10: The Johnson-Lindenstrauss Lemma*

Sariel Har-Peled

May 14, 2007
CS - Geometric Approximation Algorithms

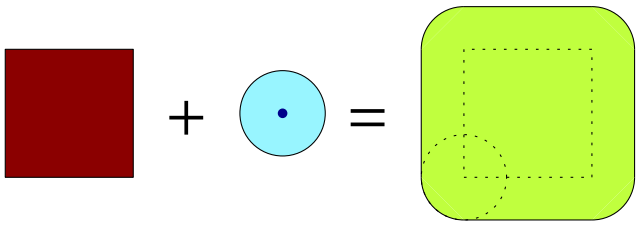
In this chapter, we will prove that given a set P of n points in $\mathbb{R}^d$, one can reduce the dimension of the points to $k = O(\varepsilon^{-2} \log n)$ and distances are $1 \pm \varepsilon$ reserved. Surprisingly, this reduction is done by randomly picking a subspace of k dimensions and projecting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

1 The Brunn-Minkowski inequality

For a set $A \subseteq \mathbb{R}^d$, an a point $\mathsf{p} \in \mathbb{R}^d$, let $A + \mathsf{p}$ denote the translation of A by p. Formally, $A + \mathsf{p} = \left\{ \mathsf{q} + \mathsf{p} \;\middle|\; \mathsf{q} \in A \right\}$.

Definition 1.1 For two sets A and B in $\mathbb{R}^n$, let $A + B$ denote the ***Minkowski sum*** of A and B. Formally,

$$A+B = \left\{ a + b \;\middle|\; a \in A, b \in B \right\} = \cup_{\mathsf{p} \in A} (\mathsf{p} + B).$$



It is easy to verify that if A', B' are translated copies of A, B (that is, $A' = A+\mathsf{p}$ and $B = B+\mathsf{q}$, for some points $\mathsf{p}, \mathsf{q} \in \mathbb{R}^d$), respectively, then $A' + B'$ is a translated copy of $A + B$. In particular, since volume is preserved under translation, we have that $\mathrm{Vol}(A'+B') = \mathrm{Vol}((A+B)+\mathsf{p}+\mathsf{q}) = \mathrm{Vol}(A+B)$.

Theorem 1.2 (Brunn-Minkowski inequality) *Let A and B be two non-empty compact sets in $\mathbb{R}^n$. Then*

$$\mathrm{Vol}(A + B)^{1/n} \geq \mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n}.$$

Definition 1.3 A set $A \subseteq \mathbb{R}^n$ is a ***brick set*** if it is the union of finitely many (close) axis parallel boxes with disjoint interiors.

It is intuitively clear, by limit arguments, that proving Theorem 1.2 for brick sets will imply it for the general case.

Lemma 1.4 (Brunn-Minkowski inequality for Brick Sets) *Let A and B be two non-empty brick sets in $\mathbb{R}^n$. Then*

$$\mathrm{Vol}(A+B)^{1/n} \geq \mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n}$$

Proof: By induction on the number k of bricks in A and B. If $k = 2$ then A and B are just bricks, with dimensions $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, respectively. In this case, the dimensions of $A + B$ are $a_1+b_1, \ldots, a_n+b_n$, as can be easily verified. Thus, we need to prove that $\left(\prod_{i=1}^n a_i\right)^{1/n} + \left(\prod_{i=1}^n b_i\right)^{1/n} \leq \left(\prod_{i=1}^n (a_i + b_i)\right)^{1/n}$. Dividing the left side by the right side, we have

$$\left(\prod_{i=1}^n \frac{a_i}{a_i+b_i}\right)^{1/n} + \left(\prod_{i=1}^n \frac{b_i}{a_i+b_i}\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^n \frac{a_i}{a_i+b_i} + \frac{1}{n}\sum_{i=1}^n \frac{b_i}{a_i+b_i} = 1,$$

by the generalized arithmetic-geometric mean inequality[①], and the claim follows for this case.

Now let $k > 2$ and suppose that the Brunn-Minkowski inequality holds for any pair of brick sets with fewer than k bricks (together). Let A, B be a pair of sets having k bricks together, and A has at least two (disjoint) bricks. However, this implies that there is an axis parallel hyperplane h that separates between the interior of one brick of A and the interior of another brick of A (the hyperplane h might intersect other bricks of A). Assume that h is the hyperplane $x_1 = 0$ (this can be achieved by translation and renaming of coordinates).

Let $\overline{A^+} = A \cap h^+$ and $\overline{A^-} = A \cap h^-$, where h^+ and h^- are the two open half spaces induced by h. Let A^+ and A^- be the closure of $\overline{A^+}$ and $\overline{A^-}$, respectively. Clearly, A^+ and A^- are both brick sets with (at least) one fewer brick than A.

Next, observe that the claim is translation invariant, and as such, let us translate B so that its volume is split by h in the same ratio A's volume is being split. Denote the two parts of B by B^+ and B^-, respectively. Let $\rho = \mathrm{Vol}(A^+)/\mathrm{Vol}(A) = \mathrm{Vol}(B^+)/\mathrm{Vol}(B)$ (if $\mathrm{Vol}(A) = 0$ or $\mathrm{Vol}(B) = 0$ the claim trivially holds).

