Parallel-ℓ_0: A fully parallel algorithm for combinatorial compressed sensing

Jared Tanner & Rodrigo Mendoza-Smith

2. Int. Matheon Conf. on CS and its applications 2015
7^{th} Dec. 2015

University of Oxford[1]
Joint with Rodrigo Mendoza-Smith

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

Combinatorial Compressed Sensing (CCS)

- Let $A \in \mathbb{R}^{m \times n}$, $x \in \chi_k^n := \{x \in \mathbb{R}^n : \|x\|_0 \leq k\}$,.
- Compressed sensing looks for the solution, with $k < m < n$, of

$$y = Ax \quad \text{s.t.} \quad x \in \chi_k^n.$$

- Most CS theory developed for A Gaussian or Partial Fourier

Combinatorial Compressed Sensing (CCS)

- Let $A \in \mathbb{R}^{m \times n}$, $x \in \chi_k^n := \{x \in \mathbb{R}^n : \|x\|_0 \leq k\}$,.
- Compressed sensing looks for the solution, with $k < m < n$, of

$$y = Ax \quad \text{s.t.} \quad x \in \chi_k^n.$$

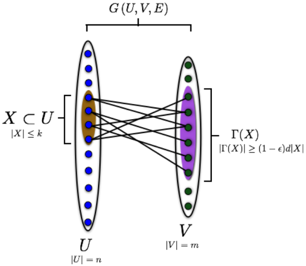
- Most CS theory developed for A Gaussian or Partial Fourier

Ensemble	Storage	Generation	$A^T y$	m
Gaussian	$\mathcal{O}(mn)$	$\mathcal{O}(mn)$	$\mathcal{O}(mn)$	$\mathcal{O}(k \log(n/k))$
Partial Fourier	$\mathcal{O}(m)$	$\mathcal{O}(n)$	$\mathcal{O}(n \log(n))$	$\mathcal{O}(k \log^5(n))$
Expander	$\mathcal{O}(dn)$	$\mathcal{O}(dn)$	$\mathcal{O}(dn)$	$\mathcal{O}(k \log(n/k))$

- In CCS A is an *expander matrix*, *i.e.* a sparse binary matrix with $d << m$ ones per column ($A \in \mathbb{E}_{k,\varepsilon,d}$).

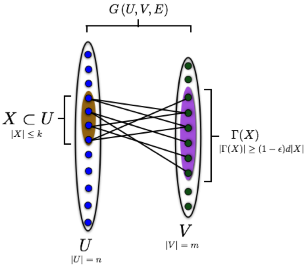
Expander matrices: $A \in \mathbb{E}_{k,\varepsilon,d}$, some notation

Edges of expander graph: $[n] = \{1, 2, \ldots, n\}$, or $[m]$
Neighbours of vertices X, $\mathcal{N}(X)$, are vertices connected by an edge



Expander matrices: $A \in \mathbb{E}_{k,\varepsilon,d}$, some notation

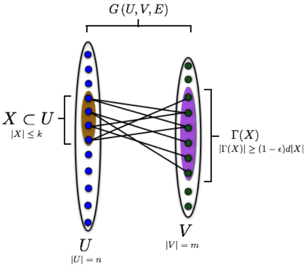
Edges of expander graph: $[n] = \{1, 2, \ldots, n\}$, or $[m]$
Neighbours of vertices X, $\mathcal{N}(X)$, are vertices connected by an edge



$$A_{ij} = \mathbb{1}_{\{i \text{ and } j \text{ are connected}\}}.$$

Expander matrices: $A \in \mathbb{E}_{k,\varepsilon,d}$, some notation

Edges of expander graph: $[n] = \{1, 2, \ldots, n\}$, or $[m]$
Neighbours of vertices X, $\mathcal{N}(X)$, are vertices connected by an edge

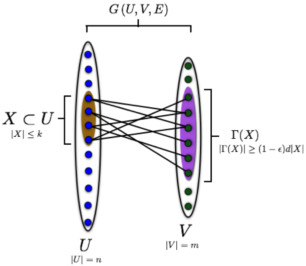


$$A_{ij} = \mathbb{1}_{\{i \text{ and } j \text{ are connected}\}}.$$

$\exists\ \varepsilon \in (0,1)$ s.t.
$|\Gamma(X)| = |\mathcal{N}(X)| > (1-\varepsilon)d|X|$
$\forall\ X \subset [n]$ with $|X| \leq k$.

