JOHNSON-LINDENSTRAUSS THEOREM

* MAKE NOTE OF CONNECTION BETWEEN FAST KNN AND APPROX. METHODS BASED ON SEARCH OVER LOW-DIMENSIONS

  * KNN CONNECTS ISOMAP/LLE W/ CLUSTERING

GIVEN N VECTORS IN $\mathbb{R}^d$ (d, N LARGE) WE CAN DEFINE THE <u>RANDOM PROJECTION</u>

$$f_K : \mathbb{R}^d \longrightarrow \mathbb{R}^k$$
$$\vec{x} \longrightarrow \begin{bmatrix} \vec{u}_1^T x \\ \vdots \\ \vec{u}_k^T \end{bmatrix} = U^T x \qquad U = [u_1, \dots u_k]$$

WHERE THE COORDINATES OF $\vec{u}_1, \dots \vec{u}_k$ ARE INDEPENDENT $N(0,1)$ RANDOM VARIABLES. WHAT HAPPENS TO DISTANCES $\|\vec{x}_i - \vec{x}_j\|$ AFTER THE PROJECTION $\|f(x_i) - f(x_j)\|$?

JOHNSON-LINDENSTRAUSS LEMMA

DERIVED LATER

FOR ANY $0 < \varepsilon < 1$ AND ANY INTEGER N, LET $k \geq \frac{3}{c\varepsilon^2} \log N$. $(c = 1/96)$

FOR <u>ANY</u> SET OF N POINTS $\vec{x}_1, \dots \vec{x}_N$ IN $\mathbb{R}^d$, THE RANDOM PROJECTION $f_K : \mathbb{R}^d \to \mathbb{R}^k$ (ABOVE) SATISFIES

$$P\left((1-\epsilon)\sqrt{k}\, \| x_i - x_j \| \leq \| f(\vec{x}_i) - f(\vec{x}_j) \| \leq (1+\epsilon)\sqrt{k}\, \|\vec{v}_i - \vec{v}_j\| ,\ 1 \leq i < j \leq N \right) \geq 1 - \frac{3}{2N}$$

$$\approx$$

$$P\left(\frac{\left| \frac{1}{k} \|f(x_i) - f(x_j)\|^2 - \|x_i - x_j\|^2 \right|}{\|\vec{x}_i - x_j\|^2} \leq 2\varepsilon ,\ \forall\, 1 \leq i < j \leq N \right)$$

Note:

1) The theorem holds for <u>all</u> x_i, x_j pairs not just most of them

2) For K-nearest neighbor searches can project to $k \propto \frac{\log d}{\varepsilon^2}$ - dimensional space and find the correct neighbors w/ high probability

- make no assumption on the distribution of $\vec{x}_1 \ldots \vec{x}_N$
- dominant term is $1/\varepsilon^2$ in applications,
  - $\frac{1}{\sqrt{k}}$ is uniform scaling of all projected distances
  - ε is maximum tolerable relative error between

$$\underbrace{\sqrt{k}\,\|\vec{x}_i - x_j\|}_{\text{rescaled original Euclidean distances}} \quad \vdots \quad \underbrace{\|f(\vec{x}_i) - f(\vec{x}_j)\|}_{\text{projected distances}}$$

3) Other versions of Thm

1) Require the components of $\vec{u}_1, \ldots \vec{v}_k$ are uncorrelated variance 1, mean 0 RVs
   - Gaussian eases proof

2) Posit the existence of map f which is Lipschitz (and can be found in polynomial time) w/ slightly better guarantees on error.

Our Goal

Prove the lemma, using properties and bounds on the the iid $N(0,1)$ components of $\vec{v}_1, \ldots \vec{v}_k$!

<u>Intuition:</u> If k is still large

$$\left.\begin{aligned} \|\vec{u}_j\|^2 &\approx k \\ \vec{u}_i \cdot \vec{u}_j &\approx k\,\delta_{ij} \end{aligned}\right\} \text{ by LLN}$$

Probability Review

Recall,

Markov's Inequality

Let Z be a random variable w/ $Z \geq 0$ and $EZ = \mu_z < \infty$

$$P(Z \geq a) \leq \frac{EZ}{a}$$

$$\left(P(|Z - EZ| \geq a) \leq \frac{E|Z - EZ|}{a} \leq \frac{2EZ}{a} \right)$$

Chebyshev's Inequality

Let Z be a random variable w/ $EZ = \mu_z < \infty$

$Var(Z) = \sigma_z^2 < \infty$

$$P\left(|Z - \mu_z| \geq a \right) = \frac{\sigma_z^2}{a^2}$$

Others Bernstein, Bennet, Hoeffding

Let $Z_1, \ldots Z_d$ be IID random variables drawn from the same distribution as Z w/ mean 0 and variance σ^2

$$P\left(|Z_1 + \ldots + Z_d| \geq a \right) \leq \frac{d\sigma^2}{a^2} \quad \text{(Chebyshev)}$$

But we'll need to do better.

