FAST & SMALL Subspace Embeddings

N. Chepurko — MIT
K. Clarkson — IBM
Praneeth Kacham — CMU
D. Woodruff — CMU

Subspace Embeddings

- Embed a d dimensional subspace V of $\mathbb{R}^n$ into $\mathbb{R}^m$, $m \ll n$

$$x \longrightarrow E(x)$$

- A useful property to preserve is that

$$\text{for all } x \in V, \|E(x)\|_2 = (1 \pm \varepsilon)\|x\|_2$$

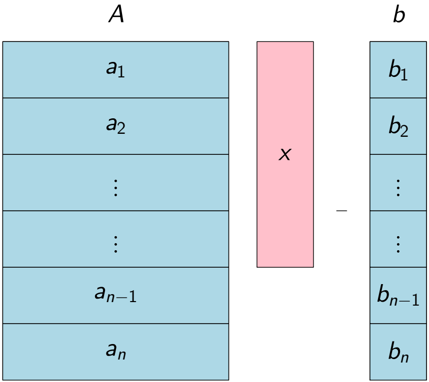
- Ideally, we also want $E(x)$ to be linear : $E(x) = Fx$ for some F
- Can think of it as an analogue of JL Transform for subspaces
- Typically, we are given a matrix $A \in \mathbb{R}^{n \times d}$ and V is defined as

$$V := \{Ax \,|\, x \in \mathbb{R}^d\}$$

An application

- Consider the least squares regression problem:

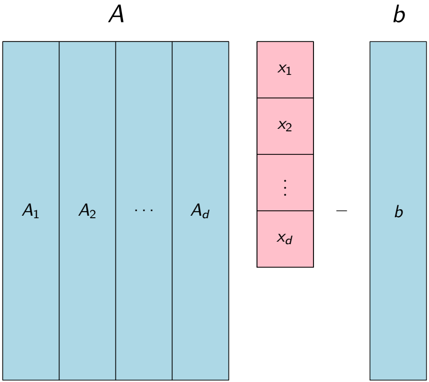
$$\min_x \|A{\color{red}x} - b\|_2^2 = \sum_{i=1}^{n} (\langle a_i, {\color{red}x} \rangle - b_i)^2$$



An application

- Consider the least squares regression problem:

$$\min_{x} \|A{\color{red}x} - b\|_2^2 = \sum_{i=1}^{n} (\langle a_i, {\color{red}x} \rangle - b_i)^2$$

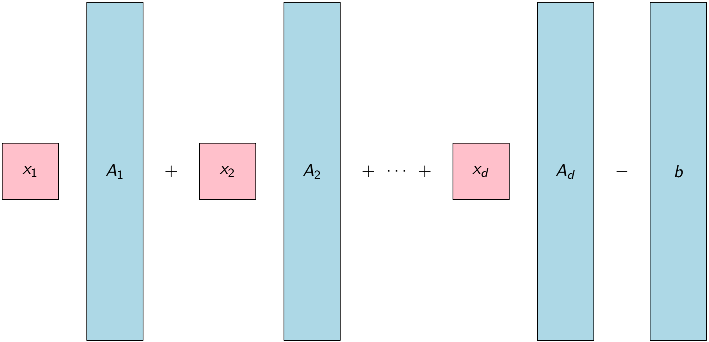


An application

- Consider the least squares regression problem:

$$\min_{x} \|A{\color{red}x} - b\|_2^2 = \sum_{i=1}^{n} (\langle a_i, {\color{red}x} \rangle - b_i)^2$$

$$Ax - b$$



An application

- Consider the least squares regression problem:

$$\min_x \|Ax - b\|_2^2 = \sum_{i=1}^{n} (\langle a_i, x \rangle - b_i)^2$$

- F is a subspace embedding for $\mathrm{colspan}([A\,b]) \implies$

$$\text{for all } x, \quad \|F(Ax - b)\|_2 = (1 \pm \varepsilon)\|Ax - b\|_2$$

- Solution to $\min_x \|FAx - Fb\|_2^2$ is a $1 + O(\varepsilon)$ approximation
- FA is much **smaller** than $A \implies$ solution can be computed quickly

Desirable Properties

We want F to simultaneously have the following properties:

- F itself must be easy to compute
- Should be able to compute FA quickly
- F should have very few rows
- Oblivious

Our transform has all these properties!

