Model-based Sketching and Recovery with Expanders

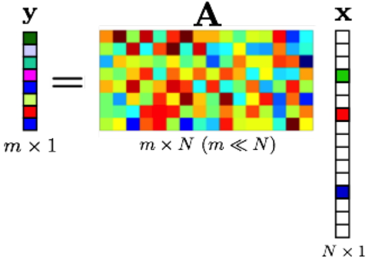
Bubacarr Bah, Luca Baldassarre, and **Volkan Cevher**

Laboratory for Information and Inference Systems (LIONS), EPFL

Information Theory and Applications
San Diego

Three key aspects of linear sketching

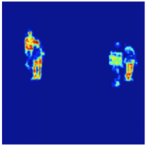
- Sparse or compressible **x**
  not sufficient alone
- Projection **A**
  information preserving
  (stable embedding)
- Recovery algorithm Δ
  tractable & correct

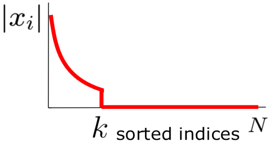


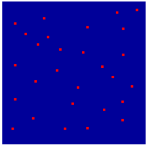
- Applications: Data streaming, compressive sensing (CS), graph sketching, machine learning, group testing, etc.

Sparsity and beyond

- Generic sparsity (or compressibility) not always enough

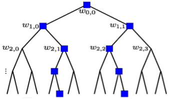






- Structured sparsity ⇒ model-based CS [Baraniuk, Cevher, Duarte, Hegde, IEEE Transactions on Information Theory 2010]:

tree-sparse

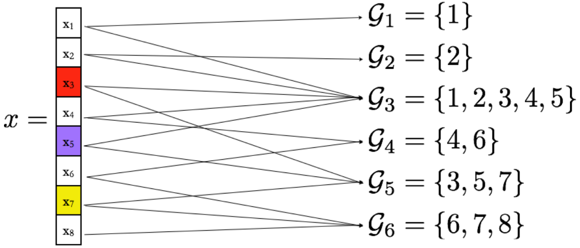


- Model-based CS exploits structure in sparsity model
  - improves interpretability
  - reduces sketch length
  - increases speed of recovery

Block-sparse

Overlapping Group Models

A natural generalization of sparsity



Group models application examples:

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

Information preserving linear embeddings **A**

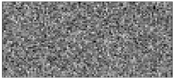
Definition (ℓ_p-norm Restricted Isometry Property (RIP-p))

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

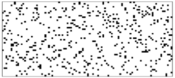
$$(1-\delta_k)\|\mathbf{x}\|_p^p \leq \|\mathbf{A}\mathbf{x}\|_p^p \leq (1+\delta_k)\|\mathbf{x}\|_p^p$$

- Subgaussian **A** $\in \mathbb{R}^{m \times N}$ (w.h.p) have RIP-2 with $m = \mathcal{O}(k \log(N/k))$, but sparse binary **A** does not have RIP-2 unless $m = \Omega\left(k^2\right)$
- Model sparsity requires fewer m for RIP-2
  - $\mathcal{O}(k)$ for tree structure
  - $\mathcal{O}(k + \log(M))$ for block structure with M blocks [Baraniuk et al. '10]
- Scaled adjacency mat. of lossless expanders have RIP-1 with $m = \mathcal{O}(k \log(N/k))$

A



A



Recovery algorithms

- Tractable recovery algorithms (Δ) with provable guarantees
  - Convex ℓ_1-minimization approaches, and
  - Discrete algorithms (OMP, IHT, CoSaMP, ALPS)
- Δ returns approximations with ℓ_p/ℓ_q-approximation error:

Definition (ℓ_p/ℓ_q-approximation error - instance optimality)

A Δ returns $\hat{\mathbf{x}} = \Delta(\mathbf{A}\mathbf{x} + \mathbf{e})$ with ℓ_p/ℓ_q-approximation error if

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

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

- The pair $(\mathbf{A}, \Delta)$ $\Rightarrow$ two types of error guarantees
  - *for each* - one pair $(\mathbf{A}, \Delta)$ for each given $\mathbf{x}$
  - *for all* - one pair $(\mathbf{A}, \Delta)$ for all $\mathbf{x}$

