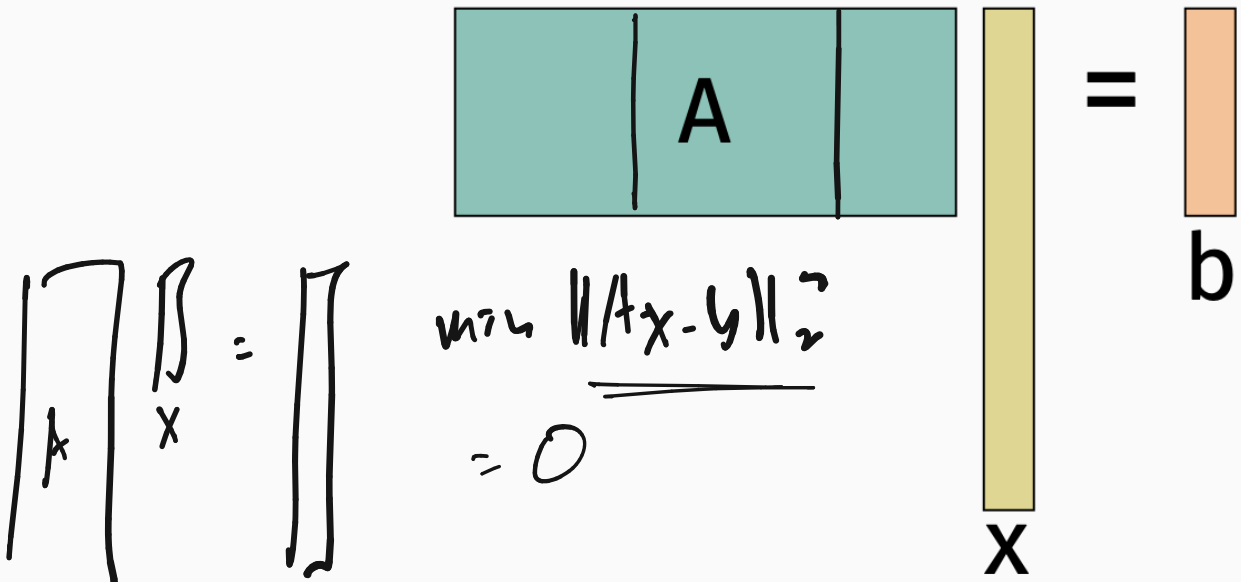
CS-GY 9223 I: Lecture 13
Compressed Sensing + Sparse Recovery

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 $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{x}$.

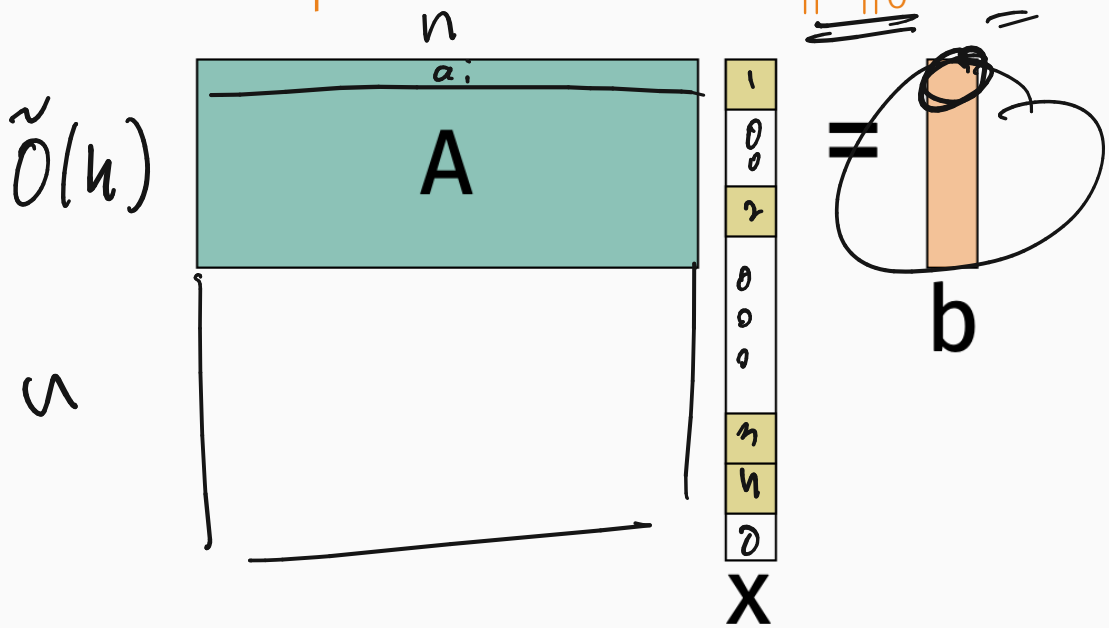


- Infinite possible solutions $\mathbf{x}$. In general, impossible to recover parameter vector.

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 (strong) assumptions about $\mathbf{A}$!

