Lecture notes on metric embeddings

Jiří Matoušek

Department of Applied Mathematics
Charles University, Malostranské nám. 25
118 00 Praha 1, Czech Republic, and

Institute of Theoretical Computer Science
ETH Zurich, 8092 Zurich, Switzerland

January 1, 2013

Preface

The area of metric embeddings, or more precisely, *approximate embeddings of metric spaces*, has been developing rapidly at least since the 1990s, when a new strong motivation for it came from computer science. By now it has many deep and beautiful results and numerous applications, most notably for approximation algorithms.

Yet, as far as I know, there is no introductory textbook and no comprehensive monograph. One of the most widely cited general sources happens to be a chapter in my 2002 book *Lectures on Discrete Geometry*. When I was invited to teach a two-week intensive doctoral course in Barcelona in 2009, I decided to do metric embeddings, and instead of just following the just mentioned chapter, I ended up writing brand new lecture notes. Not only they contain more material and newer results, but the presentation of much of the older material has also been reworked considerably. These lecture notes were further polished and extended during one-semester courses I taught in Prague and in Zurich; I would like to thank all the teaching assistants and students involved for great atmosphere and many constructive comments and corrections.

My earlier intention was to extend the notes by adding citations, remarks, more results—in short, to make them into a reasonable textbook. While this plan still exists in principle, up until now there have always been more pressing things to do. After several years of zero progress, I thus decided to make the present version publicly available, although generally I don't like publishing half-baked materials.

Thus, the current version includes almost no references, and many of the key results in the area are not mentioned at all. Still I believe that the covered material constitutes a reasonable foundation for further study of the subject, with a couple of excursions to other areas.

If you want to cite some results treated in these notes, please invest some effort to find the original sources. With the help of modern search technology, plus the several available surveys (e.g., by Linial, by Naor, by Indyk and myself) this should not be too difficult.

Contents

1

On metrics and norms

1.1 Metrics, bacteria, pictures

The concept of distance is usually formalized by the mathematical notion of a *metric*. First we recall the definition:

A **metric space** is a pair (X, d_X), where X is a set and $d_X\colon X\times X\to\mathbb{R}$ is a **metric** satisfying the following axioms (x, y, z are arbitrary points of X):

(M1) $d_X(x,y)\geq 0$,

(M2) $d_X(x,x)=0$,

(M3) $d_X(x,y)>0$ for $x\neq y$,

(M4) $d_X(y,x)=d_X(x,y)$, and

(M5) $d_X(x,y)+d_X(y,z)\geq d_X(x,z)$.

If d_X satisfies all the axioms except for (M3), i.e. distinct points are allowed to have zero distance, then it is called a **pseudometric**. The word *distance* or *distance function* is usually used in a wider sense: Some practically important distance functions fail to satisfy the triangle inequality (M5), or even the symmetry (M4).

Graph metrics. Some mathematical structures are equipped with obvious definitions of distance. For us, one of the most important examples is the **shortest-path metric** of a graph.

Given a graph G (simple, undirected) with vertex set V, the distance of two vertices u, v is defined as the length of a shortest path connecting u and v in G, where the length of a path is the number of its edges. (We need to assume G connected.)

As a very simple example, the complete graph K_n yields the n-point **equilateral space**, where every two points have distance 1.

More generally, we can consider a **weighted graph** G, where each edge $e\in E(G)$ is assigned a positive real number $w(e)$, and the length of a path is measured as the sum of the weights of its edges. (The previous case, where there are no edge weights, is sometimes referred to as an *unweighted graph*, in order to distinguish it from the weighted case.)

We will first consider graph metrics as a convenient and concise way of specifying a finite metric space. However, we should mention that several natural classes of graphs give rise to interesting classes of metric spaces. For example, the class of **tree metrics** consists of all metrics of weighted trees and all of their (metric) subspaces; here by a tree we mean a finite connected acyclic graph). Similarly one can consider **planar-graph metrics** and so on.

The relations between graph-theoretic properties of G and properties of the corresponding metric space are often nontrivial and, in some cases, not yet understood.

The importance of being metric. As we have seen in the case of graphs, some mathematical structures are equipped with obvious definitions of distance among their objects. In many other cases, mathematicians have invented clever definitions of a metric in order to prove results about the considered structures. A nice example is the application of Banach's contraction principle for establishing the existence and uniqueness of solutions for differential equations.

Metric spaces also arise in abundance in many branches of science. Whenever we have a collection of objects and each object has several numerical or non-numerical attributes (age sex salary... think of the usual examples in introduction to programming), we can come up with various methods for computing the distance of two objects.

A teacher or literary historian may want to measure the distance of texts in order to attribute authorship or to find plagiarisms. Border police of certain countries need (?!?!) to measure the distance of fingerprints in order to match your fingerprints to their database—even after your pet hamster bites your finger.

My first encounter with metric embeddings occurred through bacteria in the late 1980s. There are enormous number of bacterial species, forms,

and mutations, and only very few of them can be distinguished visually. Yet classifying a bacterial strain is often crucial for curing a disease or stopping an epidemic.

Microbiologists measure the distance, or *dissimilarity* as it is more often called, of bacterial strains using various sophisticated tests, such as the reaction of the bacteria to various chemicals or sequencing portions of their DNA. The raw result of such measurements may be a table, called a *distance matrix*, specifying the distance for every two strains. For the following tiny example, I've picked creatures perhaps more familiar than bacterial species; the price to pay is that the numbers are completely artificial:

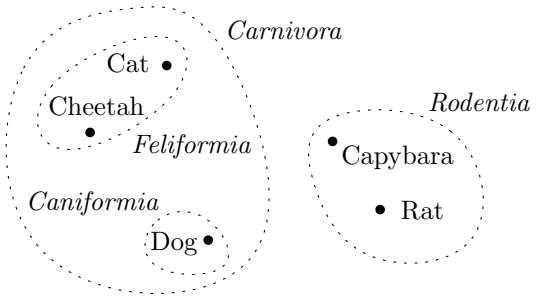
	Dog	Cat	Cheetah	Rat	Capybara
Dog	0				
Cat	0.50	0			
Cheetah	0.42	0.27	0		
Rat	0.69	0.69	0.65	0	
Capybara	0.72	0.61	0.59	0.29	0

(the entries above the diagonal are omitted because of symmetry).

It is hard to see any structure in this kind of a table. Of course, one should better think of a very large table, with tens or perhaps hundreds of rows and columns. (This is still tiny compared to some other data sets: For example, the number of proteins with known structure ranges in the *hundreds of thousand*, and there are *billions* of human fingerprints.)

Representing the distances in the plane? It would be very nice to be able to represent such data visually: Assign a point in the plane to each of the objects in such a way that the distance of two objects is equal to the Euclidean distance of the corresponding dots. In such a picture, we may be able to distinguish tight clusters, isolated points, and other phenomena of interest at a glance:[1]

[1] This particular drawing, in addition to being completely made up, bears some typical features of pseudo-science, such as using Latin names just to impress the reader, but I hope that it illustrates the point nevertheless.



Storing a distance matrix for n objects in computer memory requires storing n^2 real numbers, or rather $\binom{n}{2}$ real numbers if we omit the entries on the diagonal and above it. On the other hand, if we succeeded in representing the distances by Euclidean distances of suitable n points in the plane, it would be enough to store $2n$ real numbers, namely, the coordinates of the points. For $n = 1000$ the saving is already more than 200-fold. This is another, perhaps less obvious advantage of such a planar representation.

Moreover, a point set in the plane can be processed by various efficient geometric algorithms, which cannot work directly with a distance matrix. This advantage may be the hardest to appreciate at first, but at present it can be regarded as *the* main point of metric embeddings.

All of this sounds very good, and indeed it is too good to be (completely) true.

1.2 Distortion

Impossibility of isometric embeddings. An exact representation of one metric space in another is formalized by the notion of isometric embedding. A mapping $f\colon (X, d_X) \to (Y, d_Y)$ of one metric space into another is called an **isometric embedding** or *isometry* if $d_Y(f(x), f(y)) = d_X(x, y)$ for all $x, y \in X$.

Two metric spaces are **isometric** if there exists a bijective isometry between them.

It is easy to find examples of small metric spaces that admit no isometric embedding into the plane $\mathbb{R}^2$ with the Euclidean metric. One such example is the 4-point equilateral space, with every two points at distance 1. Here an isometric embedding fails to exist (which the reader is invited to check) for "dimensional" reasons. Indeed, this example can be isometrically embedded in Euclidean spaces of dimension 3 and higher.

Perhaps less obviously, there are 4-point metric spaces that cannot be isometrically embedded in *any* Euclidean space, no matter how high the dimension. Here are two examples, specified as the shortest-path metrics of the following graphs:

It is quite instructive to prove the impossibility of isometric embedding for these examples. Later on we will discuss a general method for doing that, but it's worth trying it *now*.

Approximate embeddings. For visualizing a metric space, we need not insist on representing distances exactly—often we don't even *know* them exactly. We would be happy with an approximate embedding, where the distances are not kept exactly but only with some margin of error. But we want to quantify, and control, the error.

One way of measuring the error of an approximate embedding is by its *distortion*.

Let (X, d_X) and (Y, d_Y) be metric spaces. An injective mapping $f\colon (X, d_X) \to (Y, d_Y)$ is called a ***D*-embedding**, where $D \geq 1$ is a real number, if there is a number $r > 0$ such that for all $x, y \in X$,

$$r \cdot d_X(x, y) \leq d_Y(f(x), f(y)) \leq Dr \cdot d_X(x, y).$$

The infimum of the numbers D such that f is a D-embedding is called the **distortion** of f.

Note that this definition permits scaling of all distances in the same ratio r, in addition to the distortion of the individual distances by factors between 1 and D (and so every isometric embedding is a 1-embedding, but not vice versa). If Y is a Euclidean space (or a normed space), we can re-scale the image at will, and so we can choose the scaling factor r at our convenience.

The distortion is not the only possible or reasonable way of quantifying the error of an approximate embedding of metric spaces, and a number of other notions appear in the literature. But the distortion is the most widespread and most fruitful of these notions so far.

Here is a piece of notation, which may sometimes be useful.

For metric spaces (X, d_X) and (Y, d_Y), let

$c_{(Y,d_Y)}(X, d_X) := \inf\{D : \text{there exists a } D\text{-embedding } (X, d_X) \to (Y, d_Y)\}$

(various parts of this notation are often amputated, e.g., we write only $c_Y(X)$ if the metrics are understood).

Determining or estimating $c_Y(X)$ for specific X and Y is often difficult and this kind of problems will occupy us for a large part of the time.

Lipschitz and bi-Lipschitz maps. Another view of distortion comes from analysis. Let us recall that a mapping $f\colon (X, d_X) \to (Y, d_Y)$ is called ***C*-Lipschitz** if $d_Y(f(x), f(y)) \leq C d_X(x, y)$ for all $x, y \in X$. Let

$$\|f\|_{\mathrm{Lip}} := \sup\left\{\frac{d_Y(f(x), f(y))}{d_X(x, y)} : x, y \in X,\ x \neq y\right\},$$

the *Lipschitz norm* of f, be the smallest possible C such that f is C-Lipschitz. Now if f is a bijective map, it is not hard to check that its distortion equals $\|f\|_{\mathrm{Lip}} \cdot \|f^{-1}\|_{\mathrm{Lip}}$. For this reason, maps with a finite distortion are sometimes called *bi-Lipschitz*.

Go to higher dimension, young man. We have used the problem of visualizing a metric space in the plane for motivating the notion of distortion. However, while research on low-distortion embeddings can be declared highly successful, this specific goal, low-distortion embeddings in $\mathbb{R}^2$, is too ambitious.

First, it is easy to construct an n-point metric space, for all sufficiently large n, whose embedding in $\mathbb{R}^2$ requires distortion at least $\Omega(\sqrt{n}\,)$,[2] and a slightly more sophisticated construction results in distortion at least $\Omega(n)$, much too large for such embeddings to be useful.

Second, it is computationally intractable (in a rigorously defined sense) to determine or approximate the smallest possible distortion of an embedding of a given metric space in $\mathbb{R}^2$.

We thus need to revise the goals—what kind of low-distortion embeddings we want to consider.

The first key to success is to replace $\mathbb{R}^2$ by a more suitable target space. For example, we may use a Euclidean space of sufficiently large dimension or some other suitable normed space. By embedding a given finite metric

[2] A reminder of asymptotic notation: $f(n) = O(g(n))$ means that there are n_0 and C such that $f(n) \leq Cg(n)$ for all $n \geq n_0$; $f(n) = o(g(n))$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$; $f(n) = \Omega(g(n))$ is the same as $g(n) = O(f(n))$, and $f(n) = \Theta(g(n))$ means that both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

space into such a target space, we have "geometrized" the problem and we can now apply geometric methods and algorithms. (This can be seen as a part of a current broader trend of "geometrizing" combinatorics and computer science.)

Moreover, we also revise what we mean by "low distortion". While for visualization distortion 1.2 can be considered reasonable and distortion 2 already looks quite large, in other kinds of applications, mainly in approximation algorithms for NP-hard problems, we will be grateful for embeddings with distortion like $O(\log n)$, where n is the number of points of the considered metric space.

We will see later how these things work in concrete examples, and so we stop this abstract discussion for now and proceed with recalling some basics on norms.

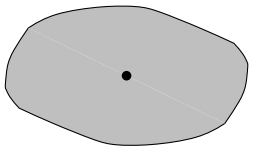
1.3 Normed spaces

A metric can be defined on a completely arbitrary set, and it specifies distances for pairs of points. A norm is defined only on a vector space, and for each point it specifies its distance from the origin.

By definition, a **norm** on a real vector space Z is a mapping that assigns a nonnegative real number $\|\mathbf{x}\|$ to each $\mathbf{x} \in Z$ so that $\|\mathbf{x}\| = 0$ implies $\mathbf{x} = 0$, $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for all $\alpha \in \mathbb{R}$, and the triangle inequality holds: $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$.

Every norm $\|\mathbf{x}\|$ on Z defines a metric, in which the distance of points $\mathbf{x}, \mathbf{y}$ equals $\|\mathbf{x} - \mathbf{y}\|$. However, by far not all metrics on a vector space come from norms.

For studying a norm $\|.\|$, it is usually good to look at its *unit ball* $\{\mathbf{x} \in Z : \|\mathbf{x}\| \le 1\}$. For a general norm in the plane it may look like this, for instance:



It is easy to check that the unit ball of any norm is a closed convex body K that is symmetric about $\mathbf{0}$ and contains $\mathbf{0}$ in the interior. Conversely, any $K \subset Z$ with the listed properties is the unit ball of a (uniquely determined) norm, and so norms and symmetric convex bodies can be regarded as two views of the same class of mathematical objects.

The ℓ_p norms. Two norms will play main roles in our considerations: the Euclidean norm and the ℓ_1 norm. Both of them are (distinguished) members of the noble family of ℓ_p norms.

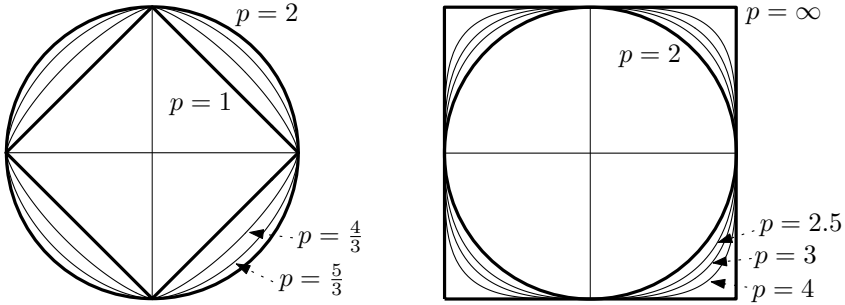
For a point $\mathbf{x} = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ and for $p \in [1, \infty)$, the **ℓ_p norm** is defined as

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{k} |x_i|^p \right)^{1/p}.$$

We denote by $\boldsymbol{\ell_p^k}$ the normed space $(\mathbb{R}^k, \|.\|_p)$.

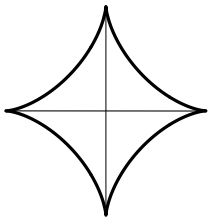
The Euclidean norm is $\|.\|_2$, the ℓ_2 norm. The ℓ_∞ norm, or *maximum norm*, is given by $\|x\|_\infty = \max_i |x_i|$. It is the limit of the ℓ_p norms as $p \to \infty$.

To gain some feeling about ℓ_p norms, let us look at their unit balls in the plane:



The left picture illustrates the range $p \in [1, 2]$. For $p = 2$ we have, of course, the ordinary disk, and as p decreases towards 1, the unit ball shrinks towards the tilted square. Only this square, the ℓ_1 unit ball, has sharp corners—for all $p > 1$ the ball's boundary is differentiable everywhere. In the right picture, for $p \ge 2$, one can see the unit ball expanding towards the square as $p \to \infty$. Sharp corners appear again for the ℓ_∞ norm.

The case $p < 1$. For $p \in (0, 1)$, the formula $\|\mathbf{x}\|_p = \big(|x_1|^p + \cdots + |x_k|^p\big)^{1/p}$ still makes sense, but it no longer defines a norm—the unit ball is not convex, as the next picture illustrates for $p = \frac{2}{3}$.



However, $d_p(\mathbf{x},\mathbf{y}) := |x_1-y_1|^p + \cdots + |x_k-y_k|^p$ does define a metric on $\mathbb{R}^k$, which may be of interest for some applications. The limit for $p=0$ is the number of coordinates in which $\mathbf{x}$ and $\mathbf{y}$ differ, a quite useful combinatorial quantity. One can regard $d_p(\mathbf{x},\mathbf{y})$ for small $p>0$ as an "analytic" approximation of this quantity.

1.4 ℓ_p metrics

For finite metric spaces, the following notion is crucial.

