Non-Convex Penalties and Applications in Data Science

Jack Xin

Department of Mathematics
UC Irvine

(Workshop on Comp. Modeling & Data Science, MSU, Oct 3, 2016, East Lansing, MI.)

Collaborators and Acknowledgement

Penghang Yin (UCLA)

Yifei Lou (University of Texas at Dallas)

Qi He (Chase Financial)

Yingyong Qi (UC Irvine)

Partially supported by the National Science Foundation.

Introduction

- Low dimensional structures in high dimensional problems arise in many settings of data science.
- Compressed sensing: reconstruct a sparse signal under a few linear measurements far less than the signal's physical dimension.
- Sparse representation in over-complete bases: human visual systems (V1) or Gabor frames, intrinsic dimension 20's.
- Human auditory systems (inner ear) or Gammatone filters, 24 critical (frequency) bands at perception level.
- Low rank matrix problems.
- Statistical and machine learning: sparse penalty to avoid overfitting and improve model generalization.

Mathematical Formulation

ℓ_0 minimization

$$\min_{x \in \mathbb{R}^N} \|x\|_0 \quad \text{s.t.} \quad A\,x = b$$

$b \in \mathbb{R}^M \setminus \{\mathbf{0}\}$ is the data, $A \in \mathbb{R}^{M \times N}$ with $M \ll N$ is of full rank.

Convex Relaxation of ℓ_0

Basis Pursuit (ℓ_1 minimization)

$$\min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{s.t.} \quad A\,x = b$$

Dual problem:

$$\max_{c \in \mathbb{R}^M} \; c \cdot b \quad \text{s.t.} \quad \|A^*\,c\|_\infty \leq 1.$$

Exact Recovery

Definition (mutual coherence)

The coherence of a matrix A is the max absolute value of the cross-correlations between its columns

$$\mu(A) = \max_{i \neq j} \frac{|A_i^{\mathrm{T}} A_j|}{\|A_i\|_2 \|A_j\|_2}.$$

Definition (Restricted Isometry Property, Candès -Tao 2005)

s-restricted isometry constant of A is the smallest $\delta_s \in (0, 1)$ *such that*

$$(1 - \delta_s)\|x\|_2^2 \leq \|A_T x\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$$

for all column index subsets T of size s and for all $x \in \mathbb{R}^s$*. A is said to satisfy the s-RIP with* δ_s*.*

Exact Recovery of Basis Pursuit

x_0 has sparsity s, $b := Ax_0$, then exact recovery is achieved via Basis Pursuit if (Donoho, Huo, Elad, 2001-2003)

$$s \leq \frac{1}{2}(1 + \frac{1}{\mu(A)})$$

or (Candès, Tao, 2005): $\delta_{3s} + 3\delta_{4s} < 2$.

Dual Certificate (iff condition):

for any s-sparse vector $v \in \mathbb{R}^N$, $|v_j| = 1$ (if $j \in \text{supp}(v)$), $v_j = 0$ (if $j \in$ else), there exists $c \in \mathbb{R}^M$ s.t. $(A^*c)_j = v_j$ ($j \in \text{supp}(v)$), $|(A^*c)_j| < 1$ ($j \in$ else).

Favorable conditions:

- Sparser x_0, larger M.
- Randomness of A: random Gaussian, random Bernoulli, random partial Fourier, ...

Non-convex Sparse Penalties

widely studied in statistics since the 1990's (Mazumder, Friedman, Hastie, JASA 2011): $P(x) = \sum_i p(|x_i|)$, interpolating ℓ_0 and ℓ_1 (LASSO, $p =$ identity),

- capped ℓ_1 (CL1), $p(t) = \min(t/\tau, 1)$, $\tau > 0$;
- transformed ℓ_1 (TL1), $p(t) = (a+1)\, t/(a+t)$, $a > 0$;
- SCAD, MCP: quadratic spline versions of TL1;
- Log penalty: $p = \frac{1}{\log(\gamma+1)} \log(\gamma t + 1)$, $\gamma > 0$;
- ℓ_q quasi-norm ($p = t^q, q \in (0,1)$, a.k.a Bridge).

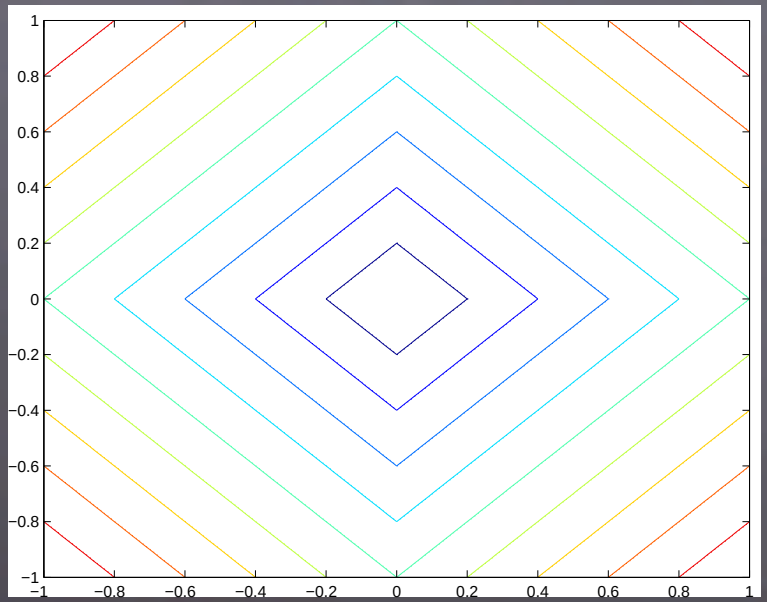
Note: p is continuous, concave, and increasing, $p'(0+) > 0$ is finite (also Bridge if smoothed, $t \to t + \epsilon$). P can be written as a difference of convex (DC) functions $P(x) = p'(0+)\|x\|_1 - g(x)$, with