QUICK ASIDE

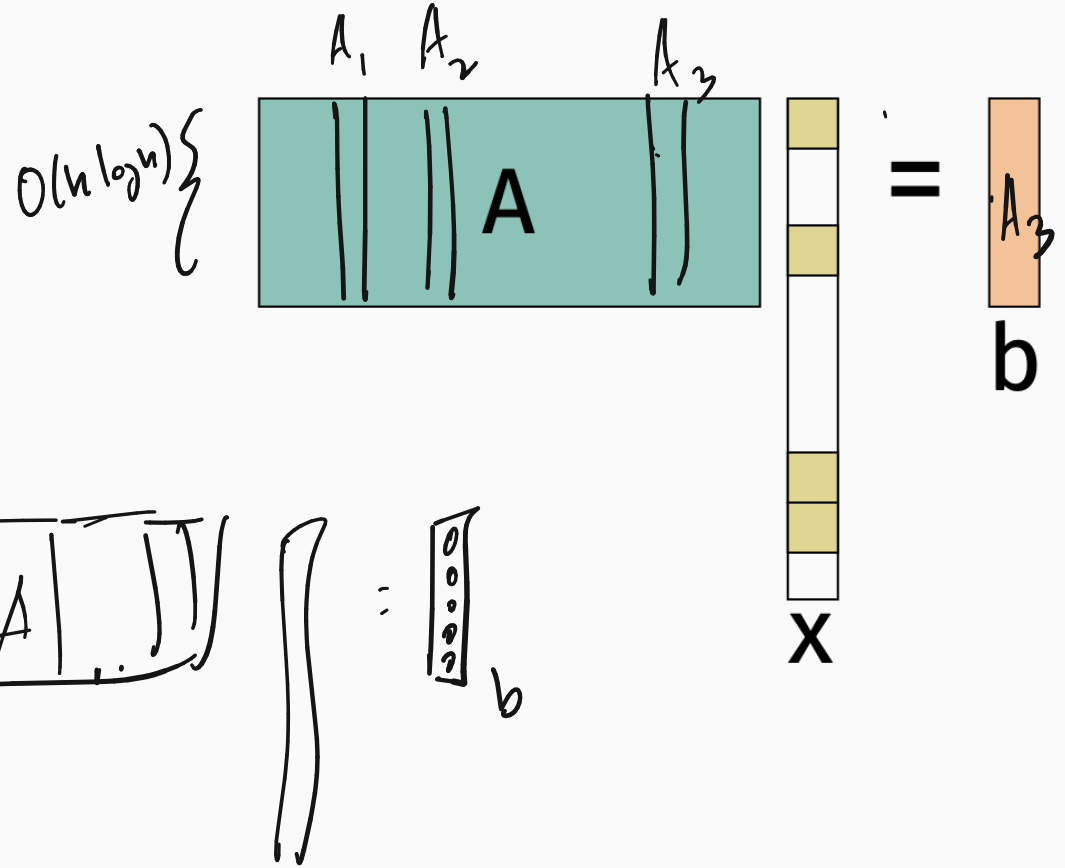
- In the past, we have thought about $\mathbf{A}$'s rows as data drawn from some universe/distribution:

	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.

ASSUMPTIONS ON MEASUREMENT MATRIX

When should this problem be difficult?



ASSUMPTIONS ON MEASUREMENT MATRIX

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 $k \leq r/2$.
- $\mathbf{A}$ 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 $k \leq 1/\mu$.
- **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:*

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

- 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.

$2^* = X$

FIRST SPARSE RECOVERY RESULT

Proof: By contradiction:

Assume by way of contradiction that

$$z^* \neq x.$$

$$\|z^*\|_0 \leq \|x\|_0 = k.$$

Consider $(z^* - x)$ $\longrightarrow$ $2k$-sparse vector.

$$Az^* = b \qquad Ax = b$$

$$\|A(z^* - x)\|_2 = 0$$

$$\|A(z^* - x)\|_2 \geq (1-\varepsilon)\|z^* - x\|_2$$

($z^* - x \neq 0$)

($z^* - x$: $2k$ sparse; $2k$-RIP)

Contradiction if $\varepsilon < 1$.

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:

Suppose $\mathbf{A}$ is $(O(1), 1/2)$-RIP

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

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 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**.

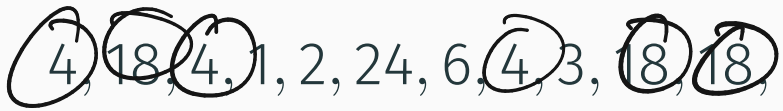
What matrices satisfy this property?

- 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 $\mathbf{A}$.

