Dimension Reduction: Johnson-Lindenstrauss Lemma[1]

- **Dimension Reduction.** In many applications, data entries are vectors in high dimension. Concretely, it could be vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ where each $\mathbf{v}_i \in \mathbb{R}^d$, and d can be of the same order as n. The term dimension reduction refers to a mapping/embedding of these vectors to a lower dimensional space such that certain properties of the data-set are preserved. In this lecture, the property we will focus on is the Euclidean distance between the points. In particular, we search for a map which takes

$$\mathbf{v}_i \in \mathbb{R}^d \mapsto \Phi(\mathbf{v}_i) \in \mathbb{R}^k \tag{1}$$

  such that

$$\forall 1 \leq i < j \leq n: \quad \|\Phi(\mathbf{v}_i) - \Phi(\mathbf{v}_j)\|_2 \in (1 \pm \varepsilon) \cdot \|\mathbf{v}_i - \mathbf{v}_j\|_2 \tag{2}$$

  We would like to know how small a k can we get away with.

  Note that a "low dimensional" intuition doesn't get us far. If we ask for dimension reduction from 2 to 1, or 3 to 2, the situation seems hopeless. The point is that the k would still be growing with n but would be much smaller.

- **The Johnson-Lindenstrauss Lemma.** One of the most influential mathematical theorems in computer science is the Johnson-Lindenstrauss lemma. Indeed, it is often stated and used without even citing the paper[2] of Johnson and Lindenstrauss. That's the mark of a general result! Also, it is called a lemma for it is Lemma 1 in their paper, and they use it to prove a theorem about extensions of functions in metric spaces. The JL-lemma says that dimension reduction satisfying (1) and (2) is possible with $k = O\left(\frac{C \log n}{\varepsilon^2}\right)$, and furthermore, the map $\Phi()$ is a *linear map*, that is, $\Phi(\mathbf{x}) = A \cdot \mathbf{x}$ for some $k \times n$ matrix A. And in fact a random matrix works with high probability.

  Theorem 1 (JL Lemma.)**.** Fix n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in $\mathbb{R}^d$. Let $k = \frac{6 \ln n}{\varepsilon^2}$ for some fixed constant C. Let A be a random $k \times d$ matrix where each $A_{ij} \sim \frac{1}{\sqrt{k}} \cdot N(0, 1)$ is drawn from a normal distribution with mean 0 and standard deviation 1. Then, with probability $1 - \frac{1}{n}$,

$$\forall 1 \leq i, j \leq n, \quad \|A\mathbf{v}_i - A\mathbf{v}_j\|_2 \in (1 \pm \varepsilon) \|\mathbf{v}_i - \mathbf{v}_j\|_2$$

- **Gaussian Random Variables.** Let us state some facts about Gaussian random variables that we will be needing in the proof. First recall that the probability distribution function of $Z \sim N(0, 1)$ is

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

  We will need two facts.

  Fact 1. If $Z_1, \ldots, Z_t$ are t independent Gaussians where each $Z_i \sim N(0, \sigma_i^2)$, then $Z := \sum_{i=1}^t a_i Z_i$ is distributed as $N(0, \sigma^2)$ where $\sigma^2 = \sum_{i=1}^t a_i^2 \sigma_i^2$.

---

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 24th May, 2023*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

[2] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz Mappings into a Hilbert Space.*

The second fact we will need is about the sum of *squares* of a normal random variable. Let $Z_1, \ldots, Z_k$ iid Gaussians $\sim N(0,1)$. Then, the random variable $X = \sum_{t=1}^{k} Z_i^2$ is said to be distributed using the χ^2-distribution with k-degrees of freedom.

