

Chapter 2

Low-Distortion Embeddings

Informally speaking, a *metric space* is a set of points for which we have a reasonable notion of distance between any two points. In this chapter, we will study questions of the following kind: How well can one kind of metric space (which we think of as “complicated”) be embedded into another one (which we think of as “simple”) in such a way that the distances are approximately preserved? Our presentation is based on Chapter 15 of Matoušek’s book [7].

2.1 Metric Spaces and Normed Spaces

This section contains the formal definitions of metric spaces and normed spaces. You may prefer to jump ahead to the following sections and refer back to the first one and look up the definitions when they are needed later.

Definition 2.1 (Metrics and Metric Spaces). *Suppose we have a set X (finite or infinite) and a function $\rho : X \times X \rightarrow \mathbb{R}$ such that the following conditions hold for all $x, y, z \in X$:*

1. $\rho(x, y) \geq 0$, and $\rho(x, y) = 0$ iff $x = y$,
2. $\rho(x, y) = \rho(y, x)$,
3. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

Then ρ is called a metric and the pair (X, ρ) is called a metric space. When the metric ρ is understood from the context, we will sometimes just speak of the metric space X and suppress ρ from the notation.

As we will see below, checking the 3rd property, called the *triangle inequality*, is usually the nontrivial part when determining whether a given function ρ is a metric or not.

Some Examples.

1. The most familiar case is $X = \mathbb{R}^d$ with the Euclidean distance $\rho(x, y) = \|x - y\|_2$. In fact, this is an example of a *normed space* (see below).
2. The *discrete cube* $X = \{0, 1\}^d$ with the hamming distance $\rho_H(x, y) = |\{i : x_i \neq y_i\}|$, which we also already encountered in Chapter 1.
3. If X is the set of all finite strings over some finite alphabet Σ , we can define the *edit distance* $\rho_{\text{edit}}(x, y)$ between two strings $x, y \in X$ to be the minimal number of substitutions (replace one letter by another one), deletions, or insertions needed to transform x into y (or vice versa).
4. If $G = (V, E)$ is a graph, we can define the *shortest path metric* on the vertices: For any two vertices $v, w \in V$, we define their distance $\rho(v, w)$ as the length of a shortest path in the graph connecting v and w .

Note that if we view the discrete cube as a graph in the usual way, i.e., $x, y \in \{0, 1\}^d$ are connected by an edge if they differ in exactly one coordinate, then the resulting shortest path metric is exactly the Hamming metric.

Usually, one assumes that the graph underlying the shortest path metric is finite, but in fact one can allow infinite vertex sets as long as any two vertices are connected by some finite path. Then also the edit distance is a special case of a shortest path metric, namely on the following graph: The vertices are all finite strings over the given alphabet, and two strings x and y are connected by an edge iff x can be obtained from y by exactly one deletion one insertion, or one substitution.

If X is a vector space, then it is quite natural to consider metrics that are in some sense compatible with the operations allowed in a vector space, addition and scalar multiplication. A metric on X is called *translation-invariant* if the distance from x to y is the same as the distance from $x + z$ to $y + z$, for all $x, y, z \in X$. For such a translation-invariant metric, it is in fact enough to know the distance from every point to some fixed point, say the origin 0 ; this already determines the metric, because $\rho(x, y) = \rho(x - y, 0)$.

If we additionally require that the metric be scaling-sensitive, then we arrive at the definition of a norm:

Definition 2.2 (Norms and Normed Spaces). *Let X be a real vector space.¹ A norm on X is a function $\|\cdot\| : X \rightarrow \mathbb{R}$ with the following properties:*

1. $\|x\| \geq 0$ for all $x \in X$, and $\|x\| = 0$ iff $x = 0$.

¹ X could be infinite-dimensional, although we will mostly be concerned with $X \cong \mathbb{R}^d$; the definition also makes sense for complex vector spaces.

2. $\|\lambda x\| = |\lambda|\|x\|$ for all $x \in X$ and $\lambda \in \mathbb{R}$.
3. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$.

Remark 2.3. Any norm defines translation-invariant metric on X by setting $\rho(x, y) := \|x - y\|$ (check this!). Conversely, if we start with a metric ρ with these additional properties, we can define a norm by $\|x\| := \rho(x, 0)$.

A viewpoint that is very useful for understanding a norm is by looking at its “unit ball” $B_1(\|\cdot\|) := \{x \in X : \|x\| \leq 1\}$. Observe that the unit ball determines a norm in the sense that

$$\|x\| = \inf\{\lambda \in \mathbb{R}_+ : \frac{1}{\lambda}x \in B_1(\|\cdot\|)\}.$$

The unit ball of a norm is a centrally symmetric convex set that is bounded in the sense that it does not contain any infinite ray. If the dimension of X is finite, then the unit ball is also compact, but in infinite-dimensional spaces, this need not be the case.

