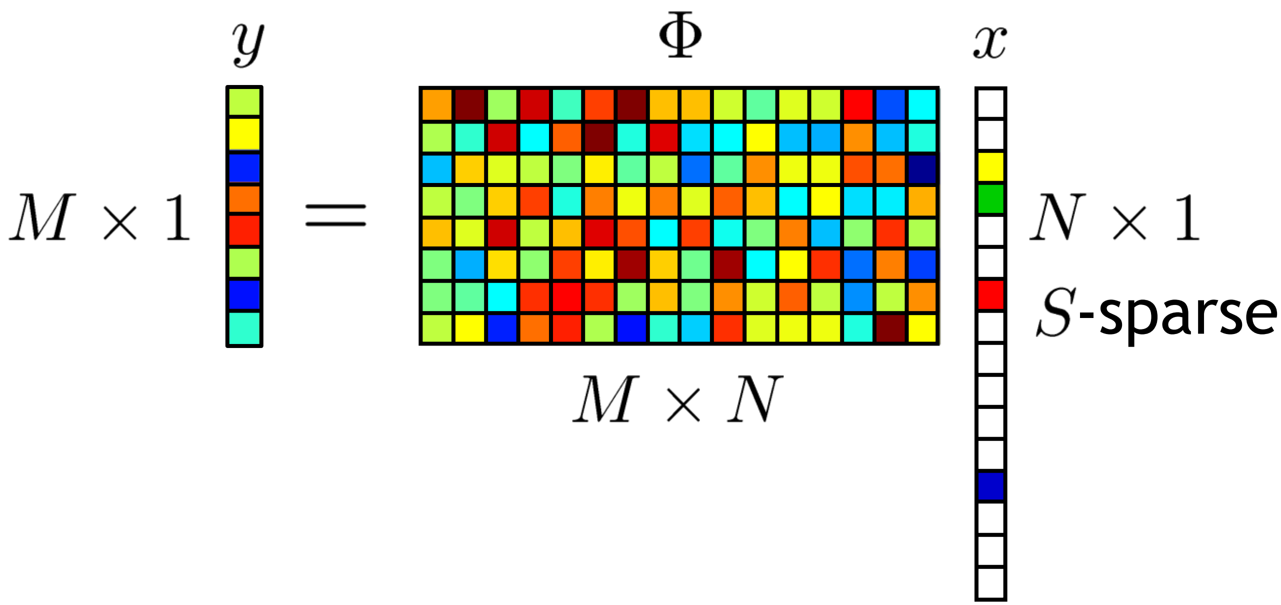
How well can we estimate a sparse vector?

Mark A. Davenport

Stanford University
Department of Statistics

Sparse Estimation



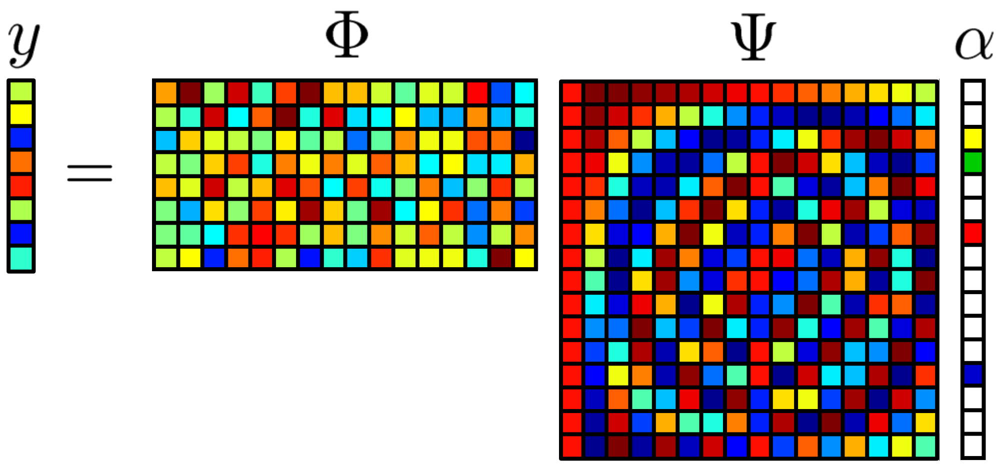
How well can we estimate x ?

