

CS-GY 6763: Lecture 10

Randomized numerical linear algebra, ϵ -net arguments.

NYU Tandon School of Engineering, Prof. Rajesh Jayaram

ANNOUNCEMENTS

- HW3 Due tonight
- HW4 out by tomorrow.
- **Final Exam:** In class, on the last class Monday May 9th (not during scheduled final slot Tuesday May 10th!)
- Reading Group this Thursday: Atsushi will discuss the Contextual Bandits problem. Dennis and Jesse are Discussion leaders (presenters from last week).
- My office hours, moved to 4:30-5:30 Wednesday (just for this week).

RANDOMIZED NUMERICAL LINEAR ALGEBRA

Today: randomized algorithms for sketching (compressing) matrices

- Given a dense $n \times n$ matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.
- Computing top eigenvectors takes $\approx O(n^2/\sqrt{\epsilon})$ time (via power method/Krylov methods from last class).

If someone asked you to speed this up and return approximate top eigenvectors, what could you do?

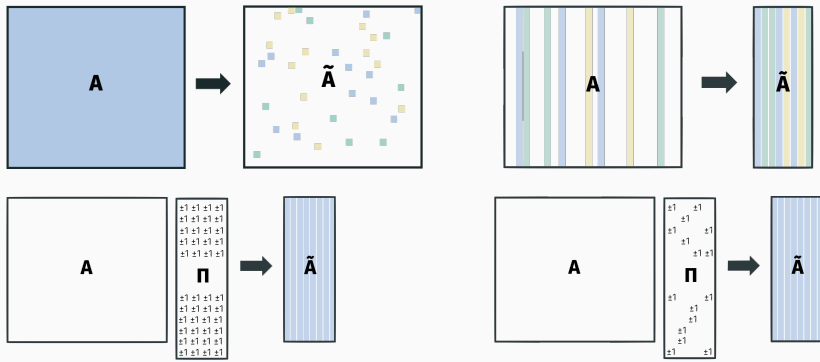
What about approximately solving the regression problem:

$$\min_x \|\mathbf{A}x - b\|_2$$

RANDOMIZED NUMERICAL LINEAR ALGEBRA

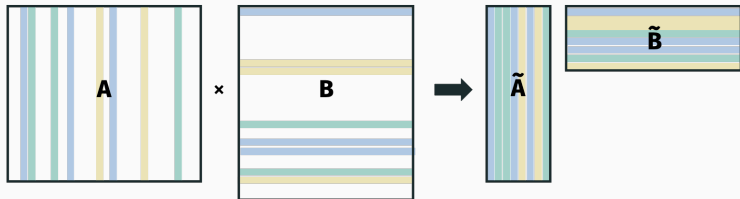
Main idea: If you want to compute singular vectors, multiply two matrices, solve a regression problem, etc.:

1. Compress your matrices using a randomized method (e.g. subsampling).
2. Solve the problem on the smaller or sparser matrix.
 - $\tilde{\mathbf{A}}$ called a “sketch” or “coreset” for $\mathbf{A}$.

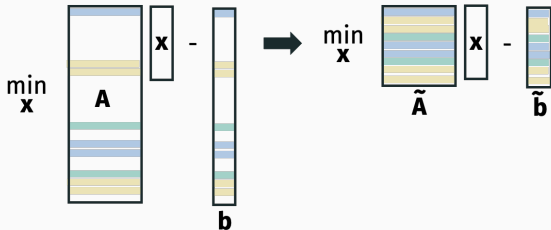


RANDOMIZED NUMERICAL LINEAR ALGEBRA

Approximate matrix multiplication:

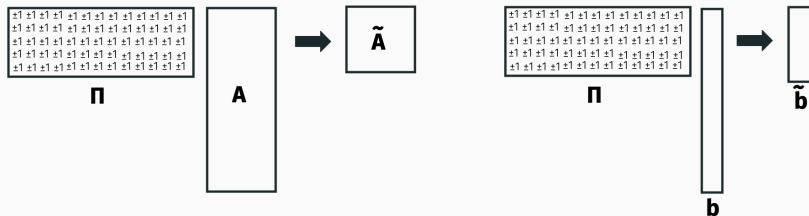


Approximate regression:



SKETCHED REGRESSION

Randomized approximate regression using a JL Matrix:



Input: $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$.

