Johnson-Lindenstrass Approximation:

- Given a set of M data points, $\{u_1, u_2, \cdots, u_M\} \in \mathbb{R}^p$.
- Want to perform dimension reduction $F: \mathbb{R}^p \to \mathbb{R}^N$ with $N \ll p$ such that for some tolerance $\delta \in (0,1)$,

$$(1-\delta)\|u_i - u_j\|_2^2 \le \|F(u_i) - F(u_j)\|_2^2 \le (1+\delta)\|u_i - u_j\|_2^2$$

for all pairs $i \neq j$.

- Form a random matrix $Z \in \mathbb{R}^{N \times p}$ with each entry $Z_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and define

$$F(u_m) := \frac{1}{\sqrt{N}} Z u_m \quad \text{for } m = 1, \cdots, M.$$

- $N\|F(u)\|_2^2 = \|Zu\|_2^2 = \sum_{i=1}^{N} \langle z_i, u \rangle^2$ (z_i is the ith row of Z)

so $N \dfrac{\|F(u)\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{N} \left\langle z_i, \dfrac{u}{\|u\|_2} \right\rangle^2$ for $u \neq 0$.

- Since $\left\langle z_i, \dfrac{u}{\|u\|_2} \right\rangle \sim \mathcal{N}(0,1)$,

$N \cdot \dfrac{\|F(u)\|_2^2}{\|u\|_2^2} \sim \chi_N^2$ because the rows of Z are independent.

- By χ_N^2 concentration bound,

$$\mathbb{P}\left(\left| \frac{\chi_N^2}{N} - 1 \right| \ge t \right) \le 2e^{-\frac{Nt^2}{8}} \quad \forall t \in (0,1)$$

we have.

$$\mathbb{P}\left(\left|\frac{\|F(u)\|_2^2}{\|u\|_2^2} - 1\right| > \delta\right) \leq 2e^{-\frac{N\delta^2}{8}} \quad \forall \delta \in (0,1)$$

$$\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [1-\delta, 1+\delta]$$

- By Union bound.

$$\mathbb{P}\left(\frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2} \notin [1-\delta,1+\delta] \text{ for some pair } i \neq j\right)$$

$$\leq 2\cdot\binom{M}{2} e^{-\frac{N\delta^2}{8}} \leq M^2 e^{-\frac{N\delta^2}{8}} \quad (*).$$

- If we choose $N > \frac{16}{\delta^2}\log\left(\frac{M}{\varepsilon}\right)$, then.

$(*) < \varepsilon^2 < \varepsilon$ for any $\varepsilon \in (0,1)$.

That is.

$$\mathbb{P}\left(\frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2} \in [1-\delta,1+\delta] \text{ for all pairs } i \neq j\right)$$

$> 1-\varepsilon$. as long as $N > \frac{16}{\delta^2}\log\left(\frac{M}{\varepsilon}\right)$.

- The dimensionality N has two properties.

(1) independent of original dimensionality P.

(2) depends on the number of samples M only logarithmically.