CS-GY 6763: Lecture 11
Sparse Recovery and Compressed Sensing

R. Teal Witter

April 18, 2022

NYU Tandon School of Engineering

Review of Last Class

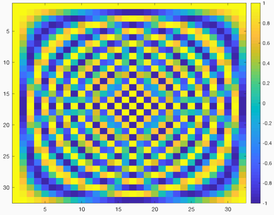
Main idea: If you want to compute singular vectors or eigenvectors, multiply two matrices, solve a regression problem, then

1. Compress matrices using randomized method
2. Solve the problem on the smaller or sparser matrix

Beyond the Hadamard Transform

The Hadamard Transform is closely related to the Discrete Fourier Transform.

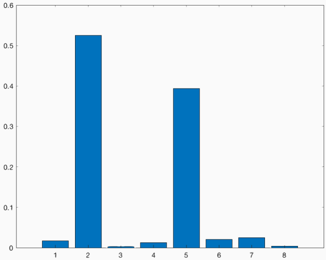
$$\mathbf{F}_{j,k} = e^{-2\pi i \frac{j \cdot k}{n}}, \qquad \mathbf{F}^*\mathbf{F} = \mathbf{I}.$$



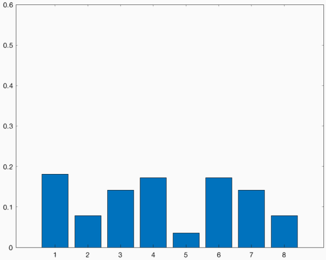
Real part of $\mathbf{F}_{j,k}$.

$\mathbf{F}\mathbf{y}$ computes the Discrete Fourier Transform of the vector $\mathbf{y}$. Can be computed in $O(n \log n)$ time using a divide and conquer algorithm (the Fast Fourier Transform).

The Uncertainty Principal (informal): A function and it's Fourier transform cannot both be concentrated.



Vector **y**.



Fourier transform **Fy**.

The Uncertainty Principal

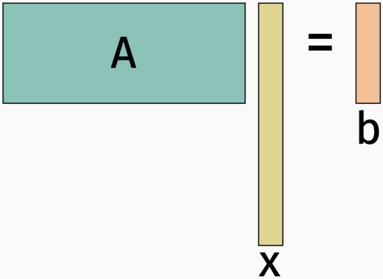
Sampling does not preserve norms, i.e. $\|\mathbf{S}\mathbf{y}\|_2 \not\approx \|\mathbf{y}\|_2$ when $\mathbf{y}$ has a few large entries.

Taking a Fourier transform exactly eliminates this hard case, without changing $\mathbf{y}$'s norm.

One of the central tools in the field of **sparse recovery** aka **compressed sensing.**

Sparse Recovery/Compressed Sensing Problem Setup

Underdetermined linear regression: Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m < n$, $\mathbf{b} \in \mathbb{R}^m$. Assume $\mathbf{b} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n$.

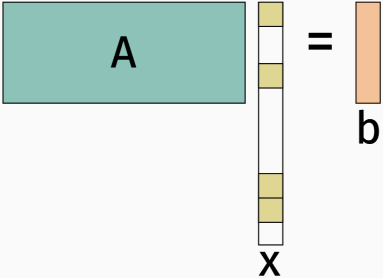


- Infinite possible solutions $\mathbf{y}$ to $\mathbf{A}\mathbf{y} = \mathbf{b}$, so in general, it is impossible to recover parameter vector $\mathbf{x}$ from the data $\mathbf{A}, \mathbf{b}$.

Sparsity Recovery/Compressed Sensing

Underdetermined linear regression: Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m < n$, $\mathbf{b} \in \mathbb{R}^m$. Solve $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{x}$.

- Assume $\mathbf{x}$ is k-sparse for small k. $\|\mathbf{x}\|_0 = k$.



- In many cases can recover $\mathbf{x}$ with $\ll n$ rows. In fact, often $\sim O(k)$ suffice.
- Need additional assumptions about $\mathbf{A}$!

