Tight Bounds For L1 Oblivious Subspace Embeddings

Ruosong Wang David Woodruff
Computer Science Department
CMU

Least Squares Regression

- Given n data points in R^d: $a_1, a_2, \ldots, a_n$
- Their corresponding values: $b_1, b_2, \ldots, b_n$

- Goal: find x in R^d to minimize $\Sigma(b_i - \langle a_i, x \rangle)^2$

- Matrix form: Given A in $R^{n \times d}$, b in R^n , find x in R^d to minimize $|Ax-b|_2$

Lp Regression

- Given A in $R^{n \times d}$, b in R^n , find x in R^d to minimize $|Ax-b|_p$

- p = 2: Least Squares Regression
- p = 1: Least Absolute Deviation Regression

- Focus on over-constrained case: n >> d

Algorithm for Least Squares Regression

- We know $x^* = A^- b$
- Calculating x^* exactly takes $O(nd^2)$ time

- Speed up by relaxing the problem
  - Allow approximation
  - Allow randomized algorithms

Subspace Embedding [Sarlos'06]

- Given A in $R^{n \times d}$
- Random matrix S in $R^{r \times n}$ is an Lp subspace embedding if
  - with constant probability, simultaneously for all x in R^d
  - $|Ax|_p \leq |SAx|_p \leq \kappa |Ax|_p$

- Algorithm for solving Lp regression
  - 1. Calculate a subspace embedding S for [A b]
  - 2. Minimize $|SAx-Sb|_p$

L2 Subspace Embedding Based on JL Lemma

- Let $r = O(d/\varepsilon^2)$
- S be a $r \times n$ matrix of i.i.d. Gaussian $N(0,1/r)$ random variables
  - Net argument + Johnson-Lindenstrauss Lemma
- Oblivious embedding
- Calculating SA requires $O(nd^2)$ time

CountSketch [CW'13, MM'13, NN'13]

- Let $r=O(d^2/\varepsilon^2)$.
- S has a random sign at a random location in each column
- With constant probability, $|SAx|_2=(1\pm\varepsilon)|Ax|_2$, for all x
- Critical observation
  - A d-dimensional subspace is different from exp(O(d)) arbitrary vectors in R^n
- Calculating SA requires only O(nnz(A)) time
- Lower bound
  - d^2 dependence is tight for L2 OSE with s=1 non-zero entry per column, even just to preserve rank [NN'13]

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

OSNAP [NN'13, BDN'15, Cohen'16]

- Let $r = O(B\, d \log d/\varepsilon^2)$, $s = O(\log_B d/\varepsilon)$
- S has s random signs at random locations in each column
- Lower bound
  - $r = \Omega(B\, d/\varepsilon^2)$ [NN'14]

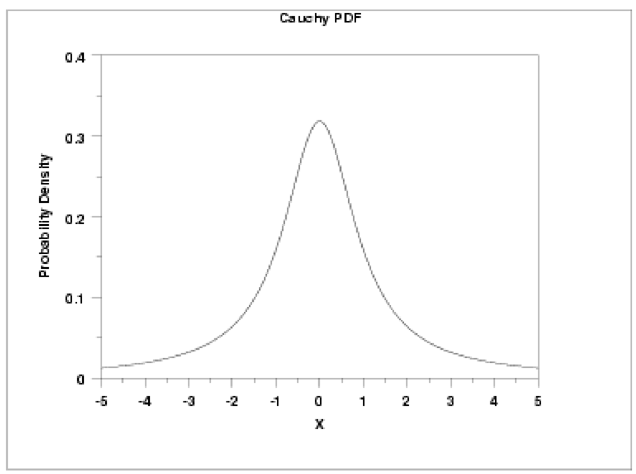
$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

L1 Subspace Embeddings

- Can we use similar constructions for the L1 norm?
- Ingredients of the Gaussian embedding
  - JL Lemma
    - 2-Stability of Gaussian distribution: $a_1G_1 + a_2G_2 + \ldots + a_nG_n \simeq |a|_2G$
    - Concentration bound of χ^2 distribution (sum of squared Gaussians)
  - Net argument for the subspace

1-Stable Distribution: Cauchy Distribution

- 1-Stability: $a_1C_1+a_2C_2+\ldots+a_nC_n \simeq |a|_1C$
- PDF: $f(x) = 1/(\pi(1+x^2))$
- Undefined mean and infinite second moment
- Tail bound: $\Pr[|C|\geq x] = 1-\Theta(1/x)$



Our Plan

- L1-JL Lemma
  - 1-Stability of Cauchy distribution
  - Concentration bound of sum of absolute values of Cauchy’s

- Net argument for the subspace

- Issue: Cauchy distribution is heavy-tailed!

