Lecture 11

Johnson - Lindenstrauss Lemma

clarification

Suppose X is a zero mean $\mathrm{SubG}(k^2)$ random variable

$\Rightarrow$

$$X \in \mathrm{subG}(k^2) \Rightarrow \text{for } \lambda < \frac{1}{k}$$

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(k^2 \lambda^2)$$

$$\exists c \quad \mathbb{E}[e^{X^2/c}] \overset{?}{\leq} 2$$

$$\leq e^{k^2/c} \leq 2$$

$\frac{1}{\sqrt{c}} < \frac{1}{k}$

$\frac{k^2}{c} \leq \ln 2$

set $c = \frac{k^2}{\ln 2}$ so both condition hold. $\square$

Norm of a vector of Gaussians

Let $\vec{v} = (v_1, \ldots, v_d)$ be a vector in $\mathbb{R}^d$

Suppose $v_i \sim N(0,1)$ are drawn from standard normal distribution.

what can we say about $\|v\|_2^2$?

we have shown $v_i^2 \in \text{SubE}(2^2, 4)$

Let $X = \sum_{i=1}^{d} v_i^2$. X is sum of d independent SubE $\Rightarrow$

$$X \in \text{SubE}(2^2 \cdot d, 4)$$

$$\Pr\left[\, \left| \|v\|_2^2 - d \right| \geq d\varepsilon \,\right]$$

$$\leq \Pr\left[\, \left| \frac{1}{d} \sum v_i^2 - 1 \right| \geq \varepsilon \,\right]$$

$$\leq 2 \exp\left(-d \cdot \min\left(\frac{\varepsilon^2}{2}, \frac{\varepsilon}{2} \right) \right)$$

$$\leq \begin{cases} 2\, e^{-\frac{d\varepsilon^2}{2}} & 0 \leq \varepsilon \leq 1 \\ 2\, e^{-\frac{d\varepsilon}{2}} & \varepsilon > 1 \end{cases}$$

using CLT:

$$E\left[\sum v_i^2\right] = d$$

$$\mathrm{Var}[v_i^2] = \underbrace{E[v_i^4]}_{3\cdot \mathrm{Var}[v_i]} - \underbrace{E[v_i^2]^2}_{=1}$$

$$\leq 3 - 1 = 2$$

$$\Rightarrow \mathrm{Var}\left(\sum v_i^2\right) = 2d$$

CLT: $$\frac{\sum v_i^2 - d}{\sqrt{2d}} \to N(0,1)$$

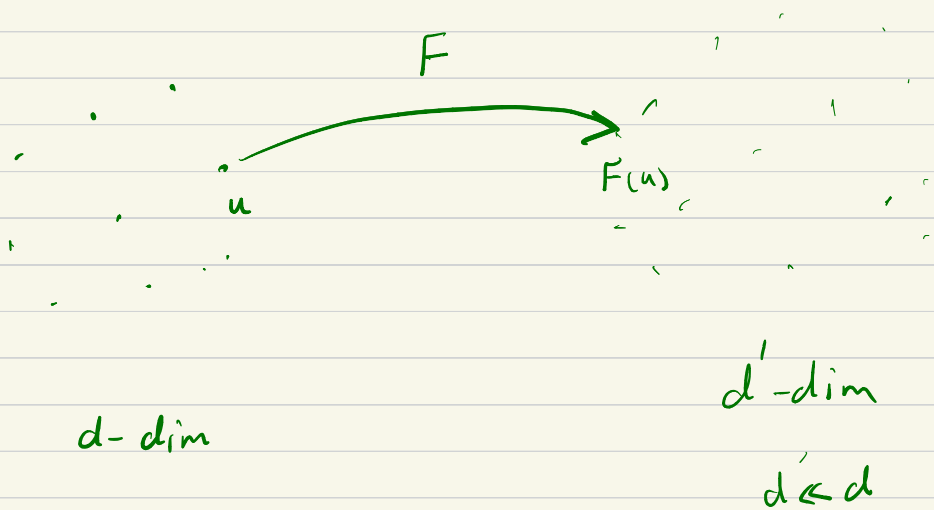
$$\Pr\left[\left|\sum v_i^2 - d\right| \geq \varepsilon d\right] \underset{d \to \infty}{\simeq}$$

$$\Pr\left[|Z| \geq \underbrace{\sqrt{\frac{d}{2}}\,\varepsilon}_{\infty ?}\right] \simeq e^{-\frac{d\varepsilon^2}{2}}$$

Is our previous bound loose?