Applications

- Statistics
  - model selection / variable selection in high-dimensional regression
- Inverse problems
- Compressive sensing (CS)
  - matrix Φ represents a sensing system
  - typically underdetermined
  - sparsity acts as a regularizer

Core Challenges in CS

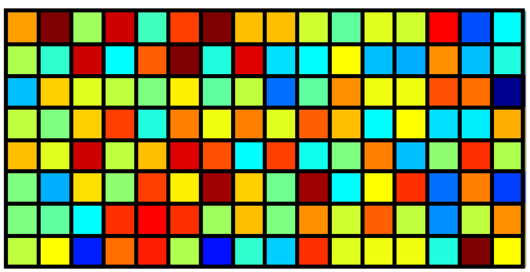
- How should we design the matrix Φ so that M is as small as possible?



- How can we recover x from the measurements y ?

Answers

- Choose a ***random matrix***
  - fill out the entries of Φ with i.i.d. samples from a sub-Gaussian distribution
  - project onto a "random subspace"



$$M = O(S \log(N/S)) \ll N$$

- Use any sparse approximation algorithm

Is this the best we can do?

Recovery from Noisy Measurements

Given $y = \Phi x + e$ or $y = \Phi(x + n)$,
find x

- Optimization-based methods
  - basis pursuit, basis pursuit de-noising, Dantzig selector

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x\|_1$$

$$\text{s.t.} \quad \|y - \Phi x\|_2 \leq \epsilon$$

- Greedy/Iterative algorithms
  - OMP, StOMP, ROMP, CoSaMP, Thresh, SP, IHT, …

Stable Signal Recovery

Suppose that we observe $y = \Phi x + e$ and that Φ satisfies the RIP of order $2S$.

$$\boxed{(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2 \qquad \|x\|_0 \leq 2S}$$

Typical (worst-case) guarantee

$$\boxed{\|\widehat{x} - x\|_2^2 \leq C\|e\|_2^2}$$

Even if $\Lambda = \mathrm{supp}(x)$ is provided by an oracle, the error can still be as large as $\|\widehat{x} - x\|_2^2 = \|e\|_2^2/(1-\delta)$.

Stable Signal Recovery: Part II

Suppose now that Φ satisfies

$$\boxed{A(1-\delta)\|x\|_2^2 \le \|\Phi x\|_2^2 \le A(1+\delta)\|x\|_2^2 \qquad \|x\|_0 \le 2S}$$

In this case our guarantee becomes

$$\boxed{\|\widehat{x} - x\|_2^2 \le \frac{C}{A}\|e\|_2^2}$$

Unit-norm rows $\Rightarrow$ $\|\widehat{x} - x\|_2^2 \le C\frac{N}{M}\|e\|_2^2$

Expected Performance

- Worst-case bounds can be pessimistic
- What about the *average* error?
  - assume e is white noise with variance σ^2

$$\mathbb{E}\left(\|e\|_2^2\right) = M\sigma^2$$

  - for (nonadaptive) oracle

$$\mathbb{E}\left(\|\widehat{x} - x\|_2^2\right) \leq \frac{S\sigma^2}{A(1-\delta)}$$

  - if e is Gaussian, then for ℓ_1 -minimization

$$\mathbb{E}\left(\|\widehat{x} - x\|_2^2\right) \leq \frac{C'}{A} S\sigma^2 \log N$$

Can We Do Better?

- Better choice of Φ ?
- Better recovery algorithm?

Assume we have a budget for $\|\Phi\|_F^2$.

