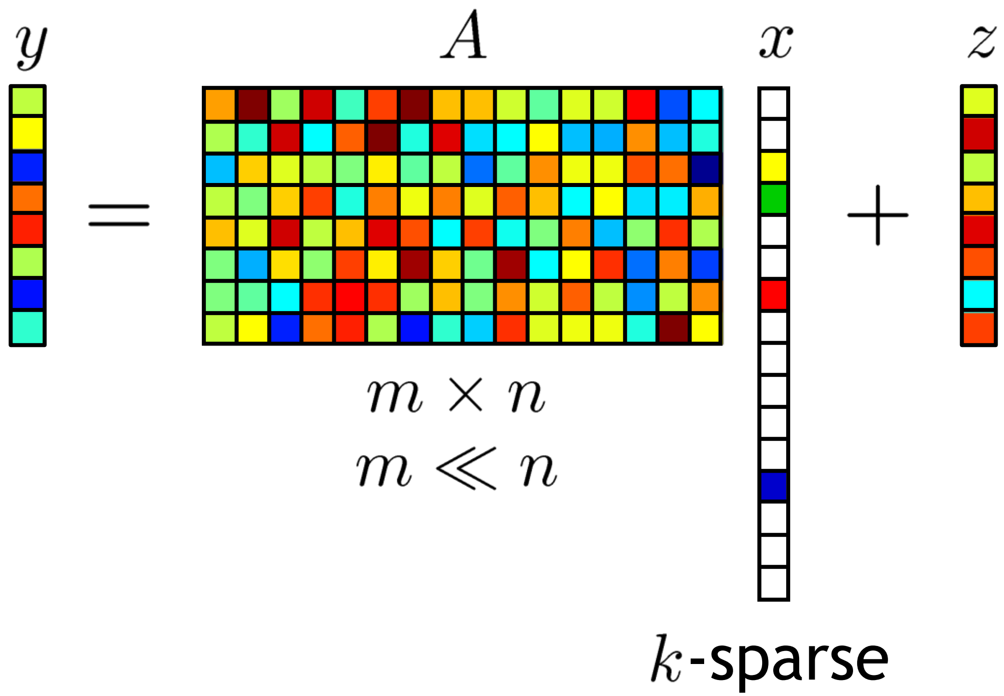
The limits of adaptive sensing

Mark A. Davenport

Stanford University
Department of Statistics

Sparse Estimation



How well can we estimate x ?

Background: Dantzig Selector

- Choose a ***random matrix***
  - fill out the entries of A with i.i.d. samples from a sub-Gaussian distribution with $\mathbb{E}\left(a_{ij}^2\right) = \frac{1}{n}$.
  - select m rows from a random unitary matrix.

- If $m = O(k \log(n/k)) \ll n$, then using ℓ_1-minimization (e.g., Dantzig selector, LASSO) we can achieve

$$\mathbb{E}\left\|\widehat{x} - x\right\|_2^2 \leq C\frac{n}{m}k\sigma^2 \log n$$

Is this the best we can do?

Can We Do Better?

Via a better choice of A ? Better recovery algorithm?

Assume that we have a "sensing power budget" that requires $\|a_i\|_2 = 1$ for $i = 1, \ldots, m,$ and that the rows a_i are selected in advance, i.e., *nonadaptively*.

Theorem

For *any* matrix A and recovery procedure $\widehat{x}$,
if $y = Ax + z$ with $z \sim \mathcal{N}(0, \sigma^2 I)$, then

$$\sup_{\|x\|_0 \le k} \mathbb{E}\, \|\widehat{x}(y) - x\|_2^2 \ge C' \frac{n}{m} k\sigma^2 \log(n/k).$$

See Raskutti, Wainwright, and Yu (2009), Ye and Zhang (2010), Candès and Davenport (2011)

