IIT KHARAGPUR

IIT GANDHINAGAR

NPTEL ONLINE CERTIFICATION COURSES

NPTEL

# Scalable Data Science

## Lecture 14a: Sparse Random Projection


Anirban Dasgupta

Computer Science and Engineering

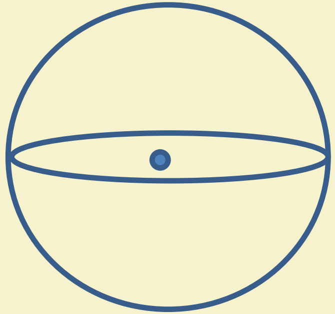
IIT GANDHINAGAR


IIT Gandhinagar

Indian Institute of Technology Gandhinagar

# Curse of Dimensionality

- Refers to rise in complexity of related issues with increasing dimensions
  - E.g. failure of intuition from low-dimensional cases: if $x \sim N(\mu, I_d)$ then $x$ does not lie “near” $\mu$



# Reducing dimension

- Basic theme
  - If we have to learn a model, or solve an optimization, reduce dimension of input to $k \ll d$
- Many variants on how to reduce dimension
  - PCA, Factor Analysis, ICA, feature selection…
- Random projection is unique among them in being data-oblivious

# Preserving pairwise distances

$x_1, x_2, \dots x_n \in R^d$ , Euclidean space

Want $x_1', x_2', \dots . x_n' \in R^k$ , possibly also Euclidean, $k \ll d$

with a guarantee that $\left|x_i' - x_j'\right| \approx \left|x_i - x_j\right|$ for every pair $(i, j)$?

# Johnson Lindenstrauss Lemma

$\epsilon, \delta > 0$ , $k \geq \frac{1}{\epsilon^2} \log(n)$ . There exists a linear mapping $A$ such that for all $(i, j)$

$$(1 - \epsilon)|x_i - x_j| \leq |Ax_i - Ax_j| \leq (1 + \epsilon)|x_i - x_j|$$

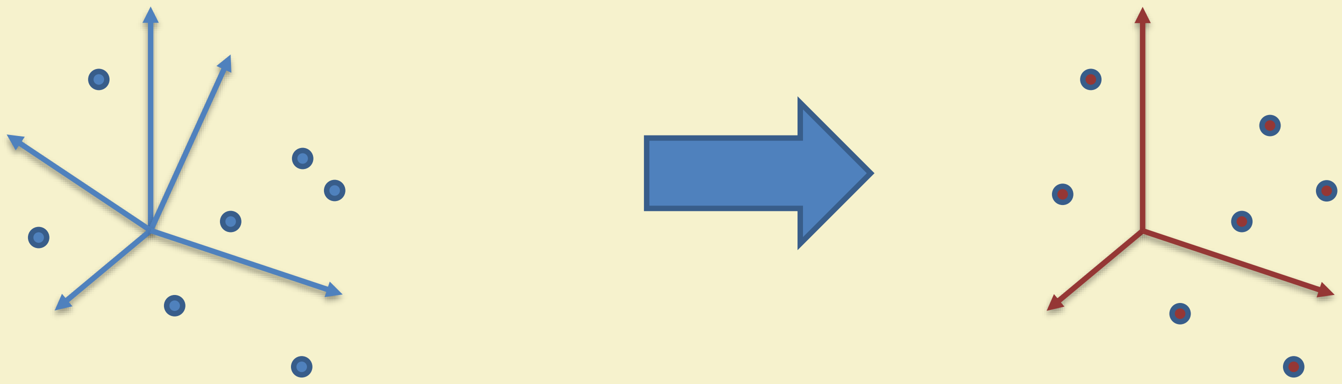
# Johnson Lindenstrauss Lemma [JL84]

$\epsilon > 0$ , $k \geq \frac{1}{\epsilon^2} \log(n)$ . There exists a linear mapping $A$ such that for all $(i, j)$

$$(1 - \epsilon)|x_i - x_j| \leq |Ax_i - Ax_j| \leq (1 + \epsilon)|x_i - x_j|$$

Can show that this actually happens with a high probability, also that such matrices $A$ are easy to construct (randomized)

# Intuition



Preserving pairwise distances preserves the inherent geometry of the points

Input dimension does not come into bound

# Why is it useful?

- Used in
  - learning classifiers efficiently
  - randomized numerical linear algebra: matrix factorizations and regressions
  - a number of streaming algos are essentially random projections
  - LSH for L2, other near neighbour data structures
  - compressed sensing

# How to create such a matrix

- $R_{ij} \sim N(0, 1)$ independently

- $\frac{1}{\sqrt{k}} R \in \Re^{k \times d}$ is the required matrix

# Why does this work?

JL lemma: $k \geq \frac{C}{\epsilon^2} \log\left(\frac{1}{\delta}\right)$ . For the previous matrix $A = \frac{1}{\sqrt{k}} R$ , $|x| = 1$

$$\Pr[(1 - \epsilon) \leq |Ax| \leq (1 + \epsilon)] \geq 1 - \delta$$

- Enough to show this since linear mapping
- Previous claim follows by taking union bound over all pairwise distances

# Proof Sketch

$Ax = \frac{1}{\sqrt{k}} Rx = \frac{1}{\sqrt{k}} (Y_1, \dots Y_k)$ $\quad Y_i = \sum_j R_{ij}\, x_j$

$|Ax|^2 = \frac{1}{k} \sum_i Y_i^2$ , we are interested in the distribution of this