Motivation

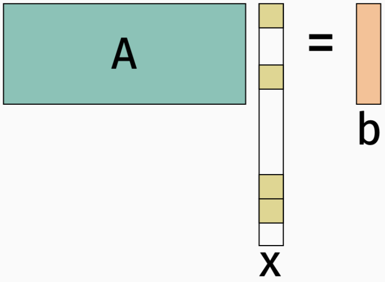
- In statistics and machine learning, we often think about **A**'s rows as data drawn from some universe/distribution:

	bedrooms	bathrooms	sq.ft.	floors	list price	sale price
home 1	2	2	1800	2	200,000	195,000
home 2	4	2.5	2700	1	300,000	310,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮
home n	5	3.5	3600	3	450,000	450,000

- In other settings, we will get to choose **A**'s rows. I.e. each $b_i = \mathbf{x}^T \mathbf{a}_i$ for some vector $\mathbf{a}_i$ that we select.
- In the later case, we often call b_i a linear measurement of $\mathbf{x}$ and we call **A** a measurement matrix.

Assumptions on Measurement Matrix

When should this problem be difficult?



Many ways to formalize our intuition

- $\mathbf{A}$ has Kruskal rank r. All sets of r columns in $\mathbf{A}$ are linearly independent.
    - Recover vectors $\mathbf{x}$ with sparsity $k = r/2$.
- $\mathbf{A}$ is μ-incoherent. $|\mathbf{A}_i^T \mathbf{A}_j| \leq \mu \|\mathbf{A}_i\|_2 \|\mathbf{A}_j\|_2$ for all columns $\mathbf{A}_i, \mathbf{A}_j$, $i \neq j$.
    - Recover vectors $\mathbf{x}$ with sparsity $k = 1/\mu$.
- **Focus today**: $\mathbf{A}$ obeys the Restricted Isometry Property.

Restricted Isometry Property

Definition ((q, ϵ)-Restricted Isometry Property)

A matrix $\mathbf{A}$ satisfies (q, ϵ)-RIP if, for all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq q$,

$$(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2.$$

- Johnson-Lindenstrauss type condition.
- $\mathbf{A}$ preserves the norm of all q sparse vectors, instead of the norms of a fixed discrete set of vectors, or all vectors in a subspace (as in subspace embeddings).

First Sparse Recovery Result

Theorem (ℓ_0-minimization)

Suppose we are given $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{b} = \mathbf{A}\mathbf{x}$ *for an unknown* k*-sparse* $\mathbf{x} \in \mathbb{R}^n$*. If* $\mathbf{A}$ *is* $(2k, \epsilon)$*-RIP for any* $\epsilon < 1$ *then* $\mathbf{x}$ *is the* *unique* *minimizer of:*

$$\min \|\mathbf{z}\|_0 \qquad \textit{subject to} \qquad \mathbf{A}\mathbf{z} = \mathbf{b}.$$

- Establishes that information theoretically we can recover $\mathbf{x}$. Solving the ℓ_0-minimization problem is computationally difficult, requiring $O(n^k)$ time. We will address faster recovery shortly.

First Sparse Recovery Result

Claim: If $\mathbf{A}$ is $(2k, \epsilon)$-RIP for any $\epsilon < 1$ then $\mathbf{x}$ is the unique minimizer of $\min_{\mathbf{Az}=\mathbf{b}} \|\mathbf{z}\|_0$.

Proof: By contradiction, assume there is some $\mathbf{y} \neq \mathbf{x}$ such that $\mathbf{Ay} = \mathbf{b}$, $\|\mathbf{y}\|_0 \leq \|\mathbf{x}\|_0$.

