Sparse Representations

Joel A. Tropp

Department of Mathematics
The University of Michigan
jtropp@umich.edu

Research supported in part by NSF and DARPA
Introduction
We consider linear systems of the form

\[
\begin{bmatrix}
\Phi
\end{bmatrix}^d \begin{bmatrix}
x
\end{bmatrix} = \begin{bmatrix}
b
\end{bmatrix}
\]

\(N\)

Assume that

- \(\Phi\) has dimensions \(d \times N\) with \(N \geq d\)
- \(\Phi\) has full rank
- The columns of \(\Phi\) have unit \(\ell_2\) norm
The Trichotomy Theorem

**Theorem 1.** For a linear system $\Phi x = b$, exactly one of the following situations obtains.

1. No solution exists.

2. The equation has a unique solution.

3. The solutions form a linear subspace of positive dimension.
Minimum-Energy Solutions

Classical approach to underdetermined systems:

\[
\min \|x\|_2 \quad \text{subject to} \quad \Phi x = b
\]

Advantages:

☑ Analytically tractable
☑ Physical interpretation as minimum energy
☑ Principled way to pick a unique solution

Disadvantages:

☑ Solution is typically nonzero in every component
☑ The wrong principle for most applications
Another approach to underdetermined systems:

$$\min \|x\|_0 \quad \text{subject to} \quad \Phi x = b$$

where $\|x\|_0 = \#\{j : x_j \neq 0\}$

**Advantages:**

- Principled way to choose a solution
- A good principle for many applications

**Disadvantages:**

- In general, computationally intractable
In practice, we solve a noise-aware variant, such as

$\min \|x\|_0 \quad \text{subject to} \quad \|\Phi x - b\|_2 \leq \varepsilon$

This is called a *sparse approximation problem*

The noiseless problem (P0) corresponds to $\varepsilon = 0$

The $\varepsilon = 0$ case is called the *sparse representation problem*
Applications
Variable Selection in Regression

❖ The oldest application of sparse approximation is linear regression

❖ The columns of $\Phi$ are *explanatory variables*

❖ The right-hand side $b$ is the *response variable*

❖ $\Phi x$ is a *linear predictor* of the response

❖ Want to use few explanatory variables
  ❖ Reduces variance of estimator
  ❖ Limits sensitivity to noise

Reference: [Miller 2002]
"In deconvolving any observed seismic trace, it is rather disappointing to discover that there is a nonzero spike at every point in time regardless of the data sampling rate. One might hope to find spikes only where real geologic discontinuities take place."

References: [Claerbout–Muir 1973]
Transform coding can be viewed as a sparse approximation problem.

Algorithms
Theorem 2. [Davis (1994), Natarajan (1995)] Any algorithm that can solve the sparse representation problem for every matrix and right-hand side must solve an NP-hard problem.
Many interesting instances of the sparse representation problem are tractable!

Basic example: \( \Phi \) is orthogonal
Algorithms for Sparse Representation

- **Greedy methods** make a sequence of locally optimal choices in hope of determining a globally optimal solution.

- **Convex relaxation methods** replace the combinatorial sparse approximation problem with a related convex program in hope that the solutions coincide.

- Other approaches include brute force, nonlinear programming, Bayesian methods, dynamic programming, algebraic techniques...

Refs: [Baraniuk, Barron, Bresler, Candès, DeVore, Donoho, Efron, Fuchs, Gilbert, Golub, Hastie, Huo, Indyk, Jones, Mallat, Muthukrishnan, Rao, Romberg, Stewart, Strauss, Tao, Temlyakov, Tewfik, Tibshirani, Willsky...]