<u>2-stability</u>: If p $\sim N(\mu, \sigma)$, $q \sim N(\alpha, \gamma)$ are iid, $p + q \sim N(\mu + \alpha, \sqrt{\sigma^2 + \gamma^2})$

# Proof Sketch

$Ax = \frac{1}{\sqrt{k}} Rx = \frac{1}{\sqrt{k}} (Y_1, \dots Y_k) \qquad Y_i = \sum_j R_{ij}\, x_j$

$|Ax|^2 = \frac{1}{k} \sum_i Y_i^2$ , we are interested in the distribution of this

<u>2-stability</u>: If p $\sim N(\mu, \sigma)$, $q \sim N(\alpha, \gamma)$ are iid, $p + q \sim N(\mu + \alpha, \sqrt{\sigma^2 + \gamma^2})$

$Y_i = \sum_j R_{ij} x_j$ , $Y_i \sim N(E[\sum_j R_{ij}\, x_j], \sigma(\sum_j R_{ij}\, x_j)) \sim N\left(0, \sqrt{\sum_j x_j^2}\right) \sim N(0,1)$

IIT Gandhinagar
Indian Institute of Technology Gandhinagar

NPTEL ONLINE CERTIFICATION COURSES

# Proof Sketch

$E\left[\frac{1}{k}\sum_j Y_i^2\right] = 1$ , also $kZ = \sum_i Y_i^2$ is $\chi_k^2$ distributed

We now apply Chernoff style tail bounds to show concentration of $kZ$ around the mean.

$$\Pr[\,Z > 1 + \epsilon] \leq \Pr\left[\exp(tkZ) > \exp\big(tk(1+\epsilon)\big)\right] \leq \frac{\mathrm{E}[\exp(tkZ)]}{\exp\big(tk(1+\epsilon)\big)}$$

$$\leq \frac{\prod_i E\left[\exp(tY_i^2)\right]}{\exp(tk(1+\epsilon))} \leq \ \ldots\ \leq \exp\left(-k\epsilon^2\, C\right)$$

Missing steps involve the MGF of the $\chi_k^2$ distribution and algebra

# Other properties

- The bound $k \geq \frac{c}{\epsilon^2} \log\left(\frac{1}{\delta}\right)$ is tight
  - Alon03, JW11
- Target dimension depends only on $\epsilon, \delta$
- Such result cannot hold in distances other than Euclidean metric

# Other constructions

[Achlioptas '03]

- Sampling and storing Gaussian random variables is expensive

$$A = \frac{1}{\sqrt{k}} R$$

- $R_{ij} = \begin{cases} +1 \ \ with\, prob\, \frac{1}{3} \\ 0 \ with\, prob\, \frac{2}{3} \\ -1 \ with\, prob\, \frac{1}{3} \end{cases}$

# Other constructions

[Achlioptas '03]

- Sampling and storing Gaussian random variables is expensive

$$A = \frac{1}{\sqrt{k}} R$$

- $R_{ij} = \begin{cases} +1 \;\; with\, prob\, \frac{1}{3} \\ 0 \; with\, prob\, \frac{2}{3} \\ -1 \; with\, prob\, \frac{1}{3} \end{cases}$

$R_{ij}$ $\sim$ any subgaussian distribution with variance 1

# Example Application: PCA

- Suppose we have $A \in \Re^{n \times d}$ , want to get a rank-k approximation $A'$ so that

  $|A \ - A'|_F$ is minimized

- Optimal low-rank approximation, $A_k = U_k \Sigma_k V_k^t$
  - Takes time $O(nd \min(n, d))$

# Low rank Approximation

- Projection $P_A^k = U_k U_k^t$ , $\left|A - P_A^k A\right|_2 = \sigma_{k+1}$
- For any $B$ and $P_B^k$,

$$\left|A - P_B^k A\right| \leq \sigma_{k+1} + \sqrt{2\left|AA^t - BB^t\right|} \quad \text{[FKV04]}$$

- Want $B$ that is
  - Efficiently computable and small
  - Leads to low error, $|AA^t - BB^t| \leq \epsilon |AA^t|$

[Example courtesy Edo Liberty]

# Cheap and effective low rank

- $A \in \Re^{n\times d}$ , create $R \in \Re^{d\times k}$ as a JL matrix
- $B = AR \in \Re^{n\times k}$

# Cheap and effective low rank

- $A \in \Re^{n \times d}$ , create $R \in \Re^{d \times k}$ as a JL matrix
- $B = AR \in \Re^{n \times k}$
- Note that $E[BB^t] = A\, E[RR^t]A^t = AA^t$

# JL in low rank approximation

- Using the JL property

$$\Pr[|yR|^2 - |y|^2 > \epsilon|y|^2] < \exp(-ck\epsilon^2)$$

Using union bound over a $\epsilon$ −net, $k = \tilde{O}\left(\frac{rank(A)}{\epsilon^2}\right)$

$$|A^tA - B^tB| = \sup_{|x|=1}\left| |xA|^2 - |xAR|^2 \right| \leq \epsilon|A^tA|$$

Time taken = $O(ndk)$

# Summary

- JL random projections a versatile tool
  - Tight bound on the number
- Number of ways to construct, will see more
- Saw a specific application

# References:

- Primary references
  - The Random Projection Method, Santosh Vempala, AMS.
  - Foundations of Data Science, Blum, Hopcroft, Kannan, *https://www.cs.cornell.edu/jeh/book.pdf*
  - Survey by Long Chen, https://www.math.uci.edu/~chenlong/RNLA/JL.pdf

# Thank You!!