}$$

and so Ψ is $n \times 2n$.

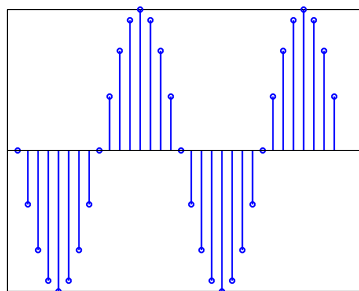
The two halves of the matrix are completely different:



In fact, they are in some sense maximally different in that

1. it takes n spikes to build up a single sinusoid
2. it takes n sinusoids to build up a single spike

Consider the case where f is a discrete harmonic sinusoid:



There are an infinite number of ways we can decompose f using atoms from Ψ . One way (the most natural way) is to use the single column from F at the corresponding frequency:

$$f = [I \quad F] \begin{bmatrix} 0 \\ \vdots \\ - \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

Another way we can get a representation for f is by applying Ψ^* and dividing by 2. Since $\Psi\Psi^* = 2I$,

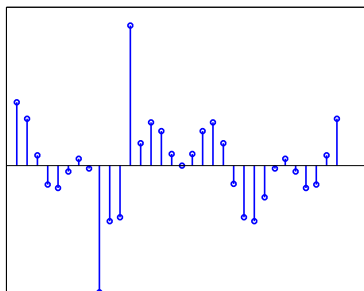
$$f = \frac{1}{2}\Psi(\Psi^*f).$$

This actually corresponds (in this case) to the **minimum energy** coefficient vector that represents f . Here is what we get for the single sinusoid above:

$$\frac{1}{2}\Psi^*f = \begin{bmatrix} \text{sinusoid} \\ \vdots \\ \text{sinusoid} \\ \vdots \end{bmatrix}$$

which is maybe not what we want

The problem is even more pronounced if we take f to be a sum of two sines and two spikes:



Of course, there exists a decomposition that only uses 4 terms. But when we look at the minimum energy decomposition, it is a mess:

$$\frac{1}{2}\Psi^* f =$$

The problem is that the minimum energy decomposition is mixing together the sinusoidal components and the spike components, resulting in a decomposition that is nowhere sparse.

We will see below that fundamentally, our ability to separate the “sines part” of f from the “spikes part” is what will determine whether or not we can find a unique sparse decomposition.

Being able to tell these two components apart, in turn, comes from a new kind of *uncertainty principle*, which states that a signal cannot be too sparse in time and frequency simultaneously.

A Discrete Uncertainty Principle

The following result is due to Donoho and Stark in a 1989 paper:

Let $f \in \mathbb{C}^n$ be a discrete signal, and let
 $\hat{f} \in \mathbb{C}^n$ be its discrete Fourier transform ($\hat{f} = Ff$).

Let T be the support of f
(locations of nonzero coefficients in time)
 $|T|$ = size of T = number of nonzero coefficients in time

Let Ω be the support of \hat{f}
(locations of nonzero coefficients in frequency)
 $|\Omega|$ = size of Ω = number of nonzero coefficients in freq

Then the following relationships hold between the sizes of these supports:

$$\begin{aligned} |T| \cdot |\Omega| &\geq n \\ \Rightarrow |T| + |\Omega| &\geq 2\sqrt{n}. \end{aligned}$$

The second statement follows from the first using the fact that arithmetic mean dominates geometric mean:

$$\sqrt{n} \leq \sqrt{|T| \cdot |\Omega|} \leq \frac{|T| + |\Omega|}{2}$$

Proof of Donoho-Stark uncertainty principle

Let I_T be the $n \times |T|$ matrix constructed by extracting the columns of I indexed by T . Let F_Ω be the $n \times |\Omega|$ matrix corresponding to the columns of F indexed by Ω . Finally, let

$$\Psi_{T\Omega} = [I_T \quad F_\Omega]$$

If f is supported on T in the time domain and $\hat{f} = F^* f$ is supported on Ω in the frequency domain, then

$$\Psi \begin{bmatrix} f \\ -\hat{f} \end{bmatrix} = \Psi_{T\Omega} \begin{bmatrix} f_T \\ -\hat{f}_\Omega \end{bmatrix} = 0,$$

where $f_T \in \mathbb{C}^{|T|}$ is the restriction of F to T (just throw away all the zeros), and similarly for $\hat{f}_\Omega \in \mathbb{C}^{|\Omega|}$.

