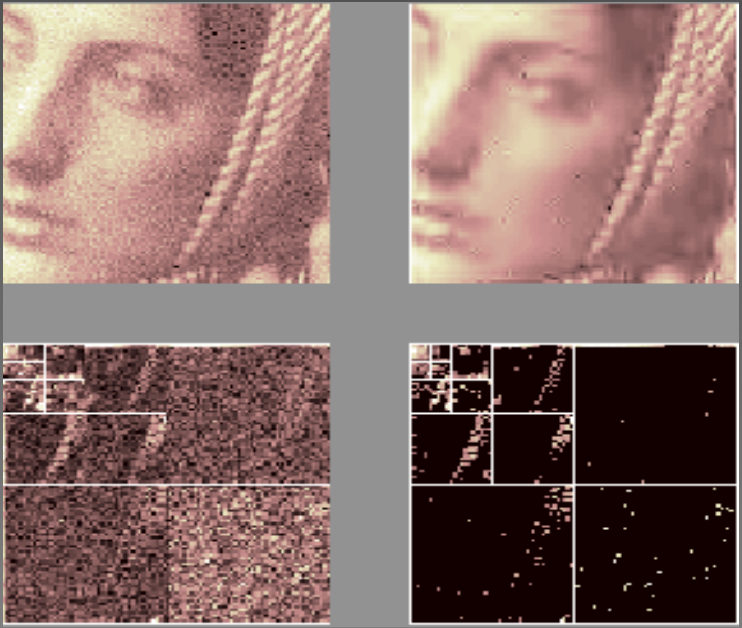
An Introduction to Sparse Approximation

Anna C. Gilbert

Department of Mathematics
University of Michigan

Basic image/signal/data compression: transform coding



Approximate signals sparsely

Compress images, signals, data

accurately (mathematics)

concisely (statistics)

efficiently (algorithms)

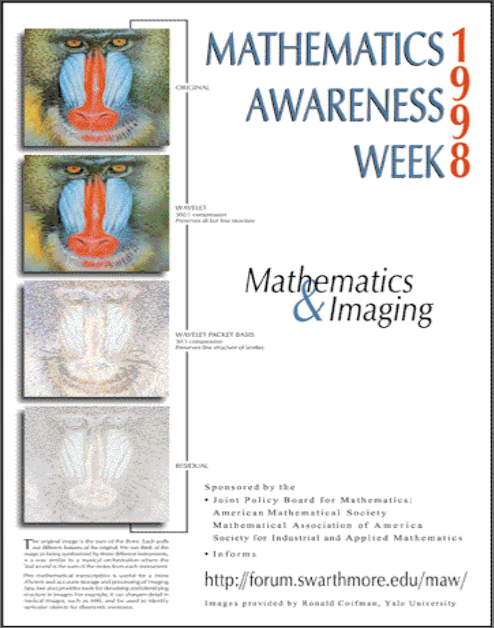
Redundancy

If one orthonormal basis is good, surely two (or more) are better...

Redundancy

If one orthonormal basis is good, surely two (or more) are better...

...especially for images


MATHEMATICS AWARENESS WEEK 1998
ORIGINAL
WAVELET
WAVELET PACKET BASIS
RESIDUAL
Mathematics & Imaging
Sponsored by the
• Joint Policy Board for Mathematics:
American Mathematical Society
Mathematical Association of America
Society for Industrial and Applied Mathematics
• Informs
http://forum.swarthmore.edu/maw/
Images provided by Ronald Coifman, Yale University

Dictionary

Definition

A dictionary *D in $\mathbb{R}^n$ is a collection $\{\varphi_\ell\}_{\ell=1}^{d} \subset \mathbb{R}^n$ of unit-norm vectors: $\|\varphi_\ell\|_2 = 1$ for all ℓ.*

- Elements are called *atoms*
- If $\text{span}\{\varphi_\ell\} = \mathbb{R}^n$, the dictionary is *complete*
- If $\{\varphi_\ell\}$ are linearly dependent, the dictionary is *redundant*

