On Adaptive Sensing for Inference of High-Dimensional Sparse Signals

Rui M. Castro

Based on joint works with Ervin Tánczos

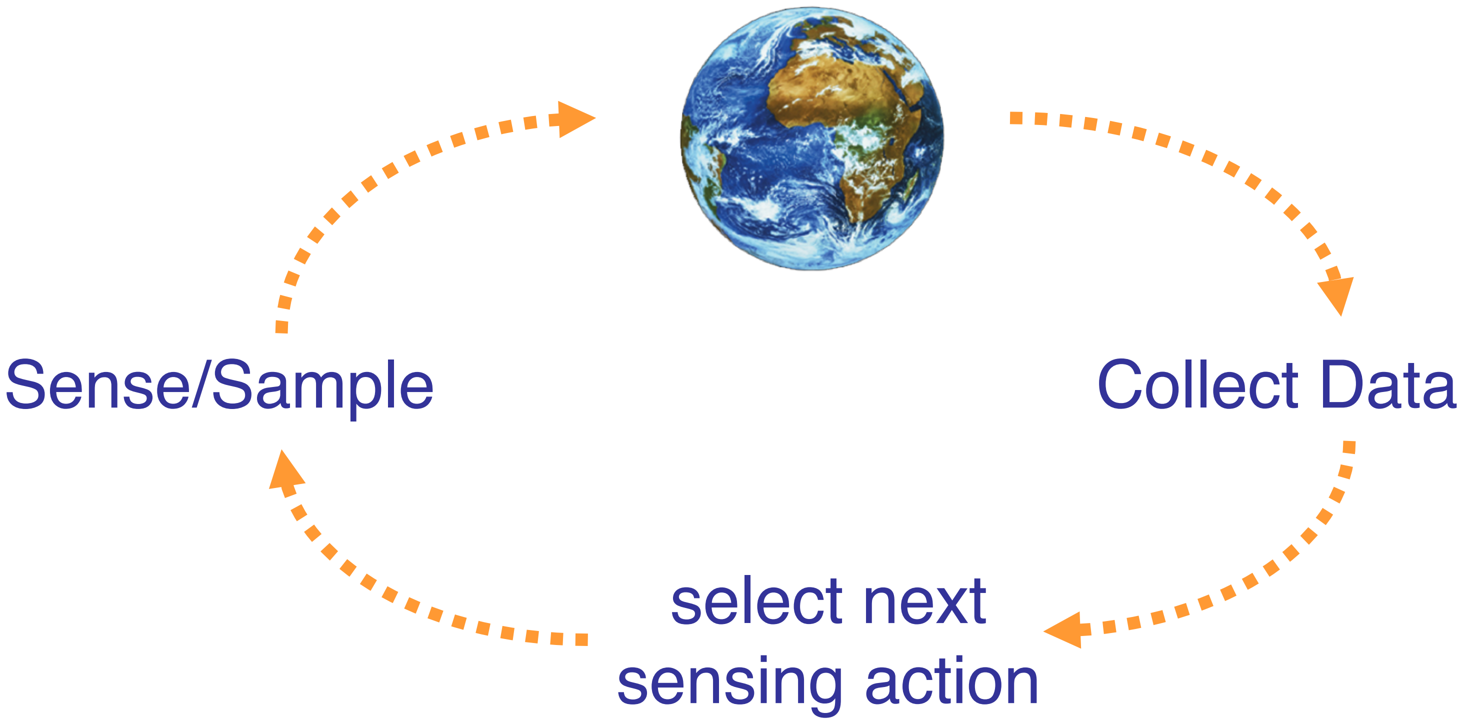
Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Technische Universiteit
Eindhoven
University of Technology

WIC Benelux – Compressed Sensing, January 25th, 2019

Learning About the World

How do we learn about the World?



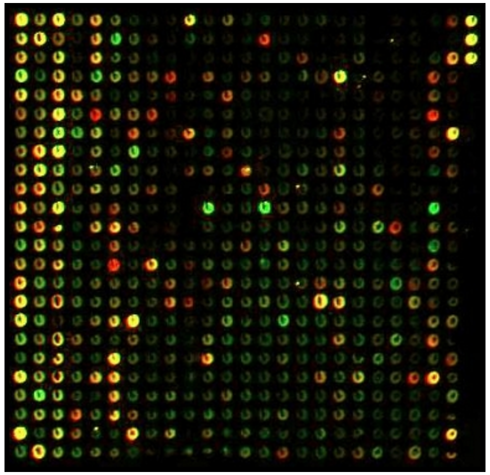
Sense/Sample

Collect Data

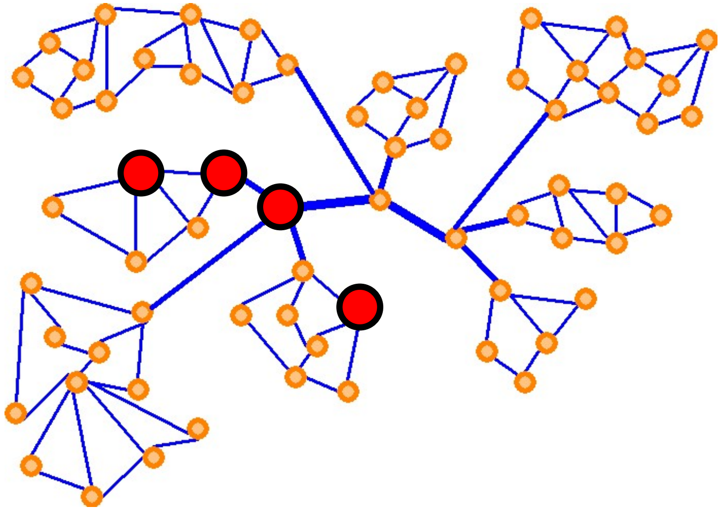
select next
sensing action

- How can we take advantage of this feedback?
- Can we quantify the gains?
- Devise **practical ways** of using this feedback?

Inference of Sparse Signals



Gene expression



Network Anomaly Detection

Can we reliably detect/identify sparse patterns in signals?