Robustness

Important note: Robust versions of this theorem and the others we will discuss exist. These are much more important practically. Here's a flavor of a robust result:

- Suppose $\mathbf{b} = \mathbf{A}(\mathbf{x} + \mathbf{e})$ where $\mathbf{x}$ is k-sparse and $\mathbf{e}$ is dense but has bounded norm.
- Recover some k-sparse $\tilde{\mathbf{x}}$ such that:

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq \|\mathbf{e}\|_1$$

  or even

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq O\left(\frac{1}{\sqrt{k}}\right) \|\mathbf{e}\|_1.$$

Robustness

We will not discuss robustness in detail, but along with computational considerations, it is a big part of what has made compressed sensing such an active research area in the last 20 years. Non-robust compressed sensing results have been known for a long time:

Gaspard Riche de Prony, *Essay experimental et analytique: sur les lois de la dilatabilite de fluides elastique et sur celles de la force expansive de la vapeur de l'alcool, a differentes temperatures.* Journal de l'Ecole Polytechnique, 24–76. **1795**.

Restricted Isometry Property

What matrices satisfy this property?

- Random Johnson-Lindenstrauss matrices (Gaussian, sign, etc.) with $m = O(\frac{k \log(n/k)}{\epsilon^2})$ rows are (k, ϵ)-RIP.

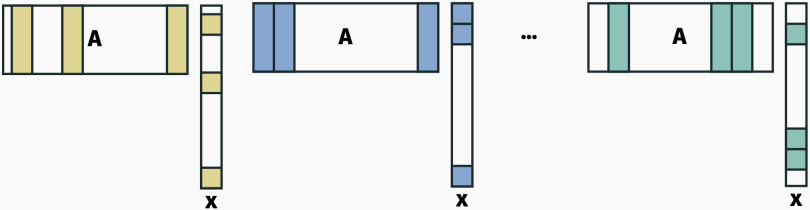
Some real world data may look random, but this is also a useful observation algorithmically when we want to design $\mathbf{A}$.

Definition ((q, ϵ)-Restricted Isometry Property – Candes, Tao '05)

A matrix $\mathbf{A}$ satisfies (q, ϵ)-RIP if, for all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq q$,

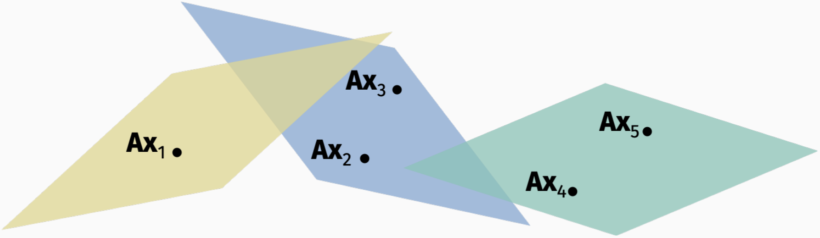
$$(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2.$$

The vectors that can be written as $\mathbf{Ax}$ for q sparse $\mathbf{x}$ lie in a union of q dimensional linear subspaces:



Restricted Isometry Property

Candes, Tao 2005: A random JL matrix with $O(q \log(n/q)/\epsilon^2)$ rows satisfies (q, ϵ)-RIP with high probability.



Any ideas for how you might prove this? I.e. prove that a random matrix preserves the norm of every $\mathbf{x}$ in this union of subspaces?

Restricted Isometry Property from JL

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a q-dimensional linear subspace in $\mathbb{R}^n$. If $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

$$(1-\epsilon)\|\mathbf{v}\|_2^2 \leq \|\Pi \mathbf{v}\|_2^2 \leq (1+\epsilon)\|\mathbf{v}\|_2^2$$

for all $\mathbf{v} \in \mathcal{U}$, as long as $m = O\left(\frac{q+\log(1/\delta)}{\epsilon^2}\right)$.

Quick argument:

Application: Return to Heavy Hitters in Data Streams

Suppose you view a stream of numbers in $1, \ldots, n$:

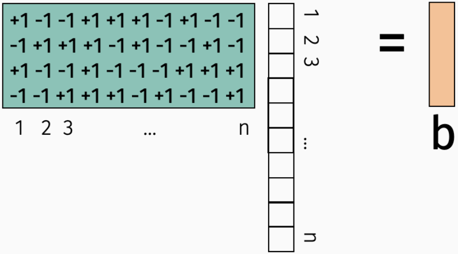
$$4, 18, 4, 1, 2, 24, 6, 4, 3, 18, 18, \ldots$$

After some time, you want to report which k items appeared most frequently in the stream.

