Compressed sensing: structured sparsity and sparse operators.

Bubacarr Bah

Inst. of Computational Engineering & Sciences, and Dept. of Mathematics
University of Texas at Austin

Mathematical and Statistical Analysis of Spatial Data
Workshop at Dept. of Mathematical Sciences
Aalborg University, June 1-3, 2015

Joint work with Volkan Cevher & Luca Baldassarre @ LIONS, EPFL
The simplicity of large data sets

Understanding and working with large data sets is built on **simple** models:

- Time series such as audio
- Images of natural scenes
- Low rank matrix approximation
- Piecewise linear embeddings

...
Understanding and working with large data sets is built on **simple** models:

- Time series such as audio
- Images of natural scenes
- Low rank matrix approximation
- Piecewise linear embeddings
  ...

---
Understanding and working with large data sets is built on **simple** models:

- Time series such as audio
- Images of natural scenes
- Low rank matrix approximation
- Piecewise linear embeddings
  
  ![Image of natural scenes](image.png)
The simplicity of large data sets

Understanding and working with large data sets is built on **simple** models:

- Time series such as audio
- Images of natural scenes
- Low rank matrix approximation
- Piecewise linear embeddings

Exploiting underlying simplicity in data is ubiquitous, but tends to be used only after detailed information is acquired.
Exploiting simplicity in the acquisition stage

For many applications data acquisition is “costly” in some way

- Tomography, Radar, Multi-spectral imaging
- Analog-to-digital converters,
- Online recommending systems (NetFlix),
- Cardiac signal processing, Drug discovery,
- Bacterial communities, Genome sensing, …

How can we exploit the **simplicity** of the data to acquire only the essential information: relevant time-frequencies, edges in images, active entries, stereotypes, few scatterers?
Compressive sampling: Data acquisition at the information rate

Implemented by linear measurements, \( y := A x \), followed by a non-linear reconstruction \( \Rightarrow \) 3 key aspects:

1. Simple (redundant) \( x \)
   - Sparsity or compressibility (e.g. \( k \)-sparse \( x \))
2. Projection \( A \)
   - Information preserving (stable embedding)
3. Recovery algorithm \( \Delta \)
   - Tractable & accurate
(1) Simplicity (redundancy)

- **k-sparse** $\mathbf{x}$ - signal, image, vectors, ...

  \[ |x_i| \]

  $k$ sorted indices $N$

  $k$ non-zero coordinates

- **k-compressible** $\mathbf{x}$

  \[ |x_i| \]

  $k$ sorted index $N$

  power-law decay

- Simplicity in **non-canonical basis** $\mathbf{x} = \Phi \alpha \Rightarrow \mathbf{y} = A\Phi \alpha$
(2) Information preserving projections

Tools of analysis

Coherence, Restricted isometry property (RIP), Nullspace property, ...

Definition (ℓ_p-norm restricted isometry property (RIP-ℓ_p))

A matrix $A$ has RIP-ℓ_p of order $k$, if for all $k$-sparse $x$, it satisfies

$$(1 - \delta_k)\|x\|_p \leq \|Ax\|_p \leq (1 + \delta_k)\|x\|_p$$
(2) Information preserving projections

**Goal**

Information preserving $\mathbf{A}$ with small number of rows $m = m(k, N)$ that guarantees efficient and provable recovery of $\mathbf{x}$

- **Optimal**: $m = O(k \log(N/k))$
  - Subgaussian matrices: Gaussian $\sim \mathcal{N}(0, 1/m)$, Bernoulli $\pm 1$ with probability $1/2$, etc
- **"Quasi"-optimal**: $m = O\left(k \log^{O(1)}(N/k)\right)$
  - Structured random matrices: partial Fourier, Toeplitz, Circulant and Bounded Orthonormal Systems
- **Sub-optimal**: $m = \Omega(k^2)$
  - Deterministic, binary $\{0, 1\}$ matrices
(3) Recovery algorithms - tractability of recovery

- **Underdetermined** system of equations, **infinite** solutions
- **Simplicity** of \( x \) like **sparsity** leads to **unique** solutions