Goal: Let $x^* = \arg \min_x \|Ax - b\|_2^2$, and
 $\tilde{x} = \arg \min_x \|\Pi Ax - \Pi b\|_2^2$

Want: $\|A\tilde{x} - b\|_2^2 \leq (1 + O(\epsilon)) \|Ax^* - b\|_2^2$

If $\Pi \in \mathbb{R}^{m \times n}$, how large does m need to be? Is it even clear this should work as $m \rightarrow \infty$?

TARGET RESULT

Theorem (Randomized Linear Regression)

Let Π be a properly scaled JL matrix (random Gaussian, sign, sparse random, etc.) with $m = O\left(\frac{d}{\epsilon^2}\right)$ rows.¹ Then with probability 9/10, for any $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$,

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_2^2 \leq (1 + \epsilon) \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2$$

where $\tilde{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\Pi \mathbf{A} \mathbf{x} - \Pi \mathbf{b}\|_2^2$.

¹This can be improved to $O(d/\epsilon)$ with a tighter analysis

PLAN

- Prove this theorem using an ϵ -net argument, which is a popular technique for applying our standard concentration inequality + union bound argument to an infinite number of events.
- These sort of arguments appear all the time in theoretical algorithms and ML research, so this lecture is as much about the technique as the final result.
- You will need to use an ϵ -net argument to prove a matrix concentration inequality on your problem set.

SKETCHED REGRESSION

Claim: Suffices to prove that for all $\mathbf{x} \in \mathbb{R}^d$,

$$(1 - \epsilon) \|\mathbf{Ax} - \mathbf{b}\|_2^2 \leq \|\Pi \mathbf{Ax} - \Pi \mathbf{b}\|_2^2 \leq (1 + \epsilon) \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

Lemma (Distributional JL)

If Π is chosen to a properly scaled random Gaussian matrix, sign matrix, sparse random matrix, etc., with $O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ rows then for any fixed $\mathbf{y}$,

$$(1 - \epsilon)\|\mathbf{y}\|_2^2 \leq \|\Pi\mathbf{y}\|_2^2 \leq (1 + \epsilon)\|\mathbf{y}\|_2^2$$

with probability $(1 - \delta)$.

Corollary: For any fixed $\mathbf{x}$, with probability $(1 - \delta)$,

$$(1 - \epsilon)\|\mathbf{Ax} - \mathbf{b}\|_2^2 \leq \|\Pi\mathbf{Ax} - \Pi\mathbf{b}\|_2^2 \leq (1 + \epsilon)\|\mathbf{Ax} - \mathbf{b}\|_2^2.$$

FOR ANY TO FOR ALL

How do we go from “for any fixed $\mathbf{x}$ ” to “for all $\mathbf{x} \in \mathbb{R}^d$ ”.

This statement requires establishing a Johnson-Lindenstrauss type bound for an infinity of possible vectors ($\mathbf{Ax} - \mathbf{b}$), which can't be tackled directly with a union bound argument.

FOR ANY TO FOR ALL

How do we go from “for any fixed $\mathbf{x}$ ” to “for all $\mathbf{x} \in \mathbb{R}^d$ ”.

This statement requires establishing a Johnson-Lindenstrauss type bound for an infinity of possible vectors ($\mathbf{Ax} - \mathbf{b}$), which can't be tackled directly with a union bound argument.

Note: all vectors of the form ($\mathbf{Ax} - \mathbf{b}$) lie in a low dimensional subspace: spanned by $d + 1$ vectors, where $\mathbf{A} \in \mathbb{R}^{n \times d}$.

Even though the set is infinite, it is only $O(d)$ -dimensional instead of $O(n)$.

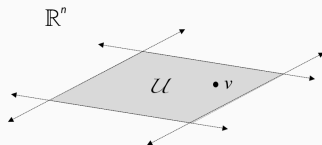
SUBSPACE EMBEDDINGS

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a d -dimensional linear subspace in $\mathbb{R}^n$. If $\Pi \in \mathbb{R}^{m \times d}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

$$(1 - \epsilon)\|\mathbf{v}\|_2^2 \leq \|\Pi\mathbf{v}\|_2^2 \leq (1 + \epsilon)\|\mathbf{v}\|_2^2$$

for all $\mathbf{v} \in \mathcal{U}$, as long as $m = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)^2$.



²It's possible to obtain a slightly tighter bound of $O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right)$. It's a nice challenge to try proving this.

SUBSPACE EMBEDDING TO APPROXIMATE REGRESSION

Corollary: If we choose Π and properly scale, then with $O(d/\epsilon^2)$ rows,

$$(1 - \epsilon) \|\mathbf{Ax} - \mathbf{b}\|_2^2 \leq \|\Pi \mathbf{Ax} - \Pi \mathbf{b}\|_2^2 \leq (1 + \epsilon) \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

for all $\mathbf{x}$ and thus

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_2^2 \leq (1 + O(\epsilon)) \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2.$$

I.e., our main theorem is proven.