If we knew the support of x *a priori*, then by adapting Φ to exploit this knowledge we could achieve

$$\mathbb{E}\left[\|\widehat{x} - x\|_2^2\right] \approx \frac{S}{\|\Phi\|_F^2} S\sigma^2 \ll C' \frac{N}{\|\Phi\|_F^2} S\sigma^2 \log N$$

Is there any way to match this performance without knowing the support of x in advance?

$$R^*_{\mathrm{mm}}(\Phi) = \inf_{\widehat{x}} \sup_{\|x\|_0 \le S} \mathbb{E}\left[\|\widehat{x}(\Phi x + e) - x\|_2^2\right]$$

No!

Theorem:

If $y = \Phi x + e$ with $e \sim \mathcal{N}(0, \sigma^2 I)$, then

$$R^*_{\mathrm{mm}}(\Phi) \geq C \frac{N}{\|\Phi\|_F^2} S\sigma^2 \log(N/S).$$

If $y = \Phi(x + n)$ with $n \sim \mathcal{N}(0, \sigma^2 I)$, then

$$R^*_{\mathrm{mm}}(\Phi) \geq C \frac{N}{M} S\sigma^2 \log(N/S).$$

$$\Phi = U\Sigma V^* \quad y' = \Sigma^{-1}U^* y = V^* x + V^* n \quad \|V^*\|_F^2 = M$$

See also: Raskutti, Wainwright, and Yu (2009)
Ye and Zhang (2010)

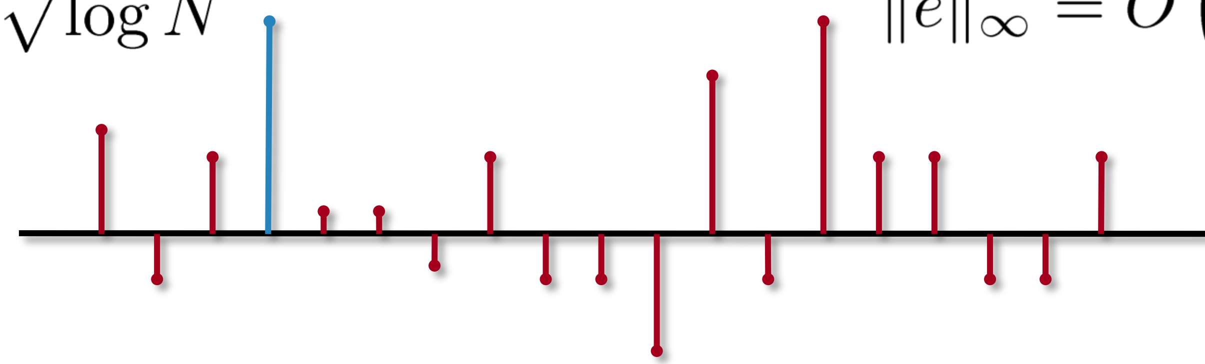
[Candès and Davenport - 2011]

Intuition

Suppose that $y = x + e$ with $e \sim \mathcal{N}(0, I)$ and that $S = 1$.

$$R^*_{\mathrm{mm}}(I) \geq C \log(N).$$

$$\mu = \sqrt{\log N} \qquad \|e\|_\infty = O\left(\sqrt{\log N}\right)$$



Proof Recipe

Ingredients [Makes $\sigma^2 = 1$ servings]

- Lemma 1: Suppose $\mathcal{X}$ is a set of S-sparse points such that $\|x_i - x_j\|_2^2 \geq 8NR^*_{\mathrm{mm}}(\Phi)$ for all $x_i, x_j \in \mathcal{X}$. Then $\frac{1}{2}\log|\mathcal{X}| - 1 \leq \frac{1}{2|\mathcal{X}|^2}\sum_{i,j}\|\Phi x_i - \Phi x_j\|_2^2$.
- Lemma 2: There exists a set $\mathcal{X}$ of S-sparse points such that
  - $|\mathcal{X}| = (N/S)^{S/4}$
  - $\|x_i - x_j\|_2 \geq \frac{1}{2}$ for all $x_i, x_j \in \mathcal{X}$
  - $\|\frac{1}{|\mathcal{X}|}\sum_i x_i x_i^* - \frac{1}{N}I\| \leq \frac{\beta}{N}$ for some $\beta > 0$

Instructions

Combine ingredients and add a dash of linear algebra.