**Nonlinear** reconstruction

Given \((A, y)\), find \(k\)-sparse \(\hat{x}\) satisfying:

\[
\hat{x} = \min_{x \in \mathbb{R}^N} \|x\|_0 \quad \text{subject to} \quad Ax = y.
\]

- **Tractable** recovery algorithms (\(\Delta\)) with **provable** guarantees
  - **Convex** \(\ell_1\)-minimization approaches:
    \[
    \hat{x} = \min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{subject to} \quad Ax = y.
    \]
  - **Discrete** algorithms (OMP, IHT, CoSaMP, EIHT, ALPS, ...)
    (IHT) iterates \(x^{n+1} = H_k (x^n + A^* (y - Ax^n))\)
(3) Recovery algorithms - examples

Convex relaxation

- $\ell_0$ problem NP-hard, recast as $\ell_1$:
  \[
  \min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{subject to} \quad Ax = y
  \]
- $\ell_2$ fails to recover sparse $x$
- $\ell_1$ does! (convex $\equiv$ LP)
(3) Recovery algorithms - examples

Convex relaxation

- $\ell_0$ problem NP-hard, recast as $\ell_1$:
  \[
  \min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{subject to} \quad Ax = y
  \]
- $\ell_2$ fails to recover sparse $x$
- $\ell_1$ does! (convex $\equiv$ LP)
(3) Recovery algorithms - examples

Convex relaxation

- $\ell_0$ problem NP-hard, recast as $\ell_1$:
  \[
  \min_{x \in \mathbb{R}^N} ||x||_1 \text{ subject to } Ax = y
  \]
- $\ell_2$ fails to recover sparse $x$
- $\ell_1$ does! (convex $\equiv$ LP)
(3) Recovery algorithms - examples

Convex relaxation

- $\ell_0$ problem NP-hard, recast as $\ell_1$:
  \[
  \min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{subject to} \quad Ax = y
  \]
- $\ell_2$ fails to recover sparse $x$
- $\ell_1$ does! (convex $\equiv$ LP)

Iterative Hard Thresholding (IHT)

- $\ell_0$ problem equivalent to
  \[
  \min_{\|x\|_0 \leq k} \|Ax - y\|_2
  \]
- Do steepest descent & enforce constraint by thresholding
  \[
  x^{n+1} = x^n + H_k (A^*(y - Ax^n))
  \]
(3) Recovery algorithms - accuracy

- Δ returns approximations with $\ell_p/\ell_q$-approximation error:

**Definition ($\ell_p/\ell_q$-approximation error - instance optimality)**

A Δ returns $\hat{x} = \Delta(Ax + e)$ with $\ell_p/\ell_q$-approximation error if

$$
\|\hat{x} - x\|_p \leq C_1 \sigma_k(x)_q + C_2 \|e\|_p
$$

for a noise vector $e$, $C_1, C_2 > 0$, $1 \leq q \leq p \leq 2$, $\sigma_k(x)_q := \min_{k\text{-sparse } x'} \|x - x'\|_q$

- The pair $(A, \Delta)$ ⇒ two types of error guarantees
  - *for each* - one pair $(A, \Delta)$ for each given $x$
  - *for all* - one pair $(A, \Delta)$ for all $x$
Tomography, Radar, Multi-spectral imaging, ... 
Analog-to-digital converters, 
Online recommending systems (NetFlix), 
Cardiac signal processing, Drug discovery, ... 
Bacterial communities, Genome sensing, ... 
Single pixel camera, ...
Single-pixel camera

- Single Photo-diode digital camera [Baraniuk and Kelly 06]

- 2% measurements compared to number of pixels in recon.
- Savings, measurement time, simple device, power of device, ...

Compressed sensing overview
Structured sparsity
Sparse sampling operators
Model expander recovery
Conclusion

Motivation for compressed sensing
Key aspects of compressed sensing
Compressed sensing applications
Selected references
A physical example

SELEX GALILEO’S SINGLE PIXEL COMPRESSED SENSING CAMERA
Signals/data, e.g. images, are huge! **Big Data!!**

*Compression* for portability eg. $\approx 10^4$, **JPEG2000**

Measure a **lot** & keep a **few**?