Observe, that $A^+ + B^+ \subseteq A + B$, and it lies on one side of h, and similarly $A^- + B^- \subseteq A + B$ and it lies on the other side of h. Thus, by induction, we have

$$\begin{aligned}
\mathrm{Vol}(A+B) &\geq \mathrm{Vol}(A^+ + B^+) + \mathrm{Vol}(A^- + B^-) \\
&\geq \left(\mathrm{Vol}(A^+)^{1/n} + \mathrm{Vol}(B^+)^{1/n}\right)^n + \left(\mathrm{Vol}(A^-)^{1/n} + \mathrm{Vol}(B^-)^{1/n}\right)^n \\
&= \left[\rho^{1/n}\,\mathrm{Vol}(A)^{1/n} + \rho^{1/n}\,\mathrm{Vol}(B)^{1/n}\right]^n \\
&\quad + \left[(1-\rho)^{1/n}\,\mathrm{Vol}(A)^{1/n} + (1-\rho)^{1/n}\,\mathrm{Vol}(B)^{1/n}\right]^n \\
&= (\rho + (1-\rho))\left[\mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n}\right]^n \\
&= \left[\mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n}\right]^n,
\end{aligned}$$

establishing the claim. ∎

*Proof of **Theorem 1.2***: Let $A_1 \subseteq A_2 \subseteq \cdots A_i \subseteq$ be a sequence of brick sets, such that $\bigcup_i A_i = A$, and similarly let $B_1 \subseteq B_2 \subseteq \cdots B_i \subseteq \cdots$ be a sequence of finite brick sets, such that $\bigcup_i B_i = B$. It is well known fact in measure theory, that $\lim_{i\to\infty} \mathrm{Vol}(A_i) = \mathrm{Vol}(A)$ and $\lim_{i\to\infty} Vol(B_i) = \mathrm{Vol}(B)$.

[①]Here is a proof of this generalized form: Let $x_1, \ldots, x_n$ be n positive real numbers. Consider the quantity $R = x_1 x_2 \cdots x_n$. If we fix the sum of the n numbers to be equal α, then R is maximized when all the x_is are equal. Thus, $\sqrt[n]{x_1 x_2 \cdots x_n} \leq \sqrt[n]{(\alpha/n)^n} = \alpha/n = (x_1 + \cdots + x_n)/n$.

We claim that $\lim_{i\to\infty} \mathrm{Vol}(A_i + B_i) = \mathrm{Vol}(A + B)$. Indeed, consider any point $z \in A + B$, and let $u \in A$ and $v \in B$ be such that $u + v = z$. By definition, there exists i, such that for all $j > i$ we have $u \in A_j$, $v \in B_j$, and as such $z \in A_i + B_i$. Thus, $\cup_i (A_i + B_i) = A + B$.

Furthermore, for any $i > 0$, since A_i and B_i are brick sets, we have

$$\mathrm{Vol}(A_i + B_i)^{1/n} \geq \mathrm{Vol}(A_i)^{1/n} + \mathrm{Vol}(B_i)^{1/n},$$

by Lemma 1.4. Thus,

$$\begin{aligned} \mathrm{Vol}(A + B) &= \lim_{i\to\infty} \mathrm{Vol}(A_i + B_i)^{1/n} \geq \lim_{i\to\infty} \Big(\mathrm{Vol}(A_i)^{1/n} + \mathrm{Vol}(B_i)^{1/n} \Big) \\ &= \mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n}. \end{aligned}$$

■

Theorem 1.5 (Brunn-Minkowski for slice volumes.) *Let $\mathcal{P}$ be a convex set in $\mathbb{R}^{n+1}$, and let $A = \mathcal{P} \cap (x_1 = a)$, $B = \mathcal{P} \cap (x_1 = b)$ and $C = \mathcal{P} \cap (x_1 = c)$ be three slices of A, for $a < b < c$. We have* $\mathrm{Vol}(B) \geq \min(\mathrm{Vol}(A), \mathrm{Vol}(C))$.

In fact, consider the function

$$v(t) = (\mathrm{Vol}(\mathcal{P} \cap (x_1 = t)))^{1/n},$$

and let $\mathcal{I} = [t_{min}, t_{max}]$ be the interval where the hyperplane $x_1 = t$ intersects $\mathcal{P}$. Then, $v(t)$ is concave in I.

Proof: If a or c are outside $\mathcal{I}$, then $\mathrm{Vol}(A) = 0$ or $\mathrm{Vol}(C) = 0$, respectively, and then the claim trivially holds.

Otherwise, let $\alpha = (b - a)/(c - a)$. We have that $b = (1 - \alpha) \cdot a + \alpha \cdot c$, and by the convexity of $\mathcal{P}$, we have $(1 - \alpha)A + \alpha C \subseteq B$. Thus, by Theorem 1.2 we have

$$\begin{aligned} v(b) = \mathrm{Vol}(B)^{1/n} &\geq \mathrm{Vol}((1 - \alpha)A + \alpha C)^{1/n} \geq \mathrm{Vol}((1 - \alpha)A)^{1/n} + \mathrm{Vol}(\alpha C)^{1/n} \\ &= (1 - \alpha) \cdot \mathrm{Vol}(A)^{1/n} + \alpha \cdot \mathrm{Vol}(C)^{1/n} \\ &\geq (1 - \alpha)v(a) + \alpha v(c). \end{aligned}$$

Namely, $v(\cdot)$ is concave on $\mathcal{I}$, and in particular $v(b) \geq \min(v(a), v(c))$, which in turn implies that $\mathrm{Vol}(B) = v(b)^n \geq \min(\mathrm{Vol}(A), \mathrm{Vol}(B))$, as claimed. ■

Corollary 1.6 *For A and B compact sets in $\mathbb{R}^n$, we have* $\mathrm{Vol}((A + B)/2) \geq \sqrt{\mathrm{Vol}(A)\,\mathrm{Vol}(B)}$.