Proof: Apply Subspace Embedding Thm. to the $(d + 1)$ dimensional subspace spanned by $\mathbf{A}$'s d columns and $\mathbf{b}$. Every vector $\mathbf{Ax} - \mathbf{b}$ lies in this subspace.

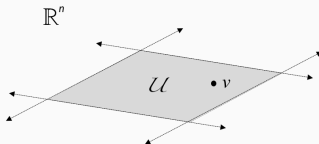
SUBSPACE EMBEDDINGS

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a d -dimensional linear subspace in $\mathbb{R}^n$. If $\Pi \in \mathbb{R}^{m \times d}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

$$(1 - \epsilon)\|\mathbf{v}\|_2^2 \leq \|\Pi\mathbf{v}\|_2^2 \leq (1 + \epsilon)\|\mathbf{v}\|_2^2 \quad (1)$$

for all $\mathbf{v} \in \mathcal{U}$, as long as $m = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$



Subspace embeddings have tons of other applications!

SUBSPACE EMBEDDING PROOF

$$(1 - \epsilon)\|\mathbf{v}\|_2^2 \leq \|\Pi\mathbf{v}\|_2^2 \leq (1 + \epsilon)\|\mathbf{v}\|_2^2 \quad (2)$$

First Observation: The theorem holds as long as (2) holds for all $\mathbf{w}$ on the unit sphere in $\mathcal{U}$. Denote the sphere $S_{\mathcal{U}}$:

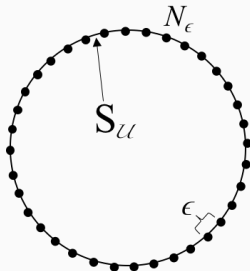
$$S_{\mathcal{U}} = \{\mathbf{w} \mid \mathbf{w} \in \mathcal{U} \text{ and } \|\mathbf{w}\|_2 = 1\}.$$

Follows from linearity: Any point $\mathbf{v} \in \mathcal{U}$ can be written as $c\mathbf{w}$ for some scalar c and some point $\mathbf{w} \in S_{\mathcal{U}}$.

- If $(1 - \epsilon)\|\mathbf{w}\|_2 \leq \|\Pi\mathbf{w}\|_2 \leq (1 + \epsilon)\|\mathbf{w}\|_2$.
- then $c(1 - \epsilon)\|\mathbf{w}\|_2 \leq c\|\Pi\mathbf{w}\|_2 \leq c(1 + \epsilon)\|\mathbf{w}\|_2$,
- and thus $(1 - \epsilon)\|c\mathbf{w}\|_2 \leq \|\Pi c\mathbf{w}\|_2 \leq (1 + \epsilon)\|c\mathbf{w}\|_2$.

SUBSPACE EMBEDDING PROOF

Intuition: There are not too many “different” points on a d -dimensional sphere:



N_ϵ is called an “ ϵ ”-net.

If we can prove

$$(1 - \epsilon)\|\mathbf{w}\|_2 \leq \|\Pi\mathbf{w}\|_2 \leq (1 + \epsilon)\|\mathbf{w}\|_2$$

for all points $\mathbf{w} \in N_\epsilon$, we can hopefully extend to all of S_U .

Lemma (ϵ -net for the sphere)

For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_{\mathcal{U}}$ with $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_{\mathcal{U}}$,

$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

Take this claim to be true for now: we will prove later.

SUBSPACE EMBEDDING PROOF

1. Preserving norms of all points in net N_ϵ .

Set $\delta' = \left(\frac{\epsilon}{4}\right)^d \cdot \delta$. By a union bound, with probability $1 - \delta$, for all $\mathbf{w} \in N_\epsilon$,

$$(1 - \epsilon)\|\mathbf{w}\|_2 \leq \|\Pi\mathbf{w}\|_2 \leq (1 + \epsilon)\|\mathbf{w}\|_2.$$

as long as Π has $O\left(\frac{\log(1/\delta')}{\epsilon^2}\right) = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$ rows.