Master Tail Bound (Blum)

Let $Z_1, \ldots, Z_d$ be mutually ind. random variables w/ mean 0 and variance at most σ^2. Suppose $a \in [0, \sqrt{2}\, d\sigma^2]$ and $s \leq d\sigma^2/2$ is a positive even integer and $|E(Z_i^r)| \leq \sigma^2 r!$ for $r = 2, 3, \ldots s$. Then

$$P\left(|Z_1 + \ldots + Z_d| \geq a \right) \leq \left(\frac{2sd\sigma^2}{a^2} \right)^{s/2}.$$

Further, if $s \geq a^2/(4d\sigma^2)$, then we also have $\quad s \in \left[\frac{a^2}{4d\sigma^2}, \frac{d\sigma^2}{2} \right]$

$$P\left(|Z_1 + \ldots + Z_d| \geq a \right) \leq 3e^{-a^2/12n\sigma^2}$$

Proof:

- First for $a \geq 0$ and even integers r

$$P\left(|Z_1 + \ldots + Z_d| \geq a \right) = P\left((Z_1 + \ldots + Z_d)^r \geq a^r \right) \leq \frac{E\left((Z_1 + \ldots + Z_d)^r \right)}{a^r}$$

So we need to work on bounds of $E\left(|Z_1 + \ldots + Z_d|^r \right)$!

- $$(Z_1 + \ldots + Z_d)^r = \sum_{r_1 + \ldots + r_d = r} \binom{r}{r_1, \ldots, r_d} Z_1^{r_1} \ldots Z_d^{r_d}$$

$$= \sum \frac{r!}{r_1! \ldots r_d!} Z_1^{r_1} \ldots Z_d^{r_d}$$

By independence,

$$E\left((Z_1 + \ldots + Z_d)^r \right) = \sum \left(\frac{r!}{r_1! \ldots r_d!} \right) E(Z_1^{r_1}) \ldots E(Z_d^{r_d})$$

IF $r_i = 1$ THEN $E z_i^{r_i} = E z_i = 0$ BY ASSUMPTION. SO WE FOCUS ON $(r_1, \dots r_d)$ RUNS OVER SETS W/

- $r_i = 0$ OR $r_i \geq 2$ $\quad i = 1, \dots d$

  * THERE ARE ONLY $\frac{d}{2}$ SUCH SETS !

- SINCE $E(z_i^{r_i}) \leq r_i! \, \sigma^2$

$$E\left((z_1 + \dots + z_d)^r \right) \leq \sum \frac{r!}{r_1! \dots r_d!} |E(z_1^{r_1})| \dots |E(z_d^{r_d})|$$

$$\leq r! \sum_{(r_1 + \dots + r_d = r)} \sigma^{2 (\# \text{ NONZERO } r_i)}$$

$$\leq r! \sum_{t=1}^{d/2} \binom{d}{t} \binom{r-t-1}{t-1} \sigma^{2t}$$

\# OF SUBSETS OF $\{1, \dots d\}$ W/ CARDINALITY t

\# OF WAYS TO ALLOCATE r TO t NONZERO r_i's W/ $r_i \geq 2$!

- $\binom{d}{t} = \frac{d(d-1) \dots (d-t+1)}{t!} \leq \frac{d^t}{t!}$

  $\binom{r-t-1}{t-1} \leq 2^{r-t-1}$ (TRIVIAL BOUNDS ON BINOMIAL COEFFICIENTS

$$\Rightarrow E\left[(z_1 + \dots + z_d)^r \right] \leq r! \sum_{t=1}^{r/2} h(t) \quad , \quad h(t) = \frac{d^t}{t!} 2^{r-t-1} \sigma^{2t}$$

$$\frac{h(t)}{h(t-1)} = \frac{d^t \, 2^{r-t-1}}{t!} \sigma^{2t} \Big/ \left(\frac{d^{t-1} \, 2^{r-t}}{(t-1)!} \sigma^{2(t-1)} \right) = \frac{d\sigma^2}{2t} \quad t \leq \frac{r}{2} \left(\leq \frac{d\sigma^2}{4} \right)$$

$$\geq 2$$

$$\Rightarrow E\left[(z_1 + \dots + z_d)^r \right] \leq r! \, h(r/2) \left(\underset{t = r/2}{1} + \underset{t = \frac{r}{2} - 1}{\frac{1}{2}} + \dots + \right) = \frac{r!}{(r/2)!} 2^{r/2} (d\sigma^2)^{r/2}$$

BY MARKOV'S INEQUALITY

$$P\left(|Z_1+\dots+Z_d|\geq a\right)=P\left((Z_1+\dots+Z_d)^r\geq a^r\right)\leq\frac{E\left[(Z_1+\dots+Z_d)^r\right]}{a^r}\leq\frac{r!\,(2d\sigma^2)^{r/2}}{(r/2)!\,a^r}=g(r)$$

$$\leq\left(\frac{2rd\sigma^2}{a^2}\right)^{r/2},\qquad \frac{r!}{(r/2)!}\leq r^{r/2}$$

WHICH HOLDS FOR $r\leq s$, r EVEN, AND APPLYING IT TO $r=s$ GIVES THE FIRST INEQUALITY!

