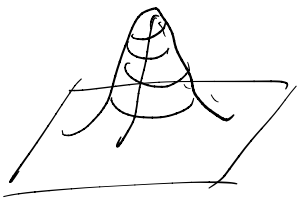
LECTURE 15

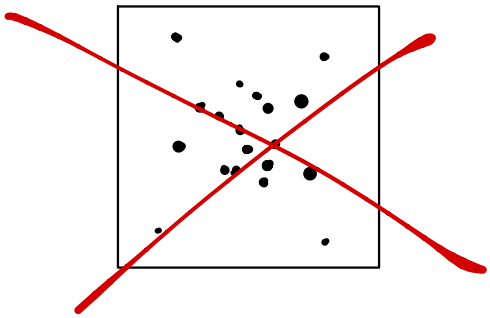
THE THIN SHELL PHENOMENON

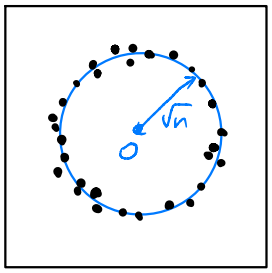
- Gaussian random vector $g \sim N(0, I_n)$: $g = (g_1, \dots, g_n)$, $g_i \sim N(0,1)$ iid
- pdf: $f(x) = f(x_1) \cdots f(x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}$

Rotation invariant :



- Sample from $N(0, I_n)$ for large n :





✓

THM (Thin shell)

For $g \sim N(0, I_n)$: $\mathbb{P}\{0.99\sqrt{n} \le \|g\|_2 \le 1.01\sqrt{n}\} \ge 1 - 2e^{-cn}$

↑ exponentially close to 1 !

Proof :

$$\|g\|_2^2 - n = \sum_{i=1}^{n} (g_i^2 - 1)$$

↑ iid, mean zero, subexponential :

$$\|g_i^2 - 1\|_{\psi_1} \le \|g_i^2\|_{\psi_1} + 1 \qquad (\Delta \text{ inequality})$$

$$= \|g_i\|_{\psi_2}^2 + 1 \le C \qquad (\text{abs. constant})$$

Bernstein's inequality $\Rightarrow$

$$\mathbb{P}\{|\|g\|_2^2 - n| \ge \underbrace{0.01n}_{t}\} \le 2\exp\left(-c \cdot \min\left(\frac{n^2}{\sigma^2}, \frac{n}{K}\right)\right) \quad (\le)$$

$|\|g\|_2^2 - n| \ge 0.001n$

where $\sigma^2 = \sum_{i=1}^n \|g_i^2 - 1\|_{\psi_1}^2 \le Cn$; $K = \max_i \|g_i^2 - 1\|_{\psi_1} \le C$.

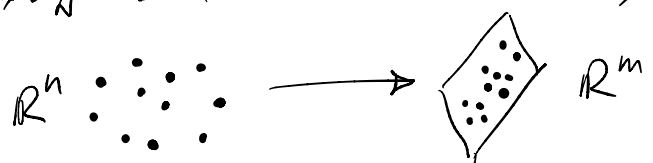
$\le 2\exp(-c'n)$.

$\Rightarrow$ with prob. $\ge 1 - 2\exp(-c'n)$,

$0.99n \le \|g\|_2^2 \le 1.01n$. QED.

APPLICATION: DIMENSION REDUCTION

- Data: $x_1, \dots, x_N \in \mathbb{R}^m \xrightarrow{?} \mathbb{R}^n$, $m \ll n$?



Data compression to save on storage, speed.

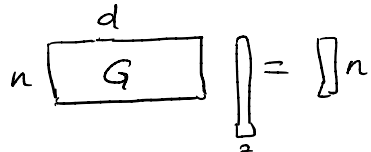
- Possible with $\boxed{n = O(\log N)}$; geometry of data preserved (pairwise distances).

THM (Johnson-Lindenstrauss lemma '1984) $\forall\, x_1, \dots, x_N \in \mathbb{R}^d$

$\exists$ linear map $T: \mathbb{R}^d \to \mathbb{R}^n$ such that $n \le C \log N$ and

$0.99\|x_i - x_j\| \le \|T(x_i) - T(x_j)\|_2 \le 1.01\|x_i - x_j\|$ $\forall\, i,j = 1, \dots, N$.

Proof. Probabilistic Method: choose random map



$G :=$ $n \times d$ Gaussian random matrix $G_{ij} \sim N(0,1)$ iid.

Claim: $\forall$ fixed $z \in \mathbb{R}^d$, $\|z\|_2 = 1$: $Gz \sim N(0, I_n)$

$(Gz)_i = \sum_{j=1}^d G_{ij} z_j \sim N(0, \sum_{j=1}^d z_j^2) = N(0,1)$ indep.

$G_{ij} z_j \sim N(0, z_j^2)$ indep.

Thin Shell Thm $\Rightarrow$ $\mathbb{P}\{0.99\sqrt{n} \le \|Gz\|_2 \le 1.01\sqrt{n}\} \ge 0.99$ (*)

Proof of JL ① Fix (i,j), use $(*)$ for $z = \frac{x_i - x_j}{\|x_i - x_j\|_2}$ $\Rightarrow$

$$P\left\{0.99\sqrt{n} \le \frac{\|G(x_i - x_j)\|_2}{\|x_i - x_j\|_2} \le 1.01\sqrt{n}\right\} \ge 1 - 2e^{-cn}.$$

$\Rightarrow$ for $T := \frac{1}{\sqrt{n}} G$,

$$P\left\{0.99\|x_i - x_j\|_2 \le \|Tx_i - Tx_j\|_2 \le 1.01\|x_i - x_j\|_2\right\} \ge 1 - 2e^{-cn}$$

$=: E_{ij}$

② Union Bound: $\exists$ N^2 pairs (i,j) $\Rightarrow$

$$P\{\forall i,j : E_{ij} \text{ holds}\} \ge 1 - N^2 \cdot 2e^{-cn} = 1 - 2\exp(2\log N - cn) \quad (\ge)$$

Choose n s.t. $\boxed{cn = 4\log N}$ $\Rightarrow$

$$(\ge)\ 1 - 2\exp(-2\log N) = 1 - \frac{2}{N^2} > 0 \quad \text{if } N \ge 2.$$

(and if $N = 1$, the thm is trivial).

$\Rightarrow$ such T exists. QED.

REMARKS 1. Why does not JL lift the "curse of high dimensionality"?
Preprocess HD data by JL $\Rightarrow$ low D data.

2. Fast JL transforms. [Ailon-Chazelle '2009]

3. JL without dependence on N = #data:
Prop 9.3.2 of book [Liaw-Mehrabian-Plan-V' 2017]

HW: JL for Ber [Achlioptas]