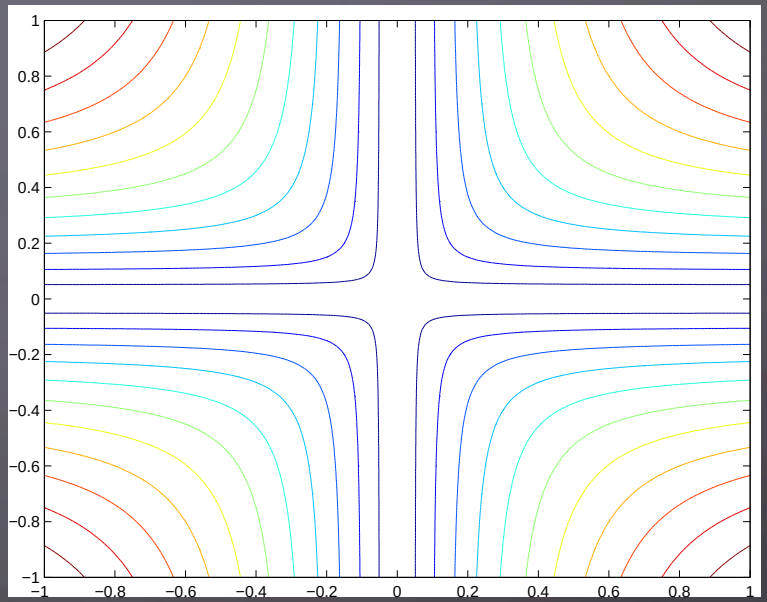
$$g = p'(0+)\|x\|_1 - P(x),$$

convex.

$\ell_1 - \ell_2$ Penalty

- $x \neq \mathbf{0}$, $\|x\|_1 - \|x\|_2 = 0 \iff \|x\|_0 = 1$ (exactly 1-sparse).
- Non-convex, non-separable, Lipschitz, non-C^1 only at $x = 0$;
- Naturally the Difference of Convex functions (DC)
- Level lines: ℓ_1 vs. $\ell_1 - \ell_2$





Connection with CL1

Capped-ℓ_1:

$$P(x) = \sum_j p(|x_j|) = \sum_j min(|x_j|/\tau, 1),$$

approaching l_0 as $\tau \to 0^+$.

Decompose:

$$p(x_j) = |x_j|/\tau - H(x_j),$$
$$H(x_j) = \max(1, |x_j|/\tau) - 1.$$

$\sum_j H(x_j)$ is a polygonal approximation of L2.

$\ell_1 - \ell_2$ Minimization

(SIAM J. Sci Computing, 2015)

$$\min_{x \in \mathbb{R}^N} \|x\|_1 - \|x\|_2 \quad \text{s.t.} \quad Ax = b$$

Theorem (Exact recovery of $\ell_1 - \ell_2$)

Let x_0 be any vector with sparsity of s satisfying $a(s) = \left(\frac{\sqrt{3s}-1}{\sqrt{s}+1}\right)^2 > 1$ (true if $s \geq 8$), and $b = Ax_0$, suppose A satisfies

$$\delta_{3s} + a(s)\delta_{4s} < a(s) - 1,$$

then x_0 is the unique solution. For $s \in [1, 7]$, modified $a(s)$ and RIP hold.

Nonconvexity helps: an example

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

The sparsest solution is $x_0 = [0, 1, 0]^{\mathrm{T}}$.

- $[t, 1 - 2t, t]^{\mathrm{T}}$ for any $t \in [0, 0.5]$ is a solution to ℓ_1 minimization.
- x_0 is the solution to $\ell_1 - \ell_2$ minimization since it is the only 1-sparse feasible vector.
- ℓ_p attains the minimum at $t = 0$ for $p < 1$ yielding the sparsest solution.

DC Algorithms (DCA)

Pham-Dinh Tao (1985), joint with Le-Thi Hoai An (since 1994).

Given $F(x)$ with D.C. decomposition:

$$F(x) = G(x) - H(x).$$

- Choose initial point x^0. Set $y^n \in \partial H(x^n)$ and iterate

$$x^{n+1} = \arg\min_{x \in \mathbb{R}^N} G(x) - H(x^n) - \langle x - x^n, y^n \rangle.$$

- $\{F(x^n)\}$ is monotonically decreasing.

DCA for the Unconstrained Problem

DC decomposition:

$$F(x) = (\frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1) - (\lambda\|x\|_2).$$

- Set $x^0 = y^0 = \mathbf{0}$, $n = 1$.
- WHILE not converged do

$$x^{n+1} = \arg\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 - \lambda\langle x, \frac{x^n}{\|x^n\|_2}\rangle$$

Convergence Results

Let $\{x^n\}$ be the sequence of iterates, then

- $\{x^n\}$ is bounded.
- $\|x^{n+1} - x^n\|_2 \to 0$.
- $\{x^n\}$ converges to a stationary point x^* satisfying the first-order optimality condition

$$\mathbf{0} \in A^{\mathrm{T}}(A\,x^* - b) + \lambda(\partial\|x^*\|_1 - \frac{x^*}{\|x^*\|_2}).$$

- Sparsity of local minimum x^* is upper bounded by $\mathrm{rank}(A)$.

Solving l_1 subproblems

Each DCA iteration is an ℓ_1 regularized subproblem

$$\min_{x\in\mathbb{R}^N} x^{\mathrm{T}}(\frac{1}{2}A^{\mathrm{T}}A + c)x + f^{\mathrm{T}}x + \lambda\|x\|_1 \Longleftrightarrow$$

$$\min_{x\in\mathbb{R}^N} x^{\mathrm{T}}(\frac{1}{2}A^{\mathrm{T}}A + c)x + f^{\mathrm{T}}x + \lambda\|z\|_1 \quad \text{s.t.} \quad x - z = 0$$

Define the augmented Lagrangian

$$L_\delta(x, z, p) := x^{\mathrm{T}}(\frac{1}{2}A^{\mathrm{T}}A+c)x+f^{\mathrm{T}}x+\lambda\|z\|_1+p^{\mathrm{T}}(x-z)+\frac{\delta}{2}\|x-z\|_2^2$$

Alternating Direction Method of Multipliers (ADMM)

Minimize L_δ with respect to x, minimize with respect to z and update the dual variable p.