Expander matrices: $A \in \mathbb{E}_{k,\varepsilon,d}$, some notation

Edges of expander graph: $[n] = \{1, 2, \ldots, n\}$, or $[m]$
Neighbours of vertices X, $\mathcal{N}(X)$, are vertices connected by an edge



$A_{ij} = \mathbb{1}_{\{i \text{ and } j \text{ are connected}\}}$.

$\exists\ \varepsilon \in (0,1)$ s.t.
$|\Gamma(X)| = |\mathcal{N}(X)| > (1-\varepsilon)d|X|$
$\forall\ X \subset [n]$ with $|X| \leq k$.

$d \equiv |\mathcal{N}(j)|\ \forall\ j \in [n]$.

$A \in \mathbb{R}^{m \times n}$ is a sparse binary matrix with $d << m$ ones per column

Structure of CCS Greedy Algorithms

Initialization: $A \in \mathbb{E}_{k,\varepsilon,d}$; $y \in \mathbb{R}^m$, $\hat{x} = 0$, $r = y$
while not converged
Compute a score s_j and an update ω_j $\forall\, j \in [n]$
Select $T \subset [n]$ based on a rule on s_j
$\hat{x}_j \leftarrow \hat{x}_j + \omega_j$ for $j \in T$
$r \leftarrow y - A\hat{x}$

- CCS algorithms differ by their score metric s_j and how many elements T is allowed to contain

Overview of CCS Greedy Algorithms

Algorithm	Score	Concurrency	Complexity
SMP (EIHT) [1]	ℓ_1 / median	parallel	$\mathcal{O}((nd + n\log n)\log \|x\|_1)$
SSMP [2]	ℓ_1 / median	serial	$\mathcal{O}((\frac{d^3 n}{m} + n)k + (n\log n)\log \|x\|_1)$
LDDSR [3] / ER [4]	ℓ_0 / mode	serial	$\mathcal{O}((\frac{d^3 n}{m} + n)k)$

Overview of CCS Greedy Algorithms

Algorithm	Score	Concurrency	Complexity
SMP (EIHT) [1]	ℓ_1 / median	parallel	$\mathcal{O}((nd + n\log n)\log \|x\|_1)$
SSMP [2]	ℓ_1 / median	serial	$\mathcal{O}((\frac{d^3 n}{m} + n)k + (n\log n)\log \|x\|_1)$
LDDSR [3] / ER [4]	ℓ_0 / mode	serial	$\mathcal{O}((\frac{d^3 n}{m} + n)k)$
Serial-ℓ_0 [5]	ℓ_0 / ℓ_0	serial	$\mathcal{O}(dn\log k)$
Parallel-ℓ_0 [5]	ℓ_0 / ℓ_0	parallel	$\mathcal{O}(dn\log k)$

Overview of CCS Greedy Algorithms

Algorithm	Score	Concurrency	Complexity
SMP (EIHT) [1]	ℓ_1 / median	parallel	$\mathcal{O}((nd + n\log n)\log\|x\|_1)$
SSMP [2]	ℓ_1 / median	serial	$\mathcal{O}((\frac{d^3 n}{m} + n)k + (n\log n)\log\|x\|_1)$
LDDSR [3] / ER [4]	ℓ_0 / mode	serial	$\mathcal{O}((\frac{d^3 n}{m} + n)k)$
Serial-ℓ_0 [5]	ℓ_0 / ℓ_0	serial	$\mathcal{O}(dn\log k)$
Parallel-ℓ_0 [5]	ℓ_0 / ℓ_0	parallel	$\mathcal{O}(dn\log k)$

