Sparse Representations

Joel A. Tropp

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

Research supported in part by NSF and DARPA

Introduction

Systems of Linear Equations

We consider linear systems of the form

$$d \left\{ \underbrace{\begin{bmatrix} & & \\ & \boldsymbol{\Phi} & \\ & & \end{bmatrix}}_{N} \right. \begin{bmatrix} \\ \\ \boldsymbol{x} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \boldsymbol{b} \\ \\ \end{bmatrix}$$

Assume that

- $\boldsymbol{\Phi}$ has dimensions $d \times N$ with $N \geq d$
- $\boldsymbol{\Phi}$ has full rank
- The columns of $\boldsymbol{\Phi}$ have unit ℓ_2 norm

The Trichotomy Theorem

Theorem 1. *For a linear system $\boldsymbol{\Phi x} = \boldsymbol{b}$, 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:

$$\min \|\boldsymbol{x}\|_2 \quad \text{subject to} \quad \boldsymbol{\Phi x} = \boldsymbol{b}$$

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

Regularization via Sparsity

Another approach to underdetermined systems:

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

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

Advantages:

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

Disadvantages:

- In general, computationally intractable

Sparse Approximation

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

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

- 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 $\mathbf{\Phi}$ are *explanatory variables*
- The right-hand side $\boldsymbol{b}$ is the *response variable*
- $\mathbf{\Phi}\boldsymbol{x}$ is a *linear predictor* of the response
- Want to use few explanatory variables
  - Reduces variance of estimator
  - Limits sensitivity to noise

Reference: [Miller 2002]

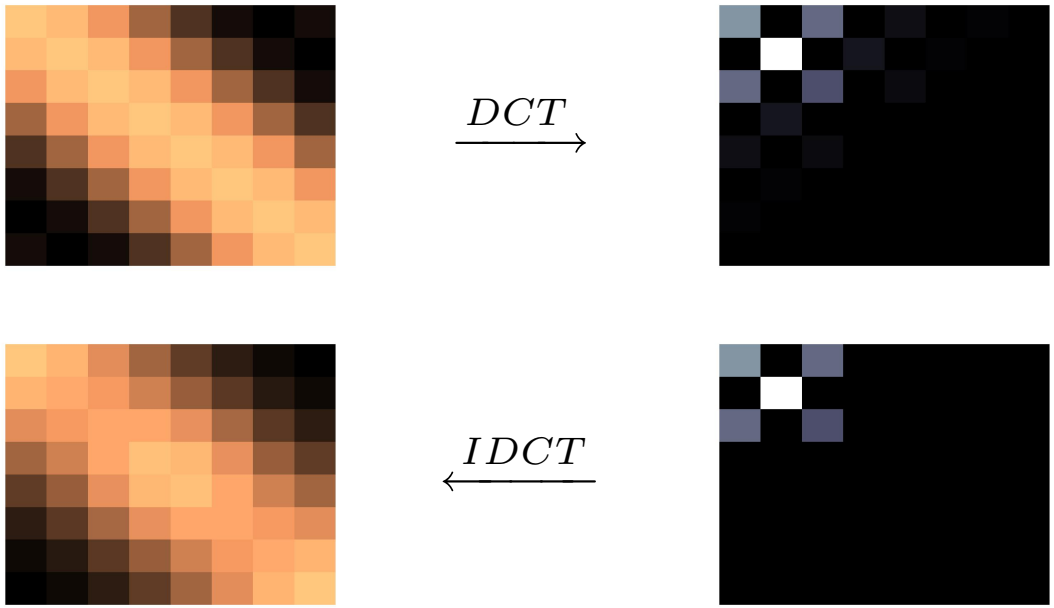
Seismic Imaging

"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

- Transform coding can be viewed as a sparse approximation problem



Reference: [Daubechies–DeVore–Donoho–Vetterli 1998]

Algorithms

Sparse Representation is Hard

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.*

But...

Many interesting instances
of the sparse representation problem
are tractable!