Proof: $\mathrm{Vol}((A + B)/2)^{1/n} = \mathrm{Vol}(A/2 + B/2)^{1/n} \geq \mathrm{Vol}(A/2)^{1/n} + \mathrm{Vol}(B/2)^{1/n} = (\mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^n)/2 \geq \sqrt{\mathrm{Vol}(A)^{1/n}\,\mathrm{Vol}(B)^{1/n}}$ by Theorem 1.2, and since $(a + b)/2 \geq \sqrt{ab}$ for any $a, b \geq 0$. The claim now follows by raising this inequality to the power n. ■

1.1 The Isoperimetric Inequality

Useless Stuff Warning!!!

The following is not used anywhere else and is provided because of its mathematical elegance. The skip-able reader can thus skip this section.

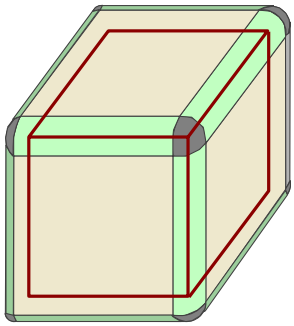
The ***isoperimetric inequality*** states that among all convex bodies of a fixed surface area, the ball has the largest volume (in particular, the unit circle is the largest area planar region with perimeter

2π). This problem can be traced back to antiquity, in particular Zenodorus (200–140 BC) wrote a monograph (which was lost) that seemed to have proved the claim in the plane for some special cases. The first formal proof for the planar case was done by Steiner in 1841. Interestingly, the more general claim is an easy consequence of the Brunn-Minkowski inequality.

Let K be a convex body in $\mathbb{R}^n$ and $\mathbf{b} = \mathbf{b}^n$ be the n dimensional ball of radius one centered at the origin. Let $\mathsf{S}(X)$ denote the surface area of a compact set $X \subseteq \mathbb{R}^n$. The ***isoperimetric inequality*** states that

$$\left(\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(\mathbf{b})}\right)^{1/n} \leq \left(\frac{\mathsf{S}(K)}{\mathsf{S}(\mathbf{b})}\right)^{1/(n-1)}. \tag{1}$$

Namely, the left side is the radius of a ball having the same volume as K, and the right side is the radius of a sphere having the same surface area as K. In particular, if we scale K so that its surface area is the same as $\mathbf{b}$, then the above inequality implies that $\mathrm{Vol}(K) \leq \mathrm{Vol}(\mathbf{b})$.



To prove Eq. (1), observe that $\mathrm{Vol}(\mathbf{b}) = \mathsf{S}(\mathbf{b})/n$[②]. Also, observe that $K + \varepsilon\mathbf{b}$ is the body K together with a small "atmosphere" around it of thickness ε. In particular, the volume of this "atmosphere" is (roughly) $\varepsilon\,\mathsf{S}(K)$ (in fact, Minkowski defined the surface area of a convex body to be the limit stated next). Formally, we have

$$\mathsf{S}(K) = \lim_{\varepsilon\to 0+} \frac{\mathrm{Vol}(K+\varepsilon\mathbf{b}) - \mathrm{Vol}(K)}{\varepsilon} \geq \lim_{\varepsilon\to 0+} \frac{\left(\mathrm{Vol}(K)^{1/n} + \mathrm{Vol}(\varepsilon\mathbf{b})^{1/n}\right)^n - \mathrm{Vol}(K)}{\varepsilon},$$

by the Brunn-Minkowski inequality. Now $\mathrm{Vol}(\varepsilon\mathbf{b})^{1/n} = \varepsilon\,\mathrm{Vol}(\mathbf{b})^{1/n}$, and as such

$$\begin{aligned}\mathsf{S}(K) &\geq \lim_{\varepsilon\to 0+} \frac{\mathrm{Vol}(K) + \binom{n}{1}\varepsilon\,\mathrm{Vol}(K)^{(n-1)/n}\,\mathrm{Vol}(\mathbf{b})^{1/n} + \binom{n}{2}\varepsilon^2\,\langle\text{whatever}\rangle\cdots + \varepsilon^n\,\mathrm{Vol}(\mathbf{b}) - \mathrm{Vol}(K)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0+} \frac{n\varepsilon\,\mathrm{Vol}(K)^{(n-1)/n}\,\mathrm{Vol}(\mathbf{b})^{1/n}}{\varepsilon} = n\,\mathrm{Vol}(K)^{(n-1)/n}\,\mathrm{Vol}(\mathbf{b})^{1/n}.\end{aligned}$$

Dividing both sides by $\mathsf{S}(\mathbf{b}) = n\,\mathrm{Vol}(\mathbf{b})$, we have

$$\frac{\mathsf{S}(K)}{\mathsf{S}(\mathbf{b})} \geq \frac{\mathrm{Vol}(K)^{(n-1)/n}}{\mathrm{Vol}(\mathbf{b})^{(n-1)/n}} \quad\Rightarrow\quad \left(\frac{\mathsf{S}(K)}{\mathsf{S}(\mathbf{b})}\right)^{1/(n-1)} \geq \left(\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(\mathbf{b})}\right)^{1/n},$$

establishing the isoperimetric inequality.

[②] Indeed, $\mathrm{Vol}(\mathbf{b}) = \int_{r=0}^{1} \mathsf{S}(\mathbf{b})\, r^{n-1} dr = \mathsf{S}(\mathbf{b})/n$.

2 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. In fact, as the dimension increases, the width of the strip around the equator $(x_1 = 0) \cap \mathbb{S}^{(n-1)}$ contains, say, 90% of the measure is of width $\approx c/n$, for some constant c. Counter intuitively, this is true for *any* equator. We are going to show that a stronger result holds: The mass is concentrated close to the boundary of any set $A \subseteq \mathbb{S}^{(n-1)}$ such that $\mathbf{Pr}[A] = 1/2$.