Main idea: If $\Psi_{T\Omega}$ does not have a null space, then it is **impossible** to find an f supported on T in time and Ω in frequency

A way to check this is to see if the matrix $\Psi_{T\Omega}^* \Psi_{T\Omega}$ is invertible.

A way to check that is to see if all the eigenvalues of the matrix $\Psi_{T\Omega}^* \Psi_{T\Omega}$ are greater than zero. Since

$$\begin{aligned} \Psi_{T\Omega}^* \Psi_{T\Omega} &= \begin{bmatrix} I_T^* \\ F_\Omega^* \end{bmatrix} [I_T \quad F_\Omega] = \begin{bmatrix} I_T^* I_T & I_T^* F_\Omega \\ F_\Omega^* I_T & F_\Omega^* F_\Omega \end{bmatrix} \\ &= I + \begin{bmatrix} 0 & M \\ M^* & 0 \end{bmatrix} \\ &= I + G \end{aligned}$$

So all the eigenvalues of $\Psi_{T\Omega}^* \Psi_{T\Omega} = I + G$ are greater than zero if all the eigenvalues of G are less than 1: $\|G\| < 1$.

(Here, $\|G\|$ is the *operator norm* of G , which is simply the magnitude of the largest eigenvalue since G is symmetric.)

We now turn our attention to bounding the operator norm of G . We start with the simple observation that since

$$G^*G = \begin{bmatrix} MM^* & 0 \\ 0 & M^*M \end{bmatrix}$$

we have

$$\|G\|^2 = \|G^*G\| = \|M^*M\|.$$

So now we are left with finding conditions on T and Ω so that $\|M^*M\| < 1$.

Since the trace of a symmetric semidefinite matrix is the sum of its eigenvalues, it is true that

$$\|M^*M\| \leq \text{trace}(M^*M) = \sum_{\omega \in \Omega} (M^*M)_{\omega,\omega}.$$

For each $\omega \in \Omega$

$$\begin{aligned} (M^*M)_{\omega,\omega} &= \frac{1}{n} \sum_{t \in T} e^{-j2\pi\omega t/n} e^{j2\pi\omega t/n} \\ &= \frac{|T|}{n} \\ \Rightarrow \text{trace}(M^*M) &= \frac{|\Omega| \cdot |T|}{n} \end{aligned}$$

So to conclude, if $|\Omega| \cdot |T| < n$, then

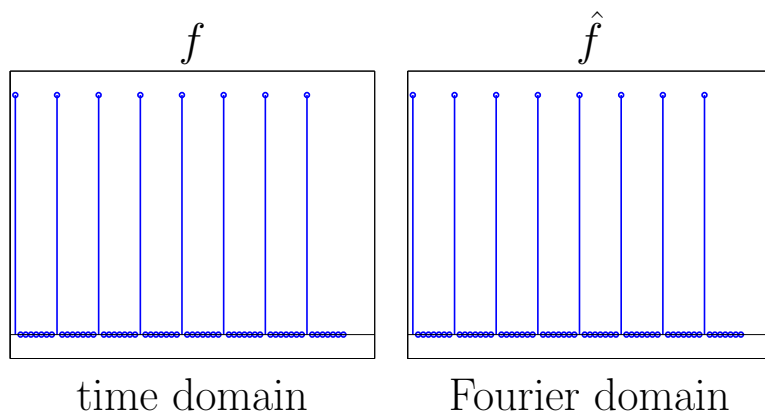
$\text{trace}(M^*M) < 1 \Rightarrow \|G\| < 1 \Rightarrow \Psi_{T\Omega}$ does not have a null space

which means it is impossible to find a f supported on T so that \hat{f} is supported on Ω .

