Sparsity, Randomness and Compressed Sensing

Petros Boufounos
Mitsubishi Electric Research Labs
petrosb@merl.com
Sparsity
Why Sparsity

• Natural data and signals exhibit structure
• Sparsity often captures that structure
• Very general signal model
• Computationally tractable
• Wide range of applications in signal acquisition, processing, and transmission
Signal Representations
Signal example: Images

- 2-D function \( f \)
- Idealized view

\[ f \in \text{some function space defined over } [0, 1] \times [0, 1] \]
Signal example: Images

- 2-D function \( f \)
- Idealized view
  \[ f \in \text{some function space defined over } [0, 1] \times [0, 1] \]
- In practice
  \[ f \in \mathbb{R}^{N \times N} \]
  ie: an \( N \times N \) matrix
Signal example: Images

- 2-D function \( f \)

- Idealized view

\[ f \in \text{some function space defined over } [0, 1] \times [0, 1] \]

- In practice

\[ f \in \mathbb{R}^{N \times N} \]

ie: an \( N \times N \) matrix (pixel average)
Signal Models

Classical Model: Signal lies in a linear vector space (e.g. bandlimited functions)

Sparse Model: Signals of interest are often sparse or compressible

\[ \mathbb{R}^3 \]

Image

Bat Sonar
Chirp

Signal

Transform

Wavelet

Gabor/STFT

i.e., very few large coefficients, many close to zero.
Sparse Signal Models

Sparse signals have few non-zero coefficients.

Compressible signals have few significant coefficients. The coefficients decay as a power law.

Compressible ($\ell_p$ ball, $p<1$)
Sparse Approximation
Computational Harmonic Analysis

- **Representation**: 
  \[ f = \sum_k a_k b_k \]
  - coefficients 
  - basis, frame

- **Analysis**: study \( f \) through *structure* of \( \{a_k\} \)
  - \( \{b_k\} \) should *extract features* of interest

- **Approximation**: \( \hat{f}_N \) uses just a few terms \( N \)
  - exploit *sparsity* of \( \{a_k\} \)
Wavelet Transform Sparsity

\[ f = \sum_{k} a_k b_k \]

• Many \( a_k \approx 0 \) (blue)
Sparseness \implies Approximation

\[ f = \sum_k a_k b_k \]

\[ \left| a_{k'} \right| \quad \text{few big} \]

\[ \text{sorted index} \quad k' \quad \text{many small} \]
Linear Approximation

\[ f = \sum_k a_k b_k \]
Linear Approximation

\[ f = \sum_{k} a_k b_k \]

- \textit{N-term approximation}: use "first" \( a_k \)

\[ \tilde{f}_N := \sum_{k=1}^{N} a_k b_k \]
Nonlinear Approximation

\[ f = \sum_k a_k b_k \]

- \( N \)-term approximation:
  use largest \( a_k \) independently

\[ \hat{f}_N := \sum_{k'=1}^{N} a_{k'} b_{k'} \]

- Greedy / thresholding

Few big
Error Approximation Rates

\[ f = \sum_k a_k b_k \]
\[ \hat{f}_N = \sum_{k'=1}^N a_{k'} b_{k'} \]

\[ \| f - \hat{f}_N \|_2^2 < CN^{-\alpha} \quad \text{as} \quad N \to \infty \]

- Optimize asymptotic error decay rate \( \alpha \)
- Nonlinear approximation works better than linear
Compression is Approximation

• Lossy compression of an image creates an approximation

\[ f = \sum_k a_k b_k \]

quantize to \( R \) total bits

\[ \hat{f}_R = \sum_k a_k^q b_k \]
Sparse approximation ≠ Compression

- Sparse approximation chooses coefficients but does *not quantize* or worry about their *locations*.

\[
f = \sum_k a_k b_k
\]

\[
\hat{f}_N = \sum_{k'=1}^N a_{k'} b_{k'}
\]
Location, Location, Location

• Nonlinear approximation selects $\mathcal{N}$ largest $\alpha_k$ to minimize error (easy – threshold)

• Compression algorithm must encode both a set of $\alpha_k$ and their locations (harder)
Exposing Sparsity
Spikes and Sinusoids example

Example Signal Model: Sinusoidal with a few spikes.

