Message Passing Algorithms for Compressed Sensing

DLD, Arian Maleki, Andrea Montanari

Stanford

September 2, 2009

Outline

Compressed Sensing – the heuristic

- Real images and signals are **compressible**
- Equivalently: Few large coefficients, eg in Wavelet basis
- Fewer than nominal degrees of freedom
  not 10^6 pixels, just 10^4 wavelet coeffs, $+$ positions of those coefficients
- Standard sampling: 10^6 measurements
- "Morally" $c \cdot 10^4$ measurements should suffice

Compressed Sensing MRI

- "Sparse MRI – The Application of Compressed Sensing in Magnetic Resonance Imaging" – Michael Lustig, DLD, John Pauly 2007, *Magnetic Resonance in Medicine*
- "Compressed Sensing MRI" – Michael Lustig, John Pauly, Juan Santos, DLD *IEEE Signal Processing* Special Issue on CS, March 2008
- Inspired by CS theory
  - Rapid Contrast-Enhanced 3D Angiography,
  - Whole-Heart Coronary Imaging,
  - Brain Imaging, and
  - Dynamic Heart Imaging.

A Theoretical Formalization of Compressed Sensing

Ingredients

- *Sparsity.*
  An object $x_0 \in \mathbf{R}^N$ with $k \ll N$ nonzero coefficients in a fixed basis
- *Random Undersampled Measurements*
  Measure $y = Ax_0$ with random n by N matrix A (eg iid Gaussian).
- *Nonlinear Reconstruction* Attempt reconstruction with x_1 solving

  $$(P_1) \qquad \min \|x\|_1 \text{ subject to } y = Ax$$

  AKA: Minimum ℓ_1, Basis Pursuit.
- Computationally Feasible; compare NP-Hard:

  $$(P_0) \qquad \min \|x\|_0 \text{ subject to } y = Ax$$

- *Surprise:* often in the $n < N$ underdetermined case (P_0) and (P_1) will have the same, unique solution.
- Voluminous literature IEEE 2001-today.

Phase Transition for $\pm$

Theorem. *There is a function $\rho_{CG}(\delta, \pm)$ with the following characteristics: Fix $\delta > 0$.*
If $n/N > \rho_{CG}(k/n; \pm)(1+\delta)$ then with overwhelming probabilty for large n, N,

$$x_1 = x_0$$

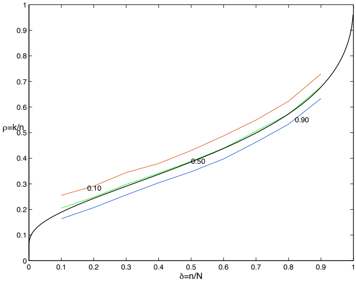
If $n/N < \rho_{CG}(k/n, \pm)(1-\delta)$, then with overwhelming probabilty for large n, N,

$$x_1 \neq x_0.$$

DLD (Discr. & Comput. Geom., 2006);
DLD and Jared Tanner (JAMS, 2009)
fully rigorous, explicit calculation of ρ_{CG} using special functions.
Methods: Combinatorial Geometry [Affentranger and Schneider (1992), Vershik and Sporyshev (1992)].

Empirical Results: $\pm$

Random Matrix A, $\delta = k/n$, $\rho = n/N$, $N = 400$.



Nonnegative coefficients

- Underdetermined system of equations:

$$y = Ax, \qquad x \geq 0.$$

- Sparsest Solution:

$$(NP_+) \qquad x_0 = \operatorname{argmin} \|x\|_0 \text{ s.t. } y = Ax, \quad x \geq 0$$

- Problem: NP-hard in general.
- Relaxation:

$$(LP_+) \qquad x_1 = \operatorname{argmin} 1'x \text{ s.t. } y = Ax, \quad x \geq 0.$$

Standard linear program

Phase Transition for $+$

Theorem *There is a function* $\rho_{CG}(\cdot,+) : (0,1] \mapsto (0,1]$ *with the following characteristics: Fix* $\epsilon > 0$. *If* $n/N > \rho_{CG}(k/n,+)(1+\epsilon)$ *then with overwhelming probabilty for large* n, N,

$$x_1 = x_0$$

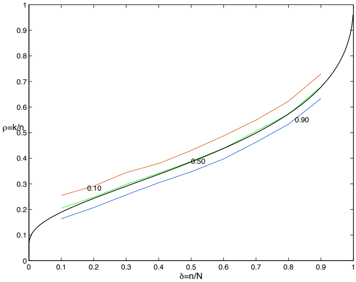
If $n/N < \rho_{CG}(k/n,+)(1-\epsilon)$, *then*

$$x_1 \neq x_0.$$

DLD and Tanner (PNAS, 2005[a,b]) *fully rigorous*, explicit calculation of ρ_{CG} using special functions. Methods: Combinatorial Geometry [Affentranger and Schneider (1992), Vershik and Sporyshev (1992)].