Example 2.4 (p-Norms). For a real number $1 \leq p < \infty$, the p -norm $\|\cdot\|_p$ on \mathbb{R}^d is defined by

$$\|x\|_p := \sqrt[p]{\sum_{i=1}^d |x_i|^p}, \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

The usual Euclidean norm is the special case $p = 2$. Observe that as p grows larger and larger, the influence of coordinates of large absolute value increases. The definition of the p -norm is extended to the limit case $p = \infty$ by setting

$$\|x\|_\infty := \max_{1 \leq i \leq d} |x_i|.$$

For $1 \leq p \leq \infty$, we write ℓ_p^d for the space \mathbb{R}^d equipped with the p -norm $\|\cdot\|_p$.

As remarked before, when proving that these functions $\|\cdot\|_p$ are indeed norms, checking the triangle inequality is the interesting part. Except for the cases $p = 1$ and $p = \infty$, this is nontrivial and boils down to the following:

Hölder’s Inequality. Let $1 \leq p, q \leq \infty$ such that $1/p + 1/q = 1$ (where we interpret $1/\infty = 0$); the numbers p and q are called conjugate exponents. Observe, for instance, that 1 and ∞ are conjugate exponents, and that 2 is its own conjugate exponent. Then, for all $x, y \in \mathbb{R}^d$,

$$|\langle x, y \rangle| \leq \|x\|_p \|y\|_q.$$

The most important p -norms are the 1-norm, the 2-norm, and the ∞ -norm. Figures 2.1 and 2.2 show their respective unit balls in dimensions 2 and 3.

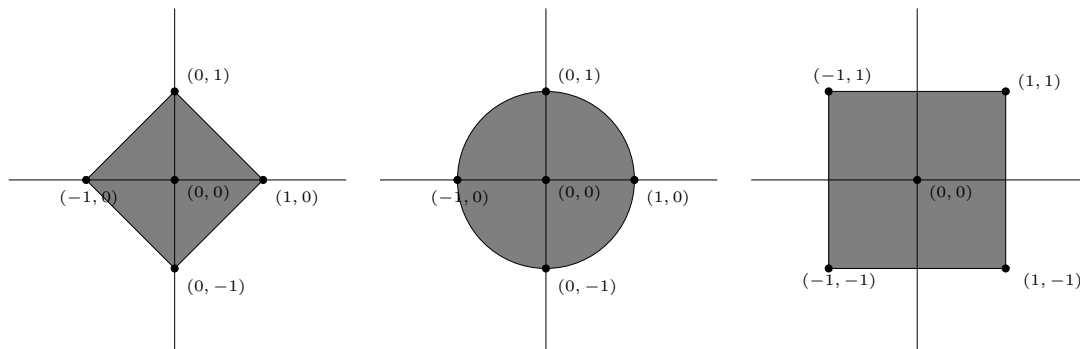


Figure 2.1: The unit balls in ℓ_1^2 , ℓ_2^2 , and ℓ_∞^2

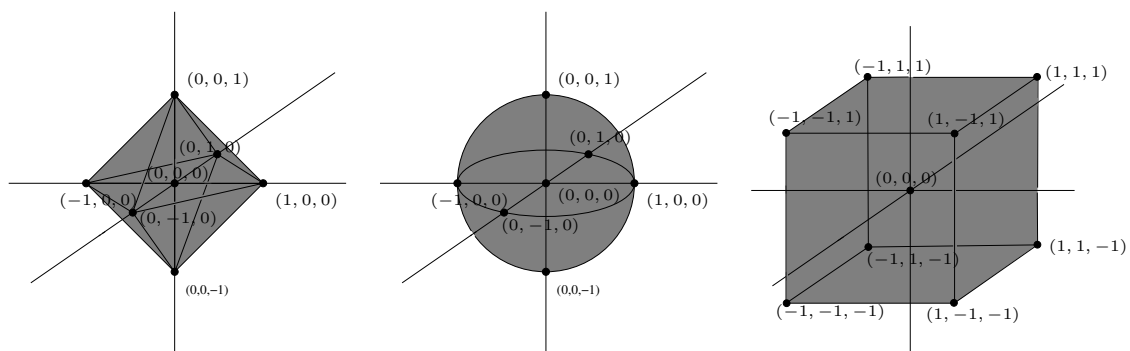


Figure 2.2: The unit balls in ℓ_1^3 , ℓ_2^3 , and ℓ_∞^3 .

2.2 Low-Distortion Embeddings

A metric ρ on an n -element set X can be specified by writing down all the $\binom{n}{2}$ pairwise distances (in a symmetric $n \times n$ -matrix, say). By contrast, if X is an n -elements subset of some space ℓ_p^d , it suffices to write down the $d \cdot n$ coordinates of the points, and we can still compute the pairwise distances in $O(d)$ steps. If d is much smaller than n , this is more space efficient. Moreover, it is generally quite difficult to discern any patterns or structure in a large table of numbers, while it is often much easier to do so in a low-dimensional normed space.