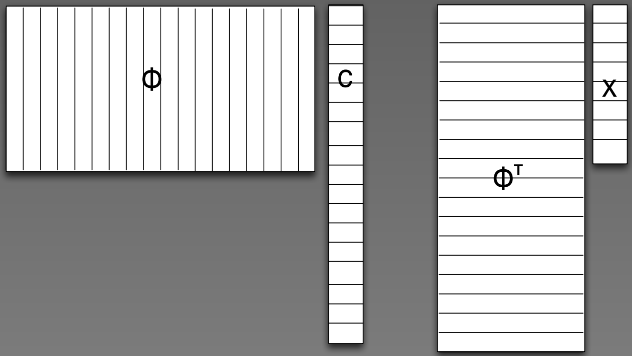
Matrix representation

Form a matrix

$$\Phi = \begin{bmatrix} \varphi_1 & \varphi_2 & \dots & \varphi_d \end{bmatrix}$$

so that

$$\Phi c = \sum_{\ell} c_\ell \varphi_\ell.$$

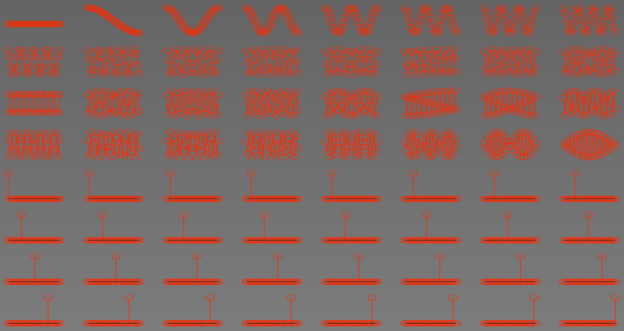


Examples: Fourier—Dirac

$\Phi = [\,\mathcal{F}\,|\,I\,]$

$$\varphi_\ell(t) = \frac{1}{\sqrt{n}} \mathrm{e}^{2\pi \mathrm{i}\ell t/n} \quad \ell = 1, 2, \ldots, n$$

$$\varphi_\ell(t) = \delta_\ell(t) \quad \ell = n+1, n+2, \ldots, 2n$$



SPARSE Problems

EXACT. Given a vector $x \in \mathbb{R}^n$ and a complete dictionary Φ, solve

$$\min_c \|c\|_0 \quad \text{s.t.} \quad x = \Phi c$$

i.e., find a sparsest representation of x over Φ.

ERROR. Given $\epsilon \geq 0$, solve

$$\min_c \|c\|_0 \quad \text{s.t.} \quad \|x - \Phi c\|_2 \leq \epsilon$$

i.e., find a sparsest approximation of x that achieves error ϵ.

SPARSE. Given $k \geq 1$, solve

$$\min_c \|x - \Phi c\|_2 \quad \text{s.t.} \quad \|c\|_0 \leq k$$

i.e., find the best approximation of x using k atoms.

NP-hardness

Theorem

Given an arbitrary redundant dictionary Φ and a signal x, it is NP-hard to solve the sparse representation problem $\textsc{Sparse}$.

[Natarajan'95,Davis'97]

Corollary

$\textsc{Error}$ *is NP-hard as well.*

Corollary

It is NP-hard to determine if the optimal error is zero for a given sparsity level k.

Exact Cover by 3-sets: X3C

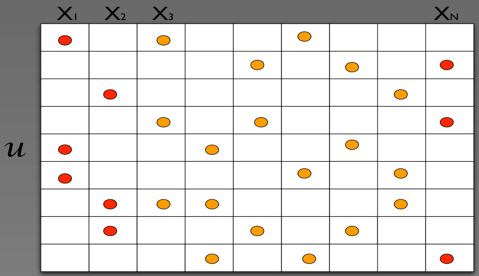
Definition

Given a finite universe $\mathcal{U}$, a collection $\mathcal{X}$ of subsets $X_1, X_2, \ldots, X_d$ s.t. $|X_i| = 3$ for each i, does $\mathcal{X}$ contain a disjoint collection of subsets whose union $= \mathcal{U}$?

