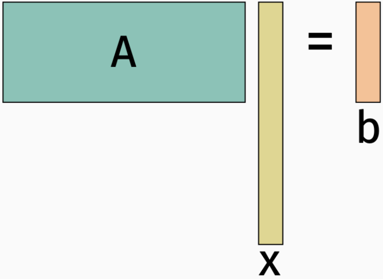
CS-GY 9223 I: Lecture 13
Compressed Sensing + Sparse Recovery

NYU Tandon School of Engineering, Prof. Christopher Musco

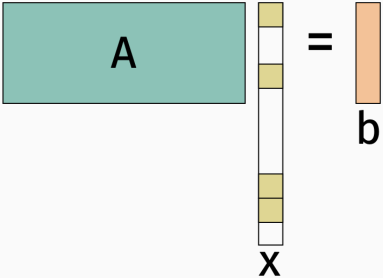
Underdetermined linear regression: Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m < n$, $\mathbf{b} \in \mathbb{R}^m$. Solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$.



- Infinite possible solutions $\mathbf{x}$. In general, impossible to recover parameter vector.

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. $\|\mathbf{x}\|_0 = k$.



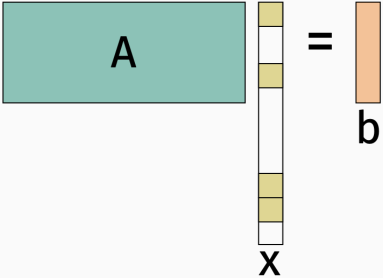
- In many cases can recover $\mathbf{x}$ with $\ll n$ rows. In fact, often $\sim O(k)$ suffice.
- Need additional (strong) assumptions about $\mathbf{A}$!

- In the past, we have thought about $\mathbf{A}$'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 many settings, we will get to choose $\mathbf{A}$'s rows. I.e. each $b_i = \mathbf{x}^T\mathbf{a}_i$ for some vector $\mathbf{a}_i$ that we select.
- In this setting, we often call b_i a linear measurement of $\mathbf{x}$ and we call $\mathbf{A}$ a 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 $\mathbf{A}_i, \mathbf{A}_j$.
    - Recover vectors $\mathbf{x}$ with sparsity $k = 1/\mu$.
- **Focus today**: $\mathbf{A}$ obeys the 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).

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 later in the lecture.

Proof:

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

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

We will not discuss robustness in detail, but 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.

What matrices satisfy this property?

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

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

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

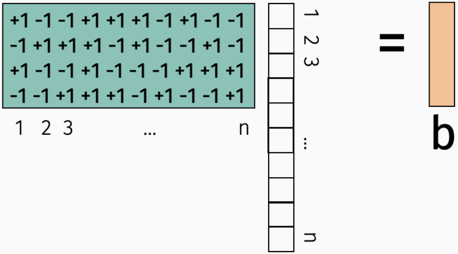
$$4, 18, 4, 1, 2, 24, 6, 4, 3, 18, 18, \ldots$$

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

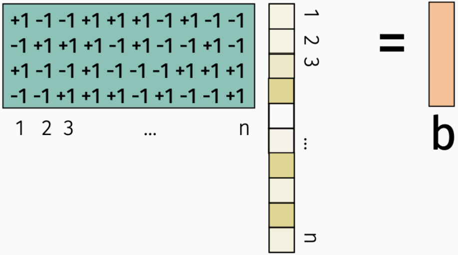
E.g. Amazon is monitoring web-logs to see which product pages people view. They want to figure out which products are viewed most frequently. $n \approx 500$ million.

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 $\mathbf{b}$.



