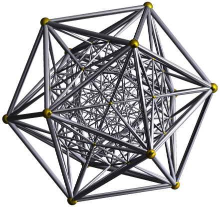
Computational Principles for High-dim Data Analysis

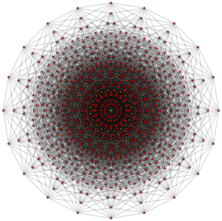
(Lecture Six)

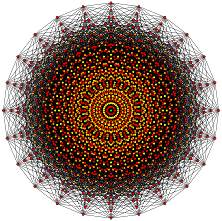
Yi Ma

EECS Department, UC Berkeley

September 16, 2021







Convex Methods for Sparse Signal Recovery
(Matrices with Restricted Isometry Property)

"Algebra is but written geometry; geometry is but drawn algebra."
– Sophie Germain

Restricted Isometry Property (Recap)

Definition (Restricted Isometry Property)

The matrix $\boldsymbol{A}$ satisfies the *restricted isometry property (RIP)* of *order* k, with constant $\delta \in [0, 1)$, if

$$\forall\, \boldsymbol{x} \;\; k\text{-sparse}, \quad (1-\delta)\, \|\boldsymbol{x}\|_2^2 \;\leq\; \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \;\leq\; (1+\delta)\, \|\boldsymbol{x}\|_2^2 \,. \tag{1}$$

The *order-k restricted isometry constant* $\delta_k(\boldsymbol{A})$ is the smallest number δ such that the above inequality holds.

Example of Gaussian Matrices: If $\boldsymbol{A}_\mathsf{I}$ is a large $m \times k$ $(k < m)$ matrix with entries independent $\mathcal{N}(0, 1/m)$,

$$\sigma_{min}(\boldsymbol{A}_\mathsf{I}^* \boldsymbol{A}_\mathsf{I}) \approx (\sqrt{1} - \sqrt{k/m})^2 \geq 1 - 2\sqrt{k/m},$$

$$\sigma_{max}(\boldsymbol{A}_\mathsf{I}^* \boldsymbol{A}_\mathsf{I}) \approx (\sqrt{1} + \sqrt{k/m})^2 \leq 1 + 3\sqrt{k/m}.$$

Length Concentration of Gaussian Vectors

Lemma (Gaussian Vectors[1])

Let $\boldsymbol{g} = [g_1, \ldots, g_m]^* \in \mathbb{R}^m$ *be an* m*-dimensional random vector whose entries are iid* $\mathcal{N}(0, 1/m)$*. Then for any* $t \in [0, 1]$*,*

$$\mathbb{P}\left[\Big|\|\boldsymbol{g}\|_2^2 - 1\Big| > t\right] \;\leq\; 2\exp\left(-\frac{t^2 m}{8}\right). \tag{2}$$

Proof (sketch): The moment generating function of a standard normal random variable is

$$\mathbb{E}\big[e^{\lambda x^2}\big] = (1 - 2\lambda)^{-1/2}, \quad \lambda < 1/2.$$

Then we have $\mathbb{E}\big[e^{\lambda g_i^2}\big] = \big(1 - \frac{2\lambda}{m}\big)^{-1/2}, \quad \lambda < \frac{m}{2}$.

[1] This result can be obtained via the Cramer-Chernoff exponential moment method (see Appendix E) or the book: *High-Dimensional Probability*, Roman Vershynin, 2018.

Length Concentration of Gaussian Vectors

Proof (continued):

$$\begin{aligned}
\mathbb{P}\Big[\sum_{i=1}^{m} g_i^2 > t+1\Big] &= \mathbb{P}\Big[\exp\Big(\lambda\sum_{i=1}^{m} g_i^2\Big) > \exp(t+1)\Big] \\
&\leq e^{-\lambda(t+1)}\mathbb{E}\Big[\exp\Big(\lambda\sum_{i=1}^{m} g_i^2\Big)\Big] \\
&= e^{-\lambda(t+1)}\prod_{i=1}^{m}\mathbb{E}\Big[e^{\lambda g_i^2}\Big] = e^{-\lambda(t+1)}\Big(1-\frac{2\lambda}{m}\Big)^{-m/2} \\
&\leq \exp(-\lambda t)
\end{aligned}$$

since for sufficiently small $\lambda = \frac{tm}{C}$, we have $\left(1-\frac{2\lambda}{m}\right)^{-m/2} \leq e^{\lambda}$.

