Compressed Sensing and Faster Variations

Martin J. Strauss
University of Michigan

Goals

Implicit:

- Noisy k-sparse vector $x \in \mathbb{C}^d$
- Parameter ϵ

We provide:

- Matrix Φ
- Decoding algo D with $D(\Phi x) = \widetilde{x} \approx x$.

Goals:

- **Uniformity:** One (randomly constructed) Φ works for all s
- **Number of measurements:** $k\text{poly}(\log(d), 1/\epsilon)$ rows in Φ
- **Runtime:** of D is $\text{poly}(k, \log(d), 1/\epsilon) \ll d$ (faster variant).
- **Error:** $\|E\|_2 = \|\widetilde{x} - x\|_2 \le \frac{\epsilon}{\sqrt{k}} \|x_m - x\|_1 = \frac{\epsilon}{\sqrt{k}} \|E_{\text{opt}}\|_1$

Error—Alternative Characterization

$$\|\widetilde{x} - x\|_2 \leq \frac{\epsilon}{\sqrt{k}} \|x_k - x\|_1$$

implies

- If $x_{(j)} = 1/j$ ("1-compressible"), then

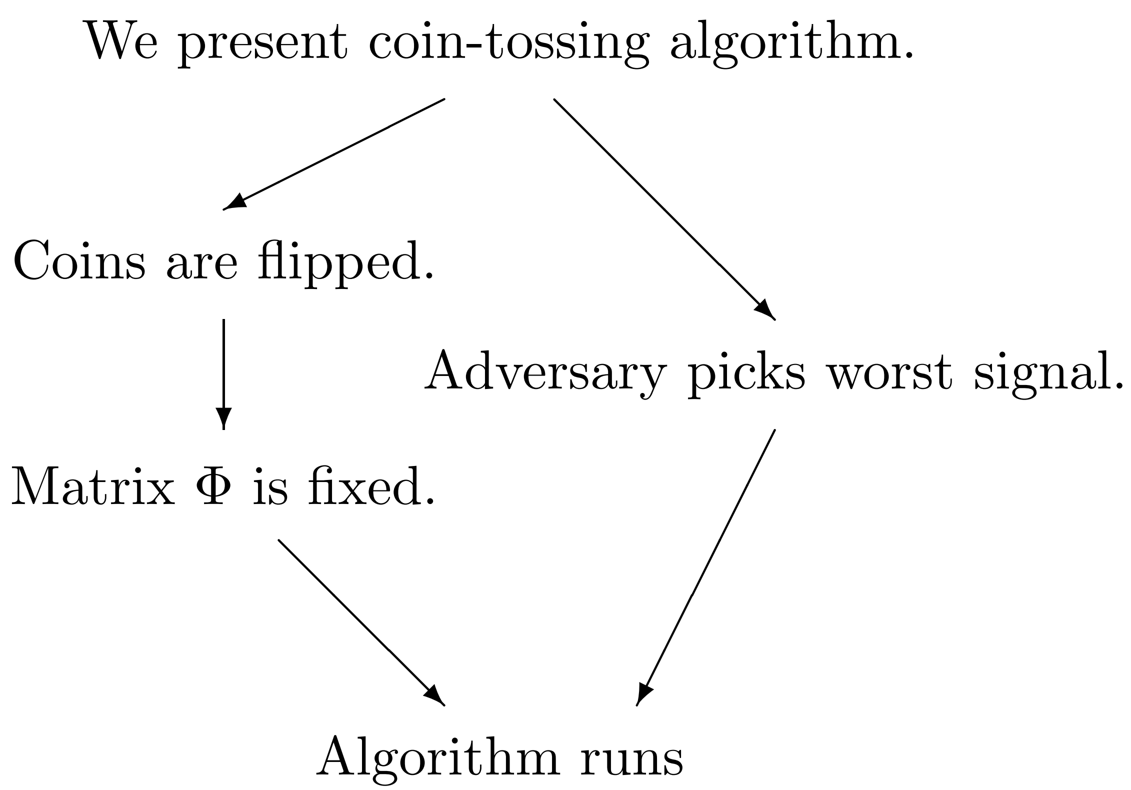
$$\|\widetilde{x}_k - x\|_2 \leq (1 + \epsilon') \|x_k - x\|_2 .$$

Role of Randomness

Signal is worst-case, not random.

Two possible models for random measurement matrix.