- Detect the presence/absence of a sparse signal
- Locate the sparse signal components (support estimation)

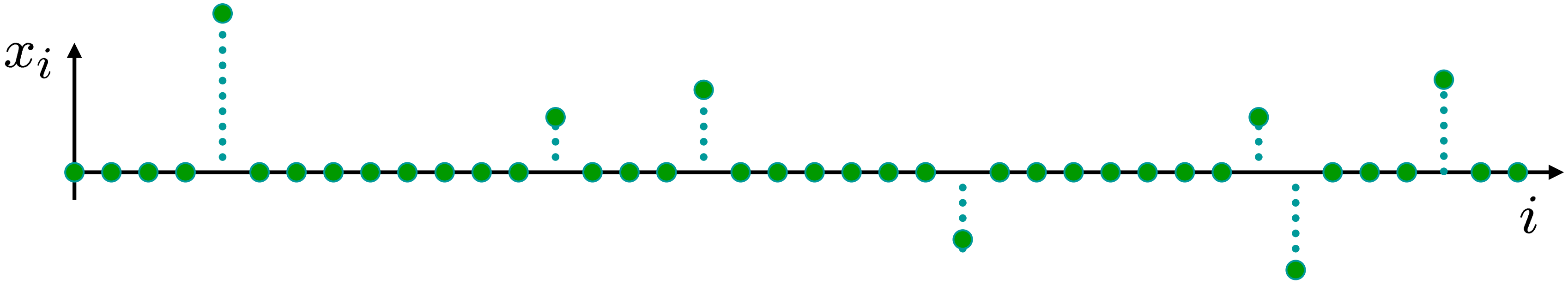
Sparse Signal Models

We are interested in learning about a signal

$$\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \ .$$

E.g.: x_i is the expression level of gene i
x_i "anomalous" traffic level in network node i

We consider situations where $\boldsymbol{x}$ is a sparse vector:



Collecting Noisy Observations

Normal Means (uniform coordinate-wise sensing):

$$Y_i = x_i + W_i, \quad i \in \{1, \ldots, n\}, \quad \text{where } W_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

Make exactly *n* measurements, all with the same precision…

Adaptive coordinate-wise sensing:

The k^{th} observation ($k \in \{1, 2, \ldots\}$) is given by

$$Y_k = x_{A_k} + (\Gamma_k)^{-1/2} W_k, \qquad \text{where } W_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

A_k - entry of x measured at kth observation

Γ_k - precision of the kth measurement

…subject to a **total precision budget**

$$\sum_{k=1}^{\infty} \underbrace{\Gamma_k}_{} \leq n$$

Precision = SNR control

Adaptive vs. Non-Adaptive Sensing

Non-Adaptive sensing:

$\{A_k, \Gamma_k\}_{k=1}^{\infty}$ must be chosen prior to the collection of any observations.

Adaptive Sensing:

A_k, Γ_k are chosen sequentially and are functions of

$$\{Y_\ell, A_\ell, \Gamma_\ell\}_{\ell=1}^{k-1}$$

Key Idea: allow future sensing location and precision to depend on past observations !!!

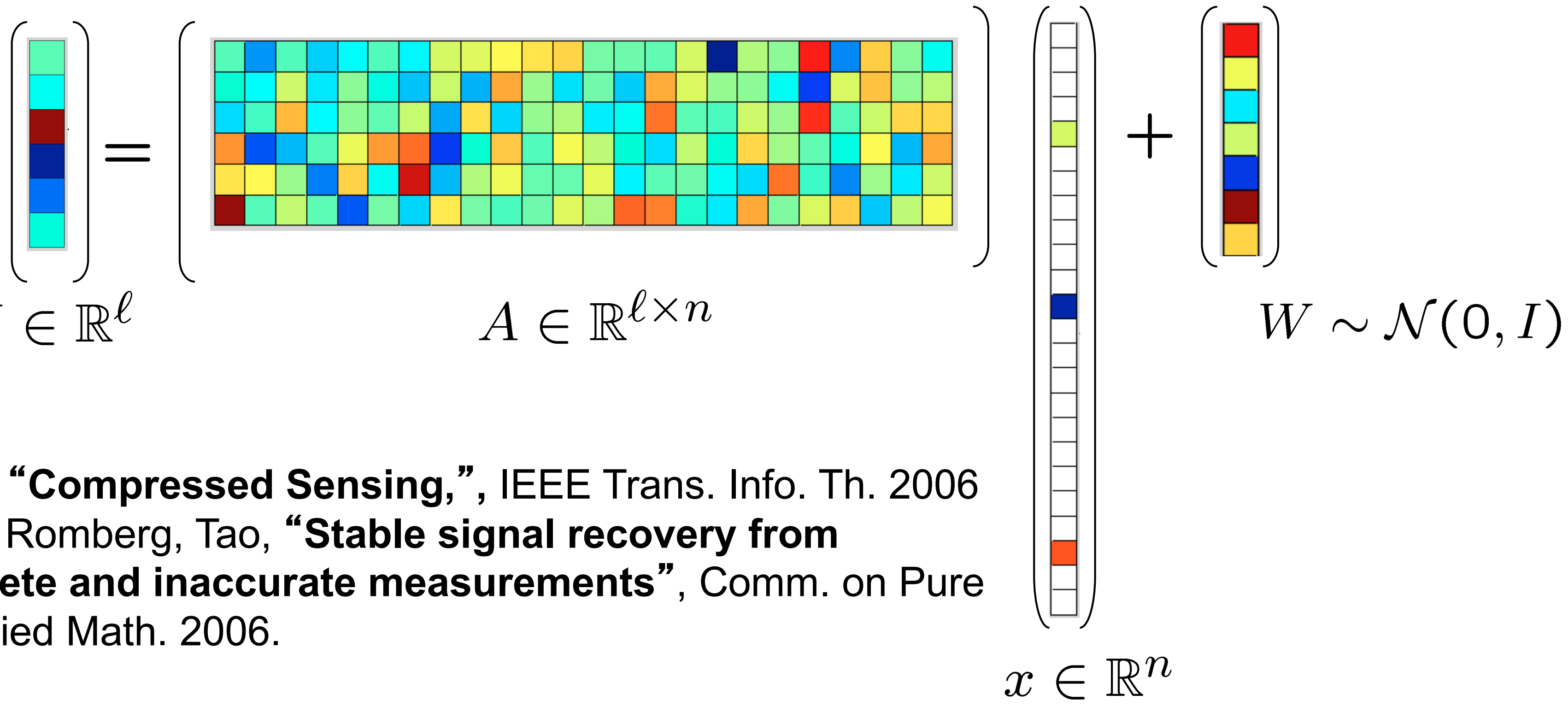
Related settings:

Zehetmayer, Bauer & Posch "**Optimized multi-stage designs controlling the false discovery or the family-wise error rate**", *Statist. Med.* 2008

Hao et al, "**Drosophila RNAi screen identifies host genes important for influenza virus replication**" Nature 2008

Our setting is closely related to that of **multi-armed bandit problems with pure exploration…**

Linear Projections/Compressed Sensing


$Y \in \mathbb{R}^\ell$
$=$
$A \in \mathbb{R}^{\ell \times n}$
$+$
$W \sim \mathcal{N}(0, I)$
$x \in \mathbb{R}^n$

Donoho, **“Compressed Sensing,”**, IEEE Trans. Info. Th. 2006
Candès, Romberg, Tao, **“Stable signal recovery from incomplete and inaccurate measurements”**, Comm. on Pure and Applied Math. 2006.

Adaptive Compressive Sensing: design the rows of A sequentially, based on previous observations.

Sensing budget: $\mathbb{E}[\|A\|_F^2] \leq n$.