E.g. Amazon is monitoring web-logs to see which product pages people view. They want to figure out which products are viewed most frequently. $n \approx 500$ million.

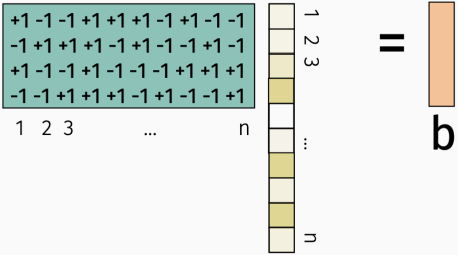
How can you do this quickly in small space?

Application: Heavy Hitters in Data Streams



- Every time we receive a number i in the stream, add column $\mathbf{A}_i$ to $\mathbf{b}$.

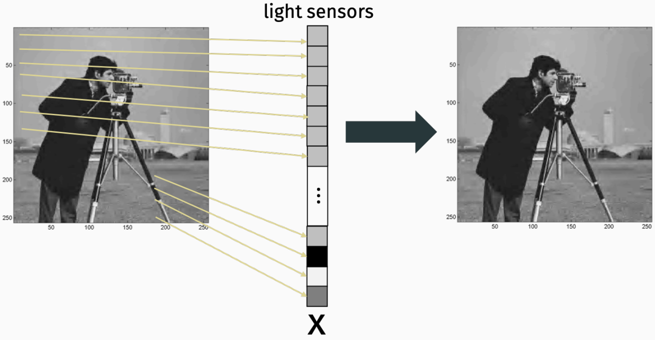
Application: Heavy Hitters in Data Streams



- At the end $\mathbf{b} = \mathbf{A}\mathbf{x}$ for an approximately sparse $\mathbf{x}$ if there were only a few "heavy hitters". Recover $\mathbf{x}$ from $\mathbf{b}$ using a sparse recovery method (like ℓ_0 minimization).

Application: Single Pixel Camera

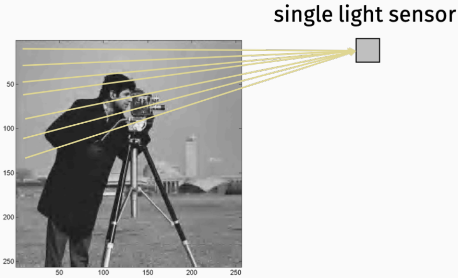
Typical acquisition of image by camera:



Requires one image sensor per pixel captured.

Application: Single Pixel Camera

Compressed acquisition of image:

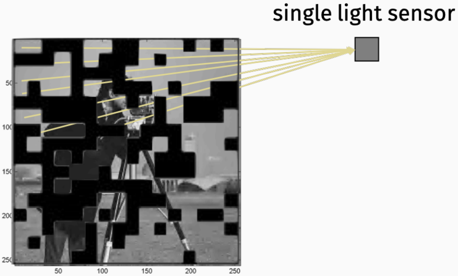


$$p = \sum_{i=1} x_i = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Does not provide very much information about the image.

Application: Single Pixel Camera

But several random linear measurements do!



$$p = \sum_{i=1} R_i x_i = \begin{bmatrix} 0 & 1 & 0 & 0 \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Application: Single Pixel Camera

Applications in:

- Imaging outside of the visible spectrum (more expensive sensors).
- Microscopy.
- Other scientific imaging.

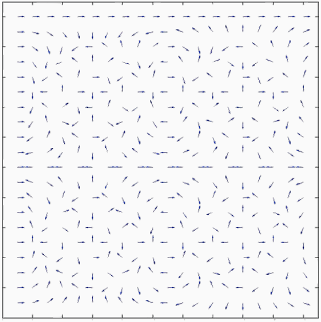
Compressed sensing theory does not exactly describe these problems, but has been very valuable in modeling them.