Basic example: $\mathbf{\Phi}$ 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...]

Orthogonal Matching Pursuit (OMP)

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

Initialize the residual $\boldsymbol{r}_0 = \boldsymbol{b}$
For $t = 1, \ldots, m$ do

A. Find a column most correlated with the residual:

$$\omega_t \;=\; \arg\max_{j=1,\ldots,N} \; |\langle \boldsymbol{r}_{t-1},\; \boldsymbol{\varphi}_j \rangle|$$

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

$$\boldsymbol{y}_t = \arg\min_{\boldsymbol{y}} \|\boldsymbol{b} - \boldsymbol{\Phi}_t\, \boldsymbol{y}\|_2$$

$$\boldsymbol{r}_t = \boldsymbol{b} - \boldsymbol{\Phi}_t\, \boldsymbol{y}_t$$

where $\boldsymbol{\Phi}_t = [\boldsymbol{\varphi}_{\omega_1} \;\; \ldots \;\; \boldsymbol{\varphi}_{\omega_t}]$

Output: Estimate $\widehat{\boldsymbol{x}}(\omega_j) = \boldsymbol{y}_m(j)$

ℓ_1 Minimization

Sparse Representation as a Combinatorial Problem

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

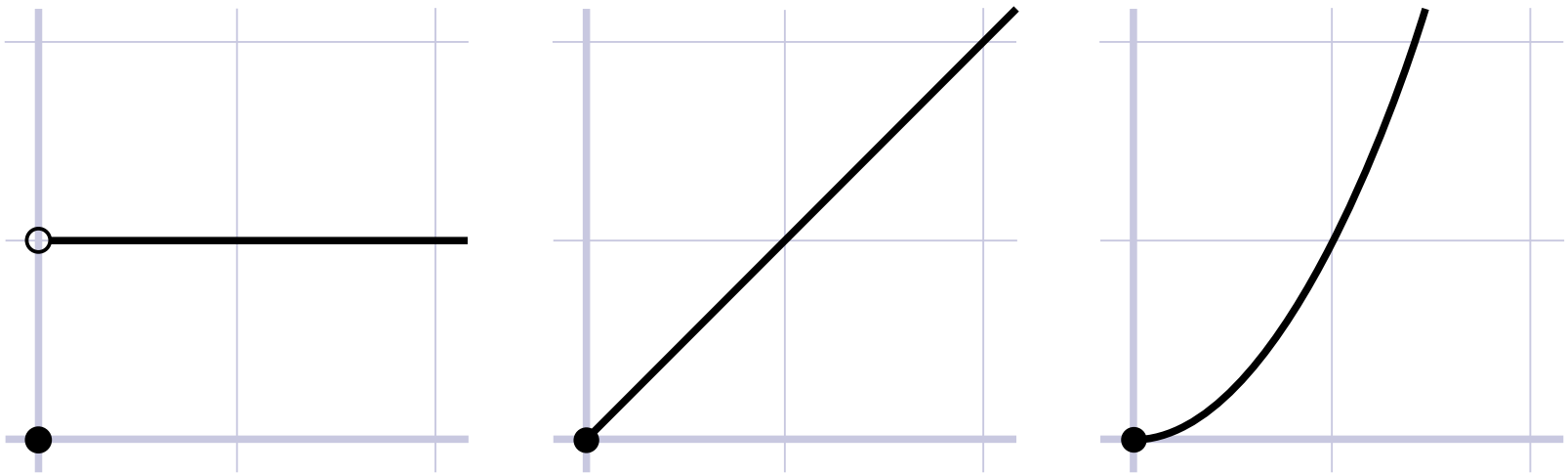
Relax to a Convex Program

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

- Any numerical method can be used to perform the minimization
- Projected gradient and interior-point methods seem to work best

