Chapter 23

Johnson-Lindenstrauss Lemma

By Sariel Har-Peled, April 6, 2023[1]

> Dixon was alive again. Consciousness was upon him before he could get out of the way; not for him the slow, gracious wandering from the halls of sleep, but a summary, forcible ejection. He lay sprawled, too wicked to move, spewed up like a broken spider-crab on the tarry shingle of the morning. The light did him harm, but not as much as looking at things did; he resolved, having done it once, never to move his eyeballs again. A dusty thudding in his head made the scene before him beat like a pulse. His mouth had been used as a latrine by some small creature of the night, and then as its mausoleum. During the night, too, he'd somehow been on a cross-country run and then been expertly beaten up by secret police. He felt bad.
>
> Lucky Jim, Kingsley Amis

In this chapter, we will prove that given a set P of n points in $\mathbb{R}^{\mathsf{d}}$, one can reduce the dimension of the points to $k = O(\varepsilon^{-2} \log n)$ such that distances are $1 \pm \varepsilon$ preserved. Surprisingly, this reduction is done by randomly picking a subspace of k dimensions and projecting the points into this random subspace. One way of thinking about this result is that we are "compressing" the input of size $n\mathsf{d}$ (i.e., n points with d coordinates) into size $O(n\varepsilon^{-2} \log n)$, while (approximately) preserving distances.

23.1. Measure concentration on the sphere

Let $\mathbb{S}^{(n-1)}$ be the unit sphere in $\mathbb{R}^n$. We assume there is a uniform probability measure defined over $\mathbb{S}^{(n-1)}$, such that its total measure is 1. Surprisingly, most of the mass of this measure is near the equator. Indeed, consider an arbitrary equator π on $\mathbb{S}^{(n-1)}$ (that it, it is the intersection of the sphere with a hyperplane passing through the center of the ball inducing the sphere). Next, consider all the points that are within distance $\approx \ell(n) = c/n^{1/3}$ from π. The question we are interested in is what fraction of the sphere is covered by this strip T (depicted in Figure 23.1.1).

Notice that as the dimension increases, the width $\ell(n)$ of this strip decreases. But surprisingly, despite its width becoming smaller, as the dimension increases, this strip contains a larger and larger fraction of the sphere. In particular, the total fraction of the sphere not covered by this (shrinking!) strip converges to zero.

Furthermore, counterintuitively, this is true for *any* equator. We are going to show that even a stronger result holds: The mass of the sphere is concentrated close to the boundary of any set $A \subseteq \mathbb{S}^{(n-1)}$ such that $\mathbb{P}[A] = 1/2$.

Before proving this somewhat surprising theorem, we will first try to get an intuition about the behavior of the hypersphere in high dimensions.

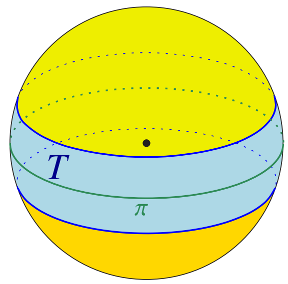


Figure 23.1.1

23.1.1. The strange and curious life of the hypersphere

Consider the ball of radius r in $\mathbb{R}^n$ (denoted by $r\,\mathsf{b}^n$), where b^n is the unit radius ball centered at the origin. Clearly, $\mathrm{Vol}(r\,\mathsf{b}^n) = r^n\,\mathrm{Vol}(\mathsf{b}^n)$. Now, even if r is very close to 1, the quantity r^n might be very close to zero if n is sufficiently large. Indeed, if $r = 1-\delta$, then $r^n = (1-\delta)^n \leq \exp(-\delta n)$, which is very small if $\delta \gg 1/n$. (Here, we used $1 - x \leq e^x$, for $x \geq 0$.) Namely, for the ball in high dimensions, its mass is concentrated in a very thin shell close to its surface.

The volume of a ball and the surface area of a hypersphere. Let $\mathrm{Vol}(r\mathsf{b}^n)$ denote the volume of the ball of radius r in $\mathbb{R}^n$, and let $\mathsf{S}(r\mathsf{b}^n)$ denote the surface area of its bounding sphere (i.e., the surface area of $r\mathbb{S}^{(n-1)}$). It is known that

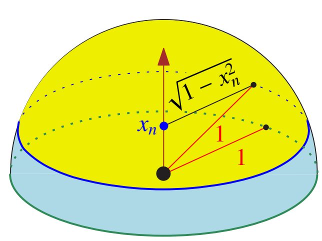
$$\mathrm{Vol}(r\mathsf{b}^n) = \frac{\pi^{n/2} r^n}{\Gamma(n/2+1)} \quad \text{and} \quad \mathsf{S}(r\mathsf{b}^n) = \frac{2\pi^{n/2} r^{n-1}}{\Gamma(n/2)},$$

where the gamma function, $\Gamma(\cdot)$, is an extension of the factorial function. Specifically, if n is even, then $\Gamma(n/2+1) = (n/2)!$, and for n odd $\Gamma(n/2+1) = \sqrt{\pi}(n!!)/2^{(n+1)/2}$, where $n!! = 1 \cdot 3 \cdot 5 \cdots n$ is the ***double factorial***. The most surprising implication of these two formulas is that, as n increases, the volume of the unit ball first increases (till dimension 5) and then starts decreasing to zero.

