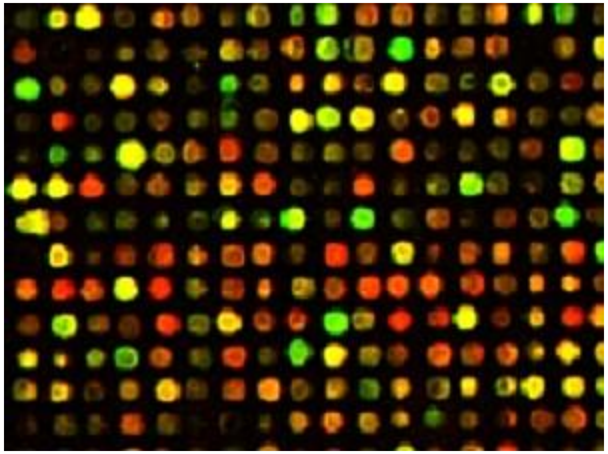
To Adapt or Not To Adapt

The Power and Limits of Adaptivity for Sparse Estimation

Mark A. Davenport

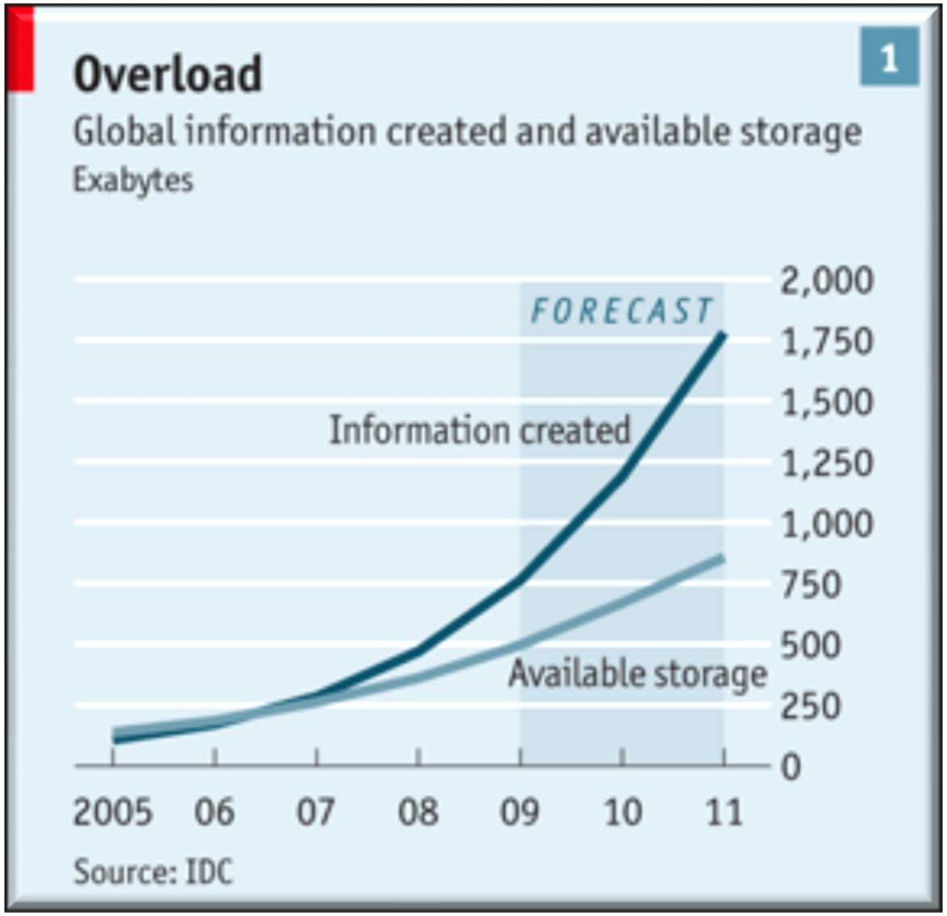
Stanford University
Department of Statistics

Sensor Explosion

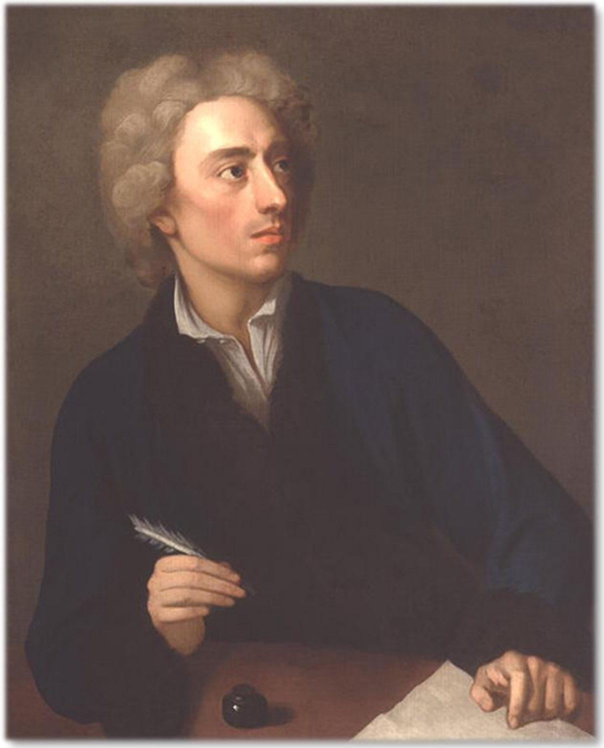


Data Deluge


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


1
Overload
Global information created and available storage
Exabytes
FORECAST
Information created
Available storage
2,000
1,750
1,500
1,250
1,000
750
500
250
0
2005
06
07
08
09
10
11
Source: IDC