- Define x^0, z^0 and p^0.
- WHILE not converged do

$$\begin{aligned} x^{n+1} &= (A^{\mathrm{T}}A + (\delta + 2c)I)^{-1}(-f + \delta z^n - p^n) \\ z^{n+1} &= \mathrm{shrink}(x^{n+1} + \frac{p^n}{\delta}, \frac{\lambda}{\delta}) \\ p^{n+1} &= p^n + \delta(x^{n+1} - z^{n+1}) \end{aligned}$$

Oversampled Partial DCT

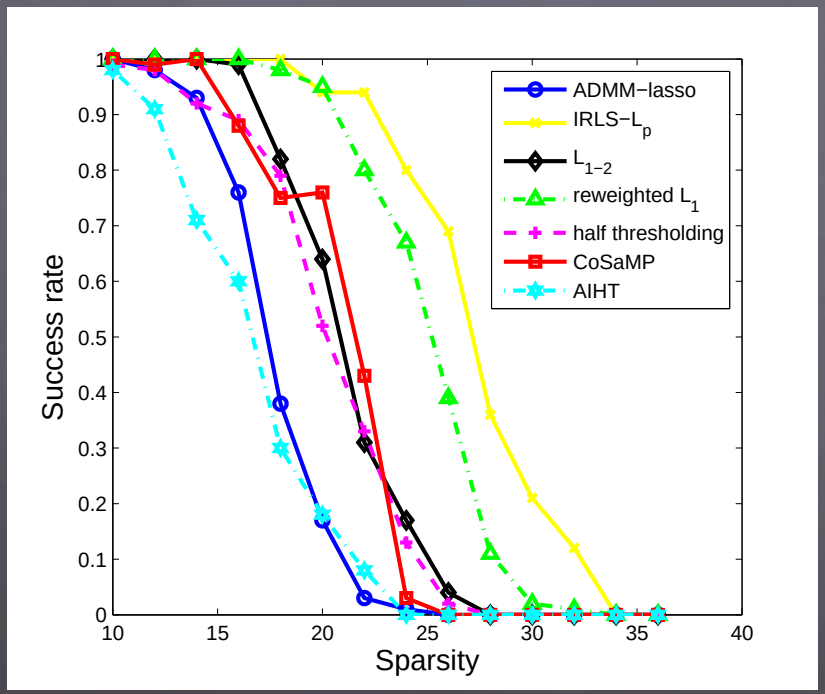
Oversampled partial DCT matrix

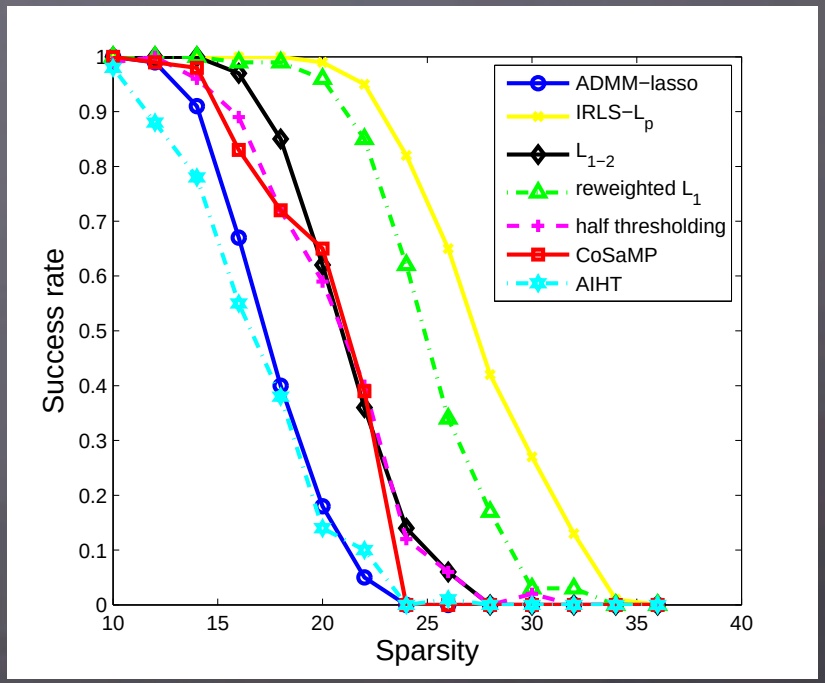
$$A_j = \frac{1}{\sqrt{M}} \cos(2j\pi\xi/F), \quad j = 1, \cdots, N$$

$\xi \in \mathbb{R}^M \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1]^M)$, $F \in \mathbb{N}$ is the refinement factor.

- A at $F = 1$ satisfies a RIP with high probability.
- $\mu(A) > 0.99$ when $F = 10$. $\mu(A) > 0.9999$ when $F = 20$.
- Minimum peak separation at least F (1 Rayleigh Length).