Dense Cauchy Embedding [SW'11]

- Let $r = O(d \log d)$. S be an $r \times n$ matrix of i.i.d. Cauchy random variables
- With constant probability, simultaneously for all x
  - $\Omega(r)|Ax|_1 \leq |SAx|_1 \leq O(rd \log d)|Ax|_1$

- Lower bound part: Net argument + Cauchy lower tail inequality

- Cauchy lower tail inequality
  - Median of absolute value of Cauchy: 1/2
  - A simple Chernoff bound

Dense Cauchy Embedding: Upper Bound

- [Auerbach'30]: Any d-dimensional subspace has a basis U
  - $|U_i|_1 = 1$ for each column U_i of U
  - $|Ux|_1 \geq |x|_\infty$

- Step 1: Show that $|SU|_1$=O(rd log (rd)) with constant probability
- Step 2: $|SUx|_1 \leq |SU|_1 |x|_\infty \leq O(rd \log(rd)) |Ux|_1$

Sparse Cauchy Embedding [MM'13]

- Let $r = O(d^5 \log^5 d)$
- S has a Cauchy at a random location in each column
- Distortion: $\Omega(1/d^2 \log^2 d)|Ax|_1 \leq |SAx|_1 \leq O(d \log d)|Ax|_1$
- Calculating SA requires O(nnz(A)) time

$$\begin{bmatrix} 0 & 0 & C & 0 & 0 & C & 0 & 0 \\ C & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C & C & 0 & C & 0 \\ 0 & C & 0 & 0 & 0 & 0 & 0 & C \end{bmatrix}$$

L1 Subspace Embeddings

- Question 1:
  - Is $\Omega(d\log d)$ distortion optimal for L1 oblivious subspace embeddings?
  - Can we achieve $(1+\varepsilon)$ distortion for L1?
    - This is possible for non-oblivious subspace embeddings (E.g., Lewis weights [CP'15])

- Question 2:
  - Can we have sparse L1 oblivious subspace embeddings with $r = O(d \log d)$ and $O(d \log d)$ distortion?
  - Can we have tradeoff between sparsity and number of rows?