FOR EVEN r, $g(r)=\frac{r!}{(r/2)!}\left(\frac{2d\sigma^2}{a^2}\right)^{r/2}$

$$\frac{g(r)}{g(r-2)}=\frac{r!\,(2d\sigma^2)^{r/2}}{(\frac{r}{2})!\,a^r}\cdot\frac{(\frac{r-2}{2})!\,a^{r-2}}{(r-2)!\,(2d\sigma^2)^{\frac{r-2}{2}}}=\frac{r(r-1)\,(2d\sigma^2)}{\frac{r}{2}}$$

$$=\frac{4(r-1)\,d\sigma^2}{a^2}\leq 1\quad\left(\Longleftrightarrow\quad r\leq\frac{a^2}{4d\sigma^2}+1\right)$$

$\Rightarrow$ $g(r)$ IS DECREASING SO LONG AS $r-1\leq\frac{a^2}{4d\sigma^2}$

IF $\boxed{s\geq}\ \frac{a^2}{4d\sigma^2}\geq r-1$ IS EVEN, WE MAY TAKE

$r^*=$ LARGEST EVEN INTEGER W/ $\frac{a^2}{6d\sigma^2}-2\leq r^*\leq\frac{a^2}{6ad^2}$.

$$\Rightarrow P\left((Z_1+\dots+Z_d)\geq a\right)\leq\left(\frac{2r^*d\sigma^2}{a^2}\right)^{r^*/2}$$

$$\leq\left(\frac{1}{3}\right)^{r^*/2}\leq\left(\frac{1}{e}\right)^{r^*/2}=e^{-r^*/2}$$

$\frac{a^2}{6d\sigma^2}-2\leq r^*$

$$\leq e^{-\left(\frac{a^2}{6d\sigma^2}-2\right)/2}=e\cdot e^{-\frac{a^2}{12d\sigma^2}}$$

$$\leq\boxed{3\,e^{-\frac{a^2}{12d\sigma^2}}}$$

$\boxed{X}$

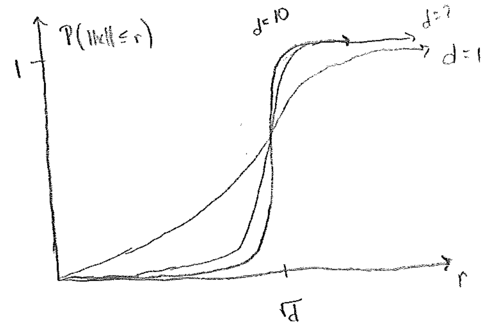
- CONSIDER $\vec{x} \sim \mathcal{N}(\vec{0}_d, I_d)$

$$P(\|x\| \le r) = \int_{\|x\| \le r} \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|x\|^2}{2\pi}} dx$$

$$= \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \cdot \frac{1}{(2\pi)^{d/2}} \int_0^r y^{d-1} e^{-y^2/2} dy$$

SPHERICAL COORDINATES

$$= \frac{2^{-\frac{d}{2}+1}}{\Gamma(d/2)} \int_0^r y^{d-1} e^{-y^2/2} dy$$



ALL PROBABILITY MASS LIES NEAR SURFACE OF UNIT SPHERE W/ RADIUS $\sqrt{d}$

GAUSSIAN ANNULUS THM

FOR $\vec{u} \sim \mathcal{N}(\vec{0}_d, I_d)$ AND $\beta \le \sqrt{d}$

$$P(\sqrt{d} - \beta \le \|u\| \le \sqrt{d} + \beta) \geq 1 - e^{-c\beta^2}$$

FOR SOME POSITIVE CONSTANT c.