*Compressed sensing*, measuring at the **compressed rate**
Selected references

- EJ Candès, Romberg, & Terry Tao - RIP
- Joel Tropp and Anna Gilbert - Coherence, OMP, sublinear
- Rich Baraniuk - Single photodiode camera
- Cohen, Dahmen, DeVore - Lindenstrauss Johnson Lemma
- Nowak, Wainwright - Behavior under noise
- Vahid Tarokh, Devore - Deterministic setting
- Michael Lustig & John Pauly, Davies - MRI
- Rauhut, Pfander - Channel estimation, Gabor, Structured
- Herrmann, Hennenfent - Seismic Imaging
- Hormati, Jovanovic, Vetterli - Acoustic Tomography
- Rice University repository on CS: http://dsp.rice.edu/cs
- Personal website: https://www.ma.utexas.edu/users/bah/
Sparsity and beyond

- **Generic sparsity** (or compressibility) not specific enough

- **Structured sparsity** $\equiv$ model-based CS [Baraniuk et al. 2010]

**Note:** $\mathcal{M}_k \subseteq \Sigma_k$
Model-based CS

- Model-based CS exploits **structure** in sparsity model
  - improves interpretability
  - reduces sketch length
  - increases speed of recovery

**tree-sparse model**

**Block-sparse model**
More models

Reduced sample complexity examples, [Baraniuk et al. ’10]

- $O(k)$ for tree structure
- $O(k + \log(M))$ for block structure with $M$ blocks

Overlapping group model

Group models application examples:

- Genetic Pathways in Microarray data analysis
- Wavelet models in image processing
- Brain regions in neuroimaging
RIP-1

Definition ($\ell_p$-norm Restricted Isometry Property (RIP-$p$))

A matrix $\mathbf{A}$ has RIP-$p$ of order $k$, if for all $k$-sparse $\mathbf{x}$, it satisfies

$$(1 - \delta_k)\|\mathbf{x}\|_p^p \leq \|\mathbf{A}\mathbf{x}\|_p^p \leq (1 + \delta_k)\|\mathbf{x}\|_p^p$$

- **Subgaussian** $\mathbf{A} \in \mathbb{R}^{m \times N}$ (w.h.p) have RIP-2 with $m = O(k \log(N/k))$, but sparse binary $\mathbf{A}$ does not have RIP-2 unless $m = \Omega(k^2)$
- Model sparsity requires fewer $m$ for RIP-2
  - $O(k)$ for **tree** structure
  - $O(k + \log(M))$ for **block** structure
- Adjacency matrices of **lossless expanders** have RIP-1 with $m = O(k \log(N/k))$
Expander Graphs

**Definition (Lossless Expander Graphs)**

A graph $G = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is an $(k, d, \epsilon)$-lossless expander if it is a bipartite graph with $|\mathcal{U}| = N$ left vertices, $|\mathcal{V}| = m$ right vertices, and has a regular left degree $d$, such that any $S \subset \mathcal{U}$ with $|S| \leq k$ has $|\Gamma(S)| \geq (1 - \epsilon) d|S|$ neighbors.

- **A** is sparse ($d$ nonzeros per col.)
  - Low storage complexity
  - Efficient application

**Computation benefits of A**

- $G = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is a $(k, d, \epsilon)$-lossless expander with $d = 3$, $\epsilon \in (0, 1/2)$.
- $S \subset \mathcal{U}$ with $|S| \leq k$ has $|\Gamma(S)| \geq (1 - \epsilon) d|S|$ neighbors.
Linear sketching

- $\mathbf{x} \in \mathbb{R}^N$ - data vector
- Compress $\mathbf{x}$ into a sketch, $\mathbf{Ax}$, $\mathbf{A}$ is binary and sparse
- $\mathbf{A} \in \mathbb{R}^{m \times N}$, $m \ll N$ but $\mathbf{Ax}$ captures essential information of $\mathbf{x}$

**Beneficial properties of linearity**

- Easy coordinate update: $\mathbf{Ax} + \mathbf{Ae}_i$ for coordinate $x_i$
- Easy decomposition: $\mathbf{A}(\mathbf{u} + \mathbf{v}) = \mathbf{Au} + \mathbf{Av}$