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

2.1 The strange and curious life of the hypersphere

Consider the ball of radius r denoted by $r\,\mathbf{b}^n$, where $\mathbf{b}^n$ is the unit radius ball centered at the origin. Clearly, $\mathrm{Vol}(r\,\mathbf{b}^n) = r^n \mathrm{Vol}(\mathbf{b}^n)$. 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 $r^n \leq (1-\delta)^n \leq \exp(-\delta n)$, which is very small if $\delta \gg 1/n$. (Here, we used the fact that $1 - x \leq e^x$, 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 hypersphere. In fact, let $\mathrm{Vol}(r\mathbf{b}^n)$ denote the volume of the ball of radius r in $\mathbb{R}^n$, $\mathrm{Area}\Big(r\mathbb{S}^{(n)}\Big)$ denote the surface area of its boundry (i.e., the surface area of $r\mathbb{S}^{(}n-1)$). It is known that

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

where the $\Gamma(\cdot)$ 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 the volume of the unit ball increases (till dimension 5 in fact) and then it 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 this compute the volume of the unit ball using an integral of its slice volume, when it is being sliced by a hyperplanes perpendicular to the nth coordinate. We have

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

Now, the integral on the right side tends to zero as n increases. In fact, for n very large, the term $\Big(1-x_n^2\Big)^{(n-1)/2}$ 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 by hyperplanes of the form $x_n = \delta$, where δ is small.

If one has to visualize how such a ball in high dimesions 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 close to its center.[3]

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

2.2 Measure Concentration on the Sphere

Theorem 2.1 (Measure concentration on the sphere.) *Let $A \subseteq \mathbb{S}^{(n-1)}$ be a measurable set with $\mathbf{Pr}[A] \geq 1/2$, and let A_t denote the set of points of $\mathbb{S}^{(n-1)}$ in distance at most t from A, where $t \leq 2$. Then $1 - \mathbf{Pr}[A_t] \leq 2\exp\left(-tn^2/2\right)$.*

Proof: We will prove a slightly weaker bound, with $-tn^2/4$ in the exponent. Let

$$\widehat{A} = \left\{ \alpha x \;\middle|\; x \in A, \alpha \in [0,1] \right\} \subseteq \mathbf{b}^n,$$

where $\mathbf{b}^n$ is the unit ball in $\mathbb{R}^n$. We have that $\mathbf{Pr}[A] = \mu\left(\widehat{A}\right)$, where $\mu\left(\widehat{A}\right) = \mathrm{Vol}\left(\widehat{A}\right) / \mathrm{Vol}(\mathbf{b}^n)$[4]

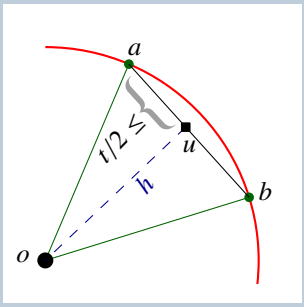
Let $B = \mathbb{S}^{(n-1)} \setminus A_t$. We have that $\|a - b\| \geq t$ for all $a \in A$ and $b \in B$.

Lemma 2.2 *For any $\widehat{a} \in \widehat{A}$ and $\widehat{b} \in \widehat{B}$, we have $\left\| \frac{a+b}{2} \right\| \leq 1 - \frac{t^2}{8}$.*

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

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

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



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

by Eq. (2) 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 $t \leq 2$. ■

By Lemma 2.2, the set $\left(\widehat{A} + \widehat{B}\right)/2$ is contained in a ball of radius $\leq 1 - t^2/8$ around the origin. Applying the Brunn-Minkowski inequality in the form of Corollary 1.6, we have

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

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

[4] 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 $\mathbf{b}^n$. Let ψ be the probability that $\mathsf{p} \in \widehat{A}$. Clearly, $\mathrm{Vol}(\widehat{A}) = \psi\,\mathrm{Vol}(\mathbf{b}^n)$. So, consider the normalized point $\mathsf{q} = \mathsf{p}/\|\mathsf{p}\|$. Clearly, $\mathsf{p} \in \widehat{A}$ if and only if $\mathsf{q} \in A$, by the definition of $\widehat{A}$. Thus, $\mu\left(\widehat{A}\right) = \mathrm{Vol}(\widehat{A})/\mathrm{Vol}(\mathbf{b}^n) = \psi = \mathbf{Pr}\left[\mathsf{p} \in \widehat{A}\right] = \mathbf{Pr}[\mathsf{q} \in A] = \mathbf{Pr}[A]$, since q has a uniform distribution on the hypersphere by the symmetry of $\mathbf{b}^n$.

3 Concentration of Lipschitz Functions

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

Lemma 3.1 *Let* $\mathbf{Pr}[f < \mathrm{med}(f)] \le 1/2$ *and* $\mathbf{Pr}[f > \mathrm{med}(f)] \le 1/2$.

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

$$\mathbf{Pr}[f < \mathrm{med}(f)] \;=\; \sup_{k\ge 1} \mathbf{Pr}\left[f \le \mathrm{med}(f) - \frac{1}{k}\right] \le \frac{1}{2}.$$

■

Definition 3.2 (*c-Lipschitz*) A function $f : A \to B$ is ***c-Lipschitz*** if, for any $x, y \in A$, we have $\|f(x) - f(y)\| \le c\,\|x - y\|$.