Similarly, the surface area of the unit sphere $\mathbb{S}^{(n-1)}$ in $\mathbb{R}^n$ tends to zero as the dimension increases. To see the above explicitly, compute the volume of the unit ball using an integral of its slice volume, when it is being sliced by a hyperplane perpendicular to the nth coordinate.

We have (see figure on the right) that

$$\mathrm{Vol}(\mathsf{b}^n) = \int_{x_n=-1}^{1} \mathrm{Vol}\Big(\sqrt{1-x_n^2}\,\mathsf{b}^{n-1}\Big) dx_n = \mathrm{Vol}\Big(\mathsf{b}^{n-1}\Big) \int_{x_n=-1}^{1} \Big(1-x_n^2\Big)^{(n-1)/2} dx_n.$$



Now, the integral on the right side tends to zero as n increases. In fact, for n very large, the term $(1-x_n^2)^{(n-1)/2}$ is very close to 0 everywhere except for a small interval around 0. This implies that the main contribution of the volume of the ball happens when we consider slices of the ball created by hyperplanes of the form $x_n = \delta$, where δ is small (roughly, for $|\delta| \leq 1/\sqrt{n}$).

If one has to visualize what such a ball in high dimensions looks like, it might be best to think about it as a star-like creature: It has very little mass close to the tips of any set of orthogonal directions we pick, and most of its mass (somehow) lies on hyperplanes passing close to its center.[2]

23.1.1.1. The stranger and curiouser life of the simplex in high dimensions

A ***regular*** n***-simplex*** Δ_n is a set of $n+1$ points in Euclidean space such that every pair of points is in distance exactly 1 from each other.

Corollary 23.1.1. *The width of the regular* n*-simplex* Δ_n *is* $\sqrt{2/(n+1)}$ *if* n *is odd. The width is* $\sqrt{\frac{2(n+1)}{n(n+2)}}$ *if* n *is even. The inradius (i.e., radius of largest ball inside* Δ_n*) is* $r_n = 1/\sqrt{2n(n+1)}$*, and the circumradius (i.e., radius of minimum ball enclosing* Δ_n *is* $R_n = \sqrt{\frac{n}{2(n+1)}}$.

23.1.2. Measure concentration on the sphere

Theorem 23.1.2 (Measure concentration on the sphere). *Let* $A \subseteq \mathbb{S}^{(n-1)}$ *be a measurable set with* $\mathbb{P}[A] \geq 1/2$*, and let* A_t *denote the set of points of* $\mathbb{S}^{(n-1)}$ *within a distance at most* t *from* A*, where* $t \leq 2$*. Then* $1 - \mathbb{P}[A_t] \leq 2\exp(-nt^2/2)$*.*

Proof: We will prove a slightly weaker bound, with $-nt^2/4$ in the exponent. Let $\widehat{A} = T(A)$, where

$$T(X) = \left\{ \alpha x \;\middle|\; x \in X, \alpha \in [0,1] \right\} \subseteq \mathsf{b}^n$$

and b^n is the unit ball in $\mathbb{R}^n$. We have that $\mathbb{P}[A] = \mu\left(\widehat{A}\right)$, where $\mu\left(\widehat{A}\right) = \mathrm{Vol}\left(\widehat{A}\right)/\mathrm{Vol}(\mathsf{b}^n)$[3].

Let $B = \mathbb{S}^{(n-1)} \setminus A_t$ and $\widehat{B} = T(B)$. We have that $\|a - b\| \geq t$ for all $a \in A$ and $b \in B$. By Lemma 23.1.3 below, the set $\left(\widehat{A} + \widehat{B}\right)/2$ is contained in the ball $r\mathsf{b}^n$ centered at the origin, where $r = 1 - t^2/8$. Observe that $\mu(r\mathsf{b}^n) = \mathrm{Vol}(r\mathsf{b}^n)/\mathrm{Vol}(\mathsf{b}^n) = r^n = \left(1 - t^2/8\right)^n$. As such, applying the Brunn-Minkowski inequality in the form of Corollary 23.7.1, we have

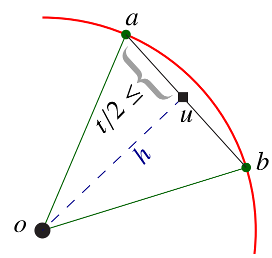
$$\left(1 - \frac{t^2}{8}\right)^n = \mu(r\mathsf{b}^n) \geq \mu\left(\frac{\widehat{A} + \widehat{B}}{2}\right) \geq \sqrt{\mu\left(\widehat{A}\right)\mu\left(\widehat{B}\right)} = \sqrt{\mathbb{P}[A]\,\mathbb{P}[B]} \geq \sqrt{\mathbb{P}[B]/2}.$$

Thus, $\mathbb{P}[B] \leq 2(1 - t^2/8)^{2n} \leq 2\exp(-2nt^2/8)$, since $1 - x \leq \exp(-x)$, for $x \geq 0$. ∎

Lemma 23.1.3. *For any* $\widehat{a} \in \widehat{A}$ *and* $\widehat{b} \in \widehat{B}$*, we have* $\left\| \dfrac{\widehat{a} + \widehat{b}}{2} \right\| \leq 1 - \dfrac{t^2}{8}$.