Classic NP-hard problem.

Proposition

Any instance of X3C is reducible in polynomial time to SPARSE.



Bad news, Good news

Bad news

Given any polynomial time algorithm for SPARSE, there is a dictionary Φ and a signal x such that algorithm returns incorrect answer

Pessimistic: worst case

Cannot hope to approximate solution, either

Bad news, Good news

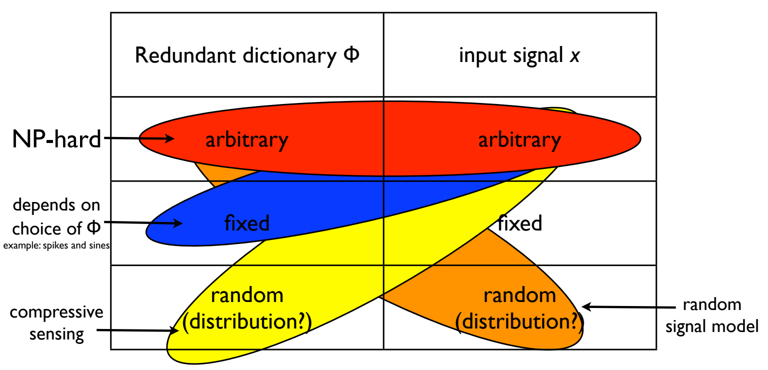
Bad news

- Given any polynomial time algorithm for SPARSE, there is a dictionary Φ and a signal x such that algorithm returns incorrect answer
- Pessimistic: worst case
- Cannot hope to approximate solution, either

Good news

- Natural dictionaries are far from arbitrary
- Perhaps natural dictionaries admit polynomial time algorithms
- Optimistic: rarely see worst case
- Leverage our intuition from orthogonal basis

Hardness depends on instance



Sparse algorithms: exploit geometry

Orthogonal case: pull off atoms one at a time, with dot products in decreasing magnitude

Sparse algorithms: exploit geometry

Orthogonal case: pull off atoms one at a time, with dot products in decreasing magnitude

Why is orthogonal case easy?

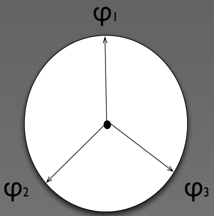
- inner products between atoms are small
- it's easy to tell which one is the best choice

When atoms are (nearly) parallel, can't tell which one is best

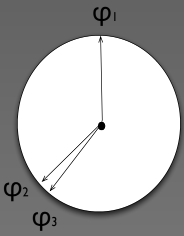
Coherence

Definition
The coherence of a dictionary

$$\mu = \max_{j \neq \ell} |\langle \varphi_j,\ \varphi_\ell \rangle|$$



Small coherence
(good)



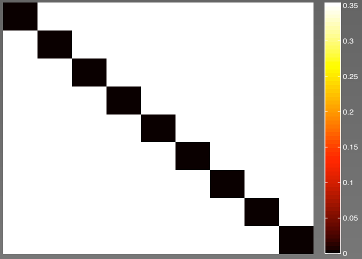
Large coherence
(bad)

Large, incoherent dictionaries

Fourier–Dirac, $d = 2n$, $\mu = \frac{1}{\sqrt{n}}$

wavelet packets, $d = n \log n$, $\mu = \frac{1}{\sqrt{2}}$

There are large dictionaries with coherence close to the lower (Welch) bound; e.g., Kerdock codes, $d = n^2$, $\mu = 1/\sqrt{n}$



Greedy algorithms

Build approximation one step at a time...

