One sketch for all: Fast algorithms for compressed sensing

Martin J. Strauss
University of Michigan

Covers joint work with
Anna Gilbert (Michigan),
Joel Tropp (Michigan), and
Roman Vershynin (UC Davis)

Heavy Hitters/Sparse Recovery

Sparse Recovery is the idea that *noisy sparse* signals *can* be *approximately* reconstructed *efficiently* from a *small number* of nonadaptive linear measurements.

Known as "Compress(ed/ive) Sensing," or the "Heavy Hitters" problem in database.

Simple Example

Measurements

Signal, s

Measurement matrix, Φ

$$\begin{pmatrix} 5.3 \\ \cdots \\ 0 \\ 5.3 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 5.3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Recover position and coefficient of single spike in signal.

In Streaming Algorithms

- Maintain vector s of frequency counts from transaction stream:
  - ◇ 2 spinach sold, 1 spinach returned, 1 kaopectate sold, ...
- Recompute top-selling items upon each new sale

Linearity of Φ:

- $\Phi(s + \Delta s) = \Phi(\Delta s)$.

Goals

- **Input:** All noisy m-sparse vectors in d dimensions
- **Output:** Locations and values of the m spikes, with
  - **Error Goal:** Error proportional to the optimal m-term error

Resources:

- **Measurement Goal:** $n \le m \text{polylog} d$ fixed measurements
- **Algorithmic Goal:** Computation time $\text{poly}(m \log(d))$
  - Time close to *output* size $m \ll d$.
- **Universality Goal:** One matrix works for all signals.

Overview

- One sketch for all
- Goals and Results
- Chaining Algorithm
- HHS Algorithm (builds on Chaining)

Role of Randomness

Signal is worst-case, not random.

Two possible models for random measurement matrix.

Random Measurement Matrix "for each" Signal

We present coin-tossing algorithm.

Coins are flipped.

Adversary picks worst signal.

Matrix Φ is fixed.

Algorithm runs

- 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.

$\downarrow$

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 s_1: 100 spinach, 5 lettuce, 2 bread, 30 back-orders for kaopectate ...
- Sketch using Φ: 98 spinach, -31 kaopectate
- Manager: Based on sketch, remove all spinach *and* lettuce; order 40 kaopectate
- New inventory s_2: 0 spinach, 0 lettuce, 2 bread, 10 kaopectate, ...

s_2 depends on measurement matrix Φ. No guarantees for Φ on s_2.

Too costly to have separate Φ per sale.

Today: Universal guarantee.

Overview

- One sketch for all ✓
- Goals and Results
- Chaining Algorithm
- HHS Algorithm (builds on Chaining)

Goals

- Universal guarantee: one sketch for all
- Fast: decoding time $\text{poly}(m \log(d))$
- Few: optimal number of measurements (up to log factors)

Previous work achieved two out of three.

Ref.	Univ.	Fast	Few meas.	technique
KM	×	✓	✓	comb'l
D, CRT	✓	×	✓✓	LP(d)
CM*	✓✓	✓	×	comb'l
Today	✓	✓	✓	comb'l

*restrictions apply

Results

Two algorithms, Chaining and HHS.

$\widetilde{O}$ hides factors of $\log(d)/\epsilon$.

	# meas.	Time	# out	error
Chg	$\widetilde{O}(m)$	$\widetilde{O}(m)$	m	$\lVert E\rVert_1 \le O(\log(m))\,\lVert E_{\text{opt}}\rVert_1$

Results

Two algorithms, Chaining and HHS.

$\widetilde{O}$ hides factors of $\log(d)/\epsilon$.

	# meas.	Time	# out	error
Chg	$\widetilde{O}(m)$	$\widetilde{O}(m)$	m	$\|E\|_1 \leq O(\log(m)) \|E_{\text{opt}}\|_1$
HHS	$\widetilde{O}(m)$	$\widetilde{O}(m^2)$	$\widetilde{O}(m)$	$\|E\|_2 \leq (\epsilon/\sqrt{m}) \|E_{\text{opt}}\|_1$

Results

Two algorithms, Chaining and HHS.

$\widetilde{O}$ hides factors of $\log(d)/\epsilon$.