Empirical Results +

Random Matrix A, $\delta = k/n$, $\rho = n/N$. $N = 400$.



Object with k-Simple, Bounded Coefficients

- Underdetermined system of equations:

$$y = Ax, \qquad -1 \leq x(i) \leq 1, \quad 1 \leq i \leq N.$$

- Simplicity: $\#\{i : |x(i)| \neq 1\}$ small.
- Simplest Solution:

$$(NP_{\square}) \qquad x_0 = \text{argmin}\#\{i : |x(i)| \neq 1\} \text{ s.t. } y = Ax, \quad x \geq 0$$

  Problem: NP-hard in general.
- Relaxation:

$$(LP_{\square}) \qquad x_1 = \text{argmin}\, 1'x \text{ s.t. } y = Ax, \quad x(i) \in [-1, 1]$$

  Standard linear program (feasibility problem)

Phase Transition for $\Box$

Theorem. *Set* $\rho_{CG}(\delta, \Box) = (2 - \delta^{-1})_+$. *Fix* $\epsilon > 0$.
If $n/N > \rho_{CG}(k/n, \Box)(1 + \epsilon)$ *then, with overwhelming probabilty for large* n, N,

$$x_1 = x_0$$

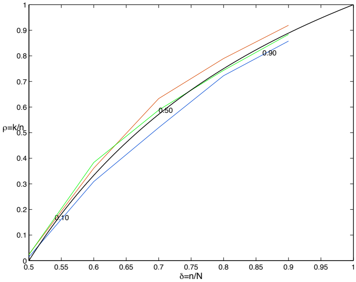
If $n/N < \rho_{CG}(k/n, \Box)(1 - \epsilon)$, *then , with overwhelming probabilty for large* n, N,

$$x_1 \neq x_0.$$

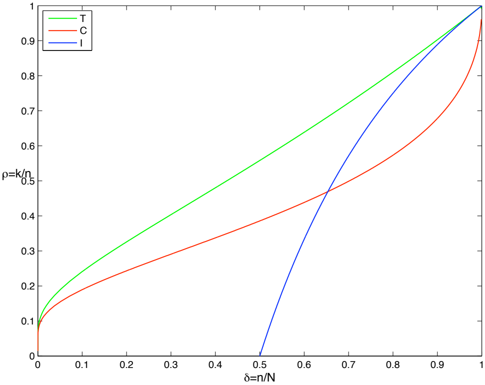
DLD and Jared Tanner (2008, in press D&CG)
Exact finite n identities in Geometric Probability
Methods: Wendel's Theorem, Winder's Theorem, Oriented Matroids.

Empirical Results □

Random Matrix A, $\delta = k/n$, $\rho = n/N$. $N = 400$.



The Three Theoretical PT's



Fast Iterative Algorithms

- Problem: interesting problem sizes for MRI application: $N \approx 10^6$, $n \approx 10^4$.
- Generic LP at least $n^2 \cdot N$ complexity.
- Alternative: iterative algorithms involving C applications $10 < C < 50$ of A and A^* to appropriate vectors
- Complexity: CnN in general eg A iid Gaussian;
- Complexity: $CN \log(N)$ for FFT-based matrices.
- Heavily used for large-scale applications: Jean-Luc Starck (2003-), Miki Elad (2004-), ...

Simple Iterative Algorithms

Assume columns of A have unit length.

$$x^{t+1} = \eta_t(A^* z^t + x^t)\,, \tag{1}$$
$$z^t = y - Ax^t. \tag{2}$$

- Thresholding $\eta_t(\,\cdot\,) = \eta(\cdot;\lambda\sigma_t,\chi)$
  - λ is a threshold control parameter;
  - $\chi \in \{+,\pm,\square\}$ controls nonlinearity shape.
  - $\sigma_t^2 = \mathrm{Ave}_j \mathbb{E}\{(x^t(j) - x_0(j))^2\}$ is MSE
- "Iterative Soft Thresholding".
- Only apply A and A^*; no need for matrix A itself.

