Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion

Jared Tanner

SPARS 2015
6^{th} July 2015

University of Oxford[1]
Joint with Blanchard & Wei

[1]Supported by: EPSRC, NVIDIA, & SELEX-Galileo

Algorithms for standard CS and Matrix Completion

Standard CS and MC recovery

- CS: Recover $x \in \mathbb{R}^n$ from $y = Ax$ for $y \in \mathbb{R}^m$ with $m \ll n$
- MC: Recover $X \in \mathbb{R}^{m \times n}$ from $y = \mathcal{A}(X) \in \mathbb{R}^p$ with $p \ll mn$ has seen rapid advances with numerous computationally efficient algorithms understood from different perspectives:
  - convex relaxations,
  - matching pursuits,
  - iterative hard thresholding,
  - approximate message passing,
  - reweighted least squares,
  - non-convex optimization techniques such as trust-region,
  - expander methods, ...
- Many of these algorithms are able to recover the same or equally valid solutions from the same problem instances
- What ingredients make hard thresholding algorithms efficient?

Three prototypical IHT algorithms for CS (similar for MC)

Alternating projection approaches to

$$\min_x \|y - Ax\|_2 \quad \text{subject to} \quad \|x\|_0 = k$$

- Normalized Iterated HT (NIHT) [Blumensath & Davies 09]

$$x_l = H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))$$

Three prototypical IHT algorithms for CS (similar for MC)

Alternating projection approaches to

$$\min_x \|y - Ax\|_2 \quad \text{subject to} \quad \|x\|_0 = k$$

- Normalized Iterated HT (NIHT) [Blumensath & Davies 09]

$$x_{l} = H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))$$

- Hard Thresholding Pursuit (HTP) [Maleki 09, Foucart 10]

$$I_l = supp(H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))) \quad \text{Descent supp. sets}$$

$$x_l = (A_{I_l}^T A_{I_l})^{-1} A_{I_l}^T y \quad \text{Pseudo-inverse}$$

Three prototypical IHT algorithms for CS (similar for MC)

Alternating projection approaches to

$$\min_x \|y - Ax\|_2 \quad \text{subject to} \quad \|x\|_0 = k$$

- Normalized Iterated HT (NIHT) [Blumensath & Davies 09]

$$x_{l} = H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))$$

- Hard Thresholding Pursuit (HTP) [Maleki 09, Foucart 10]

$$I_l = supp(H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))) \quad \text{Descent supp. sets}$$

$$x_l = (A_{I_l}^T A_{I_l})^{-1} A_{I_l}^T y \quad \text{Pseudo-inverse}$$

- Two-Stage Thres. [Milenkovic & Dai, Needell & Tropp 08]

