Nonconvex compressive sensing

Fast, easy, and better

Rick Chartrand

Los Alamos National Laboratory

March 6, 2012

In theory, there's no difference between theory and practice. In practice, there is.

–Yogi Berra

Outline

Examples (Better)

A nonconvex objective for fast minimization (Fast, Easy)

High-dimensional data modeling

Avoiding local minima (Easy)

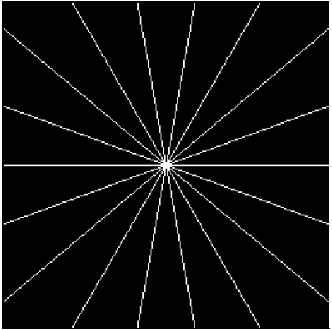
Summary

Motivating example

Suppose we want to reconstruct an image from samples of its Fourier transform. How many samples do we need?

Consider radial sampling, such as in MRI or (roughly) CT.

Shepp-Logan phantom, s



Ω

Nonconvexity is better

Fewer measurements are needed with nonconvex minimization:

$$\min_{x} H(\nabla x), \text{ subject to } (\mathcal{F}x)|_{\Omega} = (\mathcal{F}s)|_{\Omega}.$$

Nonconvexity is better

Fewer measurements are needed with nonconvex minimization:

$$\min_{x} H(\nabla x), \text{ subject to } (\mathcal{F}x)|_\Omega = (\mathcal{F}s)|_\Omega.$$

With $H(x) = \|x\|_1$, solution is $x = s$ with 18 lines ($\frac{|\Omega|}{|x|} = 6.9\%$).

backprojection, 18 lines

ℓ^1, 18 lines

Nonconvexity is better

Fewer measurements are needed with nonconvex minimization:

$$\min_{x} H(\nabla x), \text{ subject to } (\mathcal{F}x)|_\Omega = (\mathcal{F}s)|_\Omega.$$

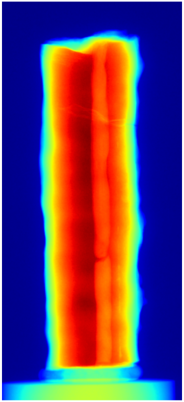
With $H(x) = \|x\|_1$, solution is $x = s$ with 18 lines ($\frac{|\Omega|}{|x|} = 6.9\%$).

With H a nonconvex functional to be described shortly, 9 lines suffice ($\frac{|\Omega|}{|x|} = 3.5\%$). (More than 10^{4500} local minima!)

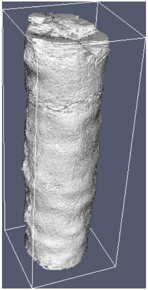
backprojection, 18 lines — ℓ^1, 18 lines — nonconvex, 9 lines — ℓ^1, 9 lines

3-D tomography

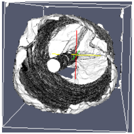
Six radiographs allow reconstruction of a stalagmite segment:



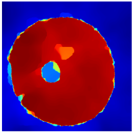
radiograph



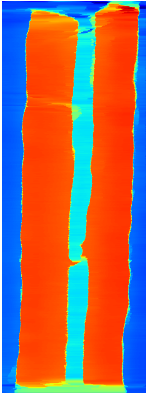
isosurface



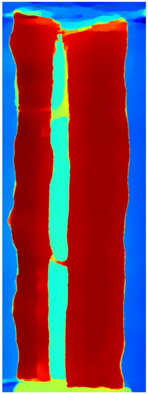
iso from end



z slice



x slice



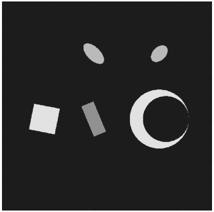
y slice

(with Gary Sandine, LANL)

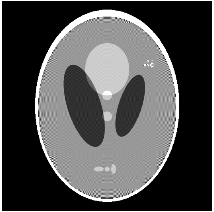
Frequency extrapolation

Consider the task of reconstructing an image with small test objects, from a 512×512 grid of samples of its continuous Fourier transform:

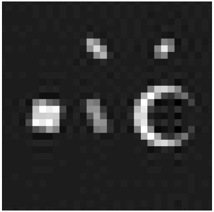
high-res. phantom



zoom-in on test objects



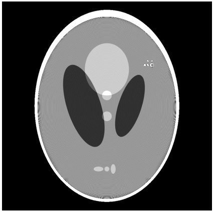
Inverse DFT of the data



zoom-in

with Emil Sidky, U. Chicago/Radiology

Frequency extrapolation



IDFT of zero-padded data

zoom

CS reconstruction, convex

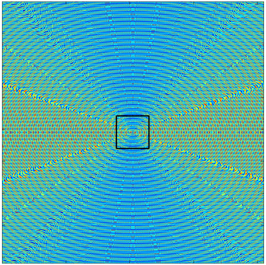
zoom

nonconvex

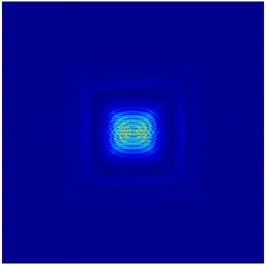
zoom

Frequency extrapolation

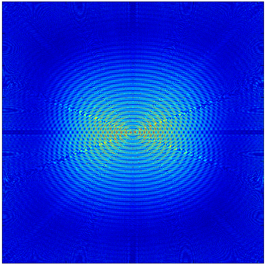
The Fourier transform of the reconstructions (scaled by $|\nu|^{3/2}$) shows that the nonconvex method results in better extrapolation.



4096×4096 FT data of object



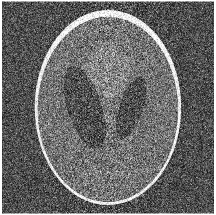
FT of convex result



FT of nonconvex result

Application: very noisy data

Independent Gaussian noise of $\sigma = 1000$ is added to the real and imaginary parts of the DFT of the 2048×2048 Shepp-Logan phantom. We exploit the greater SNR of the low-frequency portion of the data.



IFFT, SNR -10.2 dB (or $\sigma = 0.69$)

IFFT of zero-padded 256×256 data, SNR 7.1 dB

nonconvex reconstruction of 256×256 data, SNR 17.5 dB

nonconvex reconstruction from 20% of 256×256 data, SNR 14.2 dB

Interferometric imaging

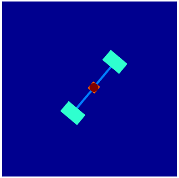
Given a network of N telescopes, the correlation between the electric field at each pair gives us $\binom{N}{2}$ Fourier samples.

Interferometric imaging

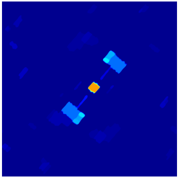
Given a network of N telescopes, the correlation between the electric field at each pair gives us $\binom{N}{2}$ Fourier samples. Radio astronomers use Earth-rotation synthesis to increase the sampling. But what about geostationary objects?

Interferometric imaging

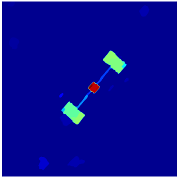
Given a network of N telescopes, the correlation between the electric field at each pair gives us $\binom{N}{2}$ Fourier samples. Radio astronomers use Earth-rotation synthesis to increase the sampling. But what about geostationary objects?



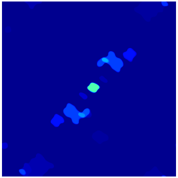
test image



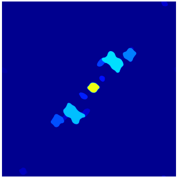
16 telescopes, convex



16 telescopes, nonconvex



10 telescopes, convex



10 telescopes, nonconvex

Outline

Examples (Better)

A nonconvex objective for fast minimization (Fast, Easy)

High-dimensional data modeling

Avoiding local minima (Easy)

Summary

Shrinkage

Many efficient algorithms rely on shrinkage (or soft thresholding). The solution of the problem

$$\min_{w} \|w\|_1 + \frac{1}{2\lambda}\|x - w\|_2^2$$

is given componentwise by:

$$w_i = \frac{x_i}{|x_i|} \max\{0, |x_i| - \lambda\}.$$

Shrinkage

Many efficient algorithms rely on shrinkage (or soft thresholding). The solution of the problem

$$\min_w \|w\|_1 + \frac{1}{2\lambda}\|x - w\|_2^2$$

is given componentwise by:

$$w_i = \frac{x_i}{|x_i|}\max\{0, |x_i| - \lambda\}.$$

Using p-shrinkage with $p < 1$ instead promotes sparsity more strongly:

$$w_i = \frac{x_i}{|x_i|}\max\{0, |x_i| - \lambda|x_i|^{p-1}\}.$$

What problem does this solve?