APPLICATION: HEAVY HITTERS IN DATA STREAMS

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

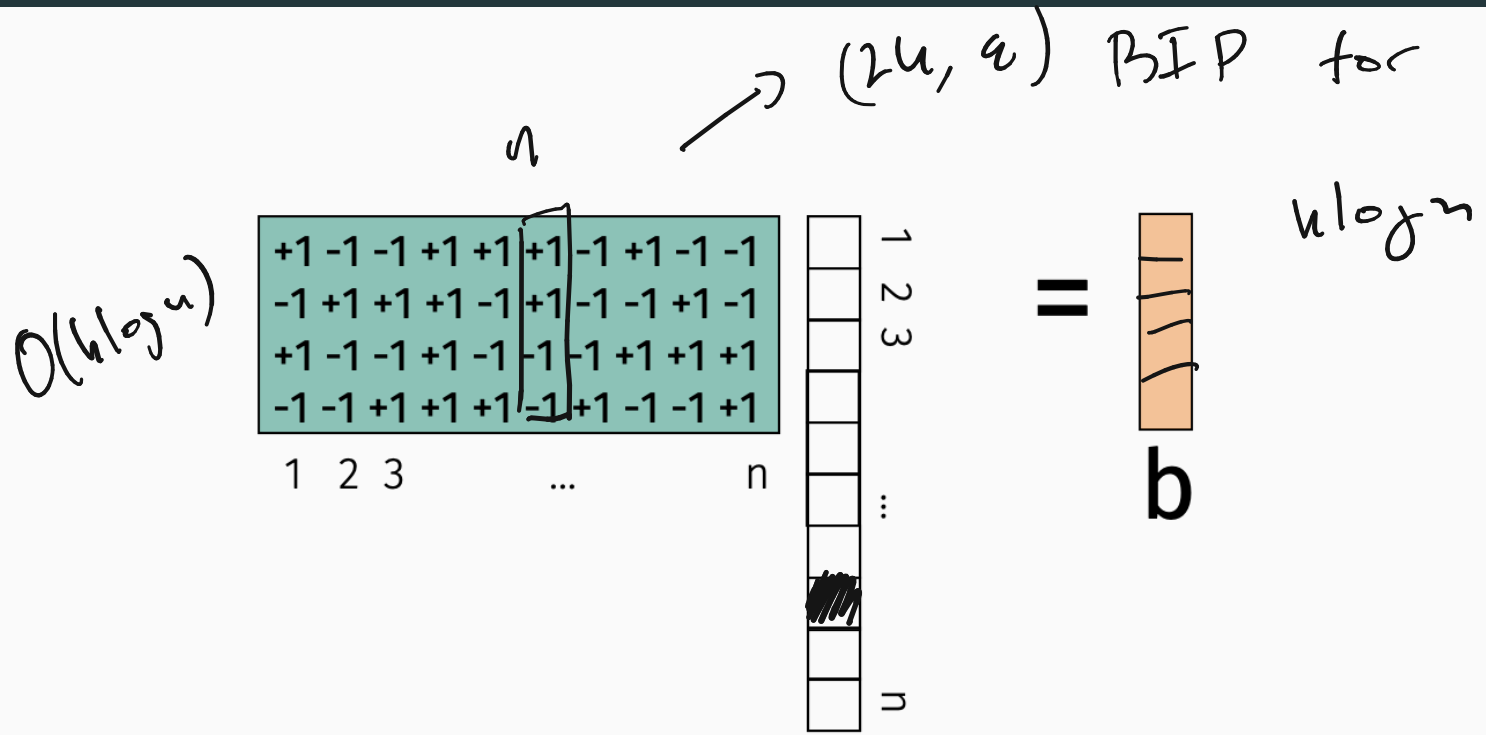


After some time, you want to report which k items appeared most frequently in the stream.

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.

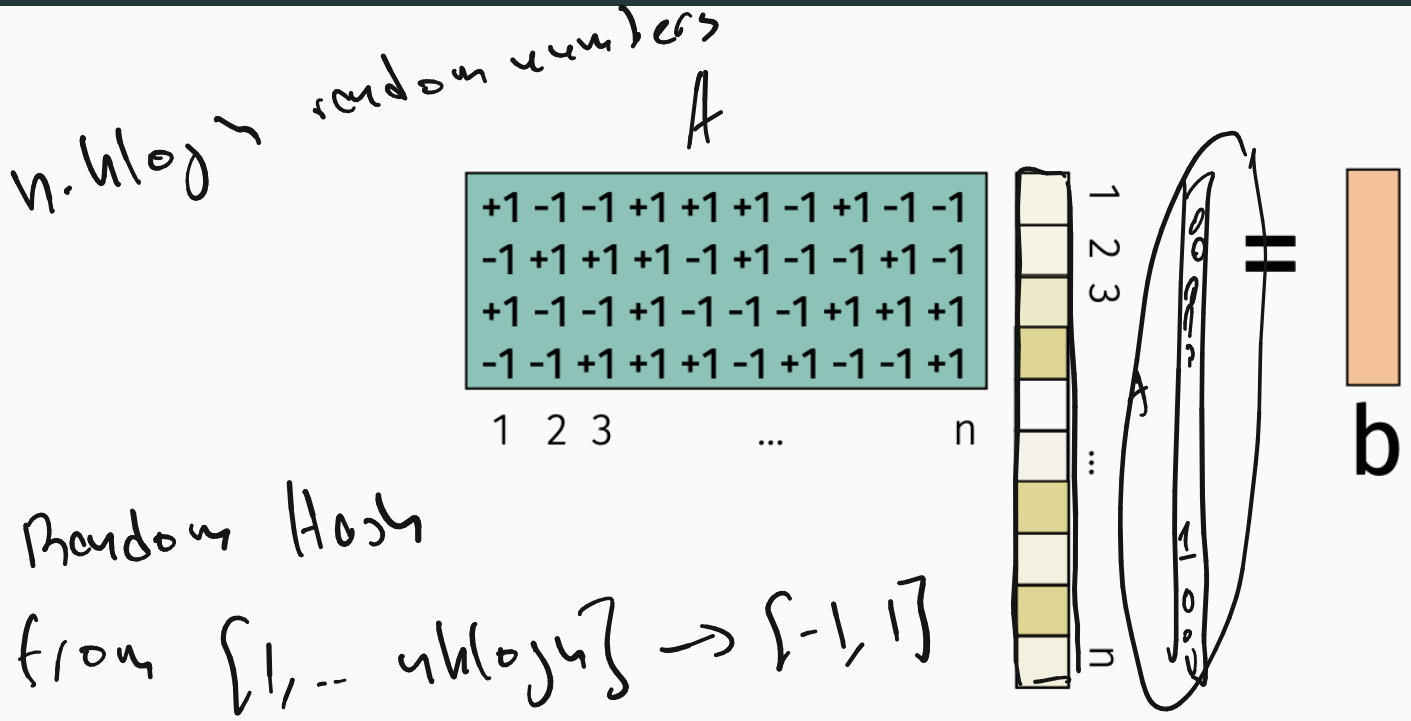
How can you do this quickly in small space?

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: Heavy hitters in data streams