[2] In short, it looks like a Boojum [Car76].

[3] This is one of these "trivial" claims that might give the reader a pause, so here is a formal proof. Pick a random point p uniformly inside the ball b^n. Let ψ be the probability that $p \in \widehat{A}$. Clearly, $\mathrm{Vol}\left(\widehat{A}\right) = \psi\,\mathrm{Vol}(\mathsf{b}^n)$. So, consider the normalized point $q = p/\|p\|$. Clearly, $p \in \widehat{A}$ if and only if $q \in A$, by the definition of $\widehat{A}$. Thus, $\mu\left(\widehat{A}\right) = \mathrm{Vol}\left(\widehat{A}\right)/\mathrm{Vol}(\mathsf{b}^n) = \psi = \mathbb{P}\left[p \in \widehat{A}\right] = \mathbb{P}[q \in A] = \mathbb{P}[A]$, since q has a uniform distribution on the hypersphere by assumption.



Proof: Let $\widehat{a} = \alpha a$ and $\widehat{b} = \beta b$, where $a \in A$ and $b \in B$. We have

$$\|u\| = \left\| \frac{a+b}{2} \right\| = \sqrt{1^2 - \left\| \frac{a-b}{2} \right\|^2} \leq \sqrt{1 - \frac{t^2}{4}} \leq 1 - \frac{t^2}{8}, \tag{23.1.1}$$

since $\|a - b\| \geq t$. As for $\widehat{a}$ and $\widehat{b}$, assume that $\alpha \leq \beta$, and observe that the quantity $\left\| \widehat{a} + \widehat{b} \right\|$ is maximized when $\beta = 1$. As such, by the triangle inequality, we have

$$\left\| \frac{\widehat{a} + \widehat{b}}{2} \right\| = \left\| \frac{\alpha a + b}{2} \right\| \leq \left\| \frac{\alpha(a+b)}{2} \right\| + \left\| (1-\alpha)\frac{b}{2} \right\| \leq \alpha\left(1 - \frac{t^2}{8}\right) + (1-\alpha)\frac{1}{2} = \tau,$$

by Eq. (23.1.1) and since $\|b\| = 1$. Now, τ is a convex combination of the two numbers $1/2$ and $1 - t^2/8$. In particular, we conclude that $\tau \leq \max(1/2, 1 - t^2/8) \leq 1 - t^2/8$, since $t \leq 2$. ■

23.2. Concentration of Lipschitz functions

Consider a function $f : \mathbb{S}^{(n-1)} \to \mathbb{R}$, and imagine that we have a probability density function defined over the sphere. Let $\mathbb{P}[f \leq t] = \mathbb{P}\Big[\Big\{ x \in S^{n-1} \,\Big|\, f(x) \leq t \Big\}\Big]$. We define the ***median*** of f, denoted by $\mathrm{med}(f)$, to be $\sup t$, such that $\mathbb{P}[f \leq t] \leq 1/2$.

We define $\mathbb{P}[f < \mathrm{med}(f)] = \sup_{x < \mathrm{med}(f)} \mathbb{P}[f \leq x]$. The following is obvious but requires a formal proof.

Lemma 23.2.1. *We have* $\mathbb{P}[f < \mathrm{med}(f)] \leq 1/2$ *and* $\mathbb{P}[f > \mathrm{med}(f)] \leq 1/2$.

Proof: Since $\bigcup_{k \geq 1} (-\infty, \mathrm{med}(f) - 1/k] = (-\infty, \mathrm{med}(f))$, we have

$$\mathbb{P}[f < \mathrm{med}(f)] = \sup_{k \geq 1} \mathbb{P}\left[f \leq \mathrm{med}(f) - \frac{1}{k} \right] \leq \sup_{k \geq 1} \frac{1}{2} = \frac{1}{2}.$$

The second claim follows by a symmetric argument. ■

Definition 23.2.2 (c-***Lipschitz***). A function $f : A \to B$ is c-***Lipschitz*** if, for any $x, y \in A$, we have that $\|f(x) - f(y)\| \leq \mathsf{c}\,\|x - y\|$.