Random Measurement Matrix "for each" Signal



- Randomness in Φ is needed to defeat the adversary.

Universal Random Measurement Matrix

We present coin-tossing algorithm.

$\downarrow$

Coins are flipped.

$\downarrow$

Matrix Φ is fixed.

$\downarrow$

Adversary picks worst signal in ℓ^1.

$\downarrow$

Algorithm runs

- Randomness is used to construct correct Φ efficiently (probabilistic method).

Why Universal Guarantee?

Often unnecessary, but needed for iterative schemes. E.g.

- Inventory x_1: 100 Thomas, 5 Barbie, 2 Lego, 30 back-orders for TSP ...
- Sketch using Φ: 98 Thomas, -31 TSP
- Manager: Based on sketch, remove all Thomas *and* Barbie; order 40 TSP
- New inventory x_2: 0 Thomas, 0 Barbie, 2 Lego, 10 TSP, ...

x_2 depends on measurement matrix Φ. No guarantees for Φ on x_2.

Too costly to have separate Φ per sale.

Today: Universal guarantee.

Compressed Sensing

- Matrix with Restricted Isometry Property
  - ◇ E.g., random Gaussian matrix
- Decoding by linear programming

Restricted Isometry Property

Matrix Φ with d columns has the k-RIP if

- Any submatrix of k columns has $\left|\frac{\sigma_1}{\sigma_k}\right| \leq 2$.

Theorem. [Donono; Candès-Tao; Rudelson-Vershynin]

1. A Gaussian matrix with $O(k \log(d))$ rows has k-RIP.
2. A random row-submatrix of the DFT with $O(k \log^4(d))$ rows has k-RIP. (Open: Improve 4 to 1.)

Theorem. [Donoho; Candès-Tao]

If Φ has $(2k)$-RIP, and x approx'ly k-sparse, then solve

$$\min \|\widetilde{x}\|_1$$
$$\text{such that } \Phi\widetilde{x} = \Phi x.$$

Use Linear Program of size d.

Linear Program

Want to solve:

$$\begin{aligned} &\min \|\widetilde{x}\|_1 \\ &\text{such that:} \quad \Phi\widetilde{x} = \Phi x \end{aligned}$$

Write $\widetilde{x} = p - n$ as difference of positive and negative parts. Then

$$\begin{aligned} &\min(p_0 + p_1 + \cdots + p_{d-1}) + (n_0 + n_1 + \cdots + n_{d-1}) \\ &\text{such that:} \quad \Phi(p - n) = \Phi x \\ &\qquad\qquad\quad p \geq 0 \\ &\qquad\qquad\quad n \geq 0 \end{aligned}$$

Advantage of Gaussian

Measurements are oblivious to basis of sparsity.

If U is orthonormal and Φ is Gaussian, then $U\Phi$ is also Gaussian.

- Measure x; get Φx.
- Decide a good U:
  - $x = Uy$, where y is sparse plus noise.
- Pretend that we've measured y by Gaussian $U\Phi$.

E.g.,

- Make few measurements of Mars-scape.
- Later, decide on a basis that's good for compressing Mars-scapes.

Gaussians have RIP

Proof sketch; slightly worse bounds than promised.

Let A be a $O(k \log(d)) \times d$ random Gaussian matrix normalized so that columns have expected Euclidean norm 1. Then, for all x, $\|Ax\| \approx \|x\|$.

Proof overview (from Vershynin):

- Cover ball with ϵ-net, N, of size $2^{O(d)}$. (Omitted.)
- Approximate x by $y \in N$.
- Show theorem holds for *each* $y \in N$ except with prob $\frac{1}{4|N|}$. (CLT; JL)
- Need only easy upper bound for $\|A(x-y)\| \leq O(\|x-y\|)$.
- Take union bound over all y in N.

RIP suffices for LP decoding

Suppose $\Phi x^{\#} = \Phi x$ and $\left\|x^{\#}\right\|_1$ is minimal. From [Candès-Romberg-Tao:]

Let

- T_0 be support of biggest k terms and T_{01} be support of top $k + M = k + 4k = 5k$ terms.
- $\eta = k^{-1/2} \|x - x_k\|_1$.
- $h = x^{\#} - x$. (Want $\|h\|_2 \le O(\eta)$.)