Discrete Fourier Matrix

The $n \times n$ discrete Fourier matrix $\mathbf{F}$ is defined:

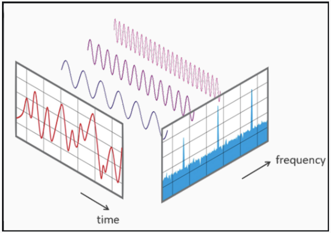
$$F_{j,k} = e^{\frac{-2\pi i}{n} j \cdot k},$$

where $i = \sqrt{-1}$. Recall $e^{\frac{-2\pi i}{n} j \cdot k} = cos(2\pi jk/n) - i \sin(2\pi jk/n)$.



Discrete Fourier Matrix

$\mathbf{Fx}$ is the Discrete Fourier Transform of the vector $\mathbf{x}$ (what an FFT computes).



Decomposes $\mathbf{x}$ into different frequencies: $[\mathbf{Fx}]_j$ is the component with frequency j/n.

Because $\mathbf{F}^*\mathbf{F} = \mathbf{I}$, $\mathbf{F}^*\mathbf{Fx} = \mathbf{x}$, so we can recover $\mathbf{x}$ if we have access to its DFT.

Restricted Isometry Property

Setting $\mathbf{A}$ to contain a random $m \sim O\left(\frac{k \log^2 k \log n}{\epsilon^2}\right)$ rows of the discrete Fourier matrix $\mathbf{F}$ yields a matrix that with high probability satisfies (k, ϵ)-RIP. [Haviv, Regev, 2016].

Improves on a long line of work: Candès, Tao, Rudelson, Vershynin, Cheraghchi, Guruswami, Velingker, Bourgain.

Proving this requires similar tools to analyzing subsampled Hadamard transforms!

Discrete Fourier Matrix

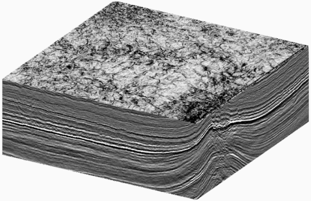
If $\mathbf{A}$ is a subset of q rows from $\mathbf{F}$, then $\mathbf{Ax}$ is a subset of random frequency components from $\mathbf{x}$'s discrete Fourier transform.

In many scientific applications, we can collect entries of $\mathbf{Fx}$ one at a time for some unobserved data vector $\mathbf{x}$.

Application: Geophysics

Warning: very cartoonish explanation of very complex problem.

Understanding what material is beneath the crust:



Think of vector $\mathbf{x}$ as scalar values of the density/reflectivity in a single vertical core of the earth.

How do we measure entries of Fourier transform $\mathbf{Fx}$?

Application: Geophysics

Vibrate the earth at different frequencies! And measure the response.

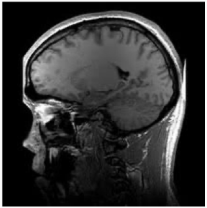
Vibroseis Truck

Can also use airguns, controlled explorations, vibrations from drilling, etc. The fewer measurements we need from $\mathbf{F}\mathbf{x}$, the cheaper and faster our data acquisition process becomes.

Application: Geophysics

Warning: very cartoonish explanation of very complex problem.

Medical Imaging (MRI)



Vector $\mathbf{x}$ here is a 2D image. Everything works with 2D Fourier transforms.

How do we measure entries of Fourier transform $\mathbf{Fx}$?

Application: Geophysics

Blast the body with sound waves of varying frequency.

The fewer measurements we need from $\mathbf{F}\mathbf{x}$, the faster we can acquire an image.

- Especially important when trying to capture something moving (e.g. lungs, baby, child who can't sit still).
- Can also cut down on power requirements (which for MRI machines are huge).

Break

Restricted Isometry Property

Definition ((q, ϵ)-Restricted Isometry Property)

A matrix $\mathbf{A}$ satisfies (q, ϵ)-RIP if, for all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq q$,

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}\|_2^2.$$

Lots of other random matrices satisfy RIP as well.

One major theoretical question is if we can deterministically construct good RIP matrices. Interestingly, if we want $(O(k), O(1))$ RIP, we can only do so with $O(k^2)$ rows (now very slightly better – thanks to Bourgain et al.).