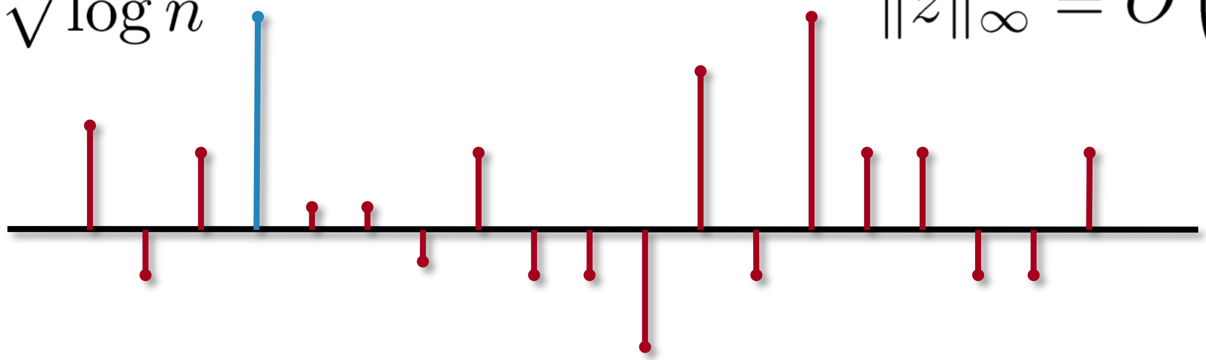
Intuition

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

$$\sup_{\|x\|_0 \le 1} \mathbb{E}\,\|\widehat{x}(y) - x\|_2^2 \ge C' \log n.$$

$\mu = \sqrt{\log n}$

$\|z\|_\infty = O\left(\sqrt{\log n}\right)$



Compressive Sensing and SNR

$$\sup_{\|x\|_0 \le k} \mathbb{E}\, \|\widehat{x}(y) - x\|_2^2 \ge C' \frac{n}{m} k \sigma^2 \log(n/k).$$

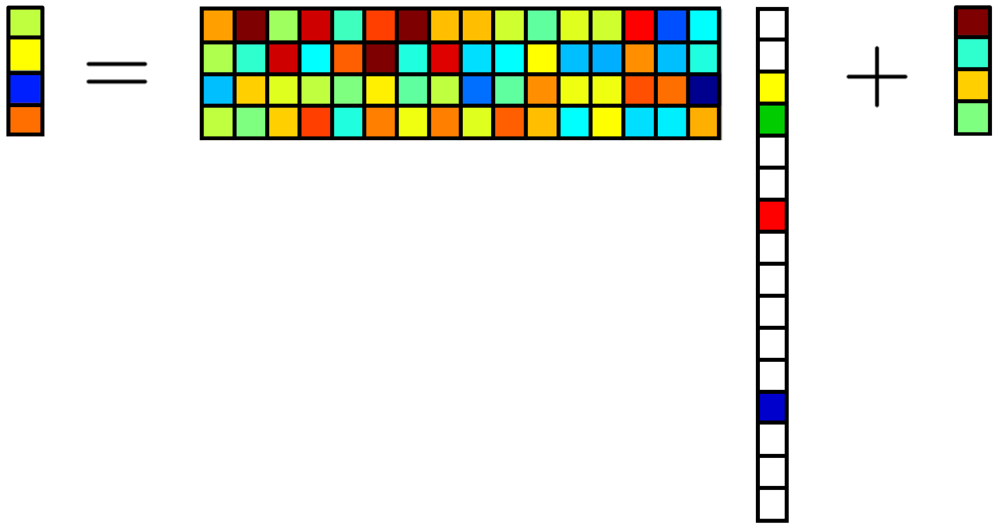
$$\langle a, x \rangle = \sum_{j=1}^{n} a_j x_j + z$$

dense sparse

- We are using most of our “sensing power” to sense entries that aren’t even there!
- Tremendous SNR loss
- Can potentially do much better if we can somehow concentrate our “sensing power” on the nonzeros

Adaptivity to the Rescue?

Think of sensing as a game of 20 questions



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

If support estimate is correct:

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

Does Adaptivity Really Help?

Sometimes…

- Information-based complexity: "Adaptivity doesn't help!"
  - assumes signal x lies in a set $\mathcal{S}$ satisfying certain conditions
  - noise-free measurements
  - adaptivity reduces minimax error over $\mathcal{S}$ 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

Main Result

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

The vector a_i can have an arbitrary dependence on the measurement history, i.e., $(a_1, y_1), \ldots, (a_{i-1}, y_{i-1})$.