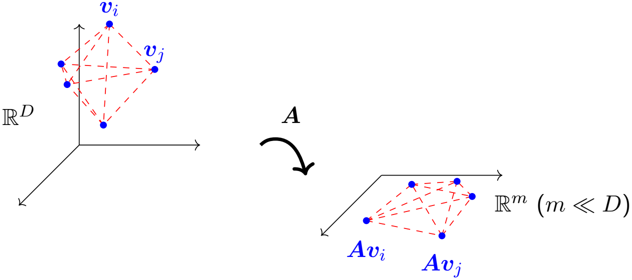
The JL Lemma: Distance Preserving Random Projections

Theorem (Johnson-Lindenstrauss Lemma)

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^D$ *and* $\boldsymbol{A} \in \mathbb{R}^{m \times D}$ *be a random matrix whose entries are i.i.d.* $\mathcal{N}(0, 1/m)$. *Then for any* $\epsilon \in (0,1)$, *with probability at least* $1 - 1/n^2$, *the following holds:*

$$\forall\, i \neq j, \;\; (1-\epsilon)\,\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 \;\leq\; \|\boldsymbol{A}\boldsymbol{v}_i - \boldsymbol{A}\boldsymbol{v}_j\|_2^2 \;\leq\; (1+\epsilon)\,\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2, \quad (3)$$

provided $m > 32\frac{\log n}{\epsilon^2}$.



The JL Lemma

Proof.

Finite cases: let $\boldsymbol{g}_{ij} = \boldsymbol{A}\frac{\boldsymbol{v}_i - \boldsymbol{v}_j}{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2}$ for any $i \neq j \ \in \{1, \ldots, n\}$.

Tail bound: $\boldsymbol{g}_{ij}$ is distributed as an iid Gaussian vector, with entries $\mathcal{N}(0, 1/m)$. Applying Lemma:

$$\mathbb{P}\left[\left|\|\boldsymbol{g}_{ij}\|_2^2 - 1\right| > t\right] \ \leq \ 2\exp\left(-t^2 m/8\right). \tag{4}$$

Union bound: Summing the probability of failure over all $i \neq j$, and then plugging in $t = \epsilon$ and $m \geq 32 \log n/\epsilon^2$, we get

$$\mathbb{P}\left[\exists\,(i,j)\ :\ \left|\|\boldsymbol{g}_{ij}\|_2^2 - 1\right| > t\right] \leq \frac{n(n-1)}{2} \times 2\exp\left(-t^2 m/8\right) \leq n^{-2}. \tag{5}$$

Hence $\left|\|\boldsymbol{g}_{ij}\|_2^2 - 1\right| \leq \epsilon$ with probability $1 - n^{-2}$.

□

The JL Lemma: Generalization to ℓ^p Norms

Locality-Sensitive Hashing[2]**:** for $p \in (0, 2]$, there exist the so-called p*-stable distributions* such that a random matrix $\boldsymbol{A}$ drawn from a p-stable distribution will preserve ℓ^p distance between vectors:

$$(1-\epsilon)\left\|\boldsymbol{v}_i - \boldsymbol{v}_j\right\|_p^2 \;\leq\; \left\|\boldsymbol{A}\boldsymbol{v}_i - \boldsymbol{A}\boldsymbol{v}_j\right\|_p^2 \;\leq\; (1+\epsilon)\left\|\boldsymbol{v}_i - \boldsymbol{v}_j\right\|_p^2. \qquad (6)$$

Example: For ℓ^1 norm, the corresponding distribution is the Cauchy distribution with density:

$$p(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}.$$

[2] Locality-sensitive hashing scheme based on p-stable distributions, M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. ACM SCG 2004.