Three ingredients:

- x feasible and $\left\|x^{\#}\right\|_1$ minimal implies $\left\|h_{T_0^c}\right\|_1 \le \|h_{T_0}\|_1 + \sqrt{k}\eta$.
- $\left\|h_{T_{01}^c}\right\|_2 \le O\left(\|h_{T_{01}}\|_2 + \eta\right)$.
- $\|h_{T_{01}}\|_2 \le O(\eta)$.

ℓ^1 Concentration

Theorem: x feasible and $\|x^{\#}\|_1$ is minimal implies $\|h_{T_0^c}\|_1 \le \|h_{T_0}\|_1 + O(\sqrt{k}\eta)$.

Proof:

$$\begin{aligned}\|x_{T_0}\|_1 - \|h_{T_0}\|_1 - \|x_{T_0^c}\|_1 + \|h_{T_0^c}\|_1 &\le \|x_{T_0} + h_{T_0}\|_1 + \|x_{T_0^c} + h_{T_0^c}\|_1 \\ &= \|x + h\|_1 = \|x^{\#}\|_1 \\ &\le \|x\|_1 \\ &= \|x_{T_0}\|_1 + \|x_{T_0^c}\|_1 ,\end{aligned}$$

so $\|h_{T_0^c}\|_1 \le \|h_{T_0}\|_1 + O(\sqrt{k}\eta)$.

Bounding the Tail

Theorem: $\left\|h^{c}_{T_{01}}\right\|_2 \le O\left(\|h_{T_{01}}\|_2 + \eta\right)$.

Proof: Markov: $j \cdot |y|_{(j)} \le \|y\|_1$, so $|h_{T_0^c}|_{(j)} \le \left\|h_{T_0^c}\right\|_1 / j$.

Thus

$$\left\|h_{T_{01}^c}\right\|_2^2 \le \left\|h_{T_0^c}\right\|_1^2 \sum_{j=M+1}^{d} \frac{1}{j^2} \le \left\|h_{T_0^c}\right\|_1^2 / M.$$

Combined with ℓ^1 Concentration,

$$\left\|h_{T_{01}^c}\right\|_2^2 \le O((\|h_{T_0}\|_1 / \sqrt{M} + \eta)^2) \le O((\|h_{T_0}\|_2 + \eta)^2).$$

Bounding the Head

Theorem: $\|h_{T_{01}}\|_2 \leq O(\eta)$.

Proof: For $j > 0$, let T_j be the support of j'th largest set of M terms *after* the first k.

$$\begin{aligned} 0 &= \left\|\Phi(x^{\#} - x)\right\|_2 = \|\Phi h\|_2 \\ &\geq \|\Phi h_{T_{01}}\|_2 - \left\|\sum_{j\geq 2} \Phi h_{T_j}\right\|_2 \\ &\geq \|\Phi h_{T_{01}}\|_2 - \sum_{j\geq 2} \left\|\Phi h_{T_j}\right\|_2 \\ &\approx \|h_{T_{01}}\|_2 - \sum_{j\geq 2} \left\|h_{T_j}\right\|_2 . \end{aligned}$$

Need to bound $\sum_{j\geq 2} \left\|h_{T_j}\right\|_2$ above.

Bounding the Head

Each term in T_{j+1} is smaller than the average term in T_j, $|h_{T_{j+1}}| \le \left\|h_{T_j}\right\|_1 / M$, so $\left\|h_{T_{j+1}}\right\|_2^2 \le \left\|h_{T_j}\right\|_1^2 / M$. (Note: Tight $1/M$ factor in 1 to 2 norm achieved by j to $j+1$.) Thus

$$\begin{aligned}
\sum_{j\ge 2} \left\|h_{T_j}\right\|_2 &\le \sum_{j\ge 1} \left\|h_{T_j}\right\|_1 / \sqrt{M} \\
&= \left\|h_{T_0^c}\right\|_1 / \sqrt{M} \\
&\le \|h_{T_0}\|_1 / \sqrt{M} + O(\eta) \\
&\le \sqrt{k/M}(\|h_{T_0}\|_2 + O(\eta)) \\
&\le \sqrt{k/M}(\|h_{T_{01}}\|_2 + O(\eta)).
\end{aligned}$$

Thus $\|h_{T_{01}}\|_2 \le \sqrt{k/M}(\|h_{T_{01}}\|_2 + \eta) \le (1/2)(\|h_{T_{01}}\|_2 + \eta)$, so

$$\|h_{T_{01}}\|_2 \le O(\eta).$$

HHS algorithm

Co-design matrix special and decoding algorithm.