PROOF:

$$P(\|u\| \notin [\sqrt{d}-\beta, \sqrt{d}+\beta]) \le P(\left|\|u\|^2 - d\right| \geq \beta\sqrt{d})$$

$$= P\left(\left|(u_1^2-1) + \dots + (u_d^2-1)\right| \geq \beta\sqrt{d}\right)$$

$$= P\left(\left|\underbrace{\frac{u_1^2-1}{2}}_{z_1} + \dots + \underbrace{\frac{u_d^2-1}{2}}_{z_d}\right| \geq \frac{\beta\sqrt{d}}{2}\right)$$

Observe

$$u_i \sim N(0,1) \quad \Rightarrow \quad E(z_i) = E\left(\frac{u_i^2-1}{2}\right) = 0$$

$$\left.\begin{aligned} Var(z_i) &= Var\left(\frac{u_i^2-1}{2}\right) = \frac{1}{4} Var(u_i^2) \le 2 \\ E(z_i)^r &\le 2\, r! \qquad r = 1, 2, \ldots \end{aligned}\right\} \text{Using properties of moments of Gaussians}$$

$\Rightarrow$ By Master Tail Bound w/ $\sigma^2 = 2$

$$P\left(\left| \frac{u_1^2-1}{2} + \dots + \frac{u_d^2-1}{2} \right| \ge \frac{\beta\sqrt{d}}{2} \right) \le 3\, e^{-\frac{\left(\frac{\beta\sqrt{d}}{2}\right)^2}{6d(2)}} = 3e^{-\frac{\beta^2}{96}}$$

$$\Rightarrow \quad P\left(\sqrt{d} - \beta \le \|u\| \le \sqrt{d} + \beta \right) \ge 1 - 3e^{-\beta^2/96}$$

$\boxtimes$

Random Projection Thm

Let $\vec{x} \in \mathbb{R}^d$ and f be the random projection to $\mathbb{R}^k$

$$P\left(\Big| \|f(\vec{x})\| - \sqrt{k}\, \|x\| \Big| \ge \varepsilon \sqrt{k}\, \|x\| \right) \le 3e^{-\frac{k\varepsilon^2}{96}}$$

$$= P\left(\|f(\vec{x})\| \notin \left[(1-\varepsilon)\sqrt{k}\, \|x\|,\ (1+\varepsilon)\sqrt{k}\, \|x\| \right] \right)$$

Proof: We may assume $\|x\| = 1$ above.

$$f(\vec{x}) = \begin{bmatrix} \vec{u}_1^T \\ \vdots \\ \vec{u}_k^T \end{bmatrix} \vec{x} \qquad \vec{u}_1, \ldots, \vec{u}_k \overset{iid}{\sim} N(0, I_d)$$

$$E(\vec{u}_i^T \vec{x}) = E[u_i^T]\, \vec{x} = 0$$

$$Var(\vec{u}_i^T \vec{x}) \overset{iid}{=} \sum_{j=1}^{d} Var(u_{ij} x_j) = \sum_{j=1}^{d} x_j^2\, Var(u_{ij}) = \sum_{j=1}^{d} x_j^2 = \|x\| = 1$$

$$\Rightarrow f(\vec{x}) \sim N(0, I_k) \quad \Rightarrow \quad P\left(\Big| \|f(\vec{x})\| - \sqrt{k}\, \|x\| \Big| \ge \varepsilon\sqrt{k}\, \|x\| \right) \le 3e^{-\frac{k\varepsilon^2}{96}}$$

$\boxtimes$

Proof (Johnson - Lindenstrauss)

$$P\left((1-\varepsilon)\sqrt{k}\,\|\underset{\vec{x}}{x_i - x_j}\| \le \|\underbrace{f(x_i) - f(x_j)}_{\underset{\vec{x}}{f(\vec{x}_i - \vec{x}_j)}}\| \le (1+\varepsilon)\sqrt{k}\,\|\underset{\vec{x}}{x_i - x_j}\| \ , \ 1 \le i < j \le N \right)$$

$$= 1 - P\left(\|f(x_i) - f(x_j)\| \notin \left[(1-\varepsilon)\sqrt{k}\,\|\vec{x}_i - \vec{x}_j\|, \ (1+\varepsilon)\sqrt{k}\,\|\vec{x}_i - \vec{x}_j\| \right] \right)$$

for some i, j

$$\geq 1 - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P\left(\|f(\vec{x}_i) - f(x_j)\| \notin \left[(1-\varepsilon)\sqrt{k}\,\|\vec{x}_i - \vec{x}_j\|, \ (1+\varepsilon)\sqrt{k}\,\|\vec{x}_i - \vec{x}_j\| \right] \right)$$

$$\geq 1 - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 3e^{-\frac{k\varepsilon^2}{96}} = 1 - 3\binom{N}{2} e^{\frac{-k\varepsilon^2}{96}}$$

$$\geq 1 - 3\binom{N}{2} e^{-\frac{3\cdot 96\,\cancel{\varepsilon^2}}{\cancel{96\,\varepsilon^2}}\log N} \qquad k \geq \frac{3\cdot 96}{\varepsilon^2}\log N$$

$$= 1 - 3\binom{N}{2}\left(\frac{1}{N^3}\right) \geq 1 - 3\left(\frac{N^2}{2}\right)\left(\frac{1}{N^3}\right) = 1 - \frac{3}{2N} \qquad \boxtimes$$