- This generalizes the previous coordinate-wise sensing model
- Compressive Sensing can naturally deal with sparsity domains different than the canonical one

Outline

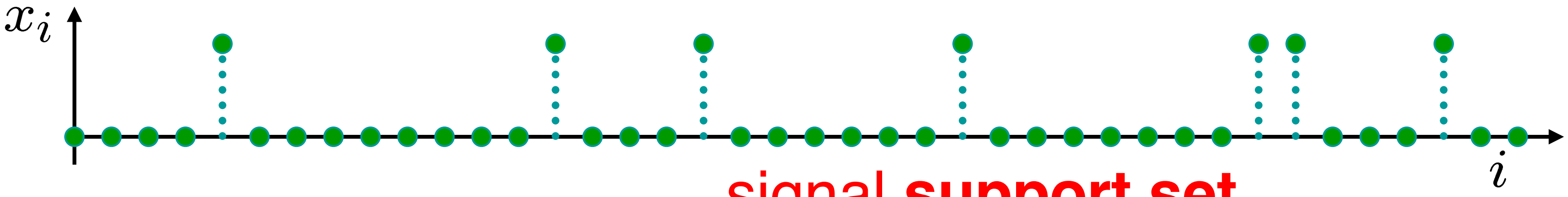
➡ Coordinate-wise Sensing (Adaptive vs. Non-adaptive)

- Dealing with structured sparsity

➡ Compressive Adaptive Sensing

- Statistical guarantees
- Sample-complexity considerations

For simplicity of presentation consider a slightly simpler signal model:



signal **support set**

$$x_i = \begin{cases} \mu & i \in S \\ 0 & i \notin S \end{cases} \quad \text{, where } |S| \ll n \ , \ \mu > 0$$

How small can μ be so that we can still perform reliable inference about S ?

Inference Goals

Assume the set S belongs to a class of subsets $\mathcal{C}$.

Estimation:

Construct a set estimator $\hat{S}$ minimizing:

Expected Hamming Distance: $\max_{S\in\mathcal{C}} \mathbb{E}_S\left[|\hat{S}\Delta S|\right]$

Error Probability: $\max_{S\in\mathcal{C}} \mathbb{P}_S\left(\hat{S} \neq S\right)$

Detection:

$H_0 : S = \emptyset$ vs. $H_1 : S \in \mathcal{C}$

(no signal present) (a signal in the class)

Given a test function $\hat{\Phi} \in \{0,1\}$ minimize

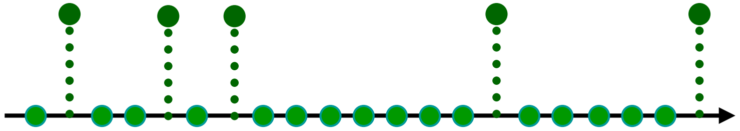
$$R_{\mathcal{C}}(\hat{\Phi}) = \mathbb{P}_{\emptyset}(\hat{\Phi} \neq 0) + \max_{S\in\mathcal{C}} \mathbb{P}_S(\hat{\Phi} \neq 1)\ .$$

Obviously, the difficulty of the problem depends on the class $\mathcal{C}$ under consideration…

Structured and Unstructured Classes

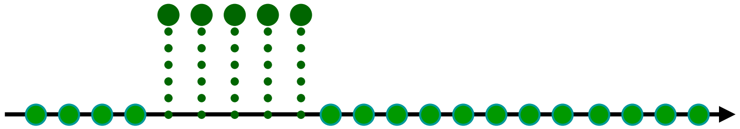
The class of ALL subsets of $\{1, \ldots, n\}$ with cardinality s. In this case we say the signal support **has no structure**:

s-**sets**:



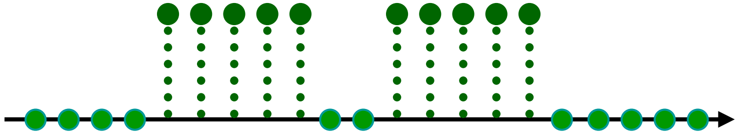
For other classes $\mathcal{C}$ we say the signal support has **structure**:

s-**intervals**:

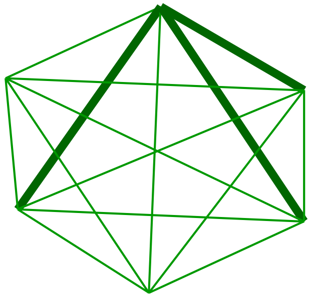


$$\mathcal{C} = \{\{1, \ldots, s\}\ ,\ \{2, \ldots, s+1\}\ ,\ \ldots\ ,\ \{n-s+1, \ldots, n\}\}\ .$$

k **disjoint**
s-**intervals**:



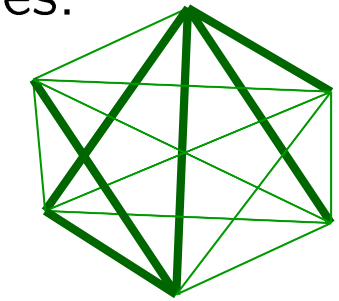
s-**stars in a complete graph with** n **edges**:



Structured and Unstructured Classes

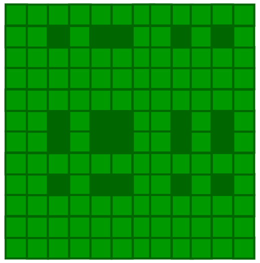
Other interesting structured classes:

unions of k s-stars (with distinct centers)



---

a size s submatrix of a $\sqrt{n} \times \sqrt{n}$ matrix



Structure can be very helpful for **non-adaptive sensing**, both for support estimation or detection:

- Addario-Berry, et al., "**On Combinatorial Testing Problems**", AoS 2010
- Arias-Castro, "**Searching for a trail of evidence in a maze**", AoS 2008
- Butucea, Ingster, "**Detection of a sparse submatrix of a high-dimensional noisy matrix**", Bernoulli, 2014
- Kolar, et al., "**Minimax Localization of Structural Information in Large Noisy Matrices**", NIPS 2011

Signal Detection

Theorem: (normal-means/non-adaptive sens.) (Addario-Berry et al., 2010) Consider the class of all s-sets. If $R_{\mathcal{C}}(\hat{\Phi}) \le \epsilon$ necessarily

$$\mu \ge \sqrt{\log\left(1+\frac{n\log(1+4(1-\epsilon)^2)}{s^2}\right)} \;.$$

If $s \ll \sqrt{n}$ this means $\mu \ge \sqrt{\log\left(n/s^2\right)}$.

Sharper bounds exist (e.g., Ingster ’99, Baraud ’02, Donoho and Jin ’04)

Theorem: (Coordinate-wise Adaptive Sens.) (C. ’12)

If $R_{\mathcal{C}}(\hat{\Phi}) \le \epsilon$ we have necessarily

$$\mu \ge \sqrt{\frac{2}{s}\log\frac{1}{2\epsilon}} \;.$$

- There is a sensing/detection algorithm achieving this bound
- **Structural assumptions cannot further improve this result!!**

Support Estimation

Theorem: (Non-Adaptive Sens.) Let $\mathcal{C}$ denote the class of all s-sets and $\varepsilon > 0$. If

$$\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\Delta S|) \leq \varepsilon \ ,$$

then necessarily $\mu \geq \sqrt{2\log n}$.