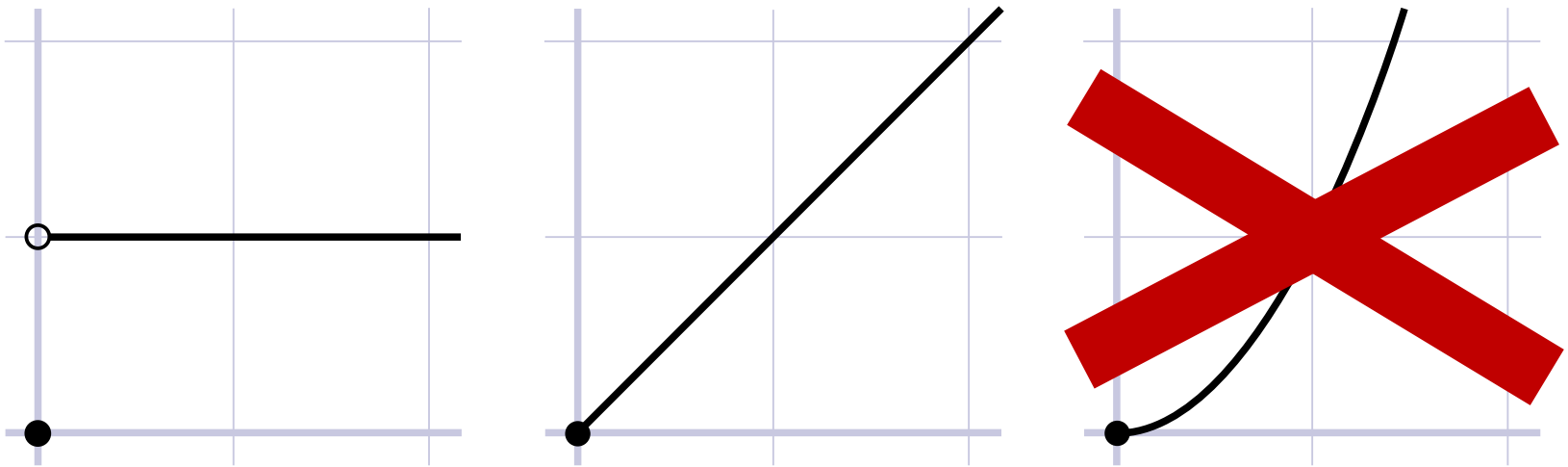
References: [Donoho et al. 1999, Figueredo et al. 2007]

Why an ℓ_1 objective?



ℓ_0 quasi-norm $\qquad$ ℓ_1 norm $\qquad$ ℓ_2 norm

Why an ℓ_1 objective?



ℓ_0 quasi-norm $\qquad$ ℓ_1 norm $\qquad$ ℓ_2 norm

Relative Merits

	OMP	(P1)
Computational Cost	✓	X
Ease of Implementation	✓	X
Effectiveness	X	✓

When do the algorithms work?

Key Insight

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

(More precisely, column submatrices
of the matrix should be well conditioned)

Quantifying Niceness

- We say $\mathbf{\Phi}$ is *incoherent* when

$$\max_{j \neq k} |\langle \boldsymbol{\varphi}_j, \ \boldsymbol{\varphi}_k \rangle| \quad \leq \quad \frac{1}{\sqrt{d}}$$

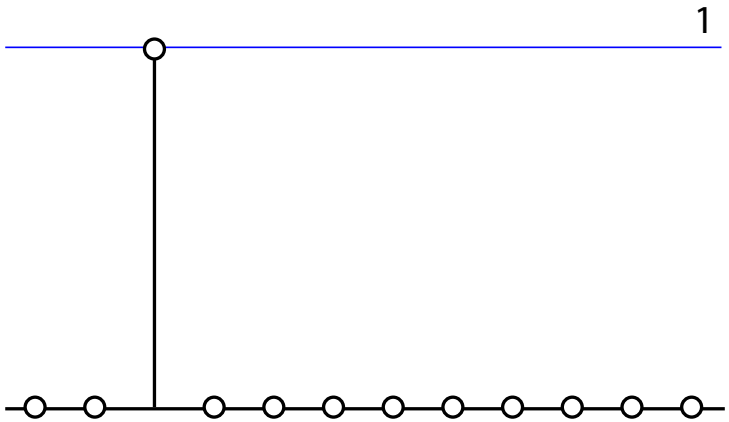
- Incoherent matrices appear often in signal processing applications
- We call $\mathbf{\Phi}$ a *tight frame* when