Thresholding Functions – scalar y, $t > 0$:

- Case $+$

$$\eta_t(y;+) = \begin{cases} y - t & y > t \\ 0 & y < t \end{cases}$$

- Case $\pm$

$$\eta_t(y;\pm) = \begin{cases} y - t & y > t \\ 0 & |y| < t \\ y + t & y < -t \end{cases}$$

- Case $\Box$

$$\eta_t(y;\Box) = \begin{cases} t & y > t \\ y & |y| < t \\ -t & y < -t \end{cases}$$

- Connection to optimization:

$$\eta_t(y;\pm) = \mathrm{argmin}_x (y - x)^2/2 + \lambda|y|.$$

- 'Shrinkage towards 0'. 'Soft Thresholding' . Used in 'Wavelet Shrinkage'

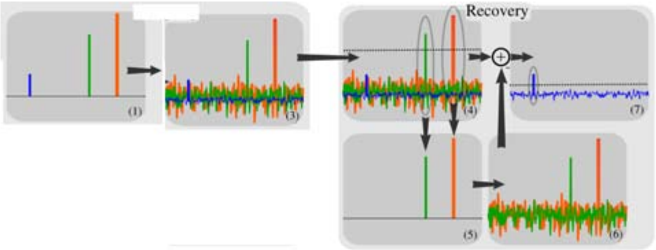
Can Iterative Thresholding Algorithms Work?

- Easy: A orthogonal: solves in one iteration.
- Hard: A nonsingular: with care, asymptotically solves LP Daubechies, De Mol, Bertero.
- Weird: A, $n < N$.

Mutual Access Interference (MAI) Heuristic

- Assume columns of A normalized to unit length.
- Set $H = A^*A - I$; call Hx_0 the MAI (aka cross-channel interference)
- $A^*y = x_0 + Hx_0 = x_0 + MAI$
- For 'typical' vector v, 'random' matrix A, Hv has marginal distribution that's $N(0, \|v\|^2)$.
- $A^*y = x_0 + Noise$.
- First Iteration: $x^1 = \eta_t(A^*z^t + x^0)$
- Shrinkage heuristic
  - Shrinkage kills 'small elements'; keeps 'large elements'
  - Small elements mostly noise; large elements mostly signal.
  - $\|x1 - x0\|_2^2 \ll \|0 - x_0\|_2^2$: off to the races!

Sketch of MAI/Thresholding Heuristics



Lustig DLD Pauly IEE Signal Processing Magazine 2008

State Evolution (SE) Definition

MSE Map:

$$\Psi(\sigma^2) \equiv \mathbb{E}\Big\{\big[\eta\big(X + \frac{\sigma}{\sqrt{\delta}} Z; \lambda\sigma\big) - X\big]^2\Big\}. \tag{3}$$

State Evolution

- *Implicit* parameters $(\chi, \delta, \rho, \lambda, F)$, $F = F_X$ marginal dist. rv X.
- *Explicit* paramater σ^2.
- *State Evolution*

  $$\sigma_t^2 \mapsto \Psi(\sigma_t^2) \equiv \sigma_{t+1}^2$$

  (Implicit parameters $(\chi, \delta, \rho, \lambda, F)$ fixed)
- Full state evolves as

$$(\sigma_t^2; \chi, \delta, \rho, \lambda, F_X) \mapsto (\Psi(\sigma_t^2); \chi, \delta, \rho, \lambda, F_X).$$

State Evolution (SE) Asymptotics

Two Regions of parameter space:

Region (I) $\Psi(\sigma^2) < \sigma^2$ for all $\sigma^2 \in (0, \mathbb{E}X^2]$.
Here $\sigma_t^2 \to 0$ as $t \to \infty$: SE evolves to $\sigma^2 = 0$.

Region (II) The complement of Region (I).
SE does *not* evolve to $\sigma^2 = 0$.

State Evolution Phase Boundary:

$$\rho_{\mathrm{SE}}(\delta; \chi, \lambda, F_X) \equiv \sup \{\rho \,:\, (\delta, \rho, \lambda, F_X) \in \text{Region (I)}\} \,. \qquad (4)$$

Independence from Object Distribution

Theorem *For every distribution $F = F_X$ with no atom at zero and finite second moment $EX^2 < \infty$,*

$$\rho_{\mathrm{SE}}(\delta; \chi, \lambda, F) = \rho_{\mathrm{SE}}(\delta; \chi, \lambda).$$

Conclude:

- δ, ρ are 'what matter' to SE.
- Graph of ρ_{SE}: boundary between *phases* of (δ, ρ) phase diagram.

