Sparsity: Compressed Sensing

Rebecca Willett
SENSORS, SENSORS EVERYWHERE
Sensing systems limited by constraints: physical size, time, cost, energy

Reduce the number of measurements needed for reconstruction

Higher accuracy data subject to constraints
Original Scene  Downsampled  Reconstruction from $\frac{1}{4}$ as many measurements
Original Scene  |  Downsampled  |  Reconstruction from $\frac{1}{4}$ as many measurements
CONVENTIONAL IMAGING

Each observation is a measurement of ONE pixel
Each observation is a measurement of ONE pixel
Images are compressible

Measuring all pixels inherently wasteful
**NEW PARADIGM FOR SENSING**

\[ y_1 = \langle f, r_1 \rangle \]

\[ = \langle \begin{array}{c}
\cdot \\
\end{array}, \begin{array}{c}
\end{array} \rangle \]

Measure sum of half the pixels

\[ \downarrow \]

Narrow down star location
NEW PARADIGM FOR SENSING

\[ y_1 = \langle f, r_1 \rangle = \langle \begin{array}{c} \cdot \\ \cdot \end{array}, \begin{array}{c|c} 0 & 1 \\ \hline 1 & 0 \end{array} \rangle \]

\[ y_2 = \langle f, r_2 \rangle = \langle \begin{array}{c} \cdot \\ \cdot \end{array}, \begin{array}{c|c|c} 0 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline 0 & 1 & 0 \end{array} \rangle \]

\[ \vdots \]

\[ y_M = \langle f, r_M \rangle = \langle \begin{array}{c} \cdot \\ \cdot \end{array}, \begin{array}{c|c|c|c} 0 & 1 & \cdots & 0 \\ \hline 1 & 0 & \cdots & 1 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 1 & \cdots & 0 \end{array} \rangle \]

Each observation is a measurement of half the pixels
These ideas extend to multiple stars and random combinations of pixels.
New Observation Model

\[ y = Rf + n \]

Observations

All \( k \) random projections

True image

Noise
ILL-POSED PROBLEM

System is underdetermined: infinitely many solutions
**SPARSITY**

Assume $f$ is $K$-sparse or $\beta$-compressible in some basis $\Psi$. That is,

$$f = \sum_{i=1}^{N} \theta_i \psi_i$$

and either

$$\|\theta\|_0 \leq K$$

or

$$\|f - f_K\| \leq K^{-\beta}$$

where $f_K$ is the best $K$-term approximation of $f$ in the basis $\Psi$. 
SPARSITY
Combining $y = Rf + n$ with $f = \psi \theta$:
**COMPRESSED SENSING**

\[
\hat{\theta} = \arg \min_{\theta} \left\| y - R\Psi \theta \right\|_2^2 + \frac{\tau}{2} \left\| \theta \right\|_1
\]

\[
\hat{f} = \Psi \hat{\theta}
\]

**Key theory:** If \( R \) meets certain conditions and \( f \) is sparse or compressible in \( \Psi \), then \( \hat{f} \) will be very accurate even when the number of measurements is small relative to \( N \).
CONVENTIONAL SENSING

Noisy Image

COMPRESSED SENSING

Random Projections
Smaller
Less Data
Cheaper
**Definition:** Restricted Isometry Property. The matrix $A$ satisfies the Restricted Isometry Property of order $K$ with parameter $\delta_K \in [0, 1)$ if

$$(1 - \delta_K)\|\theta\|_2^2 \leq \|A\theta\|_2^2 \leq (1 + \delta_K)\|\theta\|_2^2$$

holds simultaneously for all $K$-sparse vectors $\theta$. Matrices with this property are denoted $\text{RIP}(K, \delta_K)$. 

---

Candes and Tao (2006)
RIP Example

For example, if the entries of $A$ are independent and identically distributed according to

$$A_{i,j} \sim \mathcal{N}\left(0, \frac{1}{M}\right)$$

or

$$A_{i,j} = \begin{cases} M^{-1/2} & \text{with probability } \frac{1}{2} \\ -M^{-1/2} & \text{with probability } \frac{1}{2} \end{cases}$$

then $A$ satisfies RIP($K, \delta_K$) with high probability for any integer $K = O(M/\log N)$. 
Matrices which satisfy the RIP combined with sparse recovery algorithms are guaranteed to yield accurate estimates of the underlying function \( f \), as specified by the following theorem.