Theorem 23.2.3 (Lévy's lemma). *Let $f : \mathbb{S}^{(n-1)} \to \mathbb{R}$ be 1-Lipschitz. Then for all $t \in [0,1]$, we have*

$$\mathbb{P}[f > \mathrm{med}(f) + t] \leq 2\exp\!\left(-t^2 n/2\right) \quad \textit{and} \quad \mathbb{P}[f < \mathrm{med}(f) - t] \leq 2\exp\!\left(-t^2 n/2\right).$$

Proof: We prove only the first inequality; the second follows by symmetry. Let

$$A = \left\{ x \in \mathbb{S}^{(n-1)} \,\middle|\, f(x) \le \mathrm{med}(f) \right\}.$$

By Lemma 23.2.1, we have $\mathbb{P}[A] \ge 1/2$. Consider a point $x \in A_t$, where A_t is as defined in Theorem 23.1.2. Let $\mathsf{nn}(x)$ be the nearest point in A to x. We have by definition that $\|x\mathsf{nn}(x)\| \le t$. As such, since f is 1-Lipschitz and $\mathsf{nn}(x) \in A$, we have that

$$f(x) \le f(\mathsf{nn}(x)) + \|\mathsf{nn}(x)x\| \le \mathrm{med}(f) + t.$$

Thus, by Theorem 23.1.2, we get $\mathbb{P}[f > \mathrm{med}(f) + t] \le 1 - \mathbb{P}[A_t] \le 2\exp(-t^2 n/2)$. ■

23.3. The Johnson-Lindenstrauss lemma

Lemma 23.3.1. *For a unit vector $x \in \mathbb{S}^{(n-1)}$, let*

$$f(x) = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2}$$

be the length of the projection of x into the subspace formed by the first k coordinates. Let x be a vector randomly chosen with uniform distribution from $\mathbb{S}^{(n-1)}$. Then $f(x)$ is sharply concentrated. Namely, there exists $m = m(n,k)$ such that

$$\mathbb{P}[f(x) \ge m + t] \le 2\exp(-t^2 n/2) \quad \textit{and} \quad \mathbb{P}[f(x) \le m - t] \le 2\exp(-t^2 n/2),$$

for any $t \in [0,1]$. Furthermore, for $k \ge 10 \ln n$, we have $m \ge \frac{1}{2}\sqrt{k/n}$.

Proof: The orthogonal projection $p : \mathbb{R}^n \to \mathbb{R}^k$ given by $p(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$ is 1-Lipschitz (since projections can only shrink distances; see Exercise 23.6.4). As such, $f(x) = \|p(x)\|$ is 1-Lipschitz, since for any x, y we have

$$|f(x) - f(y)| = \Big| \|p(x)\| - \|p(y)\| \Big| \le \|p(x) - p(y)\| \le \|x - y\|,$$

by the triangle inequality and since p is 1-Lipschitz. Theorem 23.2.3 (i.e., Lévy's lemma) gives the required tail estimate with $m = \mathrm{med}(f)$.

Thus, we only need to prove the lower bound on m. For a random $x = (x_1, \ldots, x_n) \in \mathbb{S}^{(n-1)}$, we have $\mathbb{E}\Big[\|x\|^2\Big] = 1$. By linearity of expectations and by symmetry, we have $1 = \mathbb{E}\Big[\|x\|^2\Big] = \mathbb{E}[\sum_{i=1}^n x_i^2] = \sum_{i=1}^n \mathbb{E}[x_i^2] = n\,\mathbb{E}\Big[x_j^2\Big]$, for any $1 \le j \le n$. Thus, $\mathbb{E}\Big[x_j^2\Big] = 1/n$, for $j = 1, \ldots, n$. Thus,

$$\mathbb{E}\Big[(f(x))^2\Big] = \mathbb{E}\left[\sum_{i=1}^k x_i^2\right] = \sum_{i=1}^k \mathbb{E}\Big[x_i^2\Big] = \frac{k}{n},$$

by linearity of expectation.

We next use that f is concentrated to show that f^2 is also relatively concentrated. For any $t \ge 0$, we have

$$\frac{k}{n} = \mathbb{E}\Big[f^2\Big] \le \mathbb{P}\Big[f \le m + t\Big](m+t)^2 + \mathbb{P}\Big[f \ge m + t\Big] \cdot 1 \le 1 \cdot (m+t)^2 + 2\exp(-t^2 n/2),$$

since $f(x) \leq 1$, for any $x \in \mathbb{S}^{(n-1)}$. Let $t = \sqrt{k/5n}$. Since $k \geq 10 \ln n$, we have that $2\exp(-t^2 n/2) \leq 2/n$. We get that

$$\frac{k}{n} \leq \left(m + \sqrt{k/5n}\right)^2 + 2/n,$$

implying that $\sqrt{(k-2)/n} \leq m + \sqrt{k/5n}$, which in turn implies that $m \geq \sqrt{(k-2)/n} - \sqrt{k/5n} \geq \frac{1}{2}\sqrt{k/n}$. ∎

Next, we would like to argue that given a fixed vector, projecting it down into a random k-dimensional subspace results in a random vector such that its length is highly concentrated. This would imply that we can do dimension reduction and still preserve distances between points that we care about.