A metric d on a finite set V is called an **ℓ_p metric** if there exists a natural number k and an isometric embedding of (V,d) into the space ℓ_p^k. For $p=2$ we also speak of a **Euclidean metric**.

An ℓ_p *pseudometric* is defined similarly, but we consider isometric maps into ℓ_p^k that are not necessarily injective.

In the literature, one very often talks about ℓ_p metrics even when, strictly speaking, the considered class should also include pseudometrics that are not metrics. A similar situation prevails for the notions introduced next, such as line metrics and cut metrics. We'll gladly join this slight abuse of terminology.

In addition to ℓ_p metrics, we also introduce the following simpler classes:

A **line metric** on a set V is a (pseudo)metric isometrically embeddable in $\mathbb{R}$ with the usual metric. A **cut metric** is a line metric for which the embedding in the line attains only the values 0 and 1.

Equivalently, δ is a cut metric on a set V if there exists a nonempty proper subset $S\subset V$ such that $\delta(x,y)=1$ if one of x,y lies in S and the other outside S, and $\delta(x,y)=0$ otherwise. (So a cut metric is almost never a metric—a clear example of the abuse of terminology alluded to above.)

As we will see, line metrics and cut metrics can be used as building blocks for decomposing more complicated kinds of metrics.

Metrics as high-dimensional points. Let V be an n-point set. We can represent a metric d on V as a point $\mathbf{d}\in\mathbb{R}^N$, where $N:=\binom{n}{2}$ and

$$\mathbf{d} = \Big(d(u,v) : \{u,v\}\in\tbinom{V}{2}\Big).$$

Thus, the coordinates in $\mathbb{R}^N$ are indexed by unordered pairs of distinct points of V (in some fixed order).[3] Then a class of metrics, say all metrics on the set V, can be regarded as a subset of $\mathbb{R}^N$, and we can think about it in geometric terms, using notions such as convexity.

A much studied example is the **metric cone**

$$\mathcal{M} := \{\mathbf{d}\in\mathbb{R}^N : d \text{ is a pseudometric on } V\}\subset\mathbb{R}^N$$

(here $\mathcal{M}$ depends on n, but this is not shown in the notation). It's easy to see that $\mathcal{M}$ is a convex cone in $\mathbb{R}^N$, where a **convex cone** is a set C such that $\mathbf{x}\in C$ implies $\lambda\mathbf{x}\in C$ for every real $\lambda\ge 0$, and $\mathbf{x},\mathbf{y}\in C$ implies $\mathbf{x}+\mathbf{y}\in C$. The metric cone is a very interesting mathematical object with a complicated structure.

For our purposes, we'll need mainly the cone of ℓ_1 metrics

$$\mathcal{L}_1 := \{\mathbf{d}\in\mathbb{R}^N : d \text{ is an } \ell_1 \text{ pseudometric on } V\}.$$

1.4.1 Proposition. *The set $\mathcal{L}_1$ is a convex cone. Every $\mathbf{d}\in\mathcal{L}_1$ is a sum of line metrics, and also a nonnegative linear combination of cut metrics.*

Proof. Clearly, if $\mathbf{d}\in\mathcal{L}_1$, then $\lambda\mathbf{d}\in\mathcal{L}_1$ for all $\lambda\ge 0$, and so it suffices to verify that if $\mathbf{d},\mathbf{d}'\in\mathcal{L}_1$, then $\mathbf{d}+\mathbf{d}'\in\mathcal{L}_1$. By definition, $\mathbf{d}\in\mathcal{L}_1$ means that there is a mapping $f\colon V\to\mathbb{R}^k$ such that $d(u,v)=\|f(u)-f(v)\|_1$. Similarly, for $\mathbf{d}'$ we have a mapping $f'\colon V\to\mathbb{R}^{k'}$ with $d'(u,v)=\|f'(u)-f'(v)\|_1$. We define a new mapping $g\colon V\to\mathbb{R}^{k+k'}$ by concatenating the coordinates of f and f'; that is,

$$g(u) := \Big(f(u)_1,\dots,f(u)_k,f'(u)_1,\dots,f'(u)_{k'}\Big)\in\mathbb{R}^{k+k'}.$$

[3] Alternatively, we can represent a metric by an $n\times n$ matrix. Then it lies in a vector space of dimension n^2, but since the matrices are symmetric and have zero diagonal, we're again in an N-dimensional subspace. Choosing between these two possibilities is a matter of taste.

The point of $\mathcal{L}_1$ corresponding to g is $\mathbf{d}+\mathbf{d}'$. Thus, $\mathcal{L}_1$ is a convex cone.

Next, we want to see that every $\mathbf{d} \in \mathcal{L}_1$ is a sum of line metrics (regarded as points in $\mathbb{R}^N$, of course). But this is obvious; if $\mathbf{d}$ is represented by a mapping $f\colon V \to \ell_1^k$, then $\mathbf{d}$ is the sum of the k line metrics represented by the k coordinates of f.

Finally, we want to prove that every $\mathbf{d} \in \mathcal{L}_1$ is a nonnegative linear combination of cut metrics. We may assume that $\mathbf{d}$ is a a line metric; let $(x_v : v \in V)$ be points on the real line representing $\mathbf{d}$, i.e., $d(u,v) = |x_u - x_v|$.

We proceed by induction on the number of distinct values of the x_v. For two values, d is already a positive multiple of a cut metric.

Otherwise, let $a = \min_v x_v$ and let b be the second smallest value of the x_v. We set $x'_v := \max(x_v, b)$, $v \in V$, and let d' be the line pseudometric represented by the x'_v. Then $d = d' + (b-a)\delta$, where δ is the cut metric corresponding to the subset $S := \{v \in V : x_v = a\}$, as is easy to check. This finishes the inductive step, and the proposition is proved. □

Generalizing the definition of $\mathcal{L}_1$, for all $p \in [1,\infty)$ we define

$$\mathcal{L}_p := \Big\{ \big(d(u,v)^p : \{u,v\} \in \tbinom{V}{2}\big) : d \text{ is an } \ell_p \text{ pseudometric on } V \Big\}.$$

Thus, the elements of $\mathcal{L}_p$ are pth powers of ℓ_p metrics, rather than ℓ_p metrics themselves.

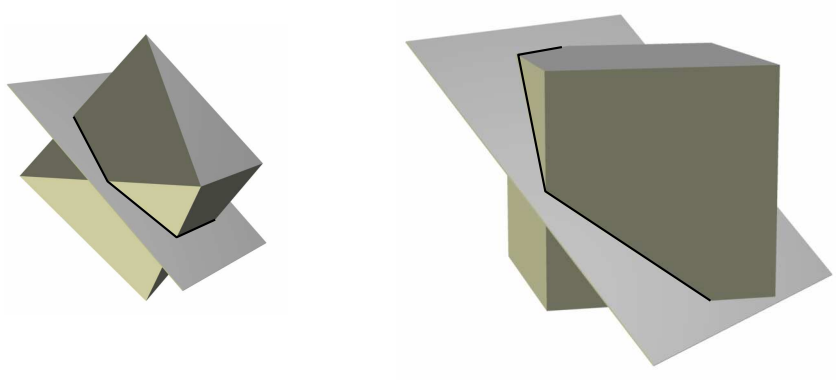
It is immediate that every element of $\mathcal{L}_p$ is a nonnegative linear combination of pth powers of line metrics. By the same trick as in the proof of Proposition 1.4.1, one can also check that $\mathcal{L}_p$ is a convex cone. (This is one of the reasons why we prefer to work with the pth powers of ℓ_p metrics: There are examples showing that the sum of two ℓ_p metrics need not be an ℓ_p metric, and so the set of all ℓ_p metrics is not convex.)

Dimension of isometric embeddings. The definition of an ℓ_p metric prompts a question: How high do we need to go with the dimension k in order to represent all possible ℓ_p metrics on n points?

For $p = 2$, the answer is easy: $k = n-1$ always suffices and it is sometimes necessary. Indeed, given any n points in $\mathbb{R}^k$, we can assume, after translation, that one of the points is $\mathbf{0}$, and then the remaining points span a linear subspace of dimension at most $n-1$. Now the restriction of the Euclidean norm to any linear subspace is again the Euclidean norm on that subspace; geometrically speaking, a central slice of the Euclidean ball is a Euclidean ball. Thus, the given n points can always be assumed to live in ℓ_2^{n-1}. On the other hand, it can be shown that the n-point

equilateral set (every two points at distance 1) cannot be isometrically embedded in a Euclidean space dimension smaller than $n-1$.

For $p \neq 2$ this kind of argument breaks down, since a central slice of the ℓ_p ball is seldom an ℓ_p ball. The picture illustrates this for 2-dimensional slices of 3-dimensional unit balls, for the ℓ_1 norm (the regular octahedron) and for the ℓ_∞ norm (the cube):



In both of the depicted cases, the slice happens to be a regular hexagon.

A completely different method is needed to show the following weaker bound on the dimension.

1.4.2 Proposition. *Every n-point space with an ℓ_p metric is isometrically embeddable in ℓ_p^N, where $N := \binom{n}{2}$.*

Proof. We use the geometry of the set $\mathcal{L}_p$ defined earlier. We know that $\mathcal{L}_p$ is a convex cone in $\mathbb{R}^N$. A version of *Carathéodory's theorem* from convex geometry tells us that if some point $\mathbf{x}$ is a nonnegative linear combination of points of a set $S \subseteq \mathbb{R}^N$, then $\mathbf{x}$ is a nonnegative linear combination of some at most N points of S. (A perhaps more familiar version of Carathéodory's theorem asserts that if a point $\mathbf{x} \in \mathbb{R}^N$ belongs to the convex hull of a set S, then $\mathbf{x}$ lies in the convex hull of some at most $N+1$ points of S. This statement also easily implies the "cone version" needed here.)

In our situation, this shows that the pth power of every ℓ_p metric on V is a nonnegative linear combination of at most N pth powers of line pseudometrics, and thus it embeds isometrically in ℓ_p^N. □

1.4.3 Corollary. *Let (V, d_V) be a finite metric space and suppose that for every $\varepsilon > 0$ there is some k such that (V, d_V) admits a $(1+\varepsilon)$-embedding in ℓ_p^k. Then d_V is an ℓ_p metric.*

Proof. Let $\Delta := \operatorname{diam}(V)$ be the largest distance in (V, d_V). For every $\varepsilon > 0$ there is a $(1+\varepsilon)$-embedding $f_\varepsilon \colon (V, d_V) \to \ell_p^N$, $N = \binom{|V|}{2}$ by Proposition 1.4.2.

By translation we can make sure that the image always lies in the 2Δ-ball around $\mathbf{0}$ in ℓ_p^N (assuming $\varepsilon \le 1$, say); here it is crucial that the dimension is the same for all ε. By compactness there is a cluster point of these embeddings, i.e., a mapping $f \colon V \to \ell_p^N$ such that for every $\eta > 0$ there is some f_ε with $\|f(v) - f_\varepsilon(v)\|_p \le \eta$ for all $v \in V$. Then f is the desired isometry. □

Infinite dimensions. The ℓ_p norms have been investigated mainly in the theory of Banach spaces, and the main interest in this area is in *infinite-dimensional* spaces. With some simplification one can say that there are two main infinite-dimensional spaces with the ℓ_p norm:

- The "small" ℓ_p, consisting of of all infinite sequences $\mathbf{x} = (x_1, x_2, \ldots)$ of real numbers with $\|\mathbf{x}\|_p < \infty$, where $\|\mathbf{x}\|_p = \bigl(\sum_{i=1}^\infty |x_i|^p\bigr)^{1/p}$.

- The "big" $L_p = L_p(0,1)$, consisting of all measurable functions $f \colon [0,1] \to \mathbb{R}$ such that $\|f\|_p := \bigl(\int_0^1 |f(x)|^p \,\mathrm{d}x\bigr)^{1/p}$ is finite. (Well, the elements of L_p are really *equivalence classes* of functions, with two functions equivalent if they differ on a set of measure zero... but never mind.)

As introductory harmonic analysis teaches us, the spaces ℓ_2 and L_2 are isomorphic, and both of them are realizations of the countable *Hilbert space*. For all $p \ne 2$, though, ℓ_p and L_p are substantially different objects.

For us, it is good to know that these infinite-dimensional spaces bring nothing new compared to finite dimensions as far as finite subspaces are concerned. Namely, an ℓ_p metric can be equivalently defined also by isometric embeddability into ℓ_p or by isometric embeddability into L_p. This follows from an approximation argument and Corollary 1.4.3. It gives us additional freedom in dealing with ℓ_p metrics: If desired, we can think of the points as infinite sequences in ℓ_p or as functions in L_p.

1.5 Inclusions among the classes of ℓ_p metrics

From the formula $\|\mathbf{x}\|_p = \bigl(|x_1|^p + \cdots + |x_k|^p\bigr)^{1/p}$ it is probably not clear the value of p should matter much for the properties of ℓ_p metrics, but one of the main facts about ℓ_p metrics is that it matters a *lot*.

We will first summarize the main facts about the relations among the classes of ℓ_p metrics for various p. Let us temporarily denote the class of all (finite!) ℓ_p metrics by $\mathbb{L}_p$.

(i) The ℓ_∞ metrics are the richest: *Every* finite metric belongs to $\mathbb{L}_\infty$.

(ii) The Euclidean metrics are the most restricted: We have $\mathbb{L}_2 \subset \mathbb{L}_p$ for every $p \in [1, \infty)$.

(iii) For $p \in [1,2]$, the richness of ℓ_p metrics grows as p decreases. Namely, $\mathbb{L}_p \subset \mathbb{L}_q$ whenever $1 \le q < p \le 2$. In particular, $\mathbb{L}_1$ is the richest in this range.

(iv) The inclusions mentioned in (i)–(iii) exhaust *all* containment relations among the classes $\mathbb{L}_p$ In particular, for $p > 2$, the classes $\mathbb{L}_p$ are great individualists: None of them contains any other $\mathbb{L}_q$ *except* for $\mathbb{L}_2$, and none of them is contained in any other $\mathbb{L}_q$ *except* for $\mathbb{L}_\infty$.

What is more, the inclusion relations of these classes doesn't change by allowing a bounded distortion: Whenever p, q are such that $\mathbb{L}_p \not\subset \mathbb{L}_q$ according to the above, then $\mathbb{L}_p$ contains metrics requiring arbitrarily large distortions for embedding into ℓ_q.

Part (i) is the only one among these statements that has a simple proof, and we will present it at the end of this section.

Dvoretzky's theorem and almost spherical slices. Part (ii) looks like something that should have a very direct and simple proof, but it doesn't.

It can be viewed as a special case of an amazing Ramsey-type result known as *Dvoretzky's theorem*. It can be stated as follows: *For every $k \ge 1$ and every $\varepsilon > 0$ there exists $n = n(k, \varepsilon)$ with the following property: Whenever $(\mathbb{R}^n, \|.\|)$ is an n-dimensional normed space with some arbitrary norm $\|.\|$, there is a linear embedding $T \colon (\mathbb{R}^k, \|.\|_2) \to (\mathbb{R}^n, \|.\|)$ with distortion at most $1+\varepsilon$. That is, we have $\|\mathbf{x}\|_2 \le \|T\mathbf{x}\| \le (1+\varepsilon)\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{R}^k$.*

In particular, for every k and ε there is some n such that ℓ_2^k can be $(1+\varepsilon)$-embedded in ℓ_p^n. It follows that for every $\varepsilon > 0$, every Euclidean

metric $(1+\varepsilon)$-embeds into ℓ_p^n for some n, and Corollary 1.4.3 tells us that every Euclidean metric is an ℓ_p metric.

If we consider the unit ball of the norm $\|.\|$ as in Dvoretzky's theorem, we arrive at the following geometric version of the theorem: *For every $k \geq 1$ and every $\varepsilon > 0$ there exists $n = n(k,\varepsilon)$ with the following property: Whenever K is a closed n-dimensional convex body in $\mathbb{R}^n$ symmetric*[4] *about $\mathbf{0}$, there exists a k-dimensional linear subspace E of $\mathbb{R}^n$ such that the slice $K \cap E$ is $(1+\varepsilon)$-spherical; that is, for some $r > 0$ it contains the Euclidean ball of radius r and is contained in the Euclidean ball of radius $(1+\varepsilon)r$.* Applying this view to ℓ_∞ and ℓ_1, we get that the n-dimensional unit cube and the n-dimensional unit ℓ_1 ball (the "generalized octahedron") have k-dimensional slices that are almost perfect Euclidean balls—certainly a statement out of range of our 3-dimensional geometric intuition.

In addition, it turns out that the cube has much *less* round slices than the ℓ_1 ball. Namely, given n and assuming ε fixed, say $\varepsilon = 0.1$, let us ask, what is the largest dimension k of a $(1+\varepsilon)$-spherical slice. It turns out that for the cube, the largest k is of order $\log n$, and this is also essentially the worst case for Dvoretzky's theorem—*every* n-dimensional symmetric convex body has $(1+\varepsilon)$-spherical slices about this big. On the other hand, for the ℓ_1 ball (and, for that matter, for all ℓ_p balls with $p \in [1,2]$), the slice dimension k is actually $\Omega(n)$ (with the constant depending on ε, of course). An intuitive reason why the ℓ_1 ball is much better than the cube is that it has many more facets: 2^n, as opposed to $2n$ for the cube.

Stated slightly differently, ℓ_2^k can be $(1+\varepsilon)$-embedded, even linearly, in ℓ_1^{Ck} for a suitable $C = C(\varepsilon)$. We will prove this later on, using probabilistic tools. The problem of constructing such an embedding explicitly is open, fascinating, related to many other explicit or pseudorandom constructions in combinatorics and computational complexity, and subject of intensive research.