Methods Comparison on Gaussian and Partial DCT Matrices: Success Rate vs. Sparsity, $(m, n) = (64, 256)$, $s = 10 : 2 : 36$

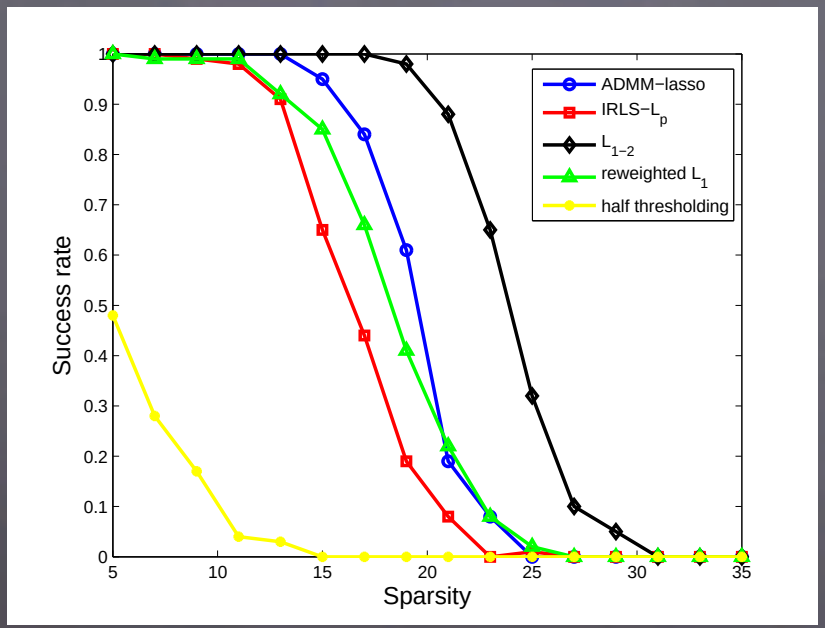




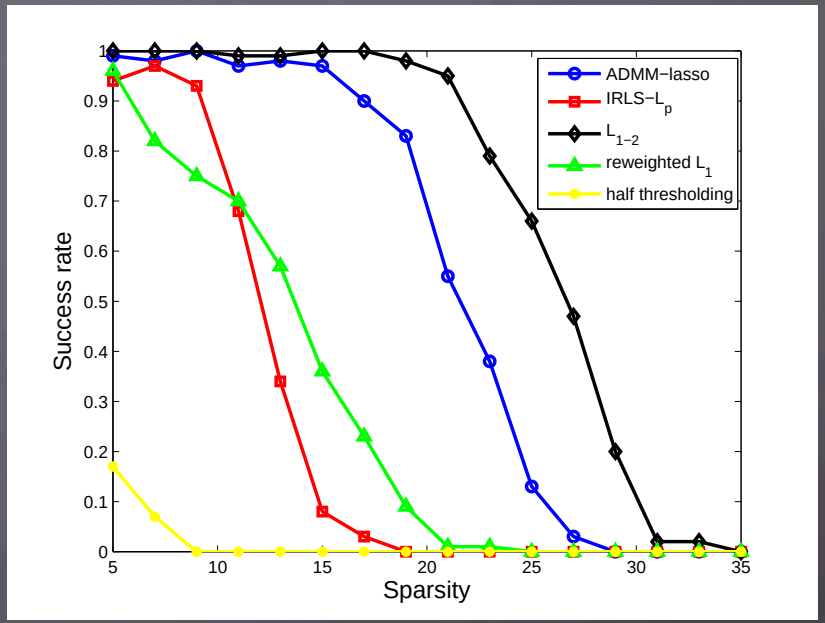
Success Rates vs. Sparsity

$M = 100, N = 2000$. Sparsity $= 5, 7, 9, \ldots, 35$. Minimum separation is $2F$.

F = 10



F = 20



MRI

Reconstruct Magnetic Resonance Imaging from a limited number of projections.

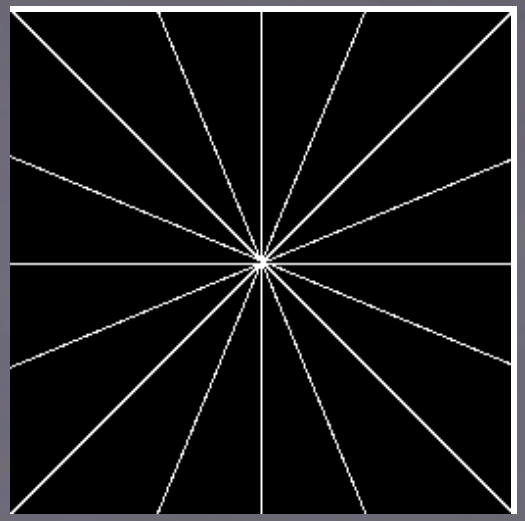
$$\min \sum_{(x,y)} |\partial_x u| + |\partial_y u| - \sqrt{(\partial_x u)^2 + (\partial_y u)^2} \quad \text{subject to } R\mathcal{F}u = b$$

$\mathcal{F}$ denotes the Fourier transform, R the sampling mask in the frequency space and b the data.

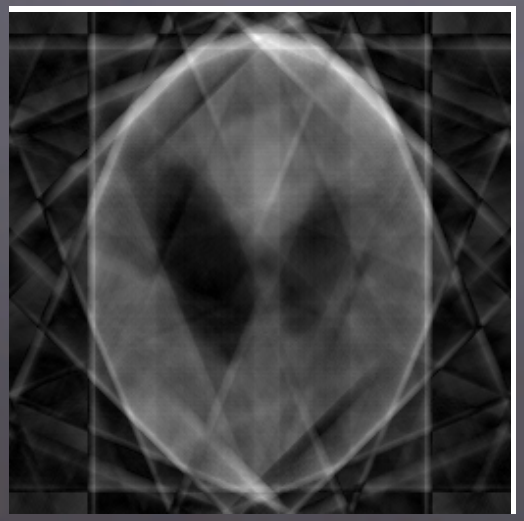
Apply DCA to the unconstrained problem

$$u^{n+1} = \arg\min_u \sum_{(x,y)} |\partial_x u| + |\partial_y u| - \frac{(\partial_x u, \partial_y u)^T (\partial_x u^n, \partial_y u^n)}{\sqrt{(\partial_x u^n)^2 + (\partial_y u^n)^2}} + \frac{\lambda}{2} \|R\mathcal{F}u - b\|_2^2$$

The 8 projections are sufficient for exact reconstruction.



FBP, ER=0.99



$\ell_1, ER = 0.1$

$\ell_1 - \ell_2, ER = 5 \times 10^{-8}$

Phase retrieval

Find $x \in \mathbb{C}^N$, s.t. $|Ax|^2 = b$.