$$v_l = H_{\alpha k}(x_{l-1} + \kappa A^T(y - Ax_{l-1}))$$

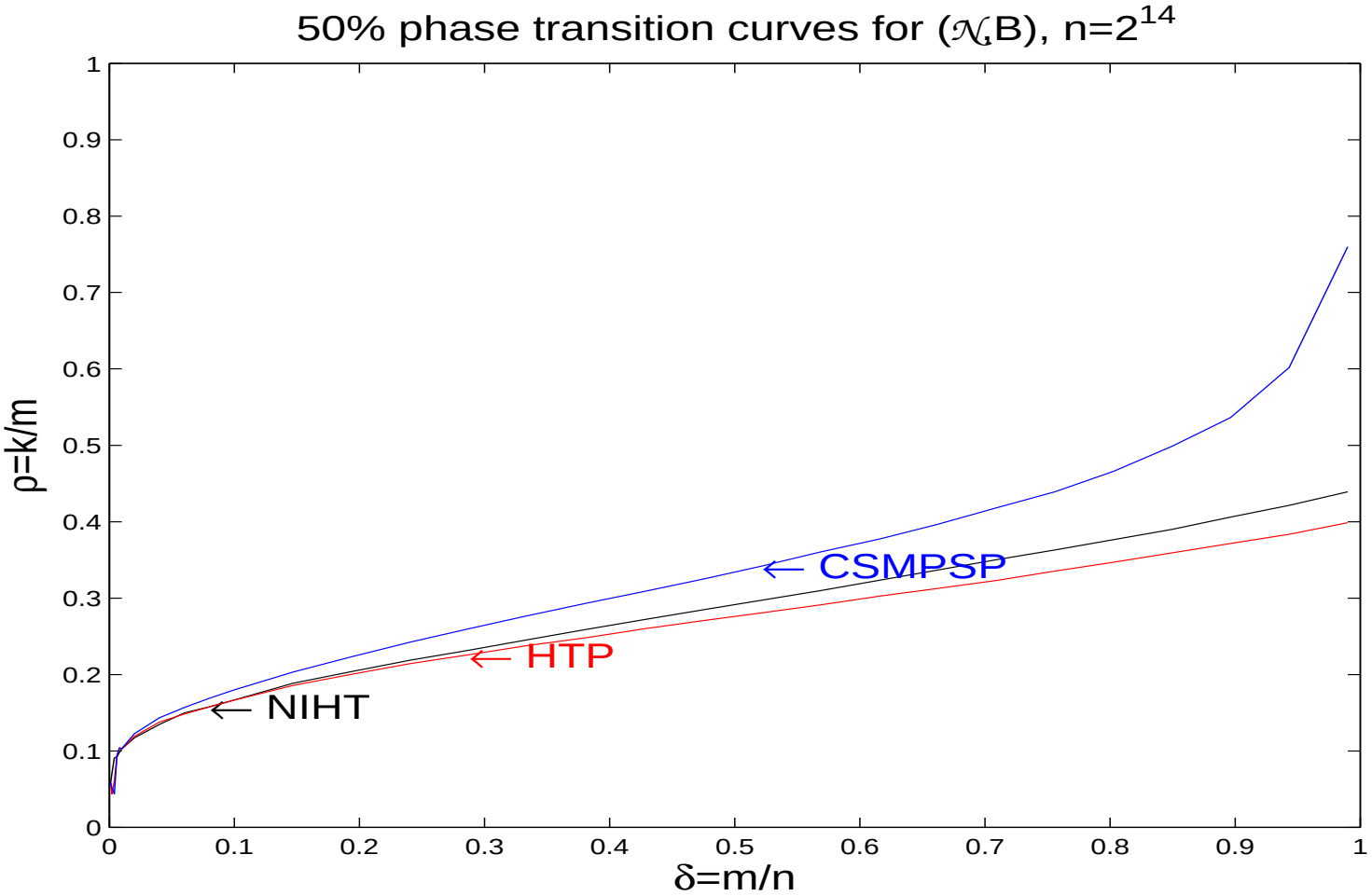
$$I_l = supp(v_l) \cup supp(x_{l-1}) \quad \text{Join supp. sets}$$

