

CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2024

Lecture 8: Sketching, types of sketching matrices

Outline

- 1 Gaussian sketching
- 2 AMM and JL moment
- 3 SRHT
- 4 Countsketch

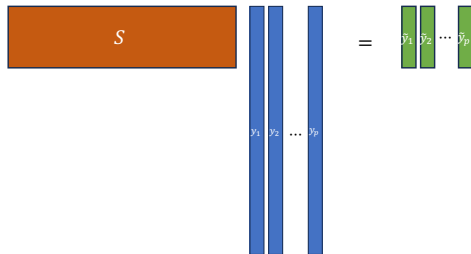
Recall: Embeddings

Embedding

A matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ is an ϵ -embedding of set $\mathcal{P} \subset \mathbb{R}^n$ if, for all $\mathbf{y} \in \mathcal{P}$,

$$\|\mathbf{S}\mathbf{y}\|_2 = (1 \pm \epsilon)\|\mathbf{y}\|_2.$$

We will call \mathbf{S} a “sketching matrix”.



Gaussian sketching matrix

Vector embedding also known as Distributional JL Lemma.

Distributional JL

Let $\mathbf{S} \in \mathbb{R}^{m \times d}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$. If $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any vector $\mathbf{y} \in \mathbb{R}^d$, $\epsilon \in (0, 1]$, with probability $(1 - \delta)$:

$$\|\mathbf{S}\mathbf{y}\|_2^2 = (1 \pm \epsilon)\|\mathbf{y}\|_2^2.$$

Gaussian sketching matrix

Vector embedding also known as Distributional JL Lemma.

Distributional JL

Let $\mathbf{S} \in \mathbb{R}^{m \times d}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$. If $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any vector $\mathbf{y} \in \mathbb{R}^d$, $\epsilon \in (0, 1]$, with probability $(1 - \delta)$:

$$\|\mathbf{S}\mathbf{y}\|_2^2 = (1 \pm \epsilon)\|\mathbf{y}\|_2^2.$$

Proof: We know that $\mathbb{E}[\|\mathbf{S}\mathbf{y}\|_2^2] = \|\mathbf{y}\|_2^2$. We have

$$\|\mathbf{S}\mathbf{y}\|_2^2 = \frac{1}{m} \sum_{i=1}^m (\langle \mathbf{s}_i, \mathbf{y} \rangle)^2 = \frac{1}{m} \sum_{i=1}^m (\mathcal{N}(0, \|\mathbf{y}\|_2^2))^2$$

Chi-squared random variable with m degrees of freedom.

Chi-squared random variable

Let z be a Chi-squared random variable with m degrees of freedom

$$\Pr\{|z - \mathbb{E}[z]| \geq \epsilon \mathbb{E}[z]\} \leq 2 \exp(-\epsilon^2 m/8).$$

We have $\mathbb{E}[z] = \mathbb{E}[\|\mathbf{S}\mathbf{y}\|_2^2] = \|\mathbf{y}\|_2^2$.

So, setting $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, we obtain the result.

Gaussian - JL property

JL Lemma

Let $\mathbf{S} \in \mathbb{R}^{m \times d}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$. If $m = O\left(\frac{\log(n)}{\epsilon^2}\right)$, then for any set of n data points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, with probability at least 9/10:

$$(1 - \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{x}_j\|_2 \leq (1 + \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

Gaussian - JL property

JL Lemma

Let $\mathbf{S} \in \mathbb{R}^{m \times d}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$. If $m = O\left(\frac{\log(n)}{\epsilon^2}\right)$, then for any set of n data points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, with probability at least 9/10:

$$(1 - \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{x}_j\|_2 \leq (1 + \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

Proof: Fix $i, j \in [d]$, let $\mathbf{y} = \mathbf{x}_i - \mathbf{x}_j$. By the Distributional JL Lemma, with probability $1 - \delta$:

$$\|\mathbf{S}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 = (1 \pm \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2.$$