- Only SMP was observed to take less computational time than non-combinatorial CS algorithms such as NIHT
- Unfortunately SMP only able to recovery $x \in \chi_k^n$ for $k/m \ll 1$
- Parallel-ℓ_0 computationally fast and recovery for $k/m \approx 0.3$
- Sudocodes is an alternative method, preprocessing to reduce n by determining locations in x that must be zero

Decoding by decreasing $\|r\|_{\ell_0}$

Parallel-ℓ_0

Initialization: $A \in \mathbb{E}_{k,\varepsilon,d}$; $y \in \mathbb{R}^m$, $\alpha \in [d-1]$, $\hat{x} = 0$, $r = y$

while not converged

$T \leftarrow \{(j, \omega_j) \in [n] \times \mathbb{R} : \|r\|_0 - \|r - \omega_j a_j\|_0 > \alpha\}$

for $(j, \omega_j) \in T$

$\hat{x}_j \leftarrow \hat{x}_j + \omega_j$ for $j \in T$

$r \leftarrow y - A\hat{x}$

Decoding by decreasing $\|r\|_{\ell_0}$

Parallel-ℓ_0
Initialization: $A \in \mathbb{E}_{k,\varepsilon,d}$; $y \in \mathbb{R}^m$, $\alpha \in [d-1]$, $\hat{x} = 0$, $r = y$
while not converged
$T \leftarrow \{(j, \omega_j) \in [n] \times \mathbb{R} : \|r\|_0 - \|r - \omega_j a_j\|_0 > \alpha\}$
for $(j, \omega_j) \in T$
$\hat{x}_j \leftarrow \hat{x}_j + \omega_j$ for $j \in T$
$r \leftarrow y - A\hat{x}$

Serial-ℓ_0
Initialization: $A \in \mathbb{E}_{k,\varepsilon,d}$; $y \in \mathbb{R}^m$, $\alpha \in [d-1]$, $\hat{x} = 0$, $r = y$
while not converged
for $j \in [n]$
$T \leftarrow \{\omega_j \in \mathbb{R} : \|r\|_0 - \|r - \omega_j a_j\|_0 > \alpha\}$
$\hat{x}_j \leftarrow \hat{x}_j + \omega_j$ for $j \in T$
$r \leftarrow y - A\hat{x}$

- Parallel-ℓ_0: computing T and updating $\hat{x}$ suitable for GPU

Theorem (Convergence of Expander ℓ_0-Decoders)

Let $A \in \mathbb{E}_{k,\varepsilon,d}$ and $\varepsilon < 1/4$. and $x \in \chi_k^n$ be a dissociated signal. Then, Serial-ℓ_0 and Parallel-ℓ_0 with $\alpha = (1-2\varepsilon)d$ can recover x from $y = Ax \in \mathbb{R}^m$ in $\mathcal{O}(dn \log k)$ operations.

Dissociated: $\sum_{j \in T_1} x_j \neq \sum_{j \in T_2} x_j \quad \forall\ T_1, T_2 \subset \mathrm{supp}(x)$ with $T_1 \neq T_2$

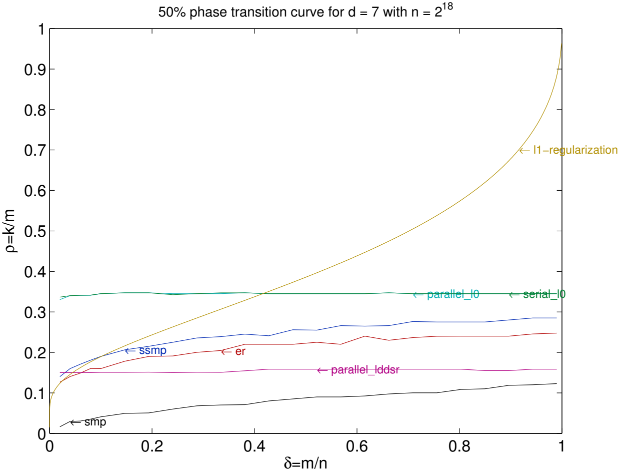
Theorem (Convergence of Expander ℓ_0-Decoders)