Whether or not a linear dependence on k is possible with a deterministic construction is unknown.

Faster Sparse Recovery

Theorem (ℓ_0-minimization)

Suppose we are given $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{b} = \mathbf{A}\mathbf{x}$ *for an unknown* k*-sparse* $\mathbf{x}$*. If* $\mathbf{A}$ *is* $(2k, \epsilon)$*-RIP for any* $\epsilon < 1$ *then* $\mathbf{x}$ *is the unique minimizer of:*

$$\min \|\mathbf{z}\|_0 \qquad \textit{subject to} \qquad \mathbf{A}\mathbf{z} = \mathbf{b}.$$

Algorithm question: Can we recover $\mathbf{x}$ using a faster method? Ideally in polynomial time.

Basis Pursuit

Convex relaxation of the ℓ_0 minimization problem:

Problem (Basis Pursuit, i.e. ℓ_1 minimization.)

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \qquad \textit{subject to} \qquad \mathbf{A}\mathbf{z} = \mathbf{b}.$$

- Objective is convex.
- Optimizing over convex set.

What is one method we know for solving this problem?

Basis Pursuit Linear Program

Equivalent formulation:

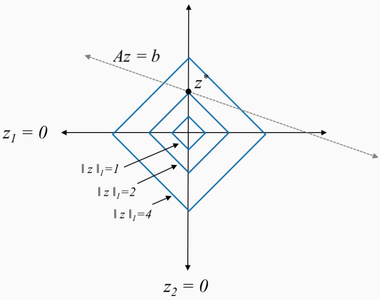
Problem (Basis Pursuit Linear Program.)

$$\min_{\mathbf{w},\mathbf{z}} \mathbf{1}^T\mathbf{w} \qquad \textit{subject to} \qquad \mathbf{A}\mathbf{z} = \mathbf{b}, \mathbf{w} \geq 0, -\mathbf{w} \leq \mathbf{z} \leq \mathbf{w}.$$

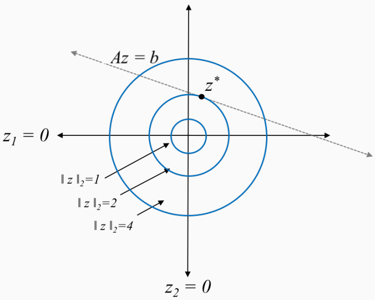
Can be solved using any algorithm for linear programming. An Interior Point Method will run in $\sim O(n^{3.5})$ time.

Basis Pursuit Intuition

Suppose **A** is 2×1, so **b** is just a scalar and **x** is a 2-dimensional vector.



Vertices of level sets of ℓ_1 norm correspond to sparse solutions.



This is not the case e.g. for the ℓ_2 norm.

Basis Pursuit Analysis

Theorem

If $\mathbf{A}$ *is* $(3k, \epsilon)$*-RIP for* $\epsilon < .17$ *and* $\|\mathbf{x}\|_0 = k$*, then* $\mathbf{x}$ *is the unique optimal solution of the Basis Pursuit LP).*

Similar proof to ℓ_0 minimization:

- By way of contradiction, assume $\mathbf{x}$ is <u>not the optimal solution</u>. Then there exists some non-zero Δ such that:
    - $\|\mathbf{x} + \Delta\|_1 \leq \|\mathbf{x}\|_1$
    - $\mathbf{A}(\mathbf{x} + \Delta) = \mathbf{A}\mathbf{x}$. I.e. $\mathbf{A}\Delta = 0$.

Difference is that we can no longer assume that Δ is sparse.

We will argue that Δ is approximately sparse.