$$w_l = (A_{I_l}^T A_{I_l})^{-1} A_{I_l}^T y \quad \text{Least squares fit}$$

$$x_l = H_k(w_l) \quad \text{Second threshold}$$

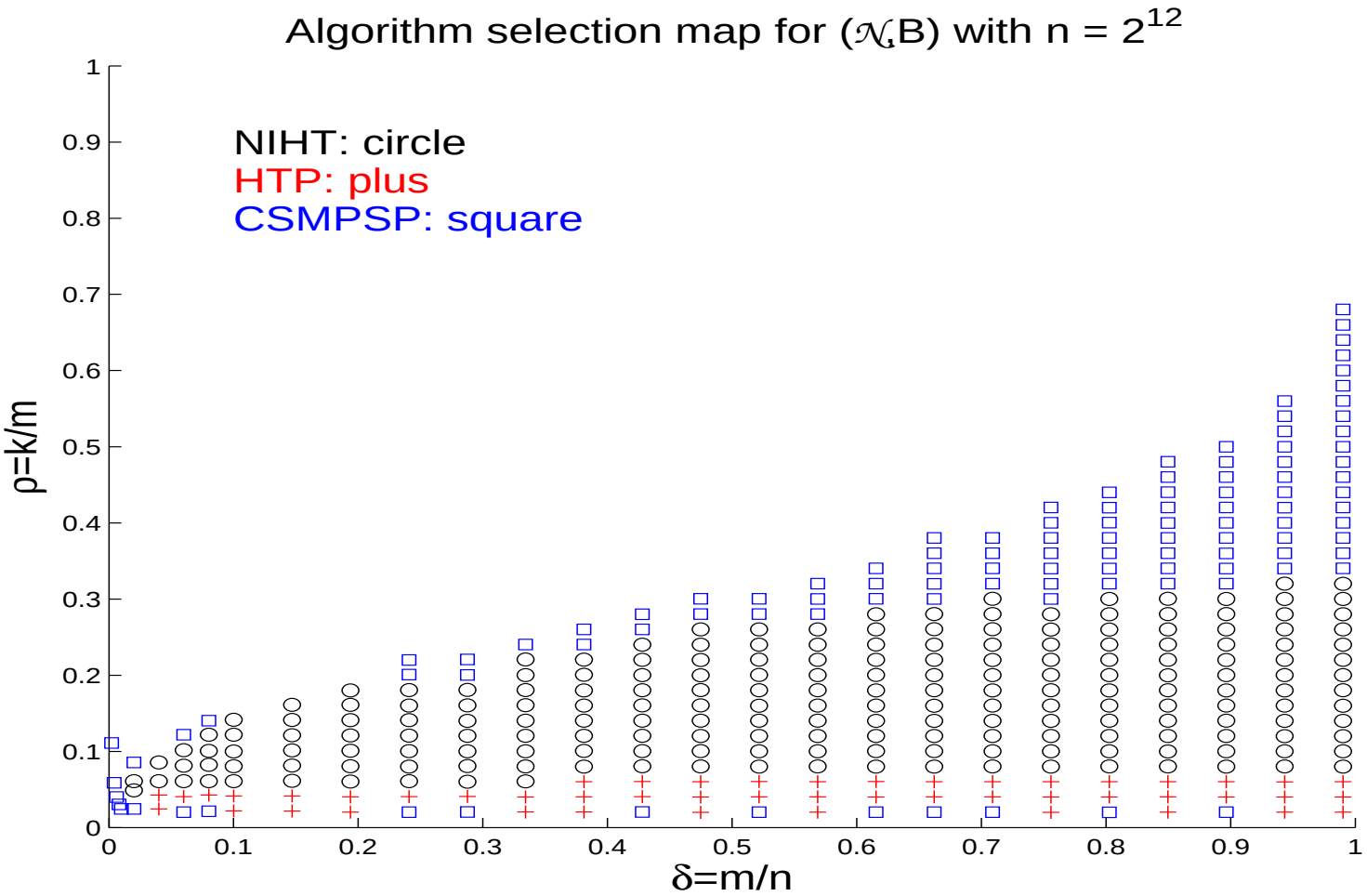
- Mixture of support set identification and local optimization

Recovery phase transitions: Gaussian matrix, sign vector, $n = 2^{14}$



Similar recovery regions, especially for $\delta \ll 1$. Which is fastest?

Algorithm Selection map:
Gaussian matrix, sign vector, $n = 2^{12}$, relative residual 10^{-3}



What goes into the design of a fast CS algorithm?

Three prototypical IHT algorithms for CS

- Normalized Iterated HT (NIHT) [Blumensath & Davies 09]
$$x_l = H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))$$
- Hard Thresholding Pursuit (HTP) [Foucart 10]
$$I_l = supp(H_k(x_{l-1} + \kappa A^T(y - Ax_{l-1}))) \quad \text{Descent supp. sets}$$
$$x_l = (A_{I_l}^T A_{I_l})^{-1} A_{I_l}^T y \quad \text{Pseudo-inverse}$$
- Two-Stage Thres. [Milenkovic & Dai, Needell & Tropp 08]
$$v_l = H_{\alpha k}(x_{l-1} + \kappa A^T(y - Ax_{l-1}))$$
$$I_l = supp(v_l) \cup supp(x_{l-1}) \quad \text{Join supp. sets}$$
$$w_l = (A_{I_l}^T A_{I_l})^{-1} A_{I_l}^T y \quad \text{Least squares fit}$$
$$x_l = H_{\beta k}(w_l) \quad \text{Second threshold}$$
- Low per iteration complexity best at early exploration phase, higher order better at later coefficient value recovery phase
- Three CGIHT variants combine low per iteration complexity and fast asymptotics via subspace confidence measure

Balancing the iteration cost with fast asymptotic rate

CGIHT Restarted [Blanchard, T & Wei 2013]

Initialization: Set $T_{-1} = \{\}$, $p_{-1} = 0$, $\nu_0 = A^* y$, $T_0 = \texttt{DetectSupport}(\nu_0)$, $x_0 = P_{T_0}(\nu_0)$, and $l = 1$.