Let $A \in \mathbb{E}_{k,\varepsilon,d}$ and $\varepsilon < 1/4$. and $x \in \chi_k^n$ be a dissociated signal. Then, Serial-ℓ_0 and Parallel-ℓ_0 with $\alpha = (1-2\varepsilon)d$ can recover x from $y = Ax \in \mathbb{R}^m$ in $\mathcal{O}(dn \log k)$ operations.

Dissociated: $\sum_{j \in T_1} x_j \neq \sum_{j \in T_2} x_j \quad \forall\ T_1, T_2 \subset \text{supp}(x)$ with $T_1 \neq T_2$

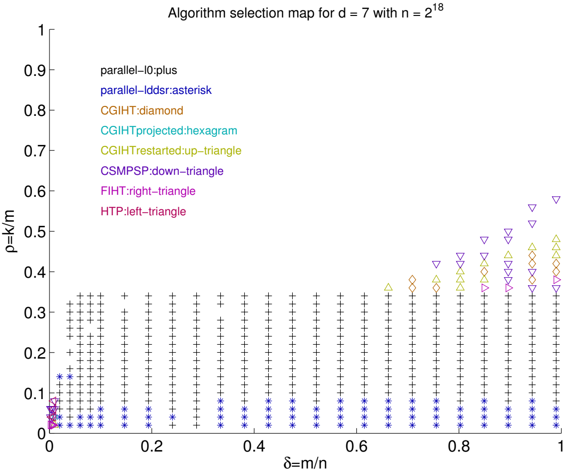
- Dissociation, the same signal model as consider by sudocodes.
- Parallel-ℓ_0 requires $\log k$ iterations of complexity $\mathcal{O}(dn)$ complexity, each of which is trivially decomposed into n independent tasks of complexity $\mathcal{O}(d)$.
- Serial-ℓ_0 requires $n \log k$ iterations of complexity $\mathcal{O}(d)$.
- Serial-ℓ_0 is faster than Parallel-ℓ_0 if both run on a single core, but Parallel-ℓ_0 substantially faster when run on high performance computing GPUs with thousands of cores.
- Serial-ℓ_0 and Parallel-ℓ_0 have nearly identical recovery regions.

Improved phase transition



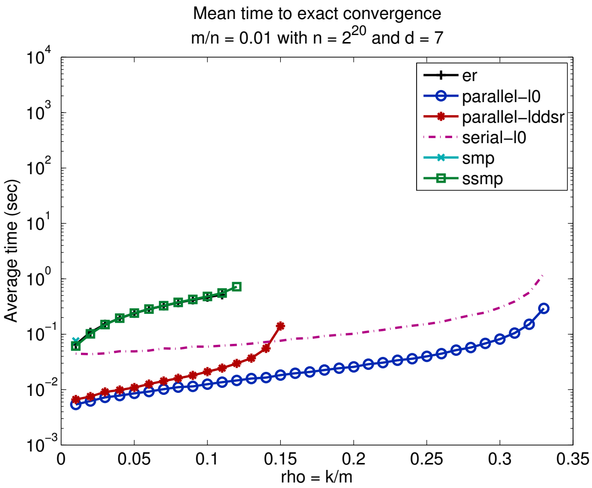
- Greater recovery region than other CCS algorithms
- No apparent decrease in phase transition for $m \ll n$

