Compressive Sensing Theory and L1-Related Optimization Algorithms

Yin Zhang

Department of Computational and Applied Mathematics
Rice University, Houston, Texas, USA

CAAM Colloquium
January 26, 2009

Outline:

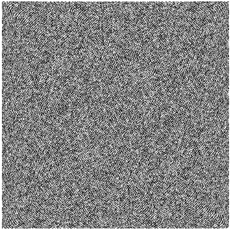
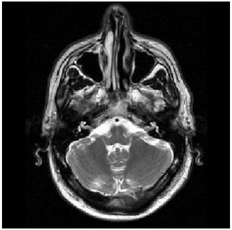
- What's Compressive Sensing (CS)?
- CS Theory: RIP and Non-RIP
- Optimization Algorithms
- Concluding Remarks

Acknowledgments

- Collaborators: Wotao Yin, Elaine Hale
- Students: Junfeng Yang, Yilun Wang
- Funding Agencies: NSF, ONR

MRI: Magnetic Resonance Imaging

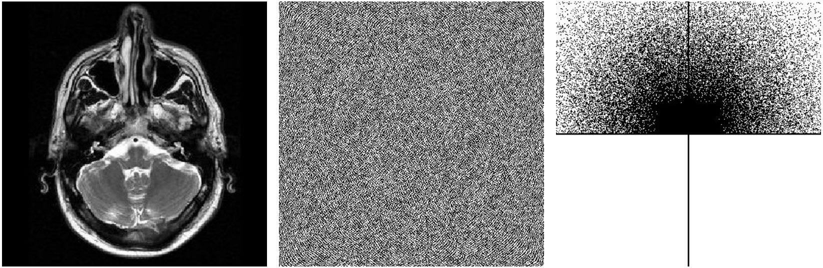
MRI Scan $\Longrightarrow$ Fourier coefficients $\Longrightarrow$ Images

Is it possible to cut the scan time into half?

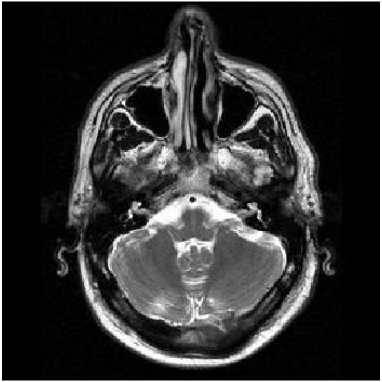
Numerical Simulation

- FFT2(image) $\Longrightarrow$ Fourier coefficients
- Pick 25% coefficients at random (with bias)
- Reconstruct image from the 25% coefficients



Simulation Result

Original vs. Reconstructed



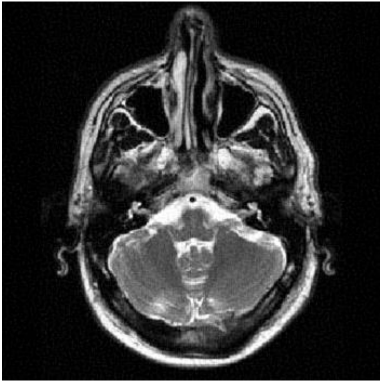


Image size: 350×350. Reconstruction time: ≤ 1s

Image Reconstruction Model

$$\min_{u} \ \alpha TV(u) + \beta \|u\|_1 + \frac{1}{2}\|F_p u - f_p\|_2^2$$

- u is the unknown image
- F_p — partial Fourier matrix
- f_p — partial Fourier coefficients
- $TV(u) = \sum_i \|(Du)_i\| = \|\text{grad magnitude}\|_1$

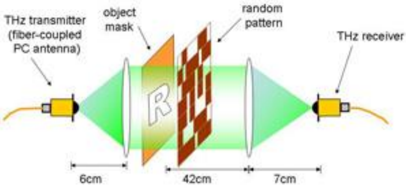
Compressing Sensing may cut scan time 1/2 or more

- Lustig, Donoho and Pauly, MR Medicine (2008)
- Research on "real-time" algorithms still needed

CS Application: Single-pixel Camera

Single-pixel Camera Has Multiple Futures

ScienceDaily (Oct. 20, 2008) A terahertz version of the single-pixel camera developed by Rice University researchers could lead to breakthrough technologies in security, telecom, signal processing and medicine. ......