Set $\delta = 1/n^2$. Since there are $< n^2$ total i, j pairs, by a union bound we have that with probability 9/10, the above will hold for *all* i, j , for:

$$m = O\left(\frac{\log(n)}{\epsilon^2}\right).$$

Gaussian - Subspace embedding

Subspace embedding

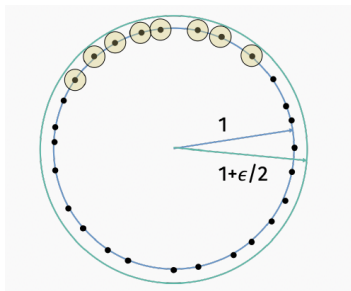
Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$. If $m = O\left(\frac{d \log(1/\delta)}{\epsilon^2}\right)$, then for a given $\mathbf{A} \in \mathbb{R}^{n \times d}$, with probability at least $1 - \delta$:

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{x}\|_2.$$

Embedding a d -dimensional subspace $\mathcal{U} \equiv \text{span}(\mathbf{A}) = \text{span}(\mathbf{U}) \subset \mathbb{R}^n$.

$$\|\mathbf{S}\mathbf{U}\mathbf{x}\|_2 = (1 \pm \epsilon)\|\mathbf{x}\|_2 \quad \text{or} \quad \|\mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U} - \mathbf{I}\|_2 \leq \epsilon$$

Recall the ϵ -Net argument.



We know $|\mathcal{N}(\epsilon)| \leq (1 + \frac{2}{\epsilon})^d$.

If \mathbf{S} is distributional JL with failure probability δ' , taking union of the ϵ -net size, we get the result, with

$$m = O\left(\frac{d \log(1/\delta)}{\epsilon^2}\right).$$

AMM to embedding

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $n \geq d$, $\text{rank}(\mathbf{A}) = r$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$. Let \mathbf{S} be chosen (with oblivious distribution) such that, with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq \epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F.$$

Then, \mathbf{S} is an $\epsilon * r$ -embedding of $\text{span}(\mathbf{A})$.

AMM to embedding

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $n \geq d$, $\text{rank}(\mathbf{A}) = r$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$. Let \mathbf{S} be chosen (with oblivious distribution) such that, with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq \epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F.$$

Then, \mathbf{S} is an $\epsilon * r$ -embedding of $\text{span}(\mathbf{A})$.

Set $\mathbf{B} = \mathbf{A}^\top$, and since \mathbf{S} is oblivious, let us assume \mathbf{A} is orthonormal. Then,

$$\|\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{I}\|_2 \leq \|\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{I}\|_F \leq \epsilon \|\mathbf{A}\|_F^2 = \epsilon r.$$

JL moment property

JL moment

A distribution on $\mathbf{S} \in \mathbb{R}^{m \times d}$, has the (ϵ, δ, ℓ) -JL moment property if for all $\mathbf{y} \in \mathbb{R}^d$ with $\|\mathbf{y}\|_2 = 1$,

$$\mathbb{E}_{\mathbf{S}}[|||\mathbf{S}\mathbf{y}\|_2^2 - 1|^\ell] \leq \epsilon^\ell \cdot \delta.$$

JL moment property

JL moment

A distribution on $\mathbf{S} \in \mathbb{R}^{m \times d}$, has the (ϵ, δ, ℓ) -JL moment property if for all $\mathbf{y} \in \mathbb{R}^d$ with $\|\mathbf{y}\|_2 = 1$,

$$\mathbb{E}_{\mathbf{S}}[|\|\mathbf{S}\mathbf{y}\|_2^2 - 1|^\ell] \leq \epsilon^\ell \cdot \delta.$$

For $\ell = 2$, and if $\mathbb{E}[\|\mathbf{S}\mathbf{y}\|_2] = 1$ we have

$$\text{Var}(\|\mathbf{S}\mathbf{y}\|_2^2) \leq \epsilon^2 \delta \quad \text{or} \quad \text{sd}(\|\mathbf{S}\mathbf{y}\|_2^2) \leq \epsilon \sqrt{\delta}.$$

JL moment and AMM

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$. Let \mathbf{S} satisfy the (ϵ, δ, ℓ) -JL moment property for $\ell \geq 2$, then with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq 3\epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F.$$

JL moment and AMM

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$. Let \mathbf{S} satisfy the (ϵ, δ, ℓ) -JL moment property for $\ell \geq 2$, then with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq 3\epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F.$$

Proof: We follow proof of Theorem 2.8 in Dr. Woodruff's monograph.

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\frac{\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} =$$

JL moment and AMM

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$. Let \mathbf{S} satisfy the (ϵ, δ, ℓ) -JL moment property for $\ell \geq 2$, then with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq 3\epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F.$$

Proof: We follow proof of Theorem 2.8 in Dr. Woodruff's monograph.