Our Result

Theorem

There is a distribution $\mathcal{S}$ over $m \times n$ matrices, $m = d \cdot \mathrm{poly}(\log \log d)$, such that given an arbitrary $n \times d$ matrix A, the random matrix $\mathbf{S} \sim \mathcal{S}$ satisifies the following property with probability $\geq 9/10$:

$$\textit{for all } x, \|Ax\|_2 \leq \|\mathbf{S}Ax\|_2 \leq \exp(\mathrm{poly}(\log \log d))\|Ax\|_2.$$

The matrix $\mathbf{S}A$ can be computed in time $O(\gamma^{-1} nnz(A) + d^{2+\gamma+o(1)})$ for any constant $\gamma > 0$.

Gaussian Embedding

- Net argument: There is a collection $\mathcal{N} \subseteq V$ of unit vectors, $|\mathcal{N}| = 2^{O(d)}$, such that if $\mathbf{G}$ preserves norms of $x \in \mathcal{N}$, then $\mathbf{G}$ preserves norms of all $x \in V$.
- JL Lemma: If $\mathbf{G}$ is $m \times n$ matrix with i.i.d. Gaussian entries, then for arbitrary $x \in \mathbb{R}^n$, with probability $\geq 1 - \delta$,

$$\|\mathbf{G}x\|_2 = (1 \pm \varepsilon)\|x\|_2$$

  if $m = O(\varepsilon^{-2} \log(1/\delta))$
- Can preserve norms of arbitrary $2^{O(d)}$ unit vectors with $m = O(\varepsilon^{-2} d)$

Properties of a Gaussian Subspace Embedding

- Easy to compute
- ~~Should be able to compute~~ $\cancel{FA}$ ~~quickly~~ $O(\mathrm{nnz}(A) \cdot d\varepsilon^{-2})$
- Should have few rows
- **Oblivious** - Don't have to know A

Other Constructions

	# of rows	Time to apply
SRHT	$d\log(d)\varepsilon^{-2}$	$nd\log(n)$
CountSketch	$d^2\varepsilon^{-2}$	$\mathrm{nnz}(A)$
OSNAP	$d^{1+\gamma}\log(d)\varepsilon^{-2}$	$\frac{1}{\gamma\varepsilon}\mathrm{nnz}(A)$
Leverage Score	$d\log(d)\varepsilon^{-2}$	$\mathrm{nnz}(A)+\mathrm{poly}(d)$

- None of these constructions have $o(d\log(d))$ rows for constant ε.
- Composing OSNAP with Gaussian - $O(d)$ rows - $O(\gamma^{-1}\mathrm{nnz}(A) + d^{2+\gamma+o(1)} + d^{\omega}\log(d))$ time.

$$n \underset{\substack{\text{OSNAP } \gamma \\ O(\gamma^{-1}\,\mathrm{nnz}(A))}}{\Longrightarrow} d^{1+\gamma}\log(d) \underset{\substack{\text{OSNAP } \gamma=\frac{1}{\log d} \\ O(d^{2+\gamma}\log^2 d)}}{\Longrightarrow} d\log(d) \underset{\substack{\text{Gaussian} \\ O(d^{\omega}\log d)}}{\Longrightarrow} d$$

Applications to Other Problems

Using our construction of subspace embeddings and a few other ideas, we obtain near-optimal running times for other problems

Application	Running time (up to constant factors)
ε Subspace Embeddings	$\mathrm{nnz}(A) + \varepsilon^{-3} d^{2.1+o(1)} + d^{\omega}\,\mathrm{poly}(\log\log(d))$
ε approximate linear regression	$\mathrm{nnz}(A) + \varepsilon^{-3} d^{2.1+o(1)} + d^{\omega}\,\mathrm{poly}(\log\log(d))$
Linearly Independent Rows	$\mathrm{nnz}(A) + k^{\omega}\,\mathrm{poly}(\log\log(k)) + k^{2+o(1)}$
0.01 Rank k Approximation	$\mathrm{nnz}(A) + (n+d)k^{\omega-1}$

What are we trying to construct?