As an often quoted example, think of a large number n of different strains of bacteria; the pairwise distances could be given by comparing the DNA (the DNA of each strain can be viewed as a word over the 4-letter alphabet $\{A, C, G, T\}$, and we could take the edit-distance) or some other biologically interesting measure of (dis)similarity. We might be interested in detecting large clusters of similar bacteria. Or maybe we would like to store the information about the numerous bacteria that we already know in such a way that when somebody claims that they found a new one, we can quickly decide whether

it is similar to one in our database.

Such tasks can be generally quite difficult. In contrast, in low-dimensional Euclidean (and other normed) spaces, efficient geometric algorithms and data structure (for clustering or nearest-neighbor queries) are available. This raises the following general question: Given a “complicated” metric space X , can we represent it in some “simpler” space Y ? For instance, in an ideal situation, we might be able to *isometrically embed* (X, ρ) into the Euclidean plane ℓ_2^2 , i.e., to assign to every point $x \in X$ a point $f(x) \in \mathbb{R}^2$ such that for any two points $x, y \in X$, their distance $\rho(x, y)$ is the same as the Euclidean distance $\|f(x) - f(y)\|_2$ between their image points. In this case, we could immediately visualize X and see clusters and other structures right away.

It is easy to see that isometric embeddings are generally too much to hope for. For instance, consider the two 4-point metric spaces defined by the shortest-path metrics on the two graphs in Figure 2.3. Neither of these two metric

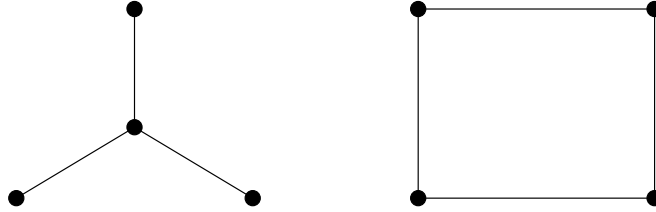


Figure 2.3: Not isometrically embeddable into Euclidean space.

spaces is isometrically embeddable into the Euclidean plane (or indeed into any ℓ_2^d ; this is Exercise 8).

However, for many purposes it suffices if we can find a mapping such that the distances are *approximately preserved*, and this leads to a rich theory, of which we will just scratch the surface.

Definition 2.5. Let (X, ρ) and (Y, σ) be metric spaces, and let $D \geq 1$ be a real number. A map $f : X \rightarrow Y$ is said to have *distortion at most D* if the following holds: There exists a real number $r > 0$ (a “scaling factor”) such that for all $x, y \in X$,

$$r \cdot \rho(x, y) \leq \sigma(f(x), f(y)) \leq D \cdot r \cdot \rho(x, y).$$

The *distortion of f* is defined as the *infimum* of all D such that the above holds; if no such D exists, then the distortion of f is ∞ .

We will focus on target spaces Y that are normed spaces. In this case, we can always take the scaling factor r to be equal to 1 (by replacing f by $\frac{1}{r} \cdot f$, if necessary), even though it will still sometimes be convenient to work with a different r .

As an example, consider the metric space defined by the second graph (the square) in Figure 2.3. If we embed it into the Euclidean plane by mapping

the nodes of the graph to the vertices of the unit square, $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ in the obvious way, then the resulting map has distortion $\sqrt{2}$: The distances between adjacent vertices are preserved exactly, and for each pair of antipodal vertices, their distance in the graph equals two, while the Euclidean distance equals $\sqrt{2}$.

A concept that is closely related to distortion is that of *Lipschitz maps*.

Definition 2.6. Let (X, ρ) and (Y, σ) be metric spaces, and let $C > 0$ be a real number. A map $f : X \rightarrow Y$ is said to be C -Lipschitz if $\sigma(f(x), f(y)) \leq C \cdot \rho(x, y)$ for all $x, y \in X$. The Lipschitz constant $\|f\|_{\text{Lip}}$ of f is defined as the infimum over all C such that f is C -Lipschitz.²

Note that a map f with distortion at most $D < \infty$ is necessarily injective, and by replacing Y by $f(X) \subseteq Y$, we may assume that f is bijective. For a bijective map $f : X \rightarrow Y$ between metric spaces, the distortion of f equals $\|f\|_{\text{Lip}} \cdot \|f^{-1}\|_{\text{Lip}}$ (Exercise 9). For this reason, maps of finite distortion are also often called *bi-Lipschitz*.

The study of low-distortion maps between metric spaces is a very active area at the borderline between mathematics with numerous applications in computer science. One of the classical results is the following:

Bourgain's Theorem. Every n -element metric space can be embedded into some Euclidean space ℓ_2^d with distortion at most $O(\log n)$. (More formally, there exists a constant $C < \infty$ such that for every n and every n -element metric space X , there exists some d and some map $f : X \rightarrow \mathbb{R}^d$ such that f has distortion at most $C \log n$.)

We will not prove Bourgain's Theorem in these lectures. We remark that the estimate is tight, in the sense that there exist n -element metric spaces such that any map into any Euclidean space has distortion at least $\Omega(\log n)$.