DCT Basis:
Spikes and Sinusoids Dictionary

\[ f = D a \]

DCT basis  Impulses

Lost Uniqueness!!
**Strategy:** Improve sparse approximation by constructing a large dictionary.

$$ f = D a $$

How do we design a dictionary?
Dictionary Design

DCT, DFT

Impulse Basis

Wavelets

Edgelets, curvelets, ...

Oversampling Frame

Dictionary \( D \)

Can we just throw in the bucket every\textbf{thing} we know?
Dictionary Design Considerations

• Dictionary Size:
  – *Computation* and *storage* increases with size

• Fast Transforms:
  – FFT, DCT, FWT, etc. dramatically *decrease computation* and *storage*

• Coherence:
  – *Similarity* in elements makes solution *harder*
Two candidate dictionaries:

\[ D_1 \]

\[ D_2 \]

BAD!

**Intuition:** \( D_2 \) has too many **similar** elements. It is very **coherent**.

**Coherence** (similarity) between elements: \( |\langle d_1, d_2 \rangle| \)

**Dictionary coherence:** \( \mu = \max_{i,j} |\langle d_i, d_j \rangle| \)
Incoherent Bases

• “Mix” well the signal components
  – Impulses and Fourier Basis
  – Anything and Random Gaussian
  – Anything and Random 0-1 basis
Computing Sparse Representations
Thresholding

Compute set of coefficients

Zero out small ones

\[ a = D^+ f \]

Computationally efficient
Good for small and very incoherent dictionaries
Matching Pursuit

- **Measure image against dictionary**
  \[ D^T f \]
  \[ \langle d_k f \rangle = \rho_k \]

- **Select largest correlation**
  \[ \rho_k \]

- **Add to representation**
  \[ a_k \leftarrow a_k + \rho_k \]

- **Compute residual**
  \[ f \leftarrow f - \rho_k d_k \]

- **Iterate using residual**
Greedy Pursuits Family

• Several Variations of MP: OMP, StOMP, ROMP, CoSaMP, Tree MP, ...
  (You can create an AndrewMP if you work on it...)

• Some have provable guarantees

• Some improve dictionary search

• Some improve coefficient selection
CoSaMP (Compressive Sampling MP)

Measure image against dictionary

$D^T f = \rho$

$\langle d_k f \rangle = \rho_k$

Select location of largest $2K$ correlations

$\Omega = \text{supp}(\rho|_{2K}) \cup T$

Add to support set

Invert over support

$B = D_{\Omega}^\dagger f$

Truncate and compute residual

$T = \text{supp}(b|_K)$

$a = b|_K$

$r \leftarrow f - Da$

Iterate using residual
Sparse approximation:

Minimize non-zeros in representation
s.t.: representation is close to signal

\[
\min \| a \|_0 \quad \text{s.t.} \quad f \approx Da
\]

Number of non-zeros (sparsity measure)  
Data Fidelity (approximation quality)

Combinatorial complexity.
Very hard problem!
Sparse approximation:

Minimize non-zeros in representation  
s.t.: representation is close to signal

\[
\min \| a \|_0 \quad \text{s.t. } f \approx Da
\]  

Convex Relaxation

\[
\min \| a \|_1 \quad \text{s.t. } f \approx Da
\]

Polynomial complexity.  
Solved using linear programming.
Why $l_1$ relaxation works

$$\min \| a \|_1 \text{ s.t. } f \approx Da$$

$l_1$ “ball”

Sparse solution

$f = Da$
Basis Pursuits

• Have **provable guarantees**
  – Finds **sparsest** solution for **incoherent** dictionaries

• Several variants in formulation:
  BPDN, LASSO, Dantzig selector, ...

• Variations on **fidelity** term and **relaxation** choice

• Several fast algorithms:
  FPC, GPSR, SPGL, ...
Compressed Sensing: Sensing, Sampling and Data Processing
Data Acquisition

• Usual acquisition methods **sample** signals uniformly
  – Time: A/D with microphones, geophones, hydrophones.
  – Space: CCD cameras, sensor arrays.