Proof Outline

$$\mu = \frac{1}{|\mathcal{X}|}\sum_i x_i \qquad Q = \frac{1}{|\mathcal{X}|}\sum_i x_i x_i^*$$

$$\begin{aligned}
\frac{S}{4}\log(N/S) - 2 &\leq \frac{1}{|\mathcal{X}|^2}\sum_{i,j}\|\Phi x_i - \Phi x_j\|_2^2 \\
&= \mathrm{Tr}\left(\Phi^*\Phi\left(\frac{1}{|\mathcal{X}|^2}\sum_{i,j}(x_i - x_j)(x_i - x_j)^*\right)\right) \\
&= \mathrm{Tr}\left(\Phi^*\Phi\left(2(Q - \mu\mu^*)\right)\right) \\
&\leq 2\mathrm{Tr}\left(\Phi^*\Phi Q\right) \\
&\leq 2\mathrm{Tr}\left(\Phi^*\Phi\right)\|Q\| \\
&\leq 2\|\Phi\|_F^2 \cdot 16 R_{\mathrm{mm}}^*(\Phi)(1+\beta)
\end{aligned}$$

$$\Rightarrow \quad R_{\mathrm{mm}}^*(\Phi) \geq \frac{S\log(N/S)}{128(1+\beta)\|\Phi\|_F^2}$$

Recall: Lemma 2

Lemma 2: There exists a set $\mathcal{X}$ of S-sparse points such that

- $|\mathcal{X}| = (N/S)^{S/4}$
- $\|x_i - x_j\|_2 \geq \frac{1}{2}$ for all $x_i, x_j \in \mathcal{X}$
- $\|\frac{1}{|\mathcal{X}|}\sum_i x_i x_i^* - \frac{1}{N} I\| \leq \frac{\beta}{N}$ for some $\beta > 0$

Strategy

Construct $\mathcal{X}$ by sampling (with replacement) from

$$\mathcal{U} = \left\{ x \in \{0, \sqrt{1/S}, -\sqrt{1/S}\}^N : \|x\|_0 \leq S \right\}$$

Repeat for $|\mathcal{X}| = (N/S)^{S/4}$ iterations.

With probability > 0, the remaining properties are satisfied.

Key: *Matrix Bernstein Inequality* [Ahlswede and Winter, 2002]

Recap

Noise added to the *measurements*

$$\mathbb{E}\left[\|\widehat{x} - x\|_2^2\right] \leq C' \frac{N}{\|\Phi\|_F^2} S\sigma^2 \log N$$