A new objective function

We can construct a function G_p so that p-shrinkage solves the analogous problem:

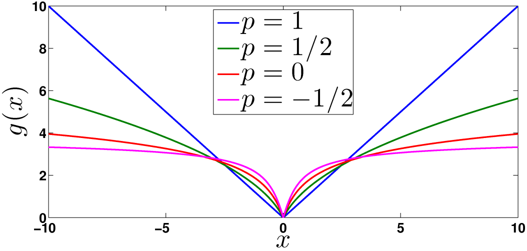
$$\min_w G_p(w) + \frac{1}{2\lambda}\|x - w\|_2^2$$

A new objective function

We can construct a function G_p so that p-shrinkage solves the analogous problem:

$$\min_{w} G_p(w) + \frac{1}{2\lambda}\|x - w\|_2^2$$

For $p \leq 1$, g is radial, radially strictly increasing, nonnegative, nonsmooth (at 0), continuous, and satisfies the triangle inequality. For large x, $g_p(x)$ grows like x^p/p.



Efficient algorithm

Our nonconvex generalization of split Bregman (or ADMM) is fast, and readily parallelizable. For example, to solve:

$$\min_{x} G_p(x), \text{ subject to } (\mathcal{F}\Phi x)|_\Omega = b$$

where the dictionary Φ gives a sparse representation of our signal, we iterate:

Efficient algorithm

Our nonconvex generalization of split Bregman (or ADMM) is fast, and readily parallelizable. For example, to solve:

$$\min_{x} G_p(x), \text{ subject to } (\mathcal{F}\Phi x)|_\Omega = b$$

where the dictionary Φ gives a sparse representation of our signal, we iterate:

1. a p-shrinkage,
2. solving a linear system with an explicit, fast inverse, and
3. updating Lagrange multipliers.

Code example

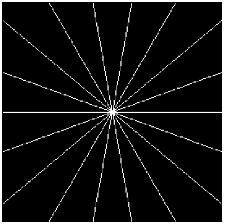
```
function x = splitFourierterse( b, M, Phi, PhiT, mu, lambda, p, ep, iter )
[ m, n ] = size( b );
w = zeros( m, n );
Lam1 = zeros( m, n );
Lam2 = zeros( m, n );
% main loop
for ii = 1 : iter
    % solve for x in the Fourier domain
    rhs = ( w + Lam1 ) / lambda + mu * PhiT( n * ifft2( M .* ( b + Lam2 ) ) );
    x = systeminverse( lambda, mu, M, Phi, PhiT,  rhs );
    % update w
    w = shrink( x - Lam1, lambda, p, ep );
    % update Lagrange multipliers, using "method of multipliers"
    Lam1 = Lam1 + w - x;
    Lam2 = Lam2 + b - M .* fft2( Phi( x ) ) / n;
end
function x = systeminverse( lambda, mu, M, Phi, PhiT,  y )
gmma = lambda^2 * mu / ( 1 + lambda * mu );
x = lambda * y - gmma * PhiT( ifft2( M .* fft2( Phi( y ) ) ) );
function y = shrink( x, lambda, p, ep )
% p-shrinkage using mollification
ax = sqrt( x .* conj( x ) );
y = max( ax - lambda * ( ax.^2 + ep ).^( p / 2 - 0.5 ), 0 );
id = ax ~= 0;
y( id ) = y( id ) .*  x( id ) ./ ax( id );
```

Phantom example

We reconstruct an image from samples of its Fourier transform:

$$\min_{\boldsymbol{x}} G_p(\nabla x), \text{ subject to } (\mathcal{F}x)|_\Omega = (\mathcal{F}s)|_\Omega.$$

test image s



9 lines/3.5% sampled

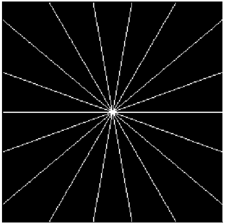
21 seconds

Phantom example

We reconstruct an image from samples of its Fourier transform:

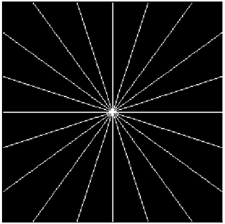
$$\min_{\boldsymbol{x}} G_p(\nabla \boldsymbol{x}), \text{ subject to } (\mathcal{F}\boldsymbol{x})|_\Omega = (\mathcal{F}\boldsymbol{s})|_\Omega.$$

test image $\boldsymbol{s}$



9 lines/3.5% sampled

21 seconds



10 lines/3.8% sampled

5 seconds

Outline

Examples (Better)