Fastest CS algorithm for $A \in \mathbb{E}_{k,\varepsilon,d}$



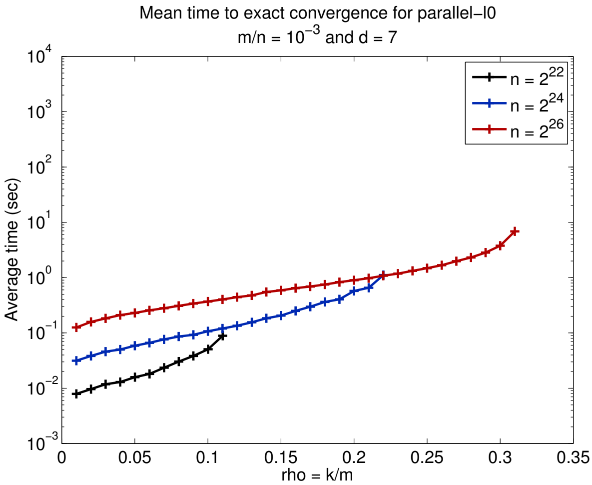
- Parallel-ℓ_0 and Parallel-LDDSR fastest when convergent
- First examples of CCS algorithms being state-of-the-art

Average timing for fixed $m/n = 1/100$



- Less computational time over Parallel-LDSR, all but $k/m \ll 1$
- Near constant speedup of Parallel-ℓ_0 over Serial-ℓ_0

High phase transition persists for $m/n \ll 1$



- Recovery ability of $m \approx 3k$ even for $m = n \times 10^{-3}$,
- Time for $n \approx 67$ million in under 2 seconds

Sketch of the complexity proof:

Lemma (Bounded frequency of values in expander measurements of dissociated signals)

Let $x \in \chi_k^n$ *be dissociated,* $A \in \mathbb{E}_{k,\varepsilon,d}$*, and* ω *a nonzero value in* Ax*. Then, there is a unique set* $T \subset \mathrm{supp}(x)$ *such that* $\omega = \sum_{j \in T} x_j$ *and the value* ω *occurs in* y *at most* d *times,*

$$|\{i \in [m] : y_i = \omega\}| \leq d \quad \forall\, \omega \neq 0.$$

Proof: The uniqueness of the set $T \subset \mathrm{supp}(x)$ such that $\omega = \sum_{j \in T} x_j$ follows by the definition of dissociated. Since $|\mathcal{N}(j)| = d$ for all $j \in [n]$, we have that,

$$|\{i \in [m] : y_i = \omega\}| = \left| \bigcap_{j \in T} \mathcal{N}(j) \right| \leq |\mathcal{N}(j_0)| = d$$

for any $j_0 \in T$.

Lemma (Pairwise column overlap)

Let $A \in \mathbb{E}_{k,\varepsilon,d}$. If $\varepsilon < 1/4$, every pair of columns of A intersect in less than $(1-2\varepsilon)d$ rows, that is, for all $j_1, j_2 \in [n]$ with $j_1 \neq j_2$

$$\left| \mathcal{N}(j_1) \bigcap \mathcal{N}(j_2) \right| < (1-2\varepsilon)d.$$

Proof: Let $S \subset [n]$ be such that $|S| = 2$ then

$$|\mathcal{N}(S)| \geq 2(1-\varepsilon)d > 2d - (1-2\varepsilon)d,$$

where the first inequality is definition of $A \in \mathbb{E}_{k,\varepsilon,d}$ and the second inequality follows from $\epsilon < 1/4$.

Lemma (Support identification)

Let $y = Ax$ for dissociated $x \in \chi_k^n$ and $A \in \mathbb{E}_{k,\varepsilon,d}$ with $\varepsilon < 1/4$. Let $\omega \neq 0$ be such that

$$|\{i \in \mathcal{N}(j) : y_i = \omega\}| > (1 - 2\varepsilon)d, \tag{1}$$

then $\omega = x_j$.

Proof: Our claim is that for any ω which is a nonzero value from y, if the cardinality condition (1) is satisfied then the value $\omega = \sum_{j \in T} x_j$ occurs for the set T being a singleton, $|T| = 1$. Frequency lemma states that T is unique and that

$$|\{i \in \mathcal{N}(j) : y_i = \omega\}| = \left| \bigcap_{j \in T} \mathcal{N}(j) \right|.$$

If $|T| > 1$ then the above is not more than the intersection of any two of the sets $\mathcal{N}(j_1)$ and $\mathcal{N}(j_2)$, and by pairwise column overlap lemma, is less than $(1 - 2\varepsilon)d$ which contradicts the cardinality condition (1) and consequently $|T| = 1$ and $\omega = x_j$.