Similar bounds can be shown for other classes:

Unions of s-intervals: $\sim \sqrt{\frac{1}{s}\log\frac{n}{s}}$

Unions s-stars: $\sim \sqrt{\log\frac{\sqrt{n}}{s}}$

(loose lower bounds)

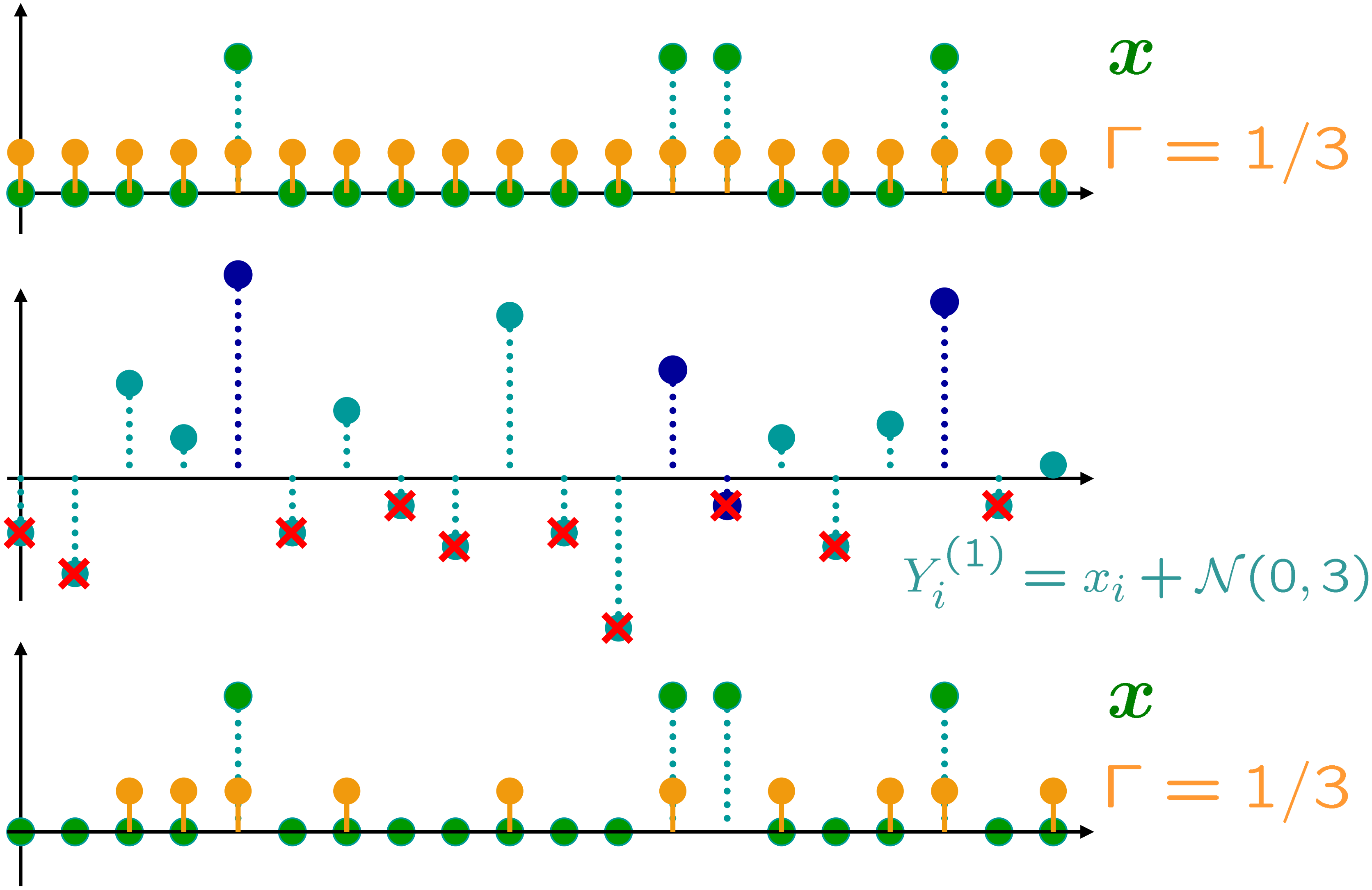
s-**submatrices of** $\sqrt{n}\times\sqrt{n}$ **matrices:** $\sim \sqrt{\frac{1}{\sqrt{s}}\log\frac{n}{s}}$

If signal is sparse ($s \ll n$) then the dependence on the extrinsic dimension is always of the form $\sim\sqrt{\log n}$, regardless of the structure.

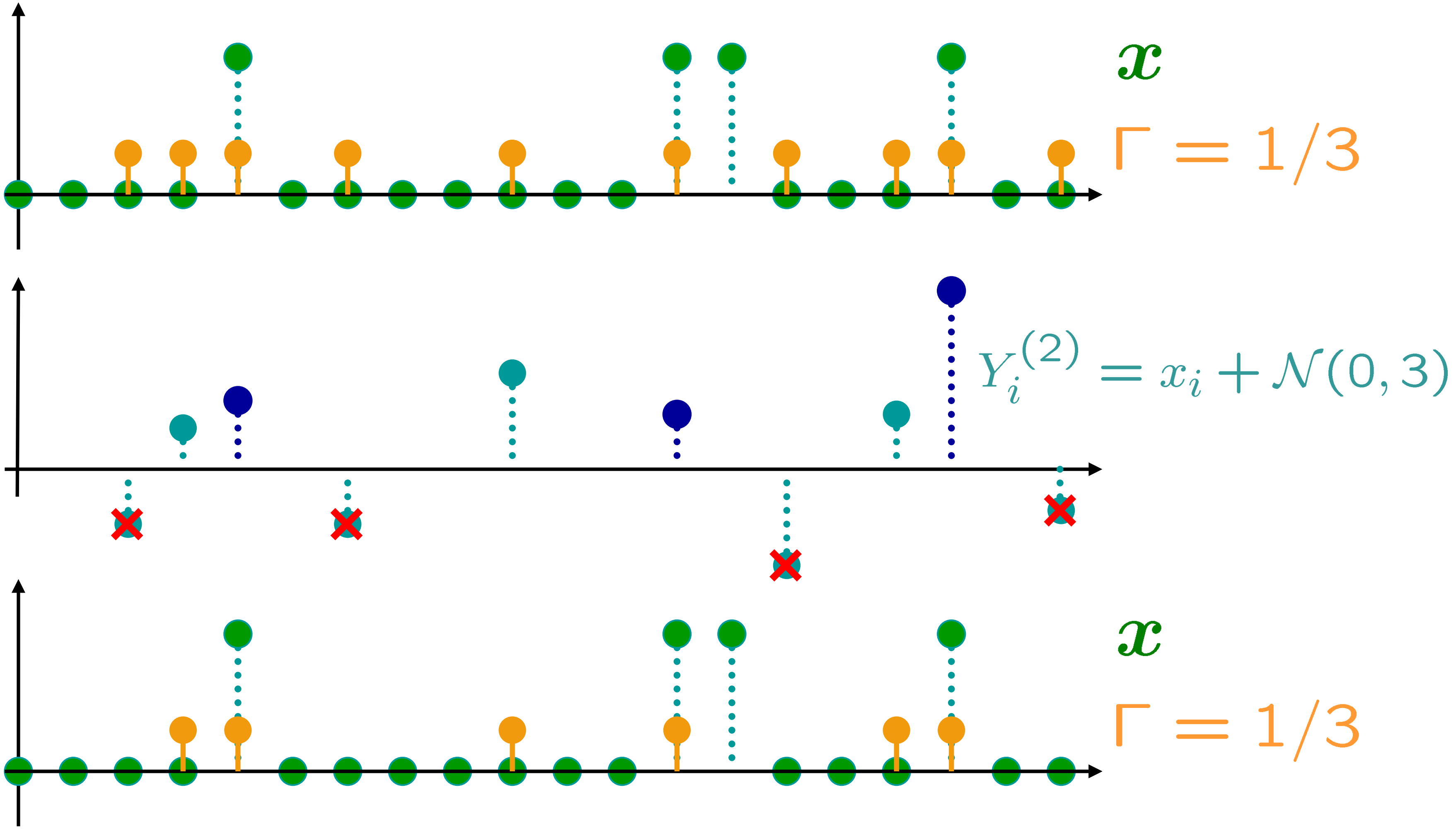
Can adaptive sensing improve upon non-adaptive sensing?