Faster decoding: time $k^2\text{poly}(\log(d)/\epsilon)$.

Fast Estimation

Have:

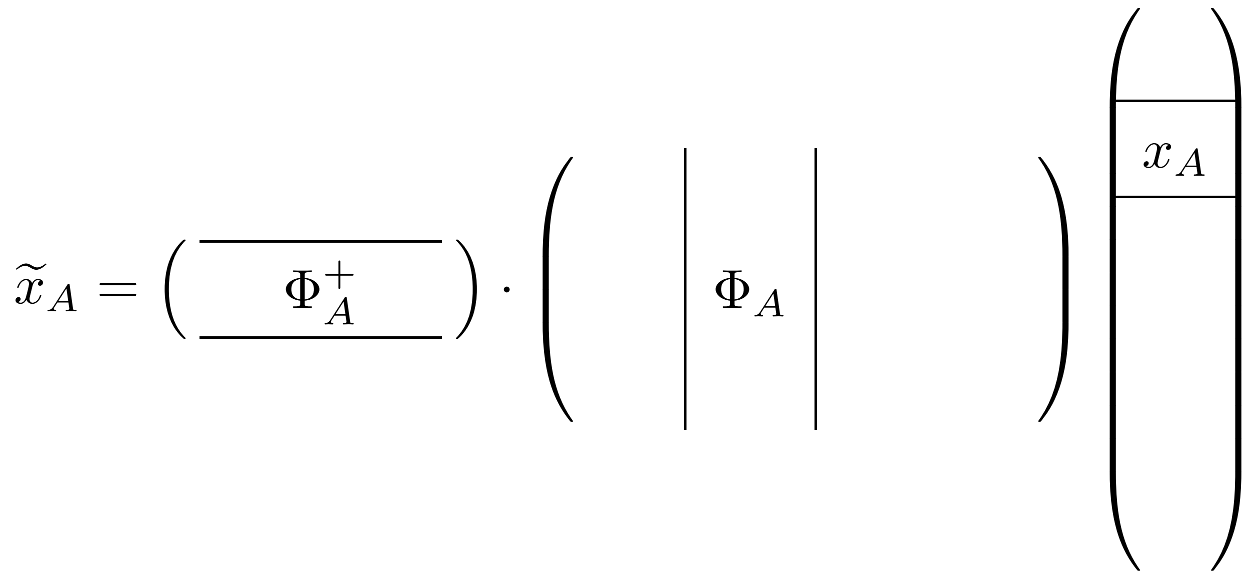
- Set A of positions in signal x.
- Measurements Φx, for random DFT-row-submatrix Φ.

Want:

- Estimate $\widetilde{x}_A$ for x_A with
- $\|\widetilde{x}_A - x_A\|_2 \le \|x - x_A\|_2 + k^{-1/2}\,\|x - x_A\|_1$.

Estimator

$\widetilde{x}_A = \Phi_A^+(\Phi x)$ (Least squares).



Get: For all x, $\|\Phi x\|_2 \le O(\|x\|_2 + k^{-1/2}\|x\|_1)$.

Proof of correctness: Similar to Compressed Sensing.

$k\text{polylog}(d) \times k\text{polylog}(d)$ matrix; time $k^2\text{polylog}(d)$.

On to Identification

How to find good candidate set of positions?

- Isolation
- Noise Reduction

HHS Algorithm, Overview

- Assume limited dyanmic range: $\|x\|_2 \le d^{\log(d)} \|E_{\text{opt}}\|_1$.
  - ◇ Previous work provides preprocessing step.
- While $\|x\|_2 > (\epsilon/\sqrt{k}) \|E_{\text{opt}}\|_1$, reduce $\|x\|_2$ by factor 2.
  - ◇ Identify some spikes
  - ◇ Estimate values.
  - ◇ ... reduce $\|x\|_2$ by a constant factor.

HHS

Our focus:

- $\approx q$ spikes with magnitude $\approx 1/t$
- Noise $\|E_{\text{opt}}\|_1 = \|\nu\|_1 = 1$.

(Try all q's and t's in a geometric progression.)

Double Hashing

Have: q spikes at magnitude $1/t$; noise 1.

Double hashing:

- Each position goes to 1 group among q.
- Within each group, each position expects to go to t/q groups among $(t/q)^2$.