The JL Lemma: Fast Nearest Neighbors

Compact Code for Fast Nearest Neighbor[3]:

```
Goal: Generate compact binary code for efficient nearest neighbor search of high-dimensional data points.
Input: x_1, ..., x_n ∈ ℝ^D and m = O(log n).
Generate a Gaussian matrix R ∈ ℝ^{m×D} with entries i.i.d. N(0, 1).
for i = 1, ..., n do
    Compute R x_i,
    Set y_i = σ(R x_i) where σ(·) is the entry-wise binary thresholding.
end for
Output: y_1, ..., y_n ∈ {0, 1}^m.
```

Instead of $O(\log n)$ real numbers, one only needs $O(\log n)$ binary bits!

[3] Compact projection: Simple and efficient near neighbor search with practical memory requirements, K. Min, J. Wright, and Y. Ma, CVPR 2010.

RIP of Gaussian Matrices

Theorem (RIP of Gaussian Matrices)

There exists a numerical constant $C > 0$ *such that if* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *is a random matrix with entries independent* $\mathcal{N}\left(0, \frac{1}{m}\right)$ *random variables, with high probability,* $\delta_k(\boldsymbol{A}) < \delta$, *provided*

$$m \;\geq\; Ck \log(n/k)/\delta^2. \tag{7}$$

Implications: ℓ^1 minimization can successfully recover k-sparse solutions $\boldsymbol{x}_o$ from about

$$m \geq Ck \log(n/k) \sim \Omega(k)$$

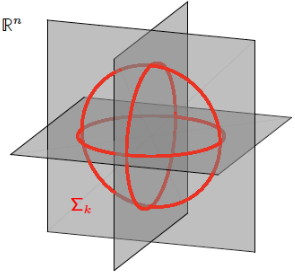
random measurements.

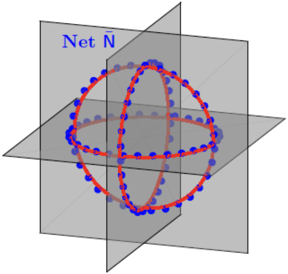
Proof: Step 1. Discretization to Finite Cases

$\delta_k(\boldsymbol{A}) \leq \delta$ if and only if $\sup_{\boldsymbol{x}\in\Sigma_k} \left| \|\boldsymbol{A}\boldsymbol{x}\|_2^2 - 1 \right| \leq \delta$ where

$$\Sigma_k = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_0 \leq k,\ \|\boldsymbol{x}\|_2 = 1\}. \tag{8}$$

Construct a finite (minimal) ϵ-**net** for Σ_k.





Proof: Step 1. Discretization to Finite Cases

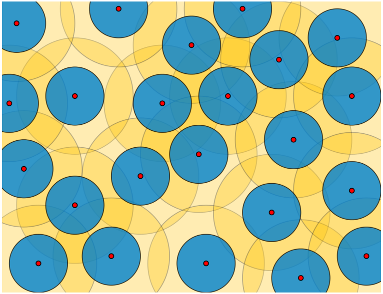
An ϵ-net (or covering) N for a given set S if

$$\forall \boldsymbol{x} \in \mathsf{S}, \quad \exists\, \bar{\boldsymbol{x}} \in \mathsf{N} \quad \text{such that} \quad \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \leq \epsilon. \tag{9}$$

A set M is **ϵ-separated** if every pair of distinct points $\boldsymbol{x}, \boldsymbol{x}'$ in M has distance at least ϵ:

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq \epsilon. \tag{10}$$

Fact: A maximal ϵ-separated subset $\mathsf{M} \subset \mathsf{S}$ is a (minimal) ϵ-net of S.



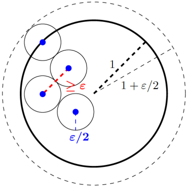
Proof: Step 1. Discretization to Finite Cases

Lemma (ϵ-Nets for the Unit Ball)

There exists an ϵ-net for the unit ball $\mathsf{B}(0,1) \subset \mathbb{R}^d$ of size at most $(3/\epsilon)^d$.