To this end, we would like to flip Lemma 23.3.1 around. Instead of randomly picking a point and projecting it down to the first k-dimensional space, we would like x to be fixed and randomly pick the k-dimensional subspace we project into. However, we need to pick this random k-dimensional space carefully. Indeed, if we rotate this random subspace, by a transformation T, so that it occupies the first k dimensions, then the point $T(x)$ needs to be uniformly distributed on the hypersphere if we want to use Lemma 23.3.1.

As such, we would like to randomly pick a rotation of $\mathbb{R}^n$. This maps the standard orthonormal basis into a randomly rotated orthonormal space. Taking the subspace spanned by the first k vectors of the rotated basis results in a k-dimensional random subspace. Such a rotation is an orthonormal matrix with determinant 1. We can generate such a matrix, by randomly picking a vector $e_1 \in \mathbb{S}^{(n-1)}$. Next, we set e_1 as the first column of our rotation matrix and generate the other $n-1$ columns, by generating recursively $n-1$ orthonormal vectors in the space orthogonal to e_1.

Remark 23.3.2 (Generating a random point on the sphere). At this point, the reader might wonder how we pick a point uniformly from the unit hypersphere. The idea is to pick a point from the multi-dimensional normal distribution $N^n(0,1)$ and normalize it to have length 1. Since the multi-dimensional normal distribution has the density function

$$(2\pi)^{-n/2} \exp\left(-(x_1^2 + x_2^2 + \cdots + x_n^2)/2\right),$$

which is symmetric (i.e., all the points at distance r from the origin have the same distribution), it follows that this indeed generates a point randomly and uniformly on $\mathbb{S}^{(n-1)}$.

Generating a vector with multi-dimensional normal distribution is no more than picking each coordinate according to the normal distribution; see Lemma 23.5.2. Given a source of random numbers according to the uniform distribution, this can be done using $O(1)$ computations per coordinate, using the Box-Muller transformation [BM58]. Overall, each random vector can be generated in $O(n)$ time.

Since projecting down the n-dimensional normal distribution to the lower-dimensional space yields a normal distribution, it follows that generating a random projection is no more than randomly picking n vectors according to the multi-dimensional normal distribution $v_1, \ldots, v_n$. Then, we orthonormalize them, using Graham-Schmidt, where $\widehat{v_1} = v_1 / \|v_1\|$ and $\widehat{v}_i$ is the normalized vector of $v_i - w_i$, where w_i is the projection of v_i to the space spanned by $v_1, \ldots, v_{i-1}$.

Taking those vectors as columns of a matrix generates a matrix A, with determinant either 1 or -1. We multiply one of the vectors by -1 if the determinant is -1. The resulting matrix is a random rotation matrix.

We can now restate Lemma 23.3.1 in the setting where the vector is fixed and the projection is into a random subspace.

Lemma 23.3.3. *Let $x \in \mathbb{S}^{(n-1)}$ be an arbitrary unit vector. Now, consider a random k-dimensional subspace $\mathcal{F}$, and let $f(x)$ be the length of the projection of x into $\mathcal{F}$. Then, there exists $m = m(n,k)$ such that*

$$\mathbb{P}\Big[f(x) \geq m + t\Big] \leq 2\exp\Big(-t^2 n/2\Big) \quad \text{and} \quad \mathbb{P}\Big[f(x) \leq m - t\Big] \leq 2\exp\Big(-t^2 n/2\Big),$$

for any $t \in [0,1]$. Furthermore, for $k \geq 10 \ln n$, we have $m \geq \frac{1}{2}\sqrt{k/n}$.

Proof: Let v_i be the ith orthonormal vector having 1 at the ith coordinate and 0 everywhere else. Let $\mathbf{M}$ be a random translation of space generated as described above. Clearly, for arbitrary fixed unit vector x, the vector $\mathbf{M}x$ is distributed uniformly on the sphere. Now, the ith column of the matrix $\mathbf{M}$ is the random vector e_i, and $e_i = \mathbf{M}^T v_i$. As such, we have

$$\langle \mathbf{M}x, v_i \rangle = (\mathbf{M}x)^T v_i = x^T \mathbf{M}^T v_i = x^T e_i = \langle x, e_i \rangle .$$

In particular, treating $\mathbf{M}x$ as a random vector and projecting it on the first k coordinates, we have that

$$f(x) = \sqrt{\sum_{i=1}^{k} \langle \mathbf{M}x, v_i \rangle^2} = \sqrt{\sum_{i=1}^{k} \langle x, e_i \rangle^2}.$$

But $e_1, \ldots, e_k$ is just an orthonormal basis of a random k-dimensional subspace. As such, the expression on the right is the length of the projection of x into a k-dimensional random subspace. As such, the length of the projection of x into a random k-dimensional subspace has exactly the same distribution as the length of the projection of a random vector into the first k coordinates. The claim now follows by Lemma 23.3.1. ■

Definition 23.3.4. The mapping $f : \mathbb{R}^n \to \mathbb{R}^k$ is called ***K-bi-Lipschitz*** for a subset $X \subseteq \mathbb{R}^n$ if there exists a constant $\mathsf{c} > 0$ such that

$$\mathsf{c}K^{-1} \cdot \|p - q\| \leq \|f(p) - f(q)\| \leq \mathsf{c} \cdot \|p - q\| ,$$

for all $p, q \in X$.