This uncertainty principle is **sharp** in that there exist signals for which it is met with equality.

Example: the Dirac comb

Say n is a square integer. Construct f by placing \sqrt{n} spikes of equal height spaced every \sqrt{n} . It turns out that \hat{f} will have exactly the same form:



Of course in this case, $|T| = |\Omega| = \sqrt{n}$ and both statements in the uncertainty principle hold with equality.

The UP and unique sparsest decompositions

An immediate consequence of the discrete uncertainty principle is that it allows us say when sparse decompositions in the $\Psi = [I \ F]$ dictionary are *unique*. That is, there is no decomposition which is sparser.

Suppose f can be written as a sum of spikes on a set T and sum of sinusoids with frequencies in Ω :

$$f(t) = \sum_{\tau \in T} \alpha_{\tau}^1 \delta(t - \tau) + \sum_{\omega \in \Omega} \alpha_{\omega}^2 \frac{1}{\sqrt{n}} e^{j2\pi\omega t/n}$$

or in matrix-vector form

$$f = \Psi\alpha = \begin{bmatrix} I & F \end{bmatrix} \begin{bmatrix} \alpha^1 \\ - \\ \alpha^2 \end{bmatrix}$$

where α^1 is nonzero on T and α^2 is nonzero on Ω .

Suppose that α has fewer than \sqrt{n} nonzero terms in it:

$$|T| + |\Omega| < \sqrt{n}$$

(so f is composed of at most a total number of \sqrt{n} spikes/sines)

Is there another way to write f using the same or fewer number of terms?

Answer: **No.**

Why not? Well, suppose there were another such way; that is, there was a $\beta \neq \alpha$ such that

$$\Psi\beta = \Psi\alpha = f$$

and

$$|\text{support } \beta^1| + |\text{support } \beta^2| < \sqrt{n}.$$

Set $h = \alpha - \beta$ and note that $h \in \text{Null}(\Psi)$, since

$$\Psi h = \Psi\alpha - \Psi\beta = f - f = 0.$$

Also note that

$$|\text{support } h| \leq |T| + |\Omega| + |\text{support } \beta^1| + |\text{support } \beta^2| \leq 2\sqrt{n}.$$

So h would be in the null space of Ψ and have fewer than $2\sqrt{n}$ nonzero terms.

But the discrete UP tells us that such a thing is impossible. We can see what vectors in the null space of Ψ look like:

$$\begin{aligned}\Psi h = 0 &\Rightarrow [I \quad F] \begin{bmatrix} h^1 \\ h^2 \end{bmatrix} = 0 \\ &\Rightarrow h^1 + Fh^2 = 0 \\ &\Rightarrow h^1 = -\hat{h}^2\end{aligned}$$

so vectors in the null space of Ψ have a structure so that the first half of the vector is the Fourier transform of the second half of the vector. The discrete UP tells us that such vectors **must** have more than $2\sqrt{n}$ nonzero terms.

$\Rightarrow \alpha$ is the *only* decomposition of f that uses \sqrt{n} or fewer terms.

Equivalently, if we observe f and there *exists* a decomposition using \sqrt{n} or fewer terms, then solving the optimization program

$$\min_{\beta} \|\beta\|_{\ell_0} \quad \text{subject to} \quad \Psi\beta = f$$

will find it.

ℓ_1 and sparsest decomposition

We will see in a later lecture that this type of uncertainty principles plays a central role in showing that the sparsest decomposition can be recovered using Basis Pursuit:

$$\min_{\beta} \|\beta\|_{\ell_1} \quad \text{subject to} \quad \Psi\beta = f$$

which of course has the advantage of being computationally tractable. Work of Donoho-Huo, Elad-Bruckstein, Nielson, and Gribonval in 2001–2003 has shown that BP recovers the sparsest decomposition when

$$|\text{support } \alpha| = |T| + |\Omega| \lesssim 0.9 \sqrt{n}$$

Pairs of Bases

All of the results above can be generalized to arbitrary pairs of orthobases. In some sense, the uncertainty principle above comes about because the spike basis and Fourier basis are very different than one another, so we will need a way to quantify this difference in the general case.