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

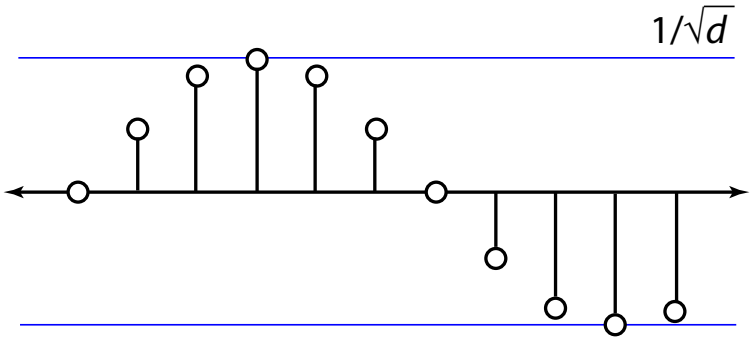
- Tight frames have minimal spectral norm among conformal matrices

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 $\boldsymbol{\Phi}$ be incoherent. Suppose that the linear system $\boldsymbol{\Phi x} = \boldsymbol{b}$ has a solution $\boldsymbol{x}_\star$ that satisfies*

$$\|\boldsymbol{x}_\star\|_0 \;<\; \tfrac{1}{2}(\sqrt{d}+1).$$

Then the vector $\boldsymbol{x}_\star$ is

1. *the unique minimal ℓ_0 solution to the linear system, and*

2. *the output of both OMP and ℓ_1 minimization.*

References: [Donoho–Huo 2001, *Greed is Good*, *Just Relax*]

The Square-Root Threshold

- 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 $\boldsymbol{b}$ that can be written as either p spikes or p sines
- By the Poisson summation formula,

$$\boldsymbol{b}(t) = \sum_{j=0}^{p-1} \delta_{pj}(t) = \frac{1}{\sqrt{d}} \sum_{j=0}^{p-1} \mathrm{e}^{-2\pi \mathrm{i} pjt/d} \qquad \text{for } t = 0, 1, \ldots, d$$

Enter Probability

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

Conditioning of Random Submatrices

Theorem 4. [T 2006] *Let $\mathbf{\Phi}$ be an incoherent tight frame with at least twice as many columns as rows. Suppose that*

$$m \;\leq\; \frac{\mathrm{c}d}{\log d}.$$

If $\boldsymbol{A}$ is a random m-column submatrix of $\mathbf{\Phi}$ then

$$\mathrm{Prob}\left\{ \|\boldsymbol{A}^*\boldsymbol{A} - \mathbf{I}\| < \frac{1}{2} \right\} \;\geq\; 99.44\%.$$

The number c is a positive absolute constant.

Reference: [*Random Subdictionaries*]

Recovering Random Sparse Vectors

Model (M) for $\boldsymbol{b} = \boldsymbol{\Phi x}$	
The matrix $\boldsymbol{\Phi}$	is an incoherent tight frame
Nonzero entries of $\boldsymbol{x}$	number $m \leq \mathrm{c}d/\log N$
	have uniformly random positions
	are independent, zero-mean Gaussian RVs

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

1. *the unique minimal ℓ_0 solution w.p. at least 99.44% and*
2. *the unique minimal ℓ_1 solution w.p. at least 99.44%.*

Reference: [*Random Subdictionaries*]

