Enhanced Compressive Sensing and More

Yin Zhang

Department of Computational and Applied Mathematics
Rice University, Houston, Texas, U.S.A.

Nonlinear Approximation Techniques Using L1
Texas A & M University
May 16th, 2008

Outline

Collaborators

- Wotao Yin, Elaine Hale
- Students: Yilun Wang, Junfeng Yang
- Acknowledgment: NSF DMS-0442065

The Outline of the Talk

- Compressive Sensing (CS): when $\ell_0 \Leftrightarrow \ell_1$?
- An accessible proof and Enhanced CS
- Rice L1-Related Optimization Project
- – An L1 Algorithm: FPC
- – A TV Algorithm: FTVd
- – Numerical Results

Compressive Sensing (CS)

Recover sparse signal from incomplete data:

- Unknown signal $x^* \in \mathbb{R}^n$
- Measurements: $b = Ax^* \in \mathbb{R}^m$, $m < n$
- x^* is sparse: #nonzeros $\|x^*\|_0 < m$

1 Solution to 2 Problems?

- ℓ_0-Prob: $\min\{\|x\|_0 : Ax = b\} \;\Rightarrow\;$ sparsest solution (hard)
- ℓ_1-Prob: $\min\{\|x\|_1 : Ax = b\} \;\Rightarrow\;$ lin. prog. solution (easy)
- **Recoverability:** When does the same x^* solve both?

RICE

CS Recoverability

When does the following happen? ($b = Ax^*$)

$$\{x^*\} = \arg\min\{\|x\|_0 : Ax = b\} = \arg\min\{\|x\|_1 : Ax = b\}$$

Answer: For a random $A \in \mathbb{R}^{m \times n}$,

$$\|x^*\|_0 < \frac{c \cdot m}{\log(n/m)}.$$

— Candes-Romberg-Tao, Donoho *et al*, 2005
— Rudelson-Vershynin, 2005, 2006
— Baraniuk-Davenport-DeVore-Wakin, 2007 ......

RICE

Recoverability Guarantees

Theoretical guarantees available:

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

(Donoho-Tanner 2005, Z 2005)

What about these convex models?

$$\min\{\|\Phi x\|_1 + \mu \mathrm{TV}(x) : Ax = b\}$$

$$\min\{\|\Phi x\|_1 + \mu\|x - \hat{x}\| : Ax = b\}$$

$$\min\{\|\Phi x\|_1 : Ax = b, Bx \leq c, x \in [l, u]\}$$

$$\cdots\cdots\cdots\cdot$$

RICE

CS Analysis

When is $\ell_0 \;\Leftrightarrow\; \ell_1$?

- Most analyses are based on the notion of **RIP**:
  —Restricted Isometry Property
- Or based on "**counting faces**" of polyhedrons
- Derivations are quite involved and not transparent
- Generalize CS analysis to more models?

A simpler, gentler, more general analysis?

Yes. Using **Kashin-Garnaev-Gluskin** (KGG) inequality.

(Extension to Z, CAAM Report TR05-09)

KGG Result

ℓ_1-norm vs. ℓ_2-norm:

$$\sqrt{n} \geq \frac{\|v\|_1}{\|v\|_2} \geq 1, \quad \forall v \in \mathbb{R}^n \setminus \{0\}$$

However, $\|v\|_1 / \|v\|_2 \gg 1$ in most subspaces of $\mathbb{R}^n$.

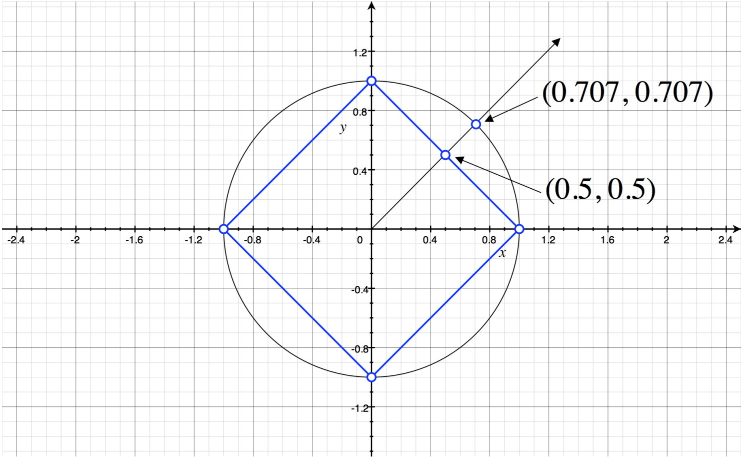
Theorem: (Kashin 77, Garnaev-Gluskin 84)

Let $A \in \mathbb{R}^{m \times n}$ be iid Gaussian. With probability $> 1 - e^{-c_1(n-m)}$,

$$\frac{\|v\|_1}{\|v\|_2} \geq \frac{c_2\sqrt{m}}{\sqrt{\log(n/m)}}, \quad \forall v \in Null(A) \setminus \{0\}$$

where c_1 and c_2 are absolute constants.