**Applications** include:

- Data streaming [Muthukrishnan 2003, Indyk 2007]
- Coding theory (LDPCs) [Muthukrishnan 2003, Indyk 2007]
- Database privacy [Dwork et al. 2007]
Construction of expanders

- **Constructions** of expanders that achieve optimal $m$ has been an open challenge.
- **Probabilistic** constructions of $(k, d, \epsilon)$-lossless expanders achieve optimal $m = \mathcal{O}(k \log(N/k))$.
- But their **deterministic** constructions are sub-optimal $m = \mathcal{O}(k^{1+\alpha})$ for $\alpha > 0$.
- We derived **sampling bounds** on random constructions of $(k, d, \epsilon)$-model expanders for more general classes.

**Standard random construction of $G = ([N], [m], \mathcal{E})$**

For every $u \in [N]$, sample a subset of $[m]$ of size $d$ and connect $u$ and all the vertices from this subset.
Model expanders

- $\mathcal{T}_k$ & $\mathcal{G}_k$ denotes $D$-ary tree & loopless overlapping groups respectively, which are jointly denoted by $\mathcal{M}_k$

**Definition ((Nested) Model sparse vectors)**
A vector $\mathbf{x}$ is $\mathcal{M}_k$-sparse if $\text{supp}(\mathbf{x}) \subseteq \mathcal{K}$ for $\mathcal{K} \in \mathcal{M}_k$

**Definition (($k$, $d$, $\epsilon$)-model expander graph)**
Let $\mathcal{K} \in \mathcal{M}_k$, $G$ is a model expander if for all $S \subseteq \mathcal{K}$, we have $|\Gamma(S)| \geq (1 - \epsilon)d|S|

**Definition (Model expander matrix)**
A matrix $\mathbf{A}$ is a model expander if it is the adjacency matrix of a ($k$, $d$, $\epsilon$)-model expander graph.
Randomized model RIP-1 constructions

**Theorem** \(((k, d, \varepsilon))-model expanders for \(D\)-ary \((D \geq 2)\) tree models\)

These exist with \(d = O\left(\frac{\log(N/k)}{\varepsilon \log \log(N/k)}\right)\) and \(m = O\left(\frac{dk}{\varepsilon}\right)\).

- **Note**: \(D\) is subsumed (as \(\log(D)\)) in the order constant for \(m\)
- This matches bounds for binary tree models by [I. & R. ’13]

**Theorem** \(((k, d, \varepsilon))-model expanders for overlapping group models\)

For \(M > 2\) number of groups of maximum size \(g_{\text{max}} = \omega(\log N)\) such that \(N \geq kg_{\text{max}}\), these exist with \(d = O\left(\frac{\log(N)}{\varepsilon \log(kg_{\text{max}})}\right)\) and \(m = O\left(\frac{dkg_{\text{max}}}{\varepsilon}\right)\).

- This matches bounds for block sparsity models by [I. & R. ’13]
- **Note**: Block sparse models are a subset of the loopless overlapping group sparsity models
Proof sketch

- **Proof technique** similar to those of [Indyk & Razenshteyn '13]
- Key ingredient of the proof is the **standard tail inequality**

Lemma (For $G = ([N], [m], \mathcal{E})$, a variant proven in [Buhrman et al. 2002])

There exist $C > 1$ and $\mu > 0$ such that, whenever $m \geq Cdt/\epsilon$, for any $T \subseteq [N]$ with $|T| = t$ we have:

\[
\text{Prob}\left[|\{j \in [m] : \exists i \in T, e_{ij} \in \mathcal{E}\}| < (1 - \epsilon)dt\right] \leq \left(\mu \frac{\epsilon m}{dt}\right)^{-\epsilon dt}
\]

- Then a **union bound** over all $M_k$-sparse sets of sparsity $t$
- The **enumeration** of the cardinality of these sets involves
  - Pfaff-Fuss-Catalan or $k$-Raney numbers for $\mathcal{T}_k$
  - a careful **counting** of such groups in $\mathcal{G}_k$
Lemma ([Bah, Baldassarre, and Cevher 2014])