Lower bound for Lp OSE

- For $1 \le p < 2$, any Lp OSE with r rows has distortion

$$\Omega\left(\frac{1}{(1/d)^{1/p}\log^{2/p} r + (r/n)^{1/p-1/2}}\right)$$

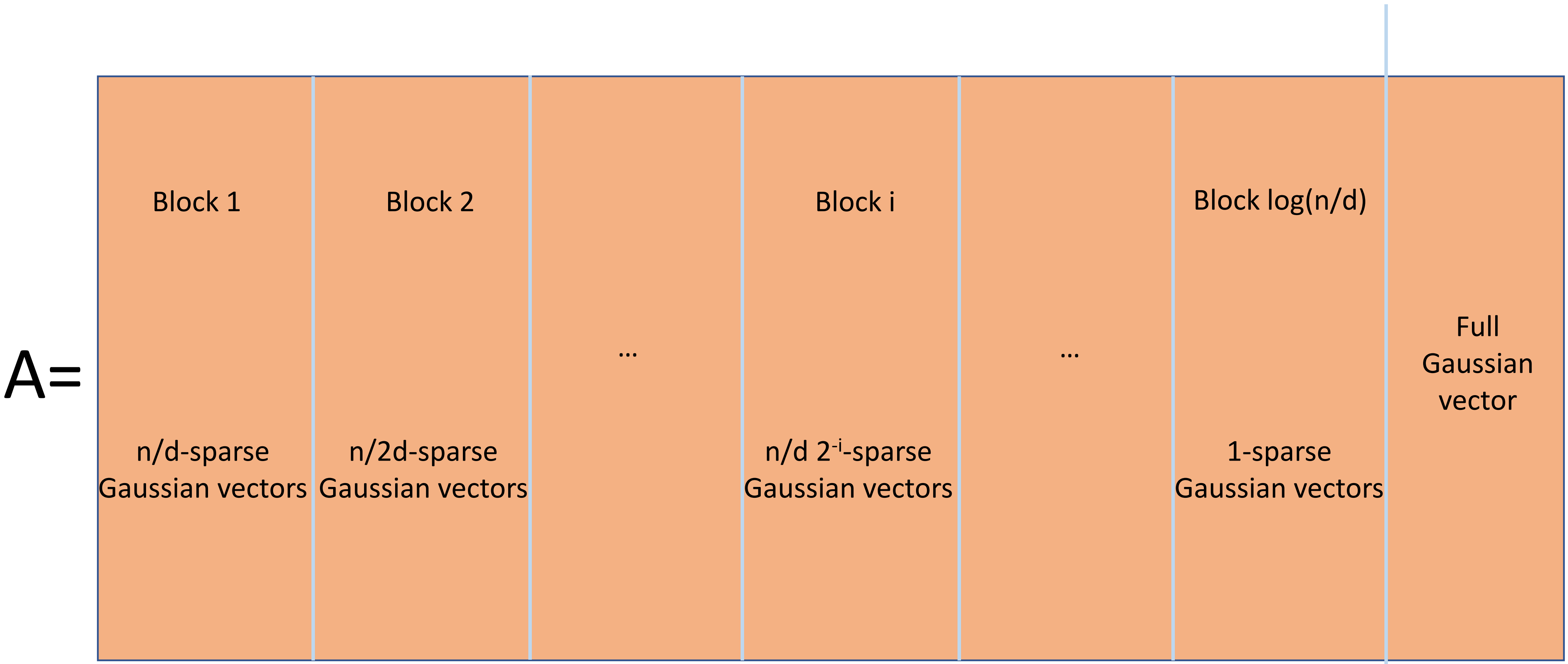
  - When r = n, the identity matrix is an OSE with no distortion
  - As p -> 2, we have OSE with (1+ε) distortion

- For p = 1
  - When r = poly(d), n >> r, the lower bound will be $\Omega(d/\log^2 d)$
  - Dense Cauchy Embedding is optimal up to an $O(\log^3 d)$ factor

The Proof

- Yao's minimax principle
  - Construct a distribution over n×d matrices A
  - Show that for any S in $R^{r \times n}$, the lower bound holds