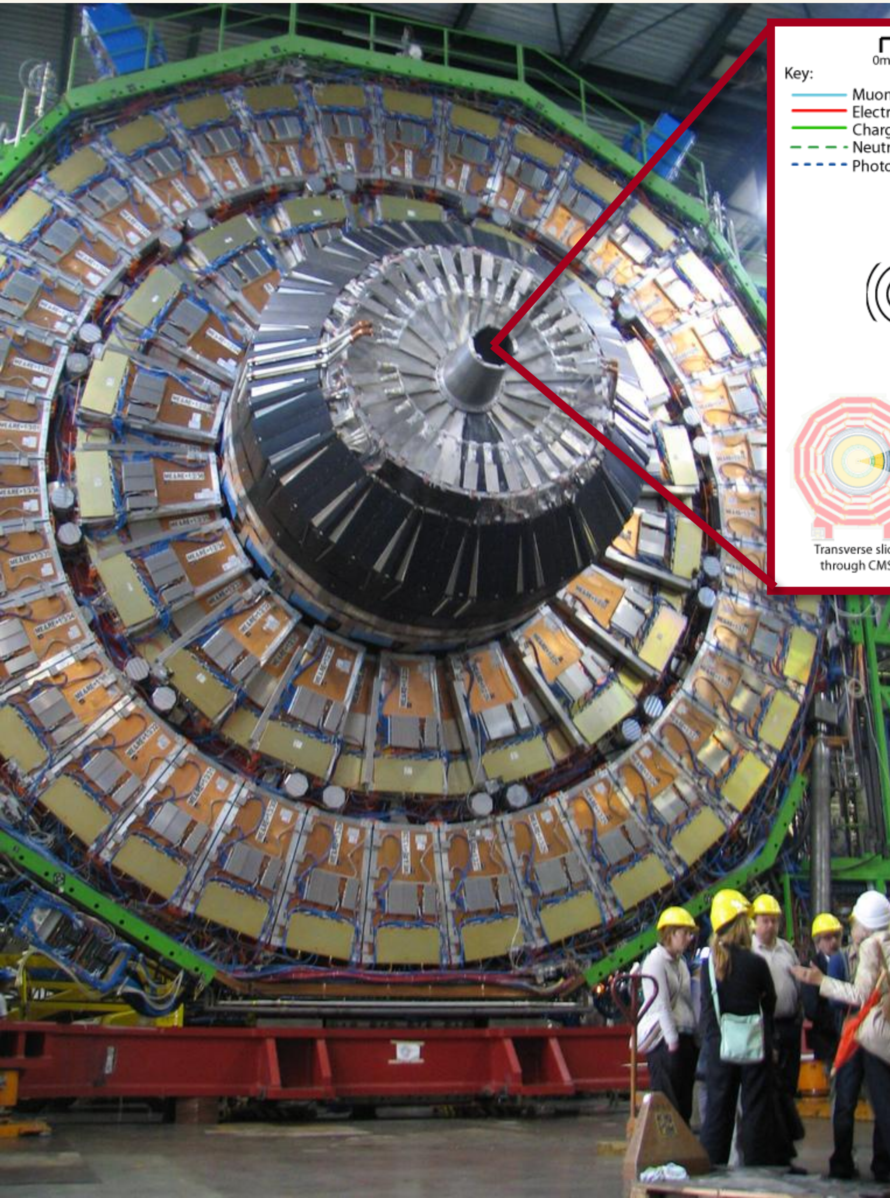
Ye Olde Data Deluge

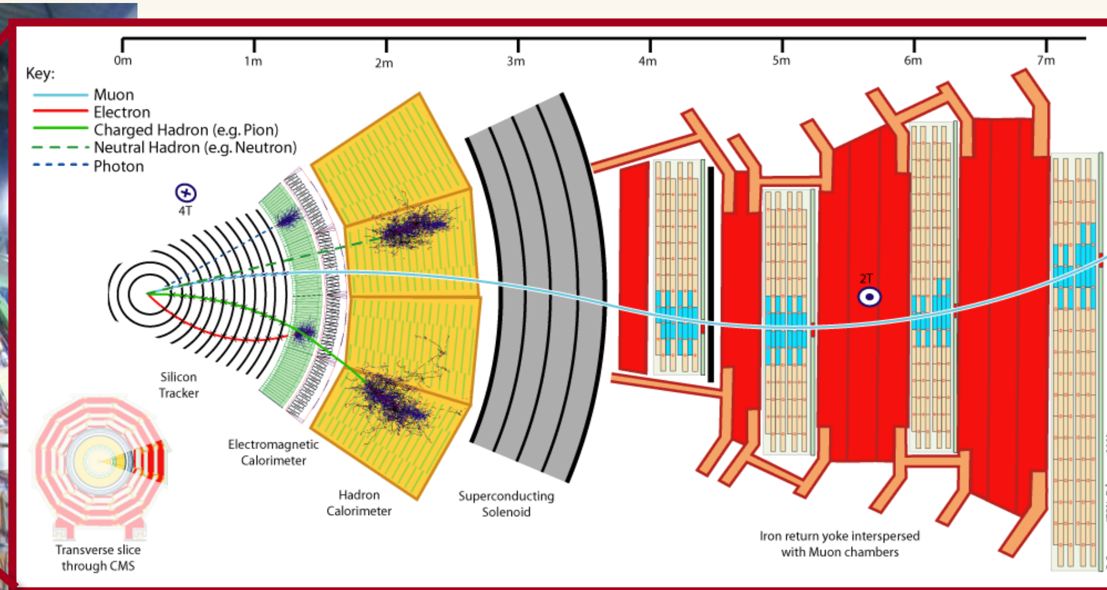


"Paper became so cheap, and printers so numerous, that a deluge of authors covered the land"

Alexander Pope, 1728

Large Hadron Collider at CERN





Compact Muon Solenoid detector

320 terabits per second raw data

Stop-gap: perform ad-hoc triage to 800 Gbps, recording only “interesting events”

Data Deluge Challenges

~~How can we get our hands on as much data as possible?~~

~~How can we extract as much information as possible from a limited amount of data?~~

How can we avoid having to acquire so much data to begin with?

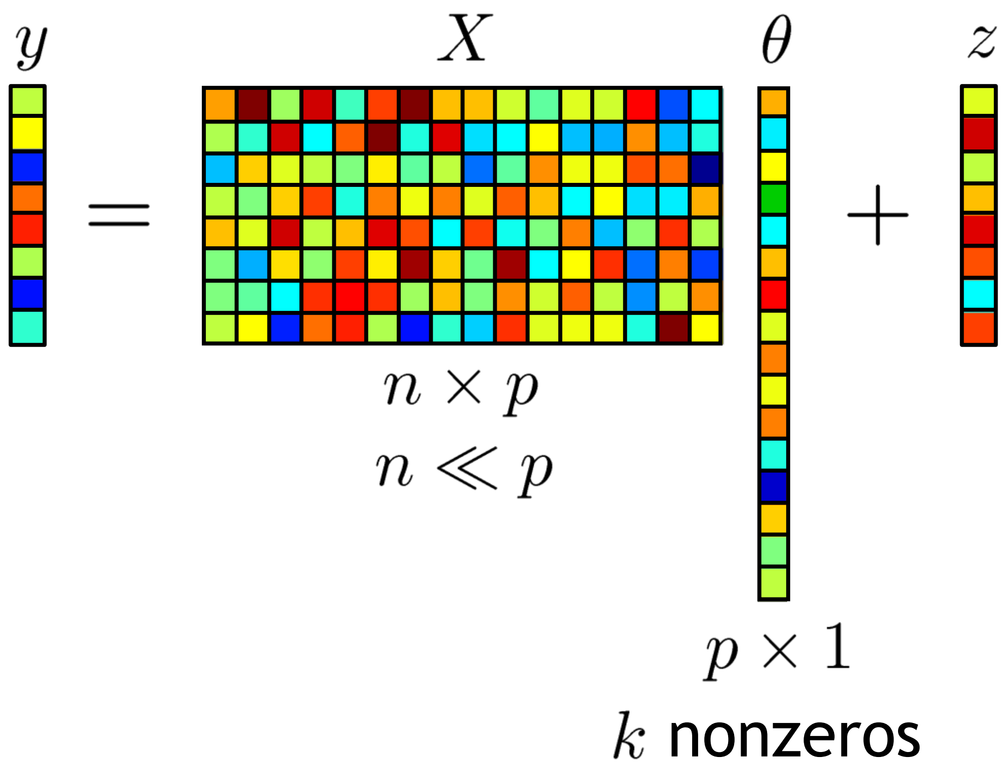
How can we extract any information at all from a massive amount of high-dimensional data?