• Foundation: **Nyquist/Shannon sampling theory**
  – Sample at **twice the signal bandwidth**.
  – Generally a **projection to a complete basis** that spans the signal space.
Data Processing and Transmission

• **Data processing steps:**
  
  – **Sample** Densely
  
  – **Transform** to an informative domain (Fourier, Wavelet)
  
  – **Process/Compress/Transmit**
    
    Sets small coefficients to zero (sparsification)
Sparsity Model

• Signals can usually be **compressed** in some basis

\[ N \text{ pixels} \]

\[ K \ll N \text{ large wavelet coefficients} \]

\[ N \text{ wideband signal samples} \]

\[ K \ll N \text{ large Gabor coefficients} \]

• Sparsity: good **prior** in picking from a lot of candidates
Compressive Sensing Principles

If a signal is \textit{sparse}, do not waste effort sampling the empty space.

Instead, use fewer samples and allow \textit{ambiguity}.

Use the \textit{sparsity} model to reconstruct and \textit{uniquely resolve} the ambiguity.

1-sparse

2-sparse
Measuring Sparse Signals
Compressive Measurements

\[ \text{null}\{\Phi\} + \mathbf{x}_1 \]

\[ \mathbb{R}^3 \]

\[ \mathbf{y} = \Phi \mathbf{x} \]

\[ y_i = \langle \phi_i, \mathbf{x} \rangle \]

Measurement (Projection)

Reconstruction

\( \Phi \) has rank \( M \ll N \)

\[ N = \text{Signal dimensionality} \]

\[ K = \text{Signal sparsity} \]

\[ M = \text{Number of measurements} \]

\( \text{(dimensionality of } \mathbf{y}) \)

\[ N \gg M \geq K \]
One Simple Question

- When is it possible to distinguish K-sparse signals?
  - require $\Phi x_1 \neq \Phi x_2$ for all K-sparse $x_1 \neq x_2$

- **Necessary:** $\Phi$ must have at least $2K$ rows
  - otherwise there exist K-sparse $x_1, x_2$ s.t. $\Phi(x_1 - x_2) = 0$

- **Sufficient:** Gaussian $\Phi$ with $2K$ rows

\[
\begin{align*}
M \times 1 & \quad \Phi \quad x \\
\text{measurements} & \quad = \quad \text{sparse signal} \\
K < M \ll N & \quad \text{2K columns} \\
\end{align*}
\]
Geometry of Sparse Signal Sets

Linear

K-plane

Sparse, Nonlinear

Union of K-planes
Geometry: Embedding in $\mathbb{R}^M$

- $\Phi(K\text{-plane}) = K\text{-plane in general}$
- $M \geq 2K$ measurements
  - necessary for injectivity
  - sufficient for injectivity when $\Phi$ Gaussian
  - but not enough for efficient, robust recovery
- See also FROI [Vetterli et al., Lu and Do]
Illustrative Example

\[ y = \Phi x \]

- \( N = 3 \): signal length
- \( K = 1 \): sparsity
- \( M = 2K = 2 \): measurements
Example: 1-sparse signal

\[ N=3 \]
\[ K=1 \]
\[ M=2K=2 \]

\[ \Phi = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \]

\[ \Phi = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \]

Bad!

Bad!
Example: 1-sparse signal

\[ N=3 \]
\[ K=1 \]
\[ M=2K=2 \]

\[ \Phi = \begin{bmatrix}
1 & 1 & 0 \\
1/2 & 0 & 1 \\
\end{bmatrix} \]

Good!

\[ \Phi = \begin{bmatrix}
1 & -1/2 & -1/2 \\
0 & \sqrt{3}/2 & -\sqrt{3}/2 \\
\end{bmatrix} \]

Better!
Restricted Isometry Property

[Candès, Romberg, Tao]

• Measurement matrix $\Phi$ has **RIP of order $K$** if

\[
(1 - \delta_K) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1 + \delta_K)
\]

for all $K$-sparse signals $x$.