The least K for which f is K-bi-Lipschitz is called the ***distortion*** of f and is denoted $\mathrm{dist}(f)$. We will refer to f as a ***K-embedding*** of X.

Remark 23.3.5. Let $X \subseteq \mathbb{R}^m$ be a set of n points, where m potentially might be much larger than n. Observe that in this case, since we only care about the inter-point distances of points in X, we can consider X to be a set of points lying in the affine subspace $\mathcal{F}$ spanned by the points of X. Note that this subspace has dimension $n-1$. As such, each point of X can be interpreted as an $(n-1)$-dimensional point in $\mathcal{F}$. Namely, we can assume, for our purposes, that the set of n points in Euclidean space we care about lies in $\mathbb{R}^n$ (i.e., $\mathbb{R}^{n-1}$).

Note that if $m < n$, we can always pad all the coordinates of the points of X by zeros, such that the resulting point set lies in $\mathbb{R}^n$.

Theorem 23.3.6 (Johnson-Lindenstrauss lemma / JLlemma). *Let X be an n-point set in a Euclidean space, and let $\varepsilon \in (0,1]$ be given. Then there exists a $(1+\varepsilon)$-embedding of X into $\mathbb{R}^k$, where $k = O(\varepsilon^{-2} \log n)$.*

Proof: By Remark 23.3.5, we can assume that $X \subseteq \mathbb{R}^n$. Let $k = 200\varepsilon^{-2} \ln n$. Assume $k < n$, and let $\mathcal{F}$ be a random k-dimensional linear subspace of $\mathbb{R}^n$. Let $P_{\mathcal{F}} : \mathbb{R}^n \to \mathcal{F}$ be the orthogonal projection operator of $\mathbb{R}^n$ into $\mathcal{F}$. Let m be the number around which $\|P_{\mathcal{F}}(x)\|$ is concentrated, for $x \in \mathbb{S}^{(n-1)}$, as in Lemma 23.3.3.

Fix two points $x, y \in \mathbb{R}^n$. We prove that

$$\left(1 - \frac{\varepsilon}{3}\right) m \|x - y\| \leq \|P_{\mathcal{F}}(x) - P_{\mathcal{F}}(y)\| \leq \left(1 + \frac{\varepsilon}{3}\right) m \|x - y\|$$

holds with probability $\geq 1 - n^{-2}$. Since there are $\binom{n}{2}$ pairs of points in X, it follows that with constant probability (say $> 1/3$) this holds for all pairs of points of X. In such a case, the mapping p is a D-embedding of X into $\mathbb{R}^k$ with $D \leq \frac{1+\varepsilon/3}{1-\varepsilon/3} \leq 1 + \varepsilon$, for $\varepsilon \leq 1$.

Let $u = x - y$. We have $P_{\mathcal{F}}(u) = P_{\mathcal{F}}(x) - P_{\mathcal{F}}(y)$ since $P_{\mathcal{F}}(\cdot)$ is a linear operator. Thus, the condition becomes $\left(1 - \frac{\varepsilon}{3}\right) m \|u\| \leq \|P_{\mathcal{F}}(u)\| \leq \left(1 + \frac{\varepsilon}{3}\right) m \|u\|$. Again, since projection is a linear operator, for any $\alpha > 0$, the condition is equivalent to

$$\left(1 - \tfrac{\varepsilon}{3}\right) m \|\alpha u\| \leq \|P_{\mathcal{F}}(\alpha u)\| \leq \left(1 + \tfrac{\varepsilon}{3}\right) m \|\alpha u\| .$$

As such, we can assume that $\|u\| = 1$ by picking $\alpha = 1/\|u\|$. Namely, we need to show that

$$\big| \|P_{\mathcal{F}}(u)\| - m \big| \leq \frac{\varepsilon}{3} m.$$

Let $f(u) = \|P_{\mathcal{F}}(u)\|$. By Lemma 23.3.1 (exchanging the random space with the random vector), for $t = \varepsilon m/3$, we have that the probability that this does not hold is bounded by

$$\mathbb{P}[|f(u) - m| \geq t] \leq 4 \exp\left(-\frac{t^2 n}{2}\right) = 4 \exp\left(\frac{-\varepsilon^2 m^2 n}{18}\right) \leq 4 \exp\left(-\frac{\varepsilon^2 k}{72}\right) < n^{-2},$$

since $m \geq \frac{1}{2}\sqrt{k/n}$ and $k = 200\varepsilon^{-2} \ln n$. ∎

23.4. Bibliographical notes

Our presentation follows Matoušek [Mat02]. The Brunn-Minkowski inequality is a powerful inequality which is widely used in mathematics. A nice survey of this inequality and its applications is provided by Gardner [Gar02]. Gardner says, “In a sea of mathematics, the Brunn-Minkowski inequality appears like an octopus, tentacles reaching far and wide, its shape and color changing as it roams from one area to the next.” However, Gardner is careful in claiming that the Brunn-Minkowski inequality is one of the most powerful inequalities in mathematics since, as a wit put it, “The most powerful inequality is $x^2 \geq 0$, since all inequalities are in some sense equivalent to it.”