Low-Dimensional Structure

How can we exploit low-dimensional structure to address the challenges posed by the "data deluge"?

- Visualization
- Feature extraction/selection
- Compression
- Regularization of ill-posed inverse problems
- Underpins *compressive sensing*

Compressive Sensing



When (and how well) can we estimate θ from the measurements y?

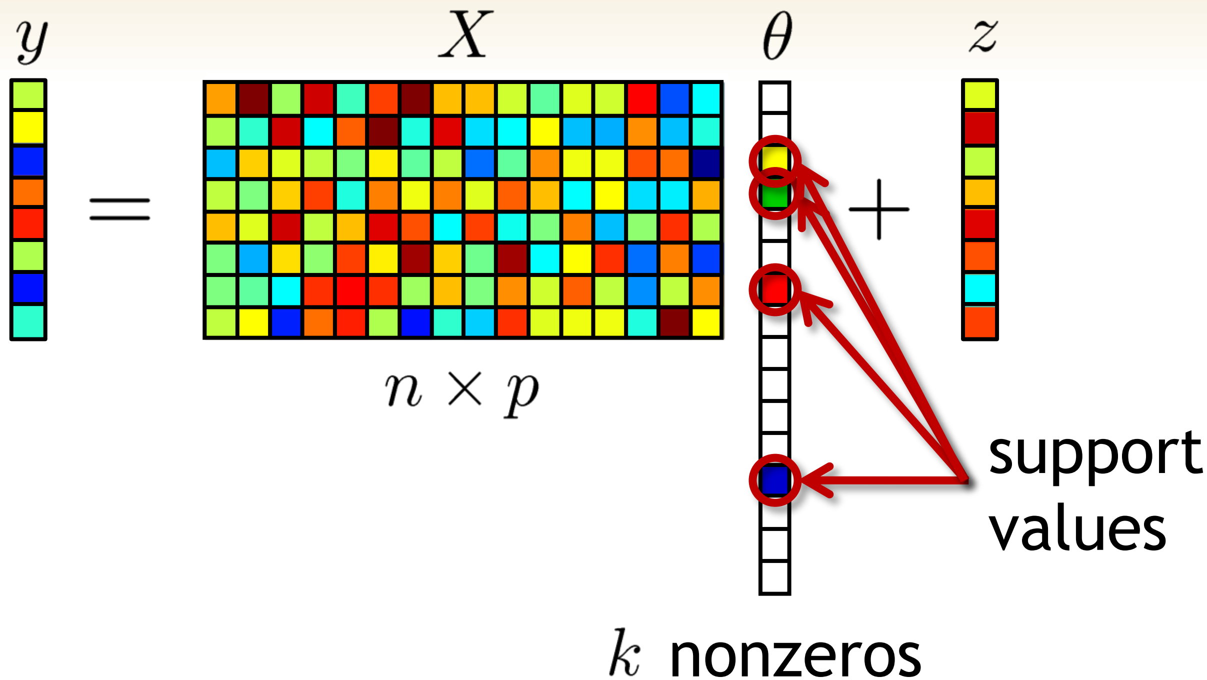
How Well Can We Estimate θ?

- What do we know via compressive sensing?
  - feasible *nonadaptive* schemes with known performance guarantees

- Can we improve upon compressive sensing?
  - lower bound on the performance of *any* nonadaptive scheme

- What are the benefits of adaptivity?
  - lower bound on the performance of *any adaptive* scheme
  - practical implications

Compressive Sensing



- How should we design X to ensure that y contains as much information about θ as possible?

- What algorithms do we have for recovering θ from y?

[Candès, Romberg, and Tao; Donoho - 2005]

How To Design X ?

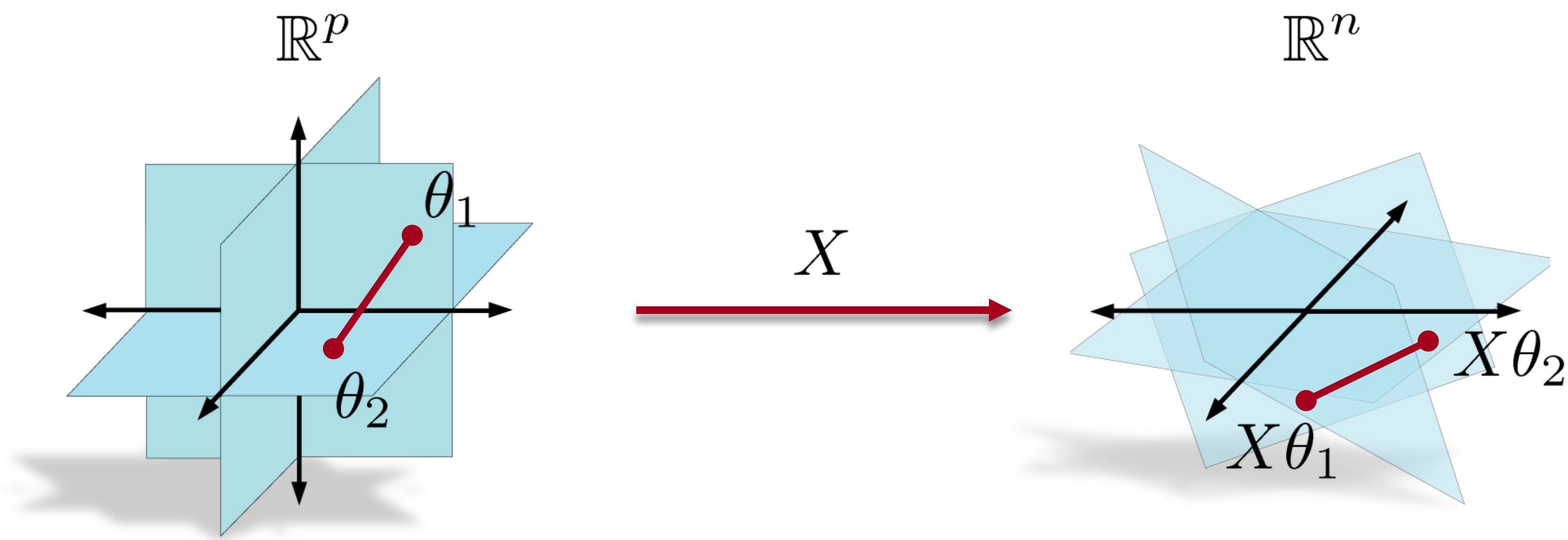
Prototypical sensing model:

$$y = X\theta + z \qquad\qquad z \sim \mathcal{N}(0, \sigma^2 I)$$

- Constrain X to have unit-norm rows
- Pick X at *random!*
  - i.i.d. Gaussian entries (with variance $1/p$)
  - random rows from a unitary matrix
- As long as $n = O(k \log(p/k))$, with high probability a random X will satisfy the *restricted isometry property*

Restricted Isometry Property (RIP)

$$\frac{\|X\theta_1 - X\theta_2\|_2^2}{\|\theta_1 - \theta_2\|_2^2} \approx \frac{n}{p} \qquad \|\theta_1\|_0, \|\theta_2\|_0 \leq k$$



How To Design X ?

Prototypical sensing model:

$$y = X\theta + z \qquad\qquad z \sim \mathcal{N}(0, \sigma^2 I)$$

- Constrain X to have unit-norm rows
- Pick X at *random!*
  - i.i.d. Gaussian entries (with variance $1/p$)
  - random rows from a unitary matrix
- As long as $n = O(k \log(p/k))$, with high probability a random X will satisfy the *restricted isometry property*
- Deep connections with *Johnson-Lindenstrauss Lemma*
  - see Baraniuk, Davenport, DeVore, and Wakin (2008)

How To Recover θ?

- Lots and lots of algorithms
  - ℓ_1-minimization (Lasso, Dantzig selector)
  - greedy algorithms (matching pursuit, forward selection)

If X satisfies the RIP, $\|\theta\|_0 \leq k$, and
$y = X\theta + z$ with $z \sim \mathcal{N}(0, \sigma^2 I)$, then

$$\widehat{\theta} = \arg\min_{\theta' \in \mathbb{R}^p} \|\theta'\|_1$$

$$\text{s.t.} \quad \|X^*(y - X\theta')\|_\infty \leq c\sqrt{\log p}\,\sigma$$

satisfies

$$\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \leq C\frac{p}{n}k\sigma^2 \log p.$$

[Candès and Tao - 2005]

How Well Can We Estimate θ?

- What do we know via compressive sensing?

  For any θ we can achieve $\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \leq C\dfrac{p}{n}k\sigma^2 \log p$

- Can we improve upon compressive sensing?

- What are the benefits of adaptivity?

Room For Improvement?

Let x_i denote the i^{th} row of X

$$y_i = \langle x_i, \theta \rangle + z_i$$

x_i and θ are almost orthogonal

- We are using most of our "sensing power" to sense entries that aren't even there!
- Tremendous loss in signal-to-noise ratio (SNR)
- It's hard to imagine any way to avoid this…

Minimax Lower Bounds

- There exists matrices X such that for *any* (sparse) θ we have
$$\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \leq C\frac{p}{n}k\sigma^2 \log p.$$

- We would like to know if there exists *any* X or *any* recovery algorithm that can do much better for *all* θ

- **Minimax lower bound:** For *any* X and *any* $\widehat{\theta}$, there exists a θ such that
$$\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \quad \geq \quad ?$$

- The bound will be determined by the *worst-case* θ

Can We Do Better?

Theorem
For *any* matrix X (with unit-norm rows) and *any* recovery procedure $\widehat{\theta}$, there exists a θ with $\|\theta\|_0 \leq k$ such that if $y = X\theta + z$ with $z \sim \mathcal{N}(0, \sigma^2 I)$, then

$$\mathbb{E}\, \|\widehat{\theta}(y) - \theta\|_2^2 \geq C' \frac{p}{n} k\sigma^2 \log(p/k).$$

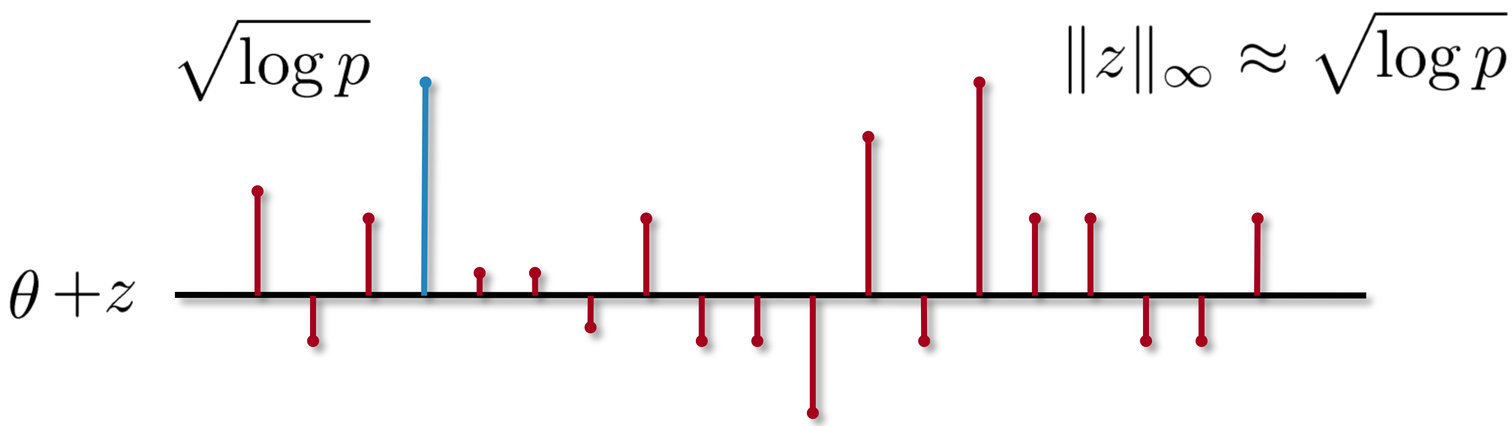
Compressive sensing is already operating at the limit

[Candès and Davenport - 2011]

Intuition

Suppose that $y = \theta + z$ with $z \sim \mathcal{N}(0, I)$ and that $k = 1$

$$\mathbb{E}\,\|\widehat{\theta}(y) - \theta\|_2^2 \geq C' \log p$$



Proof Recipe

- Construct a set Θ of k-sparse vectors such that
  - $|\Theta| = (p/k)^{k/4}$
  - $\|\theta_i - \theta_j\|_2 \geq \frac{1}{2}$ for all $\theta_i, \theta_j \in \Theta$
  - $\frac{1}{|\Theta|} \sum_i \theta_i \theta_i^* \approx \frac{1}{p} I$
- Scale this set to the worst-case amplitude and use *Fano's Inequality* to show that if θ is selected uniformly at random from Θ, then the Bayes risk is large
- Θ can be constructed simply by picking k-sparse vectors at random

How Well Can We Estimate θ ?