• Does *not* hold for $K > M$; may hold for smaller $K$.

• Implications: tractable, stable, robust recovery
RIP as a “Stable” Embedding

- RIP of order $2K$ implies: for all $K$-sparse $x_1$ and $x_2$,

$$(1 - \delta_{2K}) \leq \frac{||\Phi x_1 - \Phi x_2||_2^2}{||x_1 - x_2||_2^2} \leq (1 + \delta_{2K})$$

(if $\delta_{2K} < 1$ have injectivity; smaller $\delta_{2K}$ more stable)
Verifying RIP

How Many Measurements?

- Want RIP of order 2K (say) to hold for MxN \( \Phi \)
  - difficult to verify for a given \( \Phi \)
  - requires checking eigenvalues of each submatrix

- Prove \textit{random} \( \Phi \) will work
  - iid Gaussian entries
  - iid Bernoulli entries (+/- 1)
  - iid subgaussian entries
  - random Fourier ensemble
  - random subset of incoherent dictionary

- In each case, \( M = O(K \log N) \) suffices
  - with very high probability, usually \( 1 - O(e^{-CN}) \)
  - slight variations on log term
Universality Property

- Gaussian white noise basis is incoherent with any fixed orthonormal basis (with high probability)
- Signal sparse in time domain: $\Phi = I$
Universality Property

- Gaussian white noise basis is incoherent with any fixed orthonormal basis (with high probability)
- Signal sparse in frequency domain: $\Psi = \text{idct}$

- Product $\Phi\Psi$ remains Gaussian white noise
• Measurements are **democratic** [Davenport, Laska, Boufounos, Baraniuk]
  – They are all equally important
  – We can **loose** some *arbitrarily*, (i.e. an adversary can choose which ones)

• The $\tilde{\Phi}$ still satisfies RIP (as long as we don’t drop too many)
Reconstruction
Requirements for Reconstruction

• Let $x_1, x_2$ be $K$-sparse signals (i.e. $x_1 - x_2$ is $2K$-sparse):

• Mapping $y = \Phi x$ is invertible for $K$-sparse signals:

$$\Phi(x_1 - x_2) \neq 0 \text{ if } x_1 \neq x_2$$

• Mapping is robust for $K$-sparse signals:

$$\|\Phi(x_1 - x_2)\|_2 \approx \| x_1 - x_2 \|_2$$

  - Restricted Isometry Property (RIP):
    \[ \Phi \text{ preserves distance when projecting } K\text{-sparse signals} \]

• Guarantees there exists a unique $K$-sparse signal explains the measurements, and is robust to noise.
Reconstruction Ambiguity

• Solution should be **consistent** with measurements

\[ \hat{x} \quad \text{s.t.} \quad y = \Phi \hat{x} \quad \text{or} \quad y \approx \Phi \hat{x} \]

• Projections imply that an **infinite** number of solutions are consistent!

• Classical approach: use the pseudoinverse (minimize $l_2$ norm)

• Compressive sensing approach: pick the **sparsest**.

• **RIP guarantee**: sparsest solution **unique** and **reconstructs the signal**.

Becomes a **sparse approximation problem**!
Putting everything together
Compressed Sensing Coming Together

**Signal Structure** (sparsity)

- **Signal model**: Provides prior information; allows undersampling
- **Randomness**: Provides robustness/stability; makes proofs easier
- **Non-linear reconstruction**: Incorporates information through computation

\[ y = \Phi x \]

**Measurement**

\[ y \approx x \]

**Reconstruction using sparse approximation**

- Basis Pursuit
- Matching Pursuit
- CoSaMP
- etc...