Iteration: During iteration l, **do**

1: $r_{l-1} = A^*(y - Ax_{l-1})$ (compute the residual)

2: if $T_{l-1} \neq T_{l-2}$

$\beta_{l-1} = 0$ (set orthogonalization weight)

else

$\beta_{l-1} = \frac{\|P_{T_{l-1}} r_{l-1}\|_2^2}{\|P_{T_{l-1}} r_{l-2}\|_2^2}$ (compute orthogonalization weight)

3: $p_{l-1} = r_{l-1} + \beta_{l-1} p_{l-2}$ (define the search direction)

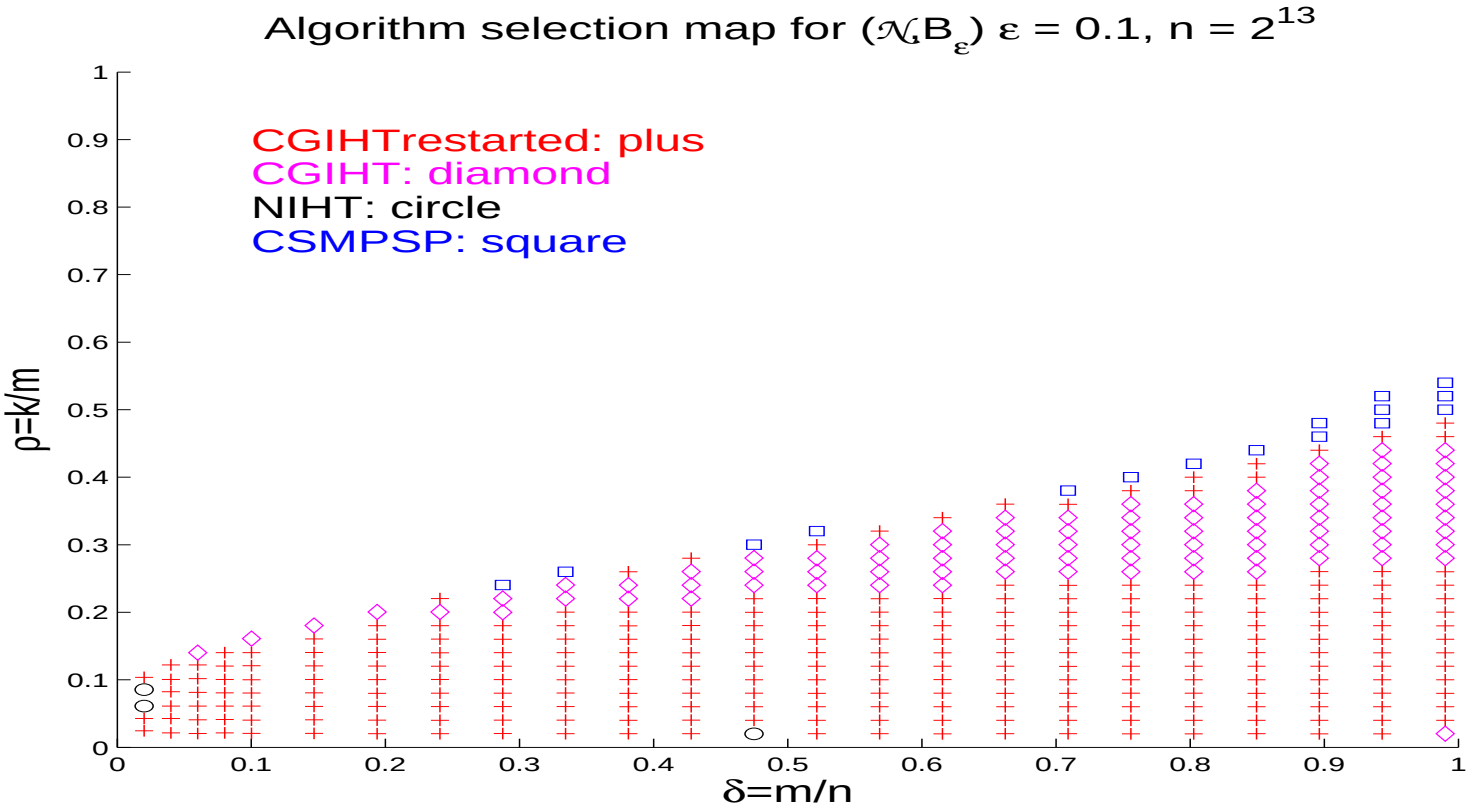
4: $\alpha_{l-1} = \frac{\|P_{T_{l-1}}(r_{l-1})\|_2^2}{\|AP_{T_{l-1}}(p_{l-1})\|_2^2}$ (optimal step size if $T_{l-1} = T_{l-2}$)