A Picture in 2D



In most subspaces, $\|v\|_1/\|v\|_2 \geq 0.8 * \sqrt{2} > 1.1$

Sparsest Point vs. ℓ_p-Minimizer, $p \in (0, 1]$

When does the following hold on $C \subset \mathbb{R}^n$?

$$\{x^*\} = \arg\min_{x \in C} \|x\|_0 = \arg\min_{x \in C} \||x|^p\|_1$$

This means: (i) "$\ell_0 \Leftrightarrow \ell_p$" on C, (ii) uniqueness of x^*.

A Sufficient Condition — entirely on sparsity

$$\sqrt{\|x^*\|_0} < \frac{1}{2} \frac{\||v|^p\|_1}{\||v|^p\|_2}, \quad \forall\, v \in (C - x^*) \setminus \{0\}$$

(10-line, elementary proof skipped)

Recoverability Proved and Generalized

For $C = \{x : Ax = b, x \in S\}$,

$$C - x^* = Null(A) \cap (S - x^*), \quad \forall S \subset \mathbb{R}^n$$

$$[\ell_0 \Leftrightarrow \ell_p] \Leftarrow \|x^*\|_0^{\frac{1}{2}} < \frac{1}{2} \frac{\||v|^p\|_1}{\||v|^p\|_2}, \; \forall \, v \in Null(A) \cap (S - x^*) \setminus \{0\}$$

For a Gaussian random A, by GKK

$$[\ell_0 \Leftrightarrow \ell_p] \;\text{ on }\; C \;\overset{\text{h.p.}}{\Longleftarrow}\; \|x^*\|_0 < \frac{c(p) \cdot m}{\log(n/m)}$$

(Stability results also available for noisy data)

Enhanced Compressive Sensing

ECS: with prior information $x \in S$

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

We have shown ECS recoverability is at least as good as CS.

More prior information (beside nonnegativity)?

$\min\{\|x\|_1 + \mu \, \mathrm{TV}(x) : Ax = b\} \;\Rightarrow\; S = \{x : \mathrm{TV}(x) \le \delta\}$

$\min\{\|x\|_1 + \mu \|x - \hat{x}\| : Ax = b\} \;\Rightarrow\; S = \{x : \|x - \hat{x}\| \le \delta\}$

...... and many more possibilities.

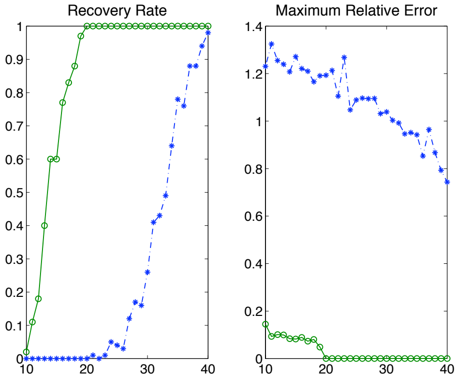
More ECS models, more algorithmic challenges for optimizers.

ECS vs. CS: A case study

Unknown signal x^* close to a prior sparse x_p:

$$\text{ECS:}\quad \min\{\|x\|_1 : Ax = b, \|x - x_p\|_1 \le \delta\}$$

With 10% differences in supports and nonzero values,



Rice L1-Related Optimization Project

Computational & Applied Math. Dept. in Engineering School:

- Y. Z., Wotao Yin (Elaine Hale, left)
- Students

Optimization Algorithmic Challenges in CS

- Large-scale, (near) real-time processing
- Dense matrices, non-smooth objectives
- Traditional (simplex, interior-point) methods have trouble.

Can convex optimization be practical in CS?

RICE

Convex Optimization Works in CS

Convex Optimization is generally more robust w.r.t noise.

Is it too slow for large-scale applications?

In many cases, it is faster than other approaches.