Theorem 3.3 (Lévy's Lemma.) *Let* $f : \mathbb{S}^{(n-1)} \to \mathbb{R}$ *be* 1*-Lipschitz. Then for all* $t \in [0, 1]$,

$$\mathbf{Pr}[f > \mathrm{med}(f) + t] \le 2\exp\left(-t^2 n/2\right) \text{ and } \mathbf{Pr}[f < \mathrm{med}(f) - t] \le 2\exp\left(-t^2 n/2\right).$$

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 3.1, we have $\mathbf{Pr}[A] \ge 1/2$. Since f is 1-Lipschitz, we have $f(x) \le \mathrm{med}(f) + t$, for any $x \in A_t$. Thus, by Theorem 2.1, we get $\mathbf{Pr}[f > \mathrm{med}(f) + t] \le 1 - \mathbf{Pr}[A_t] \le 2\exp\left(-t^2 n/2\right)$. ■

4 The Johnson-Lindenstrauss Lemma

Lemma 4.1 *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*

$$\mathbf{Pr}[f(x) \ge m + t] \le 2\exp(-t^2 n/2) \quad \text{and} \quad \mathbf{Pr}[f(x) \le m - t] \le 2\exp(-t^2 n/2).$$

Furthermore, for $k \ge 10 \ln n$*, we have* $m \ge \frac{1}{2}\sqrt{k/n}$.

Proof: The orthogonal projection $p : \ell_2^n \to \ell_2^k$ given by $p(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$ is 1-Lipschitz (since projections can only shrink distances, see Exercise 7.2). 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 3.3 (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 $\mathbf{E}\left[\|x\|^2\right] = 1$. By linearity of expectations, and symmetry, we have $1 = \mathbf{E}\left[\|x\|^2\right] = \mathbf{E}\left[\sum_{i=1}^n x_i^2\right] = \sum_{i=1}^n \mathbf{E}\left[x_i^2\right] = n\,\mathbf{E}\left[x_j^2\right]$, for any $1 \le j \le n$. Thus, $\mathbf{E}\left[x_j^2\right] = 1/n$, for $j = 1, \ldots, n$. Thus, $\mathbf{E}\left[(f(x))^2\right] = k/n$. We next use the fact that f is concentrated, to show that f^2 is also relatively concentrated.

For any $t \ge 0$, we have

$$\frac{k}{n} = \mathbf{E}\left[f^2\right] \le \mathbf{Pr}[f \le m + t]\,(m+t)^2 + \mathbf{Pr}[f \ge m + t] \cdot 1 \le 1 \cdot (m+t)^2 + 2\exp(-t^2 n/2),$$

since $f(x) \le 1$, for any $x \in \mathbb{S}^{(n-1)}$. Let $t = \sqrt{k/5n}$. Since $k \ge 10 \ln n$, we have that $2\exp(-t^2 n/2) \le 2/n$. We get that $\frac{k}{n} \le (m + k/5n)^2 + 2/n$. Implying that $\sqrt{(k-2)/n} \le m + k/5n$, which in turn implies that $m \ge \sqrt{(k-2)/n} - k/5n \ge \frac{1}{2}\sqrt{k/n}$. ∎

At this point, we would like to flip Lemma 4.1 around, and 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. However, we need to pick this k-dimensional space carefully, so that if we rotate this random subspace, by a transformation T, so that it occupies the first k dimensions, then the point $T(x)$ is uniformly distributed on the hypersphere.

To this end, we would like to randomly pick a random rotation of $\mathbb{R}^n$. This 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 is 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.

Generating a random vector from the unit hypersphere, and a random rotation. At this point, the reader might wonder how do we pick a point uniformly from the unit hypersphere. The idea is to pick a point from the multi-dimensional normal distribution $N^d(0,1)$, and normalizing 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 in distance r from the origin has the same distribution), it follows that this indeed randomly 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. Given a source of random numbers according to the uniform distribution, this can be done using a $O(1)$ computations, using the Box-Muller transformation [BM58].

Since projecting down 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 multidimensional 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.

Definition 4.2 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 $c > 0$ such that

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

for all $\mathsf{p}, \mathsf{q} \in X$.

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

Theorem 4.3 (Johnson-Lindenstrauss lemma.) *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: Let $X \subseteq \mathbb{R}^n$ (if X lies in higher dimensions, we can consider it to be lying in the span of its points, if it is in lower dimensions, we can add zero coordinates). 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 $P_{\mathcal{F}} : \mathbb{R}^n \to L$ be the orthogonal projection operator. Let m be the number around which $\|P_{\mathcal{F}}(x)\|$ is concentrated, for $x \in \mathbb{S}^{(n-1)}$, as in Lemma 4.1.

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 this holds for all pair of points of X. In such a case, the mapping p is 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 $\left(1 - \frac{\varepsilon}{3}\right) m \|u\| \leq \|P_{\mathcal{F}}(u)\| \leq \left(1 + \frac{\varepsilon}{3}\right) m \|u\|$. Since this condition is scale independent, we can assume $\|u\| = 1$. Namely, we need to show that

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

By Lemma 4.1 (exchanging the random space with the random vector), for $t = \varepsilon m/3$, we have that the probablity that this does not hold is bounded by

$$4 \exp\left(-\frac{t^2 n}{2}\right) = 4 \exp\left(\frac{-\varepsilon^2 m^2 n}{18}\right) \leq 4 \exp\left(-\frac{\varepsilon^2 k}{72}\right) < n^{-2},$$

since $m \geq \frac{1}{2}\sqrt{k/n}$. ∎

5 An alternative proof of the Johnson-Lindenstrauss lemma

5.1 Some Probability

Definition 5.1 Let $N(0, 1)$ denote the one dimensional ***normal distribution***. This distribution has density $n(x) = e^{-x^2/2}/\sqrt{2\pi}$.

Let $N^d(0, 1)$ denote the d-dimensional Gaussian distribution, induced by picking each coordinate independently from the standard normal distribution $N(0, 1)$.

Let $\mathrm{Exp}(\lambda)$ denote the ***exponential distribution***, with parameter λ. The density function of the exponential distribution is $f(x) = \lambda \exp(-\lambda x)$.