Euclidean metrics are ℓ_1 metrics. What we can do right now is a proof that every ℓ_2 metric is also an ℓ_1 metric. We actually embed all of ℓ_2^k isometrically into the infinite-dimensional space $L_1(S^{k-1})$. What is that? Similar to $L_1 = L_1(0,1)$, the elements of $L_1(S^{k-1})$ are (equivalence classes of) measurable real functions, but the domain is the $(k-1)$-dimensional unit Euclidean sphere S^{k-1}. The distance of two functions f, g is $\|f-g\|_1 = \int_{S^{k-1}} |f(\mathbf{u}) - g(\mathbf{u})| \, \mathrm{d}\mathbf{u}$, where we integrate according to the uniform (rotation-invariant) measure on S^{k-1}, scaled so that the whole of S^{k-1} has measure 1.

[4] The symmetry assumption can be dropped.

The embedding $F\colon \ell_2^k \to L_1(S^{k-1})$ is defined as $F(\mathbf{x}) := f_{\mathbf{x}}$, where $f_{\mathbf{x}}\colon S^{k-1} \to \mathbb{R}$ is the function given by $f_{\mathbf{x}}(\mathbf{u}) := \langle \mathbf{x}, \mathbf{u} \rangle$.

Let us fix some $\mathbf{v}_0 \in \ell_2^k$ with $\|\mathbf{v}_0\|_2 = 1$, and set $C := \|F(\mathbf{v}_0)\|_1 = \int_{S^{k-1}} |\langle \mathbf{v}_0, \mathbf{u} \rangle| \, \mathrm{d}\mathbf{u}$. By rotational symmetry, and this is the beauty of this proof, we have $\|F(\mathbf{v})\|_1 = C$ for every unit $\mathbf{v} \in \ell_2^k$, and hence in general $\|F(\mathbf{x})\|_1 = C\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \ell_2^k$. Since $F(\mathbf{x}) - F(\mathbf{y}) = F(\mathbf{x}-\mathbf{y})$, we see that F scales all distances by the same factor C, and so after re-scaling we obtain the desired isometry.

This is all nice, but how do we know that all finite subspaces of $L_1(S^{k-1})$ are ℓ_1 metrics? With some handwaving we can argue like this: If we choose a "sufficiently uniformly distributed" finite set $A \subseteq S^{k-1}$, then integral of every "reasonable" function f on S^{k-1}, such as our functions $f_{\mathbf{x}}$, over S^{k-1} can be approximated by the average of the function over A. In symbols, $\|f\|_1 \approx \frac{1}{|A|} \sum_{\mathbf{u} \in A} |f(\mathbf{u})|$. In this way, we can $(1+\varepsilon)$-embed a given finite subset of ℓ_2^k into the space of all real functions defined on A with the ℓ_1 norm, and the latter is isomorphic to $\ell_1^{|A|}$. As in one of the earlier arguments in this section, Proposition 1.4.2 and compactness allow us to conclude that every ℓ_2 metric is also an ℓ_1 metric.

The Fréchet embedding. We will prove that *every n-point metric space (X, d_X) embeds isometrically in ℓ_∞^n*. The proof, due to Fréchet, is very simple but it brings us to a useful mode of thinking about embeddings.

Let us list the points of X as $x_1, x_2, \ldots, x_n$. To specify a mapping $f\colon X \to \ell_\infty^n$ means to define n functions $f_1, \ldots, f_n\colon X \to \mathbb{R}$, the coordinates of the embedded points. Here we set

$$f_i(x_j) := d_X(x_i, x_j).$$

One needs to check that this indeed defines an isometry. This we leave to the reader—as the best way of understanding how the embedding works, which will be useful later on.

Which p? That is, if we have a collection of objects with a large number k (say 20 or more) attributes, such as a collection of bacterial strains in the motivating example, how should we measure their distance? We assume that the considered problem itself doesn't suggest a particular distance function and that we can reasonably think of the attributes as coordinates of points in $\mathbb{R}^k$.

An obvious suggestion is the Euclidean metric, which is so ubiquitous and mathematically beautiful. However, some theoretical and empirical studies indicate that this may sometimes be a poor choice.

For example, let us suppose that the dimension k is not very small compared to n, the number of points, and let us consider a random n-point set $X \subset \mathbb{R}^k$, where the points are drawn independently from the uniform distribution in the unit ball or unit cube, say. It turns out that with the Euclidean metric, X is typically going to look almost like an equilateral set, and thus metrically uninteresting.

On the other hand, this "equalizing" effect is much weaker for ℓ_p norms with $p < 2$, with $p = 1$ faring the best (the metrics d_p with $p \in (0, 1)$ are even better, but harder to work with). Of course, real data sets are seldom purely random, but still this can be regarded as an interesting heuristic reason for favoring the ℓ_1 norm over the Euclidean one.

1.6 Exercises

1. (a) Show that every embedding of n-point equilateral space (every two points have distance 1) into the plane $\mathbb{R}^2$ with the usual Euclidean metric has distortion at least $\Omega(\sqrt{n})$.

   (b) Give an embedding with distortion $O(\sqrt{n})$.

2. Show that every embedding of the cycle of length n (with the graph metric) into the line $\mathbb{R}^1$ with the usual metric has distortion at least $\Omega(n)$.

3. True or false? There is a function $\phi(n)$ with $\lim_{n\to\infty} \frac{\phi(n)}{n} = 0$ such that every n-vertex tree (shortest-path metric, unit-length edges) can be embedded into $\mathbb{R}^1$ with distortion at most $\phi(n)$.

4.* Show that every finite tree metric space can be embedded isometrically into ℓ_1. (Slightly less ambitiously, you can start with embedding all trees with unit-length edges.)

5. (a)* Let T be a tree on $n \geq 3$-vertices. Prove that there exist subtrees T_1 and T_2 of T that share a single vertex and no edges and together cover T, such that $\max\{|V(T_1)|, |V(T_2)|\} \leq 1 + \frac{2}{3}n$.

   (b) Using (a), prove that every n-point tree can be isometrically embedded into ℓ_∞^k with $k = O(\log n)$.

6. (a)** Show that every embedding of the graph $K_{2,3}$ (with the shortest-path metric) into ℓ_1 has distortion at least 4/3.

   (b)* Show that this bound is tight.

7. Describe an isometric embedding of ℓ_1^2 into ℓ_∞^2.

8. Show that if the unit ball K of some finite-dimensional normed space is a convex polytope with $2m$ facets, then that normed space embeds isometrically into ℓ_∞^m.

   (Using results on approximation of convex bodies by polytopes, this yields useful approximate embeddings of arbitrary norms into ℓ_∞^k.)

9. (a) Let $k \geq 1$. Give an isometric embedding of ℓ_1^k to $\ell_\infty^{2^k}$.

   (b) Devise an algorithm that, given a set X of n points in $\mathbb{R}^k$, computes the diameter of X under the ℓ_1 norm using $O(k2^k n)$ arithmetic operations.

   (c) Can you reduce the number of arithmetic operations to $O(2^k n)$ (or even further)?

10. (a) Determine the distortion of the identity mapping $(\mathbb{R}^k, \|.\|_2) \to (\mathbb{R}^k, \|.\|_1)$.

    (b) Determine the distortion of the identity mapping $(\mathbb{R}^k, \|.\|_1) \to (\mathbb{R}^k, \|.\|_\infty)$.

    (c)** Find a mapping between the spaces as in (b) with a smaller distortion. (How small can you make it?)

11.* Show that every n-vertex graph with unit-length edges can be embedded into $\mathbb{R}^1$ with distortion $O(n)$. (You may want to try solving the problem for trees first.)

12. (a)* Prove that for every four points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ in a Euclidean space of an arbitrary dimension, we have

$$\begin{aligned}\|\mathbf{x}_1 - \mathbf{x}_3\|_2^2 + \|\mathbf{x}_2 - \mathbf{x}_4\|_2^2 \quad \leq \quad & \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \|\mathbf{x}_2 - \mathbf{x}_3\|_2^2 \\ + \quad & \|\mathbf{x}_3 - \mathbf{x}_4\|_2^2 + \|\mathbf{x}_4 - \mathbf{x}_1\|_2^2.\end{aligned}$$

    (b) Using (a), find the minimum necessary distortion for embedding the 4-cycle into ℓ_2 (i.e., into a Euclidean space of arbitrary dimension).

13.* Using a method similar to the one in Exercise 12, find the minimum necessary distortion for embedding the 3-star (in other words, the complete bipartite graph $K_{1,3}$) in ℓ_2.

14.** Let S^2 denote the 2-dimensional unit sphere in $\mathbb{R}^3$. Let $X \subset S^2$ be an n-point set. Show that X with the Euclidean metric can be embedded into the Euclidean plane with distortion at most $O(\sqrt{n})$.

15.** Show that the complete binary B_m tree of height m (with the graph metric) can be embedded into ℓ_2 with distortion $O(\sqrt{\log m}\,)$.

16.* (Almost Euclidean subspaces) Prove that for every k and $\varepsilon > 0$ there exists $n = n(k, \varepsilon)$ such that every n-point metric space (X, d_X) contains a k-point subspace that is $(1+\varepsilon)$-embeddable into ℓ_2. Use Ramsey's theorem for graphs.

2

Dimension reduction by random projection

2.1 The lemma

The Johnson–Lindenstrauss lemma is the following surprising fact:[1]

2.1.1 Theorem. *Let $\varepsilon \in (0,1)$ be a real number, and let $P = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$ be a set of n points in $\mathbb{R}^n$. Let k be an integer with $k \geq C\varepsilon^{-2}\log n$, where C is a sufficiently large absolute constant. Then there exists a mapping $f\colon \mathbb{R}^n \to \mathbb{R}^k$ such that*

$$(1-\varepsilon)\|\mathbf{p}_i - \mathbf{p}_j\|_2 \leq \|f(\mathbf{p}_i) - f(\mathbf{p}_j)\|_2 \leq (1+\varepsilon)\|\mathbf{p}_i - \mathbf{p}_j\|_2$$

for all $i, j = 1, 2, \ldots, n$.

In the language acquired in the previous chapter, every n-point Euclidean metric space can be mapped in ℓ_2^k, $k = O(C\varepsilon^{-2}\log n)$, with distortion at most $\frac{1+\varepsilon}{1-\varepsilon}$. In still other words, every n-point set in any Euclidean space can be "flattened" to dimension only logarithmic in n, so that no distance is distorted by more that a factor that, for small ε, is roughly $1+2\varepsilon$.

In the formulation of the theorem we haven't used the language of distortion, but rather a slightly different notion, which we turn into a

[1] Traditionally this is called a lemma, since that's what it was in the original paper of Johnson and Lindenstrauss. But it arguably does deserve the status of a theorem.

general definition: Let us call a mapping $f\colon (X, d_X) \to (Y, d_Y)$ of metric spaces an **ε-almost isometry** if $(1-\varepsilon)d_X(x,y) \leq d_Y(f(x), f(y)) \leq (1+\varepsilon)d_X(x,y)$. For ε small this is not very different from saying that f is a $(1+2\varepsilon)$-embedding (at least if the mapping goes into a normed space and we can re-scale the image at will), but it will help us avoid some ugly fractions in the calculations.

It is known that the dependence of k on both ε and n in Theorem 2.1.1 is almost optimal—there is a lower bound of $\Omega((\log n)/(\varepsilon^2 \log \frac{1}{\varepsilon}))$. A lower-bound example is the n-point equilateral set. A volume argument (see Section 3.1) immediately gives that a 2-embedding of the equilateral set needs dimension at least $\Omega(\log n)$, which shows that the dependence on n cannot be improved. On the other hand, the argument for the dependence on ε is not that easy.

All known proofs of Theorem 2.1.1 are based on the following statement, which we call, with some inaccuracy, the random projection lemma, and which for the moment we formulate somewhat imprecisely:

2.1.2 Lemma (Random projection lemma—informal). *Let $T\colon \mathbb{R}^n \to \mathbb{R}^k$ be a "normalized random linear map" and let $\varepsilon \in (0,1)$. Then for every vector $\mathbf{x} \in \mathbb{R}^n$ we have*

$$\mathrm{Prob}\Big[(1-\varepsilon)\|\mathbf{x}\|_2 \leq \|T(\mathbf{x})\|_2 \leq (1+\varepsilon)\|\mathbf{x}\|_2\Big] \geq 1 - 2e^{-c\varepsilon^2 k},$$

where $c > 0$ is a constant (independent of n, k, ε).

The term "normalized random linear map" calls for explanation, but we postpone the discussion. For now, it is sufficient to know that there is *some* probability distribution on the set of linear maps $\mathbb{R}^n \to \mathbb{R}^k$ such that, if T is randomly drawn from this distribution, then it satisfies the conclusion. (It is also important to note what the random projection lemma *doesn't* say: It definitely doesn't claim that a random T is an ε-almost isometry—since, obviously, for $k < n$, a linear map $\mathbb{R}^n \to \mathbb{R}^k$ can't even be injective!)

Proof of Theorem 2.1.1 assuming Lemma 2.1.2. The value of k in the Johnson–Lindenstrauss lemma is chosen so large that Lemma 2.1.2 yields $\mathrm{Prob}[(1-\varepsilon)\|\mathbf{x}\|_2 \leq \|T(\mathbf{x})\|_2 \leq (1+\varepsilon)\|\mathbf{x}\|_2] \geq 1 - n^{-2}$ for every fixed $\mathbf{x}$. We apply this to the $\binom{n}{2}$ vectors $\mathbf{p}_i - \mathbf{p}_j$, $1 \leq i < j \leq n$, and use the union bound. We obtain that T restricted to our set P behaves as an ε-almost isometry with probability at least $\frac{1}{2}$. In particular, a suitable T exists. □

So, how do we choose a "normalized random linear map"? As we will see, there are many possibilities. For example:

(a) (The case of projection to a random subspace) As in the original Johnson–Lindenstrauss paper, we can pick a random k-dimensional linear subspace[2] of $\mathbb{R}^n$ and take T as the orthogonal projection on it, scaled by the factor of $\sqrt{n/k}$. This applies only for $k \le n$, while later we'll also need to use the lemma for $k > n$.

(b) (The Gaussian case) We can define T by $T(\mathbf{x}) := \frac{1}{\sqrt{k}} A\mathbf{x}$, where A is a random $k \times n$ matrix with each entry chosen independently from the standard normal distribution $N(0,1)$.

(c) (The ± 1 case) We can choose T as in (b) except that the entries of A independently attain values $+1$ and -1, each with probability $\frac{1}{2}$.

The plan is to first prove (b), where one can take some shortcuts in the proof, and then a general result involving both (b) and (c). We omit the proof of (a) here.

A random ± 1 matrix is much easier to generate and more suitable for computations than the matrix in the Gaussian case, and so the extra effort invested in proving (c) has some payoff.

2.2 On the normal distribution and subgaussian tails

We will now spend some time by building probabilistic tools.

The standard normal (or Gaussian) distribution $N(0,1)$ is well known, yet I first want to remind a beautiful computation related to it. The density of $N(0,1)$ is proportional to $e^{-x^2/2}$, but what is the right normalizing constant? In other words, what is the value of the integral $I := \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x$? It is known that the indefinite integral $\int e^{-x^2/2}\,\mathrm{d}x$ is not expressible by elementary functions.

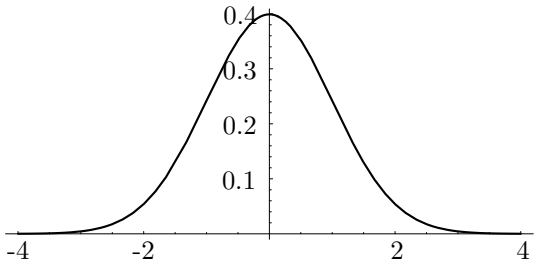
The trick is to compute I^2 as

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right) \\ &= \int_{\mathbb{R}^2} e^{-x^2/2} e^{-y^2/2}\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

[2] We won't define a random linear subspace formally; let it suffice to say that there is a unique rotation-invariant probability distribution on k-dimensional subspaces.

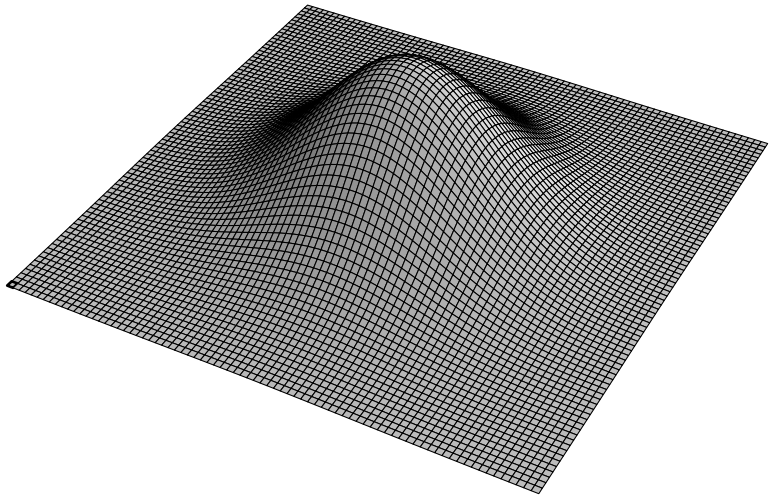
$$\begin{aligned} &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{\infty} e^{-r^2/2} 2\pi r\,\mathrm{d}r. \end{aligned}$$

To see the last equality, we consider the contribution of the infinitesimal annulus with inner radius r and outer radius $r+\mathrm{d}r$ to $\int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y$; the area of the annulus is $2\pi r\,\mathrm{d}r$ and the value of the integrand there is $e^{-r^2/2}$ (plus infinitesimal terms which can be neglected). The last integral, $\int_0^{\infty} e^{-r^2/2} 2\pi r\,\mathrm{d}r$, can already be evaluated in a standard way, by the substitution $t = r^2$, and we arrive at $I^2 = 2\pi$. Thus, the density of the normal distribution is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.