A nonconvex objective for fast minimization (Fast, Easy)

High-dimensional data modeling

Avoiding local minima (Easy)

Summary

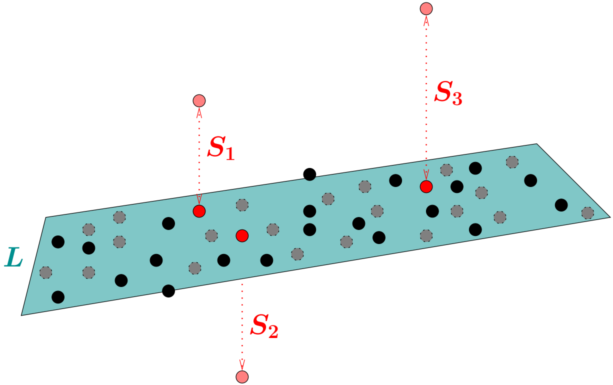
Robust data modeling

We turn the task of modeling a high-dimensional dataset into a matrix optimization problem, by forming a matrix D having each member of the dataset as a column.

Robust data modeling

We turn the task of modeling a high-dimensional dataset into a matrix optimization problem, by forming a matrix D having each member of the dataset as a column.

We seek to decompose D into a sum $L + S$, where L has low rank, and S is sparse.



Low rank + sparse decomposition

We could obtain our decomposition by solving the following:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \mathbf{rank}(\boldsymbol{L}) + \boldsymbol{\mu}\|\boldsymbol{S}\|_{\mathbf{0}}, \text{ subject to } \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{D}.$$

Low rank + sparse decomposition

We could obtain our decomposition by solving the following:

$$\min_{L,S} \operatorname{rank}(L) + \mu \|S\|_0, \text{ subject to } L + S = D.$$

An efficient approximation is to solve:

$$\min_{L,S} G_p(\sigma(L)) + \mu G_p(S) + \frac{\mu}{2\lambda} \|D - L - S - \Lambda\|_F^2.$$

Low rank + sparse decomposition

We could obtain our decomposition by solving the following:

$$\min_{L,S} \operatorname{rank}(L) + \mu\|S\|_0, \text{ subject to } L + S = D.$$

An efficient approximation is to solve:

$$\min_{L,S} G_p(\sigma(L)) + \mu G_p(S) + \frac{\mu}{2\lambda}\|D - L - S - \Lambda\|_F^2.$$

We obtain a robust, low-dimensional description of the dataset, and a set of salient features. We now examine the example of video, where each frame is a column of D.

Video example

video D, $\mathbf{240 \times 320}$ pixels, 288 frames

Video example

sparse component S

Video example

low-rank component L

Why might global minimization be possible?

Consider an ϵ-regularized ℓ^p objective, restricted to the feasible plane:

$$\sum_{i=1}^{N} (x_i^2 + \epsilon)^{p/2}.$$

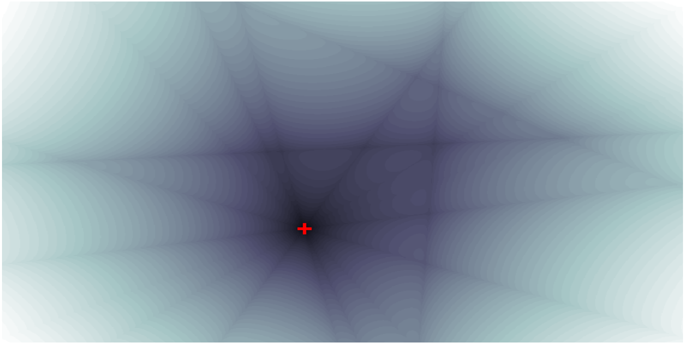
A moderate ϵ fills in the local minima.

Why might global minimization be possible?

Consider an ϵ-regularized ℓ^p objective, restricted to the feasible plane:

$$\sum_{i=1}^{N}(x_i^2+\epsilon)^{p/2}.$$

A moderate ϵ fills in the local minima.