Simple Thresholding

A simple sequential thresholding procedure



Simple Thresholding



and so on for *T* steps… For sparse signals we **remove about half** of the components from further consideration at each step

Simple Thresholding

How much of the precision budget do we use?

$$\mathbb{E}\left[\sum_{i,j}\Gamma\right]=\frac{1}{3}\mathbb{E}\left[\sum_{j=1}^{T}\left|S^{(j-1)}\right|\right]\leq\frac{1}{3}\sum_{j=1}^{T}\left(\frac{n-s}{2^{j-1}}+s\right)\leq\frac{2}{3}(n-s)+Ts\leq n$$

What is the expected number of errors we make?

$$\mathbb{E}\left[|\widehat{S}\Delta S|\right]=\mathbb{E}\left[\left|S^{(T)}\setminus S\right|\right]+\mathbb{E}\left[\left|S\setminus S^{(T)}\right|\right]$$

False positives

$$\leq\frac{n}{2^T}$$

False negatives

$$\leq\frac{Ts}{2}\exp\left(-\frac{\mu^2}{6}\right)$$

Taking $T=\log_2\frac{2n}{\varepsilon}$ ensures $\mathbb{E}\left[|\widehat{S}\Delta S|\right]\leq\varepsilon$ provided

$$\mu\geq\sqrt{6\log s+6\log\log_2\frac{2n}{\epsilon}+6\log\frac{1}{\epsilon}}\ .$$

Adaptive vs. Non-Adaptive

Requirements to ensure that $\max_{S\in\mathcal{C}} \mathbb{E}\left[\left|\widehat{S}\Delta S\right|\right] \to 0, \text{ as } n \to \infty$:

Simple Thresholding:

$$\mu \geq \sqrt{6 \log s + 6.1 \log \log_2 n}$$

Best non-adaptive sensing procedure:

$$\mu \geq \sqrt{2 \log n}\,.$$

Actually, the $\log \log n$ term is an artifact of the simple procedure, and can be removed by either using component-wise SPRT, or using a slighlty more involved thresholding procedure (Malloy-Nowak '12).

- Malloy, Nowak. “**Sequential Testing for Sparse Recovery**,” IT Trans, 2014
- Haupt, C., Nowak, “**Distilled Sensing: Adaptive Sensing for Sparse Detection and Estimation**”, IT Trans, 2011

Can we use these ideas when there is structure?

Idea for a General Procedure

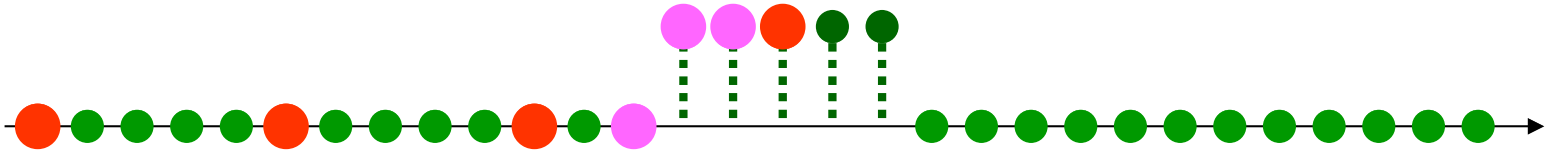
s-sets: the Simple Thresholding/Multiple SPRT approach is a coordinate-wise query of all n entries

A general approach for structured cases:

- Devise a **noiseless** support estimation procedure, making the “minimal” number of queries necessary to uniquely identify the support (and exploring the **combinatorial structure** of the signal class).

- **Robustify** the noiseless procedure to be able to deal with noisy observations, using SPRTs

Example *s*-intervals



Search phase: sequentially sample entries $1, s+1, 2s+1, \ldots$ until a significant component is found

Refinement: sample elements to the left of the significant entry until reaching the end of the interval

Once a noiseless procedure has been chosen, all we need to do is to replace the noiseless queries by SPRTs to ensure

- i) the probability of not recovering the support is small
- ii) The expected total precision spent satisfies the constraint

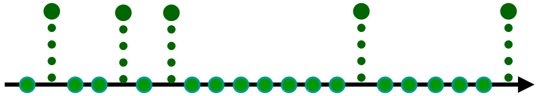
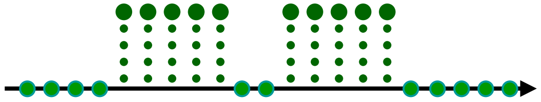
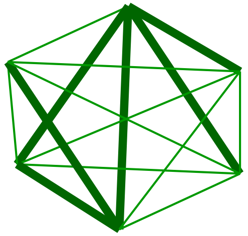
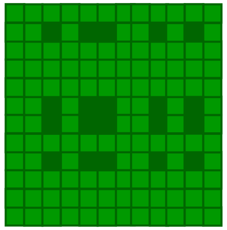
$$\mathbb{E}\left(\sum_{k=1}^{\infty} \Gamma_k\right) \leq n$$

$sk \ll \sqrt{n}$

Upper and Lower Bounds

With careful calibration of the SPRTs one can significantly improve on non-adaptive sensing. Matching adaptive sensing lower bounds can also be derived (but do require some careful work)