_Sparse Representations_ (Numerical Analysis Seminar, NYU, 20 April 2007)
Orthogonal Matching Pursuit (OMP)

Input: The matrix $\Phi$, right-hand side $b$, and sparsity level $m$

Initialize the residual $r_0 = b$
For $t = 1, \ldots, m$ do

A. Find a column most correlated with the residual:

$$\omega_t = \arg \max_{j=1,\ldots,N} |\langle r_{t-1}, \varphi_j \rangle|$$

B. Update residual by solving a least-squares problem:

$$y_t = \arg \min_y \|b - \Phi_t y\|_2$$

$$r_t = b - \Phi_t y_t$$

where $\Phi_t = [\varphi_{\omega_1} \ldots \varphi_{\omega_t}]$

Output: Estimate $\hat{x}(\omega_j) = y_m(j)$
\( \ell_1 \) Minimization

Sparse Representation as a Combinatorial Problem

\[
\min \| x \|_0 \quad \text{subject to} \quad \Phi x = b \tag{P0}
\]

Relax to a Convex Program

\[
\min \| x \|_1 \quad \text{subject to} \quad \Phi x = b \tag{P1}
\]

Any numerical method can be used to perform the minimization

Projected gradient and interior-point methods seem to work best

References: [Donoho et al. 1999, Figueredo et al. 2007]
Why an $\ell_1$ objective?

\begin{itemize}
  \item $\ell_0$ quasi-norm
  \item $\ell_1$ norm
  \item $\ell_2$ norm
\end{itemize}
Why an $\ell_1$ objective?

\begin{align*}
\ell_0 \text{ quasi-norm} & \quad \ell_1 \text{ norm} & \ell_2 \text{ norm}
\end{align*}
## Relative Merits

<table>
<thead>
<tr>
<th></th>
<th>OMP</th>
<th>(P1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational Cost</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Ease of Implementation</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
When do the algorithms work?
Key Insight

Sparse representation is tractable when the matrix $\Phi$ is sufficiently nice

(More precisely, column submatrices of the matrix should be well conditioned)
We say $\Phi$ is incoherent when

$$\max_{j \neq k} |\langle \varphi_j, \varphi_k \rangle| \leq \frac{1}{\sqrt{d}}$$

Incoherent matrices appear often in signal processing applications.

We call $\Phi$ a tight frame when

$$\Phi \Phi^T = \frac{N}{d} \mathbf{I}$$

Tight frames have minimal spectral norm among conformal matrices.

Note: Both conditions can be relaxed substantially.
Example: Identity + Fourier

Impulses

Complex Exponentials

An incoherent tight frame
Finding Sparse Solutions

**Theorem 3. [T 2004]** Let $\Phi$ be incoherent. Suppose that the linear system $\Phi x = b$ has a solution $x_\star$ that satisfies

$$
\|x_\star\|_0 < \frac{1}{2} (\sqrt{d} + 1).
$$

Then the vector $x_\star$ is

1. the unique minimal $\ell_0$ solution to the linear system, and

2. the output of both OMP and $\ell_1$ minimization.

References: [Donoho–Huo 2001, *Greed is Good, Just Relax*]
Sparse representations are not necessarily unique past the $\sqrt{d}$ threshold

**Example:** The Dirac Comb

Consider the Identity + Fourier matrix with $d = p^2$

There is a vector $b$ that can be written as either $p$ spikes or $p$ sines

By the Poisson summation formula,

$$b(t) = \sum_{j=0}^{p-1} \delta_{pj}(t) = \frac{1}{\sqrt{d}} \sum_{j=0}^{p-1} e^{-2\pi i pt/d} \quad \text{for } t = 0, 1, \ldots, d$$
Insight:
The bad vectors are atypical

- It is usually possible to identify *random* sparse vectors

- The next theorem is the first step toward quantifying this intuition
Theorem 4. [T 2006] Let $\Phi$ be an incoherent tight frame with at least twice as many columns as rows. Suppose that

$$m \leq \frac{cd}{\log d}.$$  

If $A$ is a random $m$-column submatrix of $\Phi$ then

$$\Pr \left\{ \|A^* A - I\| < \frac{1}{2} \right\} \geq 99.44\%.$$  

The number $c$ is a positive absolute constant.