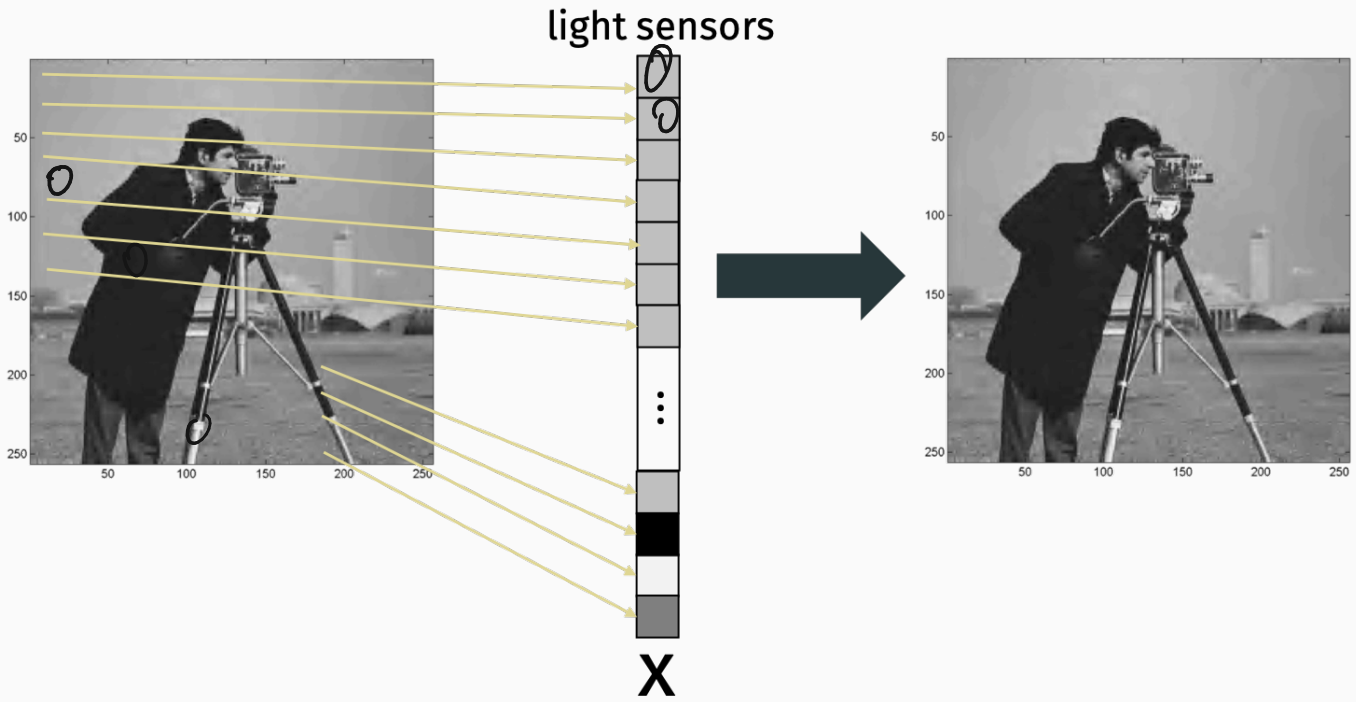
How about when there are insertions or deletions?

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

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.

APPLICATION: SINGLE PIXEL CAMERA

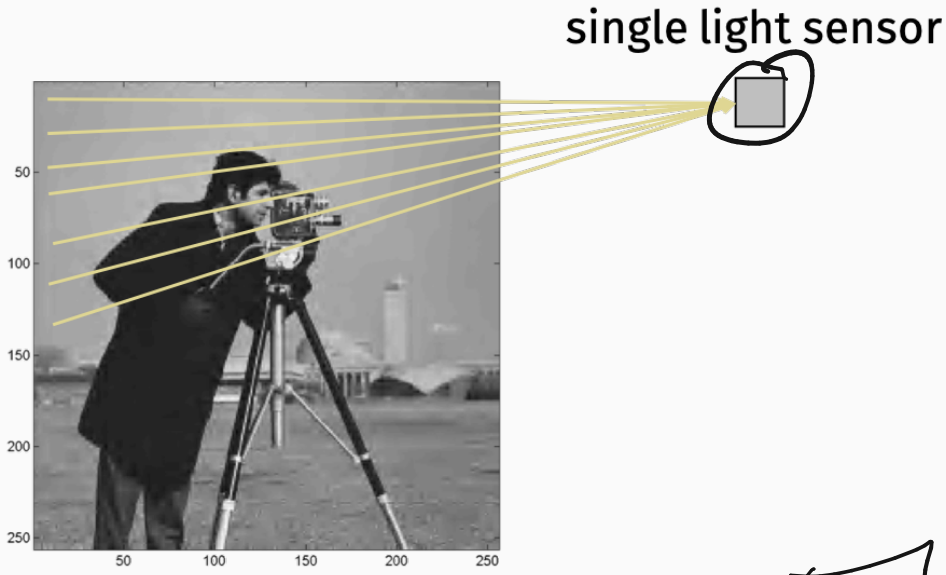
Typical acquisition of image by camera:



Requires one image sensor per pixel captured.

APPLICATION: SINGLE PIXEL CAMERA

Compressed acquisition of image:

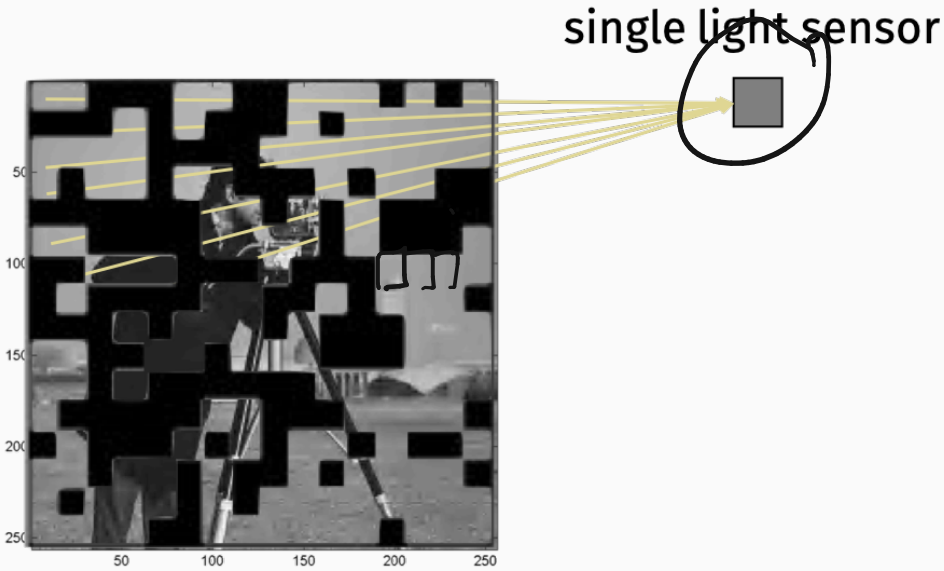


$$p = \sum_{i=1} x_i = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Does not provide very much information about the image.

APPLICATION: SINGLE PIXEL CAMERA

But several random linear measurements do!



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

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 the problem, but has been very valuable in modeling it.