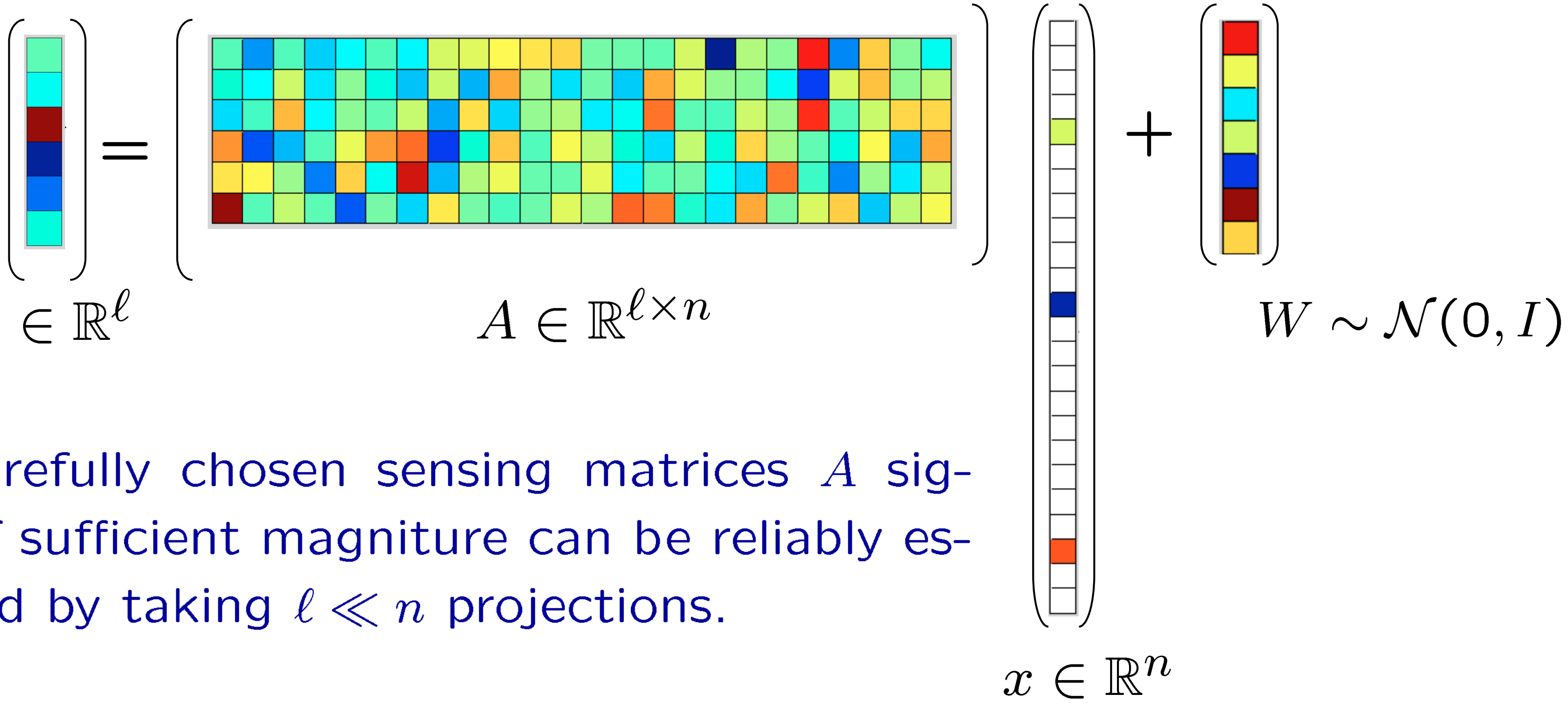
Scaling laws necessary and sufficient to ensure $\max_{S\in\mathcal{C}} \mathbb{E}_S[\widehat{S}\Delta S] \to 0$:

	Non-Adaptive	Adaptive	
	$\sim\sqrt{\log n}$	$\sim\sqrt{\log s}$	$\sim\sqrt{\log s}$
	$\sim\sqrt{\frac{1}{s}\log(n/s)}$	$\sim\sqrt{\frac{1}{s}\log ks}$	$\sim\sqrt{\frac{1}{s}\log ks}$
	$\sim\sqrt{\frac{1}{2}\log\frac{n}{s^2}}$	$\sim\sqrt{\frac{1}{s}\log^2 ks}$	$\sim\sqrt{\frac{1}{s}\log ks}$
	$\sim\sqrt{\frac{1}{\sqrt{s}}\log(n/s)}$	$\sim\sqrt{\frac{1}{s}\log^2 s}$	$\sim\sqrt{\frac{1}{s}\log s}$

Tánczos, C. "**Adaptive Sensing for Estimation of Structured Sparse Signals**," in IEEE IT Trans. , 2015

Linear Projections/Compressed Sensing

Instead of point-samples, one can consider the linear projection measurements



For carefully chosen sensing matrices A signals of sufficient magniture can be reliably estimated by taking $\ell \ll n$ projections.

In this setting the precision budget is conveniently cast as the restriction

$$\mathbb{E}[\|A\|_F^2] \leq n \ .$$

Detection using Linear Projections

Theorem: (Arias-Castro, '12) For reliable detection using adaptive compressive sensing it is necessary and sufficient for the signal magnitude to be of the order

$$\sim \sqrt{\frac{1}{s^2}}\,.$$

compare with $\sim \sqrt{\frac{1}{s}}$ for coordinate-wise sensing

As in the previous setting, **structure does not help for detection.**

Moreover, we can achieve the above bound using only a **non-adaptive sensing** procedure !!!

For estimation the story is different…

Support Estimation with Linear Proj.

Non-Adaptive Sensing: If the number of measurements ℓ is large enough it is still necessary that

$$\mu \sim \sqrt{\log n} \text{ (see e.g., Wainwright, '09)} .$$

Theorem: (Haupt, Baraniuk, C. and Nowak '12) Using $\approx s \log n$ adaptive observations we can estimate the support with probability at least $1 - o(1)$. If the minimum signal amplitude is greater than a constant times

Optimal rate (C. '12)

$$\sqrt{\log s + \log \log_2 \log n} .$$

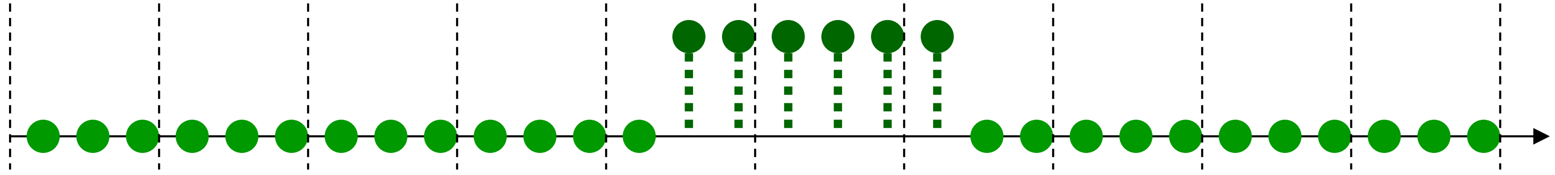
(Some) related work –>

Arias-Castro, et al., “**On the Fundamental Limits of Adaptive Sensing**”, 2011
Arias-Castro, Davenport, “**Compressive Binary Search**”, 2012
C., “**Adaptive Sensing Performance lower Bounds for Sparse Signal Estimation and Testing**,” 2012
Balakrishnan, et al., “**Recovering block-structured activations using compressive measurements**,” 2012
Malloy, Nowak, “**Near-optimal Adaptive Compressed Sensing**”, 2014

Capitalizing on Structural Properties

Same idea as before: alternate between search and refinement stages:

- **Key fact:** compressive sensing leads to a more effective search phase, as it can detect weaker signals.