Kelly Lab and Baranuik group, ECE at Rice

`http://www.dsp.ece.rice.edu/cscamera/`

What's Compressive Sensing (CS)?

Standard Paradigm in Signal Processing:

- Sample "full data" $x^* \in \mathbb{R}^n$ (subject to Nyquist limit).
- Then compress (transform + truncation)
- Decoding is simple (inverse transform)

Acquisition can become a bottleneck (time, power, speed, ...)

Paradigm Shift: Compressive Sensing

- Acquire less data $b_i = a_i^T x^*, i = 1, \cdots, m \ll n$.
- Decoding is more costly: getting x^* from $Ax = b$.

Advantage: Reducing acquisition size from n to m

CS – Emerging Methodology

Signal $x^* \in \mathbb{R}^n$. Encoding $b = Ax^* \in \mathbb{R}^m$

Fewer measurements taken ($m < n$), but no free lunch

- prior information on signal x^* required
- "good" measurement matrix A needed

Prior info is sparsity:

Ψx^* has many elements = 0 (or $\|\Psi x^*\|_0$ is small)

When does it work?

- Sparsfying basis Ψ is known
- $A \in \mathbb{R}^{m \times n}$ is "random-like"
- $m > \|\Psi x^*\|_0$ sufficiently

RICE

Sparsity is Common under Transforms

Many have sparse representations under known bases:

- Fourier, Wavelets, curvelets, DCT, ...

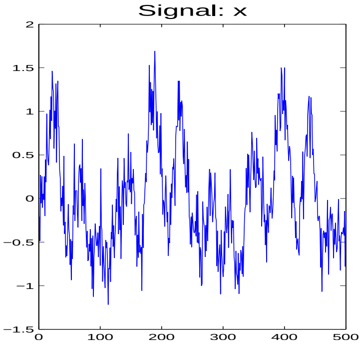


Figure: Before and After DCT Transform

Decoding in CS

Given (A, b, Ψ), find the sparsest point:

$$x^* = \arg\min\{\|\Psi x\|_0 : Ax = b\}$$

From combinatorial to convex optimization:

$$x^* = \arg\min\{\|\Psi x\|_1 : Ax = b\}$$

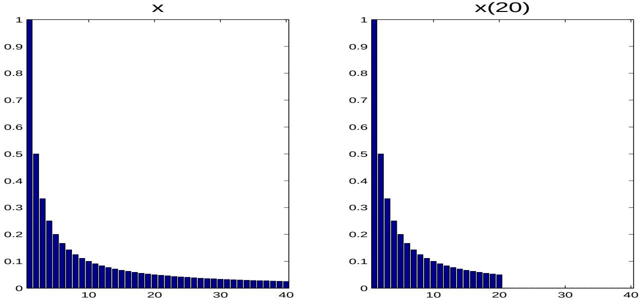
1-norm is sparsity promoting (e.g., Santosa-Symes 85)

- Basis pursuit (Donoho et al 98)
- Many variants; e.g., $\|Ax - b\|_2 \leq \sigma$ for noisy b
- Greedy algorithms (e.g., Tropp-Gilbert 05, ...)
- Big question: when is $\|\cdot\|_0 = \|\cdot\|_1$?

RICE

Notation $x(k)$: k-term Approximation

Keeping the k largest elements of x and setting the rest to 0 produce a k-term approximation of x, denoted by $x(k)$.



RICE

CS Theory – RIP Based

Assume $\Psi = I$. Let

$$x^* = \arg\min\{\|x\|_1 : Ax = A\bar{x}\}$$

A Celebrated Result:

Theorem: (Candes-Tao 2005, C-Romberg-T 2005)

If $A \in \mathbb{R}^{m \times n}$ is iid standard normal, with high probability (WHP)

$$\|x^* - \bar{x}\|_1 \leq C(\mathbf{RIP}_{2k}(A))\|\bar{x} - \bar{x}(k)\|_1$$

for $k \leq O(m/[1 + \log(n/m)])$ $(k < m < n)$.

– Donoho (2005) obtained similar RIP-like results.

– Most subsequent analyses use RIP.

RICE

What Is RIP?