A random matrix **S** such that:

- **S** has $o(d \log(d))$ rows
- For any matrix A, the matrix $\mathbf{S}A$ can be computed in time $O(\mathrm{nnz}(A) + d^c)$ for some $c < \omega$
- With probability $\geq 9/10$, for all vectors x,

$$\|Ax\|_2 \leq \|\mathbf{S}Ax\|_2 \leq \alpha \|Ax\|_2$$

with small α

Our Approach

- We go back to Gaussians and see how sparse we can make the Gaussian matrix
- For some *special* subspaces, we can set many entries of the Gaussian matrix to be 0
- Sparse Matrix $\rightarrow$ Fast Multiplication!
- Applying some embeddings, we can assume without loss of generality that A is a $d \log d \times d$ matrix

Idea

- Suppose $\mathbf{S}$ is a matrix that randomly samples d coordinates of $d\log(d)$ dimensional vector x. How large is $\|\mathbf{S}x\|_2$?
  1. If $x = e_i$: With probability $1 - 1/\log(d)$, $\|\mathbf{S}x\|_2 = 0$:(
  2. If $x = 1/\sqrt{d\log(d)}$: With probability 1, $\|\mathbf{S}x\|_2 = 1/\sqrt{\log(d)}$:)
- Having a "large" number of "large" coordinates helps in making the sketching matrix $\mathbf{S}$ sparse
- Unit vectors x that are "sketchable" by sparse matrices have $\|x\|_1 = \Omega(\sqrt{d})$

Contraction

- Consider a unit vector $x \in \mathbb{R}^{d\log(d)}$ with the property that

$$\#i \text{ such that } |x_i| \geq \tilde{\Omega}(1/\sqrt{d}) \text{ is at least } cd$$

- Consider a random matrix M with each entry 0 with probability $1 - p$ and ± 1 with probability $p/2$ each
- We want to show $\|Mx\|_2 \geq \tilde{O}(1)$ with **very high** probability
- If $p = \Theta(1/d)$, what's the probability that the 1st row hits a heavy coordinate of x?

Contraction

- Consider a unit vector $x \in \mathbb{R}^{d\log(d)}$ with the property that

$$\#i \text{ such that } |x_i| \geq \tilde{\Omega}(1/\sqrt{d}) \text{ is at least } cd$$

- Consider a random matrix M with each entry 0 with probability $1-p$ and ± 1 with probability $p/2$ each
- We want to show $\|Mx\|_2 \geq \tilde{O}(1)$ with **very high** probability
- If $p = \Theta(1/d)$, what's the probability that the 1st row hits a heavy coordinate of x? $\Theta(1)$

Contraction

- Consider a unit vector $x \in \mathbb{R}^{d\log(d)}$ with the property that

$$\#i \text{ such that } |x_i| \geq \tilde{\Omega}(1/\sqrt{d}) \text{ is at least } cd$$

- Consider a random matrix M with each entry 0 with probability $1 - p$ and ± 1 with probability $p/2$ each
- We want to show $\|Mx\|_2 \geq \tilde{O}(1)$ with **very high** probability
- If $p = \Theta(1/d)$, what's the probability that the 1st row hits a heavy coordinate of x? $\Theta(1)$
- Given that 1st row hits x, how large will $|M_{1*}x|^2$ be?

Contraction - Continued

- Consider the random sum

$$\mathbf{r}_1 x_1 + \mathbf{r}_2 x_2 + \ldots + \mathbf{r}_n x_n$$

where $\mathbf{r}_i = \pm 1$ with probability $1/2$ each independently. Also assume that $|x_n| \geq |x_i|$.