	# meas.	Time	# out	error
Chg	$\widetilde{O}(m)$	$\widetilde{O}(m)$	m	$\lVert E\rVert_1 \leq O(\log(m))\,\lVert E_{\mathrm{opt}}\rVert_1$
HHS	$\widetilde{O}(m)$	$\widetilde{O}(m^2)$	$\widetilde{O}(m)$	$\lVert E\rVert_2 \leq (\epsilon/\sqrt{m})\,\lVert E_{\mathrm{opt}}\rVert_1$
3			m	$\lVert E\rVert_2 \leq \lVert E_{\mathrm{opt}}\rVert_2 + (\epsilon/\sqrt{m})\,\lVert E_{\mathrm{opt}}\rVert_1$
4				$\lVert E\rVert_1 \leq (1+\epsilon)\,\lVert E_{\mathrm{opt}}\rVert_1$

(3) and (4) are gotten by truncating output of HHS.

Results

	# meas.	Time	error	Failure
K-M	$\widetilde{O}(m)$	$\mathrm{poly}(m)$	$\Vert E\Vert_2 \le (1+\epsilon)\Vert E_{\mathrm{opt}}\Vert_2$	"for each"
D, C-T	$O(m\log(d))$	$d^{(1\mathrm{to}3)}$	$\Vert E\Vert_2 \le (\epsilon/\sqrt{m})\Vert E_{\mathrm{opt}}\Vert_1$	univ.
CM	$\widetilde{O}(m^2)$	poly(m)	$\Vert E\Vert_2 \le (\epsilon/\sqrt{m})\Vert E_{\mathrm{opt}}\Vert_{<1}$	Det'c
Chg	$\widetilde{O}(m)$	$\widetilde{O}(m)$	$\Vert E\Vert_1 \le O(\log(m))\Vert E_{\mathrm{opt}}\Vert_1$	univ.
HHS	$\widetilde{O}(m)$	$\widetilde{O}(m^2)$	$\Vert E\Vert_2 \le (\epsilon/\sqrt{m})\Vert E_{\mathrm{opt}}\Vert_1$	univ.

$\widetilde{O}$ and poly() hide factors of $\log(d)/\epsilon$.

Overview

- One sketch for all ✓
- Goals and Results ✓
- Chaining Algorithm
- HHS Algorithm (builds on Chaining)

Chaining Algorithm—Overview

- Handle the universal guarantee
- Group testing
  - Process several spikes at once
  - Reduce noise
- Process single spike bit-by-bit as above.
- Iterate on residual.

Universal Guarantee

- Fix m spike positions
- Succeed except with probability $\exp(-m \log(d))/4$
  - succeed "for each" signal
- Union bound over all spike configurations.
  - At most $\exp(m \log(d))$ configurations of spikes.
  - Convert "for each" to universal model

Noisy Example—Isolation

Each group is defined by a mask:

signal:	0.1	0	5.3	0	0	-0.1	0.2	6.8
random mask:	1	1	1	0	1	0	1	0
product:	0.1	0	5.3	0	0	0	0.2	0

Noisy Example

$$\begin{pmatrix} 5.6 \\ \cdots \\ 0.2 \\ 5.5 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.1 \\ 0 \\ 5.3 \\ 0 \\ 0 \\ 0 \\ 0.2 \\ 0 \end{pmatrix}$$

Recover position and coefficient of single spike, even with noise.

(Mask and bit tests combine into measurements.)

Group Testing for Spikes

E.g., m spikes (i, s_i) at height $1/m$; $\|\text{noise}\|_1 = 1/20$. (For now.)

- (i, s_i) is a spike if $|s_i| \geq \left(\frac{1}{m}\right) \|\text{noise}\|_1$.

Group Testing for Spikes

E.g., m spikes (i, s_i) at height $1/m$; $\|\text{noise}\|_1 = 1/20$. (For now.)

- (i, s_i) is a spike if $|s_i| \geq \left(\frac{1}{m}\right) \|\text{noise}\|_1$.

Throw d positions into $n = O(m)$ groups, by Φ.