- Lifting: $X = xx^* \in \mathbb{C}^{N\times N}$, $\mathcal{A}(X) = \mathrm{diag}(AXA^*)$

$$\min_{X\in\mathbb{C}^{N\times N}} \|\mathcal{A}(X) - b\|_2^2 \quad \text{s.t.} \quad X \succeq 0,\ \mathrm{rank}(X) = 1.$$

- Find low-rank semi-definite matrix by trace $\mathrm{Tr}(X)$ penalty.
- If A is Gaussian, $M = O(N)$, then trace penalty recovers rank-1 solution with high probability (PhaseLift, Candès et al, 2011-2013).

PhaseLiftOff (Comm Math Sci, 2015):

$$\min_{X\in\mathbb{C}^{N\times N}} \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 + \lambda(\mathrm{Tr}(X) - \|X\|_F) \quad \text{s.t.} \quad X \succeq 0.$$

Equivalence

$\mathrm{Tr}(X) - \|X\|_F$ characterizes rank-1:

$$X \neq 0,\ X \succeq 0,\ \mathrm{Tr}(X) - \|X\|_F = 0 \Leftrightarrow \mathrm{rank}(X) = 1$$

Theorem (Yin-X, Comm Math Sci, 2015)

Let $\mathcal{A}$ be an arbitrary linear operator, and let e be the additive noise. If $\|b\|_2 > \|e\|_2$ and $\lambda > \frac{\|\mathcal{A}\|\|e\|_2}{\sqrt{2}-1}$, then PhaseLiftOff is equivalent to phase retrieval with lifting. In particular, the equivalence holds for all $\lambda > 0$ in the noiseless case.

Algorithm (DCA)

$$X^{k+1} = \begin{cases} \arg\min \ \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 + & \lambda \mathrm{Tr}(X) \quad \text{s.t.} \quad X \succeq 0 \\ & \text{if} \quad X^k = 0, \\ \arg\min \ \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 + & \lambda \langle X, I_N - \frac{X^k}{\|X^k\|_F} \rangle \quad \text{s.t.} \quad X \succeq 0 \\ & \text{otherwise.} \end{cases}$$

Subproblems can be solved by ADMM.

Exact recovery

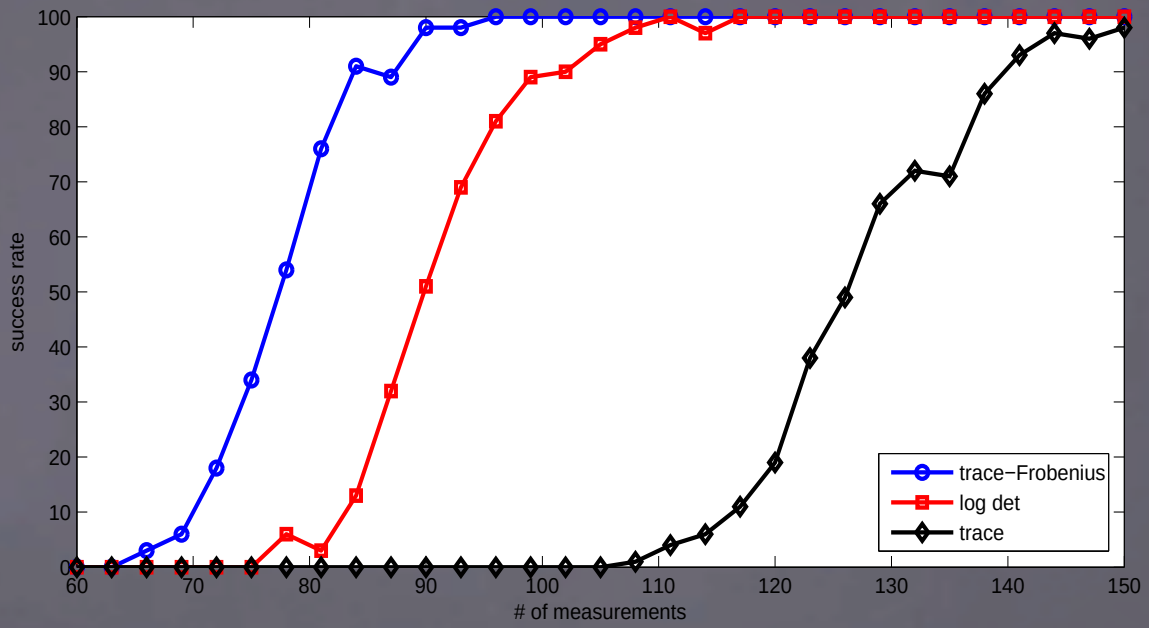
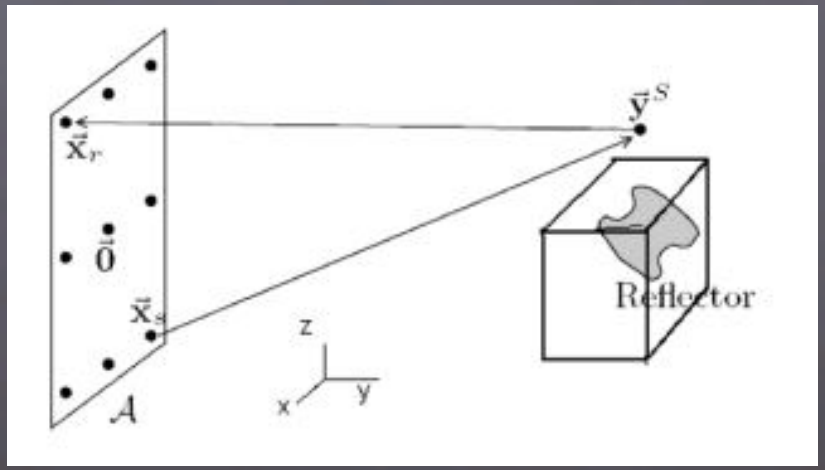


Figure: success rate v.s. number of measurements with parameters $N = 32$, $M = 60, 63, \ldots, 150$. A reconstruction is considered as a success if the relative error $< 10^{-3}$.