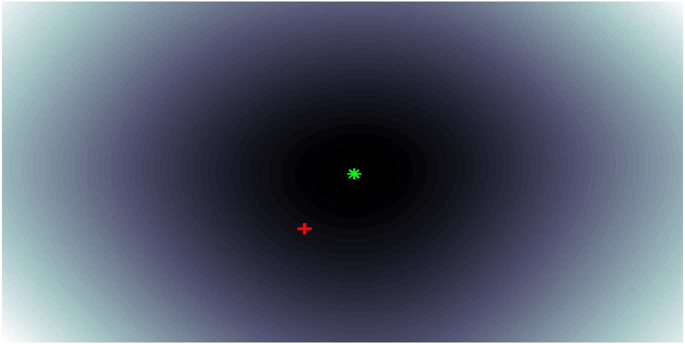
$\epsilon = 0$

Why might global minimization be possible?

Consider an ϵ-regularized ℓ^p objective, restricted to the feasible plane:

$$\sum_{i=1}^{N} \left(x_i^2 + \epsilon\right)^{p/2}.$$

A moderate ϵ fills in the local minima.



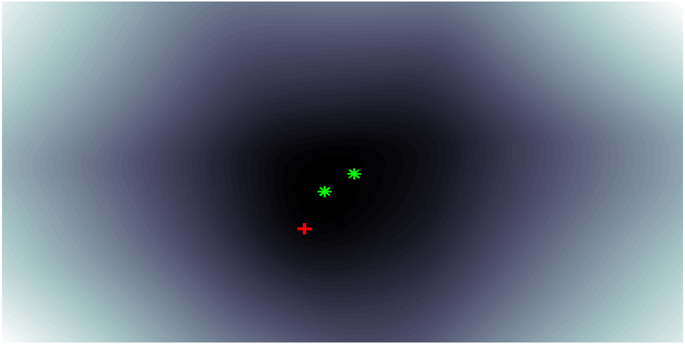
$\epsilon = 1$

Why might global minimization be possible?

Consider an ϵ-regularized ℓ^p objective, restricted to the feasible plane:

$$\sum_{i=1}^{N}(x_i^2+\epsilon)^{p/2}.$$

A moderate ϵ fills in the local minima.



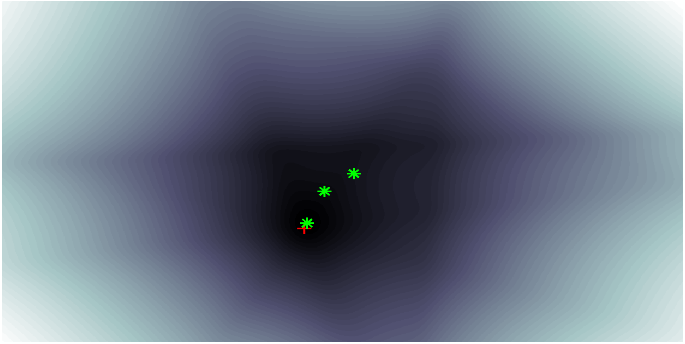
$\epsilon = 0.1$

Why might global minimization be possible?

Consider an ϵ-regularized ℓ^p objective, restricted to the feasible plane:

$$\sum_{i=1}^{N} \left(x_i^2 + \epsilon\right)^{p/2}.$$

A moderate ϵ fills in the local minima.



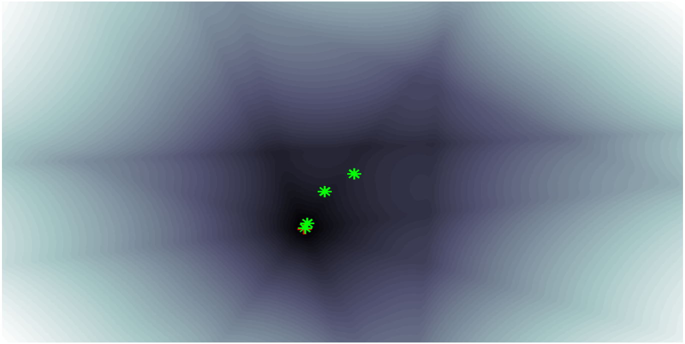
$\epsilon = 0.01$

Why might global minimization be possible?

Consider an ϵ-regularized ℓ^p objective, restricted to the feasible plane:

$$\sum_{i=1}^{N} (x_i^2 + \epsilon)^{p/2}.$$

A moderate ϵ fills in the local minima.



$\epsilon = 0.001$

Summary

- Compressive sensing allows images and signals to be recovered from many fewer measurements than previously thought possible.
- Nonconvex compressive sensing requires still fewer measurements.
- State-of-the-art convex optimization methods can be extended to the nonconvex case, giving excellent computational efficiency.
- Related matrix decomposition methods can extract interesting features from data without explicit modeling.
- New applications continue to emerge.

`math.lanl.gov/~rick`