Methods of Proof

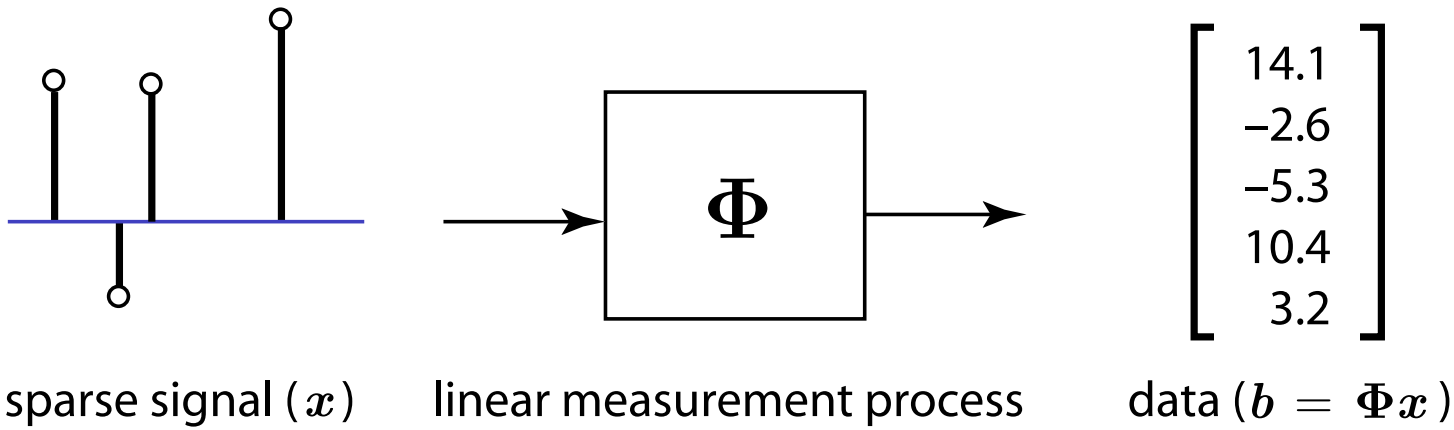
- Functional success criterion for OMP
- Duality results for convex optimization
- Banach algebra techniques for estimating matrix norms
- Concentration of measure inequalities
- Banach space methods for studying spectra of random matrices
  - Decoupling of dependent random variables
  - Symmetrization of random subset sums
  - Noncommutative Khintchine inequalities
  - Bounds for suprema of empirical processes

Compressive Sampling

Compressive Sampling I

- 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

Compressive Sampling II



- Given data $\boldsymbol{b} = \boldsymbol{\Phi x}$, must identify sparse signal $\boldsymbol{x}$
- 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*

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

- *Execute OMP with* $\boldsymbol{b} = \boldsymbol{\Phi}\boldsymbol{x}$ *to obtain estimate* $\widehat{\boldsymbol{x}}$

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

Reference: [*Signal Recovery via OMP*]

Compressive Sampling with ℓ_1 Minimization

Theorem 7. [Various] *Assume that*

- $\boldsymbol{\Phi}$ *is a* $d \times N$ *Gaussian matrix with* $d \geq \mathrm{C} m \log(N/m)$

With probability 99.44%, the following statement holds.

- *Let* $\boldsymbol{x}$ *be a vector in* $\mathbb{R}^N$ *with* m *nonzeros*
- *Execute* ℓ_1 *minimization with* $\boldsymbol{b} = \boldsymbol{\Phi}\boldsymbol{x}$ *to obtain estimate* $\widehat{\boldsymbol{x}}$

The estimate $\widehat{\boldsymbol{x}}$ *equals the vector* $\boldsymbol{x}$.

References: [Candès et al. 2004–2006], [Donoho et al. 2004–2006], [Rudelson–Vershynin 2006]

Related Directions

Sublinear Compressive Sampling

- 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 ℓ_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

$$\boldsymbol{\Phi X} = \boldsymbol{B}$$

  where $\boldsymbol{X}$ 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

Projective Packings

- The coherence statistic plays an important role in sparse representation
- What can we say about matrices $\mathbf{\Phi}$ 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
- "Simultaneous sparse approximation I and II," *J. Signal Process.*, 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