**Stable Embedding** (random projections)

\[ x \rightarrow y \rightarrow \tilde{x} \]
Beyond: Extensions, Connections, Generalizations
Sparsity Models
Block Sparsity

\[
\begin{align*}
M \times 1 & \quad \text{measurements} \\
\Phi & \quad \text{sparse signal} \\
M \times N & \\
K < M \ll N & \quad \text{nonzero blocks of } L \\
\end{align*}
\]

Mixed $l_1/l_2$ norm—sum of $l_2$ norms: $\sum_i \|x_{B_i}\|_2$

Basis pursuit becomes: $\min_x \sum_i \|x_{B_i}\|_2 \text{ s.t. } y \approx \Phi x$

Blocks are not allowed to overlap
Joint Sparsity

Mixed $l_1/l_2$ norm—sum of $l_2$ norms: $\sum_i \|x_{(i,\cdot)}\|_2$

Basis pursuit becomes: $\min_x \sum_i \|x_{(i,\cdot)}\|_2$ s.t. $y \approx \Phi x$
Randomized Embeddings
Stable Embeddings

**Recall: RIP**

- RIP of order $K$ requires: for all $K$-sparse $x$,

\[
(1 - \delta_K) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1 + \delta_K)
\]
Johnson-Lindenstrauss Lemma

[see also Dasgupta, Gupta; Frankl, Maehara; Achlioptas; Indyk, Motwani]

Consider a point set $Q \subset \mathbb{R}^N$ and random* $M \times N$ $\Phi$ with $M = O(\log(\#Q) \epsilon^{-2})$. With high prob., for all $x_1, x_2 \in Q$,

$$(1 - \epsilon) \leq \frac{\|\Phi x_1 - \Phi x_2\|_2^2}{\|x_1 - x_2\|_2^2} \leq (1 + \epsilon).$$

Proof via concentration inequality: For any $x \in \mathbb{R}^N$

$$P\left( \left| \|\Phi x\|_2^2 - \|x\|_2^2 \right| \geq \epsilon \|x\|_2^2 \right) \leq 2e^{-\frac{M}{2}(\epsilon^2/2 - \epsilon^3/3)}.$$
Favorable JL Distributions

- **Gaussian**
  \[
  \phi_{i,j} \sim \mathcal{N}\left(0, \frac{1}{M}\right)
  \]

- **Bernoulli/Rademacher [Achlioptas]**
  \[
  \phi_{i,j} := \begin{cases} 
  +\frac{1}{\sqrt{M}} & \text{with probability } \frac{1}{2}, \\
  -\frac{1}{\sqrt{M}} & \text{with probability } \frac{1}{2}
  \end{cases}
  \]

- **“Database-friendly” [Achlioptas]**
  \[
  \phi_{i,j} := \begin{cases} 
  +\sqrt{\frac{3}{M}} & \text{with probability } \frac{1}{6}, \\
  0 & \text{with probability } \frac{2}{3}, \\
  -\sqrt{\frac{3}{M}} & \text{with probability } \frac{1}{6}
  \end{cases}
  \]

- **Random Orthoprojection to } \mathbb{R}^M [Gupta, Dasgupta]**
Connecting JL to RIP

Consider effect of random JL $\Phi$ on each $K$-plane
- construct covering of points $Q$ on unit sphere
- JL: isometry for each point with high probability
- union bound $\Rightarrow$ isometry for all $q \in Q$
- extend to isometry for all $x$ in $K$-plane

$K$-plane
Connecting JL to RIP

Consider effect of random JL $\Phi$ on each $K$-plane
- construct covering of points $Q$ on unit sphere
- JL: isometry for each point with high probability
- union bound $\Rightarrow$ isometry for all $q \in Q$
- extend to isometry for all $x$ in $K$-plane
- union bound $\Rightarrow$ isometry for all $K$-planes
**Theorem:** Supposing $\Phi$ is drawn from a JL-favorable distribution,* then with probability at least $1-e^{-C^*M}$, $\Phi$ meets the RIP with

$$K \leq C \cdot \frac{M}{\log(N/M) + 1}.$$ 

* Gaussian/Bernoulli/database-friendly/orthoprojector

**Bonus:** *universality* (repeat argument for any $\Psi$)

- See also Mendelson et al. concerning subgaussian ensembles
More?
The tip of the iceberg

Today’s lecture

Compressive Sensing Repository
dsp.rice.edu/cs

Blog on CS
nuit-blancbe.blogspot.com/

Yet to be discovered…
Start working on it 😊