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\frac{\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} =$$

Minkowski's inequality : $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$.

For unit vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\|\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle\|_\ell =$$

For unit vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\|\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle\|_\ell =$$

Define a random variable

$$X_{ij} = \frac{1}{\|\mathbf{B}_{i*}\|_2 \|\mathbf{A}_{*j}\|_2} (\langle \mathbf{S}\mathbf{B}_{i*}, \mathbf{S}\mathbf{A}_{*j} \rangle - \langle \mathbf{B}_{i*}, \mathbf{A}_{*j} \rangle)$$

Then,

$$\|\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F^2\|_{\ell/2} =$$

For unit vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\|\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle\|_\ell =$$

Define a random variable

$$X_{ij} = \frac{1}{\|\mathbf{B}_{i*}\|_2 \|\mathbf{A}_{*j}\|_2} (\langle \mathbf{S}\mathbf{B}_{i*}, \mathbf{S}\mathbf{A}_{*j} \rangle - \langle \mathbf{B}_{i*}, \mathbf{A}_{*j} \rangle)$$

Then,

$$\|\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F^2\|_{\ell/2} =$$

Using

$$\mathbb{E}\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F^\ell = \|\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F^2\|_{\ell/2}^{\ell/2},$$

and Markov's inequality we get the result.

Gaussian sketch and AMM

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$.

Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$ and $m = O(\epsilon^{-2}\delta^{-1})$, then with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top\mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq \epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F.$$

Gaussian sketch and AMM

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$.

Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$ and $m = O(\epsilon^{-2}\delta^{-1})$, then with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top \mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq \epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F.$$

For Gaussian sketch, with $\ell = 2$, JL moment is

$$\text{Var}(\|\mathbf{S}\mathbf{y}\|_2^2) \leq 2/m.$$

Since $\text{Var}(\frac{1}{m}\chi_m^2) = \frac{1}{m^2}\text{Var}(\chi_m^2) = 2m/m^2 = 2/m$.

We set $2/m \leq \epsilon^2\delta/6$.

Gaussian sketch and AMM

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$; $\mathbf{B} \in \mathbb{R}^{d' \times n}$; $\epsilon, \delta > 0$.

Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ have independent entries $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0, 1)$ and $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then with probability at least $1 - \delta$:

$$\|\mathbf{B}\mathbf{S}^\top\mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq \epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F.$$

Consider $\ell = \Theta(\log(1/\delta))$. Then, the ℓ -th central moment of χ_m^2 is of the form $2^\ell(c_1m^{\ell/2} + c_2)$. So, if we choose $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, we have:

$$\frac{2^\ell}{m^{\ell/2}} = \epsilon^\ell 2^{\ell/2} (2/\ell)^{\ell/2} \leq \epsilon^\ell \delta.$$

Two approaches

We have seen two approaches to go from vector embeddings to subspace embeddings.

Let $\|\mathbf{U}\|_F^2 = d, \text{rank}(\mathbf{A})$.

- Using ϵ -nets:

$$\begin{aligned}\Pr[\|\mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U} - \mathbf{I}\|_{\mathcal{N}_\epsilon} \geq \epsilon] &\leq C^d e^{-m\epsilon^2} \\ \implies \Pr[\|\mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U} - \mathbf{I}\|_2 \geq 2\epsilon] &\leq C^d e^{-m\epsilon^2}\end{aligned}$$

- using JL moment:

$$\begin{aligned}\left\| \frac{1}{d} \|\mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U} - \mathbf{I}\|_2 \right\|_\ell^\ell &\leq \epsilon^\ell \delta \\ \implies \Pr\left[\frac{1}{d} \|\mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U} - \mathbf{I}\|_2 \geq \epsilon\right] &\leq \delta\end{aligned}$$

SHRT: Subsampled Randomized Hadamard Transform

Original JL:

- \mathbf{S} is picked to be random matrix (orthogonal columns), i.i.d entries.
- Computing \mathbf{SA} takes $O(mnd)$ time.

Faster scheme: pick a random orthogonal matrix, but:

- fewer random bits.
- faster to apply.