This computation also reminds us that if $Z_1, Z_2, \ldots, Z_n$ are independent standard normal variables, then the distribution of the vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ is spherically symmetric.[3]

[3] Which provides a good way of generating a random point on the high-dimensional Euclidean sphere S^{n-1}: Take $\mathbf{Z}/\|\mathbf{Z}\|_2$.



We also recall that if Z is a standard normal random variable, then $\mathbf{E}[Z] = 0$ (this is the 0 in $N(0,1)$) and $\mathrm{Var}[Z] = \mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}\left[Z^2\right] = 1$ (this is the 1). The random variable aZ, $a \in \mathbb{R}$, has the normal distribution $N(0, a^2)$ with variance a^2.

2-stability. We will need a fundamental property of the normal distribution called **2-stability**. It asserts that linear combinations of independent normal random variables are again normally distributed. More precisely, if X, Y are standard normal and independent, and $a, b \in \mathbb{R}$, then $aX + bY \sim N(0, a^2 + b^2)$, where $\sim$ means "has the same distribution as". More generally, of course, if $Z_1, \ldots, Z_n$ are independent standard normal and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, then $a_1 Z_1 + a_2 Z_2 + \cdots + a_n Z_n \sim \|\mathbf{a}\|_2 Z_1$, and this gives a hint why independent normal random variables might be useful for embeddings that almost preserve the Euclidean norm.

To prove the 2-stability, it suffices to prove that for X, Y independent standard normal and $a, b \in \mathbb{R}$ satisfying $a^2 + b^2 = 1$, the random variable $aX + bY$ is standard normal. We fix an angle α with $a = \cos\alpha$ and $b = \sin\alpha$. The random vector (X, Y) has the 2-dimensional normal distribution, which is spherically symmetric and thus invariant under rotations around the origin. Let us rotate the coordinate system by α; the new coordinates exressed in terms of the old ones are $(X\cos\alpha + Y\sin\alpha, -X\sin\alpha + Y\cos\alpha)$. The first coordinate is again standard normal and it equals $aX + bY$.

Subgaussian tails. There is an extensive literature concerning concentration of random variables around their expectation, and because of

phenomena related to the Central Limit Theorem, tail bounds similar to the tail of the standard normal distribution play a prominent role. We introduce the following convenient terminology.

Let X be a real random variable with $\mathbf{E}[X] = 0$. We say X has a **subgaussian upper tail** if there exists a constant $a > 0$ such that for all $\lambda > 0$,

$$\mathrm{Prob}[X > \lambda] \le e^{-a\lambda^2}.$$

We say that X has a **subgaussian upper tail up to λ_0** if the previous bound holds for all $\lambda \le \lambda_0$. We say that X has a **subgaussian tail** if both X and $-X$ have subgaussian upper tails.

If $X_1, X_2, \ldots, X_n$ is a sequence of random variables, by saying that they have a **uniform subgaussian tail** we mean that all of them have subgaussian tails with the same constant a.

A standard normal random variable has a subgaussian tail (ironically, a little proof is needed!), and the uniform ± 1 random variable clearly has a subgaussian tail.

The simplest version of the Chernoff (or rather, Bernstein) inequality provides another example of a random variable with a subgaussian tail. Namely, it tells us that if $X_1, \ldots, X_n$ are independent uniform ± 1 random variables, then $Y = Y_n := n^{-1/2}(X_1 + X_2 + \cdots + X_n)$ has a uniform subgaussian tail (the normalization by $n^{-1/2}$ is chosen so that $\mathrm{Var}[Y] = 1$, and uniformity means that the constant in the subgaussian tail is independent of n).

This inequality can be proved using the *moment generating function* of Y, which is the function that assigns to every nonnegative u the value $\mathbf{E}\left[e^{uY}\right]$.

2.2.1 Lemma (Moment generating function and subgaussian tail). *Let X be a random variable with $\mathbf{E}[X] = 0$. If $\mathbf{E}\left[e^{uX}\right] \le e^{Cu^2}$ for some constant C and for all $u > 0$, then X has a subgaussian upper tail, with $a = \frac{1}{4C}$. If $\mathbf{E}\left[e^{uX}\right] \le e^{Cu^2}$ holds for all $u \in (0, u_0]$, then X has a subgaussian upper tail up to $2Cu_0$.*

Proof. For all $u \in (0, u_0]$ and all $t \ge 0$ we have

$$\begin{aligned} \mathrm{Prob}[X \ge t] &= \mathrm{Prob}\left[e^{uX} \ge e^{ut}\right] \\ &\le e^{-ut}\mathbf{E}\left[e^{uX}\right] \qquad \text{(by the Markov inequality)} \\ &\le e^{-ut+Cu^2}. \end{aligned}$$

For $t \le 2Cu_0$ we can set $u = t/2C$, and we obtain $\mathrm{Prob}[X \ge t] \le e^{-t^2/4C}$. □

2.3 The Gaussian case of the random projection lemma

2.3.1 Lemma. **(Random projection lemma with independent Gaussian coefficients)** *Let n, k be natural numbers, let $\varepsilon \in (0,1)$, and let us define a random linear map $T\colon \mathbb{R}^n \to \mathbb{R}^k$ by*

$$T(\mathbf{x})_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{n} Z_{ij} x_j, \qquad i = 1, 2, \ldots, k,$$

where the Z_{ij} are independent standard normal random variables. Then for every vector $\mathbf{x} \in \mathbb{R}^n$ we have

$$\mathrm{Prob}\Big[(1-\varepsilon)\|\mathbf{x}\|_2 \le \|T(\mathbf{x})\|_2 \le (1+\varepsilon)\|\mathbf{x}\|_2\Big] \ge 1 - 2e^{-c\varepsilon^2 k},$$

where $c > 0$ is a constant.

Proof. Writing $Y_i = \sum_{j=1}^n Z_{ij} x_j$, we have $\|T(\mathbf{x})\|_2^2 = \frac{1}{k}\sum_{i=1}^k Y_i^2$. By the 2-stability of the normal distribution, $Y_i \sim N(0, \|\mathbf{x}\|_2^2)$ for all i. We may assume, for convenience, that $\|\mathbf{x}\|_2 = 1$, and then the Y_i are independent standard normal random variables.

We have $\mathbf{E}\big[Y_i^2\big] = \mathrm{Var}\,[Y_i] = 1$, and thus $\mathbf{E}\big[\|T(\mathbf{x})\|_2^2\big] = 1$. The expectation is exactly right, and it remains to prove that $\|T(\mathbf{x})\|_2^2$ is concentrated around 1.

We have $\mathrm{Var}[\|T(\mathbf{x})\|_2^2] = \frac{1}{k^2}\sum_{i=1}^k \mathrm{Var}[Y_i^2] = \frac{1}{k}\mathrm{Var}[Y^2]$, Y standard normal. Since $\mathrm{Var}[Y^2]$ is obviously some constant, $\mathrm{Var}[\|T(\mathbf{x})\|_2^2]$ is of order $\frac{1}{k}$. So it's natural to set $W := k^{-1/2}\sum_{i=1}^k (Y_i^2 - 1)$, so that $\mathbf{E}[W] = 0$ and $\mathrm{Var}\,[W]$ is a constant, and try to prove a subgaussian tail for W. It turns out that W doesn't have a subgaussian tail for arbitrarily large deviations, but only up to $\sqrt{k}$, but this will be sufficient for our purposes.

The core of the proof is the next claim.

2.3.2 Claim. *There exist constants C and $u_0 > 0$ such that*

$$\mathbf{E}[e^{u(Y^2-1)}] \le e^{Cu^2} \quad \textit{and} \quad \mathbf{E}[e^{u(1-Y^2)}] \le e^{Cu^2}$$

for all $u \in (0, u_0)$, where Y is standard normal.

Proof of the claim. We can directly calculate

$$\begin{aligned}
\mathbf{E}\Big[e^{u(Y^2-1)}\Big] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{u(x^2-1)} e^{-x^2/2}\,\mathrm{d}x \quad \text{(e.g., Maple...)} \\
&= \frac{1}{e^u\sqrt{1-2u}} = e^{-u-\frac{1}{2}\ln(1-2u)} \\
&= e^{u^2+O(u^3)} \quad \text{(Taylor expansion in the exponent)}
\end{aligned}$$

(the integral can actually be computed by hand, reducing it by substitution to the known integral $\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x$). It is then clear that the last expression is at most e^{2u^2} for all sufficiently small u (it can be shown that $u_0 = \frac{1}{4}$ works).

This proves the first inequality, and for the second we proceed in the same way: $\mathbf{E}\Big[e^{u(1-Y^2)}\Big] = e^u(1+2u)^{-1/2} = e^{u^2+O(u^3)}$ again. □

We can now finish the proof of the lemma. Using the claim for each Y_i, with $\tilde{u} := u/\sqrt{k}$ instead of u, and by the independence of the Y_i, we have

$$\mathbf{E}\big[e^{uW}\big] = \prod_{i=1}^{k} \mathbf{E}\Big[e^{\tilde{u}(Y_i^2-1)}\Big] \le \Big(e^{C\tilde{u}^2}\Big)^k = e^{Cu^2},$$

where $0 \le u \le u_0\sqrt{k}$, and similarly for $\mathbf{E}\big[e^{-uW}\big]$. Then Lemma 2.2.1 shows that W has a subgaussian tail up to $\sqrt{k}$ (assuming $2Cu_0 \ge 1$, which we may at the price of possibly increasing C and getting a worse constant in the subgaussian tail). That is,

$$\mathrm{Prob}\,[\,|W| \ge t] \le 2e^{-ct^2}, \quad 0 \le t \le \sqrt{k}. \tag{2.1}$$

Now $\|T(\mathbf{x})\|_2^2 - 1$ for unit $\mathbf{x}$ is distributed as $k^{-1/2}W$, and so using (2.1) with $t = \varepsilon\sqrt{k}$ gives $\mathrm{Prob}\,[1-\varepsilon \le \|T(\mathbf{x})\|_2 \le 1+\varepsilon] = \mathrm{Prob}\,\big[(1-\varepsilon)^2 \le \|T(\mathbf{x})\|_2^2 \le (1+$ $\mathrm{Prob}\,\big[1-\varepsilon \le \|T(\mathbf{x})\|_2^2 \le 1+\varepsilon\big] = \mathrm{Prob}\,\Big[|W| \le \varepsilon\sqrt{k}\Big] \ge 1 - 2e^{-c\varepsilon^2 k}$. The proof of the Gaussian version of the random projection lemma, and thus our first proof of the Johnson–Lindenstrauss lemma, are finished. □

Let us remark that tail estimates for the random variable $W = k^{-1/2}(Y_1^2 + \cdots + Y_k^2 - k)$, with the Y_i standard normal, are well known in statistics, since W has the important *chi-square distribution*. If we look up the density function of that distribution and make suitable estimates, we get another proof of the Gaussian case of the random projection lemma.

2.4 A more general random projection lemma

Replacing some of the concrete integrals in the previous lemma by general estimates, we can prove the following more general version of the

random projection lemma, where the independent $N(0,1)$ variables Z_{ij} are replaced by independent random variables R_{ij} with subgaussian tails.

2.4.1 Lemma (Random projection lemma). *Let n, k be natural numbers, let $\varepsilon \in (0, \frac{1}{2}]$, and let us define a random linear map $T\colon \mathbb{R}^n \to \mathbb{R}^k$ by*

$$T(\mathbf{x})_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{n} R_{ij} x_j, \qquad i = 1, 2, \ldots, k,$$

where the R_{ij} are independent random variables with $\mathbf{E}[R_{ij}] = 0$, $\mathrm{Var}[R_{ij}] = 1$, and a uniform subgaussian tail. Then for every vector $\mathbf{x} \in \mathbb{R}^n$ we have

$$\mathrm{Prob}\Big[(1-\varepsilon)\|\mathbf{x}\|_2 \le \|T(\mathbf{x})\|_2 \le (1+\varepsilon)\|\mathbf{x}\|_2\Big] \ge 1 - 2e^{-c\varepsilon^2 k},$$

where $c > 0$ is a constant (depending on the constant a in the uniform subgaussian tail of the R_{ij} but independent of n, k, ε).

We want to imitate the proof for the Gaussian case. The difference is that now we don't explicitly know the distribution of $Y_i := \sum_{j=1}^n R_{ij} x_j$. The plan is to first prove that Y_i has a subgaussian tail, and then use this to prove an analog of Claim 2.3.2 bounding the moment generating function of $Y_i^2 - 1$ and of $1 - Y_i^2$.

Our approach doesn't lead to the shortest available proof, but the advantage (?) is that most of the proof is rather mechanical: It is clear what should be calculated, and it is calculated in a pedestrian manner.

In order to start bounding the moment generating functions, we need the following partial converse of Lemma 2.2.1:

2.4.2 Lemma. *If X is a random variable with $\mathbf{E}[X] = 0$ and $\mathrm{Var}[X] = \mathbf{E}[X^2] = 1$, and X has a subgaussian upper tail, then $\mathbf{E}\left[e^{uX}\right] \le e^{Cu^2}$ for all $u > 0$, where the constant C depends only on the constant a in the subgaussian tail.*

We should stress that a bound of, say, $10e^{Cu^2}$ for $\mathbf{E}\left[e^{uX}\right]$ would *not* be enough for our applications of the lemma. We need to use the lemma with u arbitrarily small, and there we want $\mathbf{E}\left[e^{uX}\right]$ to be bounded by $1 + O(u^2)$ (which is equivalent to $e^{O(u^2)}$ for u small). In contrast, for subgaussian tails, a tail bound like $10e^{-at^2}$ would be as good as e^{-at^2}.

Proof of Lemma 2.4.2. Let F be the distribution function of X; that is, $F(t) = \mathrm{Prob}[X < t]$. We have $\mathbf{E}\left[e^{uX}\right] = \int_{-\infty}^{\infty} e^{ut}\,\mathrm{d}F(t)$. We split the

integration interval into two subintervals, corresponding to $ut \le 1$ and $ut \ge 1$.

In the first subinterval, we use the estimate

$$e^x \le 1 + x + x^2,$$

which is valid for all $x \le 1$ (and, in particular, for all negative x). So

$$\int_{-\infty}^{1/u} e^{ut}\,\mathrm{d}F(t) \le \int_{-\infty}^{1/u} 1 + ut + u^2t^2\,\mathrm{d}F(t) \le \int_{-\infty}^{\infty} 1 + ut + u^2t^2\,\mathrm{d}F(t)$$

$$= 1 + u\mathbf{E}[X] + u^2\mathbf{E}[X^2] = 1 + u^2.$$

The second subinterval, $ut \ge 1$, is where we use the subgaussian tail. (We proceed by estimating the integral by a sum, but if the reader feels secure in integrals, she may do integration by parts instead.)

$$\begin{aligned}
\int_{1/u}^{\infty} e^{ut}\,\mathrm{d}F(t) &\le \sum_{k=1}^{\infty} \int_{k/u}^{(k+1)/u} e^{k+1}\,\mathrm{d}F(t) \le \sum_{k=1}^{\infty} e^{k+1} \int_{k/u}^{\infty} \mathrm{d}F(t) \\
&= \sum_{k=1}^{\infty} e^{k+1} \mathrm{Prob}\left[X \ge \frac{k}{u}\right] \\
&\le \sum_{k=1}^{\infty} e^{k+1} e^{-ak^2/u^2} \quad \text{(by the subgaussian tail)} \\
&\le \sum_{k=1}^{\infty} e^{2k - ak^2/u^2}
\end{aligned}$$

($2k$ is easier to work with than $k+1$). As a function of a real variable k, the exponent $2k - ak^2/u^2$ is maximized for $k = k_0 := u^2/a$, and there are two cases to distinguish, depending on whether this maximum is within the summation range.

For $u^2 > a$, we have $k_0 \ge 1$, and the terms near k_0 dominate the sum, while going away from k_0 the terms decrease (at least) geometrically. Thus, the whole sum is $O(e^{2k_0 - ak_0^2/u^2}) = O(e^{u^2/a}) = e^{O(u^2)}$ (we recall that $u^2/a \ge 1$), and altogether $\mathbf{E}\left[e^{uX}\right] = 1 + u^2 + e^{O(u^2)} = e^{O(u^2)}$.

For $u^2 \le a$ the $k = 1$ term is the largest and the subsequent terms decrease (at least) geometrically, so the sum is of order e^{-a/u^2}, and, grossly overestimating, we have $e^{-a/u^2} = 1/e^{a/u^2} \le 1/(a/u^2) = u^2/a$. So $\mathbf{E}\left[e^{uX}\right] \le 1 + O(u^2) \le e^{O(u^2)}$ as well. □

Now, by passing from subgaussian tails to bounds for the moment generating functions and back, we can easily see that the $Y_i = \sum_{j=1}^n R_{ij}x_j$ have uniform subgaussian tails:

2.4.3 Lemma. *Let $R_1, \dots, R_n$ be independent random variables with $\mathbf{E}[R_j] = 0$, $\mathrm{Var}[R_j] = 1$, and a uniform subgaussian tail, and let $\mathbf{x} \in \mathbb{R}^n$ satisfy $\|\mathbf{x}\|_2 = 1$. Then*

$$Y := R_1x_1 + \cdots + R_nx_n$$

has $\mathbf{E}[Y] = 0$, $\mathrm{Var}[Y] = 1$, and a subgaussian tail.

This lemma can be viewed as a generalization of the usual Chernoff–Hoeffding bounds.

Proof. $\mathbf{E}[Y] = 0$ and $\mathrm{Var}[Y] = 1$ are immediate. As for the subgaussian tail, we have $\mathbf{E}\left[e^{uR_j}\right] \le e^{Cu^2}$ by Lemma 2.4.2, and so

$$\mathbf{E}\left[e^{uY}\right] = \prod_{j=1}^n \mathbf{E}\left[e^{uR_jx_j}\right] \le e^{Cu^2(x_1^2+\cdots+x_n^2)} = e^{Cu^2}.$$

Thus, Y has a subgaussian tail by Lemma 2.2.1 (and by symmetry). □

Here is the result that replaces Claim 2.3.2 in the present more general setting.