RESTRICTED ISOMETRY PROPERTY

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

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

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

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.

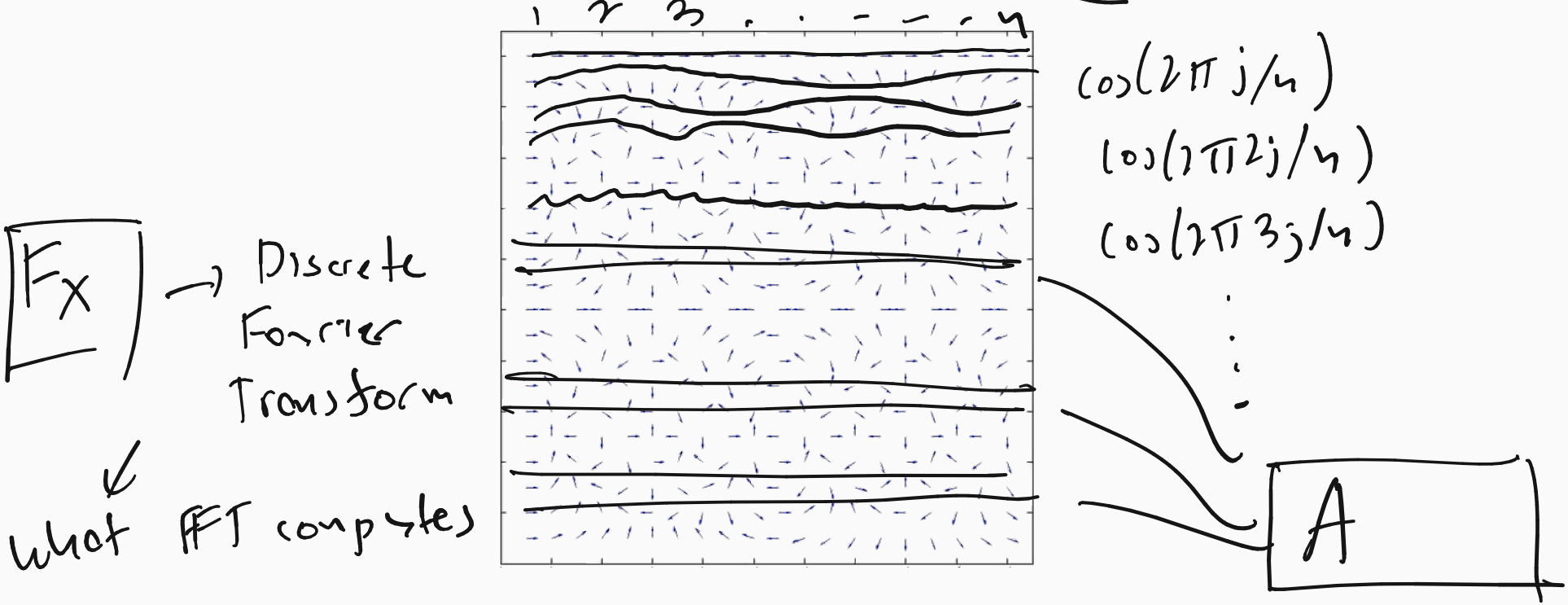
You have seen some of the tools used prove this when we proved that a subsampled Hadamard matrix, which is a type of Fourier matrix, can be used to give a *JL* guarantee.

THE DISCRETE FOURIER MATRIX

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

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

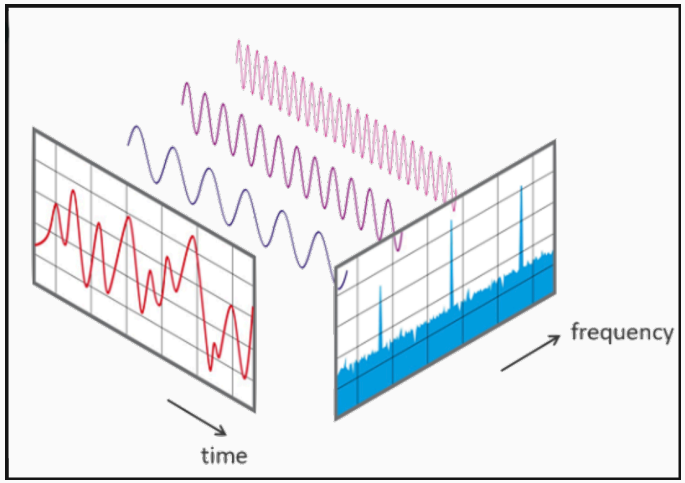
Recall that $e^{\frac{-2\pi i}{n} j \cdot k} = \cos(2\pi jk/n) - i \sin(2\pi jk/n)$.