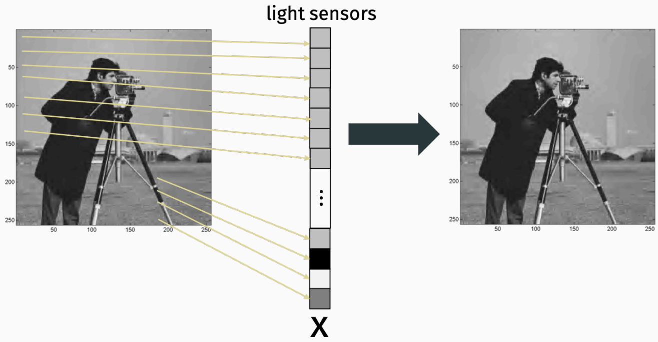
- At the end $\mathbf{b} = \mathbf{A}\mathbf{x}$ for an approximately sparse $\mathbf{x}$ if there were only a few "heavy hitters". Recover $\mathbf{x}$ from $\mathbf{b}$ using a sparse recovery method (like ℓ_0 minimization).

How about when there are insertions or deletions?

$insert(4), insert(18), remove(4), insert(1), insert(2), remove(2) \ldots$

E.g. Amazon is monitoring what products people add to their "wishlist" and wants a list of most tagged products. Wishlists can be changed over time, including by removing items.

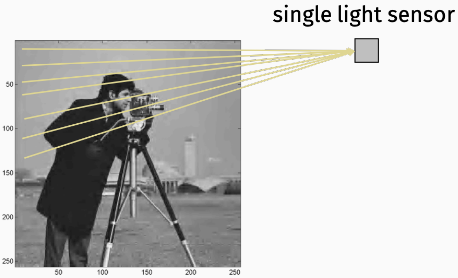
Typical acquisition of image by camera:



Requires one image sensor per pixel captured.

APPLICATION: SINGLE PIXEL CAMERA

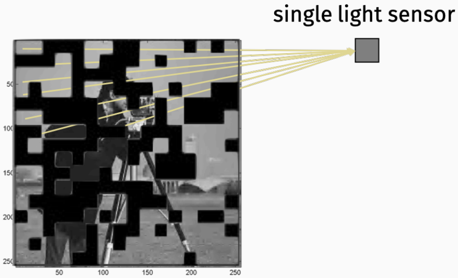
Compressed acquisition of image:



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

Does not provide very much information about the image.

But several random linear measurements do!



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

Applications in:

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

Compressed sensing theory does not exactly describe the problem, but has been very valuable in modeling it.

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

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

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

Uniformly subsampled Fourier matrices with $m \sim O\left(\frac{k \log^2 k \log n}{\epsilon^2}\right)$ rows $(O(k),\epsilon)$-RIP. [Haviv, Regev, 2016].

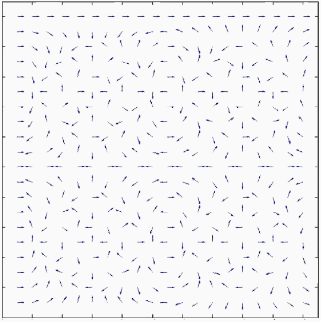
Improves on a long line of work: Candès, Tao, Rudelson, Vershynin, Cheraghchi, Guruswami, Velingker, Bourgain.

You have seen some of the tools used prove this when we proved that a subsampled Hadamard matrix, which is a type of Fourier matrix, can be used to give a *JL* guarantee.

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

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

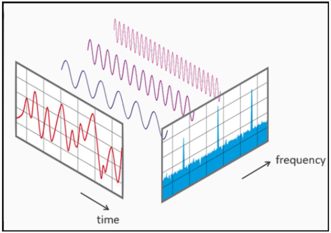
Recall that $e^{\frac{-2\pi i}{n} j \cdot k} = cos(2\pi jk/n) - i \sin(2\pi jk/n)$.



Set $\mathbf{A}$ to contain a random $\tilde{O}(k \log n)$ rows of this matrix.

THE DISCRETE FOURIER MATRIX

Fx is the Discrete Fourier Transform of the vector **x** (what an FFT computes).