Dimensionality reduction

Suppose we have n points in a d-dim space. Our hope is to embed the points to a lower dim (say d') such that the Euclidian distances between pair of points is preserved.



Example: k-means clustering

we have n points. Our goal is to partition the points into k clusters such that

sum of distances to the mean of the cluster is minimized. This is equivalent to ask for a partition $S = \{S_1, \dots, S_k\}$ that minimizes the following

$$\underset{S}{\arg\min} \sum_{S_i \in S} \sum_{x, y \in S_i} \| x - y \|_2^2$$

Generally, this problem is NP-hard. However, approximation algorithms exist with time complexity $\propto O(d)$. If we have an embedding to reduce the dimension, we can solve this problem faster.

Johnson – Lindenstrauss Lemma

Given n points in $\mathbb{R}^d$ $\rightarrow u_1, \dots, u_n$ and an integer

$d' \geq \frac{8 \log n/\delta}{\varepsilon^2}$, there exists a

randomized linear map $F: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$

such that with probability $1-\delta$:

$$(1-\varepsilon)\|u_i - u_j\|_2^2 \leq \|F(u_i) - F(u_j)\|_2^2 \leq (1+\varepsilon)\|u_i - u_j\|_2^2$$

The map:

Suppose we have an $d' \times d$ matrix

$M \in \mathbb{R}^{d' \times d}$ such that every

entry of M is drawn from

a normal distribution $N(0, \frac{1}{d'})$

$$\forall\, i,j \qquad M_{ij} \sim N\left(0, \frac{1}{d'}\right)$$

$$\text{Let} \quad F(u) = Mu$$

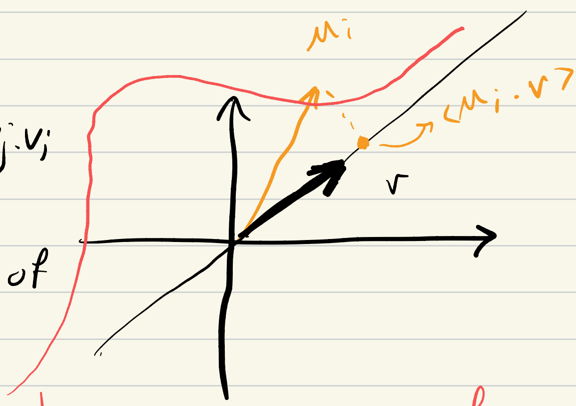
$$d' \left\{ \begin{bmatrix} F(u) \end{bmatrix} \right. = d' \left\{ \overbrace{\begin{bmatrix} \quad & M & \quad \end{bmatrix}}^{d} \right. \left. \begin{bmatrix} \\ u \\ \\ \end{bmatrix} \right\} d$$

Proof of lemma

step 1 : projection of a random vector on to a fix direction

Suppose we have a fixed vector v of size $\|v\|_2 = 1$. Let $M_i \in \mathbb{R}^d$ such that $M_{ij} \sim N(0, \frac{1}{d})$. We claim that $\langle M_i . v \rangle$ is a Gaussian random variable

$$Z = \langle M_i . v \rangle = \sum_{j=1}^{d} M_{ij} . v_j$$



Z is a linear combination of Gaussians.