Theorem
For *any* adaptive measurement strategy and *any* recovery procedure $\widehat{x}$,

$$\sup_{\|x\|_0 \leq k} \mathbb{E}\, \|\widehat{x}(y) - x\|_2^2 \geq C \frac{n}{m} k \sigma^2.$$

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

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

A Detour Down Fano's Highway

We know that feedback does not (substantially) increase the capacity of a Gaussian channel. This is very similar in flavor to our result, so can we use the same technique?

We could construct a packing set and via Fano's inequality, obtain a lower bound on

$$I(x, y) = h(y) - h(y|x) = \sum_{i=1}^{m} h(y_i|y_{[i-1]}) - h(y_i|x, y_{[i-1]})$$

where $y_{[i]} = y_1, \ldots, y_i$.

The distribution of y_i given $y_{[i-1]}$ is potentially very nasty… it is not clear how we could bound $h(y_i|y_{[i-1]})$.

Alternative Strategy

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

Step 2: Show that if you given a budget of m 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 $\pi(x)$ instead of a uniform k -sparse prior:

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

Proof of Main Result

Let $S = \{j : x_j \neq 0\}$, and $\widehat{S}$ be an estimate of S obtained via *any* adaptive measurement strategy. Set $\sigma^2 = 1$.

Lemma

Under the Bernoulli prior, if $k \leq n/2$, then

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

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

$$\begin{aligned}\|\widehat{x} - x\|_2^2 &= \sum_{j \in S} (\widehat{x}_j - x_j)^2 + \sum_{j \notin S} \widehat{x}_j^2 \\ &\geq \frac{\mu^2}{4}|S \setminus \widehat{S}| + \frac{\mu^2}{4}|\widehat{S} \setminus S| = \frac{\mu^2}{4}|\widehat{S}\Delta S|.\end{aligned}$$

Proof of Main Result

Thus, $$\mathbb{E}\,\|\widehat{x}-x\|_2^2 \geq \frac{\mu^2}{4}\mathbb{E}\,|\widehat{S}\Delta S|$$
$$\geq k\cdot\frac{\mu^2}{4}\left(1-\frac{\mu}{2}\sqrt{\frac{m}{n}}\right).$$

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

$$\mathbb{E}\,\|\widehat{x}-x\|_2^2 \geq \frac{4}{27}\cdot\frac{kn}{m} \geq \frac{1}{7}\cdot\frac{kn}{m}.$$

The hard part is proving the required lemma.

Proof of Lemma

Define $S_j = 1$ if $j \in S$ and 0 otherwise. Let $\pi_1 = k/n$ and $\pi_0 = 1 - \pi_1$.

$$\begin{aligned}\mathbb{E}\,|\widehat{S}\Delta S| &= \sum_j \mathbb{P}(\widehat{S}_j \neq S_j) \\ &= \sum_j \pi_0 \mathbb{P}(\widehat{S}_j = 1 | S_j = 0) + \pi_1 \mathbb{P}(\widehat{S}_j = 0 | S_j = 1).\end{aligned}$$

For each term in the sum, we can lower bound by the Bayes risk of the optimal detector (the LRT).

Towards this end, let $y_{[m]} = y_1, \ldots, y_m$ and define:

$$\mathbb{P}_{0,j}(y_{[m]}) = \mathbb{P}(y_{[m]} | x_j = 0)$$

$$\mathbb{P}_{1,j}(y_{[m]}) = \mathbb{P}(y_{[m]} | x_j = \mu)$$

Likelihood Ratio Test

The likelihood ratio test (LRT) will set $\widehat{S}_j = 1$ when $\pi_1 \mathbb{P}_{1,j}(y_{[m]}) > \pi_0 \mathbb{P}_{0,j}(y_{[m]})$ and has risk bounded by

$$B_j \geq \min(\pi_0, \pi_1)\left(1 - \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}\right).$$