5: $\nu_{l-1} = x_{l-1} + \alpha_{l-1} p_{l-1}$ (conjugate gradient step)

6: $T_l = \texttt{DetectSupport}(\nu_{l-1})$ (proxy to the support set)

7: $x_l = P_{T_l}((\nu_{l-1}))$ (restriction to proxy support set T_l)

Moderate noise: $n = 2^{13}$ Gaussian matrix, sign vector, $y = Ax + e$ for e drawn $\mathcal{N}\left(0, \frac{1}{10}\|Ax\|_2\right)$



CGIHT variants nearly uniformly fastest especially with additive noise.

Similar behaviour for DCT and sparse matrices, other vector distributions.

"plain CGIHT" orthogonalizes at each iteration, lacks theory

CGIHT Restarted recovery guarantee

Restricted Isometry Property: sparse near isometry

- Classical ℓ^2 eigen-analysis [Candes & Tao 05]

$$(1 - L_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U_k)\|x\|_2^2 \quad \text{for } x \;\; k\text{-sparse}$$

Theorem (CGIHT Restarted for CS)

Let A be an $m \times n$ matrix with $m < n$, and $y = Ax + e$ for any x with at most k nonzeros. If the RIC constants of A satisfy

$$\frac{(L_{3k} + U_{3k})(5 - 2L_k + 3U_k)}{(1 - L_k)^2} < 1,$$

then there exists a $K > 0$ depending only on $\|x_0 - x\|_2$ such that

$$\|x_l - x\| \leq K \cdot \gamma^l + \frac{2\kappa_\alpha(1 + U_{2k})^{1/2}}{1 - \gamma}\|e\|_2$$

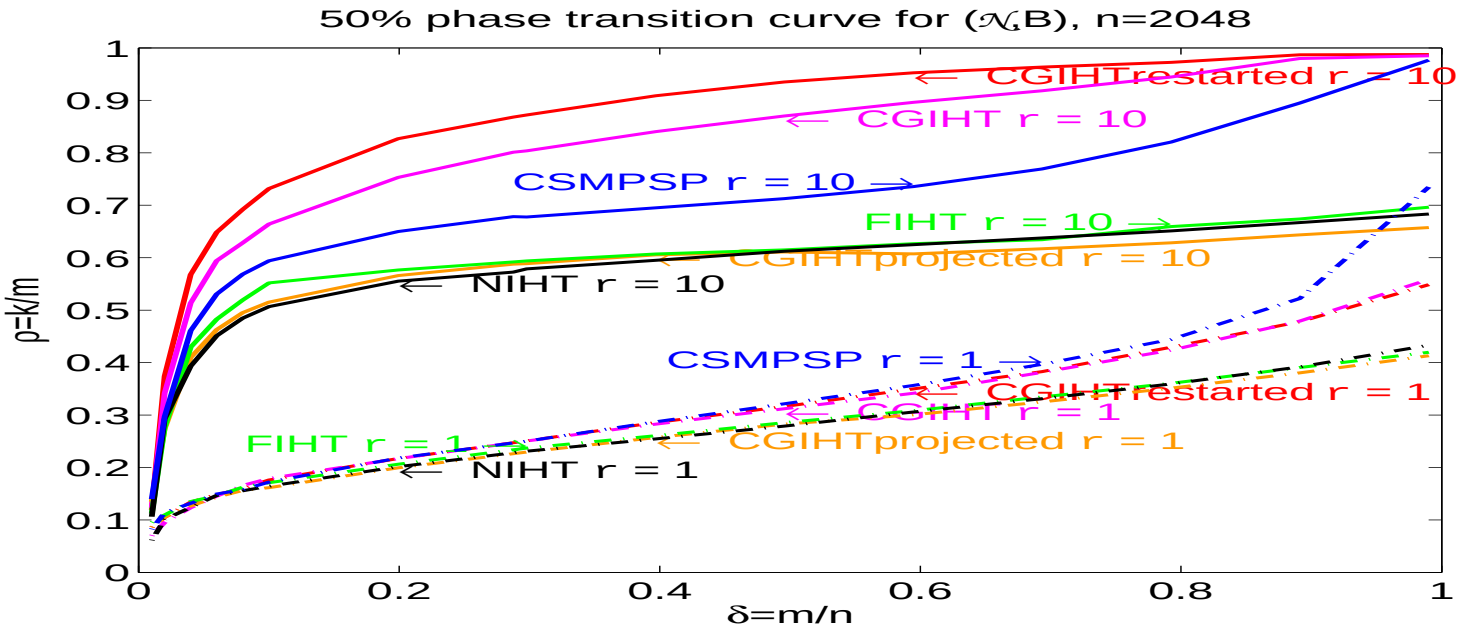
x_l is the l^{th} iteration of CGIHT and $\gamma < 1$ (formula available).