Goal of this work

To combine benefits of sparsity in **A** and benefits of model-based CS

- Prior work on model-based CS use dense **A**
- Difficult to store, creates computational bottlenecks, and not practical in real applications

Our results in perspective

	Price 2011	I. & R. 2013	this work
Models (structures)	standard	tree[1]	tree[2] & groups[3]
Error guarantees	ℓ_2/ℓ_2	ℓ_1/ℓ_1	ℓ_1/ℓ_1
Guarantee types	for each	for all	for all
Recovery algorithm	sublinear	exponential	polynomial

[1] binary trees, [2] D-ary trees for $D \geq 2$, [3] Loopless overlapping groups

Contribution summary

- **Primary**: "Tractable" algorithm with provable *for all* ℓ_1/ℓ_1 error
- **Secondary**: Existence of model expander (model-RIP-1) **A**, consistent with known sampling bounds, for more general models

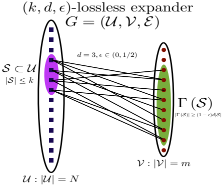
PART I: Existence of Model Expanders

Definition (RIP-1 for (k, d, ϵ)-lossless expanders)

If $\mathbf{A}$ is an adjacency matrix of a (k, d, ϵ)-lossless expanders, then $\Phi = \mathbf{A}/d$ has RIP-1 of order k, if for all k-sparse $\mathbf{x}$, it satisfies

$$(1 - 2\epsilon)\|\mathbf{x}\|_1 \leq \|\Phi\mathbf{x}\|_1 \leq \|\mathbf{x}\|_1$$

- Probabilistic constructions of expanders achieve optimal $m = \mathcal{O}\left(k \log(N/k)\right)$
- But their deterministic constructions are sub-optimal $m = \mathcal{O}\left(k^{1+\alpha}\right)$ for $\alpha > 0$



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

Models everywhere

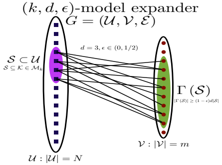
- $\mathcal{T}_k$ & $\mathfrak{G}_k$ denotes D-ary tree & loopless overlapping groups respectively, which are jointly denoted by $\mathcal{M}_k$

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, ϵ)-model expander graph)

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



Definition (Model expander matrix)

A matrix $\mathbf{A}$ is a model expander if it is the adjacency matrix of a (k, d, ϵ)-model expander graph.

Randomized model RIP-1 constructions

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

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

- ***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, ϵ)-model expanders for overlapping group models)

For $M > 2$ *number of groups of maximum size* $g_{\max} = \omega(\log N)$ *such that* $N \geq k g_{\max}$, *these exist with* $d = \mathcal{O}\left(\frac{\log(N)}{\epsilon \log(k g_{\max})}\right)$ *and* $m = \mathcal{O}\left(\frac{dk g_{\max}}{\epsilon}\right)$.

- 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

Our approach-I

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

- Then a union bound over all $\mathcal{M}_k$-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 $\mathfrak{G}_k$

Our approach-II

Lemma ([Bah, Baldassarre, and Cevher 2014])

Let $\mathcal{T}_k$-sparse & $\mathfrak{G}_k$-sparse sets with sparsity t be $\mathcal{T}_{k,t}$ & $\mathfrak{G}_{k,t}$ respectively & the Catalan no. be T_k, then $|\mathcal{T}_{k,t}| \leq \min\left[T_k\binom{k}{t}, \binom{N}{t}\right]$, $|\mathfrak{G}_{k,t}| \leq \min\left[\binom{M}{k}\binom{kg_{\max}}{t}, \binom{N}{t}\right]$

- It suffice to show that the following holds

$$\left|\mathcal{M}_{k,t}\right| \cdot \left(\mu \frac{\epsilon m}{dt}\right)^{-\epsilon dt} \leq f(N)$$

  where $f(N) =$ decaying function of N, we used $f(N) = 1/N$
- First, bound $\left|\mathcal{M}_{k,t}\right|$ using the fact that $\binom{n}{s} \leq \left(\frac{en}{s}\right)^s$
- Substitute for d & m as given with arbitrary order constants
- Finally, show that for different values of $t \in [1, k]$ for $\mathcal{T}_{k,t}$ or $t \in [1, kg_{\max}]$ for $\mathfrak{G}_{k,t}$, this bound holds