Tools Needed

First tool:

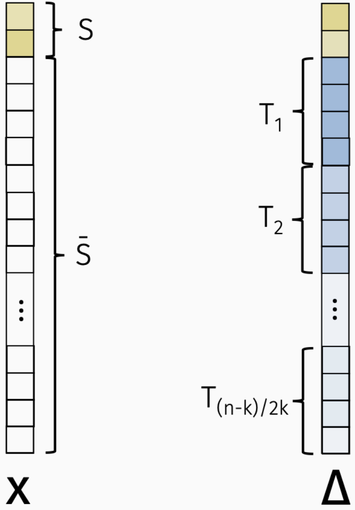
For any q-sparse vector $\mathbf{w}$, $$\|\mathbf{w}\|_2 \leq \|\mathbf{w}\|_1 \leq \sqrt{q}\|\mathbf{w}\|_2$$

Second tool:

For any norm and vectors $\mathbf{a}, \mathbf{b}$, $$\|\mathbf{a}+\mathbf{b}\| \geq \|\mathbf{a}\| - \|\mathbf{b}\|$$

Basis Pursuit Analysis

Some definitions:



Basis Pursuit Analysis

Claim 1: $\|\Delta_S\|_1 \geq \|\Delta_{\bar{S}}\|_1$

Basis Pursuit Analysis

Claim 2: $\|\Delta_S\|_2 \geq \sqrt{2}\sum_{j\geq 2}\|\mathbf{\Delta}_{T_j}\|_2$:

$$\|\mathbf{\Delta}_s\|_2 \geq \frac{1}{\sqrt{k}}\|\mathbf{\Delta}_S\|_1 \geq \frac{1}{\sqrt{k}}\|\mathbf{\Delta}_{\bar{S}}\|_1 = \frac{1}{\sqrt{k}}\sum_{j\geq 1}\|\mathbf{\Delta}_{T_j}\|_1.$$

Claim: $\|\mathbf{\Delta}_{T_j}\|_1 \geq \sqrt{2k}\|\mathbf{\Delta}_{T_{j+1}}\|_2$

Basis Pursuit Analysis

Finish up proof by contradiction: Recall that $\mathbf{A}$ is assumed to have the $(3k, \epsilon)$ RIP property.

$$0 = \|\mathbf{A}\boldsymbol{\Delta}\|_2 \geq \|\mathbf{A}\boldsymbol{\Delta}_{S \cup T_1}\|_2 - \sum_{j \geq 2} \|\mathbf{A}\boldsymbol{\Delta}_{T_j}\|_2$$

Faster Methods

A lot of interest in developing even faster algorithms that avoid using the "heavy hammer" of linear programming and run in even faster than $O(n^{3.5})$ time.

- **Iterative Hard Thresholding**: Looks a lot like projected gradient descent. Solve $\min_{\mathbf{z}} \|\mathbf{Az} - \mathbf{b}\|$ with gradient descent while continually projecting z back to the set of k-sparse vectors. Runs in time $\sim O(nk \log n)$ for Gaussian measurement matrices and $O(n \log n)$ for subsampled Fourer matrices.
- Other "first order" type methods: Orthogonal Matching Pursuit, CoSaMP, Subspace Pursuit, etc.

Faster Methods

When $\mathbf{A}$ is a subsampled Fourier matrix, there are now methods for computing a k-sparse approximation to $\mathbf{Fx}$ that run in $O(k \log^c n)$ time [Hassanieh, Indyk, Kapralov, Katabi, Price, Shi, etc. 2012+].

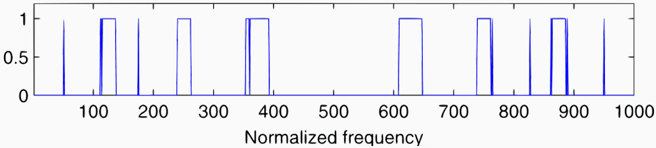
Hold up...

Sparse Fourier Transform

Corollary: When $\mathbf{x}$ is k-sparse, we can compute the inverse Fourier transform $\mathbf{F}^*\mathbf{F}\mathbf{x}$ of $\mathbf{F}\mathbf{x}$ in $O(k \log^c n)$ time!

- Randomly subsample $\mathbf{F}\mathbf{x}$.
- Feed that input into our sparse recovery algorithm to extract $\mathbf{x}$.

Fourier and inverse Fourier transforms in sublinear time when the output is sparse.



Applications in: Wireless communications, GPS, protein imaging, radio astronomy, etc. etc.