We will focus on the topic of *dimension reduction*, i.e., on low-distortion embeddings from finite subsets of high-dimensional normed spaces into low-dimensional ones. We will prove the classic among such results, the so-called Johnson-Lindenstrauss Flattening Lemma, and discuss algorithmic applications of it in Chapter 3.

2.3 The Johnson-Lindenstrauss Flattening Lemma

There are finite subsets of the d -dimensional Euclidean space ℓ_2^d that cannot be isometrically embedded into any lower-dimensional Euclidean space; the most basic example is the set of vertices of a regular d -dimensional simplex

²The notation is no accident; if Y is a normed space, then the set of all functions $f : X \rightarrow Y$ with finite Lipschitz constant forms a vector space, and $\|f\|_{\text{Lip}}$ defines a norm on this space.

(this is a set of $d + 1$ points in ℓ_2^d such that all the pairwise distances equal 1). The goal of this section is to prove the following theorem, which states that the dimension can be significantly reduced if we allow for a small distortion $(1 + \varepsilon)$.

Theorem 2.7 (Johnson-Lindenstrauss Flattening Lemma). *Let X be a set of n points in the d -dimensional Euclidean space ℓ_2^d , and let $0 < \varepsilon \leq 1$ be a parameter. Then there exists a map $g : X \rightarrow \ell_2^k$ of distortion at most $(1 + \varepsilon)$, where $k = O(\varepsilon^{-2} \log n)$.³*

Thus, if we do not need to know the precise pairwise distances but can allow an error of $\varepsilon = 5\%$, then we can embed any n -point set in Euclidean space (for instance, the n vertices of the regular simplex, whose natural habitat is in dimension $d = n - 1$) into dimension $O(\log n)$. As an immediate application, this implies a dramatic reduction in storage: To store n points in ℓ_2^n , we need n^2 numbers; similarly, we need to store $\binom{n}{2}$ numbers for the exact pairwise distances. However, after dimension reduction, it suffices to store $O(n \log n)$ numbers, and we can still reconstruct the pairwise distances up to an error of 5%.

It may be surprising at the first moment that the original dimension d does not appear in the bound for the target dimension k . The reason is that for the purposes of the theorem, we may assume that $d < n$ by restricting our attention to the affine hull of the point set X (if $|X| = n$, then the affine hull $\text{aff}(X)$ is of dimension at most $n - 1$).

The proof that we will present below will give a value of roughly 200 for the implicit constant factor in the O -notation. The currently best constant for the embedding dimension is $(4 + o(1))\varepsilon^{-2} \ln(n)$, where $o(1) \rightarrow 0$ as $\varepsilon \rightarrow 0$.

We also remark that the bound for the embedding dimension is almost sharp, for a wide range of n and ε . This follows from a construction due to Alon, which is the subject of Exercises 14, 15, and 16.

The idea for the proof of the Flattening Lemma is to take the map g as the projection onto a suitable linear subspace L of dimension k . Of course, we have to choose the subspace L right; for instance, it must not be orthogonal to any of the $\binom{n}{2}$ segments spanned by the points (otherwise, the two endpoints of such a segment will be projected to the same point in L and the distortion will be ∞). However, as we will see, this is a “rare” event: If we take the k -dimensional subspace L at random, then with high probability, the projection has distortion at most $(1 + \varepsilon)$ (for k as in the theorem).

As a first step, we need to make precise what we mean by a random subspace.

³A more precise formulation that includes all the quantifiers in the right order is the following: There exist constants $C, \varepsilon_0 > 0$ such that for every $0 < \varepsilon \leq \varepsilon_0$, there is an n_0 such that for all $n \geq n_0$, for all $d \geq 1$, for all $k \geq C \cdot \varepsilon^{-2} \log n$ and for all n -element subsets X of ℓ_2^d , there is a map $g : X \rightarrow \mathbb{R}^k$ of distortion at most $1 + \varepsilon$.

2.4 Random Orthogonal Matrices and Random Subspaces

The goal of this somewhat technical section is to define random orthogonal matrices and random subspaces and to establish some basic facts about them. Ultimately, we will only need Fact 2.8 for the proof of the Flattening Lemma. If you feel that you have a sufficiently good intuitive understanding of what a random subspace is, you might want to take this fact for granted and skip over the rest of this section at first reading.