Computational Geometry-State Evolution Agreement

Finding

For the three canonical problems $\chi \in \{+, \pm, \square\}$*:*

$$\rho_{\mathrm{SE}}(\delta; \chi) = \rho_{\mathrm{CG}}(\delta; \chi) \qquad \forall\, \delta \in (0, 1) \tag{5}$$

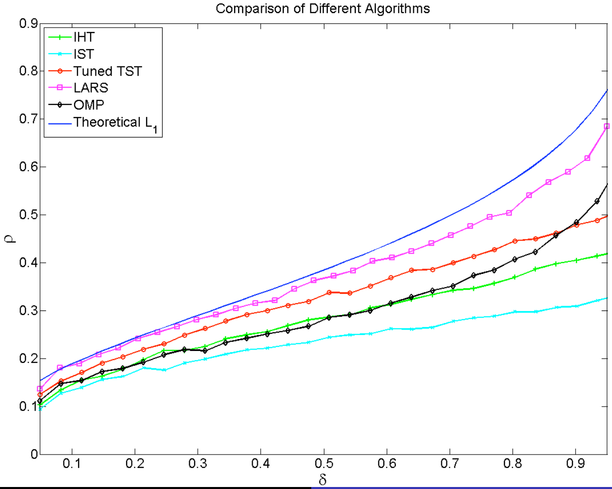
- Numerically agree to high accuracy.
- Rigorously proven equality $\chi = \square$.
- Rigorously proven asymp. equality $\chi \in \{+, \pm\}$ in limit $\delta \to 0$.

Large Scale Testing of Iterative Algorithms

Maleki and DLD (2009, submitted, IEEE Select. Topics in SP)

- Algorithms:
  IST, IHT, CoSamp, Subspace Pursuit , ...
- Matrix Ensembles A:
  Gaussian, Bernoulli, partial Fourier, partial Hadamard, ...
- Object Distributions F:
  Constant, Double Exponential, Uniform, Power Law, ...
- Problem sizes: $N = 800$ and 2^ℓ, $\ell = 10, 12, 14$.
- Goal: measure phase transitions; tune algorithms to optimize transitions.
- 3.8 CPU *years*

Empirical Results



Message Passing Iterative Thresholding

Message Passing successful in many other problems: (examples Gallagher LDPC, Yedidia, Wainwright, Willsky , ...).
Adapt to iterative Thresholding.
N thresholding reconstructions:

$$x_{i\to a}^{t+1} = \eta_t\Big(\sum_{b\in[n]\setminus a} A_{bi} z_{b\to i}^t\Big), \tag{6}$$

N residuals:

$$z_{a\to i}^t = y_a - \sum_{j\in[N]\setminus i} A_{aj} x_{j\to a}^t, \tag{7}$$

for each $(i,a)\in[N]\times[n]$.
Problem: N iterative algorithms in semi-parallel means N times as much work as simple iterative algorithms.

Approximate Message Passing (AMP)

First order Approximate Message Passing (AMP) algorithm

$$x^{t+1} = \eta_t(A^* z^t + x^t)\,, \tag{8}$$

$$z^t = y - Ax^t + \frac{1}{\delta} z^{t-1} \langle \eta'_{t-1}(A^* z^{t-1} + x^{t-1}) \rangle\,. \tag{9}$$

$\eta'_t(\,s\,) = \frac{\partial}{\partial s}\eta_t(\,s\,)$.
Feature: Essentially same cost as Iterative Soft Thresholding.