**Theorem:** Noisy Sparse Recovery with RIP Matrices. Let \( A \) be a matrix satisfying RIP \((2K, \delta_{2K})\) with \( \delta_{2K} < \sqrt{2} - 1 \), and let \( y = A\theta + n \) be a vector of noisy observations of any signal \( \theta \in \mathbb{R}^N \), where the \( n \) is a noise or error term with \( \|n\|_2 \leq \epsilon \). Let \( \theta_K \) be the best \( K \)-sparse approximation of \( \theta \). Then the estimate

\[
\hat{\theta} = \arg \min_{\theta} \|\theta\|_1 \text{ subject to } \|y - A\theta\|_2 \leq \epsilon
\]

obeys

\[
\|\theta - \hat{\theta}\|_2 \leq C_{1, K}\epsilon + C_{2, K}\frac{\|\theta - \theta_K\|_1}{\sqrt{K}},
\]

where \( C_{1, K} \) and \( C_{2, K} \) are constants which depend on \( K \) but not on \( N \) or \( M \).

Candes (2006)
This estimate can be computed in a variety of ways.

Many off-the-shelf optimization software packages are unsuitable

- Can’t handle large $N$
- Our objective isn’t differentiable
- Don’t exploit fast transforms (e.g. Fourier and wavelet)

Gradient projection methods

- Introduce additional variables and recast problem as constrained optimization with differentiable objective
- Projection onto constraint set can be done with thresholding
- More robust to noise

Orthogonal matching pursuits (OMP)

- Start with estimate $= 0$
- Greedily choose elements of estimate to have non-zero magnitude by iteratively processing residual errors
- Very fast when little noise

\[ \hat{\theta} = \arg \min_{\tilde{\theta}} \| y - A\tilde{\theta} \|^2_2 + \tau \| \tilde{\theta} \|_1. \]
 ITERATIVE HARD/SOFT THRESHOLDING

Our objective is

$$\hat{\theta} = \arg \min_{\tilde{\theta}} \| y - A\tilde{\theta} \|_2^2 + \tau \|\tilde{\theta}\|_1.$$  

The first term can be re-written as

$$y^T y - 2\tilde{\theta}^T A^T y + \tilde{\theta}^T A^T A \tilde{\theta}$$

and its gradient is

$$-2 A^T (y - A\tilde{\theta}).$$

This suggests a simple strategy for computing $\hat{\theta}$: start with an initial estimate $\tilde{\theta}$, update it by adding a step in the negative gradient direction, then apply thresholding!
**Iterative Hard/Soft Thresholding**

Start with some initial estimate \( \hat{\theta}^{(0)} \); see how well it fits \( y \):

\[
y - A\hat{\theta}^{(0)}.\]

Use this residual to update the initial estimate:

\[
\hat{\theta}^{(0)} + A^T \left( y - A\hat{\theta}^{(0)} \right).
\]

Impose sparsity via thresholding this estimate:

\[
\hat{\theta}^{(1)} = \text{threshold} \left[ \hat{\theta}^{(0)} + A^T \left( y - A\hat{\theta}^{(0)} \right) \right]
\]

Repeat until \( \| y - A\hat{\theta}^{(i)} \| \) is small:

\[
\hat{\theta}^{(i+1)} = \text{threshold} \left[ \hat{\theta}^{(i)} + A^T \left( y - A\hat{\theta}^{(i)} \right) \right].
\]
EXAMPLE

Time domain $f(t)$

Frequency domain $\hat{f}(\omega)$

Measure $M$ samples
(red circles = samples)

$K$ nonzero components
$\#\{\omega : \hat{f}(\omega) \neq 0\} = K$
**Example**

Original $\theta$, with $K = 15$

$f$ (blue) and $y$ (red circles); $M = 30$

perfect reconstruction!
Matrices which satisfy the RIP combined with sparse recovery algorithms are guaranteed to yield accurate estimates of the underlying function $f$, as specified by the following theorem.