A striking application of the Brunn-Minkowski inequality is the proof that in any partial ordering of n elements, there is a single comparison that, knowing its result, reduces the number of linear extensions that are consistent with the partial ordering, by a constant fraction. This immediately implies (the uninteresting result) that one can sort n elements in $O(n \log n)$ comparisons. More interestingly, it implies that if there are m linear extensions of the current partial ordering, we can *always* sort it using $O(\log m)$ comparisons. A nice exposition of this surprising result is provided by Matoušek [Mat02, Section 12.3].

There are several alternative proofs of the Johnson-Lindenstrauss lemma (i.e., JLlemma); see [IM98] and [DG03]. Interestingly, it is enough to pick each entry in the dimension-reducing matrix randomly out of $-1, 0, 1$. This requires a more involved proof [Ach01]. This is useful when one cares about storing this dimension reduction transformation efficiently. In particular, recently, there was a flurry of work on making the JLlemma both faster to compute (per point) and sparser. For example, Kane and Nelson [KN14] show that one can compute a matrix for dimension reduction with $O(\varepsilon^{-1} \log n)$ non-zero entries per column (the target dimension is still $O(\varepsilon^{-2} \log n)$). See [KN14] and references therein for more details.

Magen [Mag07] observed that the JL lemma preserves angles, and in fact can be used to preserve any "k-dimensional angle", by projecting down to dimension $O(k\varepsilon^{-2} \log n)$. In particular, Exercise 23.6.5 is taken from there.

Surprisingly, the random embedding preserves much more structure than distances between points. It preserves the structure and distances of surfaces as long as they are low dimensional and "well behaved"; see [AHY07] for some results in this direction.

Dimension reduction is crucial in computational learning, AI, databases, etc. One common technique that is being used in practice is to do PCA (i.e., principal component analysis) and take the first few main axes. Other techniques include independent component analysis and MDS (multi-dimensional scaling). MDS tries to embed points from high dimensions into low dimension ($\mathsf{d} = 2$ or 3), while preserving some properties. Theoretically, dimension reduction into really low dimensions is hopeless, as the distortion in the worst case is $\Omega(n^{1/(k-1)})$, if k is the target dimension [Mat90].

23.5. Normal distribution

The ***normal distribution*** has

$$f(x) = 1/\sqrt{2\pi} \exp\left(-x^2/2\right)$$

as its density function. We next verify that it is a valid density function.

Lemma 23.5.1. *We have $I = \int_{-\infty}^{\infty} f(x)dx = 1$.*

Proof: Observe that

$$\begin{aligned} I^2 &= \left(\int_{x=-\infty}^{\infty} f(x)\, dx \right)^2 = \left(\int_{x=-\infty}^{\infty} f(x)\, dx \right) \left(\int_{y=-\infty}^{\infty} f(y)\, dy \right) \\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f(x) f(y)\, dx\, dy \\ &= \frac{1}{2\pi} \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \exp\left(-\frac{x^2 + y^2}{2} \right) dx\, dy. \end{aligned}$$

Change the variables to $x = r\cos\alpha$, $y = r\sin\alpha$, and observe that the determinant of the Jacobian is

$$J = \det \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \alpha} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \alpha} \end{vmatrix} = \det \begin{vmatrix} \cos\alpha & -r\sin\alpha \\ \sin\alpha & r\cos\alpha \end{vmatrix} = r\left(\cos^2\alpha + \sin^2\alpha\right) = r.$$

As such,

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_{r=0}^{\infty} \int_{\alpha=0}^{2\pi} \exp\left(-\frac{r^2}{2}\right) |J|\; d\alpha\, dr = \frac{1}{2\pi} \int_{r=0}^{\infty} \int_{\alpha=0}^{2\pi} \exp\left(-\frac{r^2}{2}\right) r\, d\alpha\, dr \\ &= \int_{r=0}^{\infty} \exp\left(-\frac{r^2}{2}\right) r\, dr = \left[-\exp\left(-\frac{r^2}{2}\right) \right]_{r=0}^{r=\infty} = -\exp(-\infty) - (-\exp(0)) = 1. \end{aligned}$$

■

The ***multi-dimensional normal distribution***, denoted by $\mathbf{N}^{\mathsf{d}}$, is the distribution in $\mathbb{R}^{\mathsf{d}}$ that assigns a point $p = (p_1, \ldots, p_{\mathsf{d}})$ the density

$$g(p) = \frac{1}{(2\pi)^{\mathsf{d}/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{\mathsf{d}} p_i^2\right).$$