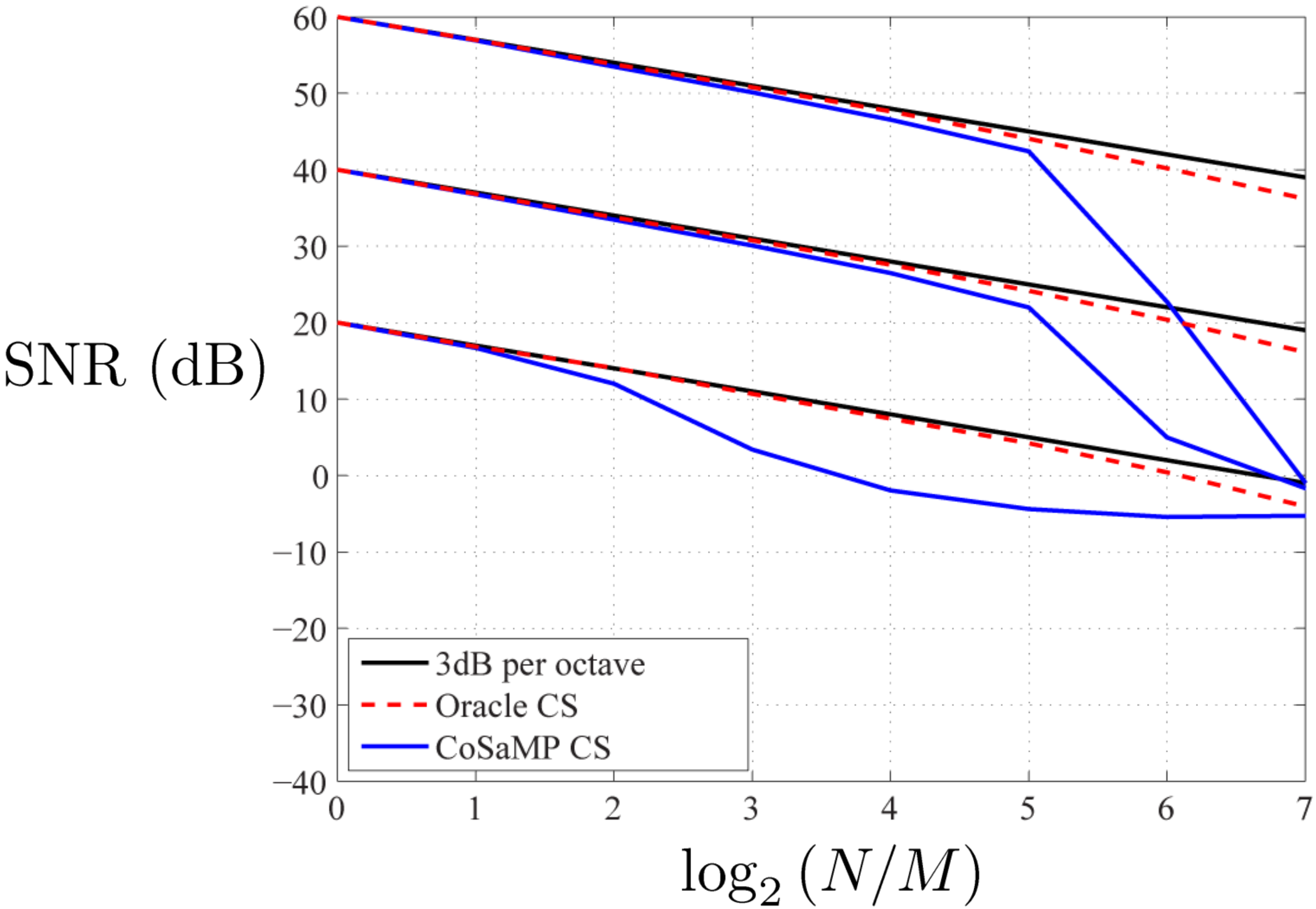
$$\mathbb{E}\left[\|\widehat{x} - x\|_2^2\right] \geq C \frac{N}{\|\Phi\|_F^2} S\sigma^2 \log(N/S)$$

Noise added to the *signal*

$$\mathbb{E}\left[\|\widehat{x} - x\|_2^2\right] \leq C' \frac{N}{M} S\sigma^2 \log N$$

$$\mathbb{E}\left[\|\widehat{x} - x\|_2^2\right] \geq C \frac{N}{M} S\sigma^2 \log(N/S)$$