**Theorem:** Noisy Sparse Recovery with RIP Matrices. Let $A$ be a matrix satisfying RIP$(2K, \delta_{2K})$ with $\delta_{2K} < \sqrt{2} - 1$, and let $y = A\theta + n$ be a vector of noisy observations of any signal $\theta \in \mathbb{R}^N$, where the $n$ is a noise or error term with $\|n\|_2 \leq \epsilon$. Let $\theta_K$ be the best $K$-sparse approximation of $\theta$. Then the estimate

$$\hat{\theta} = \arg\min_{\theta} \|\theta\|_1 \text{ subject to } \|y - A\theta\|_2 \leq \epsilon$$

obeys

$$\|\theta - \hat{\theta}\|_2 \leq C_{1,K}\epsilon + C_{2,K} \frac{\|\theta - \theta_K\|_1}{\sqrt{K}},$$

where $C_{1,K}$ and $C_{2,K}$ are constants which depend on $K$ but not on $N$ or $M$.

PROOF

Let $h \triangleq \hat{\theta} - \theta$ be our error vector.

Let $T_0$ be the indices of the largest $K$ elements of $\theta$, $T_1$ be the indices of the largest $K$ elements of $h_{T_0^c}$, $T_2$ be the indices of the next $K$ largest elements of $h_{T_0^c}$, and so on. For a vector $x$, let $x_{T_j}$ be defined via

$$x_{T_j,i} \triangleq \begin{cases} x_i, & i \in T_j \\ 0, & i \notin T_j \end{cases}$$

Then $h = h_{T_0} + h_{T_1} + h_{T_2} + \ldots$

There are two main steps to our proof:

$$||\hat{\theta} - \theta||_2 = ||h||_2 \leq ||h_{T_0 \cup T_1}||_2 + ||h_{(T_0 \cup T_1)^c}||_2$$

(STEP 1) $\leq C ||h_{T_0 \cup T_1}||_2 + CK^{-1/2}||\theta - \theta_K||_1$

(STEP 2) $\leq C\epsilon + CK^{-1/2}||\theta - \theta_K||_1$

$C$ will represent constants which may depend on $K$ but not $N$ or $M$. 
**Step 1**

\[
\| h_{(T_0 \cup T_1)^c} \|_2 = \left\| \sum_{j \geq 2} h_{T_j} \right\|_2 \\
\leq \sum_{j \geq 2} \| h_{T_j} \|_2 \quad \text{(remember me later!!)} \\
\leq \sum_{j \geq 2} K^{1/2} \| h_{T_j} \|_\infty \\
\leq \sum_{j \geq 2} K^{1/2} \| h_{T_{j-1}} \|_1 / K \\
= K^{-1/2} (\| h_{T_1} \|_1 + \| h_{T_2} \|_1 + \ldots) \\
= K^{-1/2} \underbrace{\| h_{T_0^c} \|_1}_{\text{how big??}}
\]
**Step 1**

First note

\[ \| \theta \|_1 \geq \| \hat{\theta} \|_1 = \| \theta + h \|_1 \]
\[ \geq \| \theta_{T_0} \|_1 - \| h_{T_0} \|_1 + \| h_{T_0^c} \| - \| \theta_{T_0^c} \|_1 \]

Rearranging terms we find

\[ \| h_{T_0^c} \|_1 \leq \| h_{T_0} \|_1 + 2 \| \theta_{T_0^c} \|_1 \]
\[ = \| h_{T_0} \|_1 + 2 \| \theta - \theta_K \|_1 \]

Putting everything together we have

\[ \| h_{(T_0 \cup T_1)^c} \|_2 \leq K^{-1/2}(\| h_{T_0} \|_1 + 2 \| \theta - \theta_K \|_1) \]
\[ \leq \| h_{T_0 \cup T_1} \|_2 + 2K^{-1/2} \| \theta - \theta_K \|_1 \]

as desired for Step 1.
**Step 2**

We now need to bound \( \| h_{T_0 \cup T_1} \|_2 \). Note

\[
(1 - \delta_{2K}) \| h_{T_0 \cup T_1} \|_2^2 \leq \| Ah_{T_0 \cup T_1} \|_2^2
\]

\[
= \langle Ah_{T_0 \cup T_1}, Ah \rangle - \langle Ah_{T_0 \cup T_1}, \sum_{j \geq 2} Ah_{T_j} \rangle
\]