■

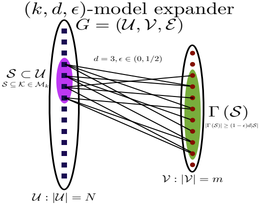
PART II: Model Expander Algorithm

Model-Expander Iterative Hard Thresholding (MEIHT)

Initialize $\mathbf{x}^0 = \mathbf{0}$, iterate

$$\mathbf{x}^{n+1} = \mathcal{P}_{\mathcal{M}_k}\left[\mathbf{x}^n + \mathfrak{M}\left(\mathbf{y} - \mathbf{A}\mathbf{x}^n\right)\right]$$

- $\mathfrak{M}(\cdot)$ is the median operator which returns a vector $\mathfrak{M}(\mathbf{u}) \in \mathbb{R}^N$ for an input $\mathbf{u} \in \mathbb{R}^m$; defined elementwise $[\mathfrak{M}(\mathbf{u})]_i := \text{median}[u_j, j \in \Gamma(i)], i \in [N]$
- ***Note***: $\mathfrak{M}$ operates like an adjoint



- $\mathcal{P}_{\mathcal{M}_k}(\mathbf{u}) \in \text{argmin}_{\mathbf{z} \in \mathcal{M}_k} \{\|\mathbf{u} - \mathbf{z}\|_1\}$ is the projection of $\mathbf{u}$ onto $\mathcal{M}_k$
- MEIHT is a fusion of various works [Berinde & Indyk 2008; Foucart & Rauhut 2013; Baldassare, Bhan, and Cevher 2013; Baraniuk, Cevher, Duarte, and Hegde 2010].

Tractability of structured sparse models

- $\min_{\mathbf{z}:\text{supp}(\mathbf{z})\in\mathcal{M}} \|\mathbf{z}-\mathbf{u}\|_1 = \max_{\text{supp}(\mathbf{z})\subseteq\mathcal{S}\in\mathcal{M}} \|\mathbf{u}_{\mathcal{S}}\|_1 \equiv$ 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. $\mathcal{T}_k$ & $\mathfrak{G}_k$) in particular, $\exists$ linear time algorithms
- These include dynamic programs that recursively compute the optimal solution via the model graph [Baldassarre, Bhan, Cevher 2013]

Runtime: *polynomial* in N for all tractable models

- Thanks to the sparsity of $\mathbf{A}$, the model projections are the dominant operation in MEIHT
- Thus, using projection complexity from [Baldassarre et al. 2013], for a fixed n MEIHT achieves linear runtime of:
  - $\mathcal{O}(knN)$ for the $\mathcal{T}_k$ model
  - $\mathcal{O}(M^2 kn + nN)$ for the $\mathfrak{G}_k$ model; M groups

Error guarantees: ℓ_1/ℓ_1 in the *for all* case

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 \leq C_1 \sigma_{\mathcal{M}_k}(\mathbf{x})_1 + C_2 \|\mathbf{e}\|_1$$

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

- Approximate solutions are in the model, $\mathcal{M}_k$; this is very useful for some applications

Lemma (Key ingredient of proof)

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

$$\| [\mathfrak{M}(\mathbf{A}\mathbf{x}_{\mathcal{S}} + \mathbf{e}) - \mathbf{x}]_{\mathcal{S}} \|_1 \leq \frac{4\epsilon_{\mathcal{M}_k}}{1 - 4\epsilon_{\mathcal{M}_k}} \|\mathbf{x}_{\mathcal{S}}\|_1 + \frac{2}{(1 - 4\epsilon_{\mathcal{M}_k}) d} \|\mathbf{e}_{\Gamma(\mathcal{S})}\|_1$$

- For $\mathcal{Q}^{n+1} := \mathcal{S} \cup \operatorname{supp}(\mathbf{x}^n) \cup \operatorname{supp}\left(\mathbf{x}^{n+1}\right)$, the triangle inequality yields

$$\|\mathbf{x}^{n+1} - \mathbf{x}_{\mathcal{S}}\|_1 \leq 2\| [\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n - \mathfrak{M}(\mathbf{A}(\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n) + \mathbf{A}\mathbf{x}_{\bar{\mathcal{S}}} + \mathbf{e})]_{\mathcal{Q}^{n+1}} \|_1$$