Restricted Isometry Property:

$\textbf{RIP}_k(A) \in (0,1) \triangleq \min\{\sigma\}$ so that for some $r > 0$

$$(1-\sigma)r \leq \left(\frac{\|Ax\|}{\|x\|}\right)^2 \leq (1+\sigma)r, \quad \forall \, \|x\|_0 = k.$$

- $\textbf{RIP}_k(A)$ measures conditioning of $\{[k \text{ columns of } A]\}$
- Candes-Tao theory requires $\textbf{RIP}_{2k}(A) < 0.414$
- $\textbf{RIP}_k(GA)$ can be arbitrarily bad for nonsingular G

RICE

Is RIP indispensable?

Invariance of solution w.r.t. nonsingular G:

$$x^* = \arg\min\{\|\Psi x\|_1 : GAx = Gb\}$$

E.g., orthogonalize rows of A so $GA = Q$ and $QQ^T = I$.

Is GA always as good an encoding matrix as A is?

Is RIP indispensable?

Invariance of solution w.r.t. nonsingular G:

$$x^* = \arg\min\{\|\Psi x\|_1 : GAx = Gb\}$$

E.g., orthogonalize rows of A so $GA = Q$ and $QQ^T = I$.

Is GA always as good an encoding matrix as A is?

"Of course".

RICE

Is RIP indispensable?

Invariance of solution w.r.t. nonsingular G:

$$x^* = \arg\min\{\|\Psi x\|_1 : GAx = Gb\}$$

E.g., orthogonalize rows of A so $GA = Q$ and $QQ^T = I$.

Is GA always as good an encoding matrix as A is?
"Of course". But Candes-Tao theory doesn't say so.

Moreover,

- RIP conditions are known to be overly stringent
- RIP analysis is not simple nor intuitive (in my view)

RICE

A Non-RIP Analysis (Z, 2008)

Lemma: For any $x, y \in \mathbb{R}^n$ and $p = 0, 1$,

$$\sqrt{\|y\|_0} < \frac{1}{2}\frac{\|x-y\|_1}{\|x-y\|_2} \implies \|y\|_p < \|x\|_p.$$

Define

$$\mathcal{F} \triangleq \{x : Ax = Ax^*\} = x^* + \mathrm{Null}(A).$$

Corollary: If above holds for $y = x^* \in \mathcal{F}$ and all $x \in \mathcal{F} \setminus \{x^*\}$,

$$x^* = \arg\min\{\|x\|_p : Ax = Ax^*\}, \quad p = 0, 1.$$

- The larger the "1 vs 2" norm ratio in $\mathrm{Null}(A)$, the better.
- What really matters is $\mathrm{Null}(A)$, not representation A.

RICE

"1 vs 2" Ratio is Mostly Large

In the entire space $\mathbb{R}^n$,

$$1 \leq \|v\|_1/\|v\|_2 \leq \sqrt{n},$$

but the ratio $\gg 1$ in "most" subspaces (or WHP).

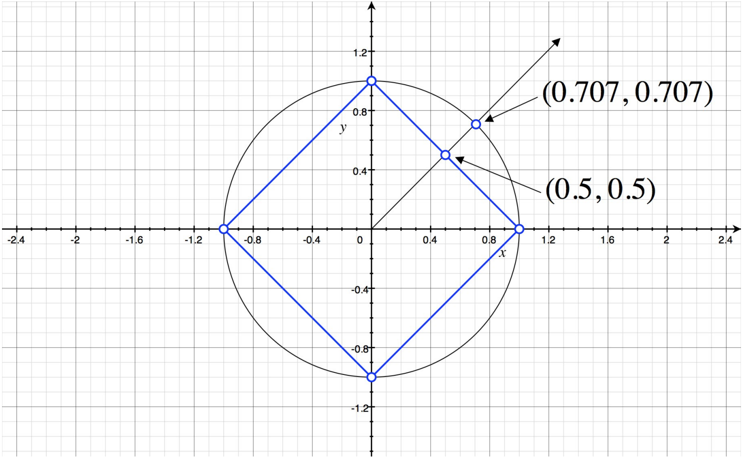
A result from Kashin-Garnaev-Gluskin (1978,84)

In "most" $(n-m)$-D subspaces $\mathcal{S} \subset \mathbb{R}^n$ (or WHP),

$$\frac{\|v\|_1}{\|v\|_2} > c_1 \sqrt{\frac{m}{\log(n/m)}}, \quad \forall v \in \mathcal{S}.$$

– In fact, Prob($\{$good $\mathcal{S}\}$) $\to 1$ as $n-m \to \infty$.
– A random space will give the best chance.
– Random-like matrices should do OK too.