Greater differences for multi-measurement CS

Between CS and MC: Multi-measurement CS

- Multi-measurement, measure r vectors, each of which are k sparse with shared suport set but different nonzero values (eg. chemical spectroscopy and video with slowly varying images)

$$\min_{Z \in \mathbb{R}^{n \times r}} \|Y - AZ\|_2 \quad \text{subject to} \quad \|Z\|_{R0} \leq k.$$



CGIHT variants have substantially higher recovery region

CGIHT (projected) for matrix completion requires a different measure of subspace confidence due to continuous subspace

CGIHT projected for matrix completion

Iteration: During iteration l, **do**

1: if $\frac{\left\|R_{l-1}-\text{Proj}_{U_{l-1}}(P_{l-1})\right\|}{\|\text{Proj}_{U_{l-1}}(R_{l-1})\|} > \theta$

$$\text{Restart\_flag} = 1,\ \alpha_{l-1} = \frac{\left\|\text{Proj}_{U_{l-1}}(R_{l-1})\right\|^2}{\left\|\mathcal{A}\left(\text{Proj}_{U_{l-1}}(R_{l-1})\right)\right\|^2}$$

$$W_{l-1} = X_{l-1} + \alpha_{l-1} R_{l-1}$$

else

$$\text{Restart\_flag} = 0,\ \alpha_{l-1} = \frac{\left\|\text{Proj}_{U_{l-1}}(R_{l-1})\right\|^2}{\left\|\mathcal{A}\left(\text{Proj}_{U_{l-1}}(P_{l-1})\right)\right\|^2}$$

$$W_{l-1} = X_{l-1} + \alpha_{l-1}\text{Proj}_{U_{l-1}}(P_{l-1})$$

2: $U_l = \texttt{PrincipalLeftSingularVectors}_r(W_{l-1})$, $X_l = \text{Proj}_{U_l}(W_{l-1})$, $R_l = \mathcal{A}^*(y - \mathcal{A}(X_l))$

3: if Restart_flag $= 1$ set $P_l = R_l$, else

$$\beta_l = \frac{\left\|\text{Proj}_{U_l}(R_l)\right\|^2}{\left\|\text{Proj}_{U_l}(R_{l-1})\right\|^2},\ P_l = R_l + \beta_l \text{Proj}_{U_l}(P_{l-1})$$

CGIHT Projected for MC recovery guarantee

Theorem (CGIHT Projected for MC)

Let $\mathcal{A}$ be a linear map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^p$ with $p < mn$, and $y = \mathcal{A}(X) + e$ for any X of rank at most r. Let $c > 0$ then for the restarting parameter, $\theta < c(L_{3r} + U_{3r})/(1 + U_{2r})$, if the RIC constants of A satisfy