Let $\Gamma_{\lambda,k}$ denote the ***gamma distribution***, with parameters λ and k. The density function of this distribution is $g_{\lambda,k}(x) = \lambda \frac{(\lambda x)^{k-1}}{(k-1)!} \exp(-\lambda x)$. The cumulative distribution function of $\Gamma_{\lambda,k}$ is $G_{\lambda,k}(x) =$

$1 - \exp(-\lambda x)\Big(1 + \frac{\lambda x}{1!} + \cdots + \frac{(\lambda x)^i}{i!} + \cdots + \frac{(\lambda x)^{k-1}}{(k-1)!}\Big)$. As we prove below, gamma distribution is how much time one has to wait till k experiments succeed, where an experiment duration distributes according to the exponential distribution.

A random variable X has the ***Poisson distribution***, with parameter $\eta > 0$, which is a discrete distribution, if $\mathbf{Pr}[X = i] = e^{-\eta}\frac{\eta^i}{i!}$.

Lemma 5.2 *The following properties hold for the d dimensional Gaussian distribution $N^d(0, 1)$:*

(i) The distribution $N^d(0, 1)$ is centrally symmetric around the origin.

(ii) If $X \sim N^d(0, 1)$ and u is a unit vector, then $X \cdot u \sim N(0, 1)$.

(iii) If $X, Y \sim N(0, 1)$ are two independent variables, then $Z = X^2 + Y^2$ follows the exponential distribution with parameter $\lambda = \frac{1}{2}$.

(iv) Given k independent variables $X_1, \ldots, X_k$ distributed according to the exponential distribution with parameter λ, then $Y = X_1 + \cdots + X_k$ is distributed according to the Gamma distribution $\Gamma_{\lambda,k}(x)$.

Proof: (i) Let $x = (x_1, \ldots, x_d)$ be a point picked from the Gaussian distribution. The density $\phi_d(x) = \phi(x_1)\phi(x_2) \cdot \phi(x_d)$, where $\phi(x_i)$ is the normal distribution density function, which is $\phi(x_i) = \exp(-x_i^2/2)/\sqrt{2\pi}$. Thus $\phi_d(x) = (2\pi)^{-n/2}\exp(-(x_1^2 \cdots + x_d^2)/2)$. Consider any two points $x, y \in \mathbb{R}^n$, such that $r = \|x\| = \|y\|$. Clearly, $\phi_d(x) = \phi_d(y)$. Namely, any two points of the same distance from the origin, have the same density (i.e., "probability"). As such, the distribution $N^d(0, 1)$ is centrally symmetric around the origin.

(ii) Consider $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^n$. Clearly, $x \cdot e_1 = x_1$, which is distributed $N(0, 1)$. Now, by the symmetry of $N^d(0, 1)$, this implies that $x \cdot u$ is distributed $N(0, 1)$. Formally, let R be a rotation matrix that maps u to e_1. We know that Rx is distributed $N^d(0, 1)$ (since $N^d(0, 1)$ is centrally symmetric). Thus $x \cdot u$ has the same distribute as $Rx \cdot Ru$, which has the same distribution as $x \cdot e_1$, which is $N(0, 1)$.

(iii) If $X, Y \sim N(0, 1)$, and consider the integral of the density function

$$A = \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}\frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)dx\,dy.$$

We would like to change the integration variables to $x(r, \alpha) = \sqrt{r}\sin\alpha$ and $y(r, \alpha) = \sqrt{r}\cos\alpha$. The Jacobian of this change of variables is

$$I(r,\alpha) = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \alpha} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \alpha} \end{vmatrix} = \begin{vmatrix} \frac{\sin\alpha}{2\sqrt{r}} & \sqrt{r}\cos\alpha \\ \frac{\cos\alpha}{2\sqrt{r}} & -\sqrt{r}\sin\alpha \end{vmatrix} = -\frac{1}{2}\Big(\sin^2\alpha + \cos^2\alpha\Big) = -\frac{1}{2}.$$

As such, we have

$$\begin{aligned}
\mathbf{Pr}[Z = z] &= \int_{x^2+y^2=\alpha}\frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) \\
&= \int_{\alpha=0}^{2\pi}\frac{1}{2\pi}\exp\left(-\frac{x(\sqrt{z},\alpha)^2 + y(\sqrt{z},\alpha)^2}{2}\right)\cdot|I(r,\alpha)| \\
&= \frac{1}{2\pi}\cdot\frac{1}{2}\cdot\int_{\alpha=0}^{2\pi}\exp\left(-\frac{z}{2}\right) = \frac{1}{2}\exp\left(-\frac{z}{2}\right).
\end{aligned}$$

As such, Z has an exponential distribution with $\lambda = 1/2$.

(iv) For $k = 1$ the claim is trivial. Otherwise, let $g_{k-1}(x) = \lambda\frac{(\lambda x)^{k-2}}{(k-2)!}\exp(-\lambda x)$. Observe that

$$\begin{aligned} g_k(t) &= \int_0^t g_{k-1}(t-x)g_1(x)\,dx = \int_0^t \left(\lambda\frac{(\lambda(t-x))^{k-2}}{(k-2)!}\exp(-\lambda(t-x))\right)(\lambda\exp(-\lambda x))\;dx \\ &= \int_0^t \lambda^2\frac{(\lambda(t-x))^{k-2}}{(k-2)!}\exp(-\lambda t)\,dx \\ &= \lambda\exp(-\lambda t)\int_0^t \lambda\frac{(\lambda x)^{k-2}}{(k-2)!}\,dx = \lambda\exp(-\lambda t)\frac{(\lambda t)^{k-1}}{(k-1)!} = g_k(x). \end{aligned}$$