- Solution sparsity helps.
- Fast transforms help.
- Structured random matrices help.
- Efficient algorithms can be built on Av and $A^T v$.
- Real-time algorithms are possible for problems with special structures (like MRI).
- 2 examples from our work: FPC and FTVd

Forward-Backward Operator Splitting

Derivation (since 1950's):

$$
\begin{aligned}
\min \|x\|_1 + \mu f(x) \quad &\Leftrightarrow \quad 0 \in \partial\|x\|_1 + \mu\nabla f(x) \\
&\Leftrightarrow \quad -\tau\mu\nabla f(x) \in \tau\partial\|x\|_1 \\
&\Leftrightarrow \quad x - \tau\mu\nabla f(x) \in x + \tau\partial\|x\|_1 \\
&\Leftrightarrow \quad (I + \tau\partial\|\cdot\|_1)x \ni x - \tau\mu\nabla f(x) \\
&\Leftrightarrow \quad \{x\} \ni (I + \tau\partial\|\cdot\|_1)^{-1}(x - \tau\mu\nabla f(x)) \\
&\Leftrightarrow \quad x = shrink(x - \tau\nabla f(x), \tau/\mu)
\end{aligned}
$$

Equivalence to Fixed Point

$$\min \|x\|_1 + \mu f(x) \iff x = Shrink(x - \tau\nabla f(x), \tau/\mu)$$

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 FB-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,

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

- 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.

FPC: Fixed-Point Continuation

$$x(\mu) := \arg\min_x \|x\|_1 + \mu f(x)$$

Idea: approximately follow the path $x(\mu)$

FPC: Set $\mu = \mu_0 < \mu_{\max}$, and x_0.
Do until $\mu \geq \mu_{\max}$

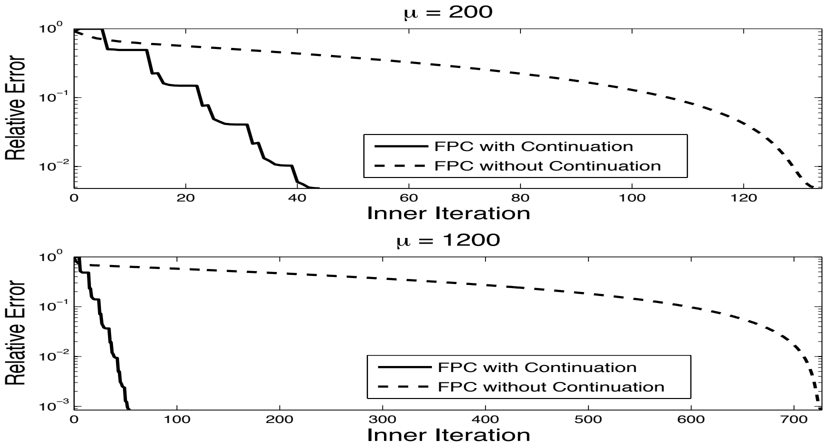
1. Starting from x_0, do shrinkage until "converged"
2. Set $\mu = 2\mu$, and x_0 to the previous "solution".

End Do

- Smaller $\mu \rightarrow$ sparser $x(\mu) \rightarrow$ faster convergence
- Converges is also fast for larger μ due to 'warm starts".
- Generally effective, may slow down near "boundary".

RICE

Continuation Makes It Kick



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

Random Kronicker Products Plus FPC

Fully random matrices are computationally costly.

Random Kronicker Products

$$A = A_1 \otimes A_2 \in \mathbb{R}^{m \times n},$$

where we only store A_1 $(m/p \times n/q)$ and A_2 $(p \times q)$.

$$Ax = \text{vec}\left(A_2 X A_1^T\right)$$

also requires much less computation.

- Trial: $n = 1\text{M}$, $m = 250\text{k}$, $\|x^*\|_0 = 25\text{k}$, $p = 500$, $q = 1000$.
- 2 (500 by 1000) matrices stored (reduction of 0.5M times).
- FPC solved the problem in 47s on a PC.

RICE

Total Variation Regularization

Discrete total variation (TV) for an image u:

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

(1-norm of the vector of 2-norms of discrete gradients)

- Advantage: able to capture sharp edges
- Rudin-Osher-Fatemi 1992, Rudin-Osher 1994
- Recent Survey: Chan and Shen 2006

Non-smoothness and non-linearity cause computational difficulties. Can TV be competitive in speed with others (say, Tikhonov-like regularizers)?

Fast TV deconvolution (FTVd)

$$(\mathrm{TV}+\mathrm{L2}) \qquad \min_u \sum \|D_i u\| + \frac{\mu}{2}\|Ku - f\|^2$$

Introducing $w_i \in \mathbb{R}^2$ and a quadratic penalty (Courant 1943):

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

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

Alternating Minimization:

- For fixed u, w_i can be solved by a 2D-shrinkage.
- For fixed w, quadratic can be minimized by 3 FFTs.

(Wang, Yang, Yin &Z, 2007, 2008)

FTVd

FTVd is a long-missing member of the half-qudratic class (Geman-Yang 95), using a 2D Huber-like approximation.