- $\geq c_1 m$ of m spikes isolated in their groups
- $\leq c_2 m$ groups have noise $\geq 1/(2m)$ (see next slide.)
- $\geq (c_1 - c_2)m$ groups have unique spike and low noise—recover!

...except with probability e^{-m}.

Repeat $O(\log(d))$ times:

$$\boxed{\text{Recover } \Omega(m) \text{ spikes except with prob } e^{-m\log(d)}.}$$

Noise

- $\|\Phi E_{\mathrm{opt}}\|_1 \le \|\Phi\|_{1\to 1} \|E_{\mathrm{opt}}\|_1$.
- We'll show $\|\Phi\|_{1\to 1} \le 1$.
- Thus *total* noise contamination is at most the signal noise.
- At most $m/10$ buckets get noise more than $(10/m)\|E_{\mathrm{opt}}\|_1$

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

We've found some spikes

We've found $(1/4)m$ spikes.

- Subtract off spikes (in sketch): $\Phi(s - \Delta s) = \Phi s - \Phi(\Delta s)$.
- Recurse on problem of size $(3/4)m$.
- Done after $O(\log(m))$ iterations.

But...

More Noise Issues

- $\geq c_1 m$ of n groups have unique spikes (of m) ✓
- $\leq c_2 m$ groups have noise $\geq 1/(2m)$ ✓
- $\leq c_3 m$ groups have false spike
  - ◇ Subtract off large phantom spike
  - ◇ Introduce new (negative) spike (to be found later)
- Other groups contribute additional noise (never to be found)
  - ◇ Spike threshold rises from m^{-1} to $\left(\frac{3m}{4}\right)^{-1}$.

More Noise Issues

- $\geq c_1 m$ of n groups have unique spikes (of m) ✓
- $\leq c_2 m$ groups have noise $\geq 1/(2m)$ ✓
- $\leq c_3 m$ groups have false spike
- Other groups contribute additional noise (never to be found)

Number of spikes:

$$m \quad \rightarrow \quad (c_1 - c_2 - c_3)m \approx (3/4)m.$$

Spike threshold increases—delicate analysis.

- Need spike (i, s_i) with $|s_i| \geq \Omega\left(\frac{1}{m\log(m)}\right)\|\text{noise}\|_1$.
  - ◇ Lets noise grow from round to round.
- Prune carefully to reduce noise.
- Get log factor in approximation.

Drawbacks with Chaining Pursuit

- log factor in error
- 1-to-1 error bound is weaker than standard 1-to-2

Drawbacks with Chaining Pursuit

- log factor in error
- 1-to-1 error bound is weaker than standard 1-to-2

Two algorithms, Chaining and HHS.

	# meas.	Time	# out	error
Chg	$\widetilde{O}(m)$	$\widetilde{O}(m)$	m	$\|E\|_1 \le O(\log(m))\,\|E_{\text{opt}}\|_1$
HHS	$\widetilde{O}(m)$	$\widetilde{O}(m^2)$	$\widetilde{O}(m)$	$\|E\|_2 \le (\epsilon/\sqrt{m})\,\|E_{\text{opt}}\|_1$
3			m	$\|E\|_2 \le \|E_{\text{opt}}\|_2 + (\epsilon/\sqrt{m})\,\|E_{\text{opt}}\|_1$
4				$\|E\|_1 \le (1+\epsilon)\,\|E_{\text{opt}}\|_1$

Overview

- Assume limited dyanmic range: $\|s\|_2 \le d^{\log(d)} \|E_{\text{opt}}\|_1$.
  - ◇ E.g., preprocess with (simplified) Chaining algorithm
- While $\|s\|_2 > (\epsilon/\sqrt{m}) \|E_{\text{opt}}\|_1$, reduce $\|s\|_2$ by factor 2.
  - ◇ Identify fraction of spikes
  - ◇ Estimate values.
    - Separation of Identification and Estimation eliminates problems caused by false positives.

2-error

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.)

Remark:

- In Chaining $(1 \leftarrow 1)$ setup, can assume $1/t \geq 1/q$. (Spike height $1/t$ is big.)
- Challenge here: Possibly $1/t = 1/\sqrt{qm}$.