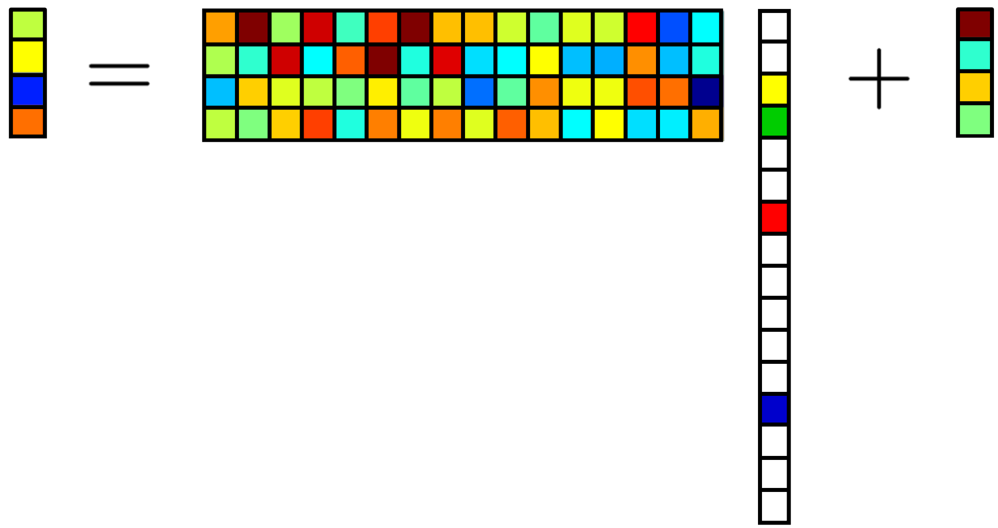
Noise Folding



[Davenport, Laska, Treichler, and Baraniuk - 2011]

Adaptivity to the Rescue?

What if we adapt the measurements to the particular signal?



If we are too greedy, our support estimate might be wrong…

Does adaptivity really help?

Sometimes…

- Information-based complexity: "Adaptivity doesn't help!"
  - assumes signal x lies in a set K satisfying certain conditions
  - noise-free measurements
  - adaptivity reduces minimax error over K by at most 2
- Nevertheless, adaptivity can still help [Indyk et al. - 2011]
  - reduced number of measurements in a probabilistic setting
  - still requires noise-free measurements
- What about noise?
  - distilled sensing (Haupt, Castro, Nowak, and others)
  - message seems to be that adaptivity really helps in noise

Adaptive Compressive Sensing

Suppose we have a budget of M measurements of the form

$$y_i = \langle \phi_i, x \rangle + e_i$$

where $\|\phi_i\|_2 = 1$ and $e_i \sim \mathcal{N}(0, \sigma^2)$.

The vector ϕ_i can have an arbitrary (but deterministic) dependence on the measurements $y_1, y_2, \ldots, y_{i-1}$.

Consider the minimax MSE

$$R^*_{\mathrm{mm}} = \inf_{\widehat{x}} \sup_{\|x\|_0 \le S} \mathbb{E}\left[\|\widehat{x}(\Phi x + e) - x\|_2^2\right]$$

Main Result

Possibilities include

- Adaptive oracle rate: $R^*_{\mathrm{mm}} \approx \frac{S}{M} S\sigma^2$
- Nonadaptive rate: $R^*_{\mathrm{mm}} \approx \frac{N}{M} S\sigma^2 \log(N/S)$
- Somewhere in-between?

$$\boxed{R^*_{\mathrm{mm}} \geq \widetilde{C} \frac{N}{M} S\sigma^2}$$

In general, adaptivity does ***not*** significantly help!!

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

Underlying Ideas

Step 1: Consider sparse signals with nonzeros of amplitude $\mu = \sqrt{N/M}$.

Step 2: Show that if you have fewer than M measurements, then with high probability you will fail to recover a significant fraction of the support.

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

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

Adaptivity in Practice

Suppose that $S = 1$ and that $x_{j^*} = \mu$.