...choose the best atom at each step

Orthogonal Matching Pursuit OMP [Mallat'93, Davis'97]

Input. Dictionary Φ, signal x, steps k

Output. Coefficient vector c with k nonzeros, $\Phi c \approx x$

Initialize. counter $t = 1$, $c = 0$

1. **Greedy selection.**

$$\ell_t = \mathrm{argmax} |\Phi^*(x - \Phi c)|$$

2. **Update.** Find $c_{\ell_1}, \ldots, c_{\ell_t}$ to solve

$$\min \left\| x - \sum_s c_{\ell_s} \varphi_{\ell_s} \right\|_2$$

new approximation $a_t \longleftarrow \Phi c$

3. **Iterate.** $t \longleftarrow t + 1$, stop when $t > k$.

Many greedy algorithms with similar outline

Matching Pursuit: replace step 2. by
$c_{\ell_t} \longleftarrow c_{\ell_t} + \langle x - \Phi c,\ \varphi_{\ell_t} \rangle$

Thresholding
Choose m atoms where $|\langle x,\ \varphi_\ell \rangle|$ are among m largest

Alternate stopping rules:
$\|x - \Phi c\|_2 \leq \epsilon$
$\max_\ell |\langle x - \Phi c,\ \varphi_\ell \rangle| \leq \epsilon$

Many other variations

Convergence of OMP

Theorem

Suppose Φ *is a complete dictionary for* $\mathbb{R}^n$*. For any vector* x*, the residual after* t *steps of* OMP *satisfies*

$$\|x - \Phi c\|_2 \leq \frac{C}{\sqrt{t}}.$$

[DEVORE-TEMLYAKOV]

Convergence of OMP

Theorem

Suppose Φ *is a complete dictionary for* $\mathbb{R}^n$*. For any vector* x*, the residual after* t *steps of* OMP *satisfies*

$$\|x - \Phi c\|_2 \leq \frac{C}{\sqrt{t}}.$$

[DeVore-Temlyakov]

Even if x can be expressed sparsely, OMP may take n steps before the residual is zero.

But, sometimes OMP correctly identifies sparse representations.

Exact Recovery Condition and coherence

Theorem (ERC)

A sufficient condition for OMP *to identify* Λ *after* k *steps is that*

$$\max_{\ell \notin \Lambda} \left\| \Phi_\Lambda^+ \varphi_\ell \right\|_1 < 1$$

where $A^+ = (A^* A)^{-1} A^*$. *[Tropp'04]*

Theorem

The ERC holds whenever $k < \frac{1}{2}(\mu^{-1} + 1)$. *Therefore,* OMP *can recover any sufficiently sparse signals.* *[Tropp'04]*

For most redundant dictionaries, $k < \frac{1}{2}(\sqrt{n} + 1)$.

Sparse representation with OMP

Suppose x has k-sparse representation

$$x = \sum_{\ell \in \Lambda} b_\ell \varphi_\ell \quad \text{where } |\Lambda| = k$$

Sufficient to find Λ—When can OMP do so?

Define

$$\Phi_\Lambda = \begin{bmatrix} \varphi_{\ell_1} & \varphi_{\ell_2} & \cdots & \varphi_{\ell_k} \end{bmatrix}_{\ell_s \in \Lambda} \quad \text{and}$$
$$\Psi_\Lambda = \begin{bmatrix} \varphi_{\ell_1} & \varphi_{\ell_2} & \cdots & \varphi_{\ell_{N-k}} \end{bmatrix}_{\ell_s \notin \Lambda}$$

Define *greedy selection ratio*

$$\rho(r) = \frac{\max_{\ell \notin \Lambda} |\langle r,\ \varphi_\ell \rangle|}{\max_{\ell \in \Lambda} |\langle r,\ \varphi_\ell \rangle|} = \frac{\left\|\Psi_\Lambda^* r\right\|_\infty}{\left\|\Phi_\Lambda^* r\right\|_\infty} = \frac{\text{max i.p. bad atoms}}{\text{max i.p. good atoms}}$$

OMP chooses good atom iff $\rho(r) < 1$

Sparse

Theorem
Assume $k \leq \frac{1}{3\mu}$. For any vector x, the approximation $\widehat{x}$ after k steps of OMP satisfies

$$\|x - \widehat{x}\|_2 \leq \sqrt{1 + 6k}\, \|x - x_k\|_2$$

where x_k is the best k-term approximation to x. [Tropp'04]