$M = 3N$ is sufficient for PhaseLiftOff to guarantee exact recovery with high probability under Gaussian measurements.

Coherent Array Optical Imaging

- Task: determine the locations and reflectivities of distributed reflectors by sending probing signals from the array and recording the backscattered fields with no phase information.
- n mesh points $\mathbf{y_i}$, $i = 1, \ldots, n$. Reflectivity vector $\mathbf{r} \in \mathbb{R}^n$ with unknown sparsity $k \ll n$ corresponding to k scatterers. m transducers located at $\mathbf{x_j}$, $j = 1, \ldots, m$.



Mathematical Model

$$\text{find} \quad \mathbf{r} \quad \text{s.t.} \quad |A(\omega)\mathbf{r}|^2 = \mathbf{b}(\omega), \ \mathbf{r} \geq 0.$$

- $\mathbf{b}$ is the response data. When ignoring multiple scattering, $A(\omega)\mathbf{r}$ has the approximation

$$(\sum_{i=1}^{n} r_i \mathbf{g}(\mathbf{y_i}, \omega)\mathbf{g}^{\mathrm{T}}(\mathbf{y_i}, \omega))\mathbf{f}(\omega),$$

$\mathbf{f}(\omega)$ is the illumination vector, and:

$$\mathbf{g}(\mathbf{y}, \omega) = (\frac{e^{i\kappa|\mathbf{x_1}-\mathbf{y}|}}{4\pi|\mathbf{x_1}-\mathbf{y}|}, \ldots, \frac{e^{i\kappa|\mathbf{x_m}-\mathbf{y}|}}{4\pi|\mathbf{x_m}-\mathbf{y}|})^{\mathrm{T}}, \quad \kappa = \frac{\omega}{c}$$

is the Green's function vector.

PhaseLift for Array Imaging

Chai, Moscoso, Papanicolaou, 2011:

- Lifting $X = \mathbf{r}\mathbf{r}^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, X rank-1.
- Solve relaxed convex problem:

$$\min_{X\in\mathbb{R}^{n\times n}} \mathrm{Tr}(X) \quad \text{s.t.} \quad \mathcal{A}(X) = \mathbf{b}(\omega),\ X \geq 0,\ X \succeq 0.$$

  for some linear operator $\mathcal{A}$ determined by A.

- Equivalence under RIP verified in the presence of strong scale separations.

Compressive PhaseLift

$$\min_{X \in \mathbb{R}^{n \times n}} \|X\|_1 \quad \text{s.t.} \quad \mathcal{A}(X) = \mathbf{b}(\omega),\ X \geq 0,\ \mathrm{rank}(X) = 1,\ X \succeq 0.$$

Compressive PhaseLift (Li, Voroninski, 2013):

$$\min_{X \in \mathbb{R}^{n \times n}} \|X\|_1 + \lambda \,\mathrm{Tr}(X) \quad \text{s.t.} \quad \mathcal{A}(X) = \mathbf{b}(\omega),\ X \geq 0,\ X \succeq 0.$$

A convex relaxation, may not be optimal. The solution is usually not rank-1.

Compressive PhaseLiftOff

$$\min_{X \in \mathbb{R}^{n \times n}} \|X\|_1 \quad \text{s.t.} \quad \mathcal{A}(X) = \mathbf{b}(\omega),\ X \geq 0,\ \mathrm{rank}(X) = 1,\ X \succeq 0.$$

- Compressive PhaseLiftOff:

$$\min_{X \in \mathbb{R}^{n \times n}} \|X\|_1 + \lambda(\mathrm{Tr}(X) - \|X\|_{\mathrm{F}}) \text{ s.t. } \mathcal{A}(X) = \mathbf{b}(\omega),\ X \geq 0,\ X \succeq 0.$$

- The two problems are equivalent as long as λ is sufficiently large. No RIP or scale separation is necessary.

Computation

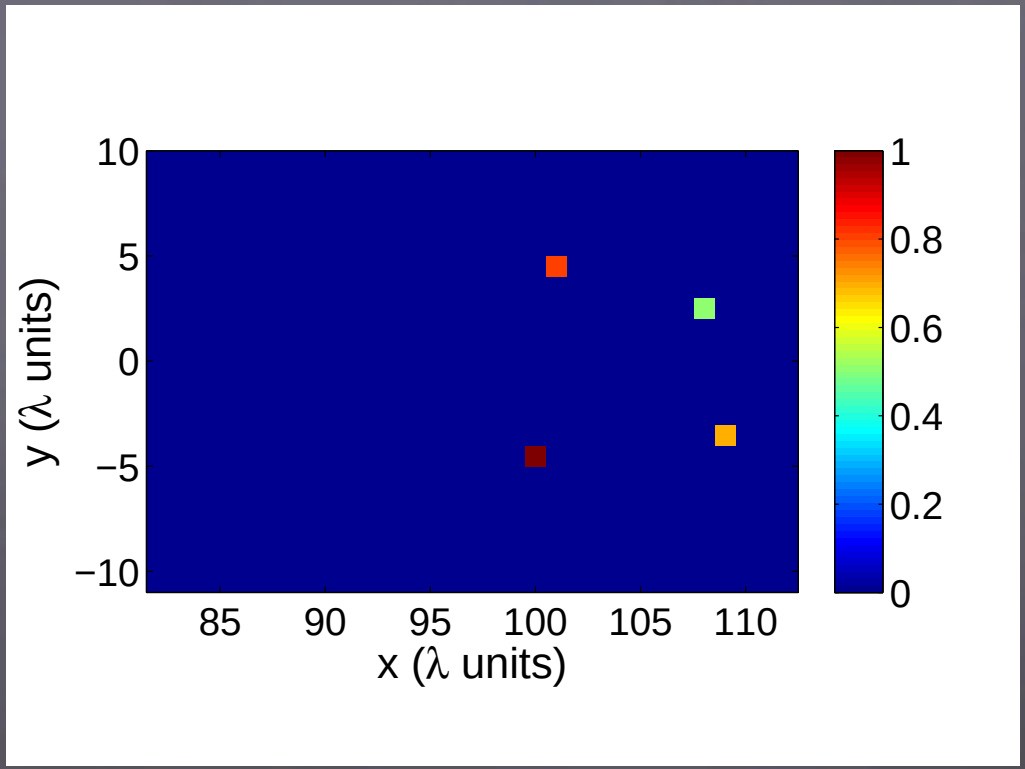
$$\min_{X\in\mathbb{R}^{n\times n}} \Phi(\mathcal{A}(X)-\mathbf{b}) + \lambda(\mathrm{Tr}(X) - \|X\|_{\mathrm{F}}) + \gamma\|X\|_1$$
$$\text{subject to} \quad X \geq 0,\ X \succeq 0.$$