Recall that an *orthonormal basis* of d -dimensional Euclidean space is a set $\{u_1, \dots, u_d\} \subset \mathbb{R}^d$ such that $\langle u_i, u_j \rangle = 0$ for $i \neq j$ and $\langle u_i, u_i \rangle = 1$ for all i , where $\langle \cdot, \cdot \rangle$ is the standard inner product on \mathbb{R}^d (which induced the Euclidean norm $\|\cdot\|_2$). A natural way of choosing an orthonormal basis at random is the following inductive procedure: First, pick a unit vector $u_1 \in \mathbb{S}^{d-1}$ uniformly at random. Next consider the set $\{x \in \mathbb{S}^{d-1} : \langle x, u_1 \rangle = 0\}$; this is the unit sphere in the $(d-1)$ -dimensional subspace orthogonal to u_1 , so it is a $(d-2)$ -dimensional sphere. We pick u_2 at random according to the uniform distribution on this $(d-2)$ -dimensional sphere. More generally, if we have already chosen orthonormal vectors u_1, \dots, u_i , then the set $\{x \in \mathbb{S}^{d-1} : \langle x, u_1 \rangle = \dots = \langle x, u_i \rangle = 0\}$ of unit vectors orthogonal to all of them forms a $(d-1-i)$ -dimensional sphere, and we pick u_{i+1} uniformly at random from this sphere, see Figure 2.4 (in particular, the last point u_d is picked from the 0-dimensional sphere, so we have exactly two choices for it).

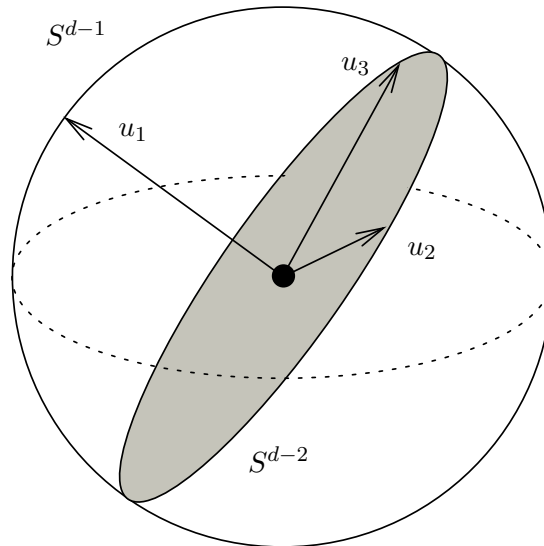


Figure 2.4: Random orthonormal vectors in \mathbb{R}^3 .

We call the resulting set $\{u_1, \dots, u_d\}$ a *random orthonormal basis* of \mathbb{R}^d , and

we call the $d \times d$ -matrix U with columns u_1, \dots, u_d a *random orthogonal matrix*. (Recall that a square matrix U is called orthogonal iff $UU^T = I$, where \cdot^T denotes the transpose and I the identity matrix; thus U is orthogonal iff its columns form an orthonormal basis, which in turn is equivalent to requiring that the rows of U be orthonormal.)

A random k -dimensional linear subspace of \mathbb{R}^d is defined as $U(L_0)$, where U is a random orthogonal matrix and L_0 is the space spanned by the first k coordinate vectors e_1, \dots, e_k . In other words, a random k -dimensional subspace is defined as the subspace spanned by the first k vectors in a random orthonormal basis.

Fact 2.8. *Consider a random k -dimensional linear subspace L of \mathbb{R}^d and a fixed unit vector $u_0 \in \mathbb{R}^d$. The Euclidean length $\|p_L(u_0)\|_2$ of the orthogonal projection of u_0 onto L is a random variable. Similarly, we get a random variable by considering the length $\|p_{L_0}(u)\|_2$ of the orthogonal projection of a random unit vector u onto any fixed k -dimensional linear subspace L_0 . These two random variables have the same distribution.*

The mathematically inclined reader may prefer the following, more abstract and arguably more elegant way of introducing random orthogonal matrices: The orthogonal $d \times d$ -matrices form a compact topological group (the topology is the one the set of matrices inherits as a subspace of \mathbb{R}^{n^2} , by considering each matrix as a point in \mathbb{R}^{n^2}). It is a general result in measure theory that on a compact topological group G , there exists a unique probability measure, called the *Haar measure*, that is invariant under left and right translations in the sense that if g is a random group element with respect to the Haar measure, and g_0 is a fixed group element, then g_0g and gg_0 are again random group elements. The procedure of successively choosing orthogonal unit vectors as described above produces a random orthogonal matrix that is distributed according to the Haar measure.⁴

We will need the following fact about random orthogonal matrices:

Fact 2.9. *If U is a random orthogonal $d \times d$ matrix, then U^{-1} (which, by definition of orthogonality, equals the transpose U^T) is again a random orthogonal matrix; that is, instead of choosing the columns of U successively as random mutually orthogonal unit vectors, we choose the rows in this way, we end up with the same distribution of matrices. Put in yet another way, the uniform probability distribution on the set of orthogonal matrices is invariant under taking inverses (or transposes).⁵*

⁴Left invariance follows from the following fact: The uniform measure on the sphere \mathbb{S}^d is invariant under multiplication by any fixed orthogonal matrix U_0 . This in turn follows from the fact that orthogonal matrices preserve the length of vectors and d -dimensional volume. It can be shown that this left invariance already uniquely determines the Haar measure (for compact topological groups, left and right invariance imply one another).