Theorem (Convergence rate of Parallel-ℓ_0)

Let $A \in \mathbb{E}_{k,\varepsilon,d}$ and let $\varepsilon < 1/4$, and $x \in \chi_k^n$ be dissociated. Then, Parallel-ℓ_0 with $\alpha = (1-2\varepsilon)d$ can recover x from $y = Ax \in \mathbb{R}^m$ in $\mathcal{O}(\log k)$ iterations of complexity $\mathcal{O}(dn)$.

Sketch of proof: Let T_ℓ be set T of vertices to update at iteration ℓ and $S_\ell = \text{supp}(x - \hat{x})$. As $A \in \mathbb{E}_{k,\varepsilon,d}$ has d nonzeros per column, the reduction in the cardinality of the residual, say $\|r^\ell\|_0 - \|r^{\ell+1}\|_0$, can be at most $d|T_\ell|$;

$$\|r^\ell\|_0 - \|r^{\ell+1}\|_0 \leq d|T_\ell|.$$

Reduction in residual bounded below by (non-obvious)

$$\|r^\ell\|_0 - \|r^{\ell+1}\|_0 \geq \alpha|T_\ell| + (|S_\ell| - |T_\ell|).$$

Combining bounds ensures linear convergence

$$|S_{\ell+1}| \leq \frac{2\varepsilon d}{1+2\varepsilon d}|S_\ell|$$

Summary

- Serial-ℓ_0 and Parallel-ℓ_0 recovery for $A \in \mathbb{E}_{k,\varepsilon,d}$ in complexity $\mathcal{O}(dn \log k)$ and observed to take less time than non-CCS algorithms
- Recovery observed, for n large enough, with $m \approx 3k$
- Theory requires either x dissociated, or x drawn independent of A and columns of A scaled by dissociated values
- Robustness to ℓ_∞ bounded additive noise follows, but unknown for other noise variants or compressible signals
- There are noise robustness techniques for sudocodes (Ma, Baron, Needell 2014) which can be applied to ℓ_0 decoders

Bibliography

[1] Radu Berinde, Piotr Indyk, and Milan Ruzic
Practical near-optimal sparse recovery in the ℓ_1 norm.
Allerton Communication, control and computing 2008
[2] Radu Berinde and Piotr Indyk
Sequential sparse matching pursuit.
Allerton Communication, control and computing 2009
[3] Weiyu Xu and Babak Hassibi
Efficient compressive sensing with deterministic guarantees using expander graphs.
Information theory workshop, 2007. ITW'07. IEEE, pages 414-419
[4] Sina Jafarpour, Weiyu Xu, Babak Hassibi, Robert Calderbank
Efficient and robust compressed sensing using high-quality expander graphs.
arXiv preprint arXiv:0806.3802, 2008.
[5] Rodrigo Mendoza-Smith and Jared Tanner
Expander ℓ_0-decoding

Alan Turing Institute (ATI): watch this space

- The UK recently (Nov. 2015) launched a new "Date Science" centre
  - Funded by 5 universities: Cambridge, Edinburgh, Oxford, UCL, Warwick and EPSRC (Eng. Phy. Sci. Res. Council)
  - Physical space in central London: British Library
  - Currently has £77million budget for five years (growing)
- The scientific programme of the ATI is currently being formed, based on a series of workshops between Oct. 2015 to Feb. 2016.
  - Currently advertising for Research Fellows (senior postdoc) with initial 3 year appointment, possibly extended to 5 years.
  - Five founding universities are advertising permanent (tenure track) positions, including Oxford...
  - Happy to answer any questions and hope to see you at the ATI

Thank you for your time