$$\mu = 2(1+c)\frac{L_{3r} + U_{3r}}{1 - L_r} < 1,$$

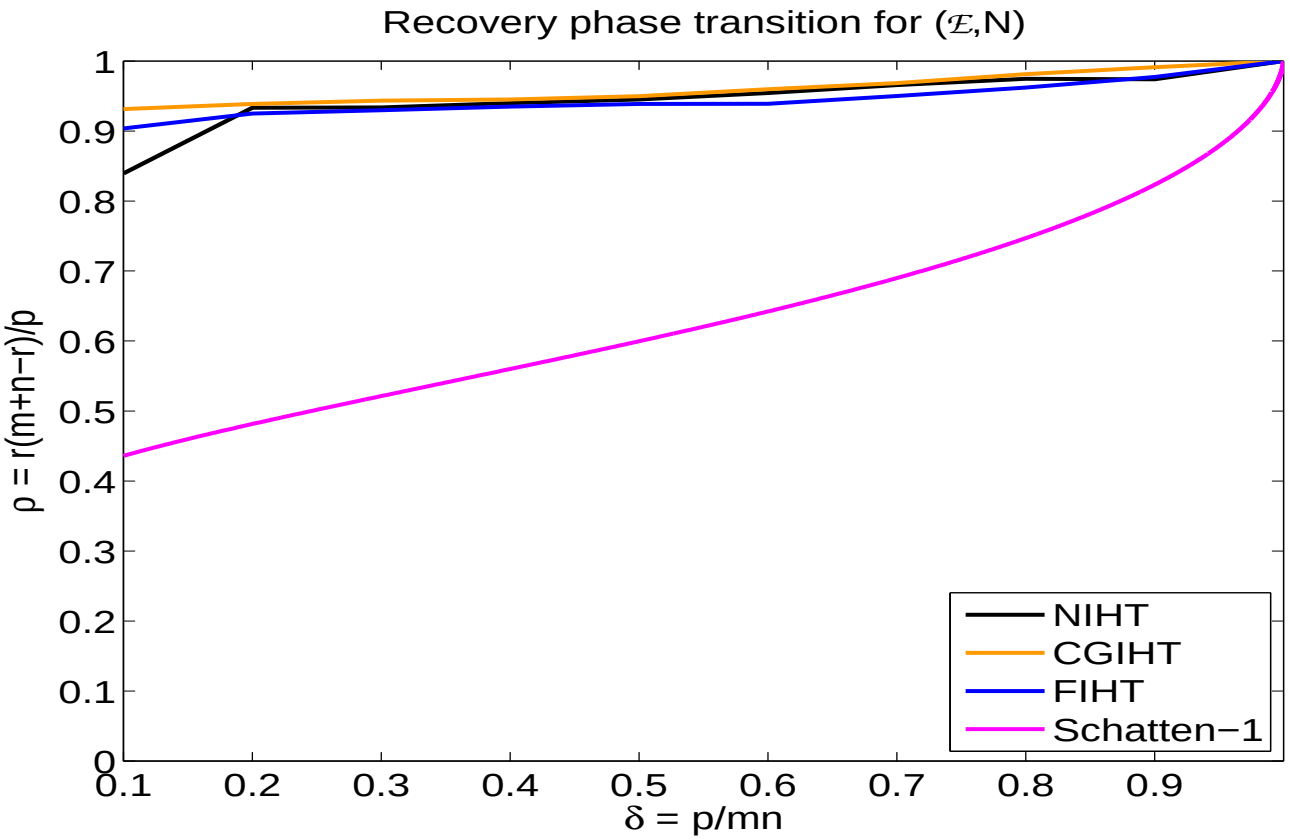
then

$$\|X_l - X\|_F \leq \mu^l \|X_0 - X\|_F + \frac{\xi}{1-\mu}\|e\|_2$$

where $\xi = 2(1+\theta)(1+U_{2r})^{1/2}/(1-L_r)$ and X_l is the l^{th} iteration of CGIHT projected.