Thus,

$$\begin{aligned}\mathbb{E}\,|\widehat{S} \Delta S| &\geq \pi_1 \sum_j \left(1 - \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}\right) \\ &\geq k - \frac{k}{\sqrt{n}} \sqrt{\sum_j \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2}.\end{aligned}$$

Our result follows from

$$\sum_j \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\mu^2}{4} m.$$

Pinsker's Inequality

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

Applying Pinsker twice we obtain

$$\|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\pi_0}{2} K(\mathbb{P}_{0,j}, \mathbb{P}_{1,j}) + \frac{\pi_1}{2} K(\mathbb{P}_{1,j}, \mathbb{P}_{0,j})$$

Consider the case of $j = 1$ and set $\mathbb{P}_0 = \mathbb{P}_{0,1}$ and $\mathbb{P}_1 = \mathbb{P}_{1,1}$.
If $x' = (x_2, \ldots, x_n)$, then we can write

$$\begin{aligned} \mathbb{P}_0(y_{[m]}) &= \sum_{x'} \mathbb{P}(x')\mathbb{P}(y_{[m]} | x_1 = 0, x') \\ &:= \sum_{x'} \mathbb{P}(x')\mathbb{P}_{0,x'}(y_{[m]}) \end{aligned}$$

and similarly for $\mathbb{P}_1$.

Bounding the KL Divergence

From the convexity of the KL divergence, we obtain

$$K(\mathbb{P}_0, \mathbb{P}_1) \leq \sum_{x'} \mathbb{P}(x') K(\mathbb{P}_{0,x'}, \mathbb{P}_{1,x'})$$

To calculate this divergence, observe that if $c_i = \sum_{j \geq 2} a_{i,j} x_j$ then $y_i = c_i + z_i$ under $\mathbb{P}_{0,x'}$ and $y_i = a_{i,1}\mu + c_i + z_i$ under $\mathbb{P}_{1,x'}$.

Moreover,

$$\mathbb{P}_{0,x'}(y_{[m]}) = \prod_{i=1}^{m} \mathbb{P}(y_i | a_i, x_1 = 0, x')$$

and similarly for $\mathbb{P}_{1,x'}$.

Bounding the KL Divergence

Combining all of this we obtain

$$
\begin{aligned}
K(\mathbb{P}_{0,x'}, \mathbb{P}_{1,x'}) &= \mathbb{E}_{0,x'} \log \frac{\mathbb{P}_{0,x'}}{\mathbb{P}_{1,x'}} \\
&= \sum_{i=1}^{m} \mathbb{E}_{0,x'} \left(\frac{1}{2}(y_1 - \mu a_{i,1} - c_i)^2 - \frac{1}{2}(y_i - c_i)^2 \right) \\
&= \sum_{i=1}^{m} \mathbb{E}_{0,x'} \left(-z_i \mu a_{i,1} + (\mu a_{i,1})^2/2 \right) \\
&= \frac{\mu^2}{2} \sum_{i=1}^{m} \mathbb{E}_{0,x'} (a_{i,1}^2)
\end{aligned}
$$

Thus, $K(\mathbb{P}_0, \mathbb{P}_1) \leq \frac{\mu^2}{2} \sum_{i=1}^{m} \mathbb{E}\left(a_{i,1}^2 | x_1 = 0\right).$

Bounding the KL Divergence

Similarly, $K(\mathbb{P}_1, \mathbb{P}_0) \leq \frac{\mu^2}{2} \sum_{i=1}^{m} \mathbb{E}\left(a_{i,1}^2 | x_1 = \mu\right)$.

Recall that we originally wanted to bound

$$\|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\pi_0}{2} K(\mathbb{P}_{0,j}, \mathbb{P}_{1,j}) + \frac{\pi_1}{2} K(\mathbb{P}_{1,j}, \mathbb{P}_{0,j}).$$

Plugging in our bound (which holds for any j) we obtain

$$\|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\mu^2}{4} \sum_{i=1}^{m} \mathbb{E}\, a_{i,j}^2 .$$

Summing over j , we finally arrive at

$$\sum_j \|\mathbb{P}_{1,j} - \mathbb{P}_{0,j}\|_{\mathrm{TV}}^2 \leq \frac{\mu^2}{4} \sum_{i,j} \mathbb{E}\, a_{i,j}^2 = \frac{\mu^2}{4} m.$$

Adaptivity in Practice

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

Algorithm 1 [Castro et al. – 2008]

- start with random (Rademacher) matrix B
- after each measurement, compute posterior distribution p
- re-weight subsequent measurements using p, i.e., set $a_i = b_i \circ \sqrt{p}$.