Theorem
Assume $4 \leq k \leq \frac{1}{\sqrt{\mu}}$. Two-phase greedy pursuit produces $\widehat{x}$ s.t.

$$\|x - \widehat{x}\|_2 \leq 3\, \|x - x_k\|_2 .$$

Assume $k \leq \frac{1}{\mu}$. Two-phase greedy pursuit produces $\widehat{x}$ s.t.

$$\|x - \widehat{x}\|_2 \leq \left(1 + \frac{2\mu k^2}{(1 - 2\mu k)^2}\right) \|x - x_k\|_2 .$$

[Gilbert, Strauss, Muthukrishnan, Tropp'03]

Alternative algorithmic approach

EXACT: non-convex optimization

$$\min \|c\|_0 \quad \text{s.t.} \quad x = \Phi c$$

Alternative algorithmic approach

EXACT: non-convex optimization

$$\min \|c\|_0 \quad \text{s.t.} \quad x = \Phi c$$

Convex relaxation of non-convex problem

$$\min \|c\|_1 \quad \text{s.t.} \quad x = \Phi c$$

ERROR: non-convex optimization

$$\arg\min \|c\|_0 \quad \text{s.t.} \quad \|x - \Phi c\|_2 \leq \epsilon$$

Alternative algorithmic approach

EXACT: non-convex optimization

$$\min \|c\|_0 \quad \text{s.t.} \quad x = \Phi c$$

Convex relaxation of non-convex problem

$$\min \|c\|_1 \quad \text{s.t.} \quad x = \Phi c$$

ERROR: non-convex optimization

$$\arg\min \|c\|_0 \quad \text{s.t.} \quad \|x - \Phi c\|_2 \leq \epsilon$$

Convex relaxation of non-convex problem

$$\arg\min \|c\|_1 \quad \text{s.t.} \quad \|x - \Phi c\|_2 \leq \delta.$$

Convex relaxation: algorithmic formulation

Well-studied algorithmic formulation [Donoho, Donoho-Elad-Temlyakov, Tropp, and many others]

Optimization problem = linear program: linear objective function (with variables c^+, c^-) and linear or quadratic constraints

Still need *algorithm* for solving optimization problem

Hard part of analysis: showing solution to convex problem = solution to original problem

Exact Recovery Condition

Theorem (ERC)

A sufficient condition for BP *to recover the sparsest representation of x is that*

$$\max_{\ell \notin \Lambda} \left\| \Phi_\Lambda^+ \varphi_\ell \right\|_1 < 1$$

where $A^+ = (A^T A)^{-1} A^T$. [Tropp'04]

Exact Recovery Condition

Theorem (ERC)

A sufficient condition for BP *to recover the sparsest representation of* x *is that*

$$\max_{\ell \notin \Lambda} \left\| \Phi_\Lambda^+ \varphi_\ell \right\|_1 < 1$$

where $A^+ = (A^T A)^{-1} A^T$. [Tropp'04]

Theorem

The ERC holds whenever $k < \frac{1}{2}(\mu^{-1} + 1)$*. Therefore,* BP *can recover any sufficiently sparse signals.* [Tropp'04]

Alternate optimization formulations

Constrained minimization:

$$\arg\min \|c\|_1 \quad \text{s.t.} \quad \|x - \Phi c\|_2 \le \delta.$$

Unconstrained minimization:

$$\text{minimize } L(c; \gamma, x) = \frac{1}{2}\|x - \Phi c\|_2^2 + \gamma \|c\|_1 .$$

Many algorithms for ℓ_1-regularization

Sparse approximation: Optimization vs. Greedy

EXACT and ERROR amenable to convex relaxation and convex optimization

SPARSE not amenable to convex relaxation

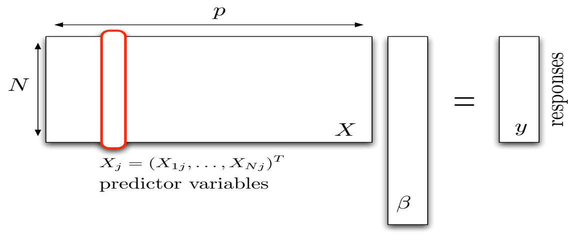
$$\arg\min \|\Phi c - x\|_2 \quad \text{s.t.} \quad \|c\|_0 \leq k$$

but appropriate for greedy algorithms

Connection between...