- $\Phi(\mathcal{A}(X)-\mathbf{b})$ measures the data fitting, being convex, e.g., $\frac{1}{2}\|\mathcal{A}(X)-\mathbf{b}\|^2$.
- DCA: linearize $\lambda\|X\|_{\mathrm{F}}$ at the k-th step and solve the convex subproblem:

$$X^{k+1} = \min_{X\in\mathbb{R}^{n\times n}} \Phi(\mathcal{A}(X)-\mathbf{b}) + \lambda\mathrm{Tr}(X) + \gamma\|X\|_1 - \lambda\langle X, \frac{X^k}{\|X^k\|_{\mathrm{F}}}\rangle$$
$$\text{s.t.} \quad X \geq 0,\ X \succeq 0.$$

Numerical experiments

4 scatterers of reflectivity 1.0, 0.8, 0.7 and 0.5, within an image window of size 10×10. 21 transducers with the aperture 100.



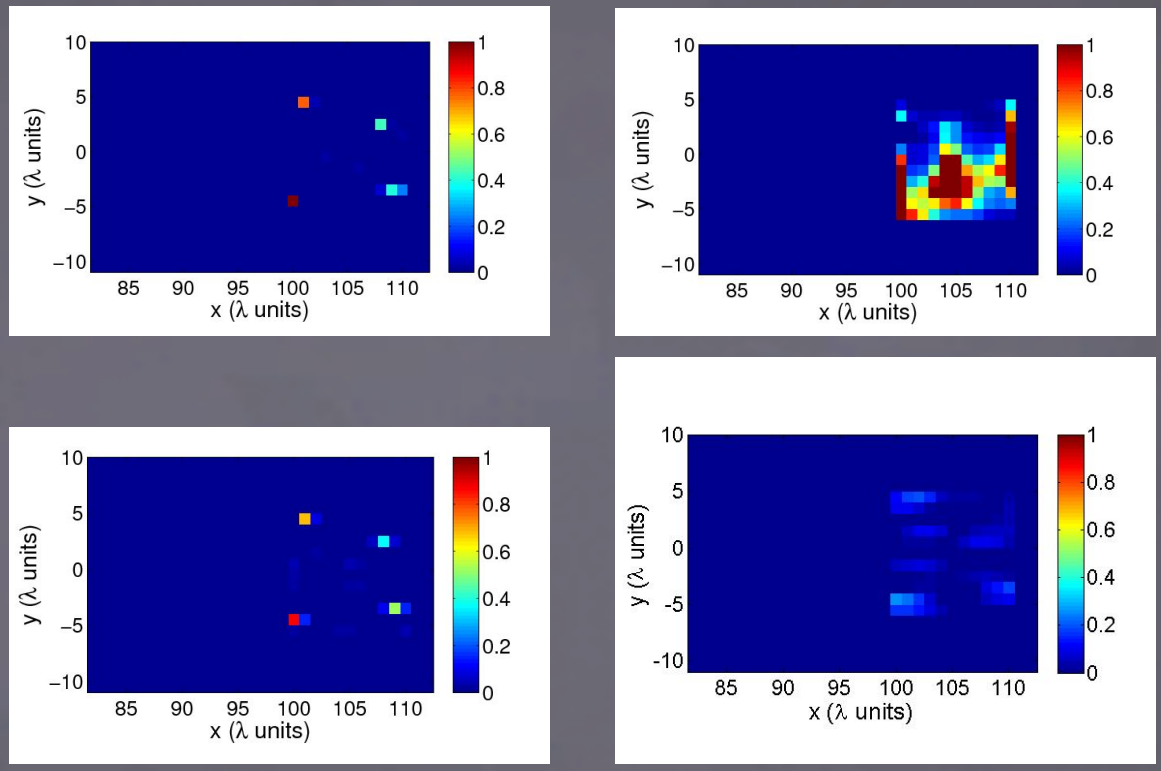
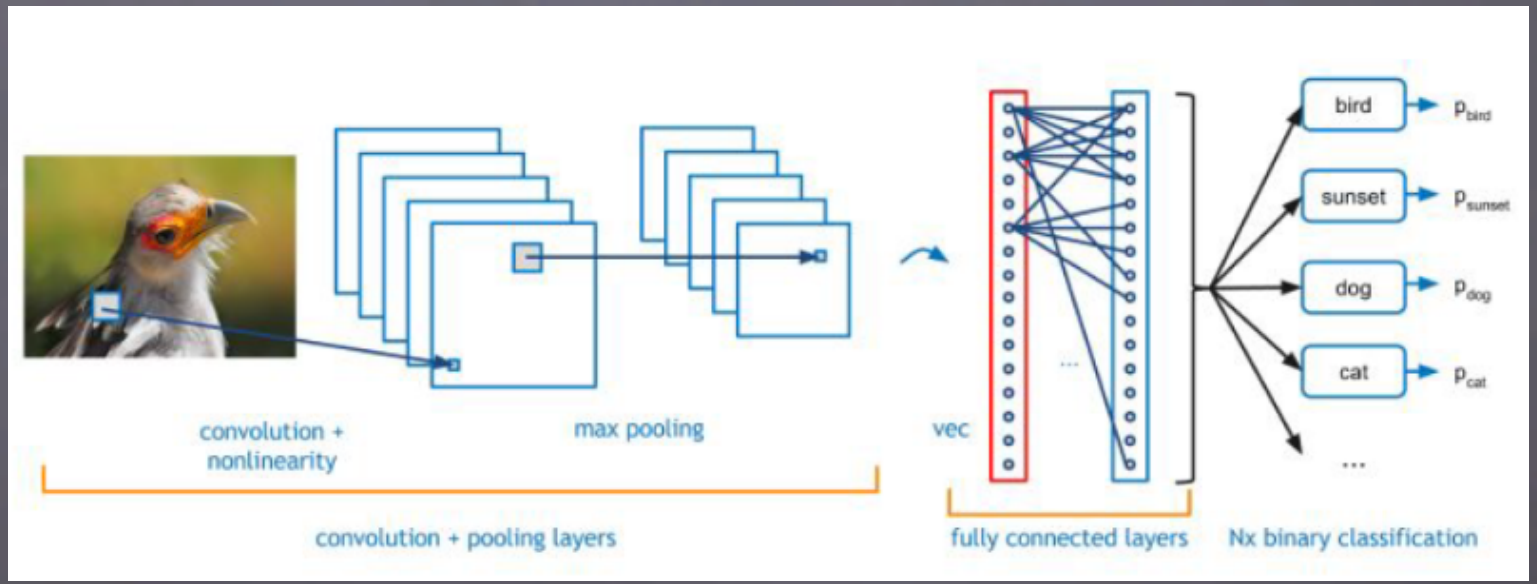