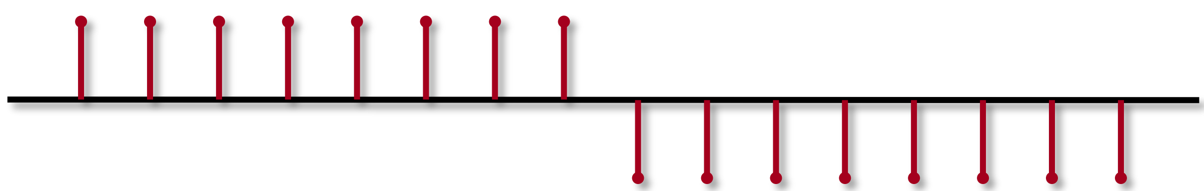
The posterior will gradually concentrate on the correct support, eventually leading to measurement vectors that use all their energy to directly measure the nonzero.

Adaptivity in Practice

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

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

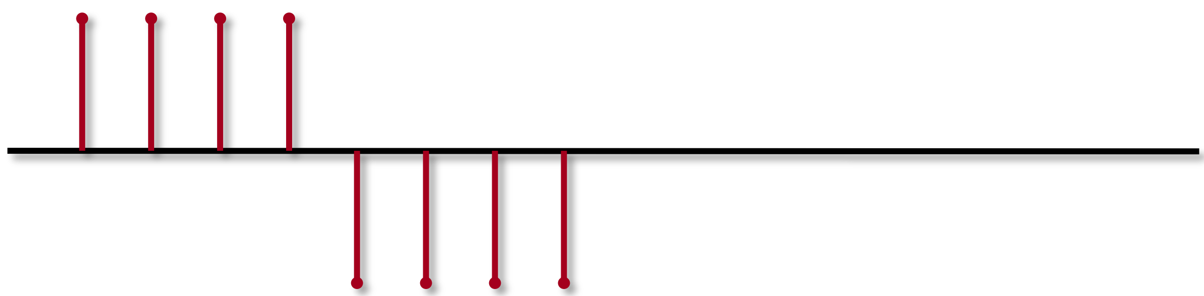


Adaptivity in Practice

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