Sparse Approximation and Statistical Learning

Sparsity in statistical learning



Goal: Given X and y, find α and coeffs. $\beta \in \mathbb{R}^p$ for linear model that minimizes the error

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \|X\beta - (y - \alpha)\|_2^2.$$

Solution:

Least squares: low bias but large variance and hard to interpret lots of non-zero coefficients
Shrink β_j's, make β sparse.

Algorithms in statistical learning

Brute force: Calculate Mallows' C_p for *every* subset of predictor variables, and choose the best one.

Greedy algorithms: Forward selection, forward stagewise, least angle regression (LARS), backward elimination.

Constrained optimization: Quadratic programming problem with linear constraints (e.g., LASSO).

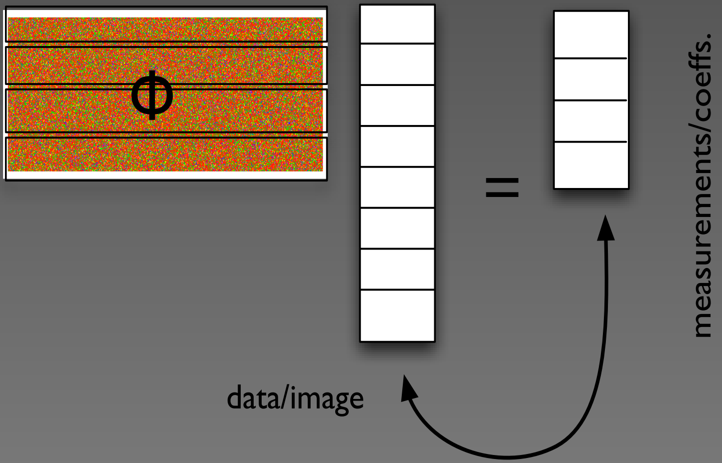
Unconstrained optimization: regularization techniques

Sparse approximation and SVM equivalence [Girosi '96]

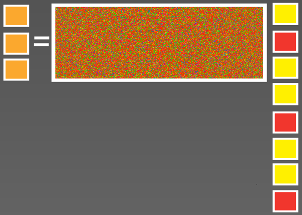
Connection between...

Sparse Approximation and Compressed Sensing

Interchange roles



Problem statement (TCS Perspective)



m as small as possible

Assume x has low complexity: x is k-sparse (with noise)

Construct

Matrix $\Phi\colon \mathbb{R}^n \to \mathbb{R}^m$

Decoding algorithm $\mathcal{D}$

Given Φx for any signal $x \in \mathbb{R}^n$, we can, with high probability, quickly recover $\widehat{x}$ with

$$\|x - \widehat{x}\|_p \leq (1+\epsilon) \min_{y\ k-sparse} \|x - y\|_q = (1+\epsilon)\|x - x_k\|_q$$

Comparison with Sparse Approximation

SPARSE: Given y and Φ, find (sparse) x such that $y = \Phi x$. Return $\widehat{x}$ with guarantee

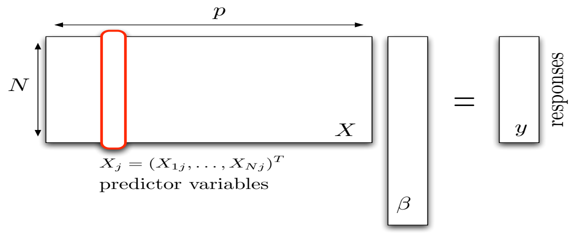
$$\|\Phi\widehat{x} - y\|_2 \quad \text{small compared with } \|y - \Phi x_k\|_2.$$

CS: Given y and Φ, find (sparse) x such that $y = \Phi x$. Return $\widehat{x}$ with guarantee

$$\|\widehat{x} - x\|_p \quad \text{small compared with } \|x - x_k\|_q.$$

p and q not always the same, not always $= 2$.