Algorithm 1 [Castro et al. – 2008]

- start with random (Rademacher) measurements
- after each measurement, compute posterior distribution p
- re-weight subsequent measurements using p

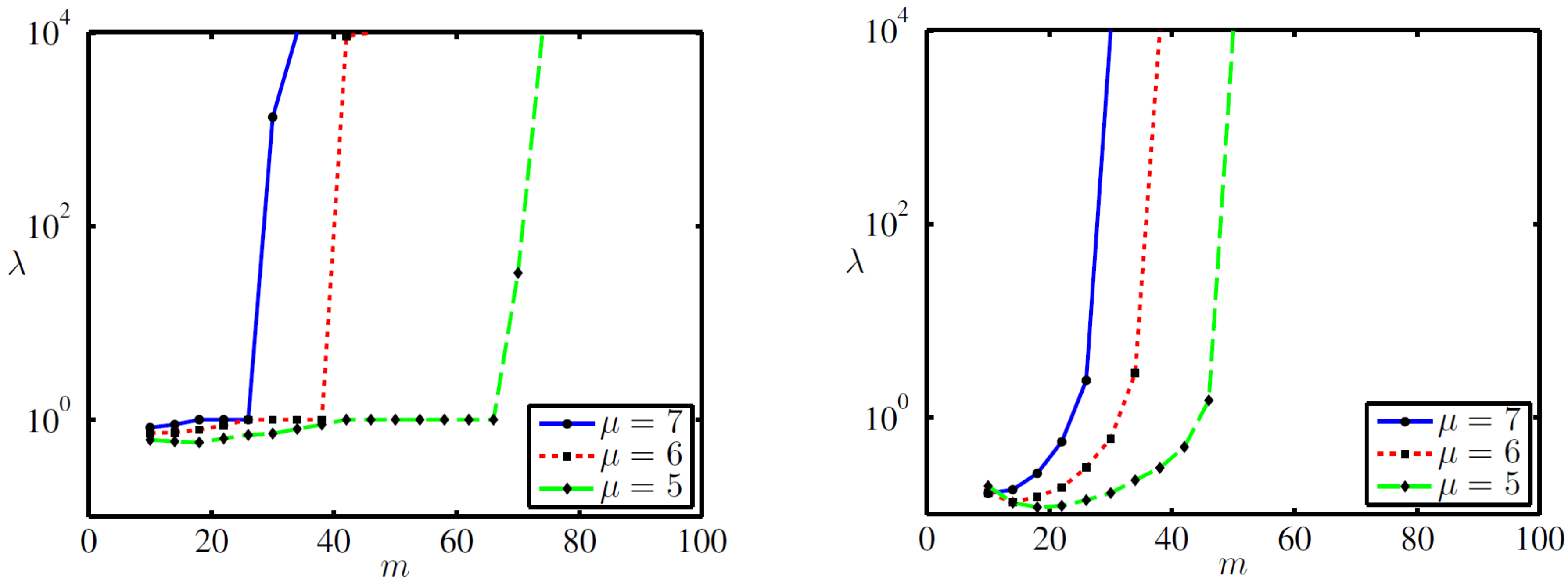
Algorithm 2 [Iwen and Tewfik – 2011]