"1 vs 2" Ratio in $\mathbb{R}^2$



E.g., in most subspaces, $\|v\|_1 / \|v\|_2 > 80\%\sqrt{2}$.

RICE

An RIP-free Result

Let $$x^* = \arg\min\{\|x\|_1 : GAx = GA\bar{x}\}$$

Theorem (Z, 2008): For all $k < k^*$ and $\bar{x} - \bar{x}(k)$ "small",

$$\|x^* - \bar{x}(k)\|_p \le C(k/k^*)\|P_r(\bar{x} - \bar{x}(k))\|_p, \quad p = 1 \text{ or } 2$$

- $P_r =$ orthogonal projection onto Range(A^T)
- $k^* \ge c_1 m/[1 + \log(n/m)]$ WHP if $A_{ij} \sim \mathcal{N}(0,1)$
- (compare C-T: $\|x^* - \bar{x}\|_1 \le C(\mathbf{RIP}_{2k}(A))\|\bar{x} - \bar{x}(k)\|_1$)

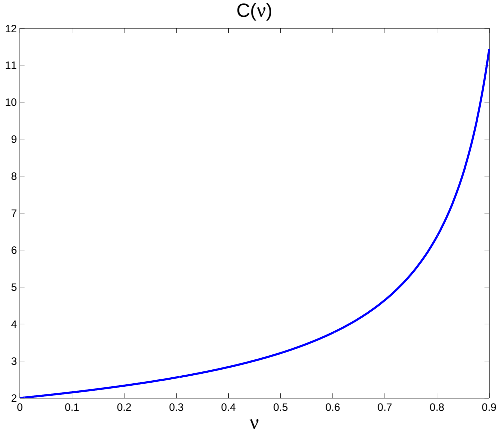
Key Difference:

- **RIP**: the smaller RIP, the more stable
- **RIP-free**: the sparser, the more stable

RICE

RIP-free Constant $C(k/k^*)$

$$C(\nu) = 1 + \frac{1 + \nu\sqrt{2 - \nu^2}}{1 - \nu^2} \geq 2, \;\; \nu \in [0, 1)$$



Models with More Prior Information

Theoretical guarantees previously existed only for

$$\min\{\|\Psi x\|_1 : Ax = b\}$$
$$\min\{\|\Psi x\|_1 : Ax = b, x \geq 0\}$$

Non-RIP analysis (Z, 2008) extends to

$$\min\{\|\Psi x\|_1 : Ax = b, x \in S\}$$

$$e.g., \quad \min\{\|\Psi x\|_1 : Ax = b, \; \|x - \hat{x}\| \leq \delta\}$$
$$e.g., \quad \min\{\|\Psi x\|_1 + \mu \mathrm{TV(x)} : Ax = b\}$$

RICE

Uniform Recoverability

What types of random matrices are good?

- Standard normal (Candes-Tao 05, Donoho 05)
- Bernoulli and a couple more (Baraniuk-Davenport-DeVore-Wakin, 07)
- Some partial orthonormal matrices (Rudelson-Vershynin, 06)

Uniform Recoverability (Z, 2008)

"All iid random matrices are asymptotically equally good"

- as long as the $4 + \delta$ moment remains bounded
- used a random determinant result by Girko (1998)

Algorithms for CS

Algorithmic Challenges in CS

- Large dense matrices, (near) real-time processing
- Standard (simplex, interior-point) methods not suitable

Optimization seems more robust than greedy methods

In many cases, it is faster than other approaches.

- Efficient algorithms can be built on Av and $A^T v$.
- Solution sparsity helps.
- Fast transforms help.
- Structured random matrices help.