Comparison with Statistical Learning



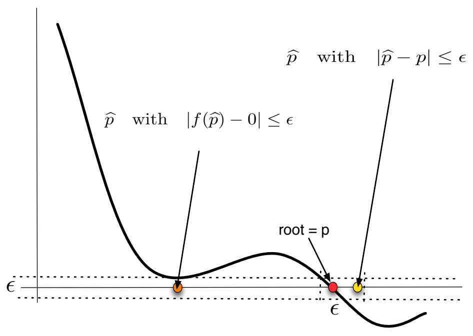
Goal: Given X and y, find α and coeffs. $\beta \in \mathbb{R}^p$ for linear model that minimizes the error

$$(\hat{\alpha}, \hat{\beta}) = \text{arg min} \|X\beta - (y - \alpha)\|_2^2.$$

Statistics: X drawn iid from distribution (i.e., cheap generation), characterize mathmematical performance as a function of distribution

TCS: X (or distribution) is constructed (i.e., expensive), characterize algorithmic performance as a function of space, time, and randomness

Analogy: root-finding



SPARSE: Given f (and $y = 0$), find p such that $f(p) = 0$. Return $\widehat{p}$ with guarantee

$$|f(\widehat{p}) - 0| \quad \text{small.}$$

CS: Given f (and $y = 0$), find p such that $f(p) = 0$. Return $\widehat{p}$ with guarantee

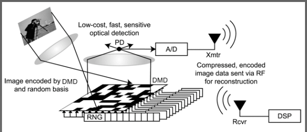
$$|\widehat{p} - p| \quad \text{small.}$$

Parameters

1. Number of measurements m
2. Recovery time
3. Approximation guarantee (norms, mixed)
4. One matrix vs. distribution over matrices
5. Explicit construction
6. Universal matrix (for any basis, after measuring)
7. Tolerance to measurement noise

Applications

* Data stream algorithms
  - x_i = number of items with index i
  - can maintain Φx under increments to x
  - recover approximation to x
* Efficient data sensing
  - digital/analog cameras
  - analog-to-digital converters
  - high throughput biological screening (pooling designs)



* Error-correcting codes
  - code $\{y \in \mathbb{R}^n | \Phi y = 0\}$
  - x = error vector, Φx = syndrome

Two approaches

Geometric [Donoho '04],[Candes-Tao '04, '06],[Candes-Romberg-Tao '05], [Rudelson-Vershynin '06], [Cohen-Dahmen-DeVore '06], and many others...

- *Dense* recovery matrices that satisfy RIP (e.g., Gaussian, Fourier)
- Geometric recovery methods (ℓ_1 minimization, LP)

$$\widehat{x} = \mathrm{argmin}\|z\|_1 \text{ s.t. } \Phi z = \Phi x$$

- Uniform guarantee: one matrix A that works for all x

Combinatorial [Gilbert-Guha-Indyk-Kotidis-Muthukrishnan-Strauss '02], [Charikar-Chen-FarachColton '02] [Cormode-Muthukrishnan '04], [Gilbert-Strauss-Tropp-Vershynin '06, '07]

- *Sparse* random matrices (typically)
- Combinatorial recovery methods or weak, greedy algorithms
- Per-instance guarantees, later uniform guarantees

Summary

* Sparse approximation, statistical learning, and compressive sensing intimately related
* Many models of computation and scientific/technological problems in which they all arise
* Algorithms for all similar: optimization and greedy
* Community progress on geometric and statistical models for matrices Φ and signals x, different problem instance types
* Explicit constructions?
* Better/different geometric/statistical models?
* Better connections with coding and complexity theory?