- Using the nestedness property of $\mathcal{M}_k$ and the lemma gives:

$$\|\mathbf{x}^{n+1} - \mathbf{x}_{\mathcal{S}}\|_1 \leq \frac{8\epsilon_{\mathcal{M}_{3k}}}{1 - 4\epsilon_{\mathcal{M}_{3k}}} \|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n\|_1 + \frac{4}{\left(1 - 4\epsilon_{\mathcal{M}_{3k}}\right) d} \|\mathbf{A}\mathbf{x}_{\bar{\mathcal{S}}} + \mathbf{e}\|_1$$

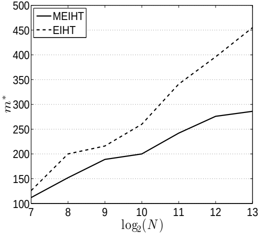
- Taking $\lim_{n \to \infty} \mathbf{x}^n = \hat{\mathbf{x}}$, using the RIP-1 property of $\mathbf{A}$ and the triangle inequality with the condition $\epsilon_{\mathcal{M}_{3k}} < 1/12$, we have:

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_1 \leq C_1 \sigma_{\mathcal{M}_k}(\mathbf{x})_1 + C_2 \|\mathbf{e}\|_1, \quad C_2 = \beta = 4\left(\left(1 - 12\epsilon_{\mathcal{M}_{3k}}\right) d\right)^{-1}, \; C_1 = 1 + \beta d$$

∎

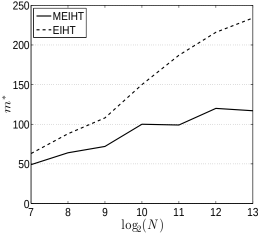
- Simulations, with different N, on group and tree models
- The median over different realizations of the minimum no. of samples for which $\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leq 10^{-5}$ is plotted for MEIHT & EIHT

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

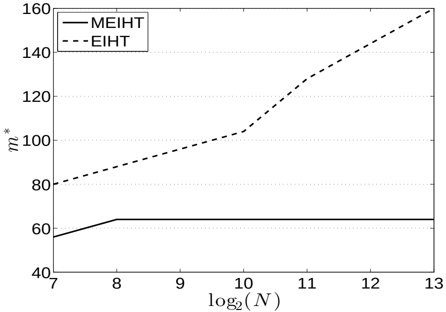


$$m \in [2k, 10\log_2(N)],\ k = \lfloor 2\log_2(N) \rfloor,$$
$$d = \lfloor 5\log(N/k)/(2\log\log(N/k)) \rfloor$$

- MEIHT requires fewer measurements than EIHT as expected

A surprising result

Constant node degree, $d = 16$



Summary

- Model expanders = model-based sketching + sparse matrices; ⇒ improvement in sampling and recovery
- Proposed an efficient algorithm with linear runtime for models considered & achieves ℓ_1/ℓ_1 guarantees in the *for all* case
- Random construction of model expanders for more a general class of models provably possible

Extensions

- Basis adaptivity for when the **x** is sparse in a basis not canonical
- Explicit construction of model expanders
- Application of model expanders to real-life sketching & compressed sensing applications

References

[1] B. Bah, L. Baldassarre, & V. Cevher. *Model-based sketching and recovery with expanders*. ACM-SIAM Symposium on Discrete Algorithms (SODA ’14)

[2] L. Baldassarre, N. Bhan, V. Cevher, & A. Kyrillidis, *Group-sparse model selection: Hardness and relaxations*, arXiv, (2013)

[3] R. Baraniuk, V. Cevher, M. Duarte, & C. Hegde, *Model-based compressive sensing*, IEEE IT. on, 56 (2010), pp. 1982-2001

[4] H. Buhrman, P. Miltersen, J. Radhakrishnan, & S. Venkatesh. *Are Bitvectors Optimal?* SIAM J. Comput., 31(6):1723ÂŰ1744, 2002. 9,13

[5] S. Foucart & H. Rauhut, *A mathematical introduction to compressive sensing*, Applied Numerical Harmonic Analysis Birkhäuser, Boston, (2013)

[6] P. Indyk & I. Razenshteyn, *On model-based RIP-1 matrices*, arXiv:1304.3604, (2013)

[7] E. Price, *Efficient sketches for the set query problem*, in Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2011, pp. 41-56