Let $\mathcal{T}_k$-sparse and $\mathcal{G}_k$-sparse sets with sparsity $t$ be $\mathcal{T}_{k,t}$ and $\mathcal{G}_{k,t}$ respectively and the Catalan number be $T_k$, then $|\mathcal{T}_{k,t}| \leq \min \left[ T_k \binom{k}{t}, \binom{N}{t} \right]$, $|\mathcal{G}_{k,t}| \leq \min \left[ \binom{M}{k} \binom{kg_{\text{max}}}{t}, \binom{N}{t} \right]$. 

- It suffice to show that the following holds:
  \[
  |\mathcal{M}_{k,t}| \cdot \left( \frac{\epsilon m}{dt} \right)^{-\epsilon dt} \leq f(N)
  \]
  where $f(N) = \text{decaying function of } N$, we used $f(N) = 1/N$.

- First, bound $|\mathcal{M}_{k,t}|$ using the fact that $\binom{n}{s} \leq \left( \frac{en}{s} \right)^s$.

- Substitute for $d$ and $m$ as given with arbitrary order constants.

- Finally, show that for different values of $t \in [1, k]$ for $\mathcal{T}_{k,t}$ or $t \in [1, kg_{\text{max}}]$ for $\mathcal{G}_{k,t}$, this bound holds.
Expander Iterative Hard Thresholding (EIHT)

Initialize $x^0 = 0$, iterate

$$x^{n+1} = \mathcal{H}_k [x^n + M(y - Ax^n)]$$

- $M(\cdot)$ is the median operator which returns a vector $M(u) \in \mathbb{R}^N$ for an input $u \in \mathbb{R}^m$; defined elementwise

  $$[M(u)]_i = \text{median}[u_j, j \in \Gamma(i)], \, i \in [N]$$

- **Note:** $M$ operates like an adjoint
Model-Expander Iterative Hard Thresholding (MEIHT)

Initialize $x^0 = 0$, iterate

$$x^{n+1} = \mathcal{P}_{M_k} \left[ x^n + M \left( y - Ax^n \right) \right]$$

- $\mathcal{P}_{M_k}(u) \in \operatorname{argmin}_{z \in M_k} \{ ||u - z||_1 \}$ is projection of $u$ onto $M_k$
- $\mathcal{P}_{M_k}(\cdot)$ replaces $\mathcal{H}_k(\cdot)$ in EIHT
Key features

**Runtime:**

*linear in* $N$ *for all tractable models*

- Thus, using projection complexity from [Baldassarre et al. 2013], for a fixed $n$ MEIHT achieves *linear* runtime of:
  - $O(knN)$ for the $T_k$ model
  - $O(M^2kn + nN)$ for the $G_k$ model; $M$ groups

**Error guarantees:** $\ell_1 / \ell_1$ *in the for all case*

\[
\|x - \hat{x}\|_1 \leq C_1 \sigma_{M_k}(x)_1 + C_2\|e\|_1
\]

where $C_1, C_2 > 0$ and $\sigma_{M_k}(x)_1 := \min_{x' \in M_k} \|x - x'\|_1$

- **Approximate solutions** are in the model, $M_k$; this is very useful for some applications
Tractability of structured sparse models

- But \( \min_{\mathbf{z}: \text{supp}(\mathbf{z}) \in \mathcal{M}} \| \mathbf{z} - \mathbf{u} \|_1 = \max_{\text{supp}(\mathbf{z}) \subseteq S \in \mathcal{M}} \| \mathbf{u}_S \|_1 \equiv \text{Weighted Max Cover (WMC)} \) for group-sparse problems
- All WMC instances can be formulated as \( \mathcal{P}_\mathcal{M}(\cdot) \)
- **Caveat:** WMC is NP-hard \( \Rightarrow \mathcal{P}_\mathcal{M}(\cdot) \) is NP-hard too
- **But:** for some \( \mathcal{M}, \mathcal{M}_k \) (i.e. \( \mathcal{T}_k \) & \( \mathcal{G}_k \)), \( \exists \) linear time algorithms
- These include **dynamic programs** that recursively compute the optimal solution via the model graph [Baldassarre, Bhan, Cevher 2013]
Proof sketch