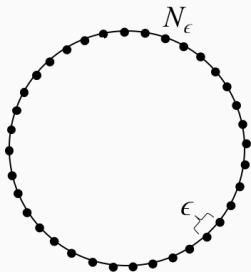
SUBSPACE EMBEDDING PROOF

2. Writing any point in sphere as linear comb. of points in N_ϵ .

For some $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2 \dots \in N_\epsilon$, any $\mathbf{v} \in S_{\mathcal{U}}$. can be written:

$$\mathbf{v} = \mathbf{w}_0 + c_1 \mathbf{w}_1 + c_2 \mathbf{w}_2 + \dots$$

for constants $c_1, c_2, \dots$ where $|c_i| \leq \epsilon^i$.



3. Preserving norm of $\mathbf{v}$.

Applying triangle inequality, we have

$$\begin{aligned}\|\Pi \mathbf{v}\|_2 &= \|\Pi \mathbf{w}_0 + c_1 \Pi \mathbf{w}_1 + c_2 \Pi \mathbf{w}_2 + \dots\| \\ &\leq \|\Pi \mathbf{w}_0\| + \epsilon \|\Pi \mathbf{w}_1\| + \epsilon^2 \|\Pi \mathbf{w}_2\| + \dots \\ &\leq (1 + \epsilon) + \epsilon(1 + \epsilon) + \epsilon^2(1 + \epsilon) + \dots \\ &\leq 1 + O(\epsilon).\end{aligned}$$

3. Preserving norm of $\mathbf{v}$.

Similarly,

$$\begin{aligned}\|\Pi \mathbf{v}\|_2 &= \|\Pi \mathbf{w}_0 + c_1 \Pi \mathbf{w}_1 + c_2 \Pi \mathbf{w}_2 + \dots\| \\ &\geq \|\Pi \mathbf{w}_0\| - \epsilon \|\Pi \mathbf{w}_1\| - \epsilon^2 \|\Pi \mathbf{w}_2\| - \dots \\ &\geq (1 - \epsilon) - \epsilon(1 + \epsilon) - \epsilon^2(1 + \epsilon) - \dots \\ &\geq 1 - O(\epsilon).\end{aligned}$$

SUBSPACE EMBEDDING PROOF

So we have proven

$$(1 - O(\epsilon)) \|\mathbf{v}\|_2 \leq \|\Pi \mathbf{v}\|_2 \leq (1 + O(\epsilon)) \|\mathbf{v}\|_2$$

for all $\mathbf{v} \in S_{\mathcal{U}}$, which in turn implies,

$$(1 - O(\epsilon)) \|\mathbf{v}\|_2^2 \leq \|\Pi \mathbf{v}\|_2^2 \leq (1 + O(\epsilon)) \|\mathbf{v}\|_2^2$$

Adjusting ϵ proves the Subspace Embedding theorem.

SUBSPACE EMBEDDINGS

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a d -dimensional linear subspace in $\mathbb{R}^n$. If $\Pi \in \mathbb{R}^{m \times d}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

$$(1 - \epsilon)\|\mathbf{v}\|_2^2 \leq \|\Pi\mathbf{v}\|_2^2 \leq (1 + \epsilon)\|\mathbf{v}\|_2^2 \quad (3)$$

for all $\mathbf{v} \in \mathcal{U}$, as long as $m = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$

Subspace embeddings have many other applications!

For example, if $m = O(k/\epsilon)$, $\Pi\mathbf{A}$ can be used to compute an approximate partial SVD, which leads to a $(1 + \epsilon)$ approximate low-rank approximation for $\mathbf{A}$.

Lemma (ϵ -net for the sphere)

For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_{\mathcal{U}}$ with $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_{\mathcal{U}}$,

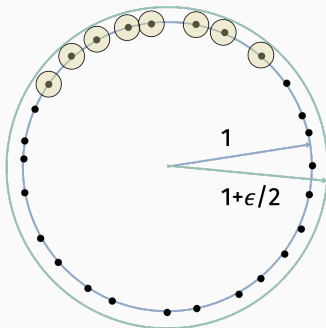
$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

Imaginary algorithm for constructing N_ϵ :

- Set $N_\epsilon = \{\}$
- While such a point exists, choose an arbitrary point $\mathbf{v} \in S_{\mathcal{U}}$ where $\nexists \mathbf{w} \in N_\epsilon$ with $\|\mathbf{v} - \mathbf{w}\| \leq \epsilon$. Set $N_\epsilon = N_\epsilon \cup \{\mathbf{w}\}$.

After running this procedure, we have $N_\epsilon = \{\mathbf{w}_1, \dots, \mathbf{w}_{|N_\epsilon|}\}$ and $\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon$ for all $\mathbf{v} \in S_{\mathcal{U}}$ as desired.