Fast JL: Using Subsampled Randomized Hadamard Transform (SRHT)

The SRHT is a matrix \mathbf{PHD} , where

- $\mathbf{D} \in \mathbb{R}^{n \times n}$ is diagonal matrix with i.i.d ± 1 on diagonal
- $\mathbf{H} \in \mathbb{R}^{n \times n}$ is a Hadamard matrix
- $\mathbf{P} \in \mathbb{R}^{m \times n}$ uniformly samples the rows of \mathbf{HD} .

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & -1 & 1 & \cdots & -1 \\ 1 & 1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & -1 & \cdots & 1 \end{bmatrix}}_{\sqrt{n}\mathbf{H}} \underbrace{\begin{bmatrix} \pm 1 & 0 & 0 & \cdots & 0 \\ 0 & \pm 1 & 0 & \cdots & 0 \\ 0 & 0 & \pm 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \pm 1 \end{bmatrix}}_{\mathbf{D}}$$

Hadamard matrices

Hadamard matrices have recursive structure.

- Let $\mathbf{H}_0 \in \mathbb{R}^{1 \times 1}$ be $[1]$.
- Let $\mathbf{H}_{i+1} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{H}_i & \mathbf{H}_i \\ \mathbf{H}_i & -\mathbf{H}_i \end{bmatrix}$ for $i \geq 0$.

So,

$$\mathbf{H}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \mathbf{H}_2 = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

In general, \mathbf{H}_k is $2^k \times 2^k$ matrix with ± 1 entries scaled by $1/2^{k/2}$.

Hadamard properties

- Hadamard matrices are orthogonal.

$$\mathbf{H}_i^\top \mathbf{H}_i = \mathbf{H}_i^2 = \mathbf{I}.$$

- For any $\mathbf{x} \in \mathbb{R}^n, n = 2^k$, we have $\|\mathbf{H}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$, also $\|\mathbf{H}\mathbf{D}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$.
- Matvecs $\mathbf{H}\mathbf{x}$ can be computed in $O(n \log n)$ time for $\mathbf{x} \in \mathbb{R}^n, n = 2^k$.

Suppose $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \in \mathbb{R}^{2^k}$, where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{2^{k-1}}$.

Then, $\mathbf{H}_{i+1}\mathbf{x} = \begin{bmatrix} \mathbf{H}_i\mathbf{x}_1 + \mathbf{H}_i\mathbf{x}_2 \\ \mathbf{H}_i\mathbf{x}_1 - \mathbf{H}_i\mathbf{x}_2 \end{bmatrix}$.

So, we can compute $\mathbf{H}_{i+1}\mathbf{x}$ in linear time from $\mathbf{H}_i\mathbf{x}_1, \mathbf{H}_i\mathbf{x}_2$.

Randomized Hadamard analysis

SHRT mixing lemma

Let \mathbf{H} be an $(n \times n)$ Hadamard matrix and \mathbf{D} a random ± 1 diagonal matrix. Let $\mathbf{z} = \mathbf{H}\mathbf{D}\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$. With probability $1 - \delta$, for all i simultaneously,

$$z_i^2 \leq \frac{c \log(n/\delta)}{n} \|\mathbf{z}\|_2^2.$$

for some fixed constant c .

The vector is very close to uniform with high probability.

$$\|\mathbf{z}\|_2^2 = \|\mathbf{H}\mathbf{D}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2.$$

Randomized Hadamard analysis

z_i is a random variable with mean 0 and variance $\|\mathbf{x}\|_2^2/n$, which is a sum of independent random variables.

Can apply Bernstein type concentration inequality to prove the bound:

Rademacher Concentration

Let r_1, \dots, r_n be Rademacher random variables (i.e. uniform ± 1 's). Then for any vector $\mathbf{a} \in \mathbb{R}^n$,

$$\Pr \left[\sum_{i=1}^n r_i a_i \geq t \|\mathbf{a}\|_2 \right] \leq e^{-t^2/2}$$

$z_i = \mathbf{h}_i^\top \mathbf{D} \mathbf{x}$ and let $\mathbf{h}_i^\top \mathbf{D} = \frac{1}{\sqrt{n}}[r_1, r_2, \dots, r_n]$, where r_i 's are random ± 1 's.

$t = \sqrt{\log(n/\delta)}$ and apply union bounds over all n entries.