Lemma (Key ingredient of proof)

Let $A \in \{0, 1\}^{m \times N}$ be a $(k, d, \epsilon_{M_k})$-model expander. If $S \subset [N]$ is $M_k$-sparse, then for all $x \in \mathbb{R}^N$ and $e \in \mathbb{R}^m$,

$$\| [M(Ax_S + e) - x]_S \|_1 \leq \frac{4\epsilon_{M_k}}{1 - 4\epsilon_{M_k}}\|x_S\|_1 + \frac{2}{(1 - 4\epsilon_{M_k})d}\|e_{\Gamma(S)}\|_1$$

- Proof of lemma uses the unique neighbor property

Proposition (Unique neighbor property)

Given a $d$-left regular bipartite graph with expansion coefficient $\epsilon$ if $S$ is a set of $s$ left indices, then the set, $\Gamma'(S)$ of right vertices connected to exactly one left vertex in $S$ has size $|\Gamma'(S)| \geq (1 - 2\epsilon)ds$. 
Lemma

Let $A \in \{0, 1\}^{m \times N}$ be a $(k, d, \epsilon_{M_k})$-model expander. If $S \subset [N]$ is $M_k$-sparse, then for all $x \in \mathbb{R}^N$ and $e \in \mathbb{R}^m$,

$$
\| [M (Ax_S + e) - x]_S \|_1 \leq \frac{4 \epsilon_{M_k}}{1 - 4 \epsilon_{M_k}} \|x_S\|_1 + \frac{2}{(1 - 4 \epsilon_{M_k}) d} \|e_{\Gamma(S)}\|_1
$$

- For $Q^{n+1} := S \cup \text{supp}(x^n) \cup \text{supp}(x^{n+1})$, the triangle inequality yields
  $$
  \|x^{n+1} - x_S\|_1 \leq 2 \| [x_S - x^n - M (A(x_S - x^n) + Ax_{\tilde{S}} + e)]_{Q^{n+1}} \|_1
  $$

- Using the nestedness property of $M_k$ and the lemma gives:
  $$
  \|x^{n+1} - x_S\|_1 \leq \frac{8 \epsilon_{M_{3k}}}{1 - 4 \epsilon_{M_{3k}}} \|x_S - x^n\|_1 + \frac{4}{(1 - 4 \epsilon_{M_{3k}}) d} \|Ax_{\tilde{S}} + e\|_1
  $$

- Taking $\lim_{n \to \infty} x^n = \hat{x}$, using the RIP-1 property of $A$ and the triangle inequality with the condition $\epsilon_{M_{3k}} < 1/12$, we have:
  $$
  \|\hat{x} - x\|_1 \leq C_1 \sigma_{M_k}(x)_1 + C_2 \|e\|_1, \quad C_2 = \beta = \frac{4}{(1 - 12 \epsilon_{M_{3k}}) d} , \quad C_1 = 1 + \beta d
  $$
Simulations, with different $N$, on group and tree models
The median over different realizations of the minimum no. of samples for which $\frac{||\hat{x} - x||_1}{||x||_1} \leq 10^{-5}$ is plotted for MEIHT & EIHT

**Group sparse**

- $M = \lfloor N / \log_2(N) \rfloor$, $g = \lfloor N / M \rfloor$, $k = 5$, $d = \lfloor 2 \log(N) / \log(kg) \rfloor$

**Tree sparse**

- $m \in [2k, 10 \log_2(N)]$, $k = \lfloor 2 \log_2(N) \rfloor$, $d = \lfloor 5 \log(N/k) / (2 \log \log(N/k)) \rfloor$

MEIHT requires fewer measurements than EIHT as expected
Summary

- Model expanders = model-based sketching + sparse matrices; ⇒ improvement in sampling and recovery
- Proposed an efficient algorithm with linear runtime for models considered & achieves $\ell_1/\ell_1$ guarantees in the for all case
- Random construction of model expanders for more a general class of models provably possible
References


THANK YOU