Proof: Let $\mathsf{N} \subset \mathsf{B}(\mathbf{0},1)$ be a *maximal* ϵ-separated set. The balls $\mathsf{B}(\boldsymbol{x},\epsilon/2)$ with $\boldsymbol{x} \in \mathsf{N}$ are contained in $\mathsf{B}(\mathbf{0},1+\epsilon/2)$. Thus,

$$|\mathsf{N}| \operatorname{vol}(\mathsf{B}(\mathbf{0},\epsilon/2)) \;\leq\; \operatorname{vol}(\mathsf{B}(\mathbf{0},1+\epsilon/2)). \tag{11}$$



Hence,

$$|\mathsf{N}| \;\leq\; \frac{\operatorname{vol}(\mathsf{B}(\mathbf{0},1+\epsilon/2))}{\operatorname{vol}(\mathsf{B}(\mathbf{0},\epsilon/2))} \tag{12}$$

$$= \left(\frac{1+\epsilon/2}{\epsilon/2}\right)^d = (1+2/\epsilon)^d \tag{13}$$

$$\leq (3/\epsilon)^d \tag{14}$$

Proof: Step 1. Discretization to Finite Cases

Lemma (Discretization)

Suppose we have a set $\bar{\mathsf{N}} \subseteq \Sigma_k$ *with the following property: for all* $\boldsymbol{x} \in \Sigma_k$*, there exists* $\bar{\boldsymbol{x}} \in \bar{\mathsf{N}}$ *such that*

- $|\mathit{supp}(\bar{\boldsymbol{x}}) \cup \mathit{supp}(\boldsymbol{x}|) \ \leq \ k$
- $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \leq \epsilon.$

set

$$\delta_{\bar{\mathsf{N}}} = \max_{\bar{\boldsymbol{x}} \in \bar{\mathsf{N}}} \left| \|\boldsymbol{A}\bar{\boldsymbol{x}}\|_2^2 - 1 \right|. \tag{15}$$

Then

$$\delta_k(\boldsymbol{A}) \ \leq \ \frac{\delta_{\bar{\mathsf{N}}} + 2\epsilon}{1 - 2\epsilon}. \tag{16}$$

Implications: RIP constant δ does not change much if we restrict our calculation to a finite ϵ-covering set $\bar{\mathsf{N}}$.

Proof: Step 1. Discretization to Finite Cases

Lemma (ϵ-Nets for Σ_k)

There exists an ϵ-net $\bar{\mathsf{N}}$ for Σ_k satisfying the two properties required in Lemma 6, with

$$|\bar{\mathsf{N}}| \;\leq\; \exp\Big(k\log(3/\epsilon) + k\log(n/k) + k\Big). \tag{17}$$

Proof.

Constructing an ϵ-Net for each ball in Σ_k and take the union. Using the Stirling's formula,[4] we can estimate

$$|\bar{\mathsf{N}}| \;\;\leq\;\; (3/\epsilon)^k \binom{n}{k} \leq (3/\epsilon)^k \left(\frac{ne}{k}\right)^k. \tag{18}$$

□

[4] Stirling's formula gives the bounds for factorials: $\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \leq k! \leq e\sqrt{k}\left(\frac{k}{e}\right)^k$.

Proof: Steps 2 and 3

Step 2: Tail Bound for Probability of Each Failure Case:
For each $\boldsymbol{x} \in \bar{\mathsf{N}}$, $\boldsymbol{A}\boldsymbol{x}$ is a random vector with entries independent $\mathcal{N}(0, 1/m)$. We have

$$\mathbb{P}\left[\left|\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - 1\right| > t\right] \;\leq\; 2\exp(-mt^2/8). \tag{19}$$

Step 3: Union Bound for Probability of All Failure Cases:
Summing over all elements of $\bar{\mathsf{N}}$, we have

$$\begin{aligned}\mathbb{P}\left[\delta_{\bar{\mathsf{N}}} > t\right] \;&\leq\; 2\left|\bar{\mathsf{N}}\right| \exp\left(-mt^2/8\right) & (20)\\ &\leq\; 2\exp\Big(-\frac{mt^2}{8} + k\log\big(\frac{n}{k}\big) + k(\log\big(\frac{3}{\epsilon}\big) + k)\Big). & (21)\end{aligned}$$

On the complement of the event $\delta_{\bar{\mathsf{N}}} > t$, we have

$$\delta_k(\boldsymbol{A}) \leq \frac{2\epsilon + t}{1 - 2\epsilon}. \tag{22}$$

Setting $\epsilon = \delta/8$, $t = \delta/4$, and ensuring that $m \geq Ck\log(n/k)/\delta^2$ for sufficiently large numerical constant C, we obtain the result.