The Fast JL Lemma

Let $\mathbf{S} = \mathbf{P}\mathbf{H}\mathbf{D} \in \mathbb{R}^{m \times n}$ be a subsampled randomized Hadamard transform with $m = O\left(\frac{\log(n/\delta) \log(1/\delta)}{\epsilon^2}\right)$. Then for any fixed $\mathbf{x} \in \mathbb{R}^n$. With probability $1 - \delta$,

$$\|\mathbf{S}\mathbf{x}\|_2^2 = (1 \pm \epsilon)\|\mathbf{x}\|_2^2.$$

Proof: Apply Hoeffding's inequality for the sum of m entries.

SRHT embeddings

SRHT - subspace embedding

For $\mathbf{S} = \mathbf{P}\mathbf{H}\mathbf{D} \in \mathbb{R}^{m \times n}$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$, if $m = O\left(\frac{d \log(n/\delta) \log(1/\delta)}{\epsilon^2}\right)$, then with probability at least $1 - \delta$:

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{x}\|_2.$$

We can compute the sketch $\mathbf{S}\mathbf{A}$ in $O(mn \log(d))$ time.

Faster Embeddings: Countsketch

- Gaussian sketching matrix is good, but is expensive to apply.
- SRHT is faster, but for dense matrices.
- **Sparse Embeddings:** Adaption of CountSketch from streaming algorithms.
- S is of the form:

$$\begin{bmatrix} 0 & -1 & 0 & 0 & \dots & 0 \\ +1 & 0 & 0 & +1 & \dots & 0 \\ 0 & 0 & -1 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \dots & -1 \end{bmatrix}$$

- One random ± 1 per column.
- Row A_{i*} of A contributes $\pm A_{i*}$ to one of the rows of SA .

Sparse Embeddings

- **Sparse sketching matrix:** For $i \in [n]$, pick uniformly and independently: $h_i \in [m]$, $s_i \in \{-1, +1\}$, and define $\mathbf{S} \in \mathbb{R}^{m \times n}$ as:

$$\mathbf{S}_{h_i, i} \rightarrow s_i \text{ for } i \in [n],$$

and $\mathbf{S}_{j, i} \rightarrow 0$ otherwise.

- \mathbf{s} is a sign (Radamacher) vector. The vector \mathbf{h} hashes to m “hash buckets”. That is,

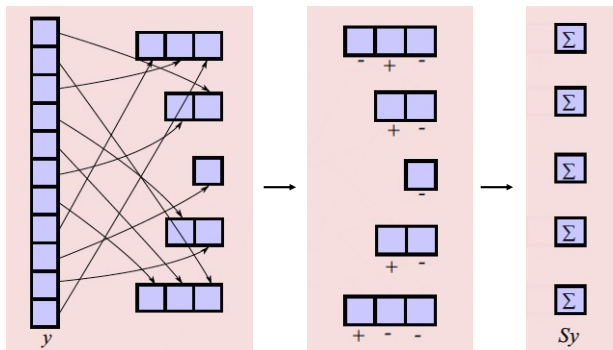
$$\mathbf{S}_{j*} = \sum_{i: h_i=j} s_i \mathbf{e}_i^\top,$$

and so

$$[\mathbf{S}\mathbf{A}]_{j*} = \sum_{i: h_i=j} s_i \mathbf{e}_i^\top \mathbf{A} = \sum_{i: h_i=j} s_i \mathbf{A}_{i*}.$$

- **Fast sketching:** Can compute $\mathbf{S}\mathbf{A}$ in $O(nnz(\mathbf{A}))$ time.

- If \mathbf{s} is a sign (Radamacher) vector, then $\mathbb{E}[\mathbf{s}^\top \mathbf{y}] = \|\mathbf{y}\|_2^2$.
- For $\mathbf{y} = \mathbf{A}\mathbf{x}$, each row of \mathbf{S} :
 - (a) collects a subset of entries y_i 's; (b) applies the signs, and (c) adds
- $\mathbb{E}[\|\mathbf{S}\mathbf{y}\|_2^2] = \|\mathbf{y}\|_2^2$.