2.4.4 Claim. *Let Y have $\mathbf{E}[Y] = 0$, $\mathrm{Var}[Y] = 1$, and a subgaussian tail. Then there exist constants C and $u_0 > 0$ such that*

$$\mathbf{E}[e^{u(Y^2-1)}] \le e^{Cu^2} \quad \text{and} \quad \mathbf{E}[e^{u(1-Y^2)}] \le e^{Cu^2}$$

for all $u \in (0, u_0)$.

Proof. We begin with the first inequality. First we note that $\mathbf{E}[Y^4]$ is finite (a constant); this follows from the subgaussian tail of Y by direct calculation, or in a simpler way, from Lemma 2.4.2 and from $t^4 = O(e^t + e^{-t})$ for all t.

Let F be the distribution function of Y^2; that is, $F(t) = \mathrm{Prob}\left[Y^2 < t\right]$. We again split the integral defining $\mathbf{E}[e^{uY^2}]$ into two intervals, corresponding to $uY^2 \le 1$ and $uY^2 \ge 1$. That is,

$$\mathbf{E}\left[e^{uY^2}\right] = \int_0^{1/u} e^{ut}\,\mathrm{d}F(t) + \int_{1/u}^{\infty} e^{ut}\,\mathrm{d}F(t).$$

The first integral is estimated, again using $e^x \le 1 + x + x^2$ for $x \le 1$, by

$$\int_0^{1/u} 1 + ut + u^2t^2\,\mathrm{d}F(t) \le \int_0^{\infty} 1 + ut + u^2t^2\,\mathrm{d}F(t)$$

$$= 1 + u\mathbf{E}[Y^2] + u^2\mathbf{E}[Y^4] = 1 + u + O(u^2).$$

The second integral can be estimated by a sum:

$$\sum_{k=1}^{\infty} e^{k+1}\mathrm{Prob}\left[Y^2 \ge k/u\right] \le 2\sum_{k=1}^{\infty} e^{2k}e^{-ak/u}.$$

We may assume that $u \le u_0 := a/4$; then $k(2 - a/u) \le -ka/2u$, and the sum is of order $e^{-\Omega(1/u)}$. Similar to the proof of Lemma 2.4.2 we can bound this by $O(u^2)$, and for $\mathbf{E}[e^{uY^2}]$ we thus get the estimate $1 + u + O(u^2) \le e^{u+O(u^2)}$.

Then we calculate $\mathbf{E}[e^{u(Y^2-1)}] = \mathbf{E}[e^{uY^2}]e^{-u} \le e^{O(u^2)}$ as required.

The calculation for estimating $\mathbf{E}[e^{-uY^2}]$ is simpler, since our favorite inequality $e^x \le 1 + x + x^2$, $x \le 1$, now gives $e^{-ut} \le 1 - ut + u^2t^2$ for all $t > 0$ and $u > 0$. Then

$$\mathbf{E}\left[e^{-uY^2}\right] = \int_0^{\infty} e^{-ut}\mathrm{d}F(t) \le \int_0^{\infty} 1 - ut + u^2t^2\mathrm{d}F(t)$$

$$= 1 - u\mathbf{E}[Y^2] + u^2\mathbf{E}[Y^4] \le 1 - u + O(u^2) \le e^{-u+O(u^2)}.$$

This yields $\mathbf{E}\left[e^{u(1-Y^2)}\right] \le e^{O(u^2)}$. □

Proof of Lemma 2.4.1. Lemmas 2.4.2 and 2.4.3 plus Claim 2.4.4 cover all that is needed to upgrade the proof of the Gaussian case (Lemma 2.3.1). □

2.5 Embedding ℓ_2^n in $\ell_1^{O(n)}$

We prove a theorem promised earlier.

2.5.1 Theorem. *Given n and $\varepsilon \in (0, 1)$, let $k \ge C\varepsilon^{-2}(\log \frac{1}{\varepsilon})n$ for a suitable constant C. Then there is a (linear) ε-almost isometry $T\colon \ell_2^n \to \ell_1^k$.*

The first and main tool is yet another version of the random projection lemma: this time the random projection goes from ℓ_2^n to ℓ_1^k.

2.5.2 Lemma (Random projection from ℓ_2 to ℓ_1). *Let n, k be natural numbers, let $\varepsilon \in (0,1)$, and let us define a random linear map $T\colon \mathbb{R}^n \to \mathbb{R}^k$ by*

$$T(\mathbf{x})_i = \frac{1}{\beta k} \sum_{j=1}^{n} Z_{ij} x_j, \quad i = 1, 2, \ldots, k,$$

where the Z_{ij} are independent standard normal random variables, and $\beta > 0$ is a certain constant ($\sqrt{2/\pi}$ if you must know). Then for every vector $\mathbf{x} \in \mathbb{R}^n$ we have

$$\mathrm{Prob}\Big[(1-\varepsilon)\|\mathbf{x}\|_2 \le \|T(\mathbf{x})\|_1 \le (1+\varepsilon)\|\mathbf{x}\|_2\Big] \ge 1 - 2e^{-c\varepsilon^2 k},$$

where $c > 0$ is a constant.

This looks almost exactly like the Gaussian version of the random projection lemma we had earlier, only the normalizing factor of T is different and the ℓ_1 norm is used in the target space. The proof is also very similar to the previous ones.

Proof. This time $\|T(\mathbf{x})\|_1 = \frac{1}{\beta k}\sum_{i=1}^{k} |Y_i|$, where $Y_i = \sum_{j=1}^{n} Z_{ij} x_j$ is standard normal (assuming $\mathbf{x}$ unit). For a standard normal Y, it can easily be calculated that $\mathbf{E}[|Y|] = \sqrt{2/\pi}$, and this is the mysterious β (but we don't really need its value, at least in some of the versions of the proof offered below). Then $\mathbf{E}[\|T(\mathbf{x})\|_1] = 1$ and it remains to prove concentration, namely, that $W := \frac{1}{\beta\sqrt{k}}\sum_{i=1}^{k}(|Y_i| - \beta)$ has a subgaussian tail up to $\sqrt{k}$. This follows in the usual way from the next claim.

2.5.3 Claim. *For Y standard normal we have*

$$\mathbf{E}[e^{u(|Y|-\beta)}] \le e^{Cu^2} \quad \text{and} \quad \mathbf{E}[e^{u(1-|Y|)}] \le e^{Cu^2}$$

with a suitable C and all $u \ge 0$ (note that we don't even need a restriction $u \le u_0$).

First proof. We can go through the explicit calculations, as we did for Claim 2.3.2:

$$\begin{aligned}
\mathbf{E}\Big[e^{u|Y|}\Big] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{u|x|-x^2/2}\,\mathrm{d}x = \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} e^{ux-x^2/2}\,\mathrm{d}x \\
&= \frac{2}{\sqrt{2\pi}} e^{u^2/2}\int_{0}^{\infty} e^{-(x-u)^2/2}\,\mathrm{d}x = 2e^{u^2/2}\cdot\frac{1}{\sqrt{2\pi}}\int_{-u}^{\infty} e^{-t^2/2}\,\mathrm{d}t \\
&= 2e^{u^2/2}\left(\frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{u} e^{-t^2/2}\,\mathrm{d}t\right) \\
&\le 2e^{u^2/2}\left(\frac{1}{2} + \frac{u}{\sqrt{2\pi}}\right) = e^{u^2/2}(1+\beta u) \le e^{\beta u + u^2/2}.
\end{aligned}$$

Thus $\mathbf{E}\big[e^{u(|Y|-\beta)}\big] \le e^{u^2/2}$. The second inequality follows analogously. □

Second proof. We can apply the technology developed in Section 2.4. The random variable $X := |Y| - \beta$ is easily seen to have a subgaussian tail, we have $\mathbf{E}[X] = 0$, and $\mathrm{Var}[X]$ is some constant. So we can use Lemma 2.4.2 for $X' := X/\sqrt{\mathrm{Var}[X]}$ and the claim follows. □

Variations and extensions. One can also prove a version of the random projection lemma where the mapping T goes from ℓ_2^n in ℓ_p^k with $1 \le p \le 2$. The same method can be used, only the calculations in the proof of appropriate claim are different. This leads to an analog of Theorem 2.5.1, i.e., a $(1+\varepsilon)$-embedding of ℓ_2^n into ℓ_p^k, $k = O(\varepsilon^{-2}(\log\frac{1}{\varepsilon})n)$. On the other hand, for $p > 2$, the method can still be used to $(1+\varepsilon)$-embed ℓ_2^n into ℓ_p^k, but the calculation comes out differently and the dimension k will no longer be linear, but a larger power of n depending on p.

An interesting feature of Lemma 2.5.2 is what *doesn't* work—namely, replacing the $N(0,1)$ variables by uniform ± 1 variables, say, a generalization analogous to Lemma 2.4.1. The concentration goes through just fine, but the *expectation* doesn't. Namely, if $Y_i := \sum_{j=1}^{n} R_{ij} x_j$ for a unit $\mathbf{x}$ and the R_{ij} are no longer Gaussian, then $\mathbf{E}[|Y_i|]$, unlike $\mathbf{E}\big[Y_i^2\big]$, may depend on $\mathbf{x}$! For example, let the R_{ij} be uniform random ± 1 and let us consider $\mathbf{x} = (1,0)$ and $\mathbf{y} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Then $\mathbf{E}[|\pm x_1 \pm x_2|] = \mathbf{E}[|\pm 1|] = 1$, while $\mathbf{E}[|\pm y_1 \pm y_2|] = \frac{1}{\sqrt{2}}$.

However, it turns out that, in the case of ± 1 random variables, the expectation can vary at most between two absolute constants, independent of the dimension n, as we will later prove (Lemma 2.7.1).

This is a special case of *Khintchine's inequality*, claiming that for every $p \in (0,\infty)$ there are constants $C_p \ge c_p > 0$ (the best values are known) with

$$c_p\|\mathbf{x}\|_2 \le \mathbf{E}\left[\left|\sum_{j=1}^{n} \epsilon_j x_j\right|^p\right]^{1/p} \le C_p\|\mathbf{x}\|_2,$$

where the ϵ_j are independent uniform random ± 1 variables. Using this fact, a random linear mapping T with ± 1 coefficients can be used to embed ℓ_2^n in ℓ_1 (or ℓ_p) with distortion bounded by a constant, but not arbitrarily close to 1.

Dense sets in the sphere. Now we know that if $T\colon \mathbb{R}^n \to \mathbb{R}^k$ is a random linear map as in Lemma 2.5.2, then it almost preserves the norm of any fixed $\mathbf{x}$ with probability exponentially close to 1. The proof of Theorem 2.5.1 goes as follows:

1. We choose a large finite set $N \subset S^{n-1}$, where $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$ is the Euclidean unit sphere, and we obtain T that is an ε-almost isometry on all of N simultaneously.

2. Then we check that any linear T with this property is a 4ε-almost isometry on the whole of ℓ_2^n.

Let us call a set $N \subseteq S^{n-1}$ **δ-dense** if every $\mathbf{x} \in S^{n-1}$ has some point $\mathbf{y} \in N$ at distance no larger than δ (the definition applies to an arbitrary metric space). For step 2 we will need that N is ε-dense. Then, in order that step 1 works, N must not be too large. We have the following (standard and generally useful) lemma:

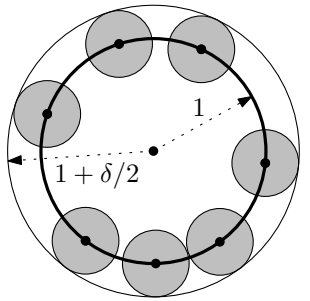
2.5.4 Lemma (Small δ-dense sets in the sphere). *For each $\delta \in (0,1]$, there exists a δ-dense set $N \subseteq S^{n-1}$ with*

$$|N| \le \left(\frac{4}{\delta}\right)^n.$$

The proof below is existential. It is hard to find explicit constructions of reasonably small dense sets in the sphere.

Proof. In order to construct a small δ-dense set, we start with the empty set and keep adding points one by one. The trick is that we do not worry about δ-density along the way, but we always keep the current set *δ-separated*, which means that every two points have distance at least δ. Clearly, if no more points can be added, the resulting set N must be δ-dense.

For each $\mathbf{x} \in N$, consider the ball of radius $\frac{\delta}{2}$ centered at $\mathbf{x}$. Since N is δ-separated, these balls have disjoint interiors, and they are contained in the ball $B(0, 1+\delta/2) \subseteq B(0,2)$.



Therefore, $\mathrm{vol}(B(0,2)) \ge |N|\mathrm{vol}(B(0,\frac{\delta}{2}))$. If a convex body in $\mathbb{R}^n$ is scaled by a factor of r, its volume is multiplied by r^n. So $\mathrm{vol}(B(0,2)) = (4/\delta)^n \mathrm{vol}(B(0,\frac{\delta}{2}))$, and the lemma follows. □

For later use let us record that exactly the same proof works for δ-dense sets in the unit sphere, or even unit ball, of an arbitrary n-dimensional normed space (where the density is measured using the metric of that space).

For large n the bound in the lemma is essentially the best possible (up to the value of the constant 4). For n small it may be important to know that the "right" exponent is $n-1$ and not n, but the argument providing $n-1$ would be technically more complicated.

For step 2 in the above outline of the proof of Theorem 2.5.1, we need the next lemma, which is slightly less trivial than it may seem.

2.5.5 Lemma. *Let $N \subset S^{n-1}$ be a δ-dense set for some $\delta \in (0, \frac{1}{2}]$ and let $T\colon \mathbb{R}^n \to \mathbb{R}^k$ be a linear map satisfying the ε-almost isometry condition $1-\varepsilon \le \|T(\mathbf{y})\|_1 \le 1+\varepsilon$ for all $\mathbf{y} \in N$. Then T is an $2(\varepsilon+\delta)$-almost isometry $\ell_2^n \to \ell_1^k$.*

Proof. Since T is linear, it suffices to prove the almost-isometry property for all $\mathbf{x} \in S^{n-1}$. So let us try to bound $\|T(\mathbf{x})\|_1$ from above. As expected, we find $\mathbf{y} \in N$ with $\|\mathbf{x}-\mathbf{y}\|_2 \le \delta$, and the triangle inequality gives

$$\|T(\mathbf{x})\|_1 \le \|T(\mathbf{y})\|_1 + \|T(\mathbf{x}-\mathbf{y})\|_1 \le 1+\varepsilon+\|T(\mathbf{x}-\mathbf{y})\|_1.$$

Letting $\mathbf{u} := (\mathbf{x}-\mathbf{y})/\|\mathbf{x}-\mathbf{y}\|_2$ and using the linearity of T and the scaling property of $\|.\|_1$, we further obtain

$$\|T(\mathbf{x})\|_1 \le 1+\varepsilon+\|\mathbf{x}-\mathbf{y}\|_2 \cdot \|T(\mathbf{u})\|_1 \le 1+\varepsilon+\delta\|T(\mathbf{u})\|_1. \qquad (2.2)$$

But now we need to bound $\|T(\mathbf{u})\|_1$, and this is the same problem as bounding $\|T(\mathbf{x})\|_1$, only with a different vector.

To get around this, we first observe that the Euclidean unit sphere S^{n-1} is a compact set, and the mapping $\mathbf{x} \mapsto \|T(\mathbf{x})\|_1$ is continuous on S^{n-1}, and hence it attains a maximum. Let this maximum be M, and let it be attained at a point $\mathbf{x}_0$. Now we apply (2.2) with $\mathbf{x} = \mathbf{x}_0$, which gives

$$M = \|T(\mathbf{x}_0)\|_1 \le 1 + \varepsilon + \delta M,$$

since $\|T(\mathbf{u})\|_1 \le M$ for all $\mathbf{u} \in S^{n-1}$, by the choice of M. From this inequality we then calculate

$$M \le \frac{1+\varepsilon}{1-\delta}.$$

A lower bound for $\|T(\mathbf{x})\|_1$ is now simple using the upper bound we already have for all $\mathbf{x}$: $\|T(\mathbf{x})\|_1 \ge \|T(\mathbf{y})\|_1 - \|T(\mathbf{x}-\mathbf{y})\|_1 \ge 1 - \varepsilon - \delta\frac{1+\varepsilon}{1-\delta}$. Estimates of some ugly fractions brings both the upper and lower bounds to the desired form $1 \pm 2(\varepsilon+\delta)$. □

Proof of Theorem 2.5.1. Let N be ε-dense in S^{n-1} of size at most $(4/\varepsilon)^n$. For $k = C\varepsilon^{-2}(\ln\frac{1}{\varepsilon})n$ the probability that a random T is not an ε-almost isometry on N is at most $|N|\cdot 2e^{-c\varepsilon^2 k} \le 2e^{-cCn\ln(1/\varepsilon)+n\ln(4/\varepsilon)} < 1$ for C sufficiently large.

If T is an ε-almost isometry on N, then it is a 4ε-almost isometry on all of ℓ_2^n. □

The proof actually shows that a random T fails to be an ε-almost isometry only with exponentially small probability (at most $e^{-\Omega(\varepsilon^2 k)}$).