It is easy to verify, using the above, that $\int_{\mathbb{R}^{\mathsf{d}}} g(p)dp = 1$. Furthermore, we have the following useful but easy properties.[④]

Lemma 23.5.2. *(A) The multi-dimensional normal distribution is symmetric; that is, for any two points $p, q \in \mathbb{R}^{\mathsf{d}}$ such that $\|p\| = \|q\|$ we have that $g(p) = g(q)$, where $g(\cdot)$ is the density function of the multi-dimensional normal distribution $\mathbf{N}^{\mathsf{d}}$.*

(B) The projection of the normal distribution on any direction is a one-dimensional normal distribution.

(C) Picking d variables $X_1, \ldots, X_{\mathsf{d}}$ using the one-dimensional normal distribution $\mathbf{N}$ results in a point $(X_1, \ldots, X_{\mathsf{d}})$ that has multi-dimensional normal distribution $\mathbf{N}^{\mathsf{d}}$.

The generalized multi-dimensional distribution is a ***Gaussian***. Fortunately, we only need the simpler notion.

23.6. Exercises

Exercise 23.6.1 (Boxes can be separated). (Easy) Let A and B be two axis parallel boxes that are interior disjoint. Prove that there is always an axis parallel hyperplane that separates the interior of the two boxes.

Exercise 23.6.2 (Brunn-Minkowski inequality, slight extension). Prove the following claim.

> **Corollary 23.6.3.** *For A and B compact sets in $\mathbb{R}^n$, we have for any $\lambda \in [0, 1]$ that $\mathrm{Vol}(\lambda A + (1 - \lambda)B) \geq \mathrm{Vol}(A)^{\lambda}\,\mathrm{Vol}(B)^{1-\lambda}$.*

Exercise 23.6.4 (Projections are contractions). (Easy) Let $\mathcal{F}$ be a k-dimensional affine subspace, and let $P_{\mathcal{F}} : \mathbb{R}^{\mathsf{d}} \to \mathcal{F}$ be the projection that maps every point $x \in \mathbb{R}^{\mathsf{d}}$ to its nearest neighbor on $\mathcal{F}$. Prove that $P_{\mathcal{F}}$ is a contraction (i.e., 1-Lipschitz). Namely, for any $p, q \in \mathbb{R}^{\mathsf{d}}$, we have that $\|P_{\mathcal{F}}(p) - P_{\mathcal{F}}(q)\| \leq \|p - q\|$.

Exercise 23.6.5 (JL lemma works for angles). Show that the Johnson-Lindenstrauss lemma also $(1 \pm \varepsilon)$-preserves angles among triples of points of P (you might need to increase the target dimension however by a constant factor). (**Hint:** For every angle, construct an equilateral triangle whose edges are being preserved by the projection (add the vertices of those triangles (conceptually) to the point set being embedded). Argue that this implies that the angle is being preserved.)

23.7. From previous lectures

Corollary 23.7.1. *For $\mathcal{A}$ and $\mathcal{B}$ compact sets in $\mathbb{R}^n$,* $\mathrm{Vol}((\mathcal{A} + \mathcal{B})/2) \geq \sqrt{\mathrm{Vol}(\mathcal{A})\,\mathrm{Vol}(\mathcal{B})}$.

[④]The normal distribution has such useful properties that it seems that the only thing normal about it is its name.

References

[Ach01] D. Achlioptas. Database-friendly random projections. *Proc. 20th ACM Sympos. Principles Database Syst.* (PODS), 274–281, 2001.

[AHY07] P. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in Euclidean space. *Proc. 23rd Annu. Sympos. Comput. Geom.* (SoCG), 381–389, 2007.

[BM58] G. E. Box and M. E. Muller. A note on the generation of random normal deviates. *Ann. Math. Stat.*, 28: 610–611, 1958.

[Car76] L. Carroll. *The hunting of the snark*. 1876.

[DG03] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Rand. Struct. Alg.*, 22(3): 60–65, 2003.

[Gar02] R. J. Gardner. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.*, 39: 355–405, 2002.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proc. 30th Annu. ACM Sympos. Theory Comput.* (STOC), 604–613, 1998.

[KN14] D. M. Kane and J. Nelson. Sparser johnson-lindenstrauss transforms. *J. Assoc. Comput. Mach.*, 61(1): 4, 2014.

[Mag07] A. Magen. Dimensionality reductions in ℓ_2 that preserve volumes and distance to affine spaces. *Discrete Comput. Geom.*, 38(1): 139–153, 2007.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*. Vol. 212. Grad. Text in Math. Springer, 2002.

[Mat90] J. Matoušek. Bi-Lipschitz embeddings into low-dimensional Euclidean spaces. *Comment. Math. Univ. Carolinae*, 31: 589–600, 1990.