Construction of the Distribution



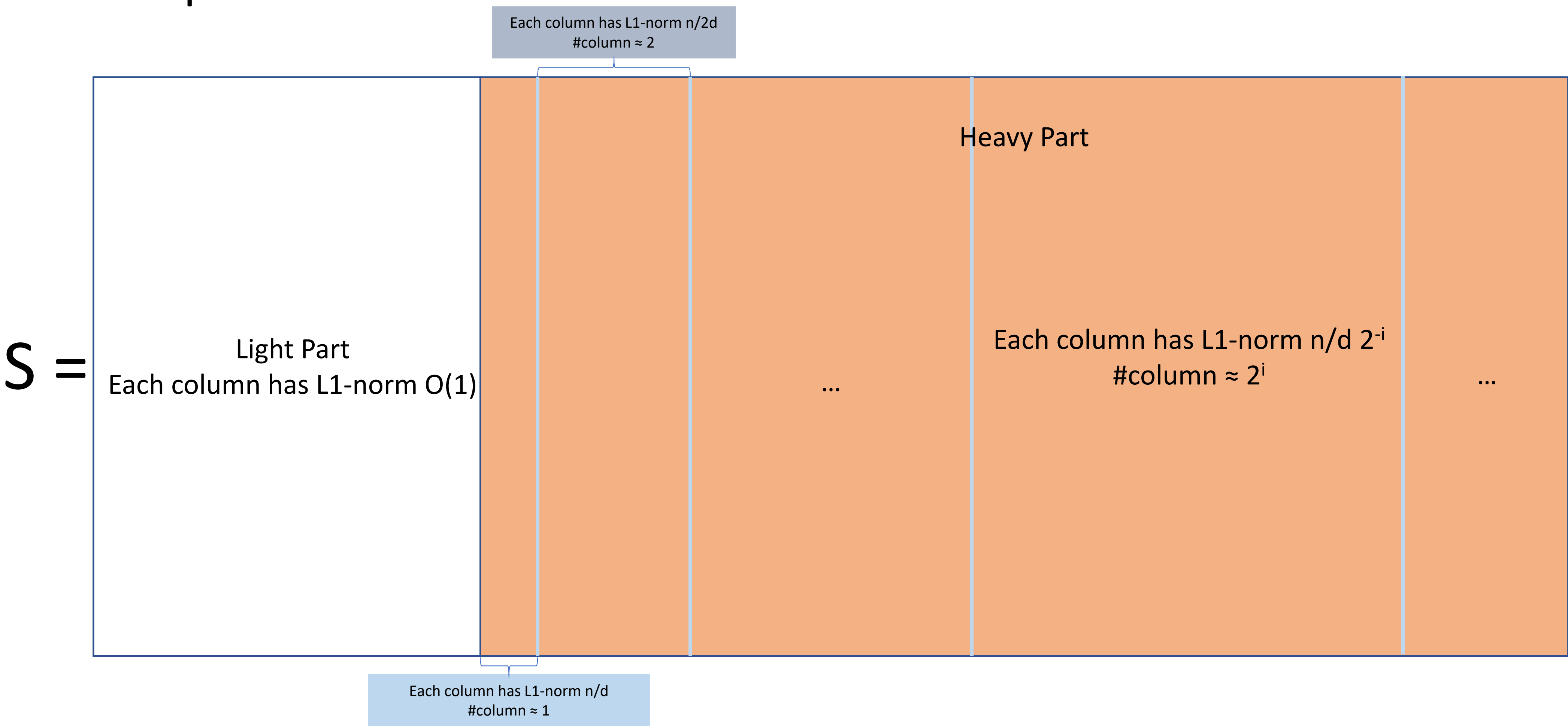
d / log(n/d) vectors in each block

A single full Gaussian vector in the last column

Why does this work?

- An L1 OSE satisfies: $(1 / \kappa)|Ax|_1 \leq |SAx|_1 \leq |Ax|_1$
- Implication of Block i
  - Each vector in Block i has L1-norm $\Theta(n/d\ 2^{-i})$ with good probability
  - If there are more than $O(2^i \text{polylog}(n))$ columns in S with L1-norm $\Omega(n/d\ 2^{-i})$, with good probability, some vector in Block i will find it
  - The condition $|SAx|_1 \leq |Ax|_1$ will be violated
- The histogram of L1-norm of columns looks like a Cauchy!

Implication of the Construction



Implication of the full Gaussian vector

- The last column in A is a full Gaussian vector
  - L1-norm = $\Theta(n)$ with good probability.

- For a full Gaussian vector g,
  - $|Sg|_1 = O(n \text{ polylog}(n) / d)$ by the histogram
  - Distortion = $\Omega(d / \text{polylog}(n))$

Lower Bound

- We have the lower bound:

$$\Omega\left(\frac{1}{(1/d)^{1/p}\log^{2/p} r + (r/n)^{1/p-1/2}}\right)$$

- This implies
  - One cannot use L1 OSE with poly(d) rows to get (1+ε) distortion.
  - It is essential to use non-oblivious subspace embeddings to get (1+ε) distortion
    - E.g., Lewis weight sampling