Viewing the embedding as a numerical integration formula. In Section 1.5 we defined the 1-embedding $F\colon \ell_2^n \to L_1(S^{n-1})$ by $F(\mathbf{x}) := f_{\mathbf{x}}$, where $f_{\mathbf{x}}(\mathbf{u}) = \langle \mathbf{x}, \mathbf{u}\rangle$. Similarly we can define an embedding G of ℓ_2^n in the space of measurable functions on $\mathbb{R}^n$ with the L_1 norm corresponding to the Gaussian measure; i.e., $\|f\|_1 := \int_{\mathbb{R}^n} |f(\mathbf{z})|\gamma(\mathbf{z})\,\mathrm{d}\mathbf{z}$, where $\gamma(\mathbf{z}) := (2\pi)^{-n/2}e^{-\|\mathbf{z}\|_2^2/2}$ is the density of the standard normal distribution. We set $G(\mathbf{x}) := f_{\mathbf{x}}$, where $f_{\mathbf{x}}$ is now regarded as a function on $\mathbb{R}^n$ (while for F, we used it as a function on S^{n-1}). By the spherical symmetry of γ we see that for all $\mathbf{x}$, $\|f_{\mathbf{x}}\|_1 = c\|\mathbf{x}\|_2$ for some normalizing constant $c > 0$, similar to the case of F, and so G is a 1-embedding as well.

The embedding $\ell_2^n \to \ell_1^{O(n)}$ discussed in the present section can now be viewed as a "discretization" of G. Namely, if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k \in \mathbb{R}^n$ are the rows of the matrix defining the embedding T in Lemma 2.5.2, or

in other words, independent random points of $\mathbb{R}^n$ drawn according to the density function γ, the results of this section show that, with high probability, the following holds for every $\mathbf{x} \in \mathbb{R}^n$:

$$\frac{1}{\beta k}\sum_{i=1}^{k} |f_{\mathbf{x}}(\mathbf{a}_i)| \approx_\varepsilon \|\mathbf{x}\|_2 = \frac{1}{c}\int_{\mathbb{R}^n} |f_{\mathbf{x}}(\mathbf{z})|\gamma(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

($\approx_\varepsilon$ means approximation up to a factor of at most $1 \pm \varepsilon$).

With this formulation, the proof of Theorem 2.5.1 thus shows that the average over a random $O(n)$-point set approximates the integral over $\mathbb{R}^n$ for each of the functions $|f_{\mathbf{x}}|$ up to $1 \pm \varepsilon$.

By projecting $\mathbb{R}^n$ radially onto S^{n-1}, we get an analogous statement for approximating $\int_{S^{n-1}} |f_{\mathbf{x}}(\mathbf{u})|\,\mathrm{d}\mathbf{u}$ by an average over a random set A in S^{n-1}. We have thus obtained a strong quantitative version of the handwaving argument from Section 1.5.

2.6 Streaming and pseudorandom generators

Stream computation is a quickly developing area of computer science motivated mainly by the gigantic amounts of data passing through the current networks. A *data stream* is a sequence of elements (numbers, letters, points in the plane, etc.), which is much larger than the available memory. The goal is to compute, at least approximately, some function of the data using only one sequential pass over the data stream.

For example, let us think of a network router, which receives packets of data and sends them further towards their destinations. Say that packets are classified into $n = 2^{64}$ types according to some of their header bits. At the end of the day we would like to know, for instance, whether some concrete type of packets has appeared in suspiciously large numbers.

This looks difficult, or perhaps impossible, since there are way too many packets and packet types to store information about each of them. (The human brain seems to be able to solve such tasks somehow, at least some people's brains—a cashier in a supermarket can't remember all customers in a day, but still she may notice if she serves someone several times.)

Let x_i denote the number of packets of the ith type that passed through the router since the morning, $i = 1, 2, \ldots, n$. The computation starts with $\mathbf{x} = (x_1, \ldots, x_n) = \mathbf{0}$, and the stream can be regarded as a sequence of instructions like

increment x_{645212} by 1
increment x_{302256} by 1
increment x_{12457} by 1
$\vdots$

For the method shown here, we will be able to accept even more general instructions, specified by an index $i \in \{1, 2, \ldots, n\}$ and an integer $\Delta \in \{\pm 1, \pm 2, \ldots, \pm n\}$ and meaning "add Δ to x_i".[4] We assume that the total number of such instructions, i.e., the length of the stream, is bounded by n^2 (or another fixed polynomial in n).

The specific problem we will consider here is to estimate the ℓ_2 norm $\|\mathbf{x}\|_2$, since the solution uses the tools we built in preceding sections. This may remind one of the man looking for his keys not in the dark alley where he's lost them but under a street lamp where there's enough light. But the square $\|\mathbf{x}\|_2^2$ is an important parameter of the stream: One can compute the standard deviation of the x_i from it, and use it for assessing how homogeneous or "random-like" the stream is (the appropriate keywords in statistics are *Gini's index of homogeneity* and *surprise index*). Moreover, as we will mention at the end of this section, an extension of the method can also solve the "heavy hitters" problem: after having gone through the stream and storing some limited amount of information, we are given an index i, and we want to test whether the component x_i is exceptionally large compared to most others.

Thus, we consider the following problem, **the ℓ_2 norm estimation**: We're given an $\varepsilon > 0$, which we think of as a fixed small number, and we go through the stream once, using memory space much smaller than n. At the end of the stream we should report a number, the norm estimate, that lies between $(1-\varepsilon)\|\mathbf{x}\|_2$ and $(1+\varepsilon)\|\mathbf{x}\|_2$.

It can be shown that this problem is *impossible* to solve by a deterministic algorithm using $o(n)$ space.[5] We describe a *randomized* solution, where the algorithm makes some internal random decisions. For every possible input stream, the output of the algorithm will be correct with probability at least $1-\delta$, where the probability is with respect to the internal random choices of the algorithm. (So we *don't* assume any kind of randomness in the input stream.) Here $\delta > 0$ is a parameter of the algorithm, which will enter bounds for the memory requirements.

[4] Choosing n both as the number of entries of $\mathbf{x}$ and as the allowed range of Δ has no deep meaning—it is just in order to reduce the number of parameters.

[5] Sketch of proof: Let us say that the algorithm uses at most $n/100$ bits of space. For every $\mathbf{x} \in \{-1, 0, 1\}^n$ let us fix a stream $S_{\mathbf{x}}$ of length n that produces $\mathbf{x}$ as the current vector at the end. For each $\mathbf{x} \in \{0,1\}^n$, run the algorithm on $S_{\mathbf{x}} \circ S_{\mathbf{0}}$, where $\circ$ means putting one stream after another, and record the contents of its memory after the first n steps, i.e., at the end of $S_{\mathbf{x}}$; let this contents be $M(\mathbf{x})$. Since there are at most $2^{n/100}$ possible values of $M(\mathbf{x})$, some calculation shows that there exist $\mathbf{x}, \mathbf{x}' \in \{0,1\}^n$ differing in at least $n/100$ components with $M(\mathbf{x}) = M(\mathbf{x}')$. Finally, run the algorithm on $S_{\mathbf{x}} \circ S_{-\mathbf{x}}$ and also on $S_{\mathbf{x}'} \circ S_{-\mathbf{x}}$. Being deterministic, the algorithm gives the same answer, but in the first case the norm is 0 and in the second case it's at least $\sqrt{n}/10$.

A random projection algorithm? Let us start by observing that some functions of $\mathbf{x}$ are easy to compute by a single pass through the stream, such as $\sum_{i=1}^n x_i$—we can just maintain the current sum. More generally, any fixed linear function $\mathbf{x} \mapsto \langle \mathbf{a}, \mathbf{x} \rangle$ can be maintained exactly, using only a single word, or $O(\log n)$ bits, of memory.

As we have seen, if A is a suitably normalized random $k \times n$ matrix, then $\|A\mathbf{x}\|_2$ is very likely to be a good approximation to $\|\mathbf{x}\|_2$ even if k is very small compared to n. Namely, we know that the probability that $\|A\mathbf{x}\|_2$ fails to be within $(1 \pm \varepsilon)\|\mathbf{x}\|_2$ is at most $2e^{-c\varepsilon^2 k}$, and so with $k := C\varepsilon^{-2} \log \frac{1}{\delta}$ we obtain the correct estimate with probability at least $1-\delta$. Moreover, maintaining $A\mathbf{x}$ means maintaining k linear functions of $\mathbf{x}$, and we can do that using k words of memory, which is even a number independent of n.

This looks like a very elegant solution to the norm estimation problem but there is a serious gap. Namely, to obey an instruction "increment x_i by Δ" in the stream, we need to add $\Delta \mathbf{a}_i$ to the current $A\mathbf{x}$, where $\mathbf{a}_i$ is the ith column of A. The same i may come many times in the stream, and we always need to use the same vector $\mathbf{a}_i$, otherwise the method breaks down. But A has kn entries and we surely can't afford to store it.

Pseudorandom numbers. To explain an ingenious way of overcoming this obstacle, we start by recalling how random numbers are generated by computers in practice.

The "random" numbers used in actual computations are not random but *pseudorandom*. One starts with an integer r_0 in range from 0 to $m-1$, where m is a large number, say 2^{64}. This r_0 is called the *seed* and we usually may think of it as truly random (for instance, it may be derived from the number of microseconds in the current time when the computer is switched on). Then a sequence $(r_0, r_1, r_2, \ldots)$ of pseudorandom numbers is computed as

$$r_{t+1} := f(r_t),$$

where f is some *deterministic* function. Often f is of the form $f(x) := (ax+b) \bmod m$, where a, b, m are large integers, carefully chosen but fixed.

One then uses the r_t as if they were *independent random* integers from $\{0, 1, \ldots, m-1\}$. Thus, each r_t brings us, say, 64 new random bits. They

are not really independent at all, but empirically, and also with some theoretical foundation, for most computations they work as if they were.

Let us now consider our matrix A, and suppose, as we may, that is a random ± 1 matrix. If we want to generate it, say column by column, we can set the first 64 entries in the first column according to r_0, the next 64 entries according to r_1, and so on. Given i and j, we can easily find which bit of which r_t is used to generate the entry a_{ij}.

Thus, if we store the seed r_0, we can re-compute the ith column of A whenever we need it, simply by starting the pseudorandom generator all over from r_0 and computing the appropriate r_t's for the desired column. This, as described, may be very slow, since we need to make about nk steps of the pseudorandom generator for a typical column. But the main purpose has been achieved—we need practically no extra memory.[6]

Although this method may very well work fine in practice, we can't provide a theoretical guarantee for all possible vectors $\mathbf{x}$.

A pseudorandom generator with guarantees. Researchers in computational complexity have developed "theoretical" versions of pseudorandom generators that provably work: For certain well-defined classes of randomized computations, and for all possible inputs, they can be used instead of truly random bits without changing the distribution of the output in a noticeable manner.

Pseudorandom generators constitute an important area of computational complexity, with many ingenious results and surprising connections to other subjects.

Here we describe only a single specific pseudorandom generator G, for *space-bounded computations*. Similar to the practically used pseudorandom generators mentioned above, G accepts a *seed* σ, which is a short sequence of truly random independent bits, and computes a much longer sequence $G(\sigma)$ of pseudorandom bits.

The particular G we will discuss, *Nisan's generator*, needs a seed of $2\ell^2+\ell$ truly random bits and outputs $\ell 2^\ell$ pseudorandom bits, exponentially many in the square root of the seed length. Formally we regard G as a mapping $\{0,1\}^{2\ell^2+\ell} \to \{0,1\}^{\ell 2^\ell}$.

To define G, we interpret the seed σ as a $(2\ell+1)$-tuple

$$(\sigma_0, a_1, b_1, \ldots, a_\ell, b_\ell),$$

where σ_0 and the a_i and b_i have ℓ bits each and they are interpreted as elements of the 2^ℓ-element finite field $\mathrm{GF}(2^\ell)$.

[6] With a generator of the form $r_{t+1} = (ar_t + b) \bmod m$ the computation can actually be done much faster.

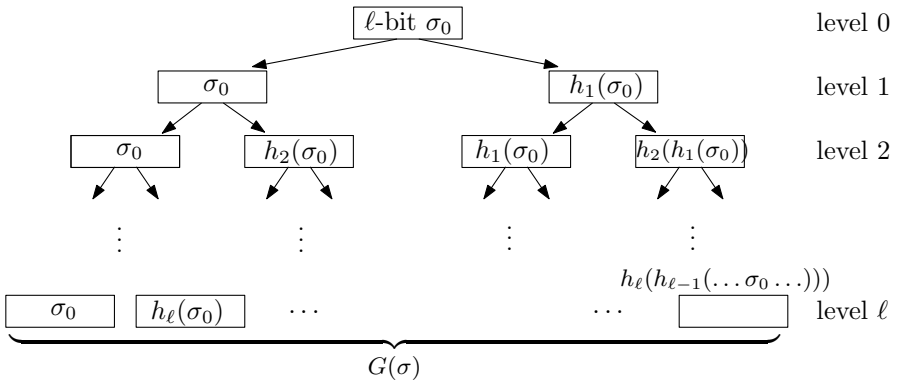
Each pair a_i, b_i defines a *hash function* $h_i\colon \mathrm{GF}(2^\ell) \to \mathrm{GF}(2^\ell)$ by $h_i(x) := a_i x + b_i$. Intuitively, the purpose of a hash function is to "mix" a given bit string thoroughly, in a random-looking fashion. Technically, the properties of the h_i needed for the construction of G are the following:

- Succinctness: h_i "mixes" 2^ℓ numbers but it is specified by only 2ℓ bits.
- Efficiency: h_i can be evaluated quickly and in small working space, $O(\ell)$ bits.[7]
- *Pairwise independence:* If $a, b \in \mathrm{GF}(2^\ell)$ are chosen uniformly at random, then the corresponding hash function h satisfies, for any two pairs $x \neq y$ and $u \neq v$ of elements of $\mathrm{GF}(2^\ell)$

$$\mathrm{Prob}[h(x) = u \text{ and } h(y) = v] = \mathrm{Prob}[h(x) = u] \cdot \mathrm{Prob}[h(y) = v] = 2^{-2\ell}.$$

Any other ensemble of hash functions with these properties would do as well.[8]

Here is the definition of $G(\sigma)$ by a picture.



[7] This assumes that we can perform addition and multiplication in $\mathrm{GF}(2^\ell)$ efficiently. For this we need a concrete representation of $\mathrm{GF}(2^\ell)$, i.e., an irreducible polynomial of degree ℓ over GF(2). Such a polynomial can be stored in ℓ bits, and it is known that it can be found deterministically in time polynomial in ℓ.

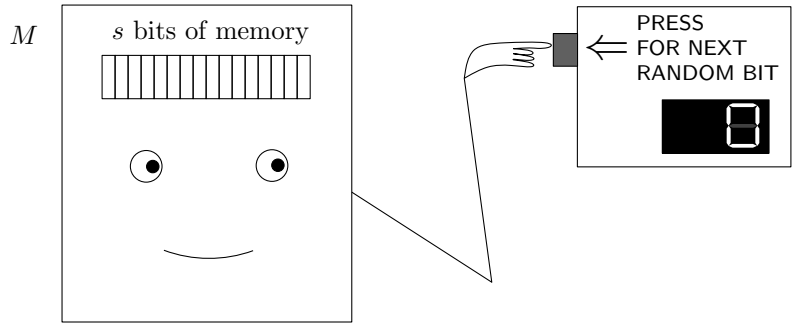
[8] Here is another suitable family: A hash function h is defined by $h(x) := a * x + b$, where $a \in \{0,1\}^{2\ell-1}$, $b \in \{0,1\}^\ell$, and "$*$" stands for convolution, i.e., $(a * x)_i = \sum_{j=1}^{n} a_{i+j-1} x_j$, with addition modulo 2. Thus, h is described by $3\ell - 1$ bits in this case. Here we need not worry about the arithmetic in $\mathrm{GF}(2^\ell)$ as in the previous case.

We construct a complete binary tree starting with a single node at level 0 with value σ_0. For a node at level i with value x, we construct two nodes at level $i+1$ with values x and $h_i(x)$. The string $G(\sigma)$ of length $\ell 2^\ell$ is the concatenation of the values of the leaves of the tree, on level ℓ, from left to right.

As we have seen, for our application in ℓ_2 norm estimation, we want a "random access" to the pseudorandom bits, and the above construction indeed provides it: Given σ and an index t of a position in $G(\sigma)$, we can compute the tth bit of $G(\sigma)$ in space $O(\ell^2)$ using $O(\ell)$ arithmetic operations, by taking the appropriate root-to-leaf path in the binary tree.

Fooling a space-bounded machine. We now describe in a semi-formal manner the theoretical guarantees offered by G. The main result says that G fools all randomized machines using space at most s, provided that $\ell \geq Cs$ for a sufficiently large constant C.

A machines M of the kind we're considering can be thought of as follows.



It has s bits of working memory, i.e., 2^s possible states. The computation begins at the state where all memory bits of M are 0.

The state may change in each step of M. The machine can also use a source of random bits: We can imagine that the source is a box with a button, and whenever M presses the button, the box displays a new random bit. In each step, M passes to a new state depending on its current state and on the random bit currently displayed on the random source. The mapping assigning the new state to the old state and to the current random bit is called the *transition function* of M.

Computers normally accept some inputs, and so the reader can ask, where is the input of M? Usually such computational models are presented as being able to read some input tape. But for our very specific purposes, we can assume that the input is hard-wired in the machine.

Indeed, we put no limits at all on the transition function of M, and so it can implicitly contain some kind of input.

We assume that for every sequence $\omega = \omega_1\omega_2\omega_3\cdots$ of random bits produced by the source M runs for at most 2^s steps and then stops with three ear-piercing beeps. After the beeps we read the current state of M, and this defines a mapping, which we also denote by M, assigning the final state to every string ω of random bits. We can assume that ω has length 2^s, since M can't use more random bits anyway.