⁵This follows immediately from the properties of the Haar measure. By uniqueness, it suffices to show that if U is a random orthogonal matrix according to the Haar measure, then the

Corollary 2.10. *If U is a random orthogonal $d \times d$ -matrix and $w \in \mathbb{S}^{d-1}$ is a fixed unit vector, then Uw is a random unit vector (according to the uniform distribution on the sphere \mathbb{S}^{d-1}).*

Proof. Let $e_1 = (1, 0, \dots, 0)$ be the first unit coordinate vector. Choose an arbitrary orthogonal matrix A such that $Aw = e_1$. (Extend w in an arbitrary way to an orthonormal basis $w = w_1, \dots, w_d$, and let A be the matrix that has the w_i 's as columns.) By invariance, UA is again a random orthogonal matrix, so Uw has the same distribution as $UAw = Ue_1$. But Ue_1 equals the first column of U , which we took as a random unit vector in \mathbb{S}^{d-1} . \square

This immediately implies Fact 2.8: The random linear subspace L is obtained as $U(L_0)$, where U is a random orthogonal matrix. Then U^{-1} is again a random orthogonal matrix, and since it preserves Euclidean length, $\|p_L(u_0)\|_2 = \|U^{-1}(p_L(u_0))\|_2 = \|p_{U^{-1}(L)}(U^{-1}(u_0))\|_2$. But $U^{-1}(L) = L_0$, and $U^{-1}(u_0)$ is a random unit vector, which completes the proof.

2.5 Levy's Lemma

We need another technical lemma for the proof of the Flattening Theorem. It is a consequence of measure concentration on the sphere (Theorem 1.11) and says that a real-valued Lipschitz function on the sphere is highly concentrated around its median. Here, the *median* of a real-valued random variable X is defined as $\text{med}(X) := \sup\{x \in \mathbb{R} : \Pr[X \leq x] \leq 1/2\}$.

First, we need the following (hopefully very intuitive) lemma:

Lemma 2.11. *For a random variable X , $\Pr[X < \text{med}(X)] \leq 1/2$ and $\Pr[X > \text{med}(X)] \leq 1/2$.*

Proof. Let $m := \text{med}(X)$. For the first inequality, we have $\Pr[X < m] = \sum_{k=1}^{\infty} \Pr[m - \frac{1}{k-1} < X \leq m - \frac{1}{k}] = \sup_{k \geq 1} \Pr[X \leq m - 1/k] \leq 1/2$. For the second inequality, note that by the definition of m , for $x > m$ we have $\Pr[X \leq x] > 1/2$, i.e., $\Pr[X > x] < 1/2$. Thus, $\Pr[X > m] = \sum_{k=1}^{\infty} \Pr[m + 1/k < X \leq m + \frac{1}{k-1}] = \sup_{k \geq 1} \Pr[X > m + 1/k] \leq 1/2$. \square

Theorem 2.12 (Levy's Lemma). *Let $f : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$ be 1-Lipschitz, and let $0 \leq t \leq 1$ be a real number. Then,*

$$\Pr[f > \text{med}(f) + t] \leq 2e^{-t^2 d/2} \quad \text{and} \quad \Pr[f < \text{med}(f) - t] \leq 2e^{-t^2 d/2}.$$

distribution of U^T is invariant under left (or right) translations. But if U_0 is a fixed orthogonal matrix, then $U^T U_0 = ((U_0)^T U)^T$. Now, $(U_0)^T$ is a fixed orthogonal matrix, hence $(U_0)^T U$ has the same distribution as U , and therefore, $U^T U_0$ has the same distribution as U^T . This shows right invariance, and the proof of left invariance is analogous.

(Where the probabilities are considered with respect to the uniform distribution on the sphere.)

Proof. Let $m := \text{med}(f)$ and $A := \{x \in \mathbb{S}^{d-1} : f(x) \leq m\}$. By Lemma 2.11, we have $P(A) = 1 - \Pr[f > m] \geq 1/2$. Moreover, since f is 1-Lipschitz, we have $f(x) \leq m + t$ for all $x \in A_t = \{x \in \mathbb{S}^{d-1} : \text{dist}(x, A) \leq t\}$. Thus, $\Pr[f > m + t] \leq 1 - P(A_t)$, so the first inequality follows from Theorem 1.11, and the second inequality is derived analogously. \square

2.6 The Proof of the Flattening Theorem

Lemma 2.13 (Concentration of the length of the projection). *Fix integers $1 \leq k \leq d$. For a unit vector $x = (x_1, \dots, x_d) \in \mathbb{S}^{d-1}$, let $f(x) := \sqrt{\sum_{i=1}^k x_i^2}$ be the length of the projection of x onto the subspace L_0 spanned by first k coordinate vectors. Then there exists a number $m = m(d, k) > 0$ such that for all $t > 0$,*

$$\Pr[f > m + t] \leq 2e^{-t^2 d/2} \quad \text{and} \quad \Pr[f < m - t] \leq 2e^{-t^2 d/2}.$$

Moreover, for d larger than some suitable constant d_0 and $k \geq 10 \ln d$, we have $m \geq \frac{1}{2} \sqrt{k/d}$.