- Fix a value for $\mathbf{r}_n$. With $1/2$ probability over $\mathbf{r}_1, \ldots, \mathbf{r}_{n-1}$, $\mathbf{r}_1 x_1 + \mathbf{r}_2 x_2 + \ldots + \mathbf{r}_{n-1} x_{n-1}$ has the same sign as $\mathbf{r}_n x_n$.
- So $|\mathbf{r}_1 x_1 + \ldots + \mathbf{r}_n x_n| \geq |\mathbf{r}_n x_n| \geq |x_n|$ with probability $1/2$
- Under the event that the first row hits a heavy coordinate of x, it contributes $\tilde{\Omega}(1/\sqrt{d})$ with probability $1/2$!
- So, with constant probability,

$$|M_{1*}x|^2 \geq \tilde{\Omega}(1/d)$$

Contraction - Continued

- Let row i be **large** if $|M_{i*}x|^2 \geq \tilde{\Omega}(1/d)$
- So, we have $\Pr[i \text{ is } \textbf{large}] = \Theta(1)$ from the previous argument
- If $m = Cd$: Chernoff bound $\implies$ with prob. $1 - \exp(-\Theta(d))$, there are $\Theta(d)$ **large** rows
- $\Theta(d)$ **large** rows $\implies$ $\|Mx\|_2^2 \geq \tilde{\Theta}(1)$

Dilation

- We also want to show that $\|Mx\|_2^2 \leq \tilde{\Theta}(1)$
- This is easy as row sums and column sums of M are $\tilde{O}(1)$ with high probability

What did we learn?

- If each unit vector of the subspace is *flat*, then the subspace can be embedded using a Sparse Sign matrix i.e.,

$$\text{for all}\, x \in V : \ \|x\|_2 = 1 \implies \|x\|_1 = \tilde{\Omega}(\sqrt{d})$$

- How do we transform any given subspace into one that has this property?
- We use a transformation in a paper of Indyk and prove new properties

New Theorem

Theorem

For any arbitrary n, let $m = n^{1+o(1)}$. Let $B_1, \ldots, B_b$ be a partition of m with $b \approx \sqrt{n}$. Then there is a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ that has the following properties:

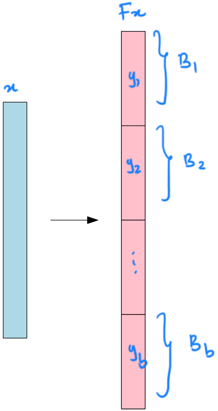
1. *For any x, Fx can be computed in $n^{1+o(1)}$ time*
2. *For any vector x,*

$$(1 - \frac{1}{100 \log\log n})\frac{1}{b}\|x\|_2^2 \le \|Fx\|_2^2 \le \frac{1}{b}\|x\|_2^2 \tag{1}$$

3. *For any vector x,*

$$(1 - \frac{1}{100 \log\log n})\|x\|_2 \le \sum_{i=1}^{b} \|(Fx)_{B_j}\|_2 \le \|x\|_2 \tag{2}$$

Indyk Embedding



$$\sum_i \|y_i\|_2^2 \approx \frac{1}{b}\|x\|_2^2$$

$$\sum_i \|y_i\|_2 \approx \|x\|_2.$$

Wrap-up

- Recursively apply the previous transform for $\Theta(\log \log n)$ times
- We end up with a transformation $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$, $m = n^{1+o(1)}$, such that for any unit vector x,

$$\|\mathcal{F}x\|_1 \geq \sqrt{n/4}$$
$$1/2 \leq \|\mathcal{F}x\|_2^2 \leq 1$$

- This means cn coordinates of $\mathcal{F}x$ have a value at least $1/n^{o(1)}$ if $\|\mathcal{F}x\|_2 = 1$
- Not as good as the property we assumed but this is enough for the construction of subspace embeddings