How many steps does this procedure take?



Can place a ball of radius $\epsilon/2$ around each $\mathbf{w}_i$ without intersecting any other balls. All of these balls live in a ball of radius $1 + \epsilon/2$.

Volume of d dimensional ball of radius r is

$$\text{vol}(d, r) = c \cdot r^d,$$

where c is a constant that depends on d , but not r . From previous slide we have:

$$\begin{aligned}\text{vol}(d, \epsilon/2) \cdot |N_\epsilon| &\leq \text{vol}(d, 1 + \epsilon/2) \\ |N_\epsilon| &\leq \frac{\text{vol}(d, 1 + \epsilon/2)}{\text{vol}(d, \epsilon/2)} \\ &\leq \left(\frac{1 + \epsilon/2}{\epsilon/2} \right)^d \leq \left(\frac{4}{\epsilon} \right)^d\end{aligned}$$

TIGHTER BOUND

You can actually show that $m = O\left(\frac{d + \log(1/\delta)}{\epsilon}\right)$ suffices to be a d dimensional subspace embedding, instead of the bound we proved of $m = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$.

The trick is to show that a constant factor net is actually all that you need instead of an ϵ factor.

RUNTIME CONSIDERATION

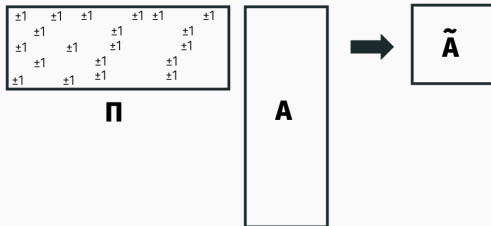
For $\epsilon, \delta = O(1)$, we need $\mathbf{\Pi}$ to have $m = O(d)$ rows.

- Cost to solve $\|\mathbf{Ax} - \mathbf{b}\|_2^2$:
 - $O(nd^2)$ time for direct method. Need to compute $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$.
 - $O(nd) \cdot (\# \text{ of iterations})$ time for iterative method (GD, AGD, conjugate gradient method).
- Cost to solve $\|\mathbf{\Pi Ax} - \mathbf{\Pi b}\|_2^2$:
 - $O(d^3)$ time for direct method.
 - $O(d^2) \cdot (\# \text{ of iterations})$ time for iterative method.

RUNTIME CONSIDERATION

But time to compute ΠA is an $(m \times n) \times (n \times d)$ matrix multiply:
 $O(mnd) = O(nd^2)$ time!

Goal: Develop faster Johnson-Lindenstrauss projections.



Typically using sparse and structured matrices.

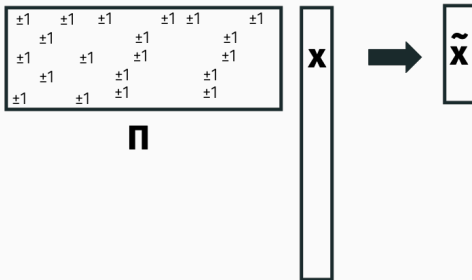
We will describe a construction where ΠA can be computed in
 $O(nd \log n)$ time.

After the break: Super-Fast JL Projections

RETURN TO SINGLE VECTOR PROBLEM

Goal: Develop methods that reduce a vector $\mathbf{x} \in \mathbb{R}^n$ down to $m \approx \frac{\log(1/\delta)}{\epsilon^2}$ dimensions in $o(mn)$ time and guarantee:

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leq \|\Pi\mathbf{x}\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}\|_2^2$$

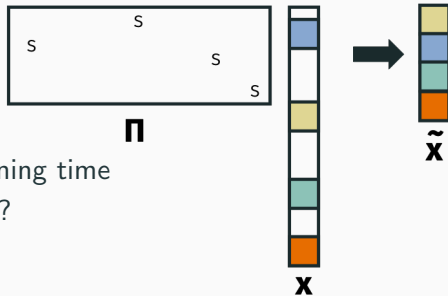


We will learn about a truly brilliant method that runs in $O(n \log n)$ time. **Preview:** Will involve Fast Fourier Transform in disguise.

FIRST ATTEMPT

Let Π be a **random sampling matrix**. Every entry is equal to $s = \sqrt{n/m}$ with probability $1/n$, and is zero otherwise.

subsampling matrix



What's the running time
to compute $\Pi\mathbf{x}$?