$\Rightarrow Z \sim N(0, ?)$

distribution of $\langle M_i . v \rangle$

$$\mathrm{Var}(Z) = \mathrm{Var}\left[\sum_{j=1}^{d} M_{ij} \cdot v_j\right]$$

independence of M_{ij}'s

$$= \sum_{j=1}^{d} v_j^2 \cdot \underbrace{\mathrm{Var}[M_{ij}]}_{= 1/d'}$$

$$= \frac{\|v\|_2^2}{d'} = \frac{1}{d'}$$

$$\Rightarrow \quad \langle M_i \cdot v \rangle \sim \mathcal{N}\left(0, \frac{1}{d'}\right)$$

step 2 if $\|v\|_2^2 = 1$, then $\|F(v)\|_2^2$ is a sub-exponential r.v.

$$\|F(v)\|_2^2 = \|M \cdot v\|_2^2 = \sum_{i=1}^{d'} (\langle M_i \cdot v \rangle)^2$$

by step 1, $\langle M_i \cdot v \rangle \sim N\left(0, \frac{1}{d'}\right)$

In distribution: $\langle M_i \cdot v \rangle \to \frac{1}{\sqrt{d'}} Z_i$

where $Z_i \sim N(0,1)$

$\Rightarrow$ In distribution $\|F(v)\|_2^2 = \sum_{i=1}^{d'} \frac{Z_i^2}{d'}$

Recall, earlier we have shown:

$$\Pr\left[\left|\|F(v)\|_2^2 - 1\right| \geq \varepsilon\right]$$

$$= \Pr\left[\left|\sum_{i=1}^{d'} \frac{Z_i^2}{d'} - 1\right| \geq \varepsilon\right]$$

$$\leq 2\exp\left(-\frac{d'\varepsilon^2}{2}\right) \qquad \text{for } \varepsilon \in (0,1]$$

step 3 Suppose $\|u\|_2 \neq 0$

$\|F(u)\|_2^2$ has a similar tail bound even when $\|u\|_2 \neq 1$

Let $v = \dfrac{u}{\|u\|_2}$. clearly $\|v\|_2^2 = 1$

$$\Pr\left[\left|\frac{\|F(u)\|_2^2}{\|u\|_2^2} - 1\right| > \varepsilon\right]$$

$$= \Pr\left[\left|\frac{\|\|u\|_2 . M . v\|_2^2}{\|u\|_2^2} - 1\right| > \varepsilon\right]$$

$$= \Pr\left[\left|\|F(v)\|_2^2 - 1\right| > \varepsilon\right]$$

$$\leq 2\exp\left(-\frac{d'\varepsilon^2}{2}\right) \quad \text{for } \varepsilon \in (0,1]$$

Step 4. proof of JL lemma

Our Goal is to show $\|F(u_i) - F(u_j)\|_2^2$ is within $(1 \pm \varepsilon)$ factor of $\|u_i - u_j\|_2^2$.

More precisely, we show that

$$\Pr\left[\exists\ i,j : \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \notin [1-\varepsilon, 1+\varepsilon]\right] \overset{?}{\leq} \delta$$

<u>proof</u>

For every $i,j \in [n]$, if $u_i = u_j$ the embedding does not change the Euclidean distance.

Assume $u_i \neq u_j$, and let

$$u_{ij} = u_i - u_j$$

Clearly $F(u_{ij}) = M(u_i - u_j) = F(u_i) - F(u_j)$

$$\Pr\left[\frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \notin [1-\varepsilon, 1+\varepsilon] \right]$$

$$= \Pr\left[\left| \frac{\|F(u_{ij})\|_2^2}{\|u_{ij}\|_2^2} - 1 \right| > \varepsilon \right]$$

$$\underset{\text{by part 3}}{\leq} 2 \exp\left(- \frac{d' \varepsilon^2}{2} \right)$$

Using union bound:

$$\Pr\left[\exists\ i,j : \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \notin [1-\varepsilon, 1+\varepsilon] \right]$$

$$\leq \binom{n}{2} 2 \exp\left(- d' \varepsilon^2 / 2 \right) \leq \delta$$

by setting $d' = \Theta\left(\frac{\log n/\delta}{\varepsilon^2} \right)$