- What do we know via compressive sensing?

For any θ we can achieve $\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \leq C\frac{p}{n}k\sigma^2\log p$

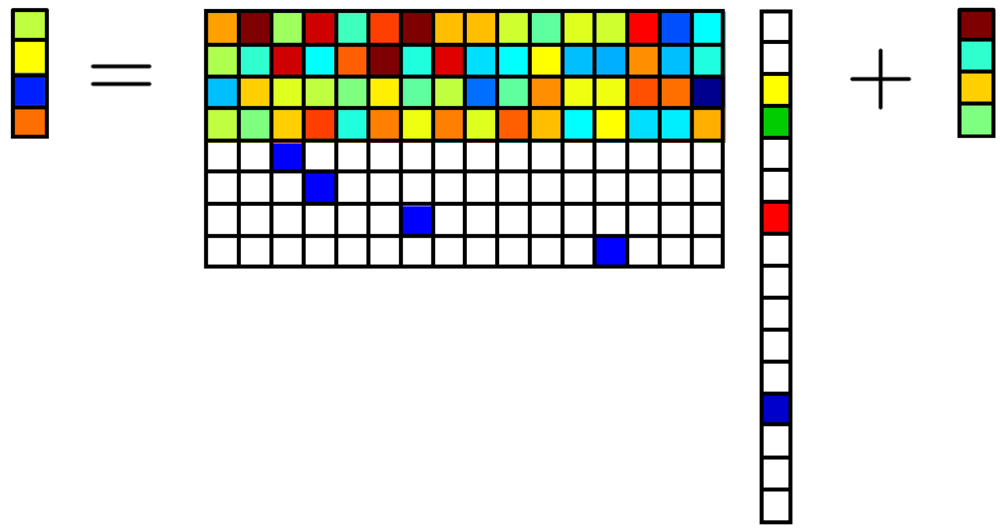
- Can we improve upon compressive sensing?

There exist θ such that $\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \geq C'\frac{p}{n}k\sigma^2\log(p/k)$

- What are the benefits of adaptivity?

Adaptivity to the Rescue?

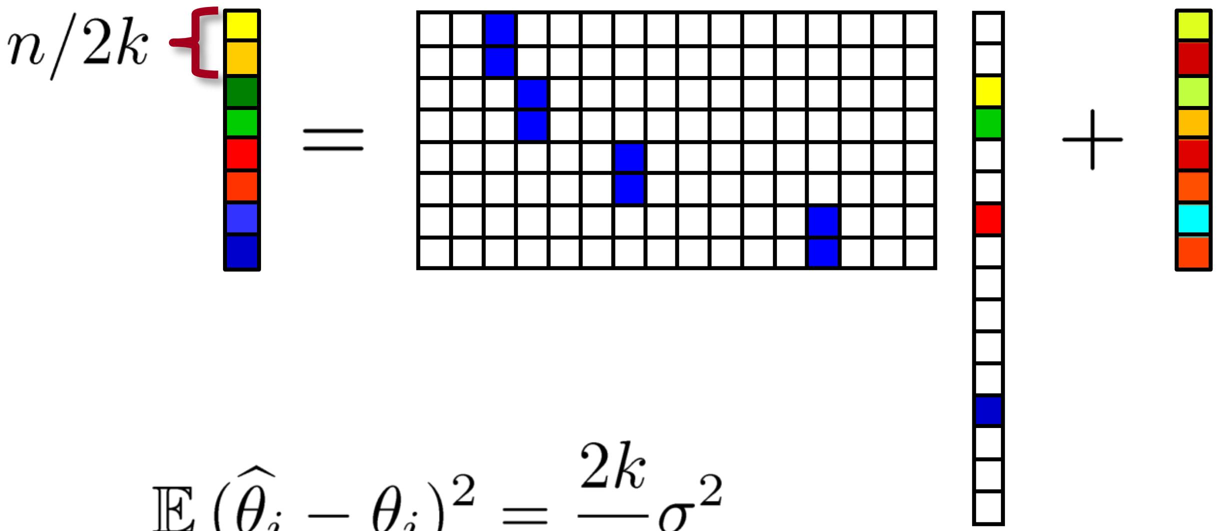
Think of sensing as a game of 20 questions



Simple strategy: Use $n/2$ measurements to find the support, and the remainder to estimate the values.

Thought Experiment

Suppose that after $n/2$ measurements we have perfectly estimated the support.



$$\mathbb{E}\left(\widehat{\theta}_i - \theta_i\right)^2 = \frac{2k}{n}\sigma^2$$

$$\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 = \frac{2k}{n}k\sigma^2 \ll \frac{p}{n}k\sigma^2 \log p$$

Does Adaptivity *Really* Help?

Sometimes…

- Noise-free measurements, but non-sparse signal
  - adaptivity doesn't help if you want a uniform guarantee
  - probabilistic adaptive algorithms can reduce the required number of measurements from $O(k \log(p/k))$ to $O(k \log \log(p/k))$ [Indyk et al. – 2011]

- Noisy setting
  - distilled sensing [Haupt et al. – 2007, 2010]
  - adaptivity can reduce the estimation error to

$$\mathbb{E}\, \|\widehat{\theta} - \theta\|_2^2 = \frac{p}{n} k\sigma^2$$

Which is it?

$$\mathbb{E}\, \|\widehat{\theta} - \theta\|_2^2 = \frac{k}{n} k\sigma^2$$

Which Is It?

Suppose we have a budget of n measurements of the form
$y_i = \langle x_i, \theta \rangle + z_i$ where $\|x_i\|_2 = 1$ and $z_i \sim \mathcal{N}(0, \sigma^2)$

The vector x_i can have an arbitrary dependence on the measurement history, i.e., $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})$

Theorem
There exist θ with $\|\theta\|_0 \leq k$ such that for *any* adaptive measurement strategy and *any* recovery procedure $\widehat{\theta}$,

$$\mathbb{E}\, \|\widehat{\theta}(y) - \theta\|_2^2 \geq C \frac{p}{n} k \sigma^2.$$

Thus, in general, adaptivity does *not* significantly help!