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

THE DISCRETE FOURIER MATRIX

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

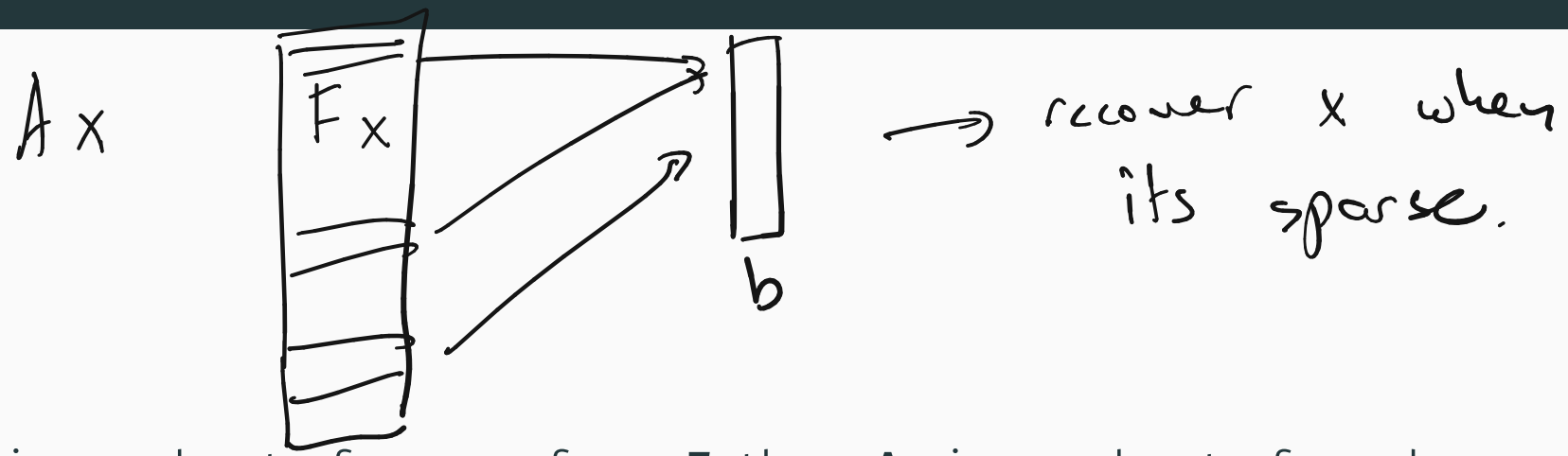


Decomposes $\mathbf{x}$ into different frequencies: $[\mathbf{Fx}]_j$ is the component with frequency j/n.

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. $\mathbf{Fx}$.

"Inverse Discrete Fourier Transform"

THE DISCRETE FOURIER MATRIX

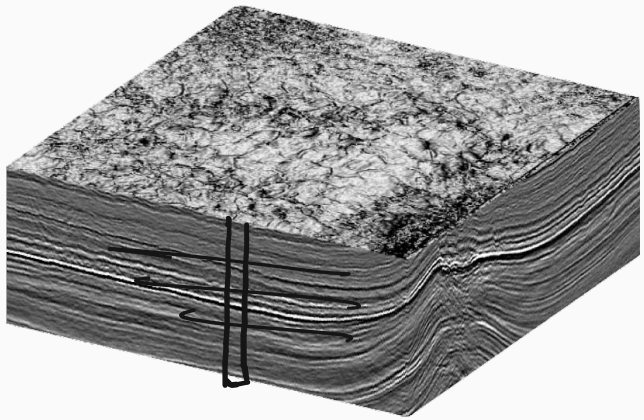


If $\mathbf{A}$ is a subset of q rows from $\mathbf{F}$, then $\mathbf{Ax}$ is a subset of random frequency components from $\mathbf{x}$'s discrete Fourier transform.

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

Warning: very cartoonish explanation of very complex problem.

Understanding what material is beneath the crust:



Think of vector **x** as scalar values of the density/reflectivity in a single vertical core of the earth.

How do we measure entries of Fourier transform **Fx**?

$[Fx]_j \rightarrow$ specific freq j.

APPLICATION: GEOPHYSICS

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

Vibroseis Truck

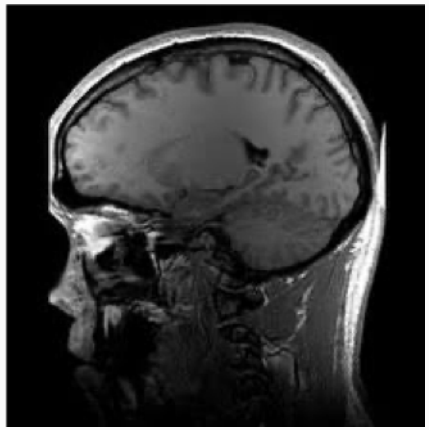
Can also use airguns, controlled explorations, vibrations from drilling, etc. The fewer measurements we need from Fx, the cheaper and faster our data acquisition process becomes.

Killer app: Oil Exploration.

APPLICATION: GEOPHYSICS

Warning: very cartoonish explanation of very complex problem.

Medical Imaging (MRI, ~~Ultrasound,~~ etc.)



Vector **x** here is a 2D image. Everything works with 2D Fourier transforms.

How do we measure entries of Fourier transform **Fx**?

APPLICATION: GEOPHYSICS

Blast the body with sounds waves ~~or ultrasonic 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 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 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:*

$O(k \log n)$ rows

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

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.)

$$\min_z \|z\|_1 \quad \textit{subject to} \quad Az = b.$$

- Objective is convex:
- Optimizing over convex set:

What is one method we know for solving this problem?

Equivalent formulation:

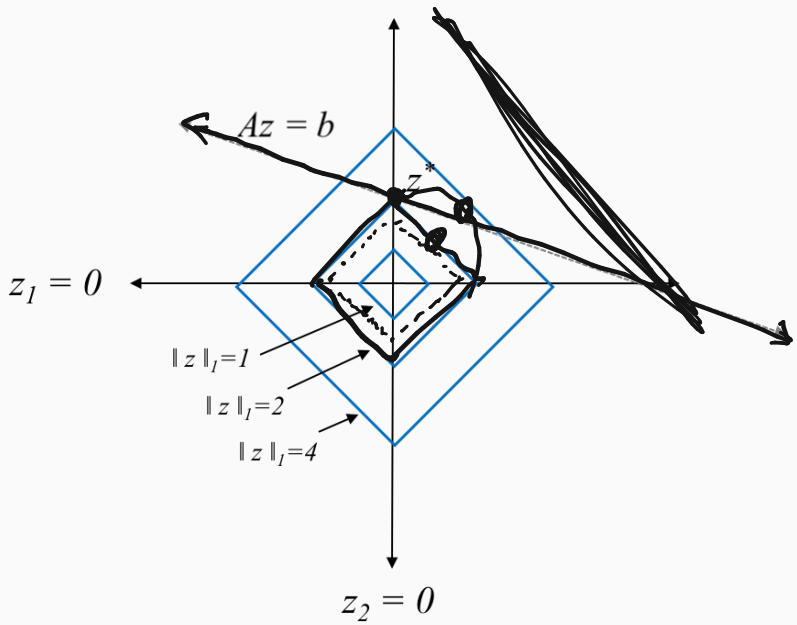
Problem (Basis Pursuit Linear Program.)