It's easy to see

$$\mathbf{Exp}[X] = \sum_{i=1}^{k} \mathbf{Exp}[Z_i^2] = k$$

and a little more work shows that

$$\mathbf{Var}[X] = 2k$$

The "rule of thumb" concentration[3] then suggests that $\mathbf{Pr}[|X-k| \geq \varepsilon k] \leq e^{-C\varepsilon^2 k}$ since $\varepsilon k = O(\varepsilon\sqrt{k})$ standard deviations. Indeed, that is a fact. The following states that in a general form (and follows by scaling)

Fact 2. If $Z_1, \ldots, Z_k$ are t independent Gaussians where each $Z_i \sim N(0,\sigma^2)$, then $X = \sum_{t=1}^{k} Z_t^2$ satisfies $\mathbf{Exp}[X] = k\sigma^2$ and

$$\mathbf{Pr}[X \notin (1 \pm \varepsilon) \cdot \mathbf{Exp}[X]] \leq 2e^{-\varepsilon^2 k/8}$$

- **Random Linear Maps.** The proof of the JL lemma follows almost immediately from the above two facts. Indeed, what one can show is this:

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \mathop{\mathbf{Pr}}_{A}\left[\, \|A\mathbf{x}\|_2 \notin (1 \pm \varepsilon) \cdot \|\mathbf{x}\|_2 \,\right] \leq 2e^{-\varepsilon^2 k/2} \tag{3}$$

  Recall, A is a random $k \times d$ matrix where each $A_{ij} \sim \frac{1}{\sqrt{k}} Z_{ij}$ for $Z_{ij} \sim N(0,1)$.

  Fix a row t with $1 \leq t \leq k$ of A and consider the dot product $Z_t := \langle A_t, \mathbf{x}\rangle$. This is simply $\sum_{j=1}^{d} \mathbf{x}_j A_{tj}$ and forms the tth entry of $A\mathbf{x}$. By Fact 1, we see that $Z_t \sim N(0, \frac{1}{k}\|\mathbf{x}\|_2^2)$. Therefore, we get

$$\|A\mathbf{x}\|_2^2 = \sum_{t=1}^{k} Z_i^2 \quad \text{where each} \quad Z_i \sim N(0, \frac{\|\mathbf{x}\|_2^2}{k})$$

  By Fact 2, we get that $\mathbf{Exp}[\|A\mathbf{x}\|_2^2] = \sum_{t=1}^{k} \frac{\|\mathbf{x}\|_2^2}{k} = \|\mathbf{x}\|_2^2$, and

$$\mathbf{Pr}\left[\, \|A\mathbf{x}\|_2^2 \notin (1 \pm \varepsilon) \cdot \|\mathbf{x}\|_2^2 \,\right] \leq 2e^{-\varepsilon^2 k/8}$$

  (3) follows by noting that $\mathbf{Pr}_A\left[\, \|A\mathbf{x}\|_2 \notin (1 \pm \varepsilon) \cdot \|\mathbf{x}\|_2 \,\right] \leq \mathbf{Pr}\left[\, \|A\mathbf{x}\|_2^2 \notin (1 \pm 2\varepsilon) \cdot \|\mathbf{x}\|_2^2 \,\right]$

- *Finishing up with union bound.* Now one considers the collection $H := \{(\mathbf{v}_i - \mathbf{v}_j) \,:\, 1 \leq i < j \leq n\}$ of $\binom{n}{2}$ vectors and applies union bound on H. More precisely,

$$\mathbf{Pr}[\exists \mathbf{u} \in H \,:\, \|A\mathbf{u}\|_2 \notin (1 \pm \varepsilon)\,\|\mathbf{u}\|_2] \,\leq 2\binom{n}{2} \cdot e^{-\varepsilon^2 k/2}$$

  If $k = \frac{6 \ln n}{\varepsilon^2}$, then the above RHS is $\leq 1/n$. Therefore, with probability $\geq 1 - 1/n$, we get that for every $1 \leq i < j \leq n$, $\|A(\mathbf{v}_i - \mathbf{v}_j)\|_2 = \|A\mathbf{v}_i - A\mathbf{v}_j\|_2 \in (1 \pm \varepsilon)\,\|\mathbf{v}_i - \mathbf{v}_j\|_2$, proving Theorem 1.

[3] If X is a sum of "nice" independent random variables, then $\mathbf{Pr}[|X - \mathbf{Exp}[X]| \geq c\sigma_X] \leq e^{-\Theta(c^2)}$. This is a vague statement since "nice" is not defined, and hence only a rule of thumb. Read the first three paras of these beautiful notes.