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

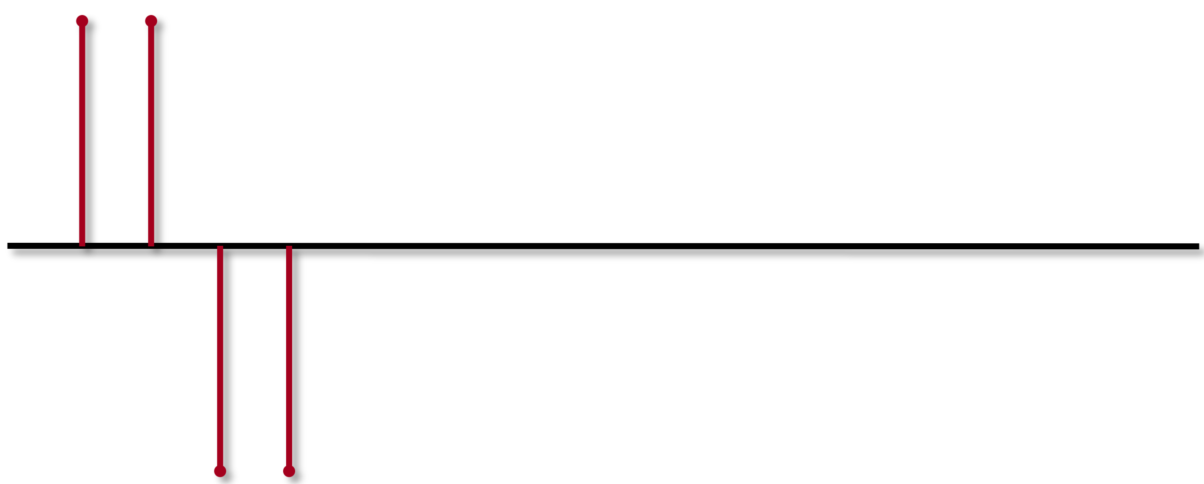


Adaptivity in Practice

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

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

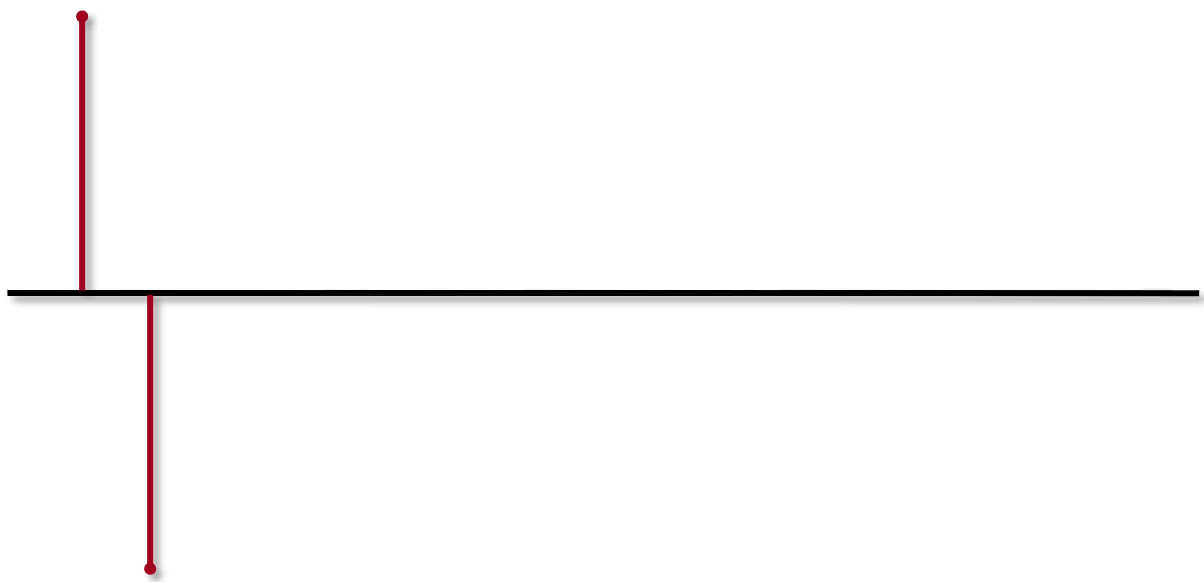


Adaptivity in Practice

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

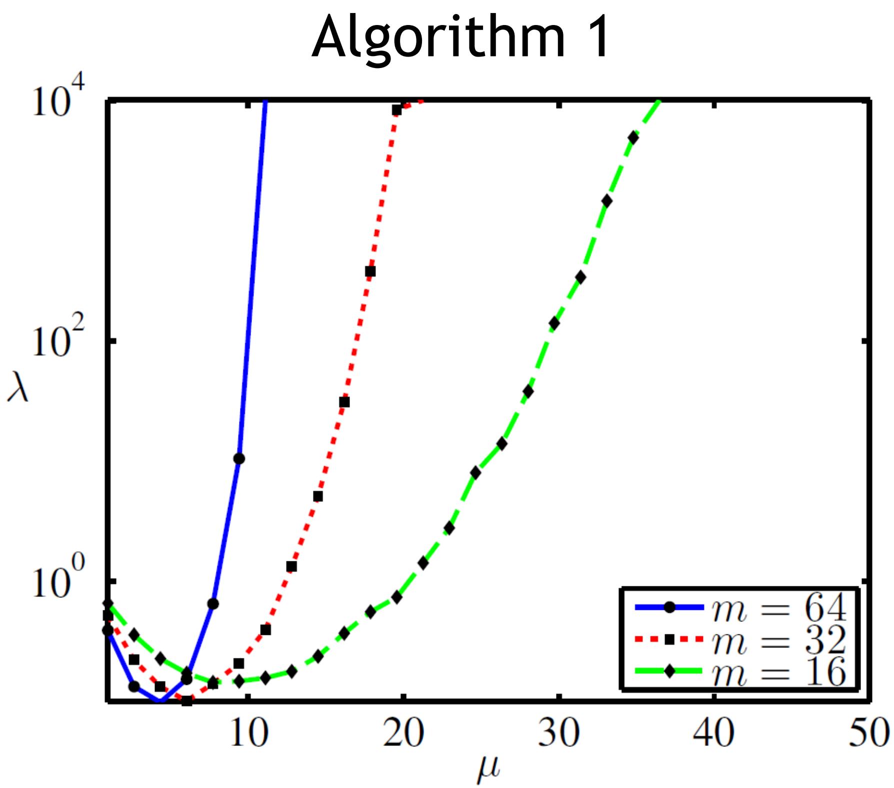
Algorithm 2 [Iwen and Tewfik – 2011]