Lower Bound

- The $\log^{2/p} r$ factor seems possible to improve
- Can we get a lower bound of

$$\Omega\left(\frac{1}{(1/d)^{1/p}+(r/n)^{1/p-1/2}}\right)?$$

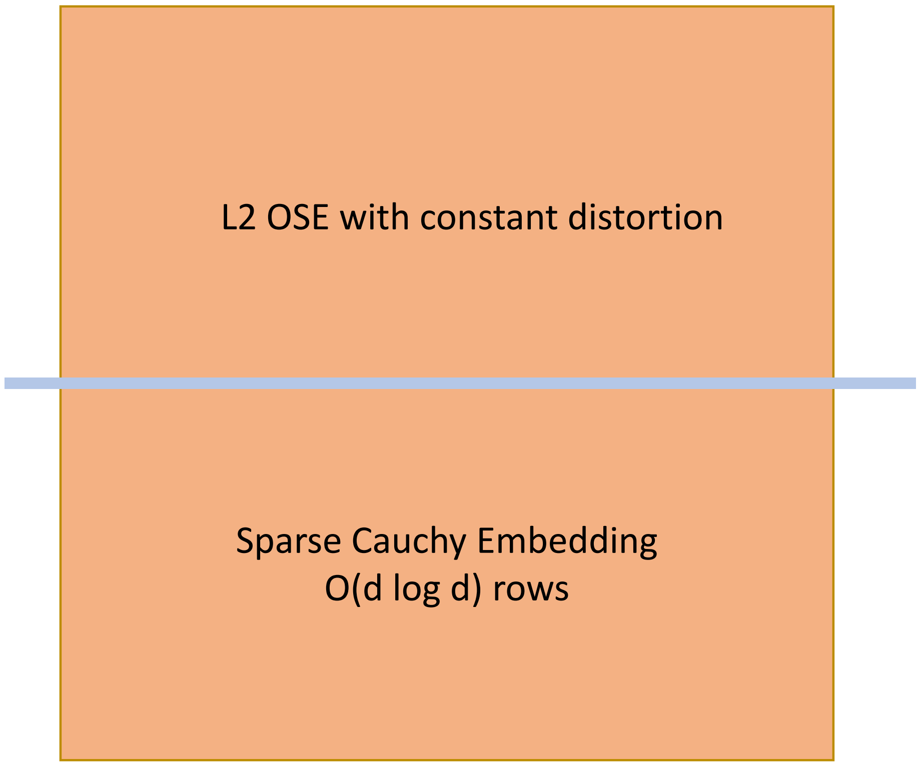
- Theorem: One can construct an L1 OSE with exp(exp(O(d))) rows and O(1) distortion.
- Technique: Standard net argument + better Cauchy tail bounds

L1 Subspace Embeddings

- Question 1:
  - Is $\Omega(d \log d)$ distortion optimal for L1 oblivious subspace embeddings?
  - Can we achieve $(1+\varepsilon)$ distortion for L1?

- Question 2:
  - Can we have sparse L1 oblivious subspace embeddings with $r = O(d \log d)$ and $O(d \log d)$ distortion?
  - Can we have a tradeoff between sparsity and number of rows (like in OSNAP)?

New Sparse L1 OSE

S = 

Distortion = O(d log d)

New Sparse L1 OSE

- Use CountSketch as the L2 OSE
  - $O(d^2)$ rows, sparsity = 2

- Use OSNAP as the L2 OSE
  - $O(B\, d\log d)$ rows, sparsity = $O(\log_B d)$

The Proof

- The upper bound is similar to previous results
  - Auerbach basis + Cauchy upper tail bound

- Let y=Ax. W.l.o.g. we assume $|y|_1=1$.
- If $\sum_{i,\,|y_i|\leq 1/d^2} |y_i| \geq \frac{1}{2}$
  - Sparse Cauchy embedding will be sufficient to prove the lower bound
- Otherwise,
  - $|Sy|_1 \geq |Sy|_2 \geq \Omega(1)|y|_2 \geq \Omega(1)|y|_1/d$

Conclusion

- Nearly optimal distortion lower bound for L1 OSE
- Nearly optimal sparse L1 OSE

Open Questions

- Is is possible to construct an L1 OSE
  - with $O(d^2)$ rows, sparsity = 1 and $O(d \log d)$ distortion?
  - with $O(d \log d)$ rows and sparsity = $O(1)$ and $O(d \log d)$ distortion?
  - with $2^{O(d)}$ rows and $O(1)$ distortion, or prove a stronger lower bound?

- Tight bounds for Lp OSEs for $1 < p < 2$