University of Edinburgh

INFR11156: Algorithmic Foundations of Data Science (2025)

Lecture 2: Johnson-Lindenstrauss Lemma

In today's lecture, we study the Johnson-Lindenstrauss lemma, which states that any n points in high dimensional Euclean space can be mapped onto k dimensions where $k = O(\log n/\varepsilon^2)$ without distorting the Euclidean distance between any pair of points more than a factor of $1 \pm \varepsilon$.

Lemma 1 (Johnson-Lindenstrauss, 1984)**.** *Let $X \subseteq \mathbb{R}^d$ be a set of n points, $\varepsilon \in (0, 1/5)$. Then, there is a random matrix $\Phi \in \mathbb{R}^{k\times d}$, such that it holds with constant probability that*

$$\forall x, y \in X : \qquad (1-\varepsilon)\|x-y\|_2 \leq \|\Phi x - \Phi y\|_2 \leq (1+\varepsilon)\|x-y\|_2, \tag{1}$$

where $k = O\left(\frac{\log n}{\varepsilon^2}\right)$.

Remark:

- The statement above holds for *all pair* of points, instead of most pairs of points.
- The number of dimensions in the projection is only a logarithmic function of n, and independent of d. Since k is usually much less than d, we sometimes call this lemma dimension reduction lemma. In applications, the dominant term is typically the $1/\varepsilon^2$ term.
- The matrix Φ is independent of the input points.
- The number of dimensions needed is shown to be optimal. It is known that there is a set of points in $\mathbb{R}^d$ such that, in order to have (1), $k = \Omega\left(\frac{\log n}{\varepsilon^2}\right)$.

The key to prove the Johnson-Lindenstrauss lemma is the following technical lemma.

Lemma 2. *Given the same hypothesis, there exists a matrix $\Phi \in \mathbb{R}^{k\times d}$ such that it holds for any $x \in \mathbb{R}^d$ that*

$$\mathbf{Pr}\left[\, \|\Phi x\|_2 \leq (1-\varepsilon)\|x\|_2 \text{ or } \|\Phi x\|_2 \geq (1+\varepsilon)\|x\|_2 \,\right] \leq 2\cdot \mathrm{e}^{-k\cdot\varepsilon^2/5}.$$

Proof of Lemma 1. For any $x, y \in X$ we define $z_{x,y} = x - y$. We apply Lemma 2 on all possible $z_{x,y}$. Hence, using the union bound the total "failure" probability is at most

$$\frac{n(n-1)}{2}\cdot 2\cdot \mathrm{e}^{-k\cdot\varepsilon^2/5},$$

which is a constant if $k = O\left(\frac{\log n}{\varepsilon^2}\right)$. □

We list several facts about the normal distributions that will be used in our proof.

Fact 3. *The following statements hold:*

1. *If $X_i \sim N(\mu_i, \sigma_i^2)$ and $a_i \in \mathbb{R}$ for any $1 \leq i \leq n$, then it holds that*

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n}(a_i\sigma_i)^2\right).$$

2. If $X_1, \ldots, X_k$ are independent, standard normal random variables, then the sum of their squares

$$Q = \sum_{i=1}^{k} X_i^2$$

is distributed according to the χ^2 distribution with k degree of freedom, denoted as $Q \sim \chi^2(k)$.

3. The moment generating function of a random variable is $M_x(t) = \mathbf{E}\left[\mathrm{e}^{tX}\right]$ for $t \in \mathbb{R}$. If $X \sim \chi^2(k)$, then it holds that $\mathbf{E}\left[\mathrm{e}^{tX}\right] = (1-2t)^{-k/2}$.

Proof of Lemma 2. Let us define Φ as a matrix

$$\Phi = \frac{1}{\sqrt{k}} \begin{bmatrix} g_{11} & g_{12} & g_{13} & \cdots & g_{1d} \\ g_{21} & g_{22} & g_{23} & \cdots & g_{2d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_{k1} & g_{k2} & g_{k3} & \cdots & g_{kd} \end{bmatrix},$$

where every $g_{i,j} \sim N(0,1)$. Let $x \in \mathbb{R}^d$ be an arbitrary vector, and we assume without loss of generality that $\|x\| = 1$. We define $y = \Phi \cdot x$. By definition, we have for each $1 \leq i \leq k$ that

$$y_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} g_{i,j} x_j.$$

We apply Fact 3 and obtain that

$$y_i \sim N\left(0, \frac{1}{k}\sum_{j=1}^{d} x_j^2\right),$$

i.e. $y_i \sim N(0, 1/k)$ due to the fact that $\|x\| = 1$. This gives us that $\sqrt{k} y_i \sim N(0,1)$, and therefore

$$\sum_{i=1}^{k} \left(\sqrt{k} y_i\right)^2 = k \sum_{i=1}^{k} y_i^2 \sim \chi^2(k).$$