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

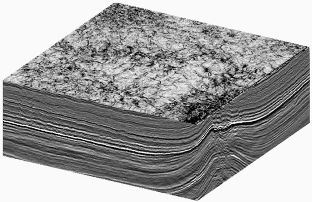
Because $\mathbf{F}^*\mathbf{F} = \mathbf{I}$, $\mathbf{F}^*\mathbf{F}\mathbf{x} = \mathbf{x}$, so we can recover **x** if we have access to its DFT. **Fx**.

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

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

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 $\mathbf{F}\mathbf{x}$?

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

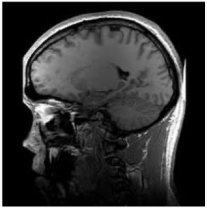
Vibroseis Truck

Can also use airguns, controlled explorations, vibrations from drilling, etc. The fewer measurements we need from **Fx**, the cheaper and faster our data acquisition process becomes.

Killer app: Oil Exploration.

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 $\mathbf{F}\mathbf{x}$?

APPLICATION: GEOPHYSICS

Blast the body with sounds waves waves of varying frequencies.

The fewer measurements we need from Fx, the faster we can acquire and 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

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 <u>deterministically construct</u> 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 Bourgain et al.).

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

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.

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?

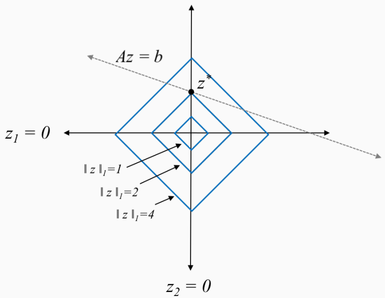
Equivalent formulation:

Problem (Basis Pursuit Linear Program.)

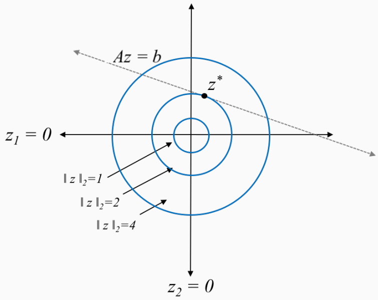
$$\min_{\mathbf{w},\mathbf{z}} \mathbf{1}^T\mathbf{w} \qquad \textit{subject to} \qquad \mathbf{A}\mathbf{z} = \mathbf{b}, -\mathbf{w} \leq \mathbf{z} \leq \mathbf{w}.$$

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

Suppose $\mathbf{A}$ is 2×1, so $\mathbf{b}$ is just a scalar and $\mathbf{x}$ 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.

Theorem

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

Similar proof to ℓ_0 minimization:

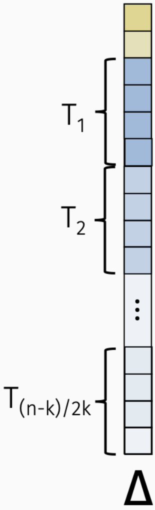
- By way of contradiction, assume $\mathbf{x}$ is <u>not the optimal solution</u>. Then there exists some non-zero Δ such that:
  - $\|\mathbf{x} + \Delta\|_1 \leq \|\mathbf{x}\|_1$
  - $\mathbf{A}(\mathbf{x} + \Delta) = \mathbf{A}\mathbf{x}$. I.e. $\mathbf{A}\Delta = 0$.

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

Only one tool needed:

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

Some definitions:



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

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

Finish up proof by contradiction:

A lot 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 $\min_{\mathbf{z}} \|\mathbf{A}\mathbf{z} - \mathbf{b}\|$ 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.

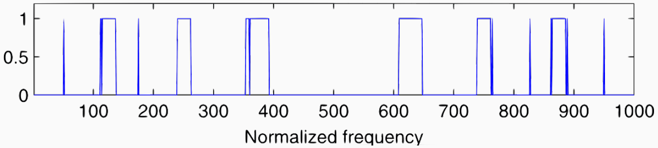
When $\mathbf{A}$ is a subsampled Fourier matrix, there are now methods that run in $O(k \log^c n)$ time [Hassanieh, Indyk, Kapralov, Katabi, Price, Shi, etc. 2012+].

Hold up...

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 <u>sublinear time</u> when the output is sparse.



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