For every probability distribution on the set of all possible values of ω, the machine M defines a probability distribution on its states. We will consider two such distributions. First, for ω *truly random*, i.e., each string of length 2^s having probability 2^{-2^s}, the probability of a state q is

$$p_{\text{truly}}(q) = \frac{|\{\omega \in \{0,1\}^{2^s} : M(\omega) = q\}|}{2^{2^s}}.$$

Now let's suppose that truly random bits are very expensive. We thus set $\ell := Cs$ and buy only $2\ell^2 + \ell$ truly random bits as the seed σ for the generator G. Then we run the machine M on the much cheaper bits from $G(\sigma)$. When σ is picked uniformly at random, this defines another probability distribution on the states of M:

$$p_{\text{pseudo}}(q) = \frac{|\{\sigma \in \{0,1\}^{2\ell^2+\ell} : M(G(\sigma)) = q\}|}{2^{2\ell^2+\ell}}.$$

The next theorem tells us that there is almost no difference; the cheap bits work just fine.

2.6.1 Theorem (Nisan's generator). *Let s be a given natural number. If the generator G is constructed as above with $\ell \geq Cs$, where C is a sufficiently large constant, then for all machines M with at most s bits of memory, the probability distributions $p_{\text{truly}}(.)$ and $p_{\text{pseudo}}(.)$ are $2^{-\ell/10}$-close. This means that*

$$\sum_q |p_{\text{truly}}(q) - p_{\text{pseudo}}(q)| \leq 2^{-\ell/10},$$

where the sum extends over all states of M.

The proof is nice and not too hard; it is not so much about machines as about random and pseudorandom walks in an acyclic graph. Here we omit it.

Now we're ready to fix the random projection algorithm.

2.6.2 Theorem. *There is a randomized algorithm for the ℓ_2 norm estimation problem that, given n, ε and δ and having read any given input stream, computes a number that with probability at least $1-\delta$ lies within $(1\pm\varepsilon)\|\mathbf{x}\|_2$. It uses $O(\varepsilon^{-2}\log\frac{n}{\varepsilon\delta}+(\log\frac{n}{\varepsilon\delta})^2)$ bits of memory, which for ε and δ constant is $O(\log^2 n)$.*

Proof. We set $s := C_0\log\frac{n}{\varepsilon\delta}$ for a suitable constant C_0, and we generate and store a random seed σ for Nisan's generator of the appropriate length (about s^2).

Then, as was suggested earlier, with $k := C\varepsilon^{-2}\log\frac{1}{\delta}$, we read the stream and maintain $A\mathbf{x}$, where A is a $k\times n$ pseudorandom ±1 matrix. This needs $O(k\log(nk))$ bits of memory, since the largest integers encountered in the computation are bounded by a polynomial in n and k.

Each entry of A is determined by the appropriate bit of $G(\sigma)$, and so when we need the ith column, we just generate the appropriate k bits of $G(\sigma)$. (For the proof to work, we need to assume that A is generated row by row; that is, the first row of A is determined by the first n bits of $G(\sigma)$, the second row by the next n bits, etc. For the algorithm itself this is admittedly somewhat unnatural, but it doesn't cause any serious harm.) At the end of the stream we output $\frac{1}{\sqrt{k}}\|A\mathbf{x}\|_2$ as the norm estimate.

As we have said, if A is truly random, then $\frac{1}{\sqrt{k}}\|A\mathbf{x}\|_2$ is a satisfactory estimate for the norm. To see that it also works when A is the pseudorandom matrix, we want to apply Theorem 2.6.1. An important point is that we don't apply it to the above algorithm (this wouldn't work, since that algorithm doesn't use the random bits sequentially). Rather, we use the theorem for a hypothetical machine M, which we now construct.

Let $\mathbf{x}$ be fixed. The machine M has the value of $\mathbf{x}$ hard-wired in it, as well as the value of ε. It reads random bits from its random source, makes them into entries of A, and computes $\|A\mathbf{x}\|_2^2$. Since A is generated row-by-row, then the entries of $A\mathbf{x}$ are computed one by one, and M needs to remember only two intermediate results, which needs $O(\log(nk))$ bits. (The machine also has to maintain a counter in range from 1 to nk in order to remember how far the computation has progressed, but this is also only $\log(nk)$ bits.)

The machine then checks whether $k^{-1/2}\|A\mathbf{x}\|_2$ lies within $(1\pm\varepsilon)\|\mathbf{x}\|_2$. (No square roots are needed since the squares can be compared.) If it does, M finishes in a state called GOOD, and otherwise, in a state called BAD.

We know that if M is fed with truly random bits, then GOOD has probability at least $1-\delta$. So by Theorem 2.6.1, if M runs on the pseudo-

random bits from Nisan's generator, it finishes at GOOD with probability at least $1-\delta-2^{-\ell/10}\ge 1-2\delta$. But this means that $k^{-1/2}\|A\mathbf{x}\|_2$ is in the desired interval with probability at least $1-2\delta$, where the probability is with respect to the random choice of the seed σ. This proves that the algorithm has the claimed properties.

Let us stress that the machine M has no role in the algorithm. It was used solely for the proof, to show that the distribution of $\|A\mathbf{x}\|_2$ is not changed much by replacing random A by a pseudorandom one. □

We've ignored another important issue, the *running time* of the algorithm. But a routine extension of the above analysis shows that the algorithm runs quite fast. For δ and ε fixed it uses only $O(\log n)$ arithmetic operations on $O(\log n)$-bit numbers per instruction of the stream.

Heavy hitters. The above method allows us to estimate x_i for a given i with (absolute) error at most $\varepsilon\|\mathbf{x}\|_2$, for a prescribed ε. The space used by the algorithm again depends on ε. Then we can detect whether x_i is exceptionally large, i.e. contributes at least 1% of the ℓ_2 norm, say.

The idea is that $x_i=\langle\mathbf{x},\mathbf{e}_i\rangle$, where $\mathbf{e}_i$ is the ith unit vector in the standard basis, and this scalar product can be computed, by the cosine theorem, from $\|\mathbf{x}\|_2$, $\|\mathbf{e}_i\|_2=1$, and $\|\mathbf{x}-\mathbf{e}_i\|_2$. We can approximate $\|\mathbf{x}\|_2$ and $\|\mathbf{x}-\mathbf{e}_i\|_2$ by the above method, and this yields an approximation of x_i. We omit the calculations.

2.7 Explicit embedding of ℓ_2^n in ℓ_1

In Section 1.5 we showed that every ℓ_2 metric embeds in ℓ_1. We used an isometric embedding $\ell_2^n\to L_1(S^{n-1})$ defined by a simple formula but going into an infinite-dimensional space. Later, in Section 2.5, we saw that a random $Cn\times n$ matrix A with independent Gaussian entries defines, with high probability, an almost-isometry $T\colon\ell_2^n\to\ell_1^{Cn}$.

Can't one just write down a *specific* matrix A for such an embedding? This question has been puzzling mathematicians for at least 30 years and it has proved surprisingly difficult.

The notion of *explicit construction* is seldom used in a precisely defined sense in classical mathematics; mathematicians usually believe they can recognize an explicit construction when they see one.

Theoretical computer science does offer a formal definition of "explicit": In our case, for example, a $k\times n$ matrix A can be regarded as given explicitly if there is an algorithm that, given n and k, outputs A in time polynomial in $n+k$. (For some purposes, computer scientists

prefer even "more explicit" constructions, which have a very fast *local* algorithm; in our case, an algorithm that, given n, k, i, j, computes the entry a_{ij} in time polynomial in $\log(n+k)$.) Taken seriously, this definition of "explicit" has led to very interesting and valuable methods and results. But, quite often, the resulting explicit constructions are very far from the intuitive idea of "something given by a formula" and when classical mathematicians see them, the most likely reaction may be "this is not what we meant!".

In any case, so far nobody has managed to construct a polynomially computable matrix A defining an ε-almost isometric embedding $\ell_2^n \to \ell_1^{C(\varepsilon)n}$. There are several weaker results, in which either the distortion is not arbitrarily close to 1, or the target dimension is not even polynomially bounded.

The current strongest results use too many tools to be presented here, but we explain some weaker results, which can serve as an introduction to the more advanced ones in the literature.

An explicit embedding in an exponential dimension. First we would like to see an explicit $O(1)$-embedding of ℓ_2^n in ℓ_1^k for some k, possibly huge but finite. We have indicated one possible route in Section 1.5, through a "discretization" the function space $L_1(S^{n-1})$. Now we take a different path.

Let $k := 2^n$, let A be the $k \times n$ matrix whose rows are all the 2^n possible vectors of $+1$'s and -1's, and let $T\colon \mathbb{R}^n \to \mathbb{R}^k$ be given by $\mathbf{x} \mapsto 2^{-n} A\mathbf{x}$. We claim that T is an $O(1)$-embedding of ℓ_2^n in ℓ_1^k.

For $\mathbf{x}$ fixed, $\|T(\mathbf{x})\|_1$ is the average of $|\pm x_1 \pm x_2 \pm \cdots \pm x_n|$ over all choices of signs. In probabilistic terms, if we set $X := \sum_{j=1}^n \epsilon_j x_j$, where $\epsilon_1, \dots, \epsilon_n$ are independent uniform ± 1 random variables, then $\|T(\mathbf{x})\|_1 = \mathbf{E}[\,|X|\,]$. Thus, the fact that T is an $O(1)$-embedding follows from the next lemma.

2.7.1 Lemma (A special case of Khintchine's inequality). *Let $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ be independent random variables, each attaining values $+1$ and -1 with probability $\frac{1}{2}$ each, let $\mathbf{x} \in \mathbb{R}^n$, and let $X := \sum_{j=1}^n \epsilon_j x_j$. Then*

$$\frac{1}{\sqrt{3}}\|\mathbf{x}\|_2 \le \mathbf{E}[\,|X|\,] \le \|\mathbf{x}\|_2.$$

Proof. The following proof is quick but yields a suboptimal constant (the optimal constant is $2^{-1/2}$). On the other hand, it contains a useful trick, and later we'll use some of its features.

We will need Hölder's inequality, which is usually formulated for vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ in basic courses: $\langle \mathbf{a}, \mathbf{b}\rangle \le \|\mathbf{a}\|_p \|\mathbf{b}\|_q$, where $1 \le p \le \infty$ and $\frac{1}{q} = 1 - \frac{1}{p}$ ($p = q = 2$ is the Cauchy–Schwarz inequality). We will use a formulation for random variables A, B: $\mathbf{E}[AB] \le \mathbf{E}[|A|^p]^{1/p}\,\mathbf{E}[|B|^q]^{1/q}$. For the case we need, where A and B attain finitely many values, this version immediately follows from the one for vectors.

We may assume $\|\mathbf{x}\|_2 = 1$. We know (or calculate easily) that $\mathbf{E}\big[X^2\big] = 1$.

The upper bound $\mathbf{E}[\,|X|\,] \le \mathbf{E}\big[X^2\big]^{1/2} = 1$ follows immediately from the Cauchy–Schwarz inequality with $A := |X|$ and $B := 1$ (a constant random variable).

For the lower bound we first need to bound $\mathbf{E}\big[X^4\big]$ from above by some constant. Such a bound could be derived from the subgaussian tail of X (Lemma 2.4.3), but we calculate directly, using linearity of expectation,

$$\mathbf{E}\big[X^4\big] = \sum_{i,j,k,\ell=1}^{n} \mathbf{E}[\epsilon_i\epsilon_j\epsilon_k\epsilon_\ell]\, x_i x_j x_k x_\ell.$$

Now if, say, $i \notin \{j,k,\ell\}$, ϵ_i is independent of $\epsilon_j, \epsilon_k, \epsilon_\ell$, and so $\mathbf{E}[\epsilon_i\epsilon_j\epsilon_k\epsilon_\ell] = \mathbf{E}[\epsilon_i]\,\mathbf{E}[\epsilon_j\epsilon_k\epsilon_\ell] = 0$. Hence all such terms in the sum vanish.

The remaining terms are of the form $\mathbf{E}\big[\epsilon_s^4\big]\, x_s^4 = x_s^4$ for some s, or $\mathbf{E}\big[\epsilon_s^2\epsilon_t^2\big]\, x_s^2 x_t^2 = x_s^2 x_t^2$ for some $s \ne t$. Given some values $s < t$, we have $\binom{4}{2} = 6$ ways of choosing two of the summation indices i, j, k, ℓ to have value s, and the other two indices get t. Hence

$$\begin{aligned}\mathbf{E}\big[X^4\big] &= \sum_{s=1}^{n} x_s^4 + \sum_{1\le s<t\le n} 6x_s^2x_t^2 \\ &< 3\Big(\sum_{s=1}^{n} x_s^4 + \sum_{1\le s<t\le n} 2x_s^2x_t^2\Big) = 3\|\mathbf{x}\|_2^4 = 3.\end{aligned}$$

Now we want to use Hölder's inequality so that $\mathbf{E}[\,|X|\,]$ shows up on the *right-hand* (larger) side together with $\mathbf{E}[X^4]$, while $\mathbf{E}[X^2]$ stands on the left. A simple calculation reveals that the right choices are $p := \frac{3}{2}$, $q = 3$, $A := |X|^{2/3}$, and $B := |X|^{4/3}$, leading to

$$\begin{aligned}1 &= \mathbf{E}[X^2] = \mathbf{E}[AB] \le \mathbf{E}[A^p]^{1/p}\mathbf{E}[B^q]^{1/q} \\ &= \mathbf{E}[\,|X|\,]^{2/3}\,\mathbf{E}\big[X^4\big]^{1/3} \le \mathbf{E}[\,|X|\,]^{2/3}\,3^{1/3},\end{aligned}$$

and $\mathbf{E}[\,|X|\,] \ge 3^{-1/2}$ follows. □

In the above we used a relation between $\mathbf{E}[X]$ and the embedding in ℓ_1^k. Before we proceed with reducing the embedding dimension, let us formulate this relation in a more general setting. The proof of the next observation is just a comparison of definitions:

2.7.2 Observation. *Let $R_1, R_2, \dots, R_n$ be real random variables on a probability space that has k elements (elementary events) $\omega_1, \omega_2, \dots, \omega_k$, and let A be the $k \times n$ matrix with $a_{ij} := \mathrm{Prob}[\omega_i]\, R_j(\omega_i)$. For $\mathbf{x} \in \mathbb{R}^n$ let us set $X := \sum_{j=1}^n R_j x_j$. Then $\mathbf{E}[\,|X|\,] = \|A\mathbf{x}\|_1$.* □

Reducing the dimension. This observation suggests that, in order to reduce the dimension 2^n in the previous embedding, we should look for suitable random variables on a smaller probability space. By inspecting the proof of Lemma 2.7.1, we can see that the following properties of the ϵ_j are sufficient:

(i) Every ϵ_j attains values $+1$ and -1, each with probability $\frac{1}{2}$.

(ii) Every 4 of the ϵ_j are independent.

Property (ii) is called **4-wise independence**. In theoretical computer science, t-wise independent random variables have been recognized as an important tool, and in particular, there is an explicit construction, for every n, of random variables $\epsilon_1, \dots, \epsilon_n$ with properties (i) and (ii) but on a probability space of size only $O(n^2)$.[9]

[9]For someone not familiar with t-wise independence, the first thing to realize is probably that 2-wise independence (every two of the variable independent) is not the same as n-wise independence (all the variables independent). This can be seen on the example of 2-wise independent random variables below.

Several constructions of t-wise independent random variables are based on the following simple linear-algebraic lemma: *Let A be an $m \times n$ matrix over the 2-element field* GF(2) *such that every t columns of A are linearly independent. Let $\mathbf{x} \in \mathrm{GF}(2)^m$ be a random vector (each of the 2^m possible vectors having probability 2^{-m}), and set $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n) := A\mathbf{x}$. Then $\epsilon_1, \dots, \epsilon_n$ are t-wise independent random variables (on a probability space of size 2^m).*

For $t = 2$, we can set $n := 2^m - 1$ and let the columns of A be all the nonzero vectors in $\mathrm{GF}(2)^m$. Every two columns are distinct, and thus linearly independent, and we obtain n pairwise independent random variables on a probability space of size $n + 1$.

Here is a more sophisticated construction of $(2r + 1)$-wise independent random variables on a probability space of size $2(n + 1)^r$ (with $r = 2$ it can be used for the proof of Proposition 2.7.3). Let $n = 2^q - 1$ and let $\alpha_1, \dots, \alpha_n$ be an enumeration of all nonzero elements of the field $\mathrm{GF}(2^q)$. In a representation of $\mathrm{GF}(2^q)$ using a degree-q irreducible polynomial over GF(2), each α_i can be regarded as a q-element column vector in $\mathrm{GF}(2)^q$. The matrix A, known as the *parity check matrix of a BCH code*, is set up as follows:

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \alpha_1^3 & \alpha_2^3 & \dots & \alpha_n^3 \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_1^{2r-1} & \alpha_2^{2r-1} & \dots & \alpha_n^{2r-1} \end{pmatrix};$$

here, e.g., $\alpha_1, \alpha_2, \dots, \alpha_n$ represents m rows of A, since each α_i is interpreted a column vector of q entries. Thus A has $m = qr + 1$ rows and $n = 2^q - 1$ columns. *If* we used the larger matrix with $2qr + 1$ rows containing all the powers $\alpha_i^1, \alpha_i^2, \dots, \alpha_i^{2r}$ in the columns, the linear independence of every $2r + 1$ columns follows easily by the nonsingularity of a Vandermonde matrix. An additional trick is needed to show that the even powers can be omitted.

In view of the above discussion, this implies the following explicit embedding:

2.7.3 Proposition. *There is an explicit $\sqrt{3}$-embedding $\ell_2^n \to \ell_1^{O(n^2)}$.* □

Getting distortions close to 1. We know that for $X := \sum_{j=1}^n Z_j x_j$, with $Z_1, \dots, Z_n$ independent standard normal, $\mathbf{E}[\,|X|\,]$ is exactly proportional to $\|\mathbf{x}\|_2$. We will now approximate the Z_j by suitable discrete random variables on a finite probability space, which will provide an embedding $\ell_2^n \to \ell_1^k$ with distortion very close to 1 but with k very large. But then we'll be able to reduce k considerably using Nisan's pseudorandom generator from Theorem 2.6.1.