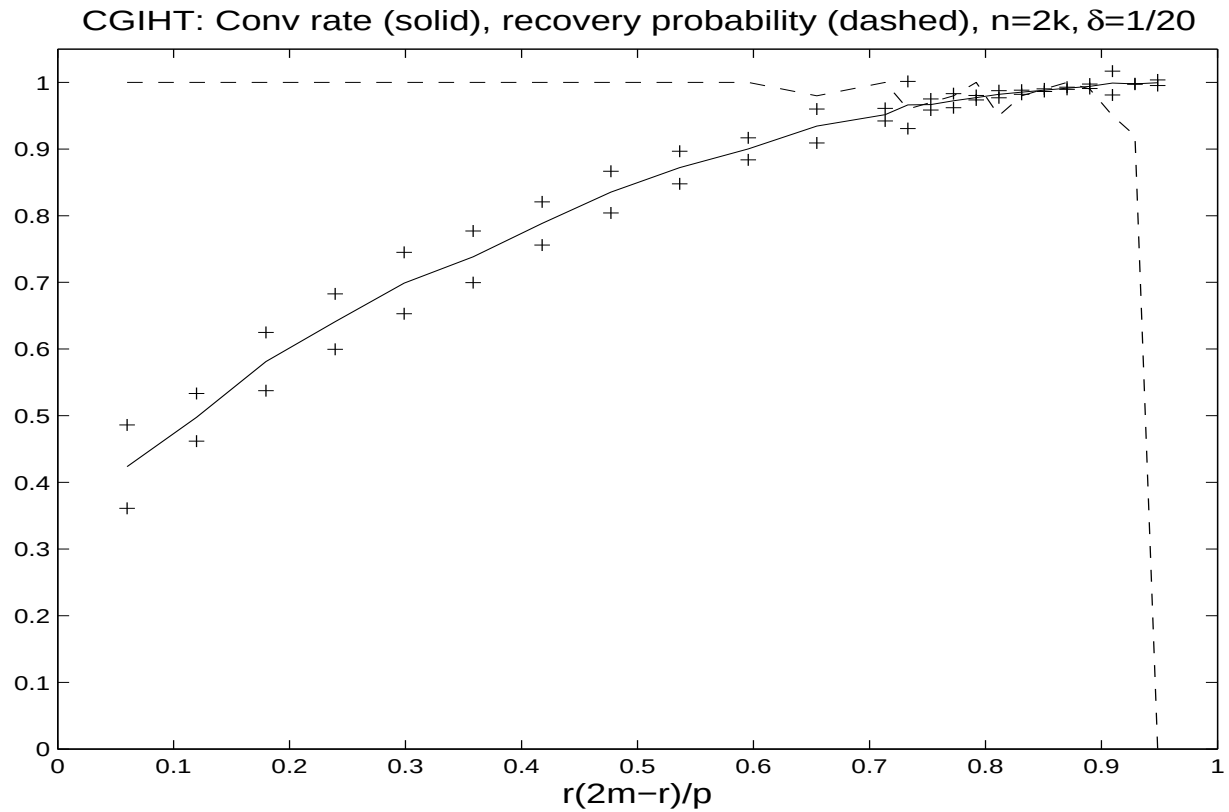
Restarting parameter c determines likelihood of restarting, with $c = 0$ recovering NIHT

NIHT, FIHT, CGIHT: entry sensing ($m = n = 2000$)



- Phase transition substantial above Schatten-1 norm
- CGIHT convergence rate is fastest in its class.
- What is happening in extreme undersampling $p \ll mn$?

CGIHT: entry sensing with $\delta = p/mn = 1/20$



- CGIHT at small $\delta = p/mn = 1/20$, 100 tests per value of r
- Recovery in at least 95 times in each of 100 tests for $\rho \leq 0.9$, whereas Schatten-1 recovery requires $\rho < 0.41$.
- Convergence rate appears to be only limit to recovery in matrix completion, even in extreme undersampling $\delta \ll 1$

A few concluding observations

- CS and MC algorithms have two phases: subspace determination and subspace data fitting
- When confidence in the subspace estimate is low, it is best to quickly search the space without minimizing local objectives
- Higher order methods can both accelerate convergence and increase recovery region
- CGIHT balances these competing aspects
- Iterative hard thresholding algorithms have substantially better average case matrix completion recovery than do convex regularizations

References

- Conjugate Gradient iterative hard thresholding for compressed sensing and matrix completion: Information and Inference (2015), Blanchard, Tanner and Wei.
- CGIHT for compressed sensing: observed noise stability: IEEE Trans. Signal Processing (2015), Blanchard, Tanner and Wei.
- Normalized iterative hard thresholding for matrix completion: SIAM J. on Scientific Computing (2012), Tanner and Wei
- GPU Accelerated Greedy Algorithms for compressed sensing; Mathematical Programming Computation (2013), Blanchard and Tanner.

Thank you for your time