RICE

Fixed-point Shrinkage

$$\min_x \|x\|_1 + \mu f(x)$$

Algorithm:

$$x^{k+1} = Shrink(x^k - \tau \nabla f(x^k), \tau/\mu)$$

where

$$Shrink(y, t) = y - \mathrm{Proj}_{[-t,t]}(y)$$

- A first-order method follows from classic operator splitting
- Discovered in signal processing by many since 2000's
- Convergence properties analyzed extensively

RICE

New Convergence Results

(Hale, Yin & Z, 2007)

How can solution sparsity help a 1st-order method?

- Finite Convergence: for all but a finite # of iterations,

$$\begin{aligned} x_j^k &= 0, &\quad \text{if } x_j^* = 0 \\ \operatorname{sign}(x_j^k) &= \operatorname{sign}(x_j^*), &\quad \text{if } x_j^* \neq 0 \end{aligned}$$

- q-linear rate depending on "reduced" Hessian:

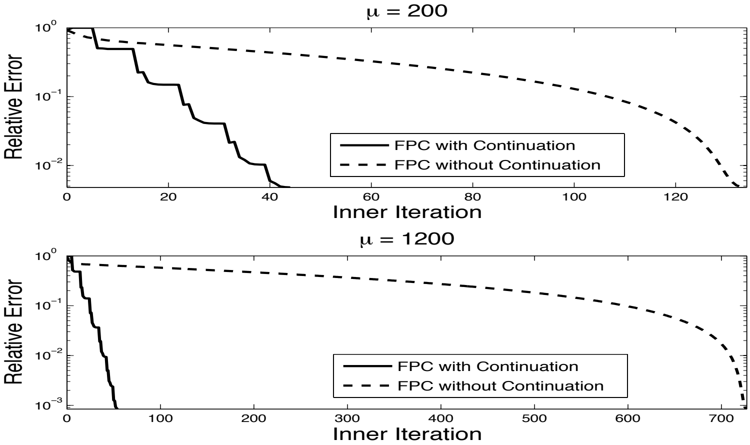
$$\limsup_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \frac{\kappa(H^*_{EE}) - 1}{\kappa(H^*_{EE}) + 1}$$

where H^*_{EE} is a sub-Hessian of f at x^* ($\kappa(H^*_{EE}) \leq \kappa(H^*)$), and $E = \mathrm{supp}(x^*)$ (under a regularity condition).

The sparser x^* is, the faster the convergence.

RICE

FPC: Fixed-Point Continuation (say, $\mu^k = 2\mu^{k-1}$)



(Numerical comparison results in Hale, Yin & Z 2007)

FPC-AS: FPC + Active Set

1st-order CS methods slow down or fail when sparsity approaches the threshold of the L0-L1 equivalence.

Can the number of measurements be pushed to limit?

Active Set: Combining 1st and 2nd orders

$$\min_x \|x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2$$

- Use shrinkage to estimate support and signs (1st order)
- Fix support and signs, solve the reduced QP (2nd order)
- Repeat until convergence

– Solved some hard problems on which other solvers failed
– Z. Wen, W. Yin, D. Goldfarb and Z, $\geq$2008

RICE

Results on a Difficult Problem

$$\min \|x\|_1, \text{ s.t. } Ax = b$$

where $A \in \mathbb{R}^{m \times n}$ is a partial DCT matrix (dense).

Table: Comparison of 6 Solvers

Problem	Solver	Rel-Err	CPU
Ameth6Xmeth2K151	FPC	4.8e-01	60.8
$n = 1024$	spg-bp	4.3e-01	50.7
$m = 512$	Cplex-primal	1.0e-12	19.8
$k = 150$	Cplex-dual	9.7e-13	11.1
$x_i = \pm 1$	Cplex-barrier	2.7e-12	22.1
	FPC-AS	7.3e-10	0.36

RICE

TV Regularization

Discrete total variation (TV) for an image:

$$TV(u) = \sum \|D_i u\| \quad \text{(sum over all pixels)}$$

(1-norm of gradient magnitude)

- Advantage: able to capture sharp edges
- Rudin-Osher-Fatemi 1992, Rudin-Osher 1994
- Also useful in CS applications (e.g., MRI)

Fast TV algorithms were in dire need for many years

FTVd: Fast TV deconvolution

$$\text{(TVL2)} \qquad \min_u \sum \|D_i u\| + \frac{\mu}{2}\|Ku - f\|^2$$

Introducing $w_i \in \mathbb{R}^2$ (grayscale) and quadratic penalty:

$$\min_{u,w} \sum \left(\|w_i\| + \frac{\beta}{2}\|w_i - D_i u\|^2 \right) + \frac{\mu}{2}\|Ku - f\|^2$$

In theory, $u(\beta) \to u^*$ as $\beta \to \infty$. In practice, $\beta = 200$ suffices.