For ease of analysis we introduce $h_1, \ldots, h_k$, which are independent and identically distributed random variables such that

$$\sum_{i=1}^{k} h_i^2 = k \sum_{i=1}^{k} y_i^2. \tag{2}$$

Hence, it holds that

$$\mathbf{Pr}\left[\|\Phi x\| \geq 1+\varepsilon\right] \leq \mathbf{Pr}\left[\|\Phi x\|^2 \geq 1+\varepsilon\right] = \mathbf{Pr}\left[\sum_{i=1}^{k} y_i^2 \geq 1+\varepsilon\right].$$

By (2) we have for any $\lambda > 0$ that

$$\mathbf{Pr}\left[\|\Phi x\| \geq 1+\varepsilon\right] \leq \mathbf{Pr}\left[\sum_{i=1}^{k} h_i^2 \geq (1+\varepsilon)\cdot k\right] = \mathbf{Pr}\left[\mathrm{e}^{\lambda\cdot\sum_{i=1}^{k} h_i^2} \geq \mathrm{e}^{\lambda(1+\varepsilon)\cdot k}\right].$$

Since the h_i's are independent to each other, we apply Markov's inequality and obtain

$$\mathbf{Pr}\left[\mathrm{e}^{\lambda\cdot\sum_{i=1}^{k} h_i^2} \geq \mathrm{e}^{\lambda(1+\varepsilon)\cdot k}\right] \leq \frac{\mathbf{E}\left[\mathrm{e}^{\lambda\cdot\sum_{i=1}^{k} h_i^2}\right]}{\mathrm{e}^{\lambda(1+\varepsilon)\cdot k}} = \frac{\prod_{i=1}^{k}\mathbf{E}\left[\mathrm{e}^{\lambda\cdot h_i^2}\right]}{\mathrm{e}^{\lambda(1+\varepsilon)\cdot k}}.$$

Since $h_i \sim N(0,1)$ and $\mathbf{E}\left[\mathrm{e}^{\lambda h_i^2}\right] = (1-2t)^{-1/2}$ by using the moment generating function of χ^2 distributions, we have

$$\mathbf{Pr}\left[\mathrm{e}^{\lambda\cdot\sum_{i=1}^{k} h_i^2} \geq \mathrm{e}^{\lambda(1+\varepsilon)\cdot k}\right] \leq \frac{\left(1/\sqrt{1-2\lambda}\right)^k}{\mathrm{e}^{\lambda(1+\varepsilon)\cdot k}} = \frac{\mathrm{e}^{-\frac{k}{2}\cdot\log(1-2\lambda)}}{\mathrm{e}^{\lambda(1+\varepsilon)\cdot k}}.$$

Since $\log(1-x) \geq -x - x^2/2 - x^3/2$ for $x \in (0, 1/5)$, we assume $\lambda \leq 1/10$ and have

$$\mathbf{Pr}\left[\mathrm{e}^{\lambda \cdot \sum_{i=1}^{k} h_i^2} \geq \mathrm{e}^{\lambda(1+\varepsilon)\cdot k} \right] \leq \frac{\mathrm{e}^{\frac{k}{2}\cdot(2\lambda+2\lambda^2+4\lambda^3)}}{\mathrm{e}^{\lambda(1+\varepsilon)\cdot k}} \leq \mathrm{e}^{-k\varepsilon^2/5}$$

by setting $\lambda = \varepsilon/2$. Combining all the calculations above gives us that

$$\mathbf{Pr}\left[\, \|\Phi x\| \geq 1+\varepsilon \,\right] \leq \mathrm{e}^{-k\varepsilon^2/5}.$$

By the symmetry of random variables y_i's and the union bound, we have

$$\mathbf{Pr}\left[\, \|\Phi x\|_2 \leq (1-\varepsilon)\|x\|_2 \text{ or } \|\Phi x\|_2 \geq (1+\varepsilon)\|x\|_2 \,\right] \leq 2 \cdot \mathrm{e}^{-k\varepsilon^2/5},$$

which finishes the proof. □