Given a dictionary composed of two orthobases

$$\Psi = [\Psi_1 \quad \Psi_2],$$

we define the **coherence** as

$$\mu = \sqrt{n} \cdot \max_{\substack{\psi_1 \in \Psi_1 \\ \psi_2 \in \Psi_2}} |\langle \psi_1, \psi_2 \rangle|$$

It is a fact that

$$1 \leq \mu \leq \sqrt{n},$$

where the lower bound is achieved when an atom in Ψ_1 is maximally diffuse in the Ψ_2 domain (e.g. spikes and sines), and the upper bound is achieved when Ψ_1 and Ψ_2 share an element.

The following uncertainty principle was established by Elad and Bruckstein in 2002:

General Uncertainty Principle: Let $f \in \mathbb{C}^n$ be a given signal, and set

$$\begin{aligned}\beta^1 &= \Psi_1^* f \quad (\Psi_1 \text{ transform}) \\ \beta^2 &= \Psi_2^* f \quad (\Psi_2 \text{ transform}),\end{aligned}$$

If Γ_1 and Γ_2 are the supports of β^1 and β^2 , then

$$\begin{aligned}|\Gamma_1| \cdot |\Gamma_2| &\geq \frac{n}{\mu^2} \\ \Rightarrow |\Gamma_1| + |\Gamma_2| &\geq \frac{2\sqrt{n}}{\mu}.\end{aligned}$$

This is proved in almost the exact same way as the spikes and sines case (since that argument only really depended on the sizes of inner products between the bases).

The uniqueness of the sparsest decomposition follows: Suppose $f = \Psi\alpha$, where

$$|\text{support } \alpha| < \sqrt{n}/\mu.$$

Then α is the **only** decomposition of f in Ψ that uses fewer than \sqrt{n}/μ terms, and is the solution to

$$\min \|\beta\|_{\ell_0} \quad \text{subject to} \quad \Psi\beta = f.$$

There is also a corresponding result (again due to E& B '02) for recovery using the ℓ_1 norm (again, we will see why this is true in a later lecture):

Suppose $f = \Psi\alpha$, where

$$|\text{support } \alpha| < \frac{1}{2}(1 + \sqrt{n}/\mu).$$

Then α is the unique solution to the convex program

$$\min_{\beta} \|\beta\|_{\ell_1} \quad \text{subject to} \quad \Psi\beta = f.$$

Generic supports. We have seen that the UP for spikes and sines is tight, since there exists a signal (the Dirac comb) that meets the bound with equality.

But it turns out we can derive **much stronger** uncertainty principles which hold for most supports in time and frequency.

Robust Uncertainty Principle (Candes and R, '06):

Select a set T in the time domain and a set Ω in the frequency domain uniformly at random with

$$|T| + |\Omega| \leq \text{Const} \cdot \frac{n}{\sqrt{\log n}} \quad (\text{can take Const} \sim 1/8)$$

The with *high probability* it is **impossible** to find a signal $f \in \mathbb{C}^n$ supported on T such that \hat{f} is supported on Ω .

(In fact, it is impossible that even *half* of the energy of \hat{f} can be supported on Ω .)

So now we know the uncertainty principle holds for:

all supports with $|T| + |\Omega| < 2\sqrt{n}$, and

most supports with $|T| + |\Omega| \lesssim n/\sqrt{\log n}$

Proving the robust UP is more involved than the deterministic case, and we will not prove it here.

There is also an associated ℓ_1 recovery result which requires a little more sparsity:

Choose T and Ω uniformly at random so that

$$|T| + |\Omega| \leq \text{Const} \cdot \frac{n}{\log n}$$

Then for the vast majority of α supported on $T \cup \Omega$, α is the unique minimizer of

$$\min_{\beta} \|\beta\|_{\ell_1} \quad \text{subject to} \quad \Psi\beta = f.$$

Both the robust uncertainty principle and the ℓ_1 recovery results can be extended to general bases (C&R, Tropp '06)