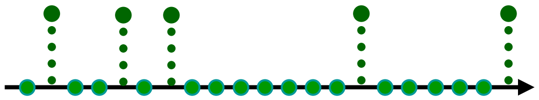
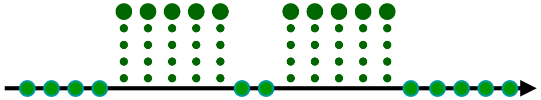
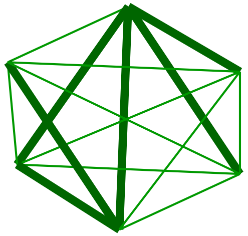
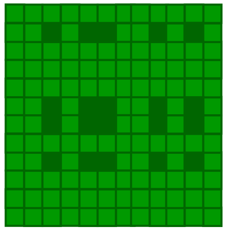
Search phase: Divide the domain is bins of size $s/2$ and use a (actually non-adaptive) CS test to find a bin entirely contained in the interval

Refinement: Focus on the "central" bin above and the neighboring bins, and do coordinate-wise adaptive sensing

$sk \ll \sqrt{n}$

Adaptive Compressed Sensing

Scaling laws necessary and sufficient to ensure $\max_{S\in\mathcal{C}_f} \mathbb{E}_S[\widehat{S}\Delta S] \to 0$:

	Non-Adaptive	**Adaptive**	
	$\sim\sqrt{\log n}$	$\sim\sqrt{\log s}$	$\sim\sqrt{\log s}$
	$\sim\sqrt{\frac{1}{s^2}\log n/s}$	$\sim\sqrt{\frac{1}{s^2}\log ks}$	$\sim\sqrt{\frac{1}{s^2}\log ks}$
	$\sim\sqrt{\log\frac{\sqrt{n}}{s}}$	$\sim\sqrt{\frac{1}{s^2}\log ks}$	$\sim\sqrt{\frac{1}{s^2}\log ks}$
 $\sqrt{s}\times\sqrt{s}$ submatrix	$\sim\sqrt{\frac{1}{\sqrt{s}}\log n/s}$	$\sim\sqrt{\frac{1}{s\sqrt{s}}\log s}$	$\sim\sqrt{\frac{1}{s^2}\log s}$

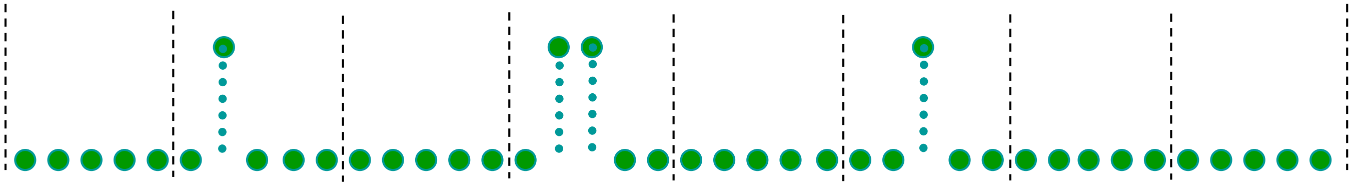
Tánczos, C. "**Adaptive Compressed Sensing for Estimation of Structured Sparse Signals**," IT Trans, 2017

These results characterize the statistical difficulty of the problem, but completely disregard sample complexity...

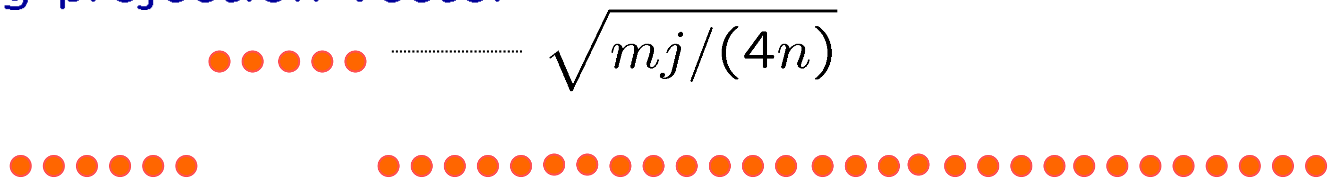
Minimizing Sample Complexity - CASS

For the case of *s*-sparse sets there is a procedure (Compressive Adaptive Sense and Search – CASS) attaining the results in the previous slide, but using only on the order of ~*s* log *n* projections

Malloy, Nowak, "**Near-Optimal Adaptive Compressed Sensing**", IEEE IT, 2014



- $(j = 1)$ partition domain into $2s$ bins (at most s bins will contain signal)
- For each bin test H_0 : bin empty vs. H_1: bin non-empty using projection vector



- $(j \leq \log_2 \frac{n}{2s})$ Split each significant bin in two and repeat the above procedure up to a maximal depth

CASS - Guarantees

Theorem (Malloy and Nowak, 2014): Provided $\mu \geq \sqrt{32 \log \frac{2s}{\varepsilon}}$ we have

$$\mathbb{P}(\widehat{S} \neq S) \leq \varepsilon \ ,$$

and the CASS procedure uses at most $2s \log_2 \frac{n}{2s}$ projections.

The CASS idea of binary bisection can be combined with the search and refinement approach and used to deal with structure as well:

Unions of *k* *s*-intervals:

search: use CASS to find the "middle" of an interval

refinement: use CASS only "around" each found interval

Proposition: Provided $\mu \geq \sqrt{\frac{768}{s^2} \log \frac{3\sqrt{2}ks}{\varepsilon}}$ we have

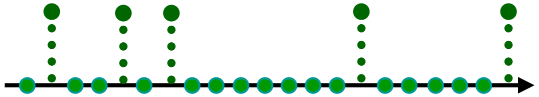
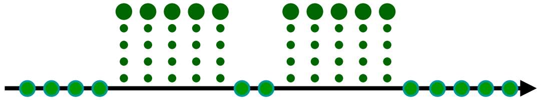
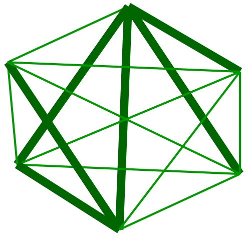
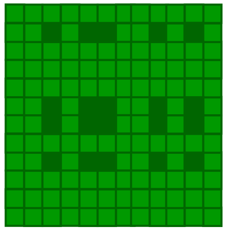
$$\mathbb{E}[|\widehat{S} \Delta S|] \leq \varepsilon \ ,$$

and collects at most $3k\left(\log_2 \frac{n}{2ks} + \frac{3}{2}s\right)$ projections.