Proof. The first part of the lemma follows immediately from Levy's Lemma if we set $m := \text{med}(f)$. It remains to prove the lower bound for m . For this, we use the following trick: Consider a random point $x = (x_1, \dots, x_d) \in \mathbb{S}^{d-1}$. We have $1 = \mathbf{E}[\|x\|_2^2] = \sum_{i=1}^d \mathbf{E}[x_i^2]$, by linearity of expectation. By symmetry, all $\mathbf{E}[x_i^2]$ are equal, so we have $\mathbf{E}[x_1^2] = \dots = \mathbf{E}[x_d^2] = 1/d$. It follows that $\mathbf{E}[f^2] = k/d$. We show that by concentration, the last expectation cannot be much larger than m^2 . We have

$$\frac{k}{d} = \mathbf{E}[f^2] \leq \underbrace{\Pr[\leq m + t]}_{\leq 1} (m + t)^2 + \underbrace{\max_{x \in \mathbb{S}^{d-1}} f(x)^2}_{=1} \underbrace{\Pr[f > m + t]}_{\leq 2e^{-t^2 d/2}}.$$

Setting $t = \sqrt{k/(5d)}$ and using $k \geq 10 \ln d$ we get $2e^{-t^2 d/2} \leq 2/d$, hence $k/d \leq (m + \sqrt{k/(5d)})^2 + 2/d$, and therefore $m \geq \sqrt{\frac{k-2}{d}} - \sqrt{\frac{k}{5d}} \geq \frac{1}{2} \sqrt{k/d}$. \square

Proof of the Flattening Theorem 2.7. Let X be a set of n points in ℓ_2^d . As remarked after the statement of the theorem, we may assume that $d = n$ (by restricting ourselves to the affine hull of X , if necessary). Moreover, since we are trying to prove an asymptotic statement for large n and small ε , we may assume that n is at least some sufficiently large constant n_0 and that $200\varepsilon^{-2} \geq 10$. Set $k := \lceil 200\varepsilon^{-2} \ln n \rceil$.

We want to show that there exists a map $p : X \rightarrow \ell_2^k$ of distortion at most $(1 + \varepsilon)$. If $k \geq n$, there is nothing to prove. Otherwise, let L be a random k -dimensional subspace of \mathbb{R}^n , and let $p : \mathbb{R}^d \rightarrow L$ be the orthogonal projection onto L .

By Fact 2.8 and Lemma 2.13, we know that for any fixed unit vector $u \in \mathbb{S}^{n-1}$, the length of $p(u)$ is concentrated around a certain number $m \geq \frac{1}{2}\sqrt{k/d}$.

For any two distinct points x and y in X , let us say that a particular choice of the random subspace L is *good* for the pair $\{x, y\}$ if

$$(1 - \varepsilon/3)m\|x - y\|_2 \leq \|p(x) - p(y)\|_2 \leq (1 + \varepsilon/3)m\|x - y\|_2.$$

Claim. For any given pair $\{x, y\}$, the probability that L is bad for $\{x, y\}$ is at most n^{-2} .

Suppose that we had already proved the claim. Since there are at most $\binom{n}{2}$ pairs of points in X , the probability that a random subspace L is good for all pairs is at least $1/2$. But if L is good, then p has distortion at most $\frac{1+\varepsilon/3}{1-\varepsilon/3} \leq 1 + \varepsilon$ (for $\varepsilon \leq 1$). Thus, it suffices to prove the claim.

Fix a particular pair $\{x, y\}$ of points in x and set $u := \frac{x-y}{\|x-y\|_2}$. By linearity of p , L is good for $\{x, y\}$ iff $(1 - \varepsilon/3)m \leq \|p(u)\|_2 \leq (1 + \varepsilon/3)m$, i.e., iff $|\|p(u)\|_2 - m| \leq \frac{\varepsilon}{3}m$. By concentration of the length of the projection (with $t = \varepsilon m/3$), we have that the probability that L is bad for $\{x, y\}$ is at most $4e^{-\frac{\varepsilon^2 m^2 d}{18}} \leq 4e^{-\varepsilon^2 k/72}$, where we use $m \geq \frac{1}{2}\sqrt{k/d}$. Finally, by taking logarithms, we see that $4e^{-\varepsilon^2 k/72} < n^{-2}$, as claimed, because of our choice of k . \square

2.7 Exercises

Exercise 8. Consider the two 4-point metric spaces given by the shortest path metrics on the two graphs in Figure 2.3. Show that neither of these metric spaces can be isometrically embedded into ℓ_2^d , for any d .

Exercise 9. Let $f : X \rightarrow Y$ be a bijective map between vector spaces. Show that the distortion of f equals $\|f\|_{Lip} \cdot \|f^{-1}\|_{Lip}$.