$$\min_{w,z} 1^T w \quad \textit{subject to} \quad Az = b, -w \leq z \leq w.$$

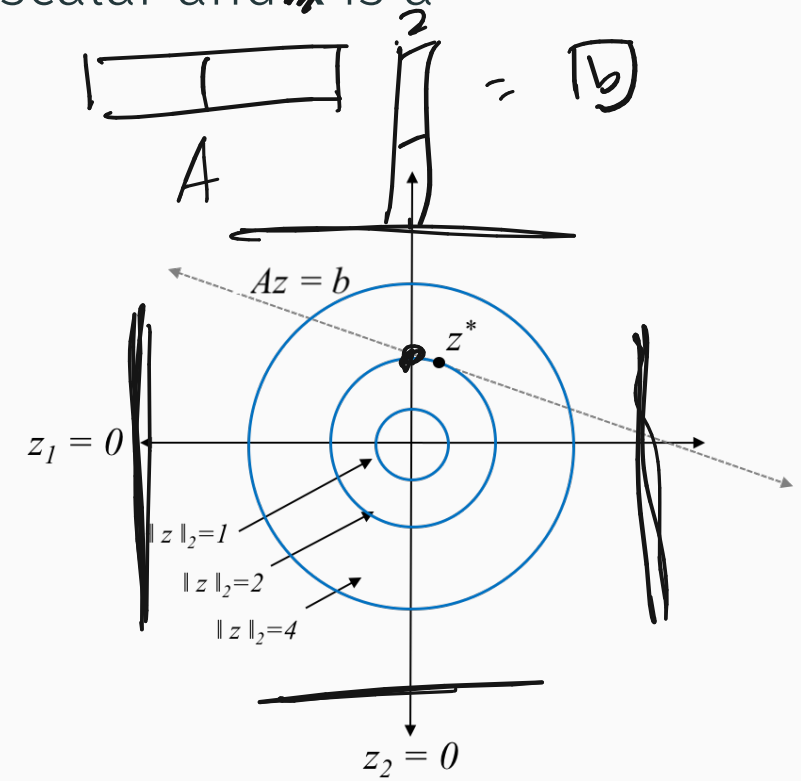
Can be solved using any algorithm for linear programming. An Interior Point Method will run in at worst $\sim O(n^{3.5})$ time.

BASIS PURSUIT INTUITION

Suppose **A** is 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.

$(2k, 4)$ for ℓ_0 minimization

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 not the optimal solution. 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. $\mathbf{A}\Delta = 0$.

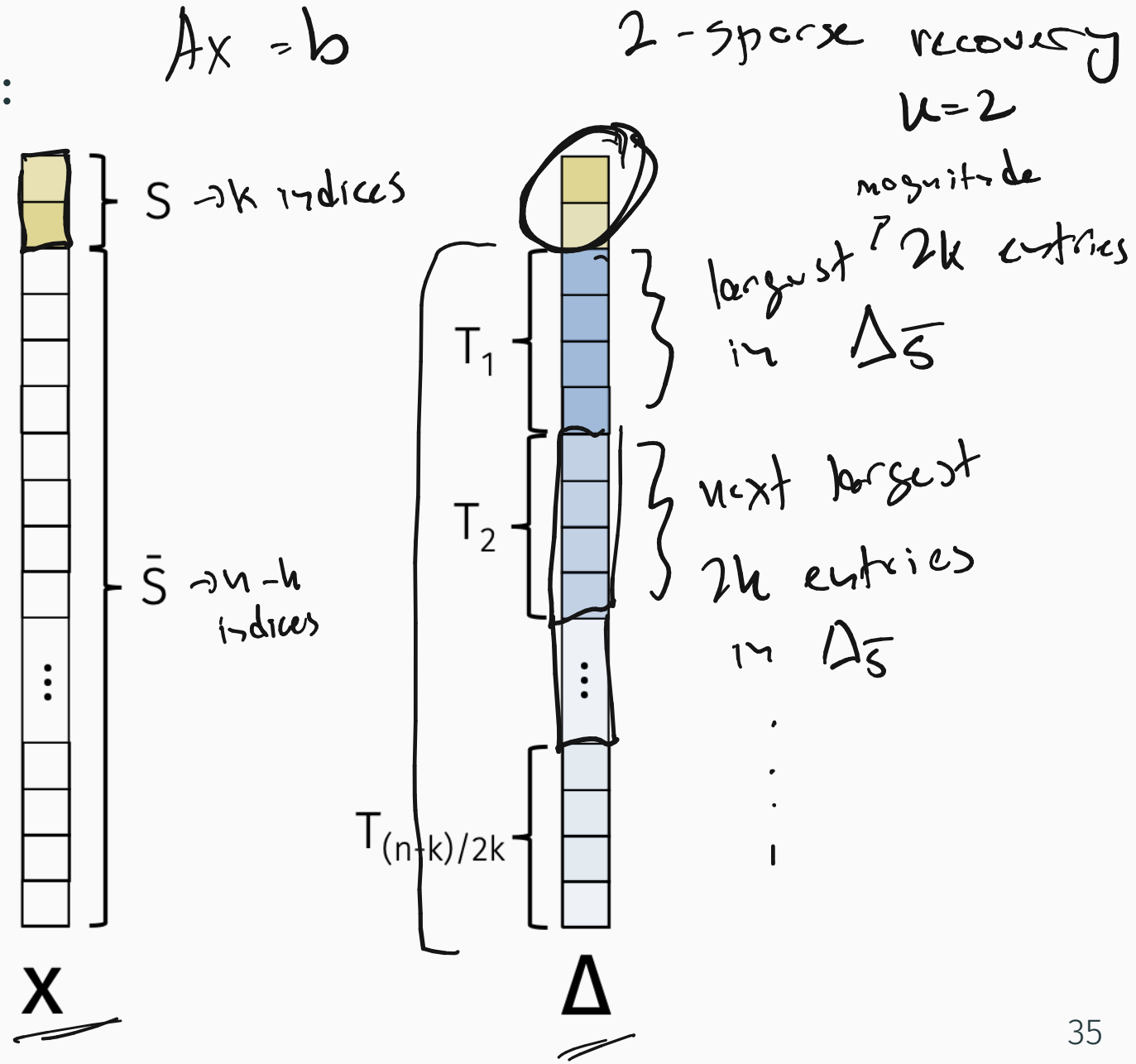
Difference is that we can no longer assume that Δ is sparse.