■

5.2 The proof

Lemma 5.3 *Let u be a unit vector in $\mathbf{R}^d$. For any even positive integer k, let $U_1, \ldots, U_k$ be random vectors chosen independently from the d-dimensional Gaussian distribution $N^d(0,1)$. For $X_i = u \cdot U_i$, define $W = W(u) = (X_1, \ldots, X_k)$ and $L = L(u) = \|W\|^2$. Then, for any $\beta > 1$, we have:*

1. $\mathbf{E}[L] = k$.
2. $\mathbf{Pr}[L \geq \beta k] \leq \frac{k+3}{2}\exp\left(-\frac{k}{2}(\beta - (1+\ln\beta))\right)$.
3. $\mathbf{Pr}[L \leq k/\beta] \leq 6k\exp\left(-\frac{k}{2}(\beta^{-1} - (1-\ln\beta))\right)$.

Proof: By Lemma 5.2 (ii) each X_i is distributed as $N(0,1)$, and $X_1, \ldots, X_k$ are independent. Define $Y_i = X_{2i-1}^2 + X_{2i}^2$, for $i = 1, \ldots, \tau$, where $\tau = k/2$. By Lemma 5.2 (iii) Y_i follows the exponential distribution with parameter $\lambda = 1/2$. Let $L = \sum_{i=1}^{\tau} Y_i$. By Lemma 5.2 (iv), the variable L follows the Gamma distribution $(k/2, 1/2)$, and its expectation is $\mathbf{E}[L] = \sum_{i=1}^{k/2}\mathbf{E}[Y_i] = 2\tau = k$.

Now, let $\eta = \beta\tau = \beta k/2$, we have

$$\mathbf{Pr}[L \geq \beta k] = 1 - \mathbf{Pr}[L \leq \beta k] = 1 - G_{1/2,\tau}(\beta k) = \sum_{i=0}^{\tau} e^{-\eta}\frac{\eta^i}{i!} \leq (\tau+1)e^{-\eta}\frac{\eta^\tau}{\tau!},$$

since $\eta = \beta\tau > \tau$, as $\beta > 1$. Now, since $\tau! \geq (\tau/e)^\tau$, as can be easily verified[5], and thus

$$\begin{aligned} \mathbf{Pr}[L \geq \beta k] &\leq (\tau+1)e^{-\eta}\frac{\eta^\tau}{\tau^\tau/e^\tau} = (\tau+1)e^{-\eta}\left(\frac{e\eta}{\tau}\right)^\tau = (\tau+1)e^{-\beta\tau}\left(\frac{e\beta\tau}{\tau}\right)^\tau \\ &= (\tau+1)e^{-\beta\tau}\cdot\exp(\tau\ln(e\beta)) = (\tau+1)\exp(-\tau(\beta - (1+\ln\beta))) \\ &= \frac{k+3}{2}\exp\left(-\frac{k}{2}(\beta - (1+\ln\beta))\right). \end{aligned}$$

Arguing in a similar fashion, we have, for $\nu = \lceil 2e\tau\rceil$, that

$$\begin{aligned} \mathbf{Pr}[L \leq k/\beta] &= \sum_{i=\tau}^{\infty} e^{-\tau/\beta}\frac{(\tau/\beta)^i}{i!} \leq e^{-\tau/\beta}\sum_{i=\tau}^{\infty}\left(\frac{e\tau}{i\beta}\right)^i = e^{-\tau/\beta}\left[\sum_{i=\tau}^{\nu}\left(\frac{e\tau}{i\beta}\right)^i + \sum_{i=\nu+1}^{\infty}\left(\frac{e\tau}{i\beta}\right)^i\right] \\ &\leq e^{-\tau/\beta}\left[\sum_{i=\tau}^{\nu}\left(\frac{e\tau}{i\beta}\right)^i + \frac{1}{(2\beta)^\nu}\right] \leq 2e^{-\tau/\beta}\sum_{i=\tau}^{\nu}\left(\frac{e\tau}{i\beta}\right)^i, \end{aligned}$$

[5] Indeed, $\ln\tau! = \sum_{i=1}^{\tau}\ln i \geq \int_{x=1}^{n}\ln x\,dx = \Big[x\ln x - x\Big]_{x=1}^{n} = n\ln n - n + 1 \geq n\ln n - n = \ln((n/e)^n)$.

since $(e\tau/\tau\beta)^\tau \geq 1/(2\beta)^\nu$. As the sequence $(e\tau/i\beta)^i$ is decreasing for $i > \tau/\beta$, as can be easily verified[6], we can bound the (decreasing) summation above by

$$\sum_{i=\tau}^{\nu}\left(\frac{e\tau}{i\beta}\right)^i \leq \nu\left(\frac{e}{\beta}\right)^\tau = \lceil 2e\tau\rceil \exp(\tau(1-\ln\beta))\,.$$

We conclude

$$\mathbf{Pr}[L \leq k/\beta] \leq 2\lceil 2e\tau\rceil \exp(-\tau/\beta + \tau(1-\ln\beta)) \leq 6k\exp\left(-\frac{k}{2}(\beta^{-1} - (1-\ln\beta))\right).$$

■

Next, we show how to interpret the inequalities of Lemma 5.3 in a somewhat more intuitive way. Let $\beta = 1+\varepsilon$, for ε such that $1 > \varepsilon > 0$. From the Taylor expansion of $\ln(1+x) = \sum_{i=0}^{\infty}\frac{(-1)^i}{i+1}x^{i+1}$, it follows that $\ln\beta \leq \varepsilon - \varepsilon^2/2 + \varepsilon^3/3$. By plugging it into the upper bound for $\mathbf{Pr}[L \geq \beta k]$ we get

$$\mathbf{Pr}[L \geq \beta k] \leq \frac{k+3}{2}\exp\left(-\frac{k}{2}(1+\varepsilon-1-\varepsilon+\varepsilon^2/2-\varepsilon^3/3)\right) \leq \frac{k+3}{2}\exp\left(-\frac{k}{2}(\varepsilon^2/2-\varepsilon^3/3)\right)$$