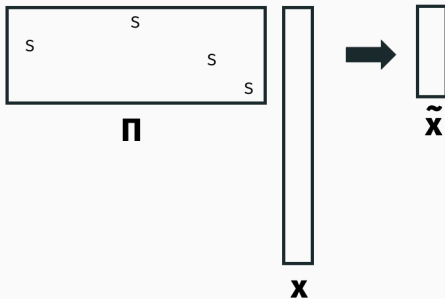
$$\|\Pi\mathbf{x}\|_2^2 = \sum_{i=1}^m \sum_{j=1}^n Z_i \cdot s^2 x_j^2$$

$$\mathbb{E}[\|\Pi\mathbf{x}\|_2^2] =$$

FIRST ATTEMPT

So $\mathbb{E}\|\Pi\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$ in expectation. To show it is close with high probability we would need to apply a concentration inequality. How do you think this will work out?

subsampling matrix



VARIANCE ANALYSIS

$$\|\mathbf{p}\mathbf{x}\|_2^2 = \sum_{i=1}^m \sum_{j=1}^n Z_i \cdot s^2 x_i^2$$

$$\begin{aligned}\sigma^2 &= \text{Var}[\|\mathbf{p}\mathbf{x}\|_2^2] \\&= \sum_{i=1}^m \sum_{j=1}^n s^4 x_i^4 \text{Var}[Z_i] \\&= \frac{n^2}{m^2} \sum_{i=1}^m \sum_{j=1}^n \frac{1}{n} x_i^4 \\&= \frac{n}{m^2} \sum_{i=1}^m \|\mathbf{x}\|_4^4 = \frac{n}{m} \|\mathbf{x}\|_4^4\end{aligned}$$

VARIANCE ANALYSIS

$$\begin{aligned}\|\Pi \mathbf{x}\|_2^2 &= \sum_{i=1}^m \sum_{j=1}^n Z_i \cdot s^2 x_i^2 \\ \sigma^2 &\leq \frac{n}{m} \|\mathbf{x}\|_4^4\end{aligned}$$

Recall Chebyshev's Inequality:

$$\Pr[|\|\Pi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2| \leq 10 \cdot \sigma] \leq \frac{1}{100}$$

We want additive error $|\|\Pi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2| \leq \epsilon \|\mathbf{x}\|_2^2$

VARIANCE ANALYSIS

We need to choose m so that:

$$10\sqrt{\frac{n}{m}}\|\mathbf{x}\|_4^2 \leq \epsilon\|\mathbf{x}\|_2^2.$$

How do these two norms compare?

$$\|\mathbf{x}\|_4^2 = \left(\sum_{i=1}^n x_i^4 \right)^{1/2} \qquad \|\mathbf{x}\|_2^2 = \sum_{i=1}^n x_i^2$$

Consider 2 extreme cases:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

VARIANCE FOR SMOOTH FUNCTIONS

We need to choose m so that:

$$\frac{1}{10} \sqrt{\frac{n}{m}} \|\mathbf{x}\|_4^2 \leq \epsilon \|\mathbf{x}\|_2^2.$$

Suppose $\mathbf{x}$ is very evenly distributed. I.e., for all $i \in 1, \dots, n$,

$$x_i^2 \leq \frac{c}{n} \sum_{i=1}^n x_i^2 = \frac{c}{n} \|\mathbf{x}\|_2^2$$

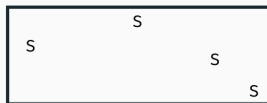
Claim: $\|\mathbf{x}\|_4^2 \leq \frac{c}{\sqrt{n}} \|\mathbf{x}\|_2^2$. So $m = O(c/\epsilon^2)$ samples suffices.³

³Using the right Bernstein bound we can prove $m = O(c \log(1/\delta)/\epsilon^2)$ suffices for failure probability δ .

VECTOR SAMPLING

So sampling does work to preserve the norm of $\mathbf{x}$, but only when the vector is relatively “smooth” (not concentrated). Do we expect to see such vectors in the wild?

subsampling matrix



$\mathbf{P}$



$\mathbf{x}$



$\tilde{\mathbf{x}}$

THE FAST JOHNSON-LINDENSTRAUSS TRANSFORM

Subsampled Randomized Hadamard Transform (SHRT) (Ailon-Chazelle, 2006)

Key idea: First multiply $\mathbf{x}$ by a “mixing matrix” $\mathbf{M}$ which ensures it cannot be too concentrated in one place.

$\mathbf{M}$ should have the property that $\|\mathbf{M}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$ exactly, or is very close. Then we will multiply by a subsampling matrix $\mathbf{S}$ to do the actual dimensionality reduction:

$$\Pi\mathbf{x} = \mathbf{S}\mathbf{M}\mathbf{x}$$

Oh... and $\mathbf{M}$ needs to be fast to multiply by!