Analysis of sparse embeddings

Variance of Countsketch

For $\mathbf{S} \in \mathbb{R}^{m \times n}$ a sparse sketching distribution, and $\mathbf{y} \in \mathbb{R}^n$ a unit vector,

$$\text{Var}[\|\mathbf{S}\mathbf{y}\|_2^2] \leq \frac{3}{m}.$$

Analysis of sparse embeddings

Variance of Countsketch

For $\mathbf{S} \in \mathbb{R}^{m \times n}$ a sparse sketching distribution, and $\mathbf{y} \in \mathbb{R}^n$ a unit vector,

$$\text{Var}[\|\mathbf{S}\mathbf{y}\|_2^2] \leq \frac{3}{m}.$$

Proof: Let $\mathbf{z} = \mathbf{S}\mathbf{y}$. We have $\mathbb{E}[\|\mathbf{z}\|_2^2] =$

$$\text{Var}[\|\mathbf{z}\|_2^2] =$$

Analysis of sparse embeddings

Variance of Countsketch

For $\mathbf{S} \in \mathbb{R}^{m \times n}$ a sparse sketching distribution, and $\mathbf{y} \in \mathbb{R}^n$ a unit vector,

$$\text{Var}[\|\mathbf{S}\mathbf{y}\|_2^2] \leq \frac{3}{m}.$$

Proof: Let $\mathbf{z} = \mathbf{S}\mathbf{y}$. We have $\mathbb{E}[\|\mathbf{z}\|_2^2] =$

$$\text{Var}[\|\mathbf{z}\|_2^2] =$$

$$\mathbb{E}[\|\mathbf{z}\|_2^4] =$$

Analysis of sparse embeddings

Variance of Countsketch

For $\mathbf{S} \in \mathbb{R}^{m \times n}$ a sparse sketching distribution, and $\mathbf{y} \in \mathbb{R}^n$ a unit vector,

$$\text{Var}[\|\mathbf{S}\mathbf{y}\|_2^2] \leq \frac{3}{m}.$$

Proof: Let $\mathbf{z} = \mathbf{S}\mathbf{y}$. We have $\mathbb{E}[\|\mathbf{z}\|_2^2] =$

$$\text{Var}[\|\mathbf{z}\|_2^2] =$$

$$\mathbb{E}[\|\mathbf{z}\|_2^4] =$$

$$\mathbb{E}_{s,h}[z_j^4] =$$

Countsketch Embedding

Countsketch - subspace embedding

For $\mathbf{S} \in \mathbb{R}^{m \times n}$ a countsketch matrix and $\mathbf{A} \in \mathbb{R}^{n \times d}$, if $m = O\left(\frac{d^2}{\delta \epsilon^2}\right)$, then with probability at least $1 - \delta$:

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{x}\|_2.$$

Countsketch Embedding

Countsketch - subspace embedding

For $\mathbf{S} \in \mathbb{R}^{m \times n}$ a countsketch matrix and $\mathbf{A} \in \mathbb{R}^{n \times d}$, if $m = O\left(\frac{d^2}{\delta \epsilon^2}\right)$, then with probability at least $1 - \delta$:

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{x}\|_2.$$

We use the AMM and JL moment result.

We have $\text{Var}[\|\mathbf{S}\mathbf{y}\|_2^2] \leq \frac{K}{m}$.

If $\frac{K}{m} \leq \epsilon^2 \delta$, we know \mathbf{S} is ϵd -embedding with probability at least $1 - \delta$.

Types of sketching matrices

Sketching matrix	Sketch size m	Cost to sketch \mathbf{SA}
JL - i.i.d subGaussians	$m = O\left(\frac{d \log(1/\delta)}{\epsilon^2}\right)$	$O(mnd)$
Fast JL -SRHT	$m = O\left(\frac{d \log(d) \log(1/\delta)}{\epsilon^2}\right)$	$O(mn \log(d))$
Countsketch	$m = O\left(\frac{d^2}{\delta \epsilon^2}\right)$	$O(nnz(\mathbf{A}))$

We have other sparse embeddings where nnz per column is > 1 , e.g, OSNAPs, sparse graphs.

Can improve $m = O\left(\frac{d \log(d) \log(1/\delta)}{\epsilon^2}\right)$ with $s = \Theta(\log(1/\delta))$ nonzero entries per column.

Further Reading:

- Woodruff, David P. “Sketching as a tool for numerical linear algebra.” Foundations and Trends® in Theoretical Computer Science 10.1–2 (2014): 1-157.
- Kane, Daniel M., and Jelani Nelson. “Sparsifier johnson-lindenstrauss transforms.” Journal of the ACM (JACM) 61.1 (2014): 1-23.
- Tropp, Joel A. “Improved analysis of the subsampled randomized Hadamard transform.” Advances in Adaptive Data Analysis 3.01n02 (2011): 115-126.

Questions?