[Arias-Castro, Candès, and Davenport - 2011]

Proof Strategy

Step 1: Consider sparse signals with nonzeros of amplitude $\mu \approx \sigma\sqrt{p/n}$

Step 2: Show that if given a budget of n measurements, you cannot detect the support very well

Step 3: Immediately translate this into a lower bound on the MSE

To make things simpler, we will consider a Bernoulli prior $\pi(\theta)$ instead of a uniform k-sparse prior:

$$\theta_j = \begin{cases} 0 & \text{with probability } 1 - k/p \\ \mu > 0 & \text{with probability } k/p \end{cases}$$

Proof of Main Result

Let $S = \{j : \theta_j \neq 0\}$ and set $\sigma^2 = 1$

For any estimator $\widehat{\theta}$, define $\widehat{S} := \{j : |\widehat{\theta}_j| \geq \mu/2\}$

Whenever $j \in S \setminus \widehat{S}$ or $j \in \widehat{S} \setminus S$, $|\widehat{\theta}_j - \theta_j| \geq \frac{\mu^2}{4}$

$$\|\widehat{\theta} - \theta\|_2^2 \geq \frac{\mu^2}{4}|S \setminus \widehat{S}| + \frac{\mu^2}{4}|\widehat{S} \setminus S| = \frac{\mu^2}{4}|\widehat{S} \Delta S|$$

$$\Rightarrow \quad \mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \geq \frac{\mu^2}{4}\mathbb{E}\,|\widehat{S} \Delta S|$$

Proof of Main Result

Lemma
Under the Bernoulli prior, *any* estimate $\widehat{S}$ satisfies

$$\mathbb{E}\,|\widehat{S}\Delta S| \geq k\left(1 - \frac{\mu}{2}\sqrt{\frac{n}{p}}\right).$$

Thus, $\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \geq \frac{\mu^2}{4}\mathbb{E}\,|\widehat{S}\Delta S|$

$$\geq k \cdot \frac{\mu^2}{4}\left(1 - \frac{\mu}{2}\sqrt{\frac{n}{p}}\right)$$

Plug in $\mu = \frac{8}{3}\sqrt{\frac{p}{n}}$ and this reduces to

$$\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \geq \frac{4}{27} \cdot \frac{kp}{n} \geq \frac{1}{7} \cdot \frac{kp}{n}$$

Key Ideas in Proof of Lemma

$$\mathbb{P}_{0,j}(y_1,\ldots,y_n) = \mathbb{P}(y_1,\ldots y_n|\theta_j = 0)$$
$$\mathbb{P}_{1,j}(y_1,\ldots,y_n) = \mathbb{P}(y_1,\ldots,y_n|\theta_j = \mu)$$

$$\begin{aligned}\mathbb{E}\,|\widehat{S}\Delta S| &\geq \frac{k}{p}\sum_j (1 - \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}) \\ &\geq k - \frac{k}{\sqrt{p}}\sqrt{\sum_j \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2}\end{aligned}$$

$$\sum_j \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\mu^2}{4}n \quad \Rightarrow \quad \mathbb{E}\,|\widehat{S}\Delta S| \geq k\left(1 - \frac{\mu}{2}\sqrt{\frac{n}{p}}\right)$$

Key Ideas in Proof of Lemma

Pinsker's Inequality

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{K(\mathbb{P}, \mathbb{Q})/2}$$

$$\begin{aligned}\|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 &\leq \frac{\pi_0}{2} K(\mathbb{P}_{0,j}, \mathbb{P}_{1,j}) + \frac{\pi_1}{2} K(\mathbb{P}_{1,j}, \mathbb{P}_{0,j}) \\ &\leq \frac{\mu^2}{4} \sum_i \mathbb{E}\, x_{i,j}^2\end{aligned}$$

$$\Rightarrow \quad \sum_j \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\mu^2}{4} \sum_{i,j} \mathbb{E}\, x_{i,j}^2 = \frac{\mu^2}{4} n$$

How Well Can We Estimate θ?

- What do we know via compressive sensing?

For any θ we can achieve $\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \leq C\frac{p}{n}k\sigma^2 \log p$

- Can we improve upon compressive sensing?

There exist θ such that $\mathbb{E}\,\|\widehat{\theta} - \theta\|_2^2 \geq C'\frac{p}{n}k\sigma^2 \log(p/k)$

- What are the benefits of adaptivity?

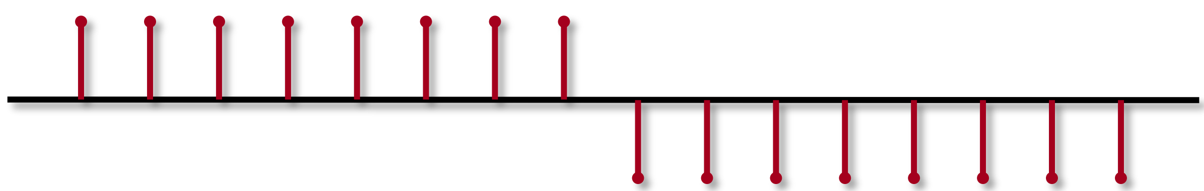
Minimal?

Adaptivity In Practice

Suppose that $k = 1$ and that $\theta_{j^*} = \mu$

Recursive Bisection [Iwen and Tewfik – 2011]

- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the "active set"
- after subdividing $\log p$ times, return support

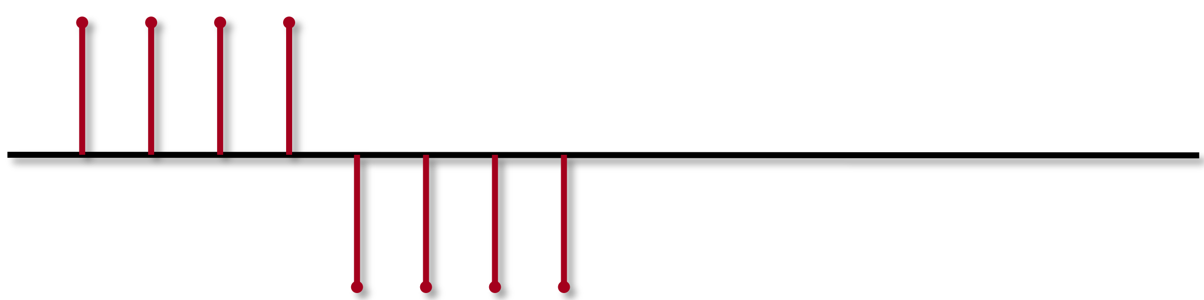


Adaptivity In Practice

Suppose that $k = 1$ and that $\theta_{j^*} = \mu$

Recursive Bisection [Iwen and Tewfik – 2011]

- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the “active set”
- after subdividing $\log p$ times, return support

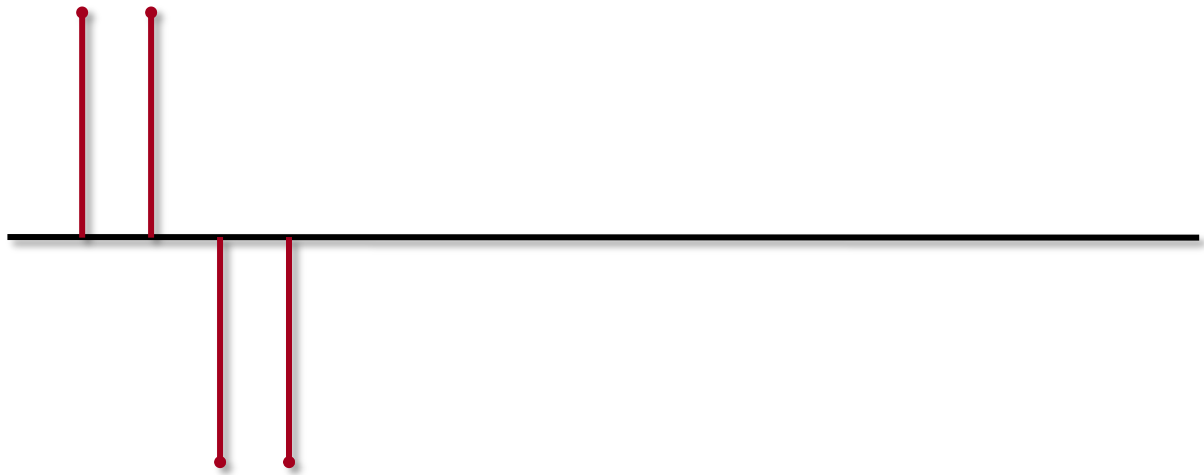


Adaptivity In Practice

Suppose that $k = 1$ and that $\theta_{j^*} = \mu$

Recursive Bisection [Iwen and Tewfik – 2011]

- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the "active set"
- after subdividing $\log p$ times, return support

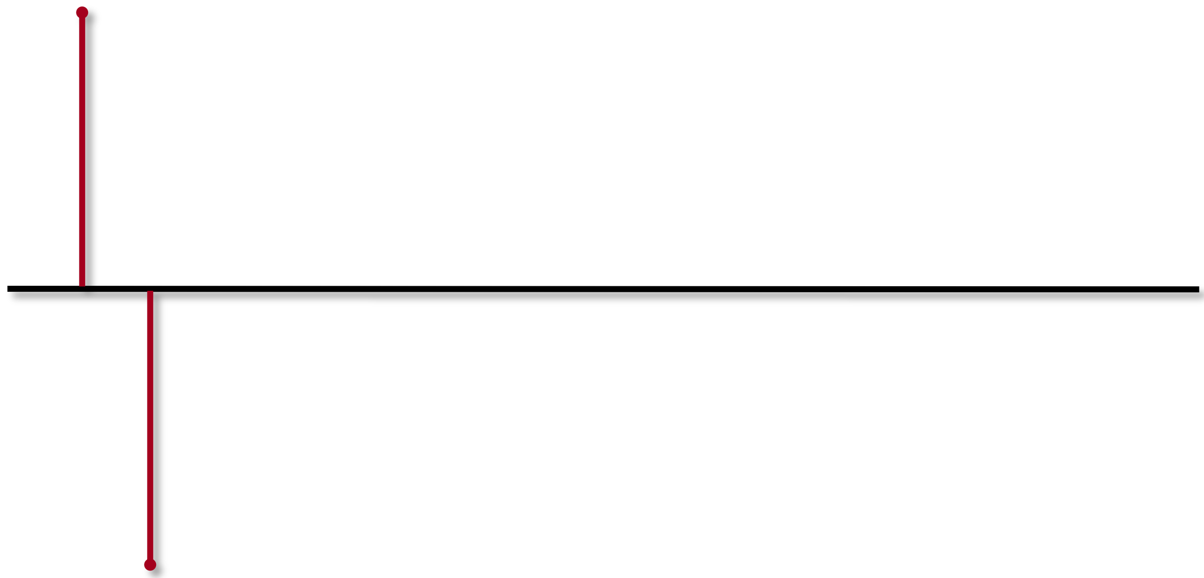


Adaptivity In Practice

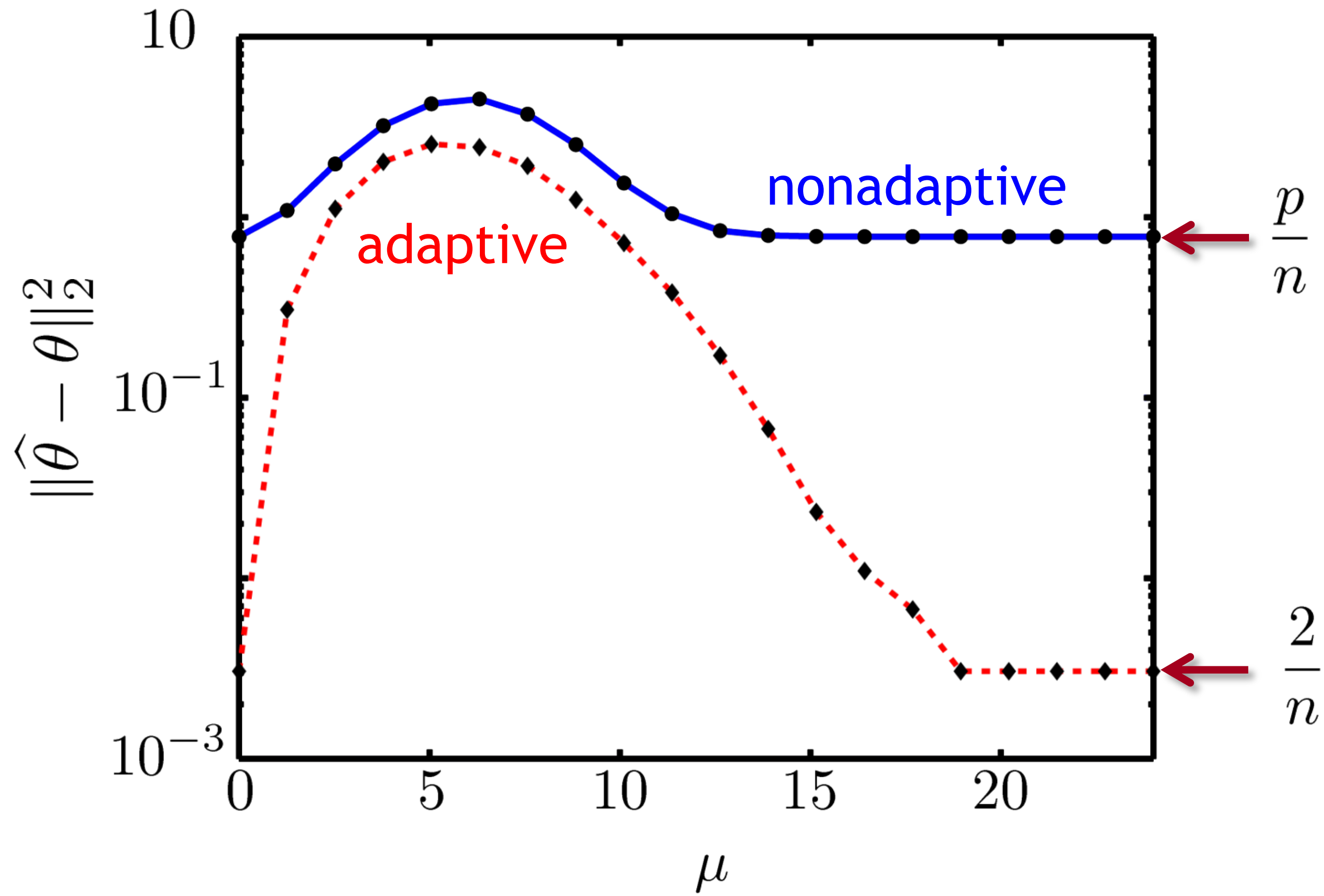
Suppose that $k = 1$ and that $\theta_{j^*} = \mu$