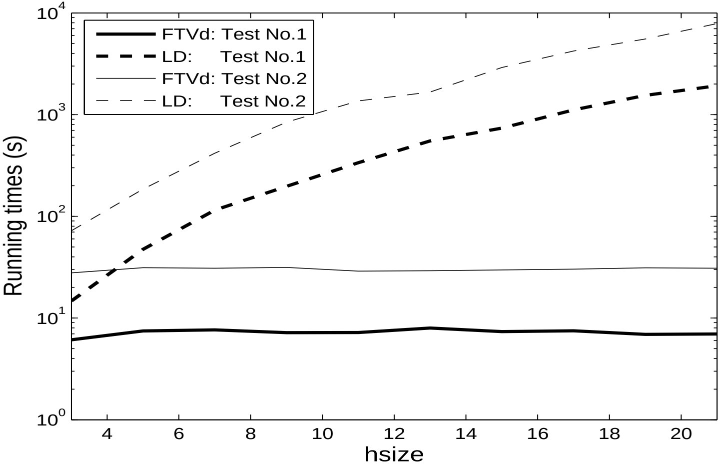
Alternating Minimization:

- For fixed u, $\{w_i\}$ solved by 2D-shrinkage at $O(N)$
- For fixed $\{w_i\}$, u solved by 2 FFTs at $O(N \log N)$

– Extended to color ($2 \to 6$), TVL1, and CS
(Yang-Wang-Yin-Z, 07-08)

RICE

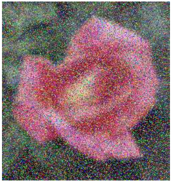
FTVd vs. Lagged Diffusivity



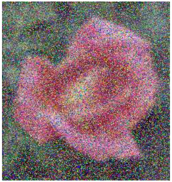
(Test 1: Lena 512 by 512; Test 2: Man 1024 by 1024)

Color Deblurring with Impulsive Noise

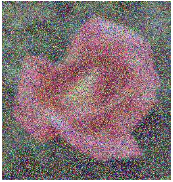
Bn. RV 40%



Bn. RV 50%



Bn. RV 60%



μ: 8, t: 161s, SNR: 16dB

μ: 4, t: 181s, SNR: 14dB

μ: 2, t: 204s, SNR: 10dB

Extension to CS

RecPF: Reconstruction from Partial Fourier Data

$$\min_{u} \; TV(u) + \lambda \|\Psi u\|_1 + \mu \|\mathcal{F}_p(u) - f_p\|^2$$

Based on same splitting and alternating idea (Yang-Z-Yin, 08)

Matlab Code:

http://www.caam.rice.edu/~optimization/L1/RecPF

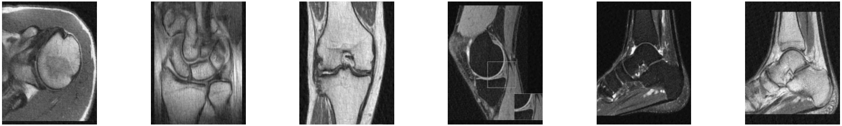


Figure: 250x250 images, 17% coefficients, CPU $\approx$ 0.2s

Concluding Remarks

What I have learned so far:

- CS can reduce data acquisition costs:
  less data, (almost) same content
- CS poses many algorithmic challenges:
  optimization algorithms for various models
- Case-by-case studies are often required:
  finding and exploiting structures
- Still a long way to go from theory to practice:
  but potentials and promises are real

Will CS be revolutionary?

Compressive Sensing Resources:

`http://www.dsp.ece.rice.edu/cs/`

Papers and Software (FPC, FTVd and RecPF) available at:

`http://www.caam.rice.edu/~optimization/L1`

Thank You!

Happy Lunar New Year!