Exercise 10. Show that for every $d \geq 1$, there is an isometry $f : \ell_1^d \rightarrow \ell_\infty^{2^d}$. (An isometry is a distance preserving or distortion 1 map, i.e., in our case we require that $\|f(x) - f(y)\|_\infty = \|x - y\|_1$ hold for all $x, y \in \mathbb{R}^d$.)

Exercise 11. Suppose you wish to design an algorithm that solves the following problem: Given a finite set $X \subseteq \ell_1^d$, $|X| = n$, compute the diameter $\text{diam}(X) := \max_{x, y \in X} \|x - y\|_1$. What is the runtime of the “naive” algorithm that just computes pairwise distances? Show, using Exercise 10, that there is an algorithm that computes the diameter of a set of n points in ℓ_1^d in time $O(d2^d n)$. (If d is fixed and n is large, this improves upon the naive algorithm.)

Exercise 12. Let (X, ρ) be an arbitrary metric space, $|X| = n$. Show that there is an isometry $f : X \rightarrow \ell_\infty^n$.

Exercise 13. Show that the following inequality holds for all $v_1, v_2, v_3, v_4 \in \mathbb{R}^d$:

$$\|v_1 - v_3\|_2^2 + \|v_2 - v_4\|_2^2 \leq \|v_1 - v_2\|_2^2 + \|v_2 - v_3\|_2^2 + \|v_3 - v_4\|_2^2 + \|v_4 - v_1\|_2^2.$$

(Hint: The above inequality can be expressed as a sum of 1-dimensional inequalities, one for each coordinate.)

The goal of the next three exercises is to show that the bound on the target dimension in the Johnson-Lindenstrauss Flattening Theorem cannot be improved much.

Exercise 14. (a) Let $A = [a_{ij}]$ be a symmetric real $(n \times n)$ -matrix such that $a_{ii} = 1$ for all i and $|a_{ij}| \leq 1/\sqrt{n}$ for $i \neq j$. Show that $\text{rank } A > n/2$. (Hint: Let $\lambda_1, \dots, \lambda_r$ be the non-zero eigenvalues of A . Show, using traces of matrices, that $\sum_{i=1}^r \lambda_i = n$ and $\sum_{i=1}^r \lambda_i^2 < 2n$, and apply the Cauchy-Schwarz Inequality in a clever way.)

(b) Suppose that A satisfies the assumptions from Part (a), except that A is not necessarily symmetric. Show that we can still conclude $\text{rank}(A) > n/4$. (Hint: How do you symmetrize a matrix? By which factor can its rank increase at most in the process?)

Exercise 15. Let $B = [b_{ij}]$ be a real $(n \times n)$ -matrix, and let $f(x) \in \mathbb{R}[x]$ be a polynomial (in one variable) of degree k . Define a matrix $C = [c_{ij}]$ by applying f to each entry of B separately, i.e., $c_{ij} = f(b_{ij})$ for $1 \leq i, j \leq n$. Show that

$$\text{rank}(C) \leq \binom{k+r}{r} \leq \left(\frac{e(k+r)}{k} \right)^k.$$

(Hint: Suppose without loss of generality that the first r rows of B linearly span all other rows. For integers $0 \leq k_1, \dots, k_r$ with $\sum_{i=1}^r k_i \leq k$, define a vector $v^{(k_1, \dots, k_r)} \in \mathbb{R}^n$ by $v_j^{(k_1, \dots, k_r)} := \prod_{i=1}^r b_{ij}^{k_i}$ and show that the span of these vectors contains all rows of C .)

Exercise 16. (a) Let $B = [b_{ij}]$ be a real $(n \times n)$ -matrix such that $b_{ii} = 1$ for all i and $|b_{ij}| \leq \varepsilon$ for $i \neq j$, where $1/\sqrt{n} \leq \varepsilon \leq 1/2$. Show that

$$\text{rank}(B) \geq \Omega \left(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)} \right).$$

(Hint: Apply Exercises 14 (b) and 15. To what power do you have to raise ε to make it smaller than $1/\sqrt{n}$?)

- (b) Consider the set $X = \{0, e_1, \dots, e_n\} \subset \mathbf{R}^n$ (where the e_i 's are the vectors of the standard orthonormal basis). Suppose that this set of points (with their Euclidean distances) can be mapped with distortion at most $(1 + \varepsilon)$ into ℓ_2^k (i.e., into \mathbf{R}^k with Euclidean distances). Show that then there exist $v_1, \dots, v_n \in \mathbf{R}^k$ that are "almost orthogonal" unit vectors, i.e., $\|v_i\|_2 = 1$ for all i and $|\langle v_i, v_j \rangle| \leq 100\varepsilon$ (the constant 100 could be improved).
- (c) Assuming that there is a low-distortion map as in Part (b) and $\frac{1}{100\sqrt{n}} \leq \varepsilon \leq 1/2$, show that

$$k \geq \Omega \left(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)} \right).$$