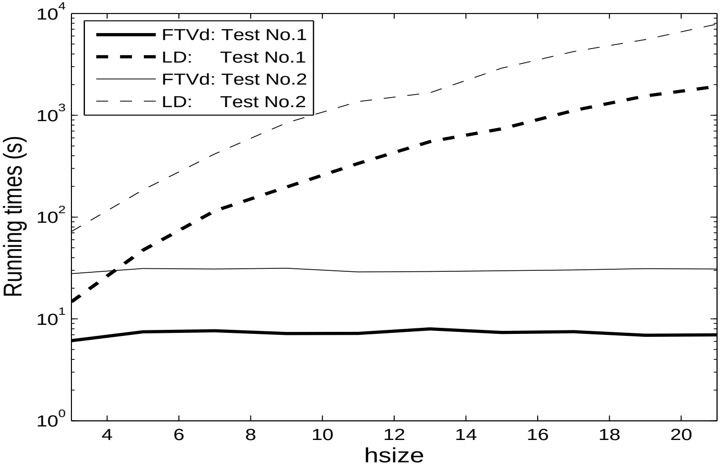
FTVd Convergence

- Finite convergence for $w_i^k \to 0$ (sparsity helps).
- Strong q-linear convergence rates for the others
- Rates depend on submatrices (sparsity helps).
- Continuation accelerates practical convergence.

FTVd Performance

- Orders of magnitude faster than Lagged Diffusivity.
- Comparable speed with Matlab deblurring, with better quality. TV models has finally caught up in speed.

FTVd vs. Lagged Diffusivity



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

FTVd vs. Others

Blurry&Noisy. SNR: 5.19dB

ForWaRD: SNR: 12.46dB, t = 4.88s

FTVd: $\beta = 2^5$, SNR: 12.58dB, t = 1.83s

deconvwnr: SNR: 11.51dB, t = 0.05s

deconvreg: SNR: 11.20dB, t = 0.34s

FTVd: $\beta = 2^7$, SNR: 13.11dB, t = 14.10s

RICE

FTVd Extensions

Multi-channel Image Deblurring (paper forthcoming)

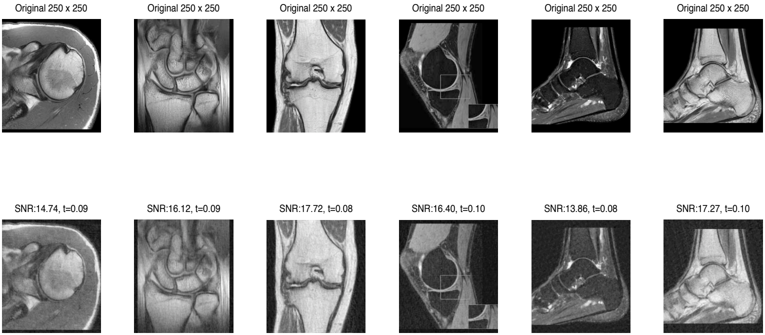
- cross-channel or within-channel blurring
- a "small" number of FFTs per iteration
- convergence results have been generalized
- TV+L^1 deblurring models (codes hard to find)

Other Extensions

- higher-order TV regularizers (reducing stair-casing)
- multi-term regularizations $\longrightarrow$ multiple splittngs
- locally weighted TV $\longrightarrow$ weighted shrinkage
- reconstruction from partial Fourier coefficients (MRI)

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

MRI Construction from 15% Coefficients

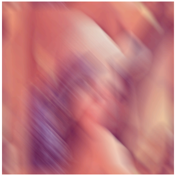


250 by 250 Images: time $\leq$ 0.1s on a PC (3 GHz Pentium D).

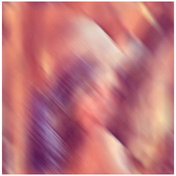
Color Image Deblurring

Original image: 512x512

Blurry & Noisy SNR: 5.1dB.



deconvlucy: SNR=6.5dB, t=8.9



deconvreg: SNR=10.8dB, t=4.4

deconvwnr: SNR=10.8dB, t=1.4

MxNopt: SNR=16.3dB, t=1.6

Comparison to Matlab Toolbox: 512×512 Lena

Summary

Take-Home Messages

- CS recoverability can be proved in 1 page via KGG.
- Prior information can never degrade CS recoverability, but may significantly enhance it.
- 1st-order methods can be fast thanks to solution sparsity (finite convergence, rates depending on sub-Hessians).
- TV models can be solved quickly if structures exploited.
- Continuation is necessary to make algorithms practical.
- Rice has a long tradition in optim. algorithms/software.

RICE

The End

Software FPC and FTVd available at:

http://www.caam.rice.edu/˜optimization/L1

Thank You!