Figure: Adding 5% noise to the measurements. Top left: Compressive PhaseLiftOff, 2 illuminations. Top right: PhaseLiftOff without sparsity, 2 illuminations. Bottom left: PhaseLiftOff without sparsity, 3 illuminations. Bottom right: PhaseLift without sparsity, 10 illuminations.

Linear Feature Transform (LFT)

Multiscale features arise in deep neural networks. Let Y_i be the feature matrix (subspace) of training data for the i-th class, Y the concatenation of all Y_i's, $i = 1, 2, \cdots, c$.



Goal: Learn a linear transform T to preserve the low-rank structure for data within the same subspace, and introduce a maximally separated structure for data from different subspaces.

Non-Convex Nuclear Norm Models

- (Qiu, Sapiro, JMLR 2015)

$$T^* = \text{argmin}_T \sum_{i=1}^{c} \|TY_i\|_* - \|TY\|_* \quad \text{s.t.} \quad \|T\|_2 = 1.$$

  however, rank(T^*) may degenerate to below c.

- (Yin,X,Qi,2016) weighted, convex regularized model:

$$T^* = \text{argmin}_T \sum_{i=1}^{c} \|TY_i\|_* - w\|TY\|_* + \frac{\lambda}{2}\|T - P\|_F^2$$

  $w > 1$, P by PCA if T reduces dimension, else $P =$ identity.

- DCA on the weighted and convex regularized model is descending and convergent.

Ten Class Image Classification: CIFAR-10

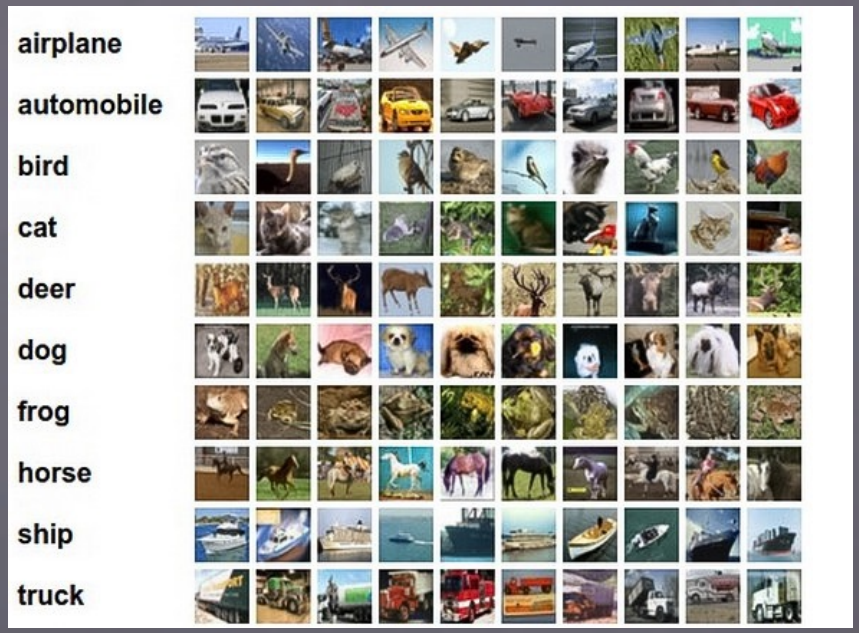


Table: Accuracy (%) at reduced feature dim down from 1024.

Reduced dim	PCA	PCA + LFT
64	80.21	80.90
32	77.91	79.95

MNIST

Table: Accuracy (%) at reduced feature dim down from 500.

Reduced dim	PCA	PCA + LFT
16	98.14	98.33
8	95.31	97.1

Conclusions

- $\ell_1 - \ell_2$ is an effective sparsity promoting metric for highly coherent dictionaries (other non-convex penalties similar if put in DC form), consistently improving ℓ_1.
- The difference of nuclear and Frobenius norms is more effective for rank-1 minimization than nuclear norm.
- Applications in MRI, array imaging, and image classification show merits of non-convex ℓ_1 based methods.
- Study fine properties of stationary points of DCA, and explore other problems of data science in future work.