For the first term

\[
\langle Ah_{T_0 \cup T_1}, Ah \rangle \leq \| Ah_{T_0 \cup T_1} \|_2 \| Ah \|_2
\]

\[
\leq (\sqrt{1 + \delta_{2K} \| h_{T_0 \cup T_1} \|_2}) \| A(\hat{\theta} - \theta) \|_2
\]

\[
\leq (\sqrt{1 + \delta_{2K} \| h_{T_0 \cup T_1} \|_2})(\| A\hat{\theta} - y \|_2 + \| y - A\theta \|_2)
\]

\[
\leq (\sqrt{1 + \delta_{2K} \| h_{T_0 \cup T_1} \|_2})2\epsilon
\]

The second term is bounded similarly by

\[-\langle Ah_{T_0 \cup T_1}, \sum_{j \geq 2} Ah_{T_j} \rangle \leq \sqrt{2}\delta_{2K} \sum_{j \geq 2} \| h_{T_j} \|_2 \| h_{T_0 \cup T_1} \|_2\]
Thus

\[(1 - \delta_{2K}) \| h_{T_0 \cup T_1} \|_2^2 \leq \| h_{T_0 \cup T_1} \|_2 \left(2\epsilon \sqrt{1 + \delta_{2K}} + \sqrt{2}\delta_{2K} \sum_{j \geq 2} \| h_{T_j} \|_2 \right)\]

\[\| h_{T_0 \cup T_1} \|_2 \leq C\epsilon + CK^{-1/2} \| \theta - \theta_K \|_1.\]

Putting it all together we have

\[\| \hat{\theta} - \theta \|_2 \leq C\epsilon + CK^{-1/2} \| \theta - \theta_K \|_1\]

as desired.
In other words, the accuracy of the reconstruction of a general image \( f \) from measurements collected using a system which satisfies the RIP depends on (a) the amount of noise present and (b) how well \( f \) may be approximated by an image sparse in \( \Psi \).

If we have no noise (\( \epsilon = 0 \)) and our signal is \( K \)-sparse, then we have

\[
\theta = \hat{\theta};
\]

i.e., we can perfectly reconstruct the original signal!
**Solvability Boundary**

Unsolvable; too little data or too little sparsity

Solvable; sufficient data and sparsity

Donoho and Tanner (2010)
ANOTHER PERSPECTIVE

Consider the worst-case coherence of $A \equiv R\Psi$. Formally, one denotes the Gram matrix $G \triangleq A^T A$ and let

$$\mu(A) \triangleq \max_{1 \leq i, j \leq N, i \neq j} |\langle G_{i,j} \rangle|$$

be the largest off-diagonal element of the Gram matrix. A good goal in designing a sensing matrix is to therefore choose $R$ and $\Psi$ so that $\mu$ is as close as possible to $N^{-1/2}$.

**Theorem:** Noisy Sparse Recovery with Incoherent Matrices. Let $y = A\theta + n$ be a vector of noisy observations of any $K$-sparse signal $\theta \in \mathbb{R}^N$, where $K \leq (\mu(A)^{-1} + 1)/4$ and the $n$ is a noise or error term with $\|n\|_2 \leq \epsilon$. Then our estimate obeys

$$\|\theta - \hat{\theta}\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(A)(4K - 1)}.$$

INCOHERENT MEASUREMENT

Sparse vector

Projection vectors

Signal is locally concentrated, measurements are global

⇓

Each measurement contains a little information about each component
MAGNETIC RESONANCE IMAGING

Space domain

Backprojection

Fourier sampling

CS [Candès, Romberg]
Next time...

- What are the major open problems and areas of research?
- In what ways can these concepts be generalized to other problem domains?