Double Hashing

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

Double hashing:

- Each position goes to 1 group among q. (As in Chaining.)
- 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 $1/t$; noise 1.

Throw positions into q buckets, by Φ. As in Chaining, except with prob $e^{-q\log(d)} = \binom{d}{q}^{-1}$,

- $\Omega(q)$ spikes are isolated from other spikes
- $\|\Phi\|_{1\to 1} \le 1$.
  - ◇ Thus only $O(q)$ buckets get noise more than $1/q$.

Second Hashing

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

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

$$
\begin{array}{c}
\begin{array}{cccccccc} & \downarrow & & & & & & \end{array} \\
\begin{pmatrix}
0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0
\end{pmatrix}
\end{array}
\begin{pmatrix}
1/dq \\ 1/t \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq \\ 1/dq
\end{pmatrix}
$$

Second Hashing

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

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

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

- 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).

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

- 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 $\|\nu\|_1 \leq 1/q$.

With prob $1/d^3$,

- Our spike survives $r' = r \cdot (q/t) = t/q$ times.
- On surviving submatrix, expect $r' \cdot (q/t) =$ one 1 per 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: $q(t/q)^2 \log(d) = \mathrm{poly}(\log(d)/\epsilon)t^2/q$, 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: $q/t^2 = \|s\|_2^2 > (m^{-1/2}\|E_{\mathrm{opt}}\|_1)^2 = 1/m$.

So number of measurements is $t^2/q \le m$.

Cost

Re-measure $\widetilde{O}(m)$-sparse vector by matrix with at most $\widetilde{O}(m)$ rows:

- Time: $m^2\mathrm{poly}(\log(d)/\epsilon)$.

Matrix generation, first hashing:

- Generate m rv's from m-wise independent family
- Time $m\mathrm{polylog}(d)$.

Matrix generation, second hashing:

- m times, generate m rv's from 2-wise independent family
- Time $m^2\mathrm{polylog}(d)$.

Improvement to $m^{3/2}$ possible here; bottleneck of m^2 in Estimation.

Estimation

Have:

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

Want:

- Estimate $\widetilde{s}_A$ for s_A with
- $\|\widetilde{s}_A - s_A\|_2 \leq \|s - s_A\|_2 + m^{-1/2}\,\|s - s_A\|_1$.

Note: Can assume by $\|s - s_A\|_2$ small, by goodness of identification.

Estimator

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

$$\widetilde{s}_A = \left(\Phi_A^+ \right) \cdot \left(\Phi_A \right) \left(s_A \right)$$

- Correctness mostly follows from Candès-Tao, Rudelson-Vershynin.
- Small space and runtime $\widetilde{O}(m^2)$ immediate.
- Open: $m \times m$ DFT submatrix times vector faster than m^2.

Recap

New compressed sensing/heavy hitter algorithms that get

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

Chaining material based on paper:

> Algorithmic Linear Dimension Reduction in the ℓ_1 Norm for Sparse Vectors (available from my homepage)

HHS material based on paper:

> One sketch for all: Fast algorithms for compressed sensing (submitted; available soon.)

by Gilbert, Strauss, Tropp, Vershynin

Euclid v. Taxicab

Optimal error vector $E_{\mathrm{opt}} = s - s_m$ is s with m heavy hitters *zeroed out*.

Our error vector is $E = s - \widetilde{s}$.

- Ideally, $\|E\|_2 \leq (1+\epsilon)\,\|E_{\mathrm{opt}}\|_2$.
  - ◇ Achievable with "for each" guarantee
  - ◇ Impossible with universal guarantee (Cohen-Dahmen-DeVore, 2006)
- Best with universal guarantee is $\|E\|_2 \leq \frac{\epsilon}{\sqrt{m}}\,\|E_{\mathrm{opt}}\|_1$ (and related).

Alternative Characterization

- $\|E\|_2 \le (1+\epsilon)\,\|E_{\text{opt}}\|_2$ vacuous unless $E_{\text{opt}} \in B_2(1)$.
- $\|E\|_2 \le \frac{\epsilon}{\sqrt{m}}\,\|E_{\text{opt}}\|_1$ vacuous unless $E_{\text{opt}} \in B_1(\sqrt{m}/\epsilon)$.