SE is Correct for AMP

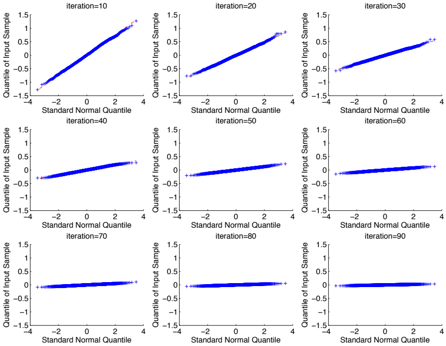
Finding

For the AMP algorithm, and large dimensions N, n, we observe:

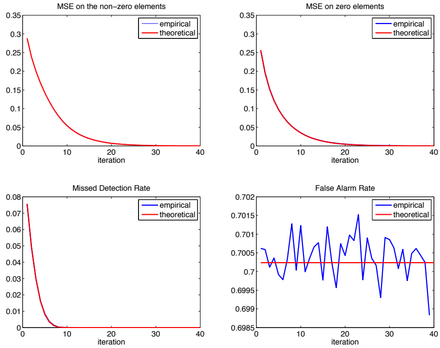
I. SE correctly predicts evolution of numerous statistical properties of x^t with t. The MSE, the number of nonzeros in x^t, the number of false alarms, the number of missed detections, and several other measures all evolve in way consistent with the state evolution formalism to within experimental accuracy.

II. SE correctly predicts the success/failure to converge to the correct result. In particular, SE predicts no convergence when $\rho > \rho_{\rm SE}(\delta; \chi, \lambda)$, and convergence if $\rho < \rho_{\rm SE}(\delta; \chi, \lambda)$. This is indeed observed empirically.

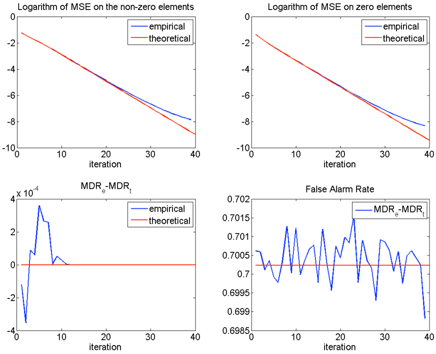
Finding 1.a MAI Heuristic is Correct for AMP



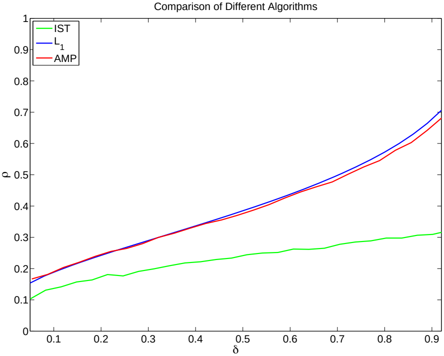
Finding 1.b State Evolution is Correct for AMP Observables, 1



Finding 1.b State Evolution is Correct for AMP Observables, 2



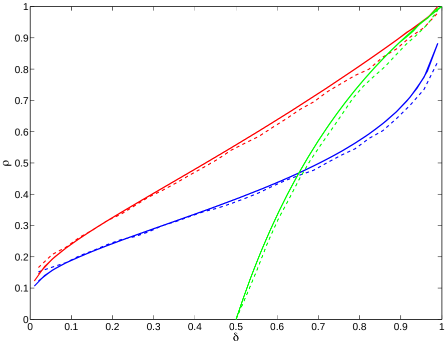
Finding 2. SE Transition agrees/w AMP Transition, 1



Random Matrix A, $\delta = k/n$, $\rho = n/N$.

$N = 1000$; Blue: CG; Red:AMP; Green: IST.

Finding 2. SE Transition agrees/w AMP Transition, 2



Random Matrix A, $\delta = k/n$, $\rho = n/N$.

Dashed Empirical; Solid Asymptotic; $N = 1000$; Red: $+$; Blue: $\pm$; Green: $\Box$.

Universality

- Same experiment, different coefficient distributions
  - Constant amplitude nonzeros
  - Power Law
  - Gaussian
- Same experiment, with different matrix ensembles
  - Bernoulli: Fair Coin Tossing ± 1.
  - Partial Fourier: sample n rows from $N \times N$ Fourier
  - Partial Hadamard: sample n rows from $N \times N$ Fourier
  - Gaussian: $A_{i,j} \sim N(0,1)$
  - Uniform Random Projection: A uniform on Haar measure for orthoprojectors

Empirically: same behavior. Matches findings for LP optimization in Tanner and DLD (2009, In Press, Phil Trans Roy Soc A)

Conclusions

- Phase Transitions for LP: rigorous foundation in Combinatorial Geometry
- Standard Iterative Algorithms heuristically like LP but not CG PT's
- AMP Algorithm slight modification of SIA, does achieve CG PT's
- State Evolution explains properties of AMP
- SE generates new formulas for PTs, for optimal thresholds in algorithms.