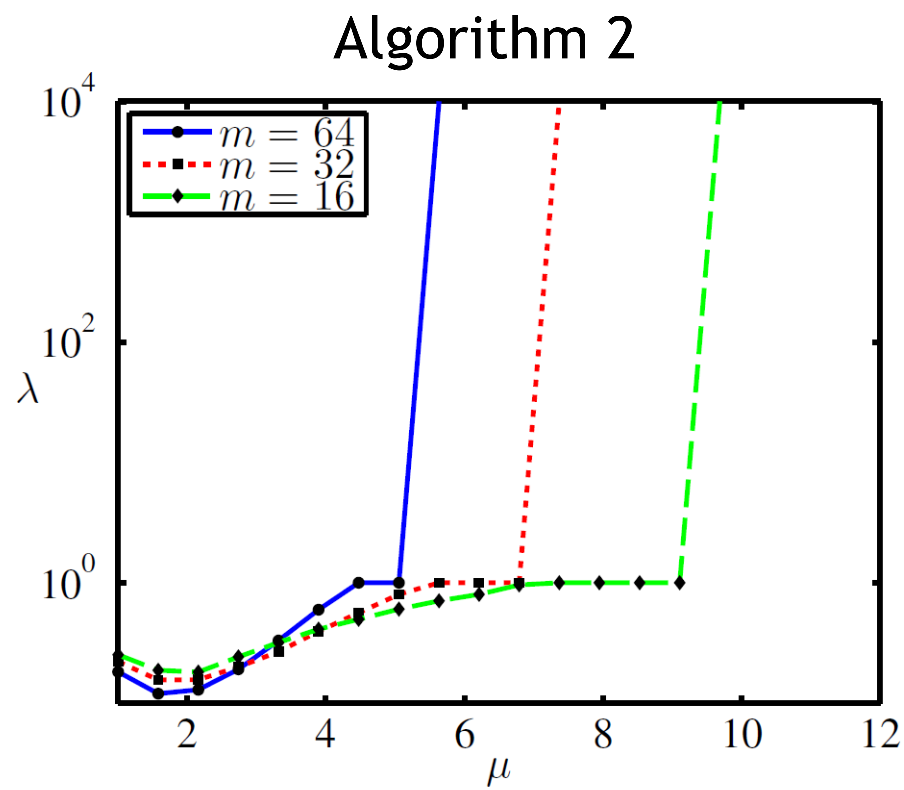
- split measurements into $\log n$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the "active set"
- after subdividing $\log n$ times, return support