$sk \ll \sqrt{n}$

Adaptive Compressed Sensing

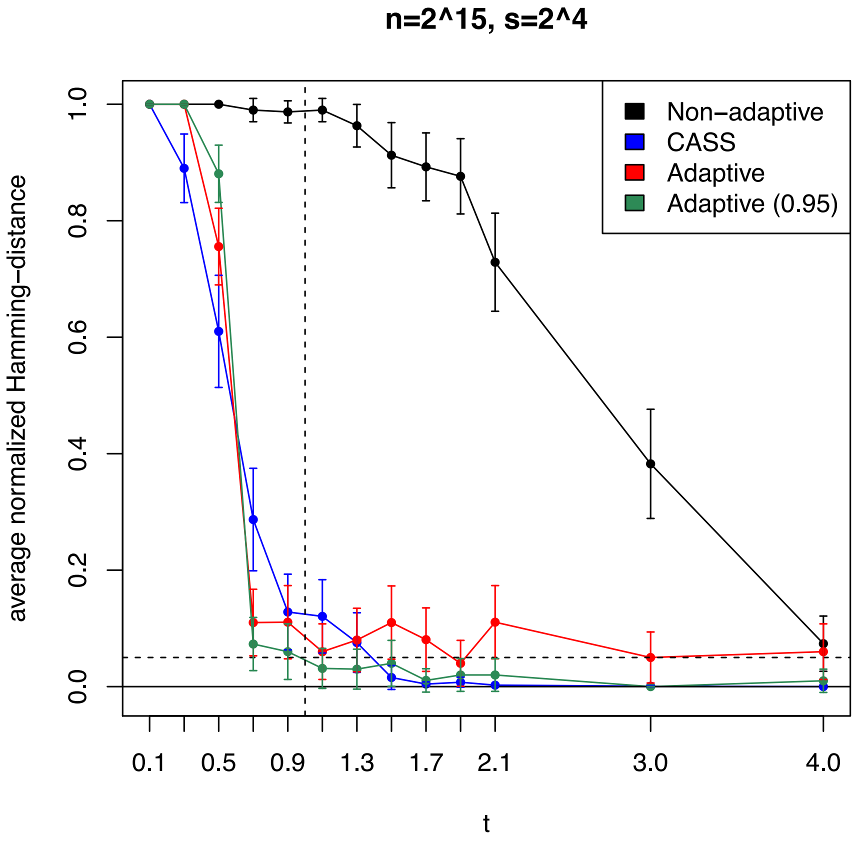
Scaling laws necessary and sufficient to ensure $\max_{S\in\mathcal{C}_f} \mathbb{E}_S[\widehat{S}\Delta S] \to 0$:

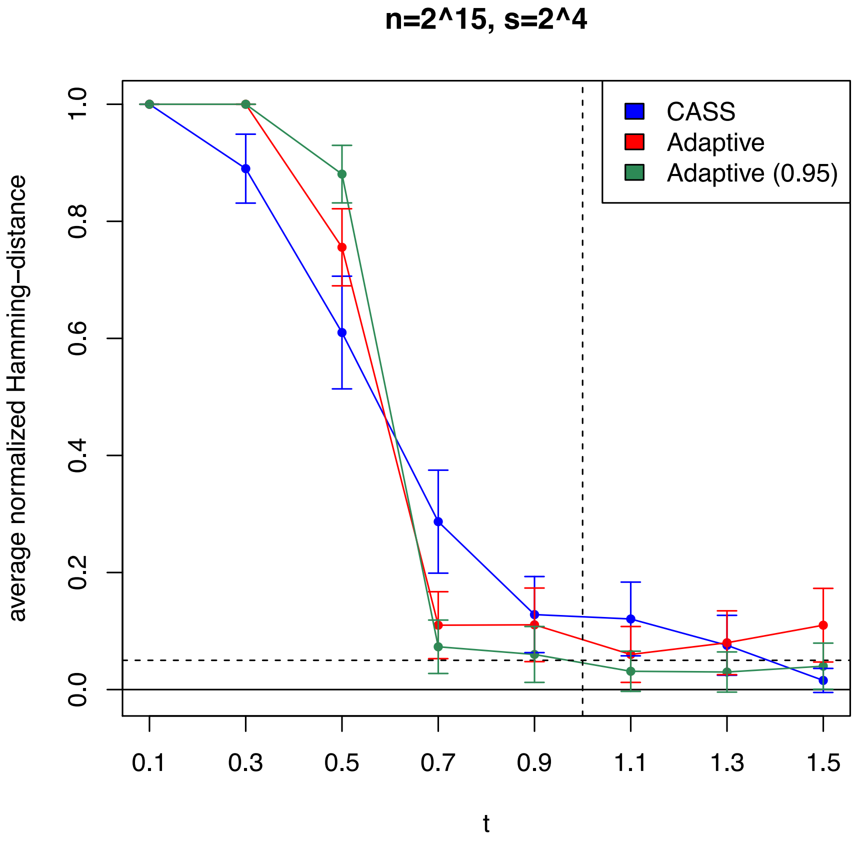
	Non-Adaptive	**Adaptive**	
	$\sim \sqrt{\log n}$	$\sim \sqrt{\log s}$	$\sim \sqrt{\log s}$
	$\sim \sqrt{\frac{1}{s^2}\log n/s}$	$\sim \sqrt{\frac{1}{s^2}\log ks}$	$\sim \sqrt{\frac{1}{s^2}\log ks}$
	$\sim \sqrt{\log \frac{\sqrt{n}}{s}}$	$\sim \sqrt{\frac{1}{s^2}\log ks}$	$\sim \sqrt{\frac{1}{s^2}\log ks}$
 $\sqrt{s}\times\sqrt{s}$ submatrix	$\sim \sqrt{\frac{1}{\sqrt{s}}\log n/s}$	$\sim \sqrt{\frac{1}{s\sqrt{s}}\log s}$	$\sim \sqrt{\frac{1}{s^2}\log s}$

Tánczos, C. "**Adaptive Compressed Sensing for Estimation of Structured Sparse Signals**," IT Trans, 2017

These limits can be attained taking only $\ell \sim s\log n$ projections, by the same principles of the CASS algorithm.

Numerical Results





Even though these procedures (particularly CASS) were designed to ensure worst-case guarantees they still seem to have very reasonable performance

Remarks

- The submatrix case is more delicate than presented here. For non-square matrices one encounters different inference regimes, and there are some gaps between upper and lower bounds…

- Good lower bounds on the sample complexity of adaptive compressive sensing are still needed.

Non-Adaptive Sample Complexity: In order to have vanishing error we must take $\Omega\left(\frac{s\log(n/s)}{\log(\mu^2+1)}\right)$ projections.

Adaptive Sensing Conjecture: Near the estimation threshold we need $\Omega(s\log(n/s))$ projections
(best existing lower bound we are aware of is $\Omega(s)$).

Aksoylar, C., Saligrama, V “**Information-theoretic bounds for adaptive sparse recovery**” (2014)

Final Remarks

- Deriving good sample complexity and performance lower bounds for adaptive (compressed) sensing is tricky. A common challenge (still open): characterize accurately how information contracts when using adaptive sensing…

- There are other interesting settings one can consider:
  - Detection of correlations

    Castro, Savalle, Lugosi, "**Detection of Correlations with Adaptive Sensing**", in IEEE IT Trans., 2014

  - Detection of evolving signals

    Tánczos, C. "**Are there needles in a moving haystack? Adaptive sensing**
    for detection of dynamically evolving signals", to appear in Bernoulli

Thank You