On the other hand, since $\ln\beta \geq \varepsilon - \varepsilon^2/2$, we have $\mathbf{Pr}[L \leq k/\beta] \leq 6k\exp(\Delta)$, where

$$\begin{aligned}\Delta &= -\frac{k}{2}(\beta^{-1} - (1-\ln\beta)) \leq -\frac{k}{2}\left(\frac{1}{1+\varepsilon} - 1 + \varepsilon - \frac{\varepsilon^2}{2}\right)\\ &\leq -\frac{k}{2}\left(\frac{\varepsilon^2}{1+\varepsilon} - \frac{\varepsilon^2}{2}\right) = -\frac{k}{2}\cdot\frac{\varepsilon^2-\varepsilon^3}{2(1+\varepsilon)}\end{aligned}$$

Thus, the probability that a given unit vector gets distorted by more than $(1+\varepsilon)$ in any direction[7] grows roughly as $\exp(-k\varepsilon^2/4)$, for small $\varepsilon > 0$. Therefore, if we are given a set P of n points in l_2, we can set k to roughly $8\ln(n)/\varepsilon^2$ and make sure that with non-zero probability we obtain projection which does not distort distances[8] between *any* two different points from P by more than $(1+\varepsilon)$ in each direction.

Theorem 5.4 *Let P be a set of n points in $\mathbb{R}^d$, $0 < \varepsilon, \delta < 1/2$, and $k = 16\ln(n/\delta)/\varepsilon^2$. Let $U_1, \ldots, U_k$ be random vectors chosen independently from the d-dimensional Gaussian distribution $N^d(0,1)$, and let $T(x) = (U_1 \cdot x, \ldots, U_k \cdot x)$ be a linear transformation. Then, with probablity $\geq 1-\delta$, for any $\mathsf{p}, \mathsf{q} \in \mathsf{P}$, we have that*

$$\frac{1}{(1+\varepsilon)}\sqrt{k}\,\|\mathsf{p}-\mathsf{q}\| \leq \|T(\mathsf{p}) - T(\mathsf{q})\| \leq (1+\varepsilon)\sqrt{k}\,\|\mathsf{p}-\mathsf{q}\|\,.$$

Sometime it is useful to be able to handle high distortion.

Corollary 5.5 *Let k be the target dimension of the transformation T of Theorem 5.4 and $\beta \geq 3$ a parameter. We have that*

$$\|T(\mathsf{p}) - T(\mathsf{q})\| \leq \beta\sqrt{k}\,\|\mathsf{p}-\mathsf{q}\|\,,$$

for any two points $\mathsf{p}, \mathsf{q} \in \mathsf{P}$, and this holds with probability $\geq 1 - \exp\left(-\frac{k\beta^2}{32\ln n}\right)$.

[6] Indeed, consider the function $f(x) = x\ln(c/x)$, its derivative is $f'(x) = \ln(c/x) - 1$, and as such $f'(x) = 0$, for $x = c/e$. Namely, for $c = e\tau/\beta$, the function $f(x)$ achieves its maximum at $x = \tau/\beta$, and from this point on the function is decreasing.

[7] Note that this implies distortion $(1+\varepsilon)^2$ if we require the mapping to be a contraction.

[8] In fact, this statement holds even for the *square* of the distances.

6 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].

The probability review of Section 5.1 can be found in Feller [Fel71]. The alternative proof of the Johnson-Lindenstrauss lemma of Section 5.2 is described by Indyk and Motwani [IM98] but earlier proofs are known [Dur95]. It exposes the fact that the Johnson-Lindenstrauss lemma is no more than yet another instance of the concentration of mass phenomena (i.e., like the Chernoff inequality). The alternative proof is provided since it is conceptually simpler (although the computations are more involved), and it is technically easier to use. Another alternative proof is provided by Dasgupta and Gupta [DG03].

Interestingly, it is enough to pick each entry in the dimension reducing matrix randomly out of $-1, 0, 1$. This requires more involved proof [Ach01]. This is useful when one care about storing this dimension reduction transformation efficiently.

Magen [Mag02] observed that in fact 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 7.3 is taken from there.

In fact, the random embedding preserves much more structure than just 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 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 axises. Other techniques include independent component analysis, and MDS (multidimensional scaling). MDS tries to embed points from high dimensions into low dimension ($d = 2$ or 3), which 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].

7 Exercises

Exercise 7.1 (Boxes can be separated.) [1 Points]

(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 7.2 (Projections are contractions.) [1 Points]

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

Exercise 7.3 (JL Lemma works for angles.) [10 Points]

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 a equilateral triangle that its 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.**]**

References

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

[AHY07] P. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in euclidean space. In *Proc. 23rd Annu. ACM Sympos. Comput. Geom.*, page to appear, 2007.

[BM58] G. E.P. Box and M. E. Muller. A note on the generation of random normal deviates. *Annl. 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.

[Dur95] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, August 1995.

[Fel71] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. John Wiley & Sons, NY, 1971.

[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. In *Proc. 30th Annu. ACM Sympos. Theory Comput.*, pages 604–613, 1998.

[Mag02] A. Magen. Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications. In *The 6th Intl. Work. Rand. Appr. Tech. Comp. Sci.*, pages 239–253, 2002.

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

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*. Springer, 2002.