Phase Transition in the Posterior

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

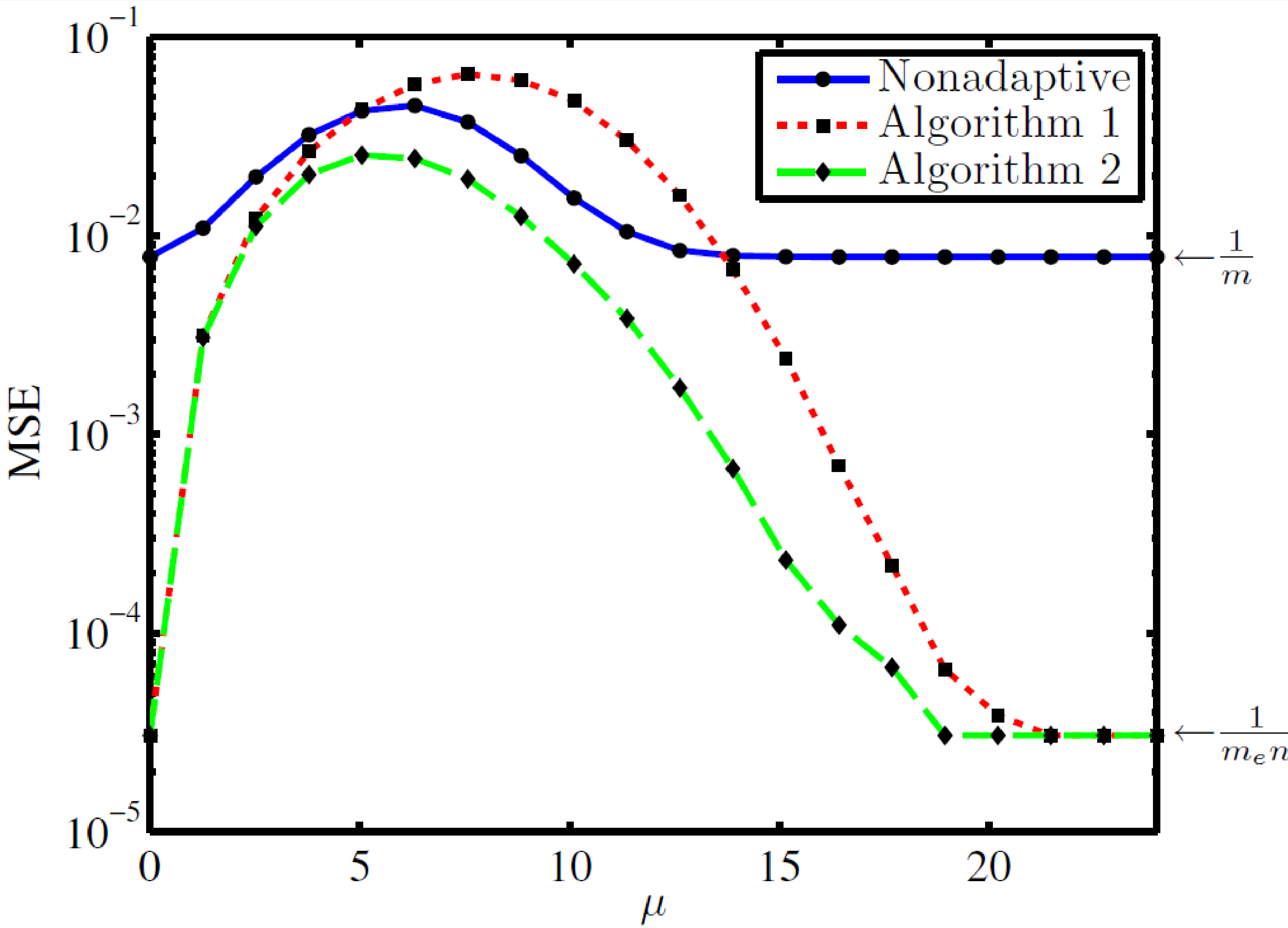




$$n = 512 \qquad \sigma^2 = 1$$

[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

- Surprisingly, adaptive algorithms, no matter how intractable, cannot significantly improve over seemingly naively simple nonadaptive strategies

- Adaptivity might still be very useful in practice
  - for a given value of μ, how many additional measurements are required to transition from the regime where adaptivity doesn't help to where it does?
  - practical adaptive algorithms that achieve the minimax rate for all values of μ ?
  - practical architectures for implementing adaptive measurements in real-world signals?