THE FAST JOHNSON-LINDENSTRAUSS TRANSFORM

Good mixing matrices should look random:

+1	-1	+1	+1	+1	-1	+1	-1
-1	-1	-1	+1	+1	+1	-1	-1
+1	-1	+1	+1	+1	-1	-1	-1
+1	+1	+1	+1	-1	+1	-1	+1
-1	-1	+1	+1	-1	+1	+1	-1
-1	+1	-1	-1	-1	+1	-1	-1
-1	+1	-1	+1	-1	-1	-1	+1

M**x**

For this approach to work, we need to be able to compute $\mathbf{M}\mathbf{x}$ very quickly. So we will use a **pseudorandom** matrix instead.

THE FAST JOHNSON-LINDENSTRAUSS TRANSFORM

Subsampled Randomized Hadamard Transform (SHRT) (Ailon-Chazelle, 2006)

$\Pi = \mathbf{SM}$ where $\mathbf{M} = \mathbf{HD}$:

- $\mathbf{D} \in n \times n$ is a diagonal matrix with each entry uniform ± 1 .
- $\mathbf{H} \in n \times n$ is a Hadamard matrix.

The Hadarmard matrix is an orthogonal matrix closely related to the discrete Fourier matrix. It has two critical properties:

1. $\|\mathbf{H}\mathbf{v}\|_2^2 = \|\mathbf{v}\|_2^2$ exactly. Thus $\|\mathbf{HD}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$
2. $\|\mathbf{H}\mathbf{v}\|_2^2$ can be computed in $O(n \log n)$ time.

HADAMARD MATRICES RECURSIVE DEFINITION

Assume that n is a power of 2. For $k = 0, 1, \dots$, the k^{th} Hadamard matrix $\mathbf{H}_k$ is a $2^k \times 2^k$ matrix defined by:

$$H_0 = 1 \quad H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad H_2 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

$$H_k = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}$$

The $n \times n$ Hadamard matrix has all entries as $\pm \frac{1}{\sqrt{n}}$.

HADAMARD MATRICES ARE ORTHOGONAL

Property 1: For any $k = 0, 1, \dots$, we have $\|\mathbf{H}_k \mathbf{v}\|_2^2 = \|\mathbf{v}\|_2^2$ for all $\mathbf{v}$. I.e., $\mathbf{H}_k$ is orthogonal.

$$\begin{aligned} H_k H_k^T &= \frac{1}{2} \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix} \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix} \\ &= \end{aligned}$$

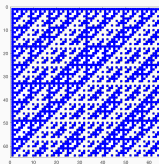
HADAMARD MATRICES

Property 2: Can compute $\mathbf{Px} = \mathbf{SHDx}$ in $O(n \log n)$ time.

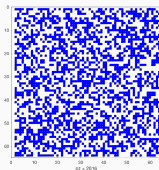
This is a nice exercise...can use recursion.

RANDOMIZED HADAMARD TRANSFORM

Property 3: The randomized Hadamard matrix is a good “mixing matrix” for smoothing out vectors.



Deterministic
Hadamard matrix.



Randomized Hadamard
PHD.



Fully random sign
matrix.

Blue squares are $1/\sqrt{n}$'s, white squares are $-1/\sqrt{n}$'s.

RANDOMIZED HADAMARD ANALYSIS

Lemma (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 $1 - \delta$,

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

for some fixed constant c .

The vector is very close to uniform with high probability. As we saw earlier, we can thus argue that $\|\mathbf{S}\mathbf{z}\|_2^2 \approx \|\mathbf{z}\|_2^2$. I.e. that:

$$\|\mathbf{\Pi}\mathbf{x}\|_2^2 = \|\mathbf{S}\mathbf{H}\mathbf{D}\mathbf{x}\|_2^2 \approx \|\mathbf{x}\|_2^2$$

JOHNSON-LINDENSTRAUSS WITH SHRTS

Theorem (The Fast JL Lemma)

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

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{x}\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}\|_2^2$$

with probability $(1 - \delta)$.

Very little loss in embedding dimension compared to full random matrix, and $\mathbf{\Pi}$ can be multiplied by $\mathbf{x}$ in $O(n \log n)$ (nearly linear) time.

RANDOMIZED HADAMARD ANALYSIS

SHRT mixing lemma proof: Need to prove $(z_i)^2 \leq \frac{c \log(n/\delta)}{n} \|\mathbf{z}\|_2^2$ for all i .