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

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

Phase Transition in the Posterior

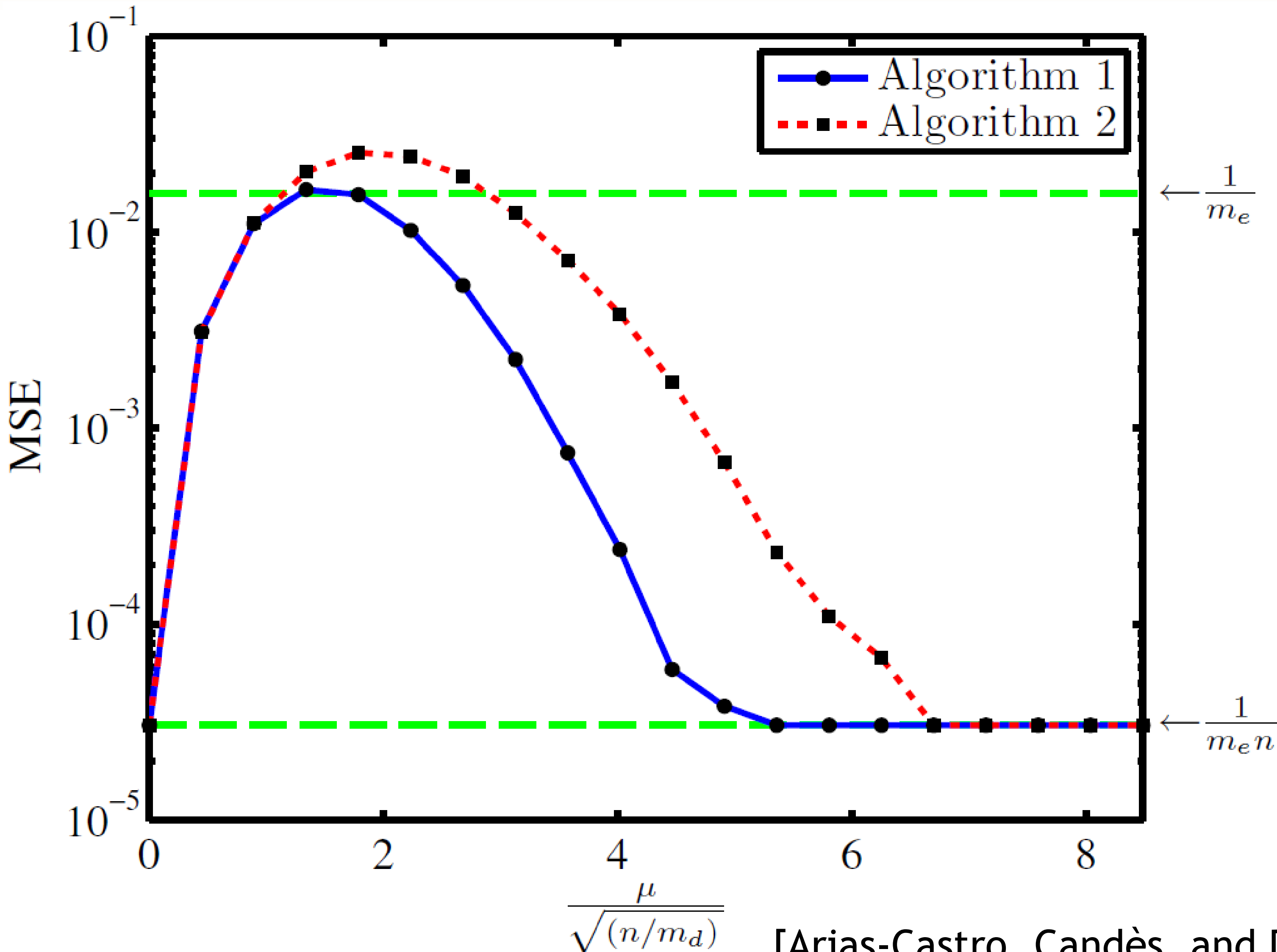
$$\lambda = \frac{p_{j^*}}{\max_{j \neq j^*} p_j}$$



$$n = 512 \qquad \sigma^2 = 1$$

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

Phase Transition in the MSE



$n = 512$

$\sigma^2 = 1$

$m_d = 128$

$m_e = 128$

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

Conclusions

- In some scenarios, CS can be sensitive to noise
  - inherent lower bound that applies to any possible sensing scheme
  - if you can average out noise, that will always help
  - sparsity is still helping a lot

- Surprisingly, adaptive algorithms cannot overcome this obstacle!

- Adaptivity might still be very useful in practice
  - practical adaptive algorithms that achieve the minimax rate for all values of μ ?