Defeat Φ by finding s with $s \in \text{null}(\Phi)$.

- Any Φ: There's $s \in \text{null}(\Phi)$ with $E_{\text{opt}} \in B_2(1)$.
- Our Φ: There's no $s \in \text{null}(\Phi)$ with $E_{\text{opt}} \in B_1(\sqrt{m}/\epsilon)$.

Today: Universal failure guarantee, with ℓ^1 noise.

Cor.: Algorithmic Dimension Reduction

Goal: $(\mathbb{R}^d, \ell^1) \to (\mathbb{R}^n, \ell^1)$, for $n \ll d$.

Impossibility results, in general (Brinkman and Charikar, 2003)

Chaining algorithm:

$$(\mathbb{X}_m^d, \ell^1) \to (\mathbb{R}^n, \ell^1),$$

for $n = m\mathrm{polylog} d$, and $\mathbb{X}_m^d \subseteq \mathbb{R}^d$ is m-sparse signals.

- Robust to perturbations
- Compute and invert in time $m\mathrm{polylog} d$.
- Distortion $\mathrm{polylog}(m)$.

cf. Charikar and Sahai: Distortion $(1 + \epsilon)$ but
$n = \Theta((m/\epsilon)^2 \log(d))$.

Analysis

$$\begin{aligned}
\|\widetilde{s}_A - s_A\|_2 &= \left\|\Phi_A^+ \Phi s - s_A\right\|_2 \\
&= \left\|\Phi_A^+ \Phi (s - s_A)\right\|_2 \\
&\leq O(\|s - s_A\|_K) \quad \text{(Need this!)} \\
&= O(\|s - s_A\|_2 + m^{-1/2}\|s - s_A\|_1).
\end{aligned}$$

We'll bound $\left\|\Phi_A^+ \Phi\right\|_{K \to 2}$ by bounding

- $\left\|\Phi_A^+\right\|_{2 \to 2}$
- $\|\Phi\|_{K \to 2}$

Operator bounds

Need to bound

- $\|\Phi_A^+\|_{2\to 2}$
- $\|\Phi\|_{K\to 2}$

Candès-Tao, Rudelson-Vershynin:

- All size-$(2m)$ column submatrices are near-isometries (RIC)
- ...so $\|\Phi^+\|_{2\to 2} \le 2$ immediately

We show RIC implies bound on $\|\Phi\|_{K\to 2}$.

K to 2 Bound

If s is q spikes of (near-)equal size, $m \le q \le 2m$, then $\|\Phi s\|_2 \le m^{-1/2}\|s\|_1$.

Suppose $\|\Phi x\|_2 \le m^{-1/2}\|\Phi x\|_1$ and $\|\Phi y\|_2 \le m^{-1/2}\|\Phi y\|_1$, for x and y disjointly supported. Then

$$\begin{aligned}
\|\Phi(x+y)\|_2 &\le \|\Phi x\|_2 + \|\Phi y\|_2 \\
&\le m^{-1/2}(\|x\|_1 + \|y\|_1) \\
&= m^{-1/2}\|x+y\|_1 \\
&\le m^{-1/2}\|x+y\|_K
\end{aligned}$$

Combine all groups of size $\ge m$ this way.

K to 2 Bound

If s is $q \leq m$ spikes of (near-)equal size t, then $\|\Phi s\|_2 \leq \|s\|_2$.

Do all $q = 1, 2, 4, 8, \ldots, m$ and $O(\log(d))$ relevant values of t.
Suppose $\|\Phi x\|_2 \leq \|x\|_2$ and $\|\Phi y\|_2 \leq \|y\|_2$, for x and y disjointly supported. Then

$$\begin{aligned} \|\Phi(x+y)\|_2 &\leq \|\Phi x\|_2 + \|\Phi y\|_2 \\ &\leq \|x\|_2 + \|y\|_2 \\ &= \sqrt{2}\,\|x+y\|_2\,, \end{aligned}$$

by Cauchy-Schwarz. Give up factor polylog(d) in this proof.

Slicker proof gives no overhead from RIC to $K \to 2$ norm.