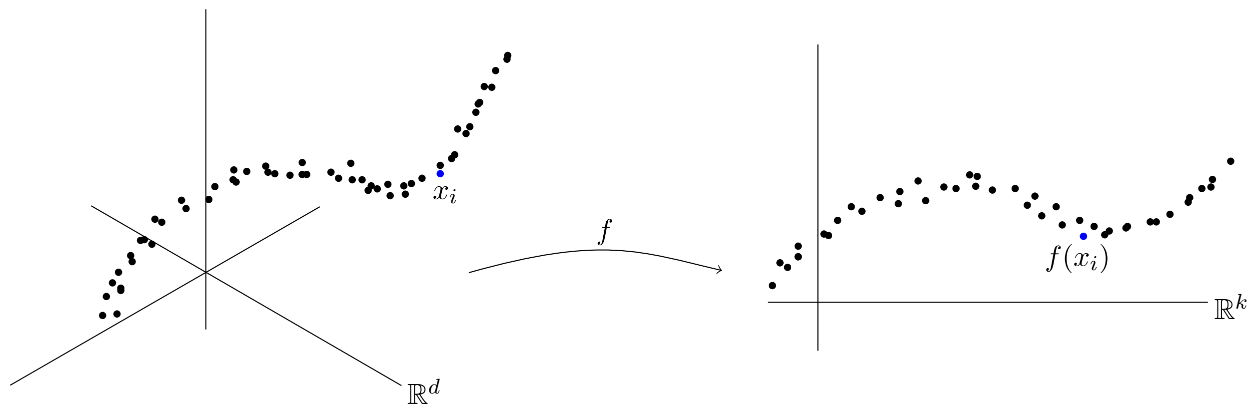
Johnson-Lindenstrauss Theory

Let $Q = \{x_i\}_{i=1}^n$ be a set of points in $\mathbb{R}^d$. Think of d as being large, so that tree-based methods may not perform well. Suppose we are interested in analyzing the geometry of the set Q. For example, we could be interested in nearest-neighbors search, finding low-dimensional structure, etc.



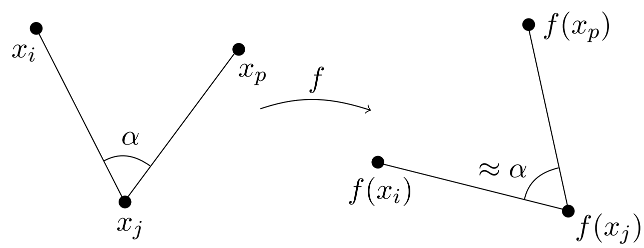
It is natural to look for a map $f : \mathbb{R}^d \to \mathbb{R}^k$ that maps the points to $\mathbb{R}^k$, where $k < d$. Desirable properties of f include:

- nice continuity (J-L gives a linear, Lipschitz map)
- We would like k to be reasonably small (J-L gives $k \sim \log n$ independent of d)
- We want to approximately preserve pairwise distances:

$$\|x_i - x_j\| \approx \|f(x_i) - f(x_j)\| \quad \forall x_i, x_j \in Q.$$

- We want to approximately preserve angles:

$$\langle x_i - x_j, x_p - x_j\rangle \approx \langle f(x_i) - f(x_j), f(x_p) - f(x_j)\rangle \quad \forall x_i, x_p, x_j \in Q.$$



The Johnson-Lindenstrauss theorem asserts that there exists a linear map f and that image dimension k will scale as $\log n$ with no dependence on the original dimension d. From a practical perspective, we often choose f as a random projection (e.g. a "short fat matrix").

1. Brief review of basic probability

Let us briefly review basic probability and introduce our notation. Let $X \in \mathbb{R}$ be a random variable with probability density function p. The **mean** of X is

$$\mu = \mathbb{E}[X] = \int_{\mathbb{R}} x p(x) \mathrm{d}x.$$

The **variance** of X is

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \int_{\mathbb{R}} (x-\mu)^2 p(x) \mathrm{d}x.$$

Example 1. Let $X \sim \mathcal{N}(0,1)$ be sampled from the standard normal distribution. We have $p(x) = (2\pi)^{-1/2}e^{-x^2/2}$. Using a symmetry argument,

$$\mu = \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = 0.$$

With a bit more work, one can show

$$\sigma^2 = \int_{\mathbb{R}} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = 1.$$

Example 2. Let

$$\mathbf{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

be a 2×2 random matrix where the entries a, b, c, d are independent and have mean 0 and variance 1. Fix $x \in \mathbb{R}^2$; note that x is not a random variable, but an arbitrary vector. Set $y = \mathbf{A}x$; note that y is a random variable. Let's compute $\mathbb{E}[\|y\|^2]$. We have

$$\|y\|^2 = y_1^2 + y_2^2 = \frac{1}{2}(ax_1 + bx_2)^2 + \frac{1}{2}(cx_1 + dx_2)^2 .$$