Let $\mathbf{h}_i^T$ be the i^{th} row of $\mathbf{H}$. $z_i = \mathbf{h}_i^T \mathbf{D} \mathbf{x}$ where:

$$\mathbf{h}_i^T \mathbf{D} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \dots & -1 & -1 \end{bmatrix} \begin{bmatrix} D_1 & & & & \\ & D_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & D_n \end{bmatrix}$$

where $D_1, \dots, D_n$ are random ± 1 's.

This is equivalent to

$$\mathbf{h}_i^T \mathbf{D} = \frac{1}{\sqrt{n}} \begin{bmatrix} R_1 & R_2 & \dots & R_n \end{bmatrix},$$

where $R_1, \dots, R_n$ are random ± 1 's.

RANDOMIZED HADAMARD ANALYSIS

So we have, for all i , $\mathbf{z}_i = \mathbf{h}_i^T \mathbf{D} \mathbf{x} = \frac{1}{\sqrt{n}} \sum_{i=1}^n R_i x_i$.

- $\mathbf{z}_i$ is a random variable with mean 0 and variance $\frac{1}{n} \|\mathbf{x}\|_2^2$, and is a sum of independent random variables.
- By Central Limit Theorem, we expect that:

$$\Pr[|\mathbf{z}_i| \geq t \cdot \frac{\|\mathbf{x}\|_2}{\sqrt{n}}] \leq e^{-O(t^2)}.$$

- Setting $t = \sqrt{\log(n/\delta)}$, we have for constant c ,

$$\Pr\left[|\mathbf{z}_i| \geq c \sqrt{\frac{\log(n/\delta)}{n}} \|\mathbf{y}\|_2\right] \leq \frac{\delta}{n}$$

- Applying a union bound to all n entries of $\mathbf{z}$ gives the SHRT mixing lemma.

RADEMACHER CONCENTRATION

Formally, need to use Bernstein type concentration inequality to prove the bound:

Lemma (Rademacher Concentration)

Let $R_1, \dots, R_n$ be Rademacher random variables (i.e. uniform ± 1 's). Then for any vector $\mathbf{a} \in \mathbb{R}^n$,

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

This is called the Khinchine Inequality. It is specialized to sums of scaled ± 1 's, and is a bit tighter and easier to apply than using a generic Bernstein bound.

FINISHING UP

With probability $1 - \delta$, we have that all $\mathbf{z}_i \leq \sqrt{\frac{c \log(n/\delta)}{n}} \|\mathbf{c}\|_2$.

As shown earlier, we can thus guarantee that:

$$(1 - \epsilon) \|\mathbf{z}\|_2^2 \leq \|\mathbf{S}\mathbf{z}\|_2^2 \leq (1 + \epsilon) \|\mathbf{z}\|_2^2$$

as long as $\mathbf{S} \in \mathbb{R}^{m \times n}$ is a random sampling matrix with

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

$\|\mathbf{S}\mathbf{z}\|_2^2 = \|\mathbf{S}\mathbf{H}\mathbf{D}\mathbf{x}\|_2^2 = \|\mathbf{\Pi}\mathbf{x}\|_2^2$ and $\|\mathbf{z}\|_2^2 = \|\mathbf{x}\|_2^2$, so we are done.

JOHNSON-LINDENSTRAUSS WITH SHRTS

Theorem (The Fast JL Lemma)

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

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leq \|\Pi\mathbf{x}\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}\|_2^2$$

with probability $(1 - \delta)$.

Upshot for regression: Compute $\Pi\mathbf{A}$ in $O(nd \log n)$ time instead of $O(nd^2)$ time. Compress problem down to $\tilde{\mathbf{A}}$ with $O(d^2)$ dimensions.

BRIEF COMMENT ON OTHER METHODS

$O(nd \log n)$ is nearly linear in the size of $\mathbf{A}$ when $\mathbf{A}$ is dense.

Clarkson-Woodruff 2013, STOC Best Paper: Possible to compute $\tilde{\mathbf{A}}$ with $\text{poly}(d)$ rows in:

$O(\text{nnz}(\mathbf{A}))$ time.

- Π is chosen to be an ultra-sparse random matrix (spoiler: Π is count-sketch!).
- Uses totally different techniques (you can't do JL + ϵ -net).

Lead to a whole class of matrix algorithms (for regression, SVD, etc.) which run in time:

$O(\text{nnz}(\mathbf{A})) + \text{poly}(d, \epsilon)$.