RIP of Order k for Gaussian Matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$

From the above derivation, especially from equation (21), we see that **a slight more tight bound** for m is of the form

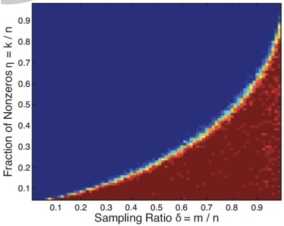
$$m \geq 128k \log(n/k)/\delta^2 + (\log(24/\delta) + 1)k/\delta^2 \doteq C_1 k \log(n/k) + C_2 k.$$

For a small δ, the constants C_1 and C_2 can be rather large.

A much tighter bound (one of the best known) for m is given as[5]:

$$m \geq 8k \log(n/k) + 12k.$$

A precise (phase transition) expression of m as function of k, n exists (Lecture Eight: Section 3.6 or Chapter 6).



[5] *On sparse reconstruction from Fourier and Gaussian measurements*, M. Rudelson and R. Vershynin. Comm. on Pure and Applied Mathematics, 2008.

RIP of Random Unitary Matrices

Motivating example: recall the MRI sensing model:

$$\boldsymbol{y} = \boldsymbol{F}_{\Omega} \boldsymbol{\Psi} \boldsymbol{x}, \quad \text{with } \boldsymbol{F} \text{ Fourier and } \boldsymbol{\Psi} \text{ wavelet.}$$

Theorem

Let $\boldsymbol{U} \in \mathbb{C}^{n \times n}$ be unitary ($\boldsymbol{U}^* \boldsymbol{U} = \boldsymbol{I}$) and Ω is a random subset of m elements from $\{1, \ldots, n\}$. Suppose that

$$\|\boldsymbol{U}\|_{\infty} \;\le\; \zeta / \sqrt{n}. \tag{23}$$

If

$$m \;\ge\; \frac{C\zeta^2}{\delta^2} k \log^4(n), \tag{24}$$

then with high probability, $\boldsymbol{A} = \sqrt{\frac{n}{m}} \boldsymbol{U}_{\Omega, \bullet}$ satisfies the RIP of order k, with constant $\delta_k(\boldsymbol{A}) \le \delta$.

Circulant Convolution Matrices

A (random) circulant convolution:

$$\boldsymbol{r} * \boldsymbol{x} \quad = \quad \begin{bmatrix} r_0 & r_{n-1} & \cdots & r_2 & r_1 \\ r_1 & r_0 & r_{n-1} & & r_2 \\ \vdots & r_1 & r_0 & \ddots & \vdots \\ r_{n-2} & & \ddots & \ddots & r_{n-1} \\ r_{n-1} & r_{n-2} & \cdots & r_1 & r_0 \end{bmatrix} \boldsymbol{x} \quad \doteq \quad \boldsymbol{R}\boldsymbol{x}. \tag{25}$$

Fact: any circulant matrix can be diagonalized by the discrete Fourier transform[6]:

$$\boldsymbol{R} = \boldsymbol{F}\boldsymbol{D}\boldsymbol{F}^*.$$

Select a (random) subset of the measurements:

$$\boldsymbol{y} = \mathcal{P}_\Omega[\boldsymbol{r} * \boldsymbol{x}] = \boldsymbol{A}\boldsymbol{x}, \tag{26}$$

[6] See Appendix A.7.

RIP of Random Circulant Convolution Matrices

Let $\boldsymbol{r}$ be a random vector with independent zero-mean, subgaussian random variables of variance one.

Theorem

Let $\Omega \subseteq \{1, \ldots, n\}$ be any fixed subset of size $|\Omega| = m$. Then if

$$m \geq \frac{Ck \log^2(k) \log^2(n)}{\delta^2}, \tag{27}$$

then with high probability, $\boldsymbol{A}$ has RIP of order k with $\delta_k(\boldsymbol{A}) \leq \delta$.

Approximate isometric property is the key to deep convolution neural networks![7]

[7] *Deep Isometric Learning for Visual Recognition*, H. Qi, C. You, X. Wang, Yi Ma, and J. Malik, ICML 2020.

Assignments

- Reading: Section 3.4 of Chapter 3.