Observe that y_1 is independent from y_2, since x is fixed and the entries of $\mathbf{A}$ are independent. Expectation is linear, so we may write

$$\mathbb{E}\left[\|y\|^2\right] = \mathbb{E}[y_1^2] + \mathbb{E}[y_2^2].$$

Observe that all the entries of $\mathbf{A}$ are independent with mean 0 and variance 1. Therefore we have

$$\begin{aligned} \mathbb{E}[y_1^2] &= \frac{1}{2}\mathbb{E}[a^2x_1^2 + 2abx_1x_2 + b^2x_2^2] = \frac{1}{2}x_1^2\mathbb{E}[a^2] + x_1x_2\mathbb{E}[ab] + \frac{1}{2}x_x^2\mathbb{E}[b^2] \\ &= \frac{1}{2}x_1^2 + x_1x_2\mathbb{E}[a]\mathbb{E}[b] + \frac{1}{2}x_2^2 = \frac{1}{2}x_1^2 + 0 + \frac{1}{2}x_2^2 \\ &= \frac{1}{2}\|x\|^2. \end{aligned}$$

Analogously, we know $\mathbb{E}[y_2^2] = \|x\|^2/2$, and so we have $\mathbb{E}[\|y\|^2] = \|x\|^2$. In other words, the expected value of $\|y\|^2$ is $\|x\|^2$, so the random matrix $\mathbf{A}$ preserves the squared 2-norm in expected value. Note that we do not yet know anything about the variance of $\|y\|^2$.

The above example generalizes to the case of a random $k \times d$ matrix.

Theorem 1. *Let $\mathbf{A}$ be a $k \times d$ random matrix with entries that are independent and have mean 0 and variance 1. Given $x \in \mathbb{R}^d$, set $y = \frac{1}{\sqrt{k}}\mathbf{A}x$. Then $\mathbb{E}[\|y\|^2] = \|x\|^2$.*

Proof. Due to linearity,

$$\mathbb{E}[\|y\|^2] = \sum_{i=1}^{k} \mathbb{E}[y_i^2].$$

Following the example, we have

$$\mathbb{E}[y_i^2] = \frac{1}{k}\mathbb{E}\left[\left(\sum_{j=1}^{d} a_{ij}x_j\right)\right] = \frac{1}{k}\mathbb{E}\left[\sum_{j,p=1}^{d} a_{ij}x_j a_{ip}x_p\right].$$

Since the entries a_{ij} are independent, mean 0, and variance 1, we know $\mathbb{E}[a_{ij}a_{ip}] = \delta_{jp}$, where δ_{jp} is the Kronecker delta ($\delta_{jp} = 1$ if $j = p$ and 0 otherwise). Using this to simplify the double sum, we have

$$\mathbb{E}[y_i]^2 = \frac{1}{k}\mathbb{E}\left[\sum_{j=1}^{d} a_{ij}^2 x_j^2\right] = \frac{1}{k}\sum_{j=1}^{d} x_j^2 \mathbb{E}[a_{ij}^2] = \frac{1}{k}\|x\|^2.$$

Combining everything, we have the desired result. □

There are many matrices that satisfy the conditions of Theorem 1. For instance, if the entries a_{ij} are sampled independently from $N(0,1)$, or ± 1 with probability $1/2$ (Bernoulli variables), then the conditions are satisfied. Again note that we know the expected value of $\|y\|^2$, but nothing about the variance (which could be unreasonably large).

Theorem 2. *Let* A *be a* $k \times d$ *random matrix with entries sampled independently from* $N(0,1)$*. Fix* $\epsilon \in (0, 1/2)$*. Then*

$$(1-\epsilon)\|x\|^2 \le \|y\|^2 \le (1+\epsilon)\|x\|^2$$

with probability at least $1 - 2e^{-(\epsilon^2 - \epsilon^3)k/4}$.

Proof. Set $z = \frac{\sqrt{k}}{\|x\|} y$. Then

$$z_i = \frac{\sqrt{k}}{\|x\|^2} y_i = \frac{1}{\|x\|}\sum_{j=1}^{d} a_{ij}x_j.$$

Notice that z_i is a Gaussian random variable, since it is a linear combination of Gaussian random variables. We easily compute

$$\mathbb{E}[z_i] = \frac{1}{\|x\|}\sum_{j=1}^{d} x_j \mathbb{E}[a_{ij}] = 0.$$

In the proof of Theorem 1, we showed $\mathbb{E}[z_i]^2 = 1$. Thus, since Gaussian random variables are completely determined by their mean and (co)variance, each $z_i \sim N(0,1)$, and they are all independent.

The proof will continue next time. After some algebra, we find

$$\text{Prob}\left[\|y\|^2 > (1+\epsilon)\|x\|^2\right] = \text{Prob}\left[\sum_{i=1}^{k} z_i^2 > (1+\epsilon)k\right].$$

□