Only one tool needed:

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

$w = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ $\|w\|_1 = \|w\|_2 = 1$ $w = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix}$ 1's in k locations $\|w\|_1 = q$ $\|w\|_2 = \sqrt{q}$

Some definitions:

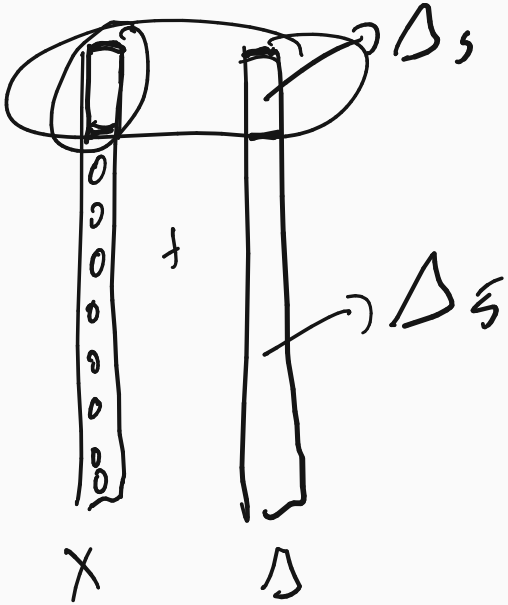


Claim 1: $\|\Delta_S\|_1 \geq \|\Delta_{\bar{S}}\|_1$

$A\Delta = 0$

$x+\Delta$ is a solution better than x, conjectured to exist by way of contradiction

$\|x+\Delta\|_1 \leq \|x\|_1$.



$$\|x+\Delta\|_1 = \|x+\Delta_S\|_1 + \|\Delta_{\bar{S}}\|_1$$

$$\geq \|x\|_1 - \|\Delta_S\|_1 + \|\Delta_{\bar{S}}\|_1$$

$$= \|x\|_1 + (\|\Delta_{\bar{S}}\|_1 - \|\Delta_S\|_1)$$

if positive

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

$$\sqrt{k}\|\Delta_S\|_2 \geq \|\Delta_S\|_1$$

$$\|\Delta_S\|_1 \geq \sqrt{2k} \sum_{j \geq 2} \|\Delta_{T_j}\|_2$$

1. $\|\Delta_S\|_1 = \sum_{j \geq 1} \|\Delta_{T_j}\|_1$

2. For all j, $\|\Delta_{T_j}\|_1 \geq \sqrt{2k}\, \|\Delta_{T_{j+1}}\|_2$

$\|\Delta_{T_j}\|_1 \geq 2k \cdot \min(|\Delta_{T_j}|)$

$\geq \|\Delta_{T_{j+1}}\|_2$

$\|\Delta_{T_{j+1}}\|_2 \leq \sqrt{2k \cdot \min(|\Delta_{T_j}|)^2}$

$\leq \sqrt{2k} \cdot \min(|\Delta_{T_j}|)$

$\|\Delta_{T_j}\|_1 \geq \sqrt{2k}\, \|\Delta_{T_{j+1}}\|_2$

Finish up proof by contradiction:

$(3k, \varepsilon)$ - RIP

$$\|\Delta_S\|_1 \geq \sqrt{2} \sum_{j \geq 2} \|\Delta_{T_j}\|_2$$

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

$$\geq (1-\varepsilon)\|\Delta_{S \cup T_1}\|_2 - (1+\varepsilon) \sum_{j \geq 2} \|\Delta_{T_j}\|_2$$

$$\geq (1-\varepsilon)\|\Delta_S\|_2 - (1+\varepsilon)\frac{1}{\sqrt{2}}\|\Delta_S\|_2$$

$$= \left(1 - \varepsilon - \frac{(1+\varepsilon)}{\sqrt{2}}\right)\|\Delta_S\|_2$$

contradiction when positive

FASTER METHODS

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 $\min_{\mathbf{z}} \|\mathbf{Az} - \mathbf{b}\|$ 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.

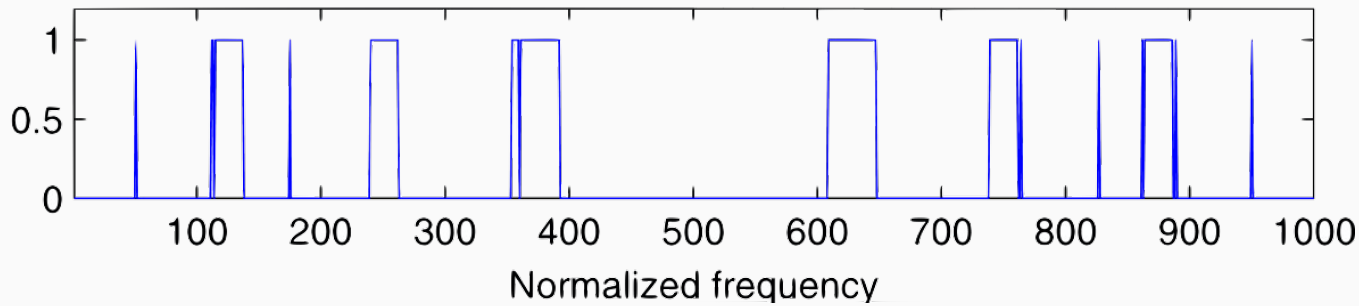
When $\mathbf{A}$ is a subsampled Fourier matrix, there are now methods that run in $O(k \log^c n)$ time [Hassanieh, Indyk, Kapralov, Katabi, Price, Shi, etc. 2012+].

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 sublinear time when the output is sparse.



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