Reference: [Random Subdictionaries]
Recovering Random Sparse Vectors

<table>
<thead>
<tr>
<th>Model (M) for $b = \Phi x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>The matrix $\Phi$ is an incoherent tight frame</td>
</tr>
<tr>
<td>Nonzero entries of $x$ number $m \leq cd / \log N$</td>
</tr>
<tr>
<td>have uniformly random positions</td>
</tr>
<tr>
<td>are independent, zero-mean Gaussian RVs</td>
</tr>
</tbody>
</table>

**Theorem 5. [T 2006]** Let $b = \Phi x$ be a random vector drawn according to Model (M). Then $x$ is

1. the unique minimal $\ell_0$ solution w.p. at least 99.44% and
2. the unique minimal $\ell_1$ solution w.p. at least 99.44%.

Reference: [Random Subdictionaries]
Methods of Proof

- Functional success criterion for OMP
- Duality results for convex optimization
- Banach algebra techniques for estimating matrix norms
- Concentration of measure inequalities
- Banach space methods for studying spectra of random matrices
  - Decoupling of dependent random variables
  - Symmetrization of random subset sums
  - Noncommutative Khintchine inequalities
  - Bounds for suprema of empirical processes
Compressive Sampling
In many applications, signals of interest have sparse representations.

Traditional methods acquire entire signal, then extract information.

Sparsity can be exploited when acquiring these signals.

Want number of samples proportional to amount of information.

Approach: Introduce randomness in the sampling procedure.

Assumption: Each random sample has unit cost.
Given data $b = \Phi x$, must identify sparse signal $x$

This is a sparse representation problem with a random matrix

References: [Candès–Romberg–Tao 2004, Donoho 2004]
**Compressive Sampling and OMP**

**Theorem 6. [T, Gilbert 2005]** Assume that

- $x$ is a vector in $\mathbb{R}^N$ with $m$ nonzeros and
- $\Phi$ is a $d \times N$ Gaussian matrix with $d \geq Cm \log N$

- Execute OMP with $b = \Phi x$ to obtain estimate $\hat{x}$

The estimate $\hat{x}$ equals the vector $x$ with probability at least 99.44%.

Reference: [Signal Recovery via OMP]
Theorem 7. [Various] Assume that

- $\Phi$ is a $d \times N$ Gaussian matrix with $d \geq Cm \log(N/m)$

With probability 99.44%, the following statement holds.

- Let $x$ be a vector in $\mathbb{R}^N$ with $m$ nonzeros
- Execute $\ell_1$ minimization with $b = \Phi x$ to obtain estimate $\hat{x}$

The estimate $\hat{x}$ equals the vector $x$.

Related Directions
There are algorithms that can recover sparse signals from random measurements in time proportional to the number of measurements.

This is an *exponential* speedup over OMP and $\ell_1$ minimization.

The cost is a *logarithmic* number of additional measurements.

References: [Algorithmic dimension reduction, One sketch for all]

Joint with Gilbert, Strauss, Vershynin.
Simultaneous Sparsity

In some applications, one seeks solutions to the matrix equation

$$\Phi X = B$$

where $X$ has a minimal number of *nonzero rows*

We have studied algorithms for this problem

References: [Simultaneous Sparse Approximation I and II]
Joint with Gilbert, Strauss
The coherence statistic plays an important role in sparse representation

What can we say about matrices $\Phi$ with minimal coherence?

Equivalent to studying packing in projective space

We have theory about when optimal packings can exist

We have numerical algorithms for constructing packings

References: [Existence of ETFs, Constructing Structured TFs, ...]

Joint with Dhillon, Heath, Sra, Strohmer, Sustik
To learn more...

Web: http://www.umich.edu/~jtropp
E-mail: jtropp@umich.edu

Partial List of Papers

- “Greed is good,” *Trans. IT*, 2004
- “Constructing structured tight frames,” *Trans. IT*, 2005
- “Just relax,” *Trans. IT*, 2006
- “One sketch for all,” to appear, STOC 2007
- “Existence of equiangular tight frames,” submitted, 2004
- “Signal recovery from random measurements via OMP,” submitted, 2005
- “Algorithmic dimension reduction,” submitted, 2006
- “Random subdictionaries,” submitted, 2006
- “Constructing packings in Grassmannian manifolds,” submitted, 2006

Coauthors: Dhillon, Gilbert, Heath, Muthukrishnan, Rice DSP, Sra, Strauss, Strohmer, Sustik, Vershynin