(Some log factors suppressed.)

First Hashing

Have: q spikes at magnitude $1/t$; noise 1.

Throw positions into $\approx q$ buckets, by Φ. Repeat $\log(d)$ times. Except with prob $e^{-q\log(d)} = \binom{d}{q}^{-1}$,

- $\Omega(q)$ spikes are isolated from other spikes

Take union bound over all $\binom{d}{q}$ possible configurations of spikes. Get one spike at $1/t$.

Noise, Part I

Have one spike at $1/t$. Noise?

We'll show $\|\Phi E_{\mathrm{opt}}\|_1 \le \|E_{\mathrm{opt}}\|_1$. (Next slide.)

- Property of Φ; no union bound over E_{opt}.
- At most $n/10$ of n buckets get noise more than $(10/n)\|E_{\mathrm{opt}}\|_1 \approx (1/q)\|E_{\mathrm{opt}}\|$.

Get 1 spike at $1/t$ and noise $1/q$.

- Need further q/t factor of noise reduction.

Noise, Part I, Illustrated

Throw d positions into $n = q\log(d)$ buckets, by Φ.

- Want $\|\Phi E_{\mathrm{opt}}\|_1 \le \|E_{\mathrm{opt}}\|_1$; we'll show $\|\Phi x\|_1 \le \|x\|_1$ for *all* x.
- At most $n/10$ buckets get noise more than $(10/n)\,\|E_{\mathrm{opt}}\|_1 \approx (1/q)\,\|E_{\mathrm{opt}}\|$.

$$\begin{pmatrix} 7 \\ 9 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

Second Hashing

Have 1 spike at $1/t$; noise $\|\nu\|_1 \le 1/q$.

Use $r = (t/q)^2$ rows of Bernoulli(q/t).

$$\begin{pmatrix} & \downarrow & & & & & & \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/dq \\ 1/t \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \end{pmatrix} \leftarrow$$

Second Hashing

Have 1 spike at $1/t$; noise $\|\nu\|_1 \leq 1/q$.

Use $r = (t/q)^2$ rows of Bernoulli(q/t).

$$\begin{pmatrix} & \downarrow & & & & & & \\ & & & & & & & \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ & & & & & & & \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/dq \\ 1/t \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \end{pmatrix} \leftarrow$$

- Our spike survives $r' = r \cdot (q/t) = t/q$ times.

Second Hashing

Have 1 spike at $1/t$; noise $\|\nu\|_1 \leq 1/q$.

Use $r = \widetilde{O}((t/q)^2)$ rows of Bernoulli(q/t).

$$\begin{pmatrix} & \color{red}{\downarrow} & & & & & \\ \\ 0 & & 1 & 0 & 0 & 0 & 1 & 1 \\ \\ 1 & & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/dq \\ \color{red}{1/t} \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \end{pmatrix} \color{red}{\leftarrow}$$

- Our spike survives $r' = r \cdot (q/t) = t/q$ times.
- On surviving submatrix, expect $r' \cdot (q/t) =$ one 1 per other column.

Second Hashing

Have 1 spike at $1/t$; noise $\|\nu\|_1 \leq 1/q$.

Except with prob $1/d^3$ (with cost factor $O(\log(d))$),

- Our spike survives $r' = r \cdot (q/t) = t/q$ times.
- In surviving submatrix, $r' \cdot (q/t) =$ one 1 per each other column.

Take union bound over d spikes and d matrix columns.

For any noise $\|\nu\|_1 = 1/q$, some row gets average noise, $(1/q)/r' = 1/t$.

Can recover spike of magnitude $1/t$ from noise $1/(2t)$.

Number of Measurements

Number of measurements: $q(t/q)^2 \log(d) \approx t^2/q$, for

- First hashing (q rows)
- Second hashing ($(t/q)^2$ rows)
- Bit tests ($\log(d)$ rows)
- (Several!) omitted factors of $\log(d)$ and $1/\epsilon$.

Note: $q/t^2 = \|s\|_2^2 > (k^{-1/2}\|E_{\text{opt}}\|_1)^2 = 1/k$.

So number of measurements is $t^2/q \leq k$.

Recap

New compressed sensing/heavy hitter algorithms that get

- Appropriate error
- Universal guarantee
- Optimal number of measurements (up to log factors)
- Decoding time poly$(k \log(d))$

Also

- Efficient pseudorandom constructions suffice.