Recursive Bisection [Iwen and Tewfik – 2011]

- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the "active set"
- after subdividing $\log p$ times, return support



Experimental Results



$p = 512$

$n = 256$

$\sigma^2 = 1$

[Arias-Castro, Candès, and Davenport - 2011]

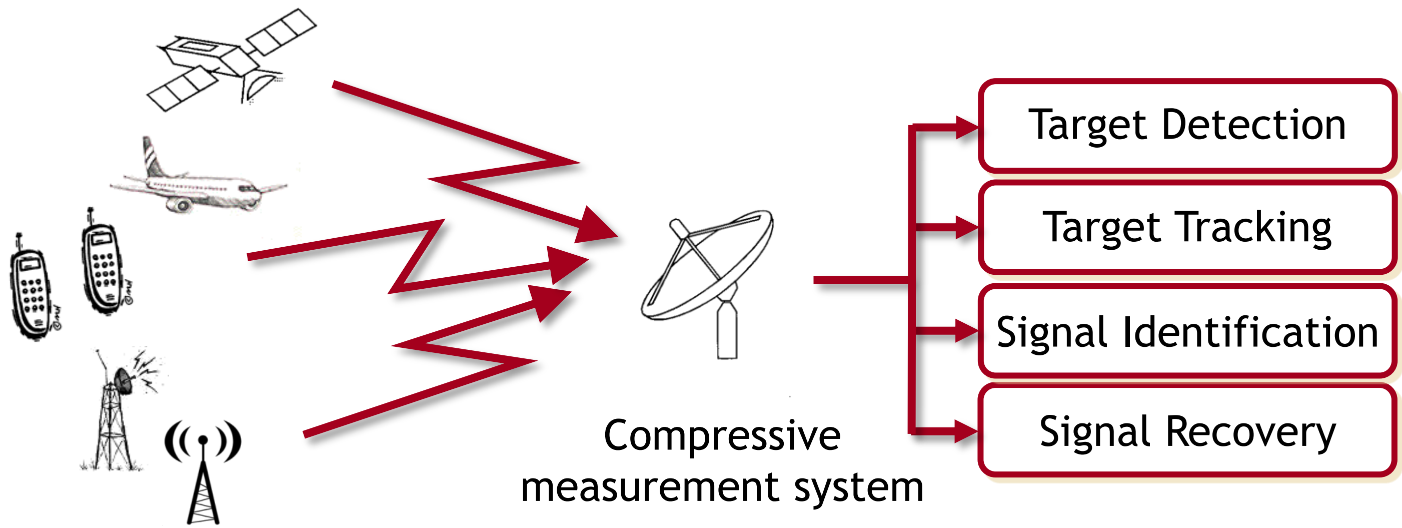
Looking Forward

Adaptivity in Practice

- Sharp bounds to differentiate the regions where adaptivity helps and where it doesn't
- Practical algorithms that work well for all values of μ
- New theory for restricted adaptive measurements
  - single-pixel camera: 0/1 measurements
  - magnetic resonance imaging (MRI): Fourier measurements
  - analog-to-digital converters: linear filter measurements
- New sensors and architectures that can actually acquire adaptive measurements

Beyond Recovery

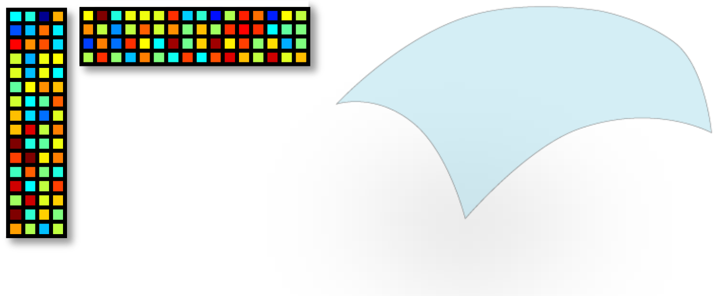
When and how can we directly solve inference problems directly from measurements?



- "Compressive signal processing"
- Links with machine learning
  - Johnson-Lindenstrauss lemma and geometry preservation
  - quantized compressive sensing and logistic regression

Beyond Sparsity

- Learned dictionaries, structured sparsity, models for continuous-time signals
- Multi-signal models
  - e.g., sensor networks/arrays, multi-modal data, …
- Low-rank matrix models
- Manifold/parametric models



Acquisition

- how to design X
- practical devices
- adaptivity

Recovery

- practical algorithms
- robust
- stable

Inference

- classification
- estimation
- learning

Acknowledgements

- Richard Baraniuk (Rice)
- Emmanuel Candès (Stanford)
- Ery Arias-Castro (UC San Diego)
- Ronald DeVore (Texas A&M)
- Marco Duarte (UMass Amherst)
- Kevin Kelly (Rice)
- Michael Wakin (Colorado School of Mines)

More Information

http://stat.stanford.edu/~markad

markad@stanford.edu