There are many possible ways of "discretizing" the standard normal random variables. Here we use one for which Nisan's generator is very easy to apply and which relies on a generally useful theorem.

Namely, for an integer parameter b, we set $Z_j' := b^{-1/2} \sum_{\ell=1}^b \epsilon_{j\ell}$, where the $\epsilon_{j\ell}$ are independent uniform ± 1 random variables. So Z_j' has a binomial distribution which, by the central limit theorem, approaches the standard normal distribution as $b \to \infty$. But we won't use this directly. What we really need is that for $X' := \sum_{j=1}^n Z_j' x_j$ with $\mathbf{x}$ unit, $\mathbf{E}[\,|X'|\,]$ is close to $\mathbf{E}[\,|Z|\,]$ for Z standard normal.

The Berry–Esséen theorem from probability theory quantifies how the distribution of a sum of n independent random variables approaches the standard normal distribution; one can find numerous variants in the literature. We will use the following Berry–Esséen-type result.

2.7.4 Theorem. *Let $\epsilon_1, \dots, \epsilon_n$ be independent uniform ± 1 random variables and let $\boldsymbol{\alpha} \in \mathbb{R}^n$ satisfy $\|\boldsymbol{\alpha}\|_2 = 1$. Then for $Y := \sum_{j=1}^n \epsilon_j \alpha_j$*

$$\Big|\mathbf{E}[\,|Y|\,] - \beta\Big| \le C\|\boldsymbol{\alpha}\|_\infty = C \max|\alpha_j|,$$

where C is an absolute constant and $\beta := \mathbf{E}[\,|Z|\,]$ with Z standard normal.

This can be viewed as a strengthening of Khintchine's inequality (e.g., of Lemma 2.7.1)—it tells us that if none of the coefficients α_j is too large, then $\mathbf{E}[|\sum_{j=1}^n \epsilon_j \alpha_j|]$ is almost determined by $\|\boldsymbol{\alpha}\|_2$.

2.7.5 Corollary. *Let the $Z'_j = b^{-1/2}\sum_{\ell=1}^b \epsilon_{j\ell}$ and $X' = \sum_{j=1}^n Z'_j x_j$ be as above, $\|\mathbf{x}\|_2 = 1$. Then $\mathbf{E}[\,|X'|\,] = \beta + O(b^{-1/2})$.*

Proof. We use the theorem with

$$\boldsymbol{\alpha} := b^{-1/2}(\underbrace{x_1, x_1, \ldots, x_1}_{b \text{ times}}, \underbrace{x_2, \ldots, x_2}_{b \text{ times}}, \ldots, \underbrace{x_n, \ldots, x_n}_{b \text{ times}}).$$

□

The corollary as is provides, via Observation 2.7.2, an explicit embedding $\ell_2^n \to \ell_1^k$ with $k = 2^{bn}$ and with distortion $1 + O(b^{-1/2})$. The dimension can be reduced considerably using Nisan's generator:

2.7.6 Proposition. *There is an explicit embedding $\ell_2^n \to \ell_1^k$ with $k = n^{O(\log n)}$ and with distortion $1 + O(n^{-c})$, where the constant c can be made as large as desired.*

Proof. We can think of each Z'_j in Corollary 2.7.5 as determined by a block of b of truly random bits. Instead, let us set $s := \lceil C_1 \log_2(nb)\rceil$ for a suitable constant C_1, let $\ell := Cs$ as in Theorem 2.6.1, and let σ be a string of $2\ell^2+\ell$ truly random bits. Let us define $\tilde{Z}_j$ using the appropriate block of b bits from $G(\sigma)$, and let $\tilde{X} := \sum_{j=1}^n \tilde{Z}_j x_j$. It suffices to set $b := n^{2c}$ and to show that $\bigl|\mathbf{E}[\,|\tilde{X}|\,] - \mathbf{E}[\,|X'|\,]\bigr| = O(b^{-1/2})$.

Let M be a hypothetical machine with working space s, of the kind considered in Theorem 2.6.1, that with a source of truly random bits approximates X' with accuracy at most $b^{-1/2}$. That is, the final state of M encodes a number (random variable) Y' such that $|X' - Y'| \le b^{-1/2}$. For such task, working space s is sufficient.

If M is fed with the pseudorandom bits of $G(\sigma)$ instead, its final state specifies a random variable $\tilde{Y}$ with $|\tilde{X} - \tilde{Y}| \le b^{-1/2}$. Theorem 2.6.1 guarantees that

$$\sum_y \Bigl|\mathrm{Prob}[Y' = y] - \mathrm{Prob}[\tilde{Y} = y]\Bigr| \le 2^{-\ell/10}.$$

Since Y' and $\tilde{Y}$ obviously cannot exceed $2n$ (a tighter bound is $\sqrt{n} + O(b^{-1/2})$ but we don't care), we have

$$\begin{aligned}\Bigl|\mathbf{E}[\,|Y'|\,] - \mathbf{E}[\,|\tilde{Y}|\,]\Bigr| &\le \sum_y |y| \cdot \Bigl|\mathrm{Prob}[Y' = y] - \mathrm{Prob}\Bigl[\tilde{Y} = y\Bigr]\Bigr| \\ &\le 2n \cdot 2^{-\ell/10} \le b^{-1/2}.\end{aligned}$$

So $\mathbf{E}[\,|X'|\,]$ and $\mathbf{E}[\,|\tilde{X}|\,]$ indeed differ by at most $O(b^{-1/2})$.

The random variable $\tilde{X}$ is defined from $2\ell^2 + \ell = O(\log^2 n)$ random bits, and thus we obtain an embedding in ℓ_1^k with $k = \exp(O(\log^2 n)) = n^{O(\log n)}$. □

Currently there are two mutually incomparable best results on explicit embeddings $\ell_2^n \to \ell_1^k$. One of them provides distortions close to 1, namely, $1 + O(\frac{1}{\log n})$, and a slightly superlinear dimension $k = n2^{O((\log\log n)^2)}$. The other has a sublinear distortion $n^{o(1)}$ but the dimension is only $k = (1 + o(1))n$.

2.8 Error correction and compressed sensing

Error-correction over the reals. A cosmic probe wants to send the results of its measurements, represented by a vector $\mathbf{w} \in \mathbb{R}^m$, back to Earth. Some of the numbers may get corrupted during the transmission. We assume the possibility of *gross errors*; that is, if the number 3.1415 is sent and it gets corrupted, it can be received as 3.1425, or 2152.66, or any other real number.

We would like to convert (encode) $\mathbf{w}$ into another vector $\mathbf{z}$, so that if no more than 8%, say, of the components of $\mathbf{z}$ get corrupted, we can still recover the original $\mathbf{w}$ exactly.

This problem belongs to the theory of *error-correcting codes*. In this area one usually deals with encoding messages composed of letters of a finite alphabet, while our "letters" are arbitrary real numbers.

In order to allow for error recovery, the encoding $\mathbf{z}$ has to be longer than the original $\mathbf{w}$. Let its length be n, while $k := n - m$ is the "excess" added by the coding.

We will use a linear encoding, setting $\mathbf{z} := G\mathbf{w}$ for a suitable $n \times m$ matrix G (analogous to the generator matrix for linear error-correcting codes).

Let $\tilde{\mathbf{z}}$ be the received vector. Let r be the maximum number of errors that the code should still be able to correct. That is, we assume that the

error vector $\mathbf{x} := \mathbf{z} - \tilde{\mathbf{z}}$ has at most r nonzero components. We call such an $\mathbf{x}$ **r-sparse**, or just *sparse* when r is understood.

How can we hope to recover the original message $\mathbf{w}$ from $\tilde{\mathbf{z}}$? We concentrate on finding the error vector $\mathbf{x}$ first, since then $\mathbf{w}$ can be computed by solving a system of linear equations. Let us assume that the matrix G has the full rank m, i.e., its columns span an m-dimensional linear subspace L of $\mathbb{R}^n$.

Then the kernel $\mathrm{Ker}(G^T) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T G = \mathbf{0}\}$, i.e., the orthogonal complement of L, has dimension $k = n-m$. Let A be a $k \times n$ matrix whose rows span $\mathrm{Ker}(G^T)$ (this is an analog of the *parity check matrix* for linear codes). Then AG is the zero matrix, and we have $A\tilde{\mathbf{z}} = A(G\mathbf{w} + \mathbf{x}) = 0\mathbf{w} + A\mathbf{x}$. Hence the unknown error vector $\mathbf{x}$ is a solution to $A\mathbf{x} = \mathbf{b}$, where A and $\mathbf{b} := A\tilde{\mathbf{z}}$ are known.

There are more unknowns than equations in this system, so it has infinitely many solutions. But we're not interested in all solutions—we're looking for one with at most r nonzero components.

Later in this section, we will show that if A is a random matrix as in the random projection lemma, then a sparse solution of $A\mathbf{x} = \mathbf{b}$ can be efficiently computed, provided that one exists, and this provides a solution to the decoding problem.

Naturally, the encoding length n has to be sufficiently large in terms of the message length m and the number r of allowed errors. It turns out that we will need k, the "excess", at least of order $r \log \frac{n}{r}$. As a concrete numerical example, it is known that when we require $r = 0.08n$, i.e., about 8% of the transmitted data may be corrupted, we can take $n = 1.33m$, i.e., the encoding expands the message by 33%.

Compressed sensing (or *compressive sensing* according to some authors) is an ingenious idea, with great potential of practical applications, which also leads to the problem of finding sparse solutions of systems of linear equations. To explain the idea, we begin with a slightly different topic—encoding of digital images.

A digital camera captures the image by means of a large number n of sensors; these days one may have n around ten millions in more expensive cameras. The outputs of these sensors can be regarded as a vector $\mathbf{s} \in \mathbb{R}^n$ (the components are known only approximately, of course, but let's ignore that).

The picture is usually stored in a compressed format using a considerably smaller amount of data, say a million of numbers (and this much is needed only for large-format prints—hundreds of thousand numbers amply suffice for a computer display or small prints).

The compression is done by complicated and mathematically beautiful methods, but for now, it suffices to say that the image is first expressed as a linear combination of suitable basis vectors. If we think of the image $\mathbf{s}$ as a real function defined on a fine grid of n points in the unit square, then the basis vectors are usually obtained as restrictions of cleverly chosen continuous functions to that grid. The usual JPEG standard uses products of cosine functions, and the newer JPEG2000 standard uses the fancier Cohen–Daubechies–Feauveau (or LeGall) wavelets.

But abstractly speaking, one writes $\mathbf{s} = \sum_{i=1}^n x_i \mathbf{b}_i$, where $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is the chosen basis. For an everyday picture $\mathbf{s}$, most of the coefficients x_i are zero or very small. (Why? Because the basis functions have been chosen so that they can express well typical features of digital images.) The very small coefficients can be discarded, and only the larger x_i, which contain almost all of the information, are stored.

We thus gather information by 10^7 sensors and then we reduce it to, say, 10^6 numbers. Couldn't we somehow acquire the 10^6 numbers right away, without going through the much larger raw image?

Digital cameras apparently work quite well as they are, so there is no urgency in improving them. But there are applications where the number of sensors matters a lot. For example, in medical imaging, with fewer sensors the patient is less exposed to harmful radiation and can spend less time inside various unpleasant machines. Compressed sensing provides a way of using much fewer sensors. Similarly, in astronomy light and observation time of large telescopes are scarce resources, and compressed sensing might help observers gain the desired information faster. More generally, the idea may be applicable whenever one wants to measure some signal and then extract information from it by means of linear transforms.

We thus consider the expression $\mathbf{s} = \sum_i x_i \mathbf{b}_i$. Each coefficient x_i is a linear combination of the entries of $\mathbf{s}$ (we're passing from one basis of $\mathbb{R}^n$ to another). It is indeed technically feasible to make sensors that acquire a given x_i directly, i.e., they measure a prescribed linear combination of light intensities from various points of the image.

However, a problem with this approach is that we don't know in advance which of the x_i are going to be important for a given image, and thus which linear combinations should be measured.

The research in compressed sensing has come up with a surprising solution: Don't measure any particular x_i, but measure an appropriate number of *random linear combinations* of the x_i (each linear combination of the x_i corresponds to a uniquely determined combination of the s_i and

so we assume that it can be directly "sensed").

Then, with very high probability, whenever we measure these random linear combinations for an image whose corresponding $\mathbf{x}$ is r-sparse, we can exactly reconstruct $\mathbf{x}$ from our measurements. More generally, this works even if $\mathbf{x}$ is *approximately sparse*, i.e., all but at most r components are very small—then we can reconstruct all the not-so-small components.

Mathematically speaking, the suggestion is to measure the vector $\mathbf{b} := A\mathbf{x}$, where A is a random $k \times n$ matrix, with k considerably smaller than n. The problem of reconstructing a sparse $\mathbf{x}$ is precisely the problem of computing a sparse solution of $A\mathbf{x} = \mathbf{b}$. (Or an approximately sparse solution—but we will leave the approximately sparse case aside, mentioning only that it can be treated by extending the ideas discussed below.)

Sparse solutions of linear equations. We are thus interested in matrices A with n columns such that for every right-hand side $\mathbf{b}$, we can compute an r-sparse solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$, provided that one exists. Moreover, we want k, the number of rows, small.

If every at most $2r$ columns of A are linearly independent, then the sparse solution is guaranteed to be unique—showing this is an exercise in linear algebra. Unfortunately, even if A satisfies this condition, computing the sparse solution is computationally intractable (NP-hard) in general.

Fortunately, methods have been invented that find the sparse solution efficiently for a wide class of matrices. Roughly speaking, while the condition above for uniqueness of a sparse solution requires every $2r$ columns of A to be linearly independent, a sufficient condition for efficient computability of the sparse solution is that every $3r$ columns of A are nearly orthogonal. In other words, the linear mapping $\mathbb{R}^{3r} \to \mathbb{R}^k$ defined by these columns should be a (Euclidean) ε_0-almost isometry for a suitable small constant ε_0.

Basis pursuit. As we will prove, for a matrix A satisfying the condition just stated, a sparse solution $\mathbf{x}$ can be found as a solution to the following minimization problem:

$$\text{Minimize } \|\mathbf{x}\|_1 \text{ subject to } \mathbf{x} \in \mathbb{R}^n \text{ and } A\mathbf{x} = \mathbf{b}. \qquad \text{(BP)}$$

That is, instead of looking for a solution $\mathbf{x}$ with the smallest number of nonzero components, we look for a solution with the smallest ℓ_1 norm. This method of searching for sparse solutions is called the *basis pursuit* in the literature, for reasons which we leave unexplained here.

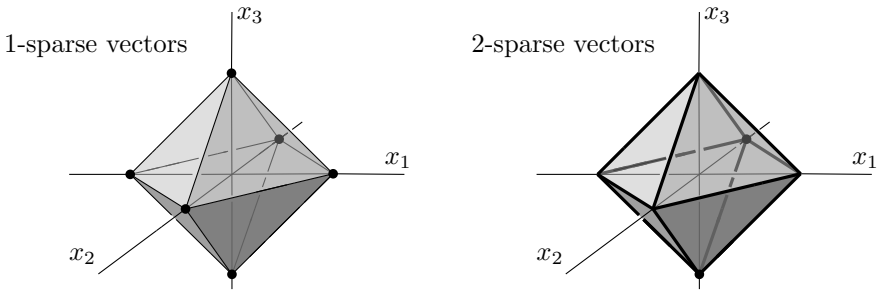
Let us call the matrix A **BP-exact** (for sparsity r) if for all $\mathbf{b} \in \mathbb{R}^m$ such that $A\mathbf{x} = \mathbf{b}$ has an r-sparse solution $\tilde{\mathbf{x}}$, the problem (BP) has $\tilde{\mathbf{x}}$ as the unique minimum.

The problem (BP) can be re-formulated as a linear program, i.e., as minimizing a linear function over a region defined by a system of linear equations and inequalities. Indeed, we can introduce n auxiliary variables $u_1, u_2, \ldots, u_n$ and equivalently formulate (BP) as finding

$$\min\{u_1 + u_2 + \cdots + u_n : \mathbf{u}, \mathbf{x} \in \mathbb{R}^n, A\mathbf{x} = \mathbf{b}, \\ -u_j \le x_j \le u_j \text{ for } j = 1, 2, \ldots, n\}.$$

Such linear programs can be solved quite efficiently.[10]

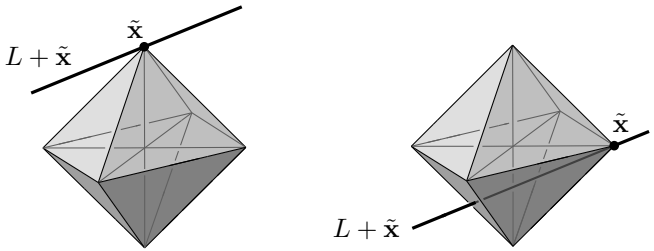
Geometric meaning of BP-exactness. The set of all r-sparse vectors in $\mathbb{R}^n$ is a union of r-dimensional coordinate subspaces. We will consider only r-sparse $\tilde{\mathbf{x}}$ with $\|\tilde{\mathbf{x}}\|_1 = 1$ (without loss of generality, since we can re-scale the right-hand side $\mathbf{b}$ of the considered linear system). These vectors constitute exactly the union of all $(r-1)$-dimensional faces of the unit ℓ_1 ball B_1^n (generalized octahedron), as the next picture illustrates for $n = 3$ and $r = 1, 2$.



Let A be a $k \times n$ matrix of rank k and let $L := \operatorname{Ker} A$; then $\dim L = n - k = m$. A given r-sparse vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$ satisfies the linear system $A\mathbf{x} = \mathbf{b}_{\tilde{\mathbf{x}}}$, where $\mathbf{b}_{\tilde{\mathbf{x}}} := A\tilde{\mathbf{x}}$, and the set of all solutions of this system is a translate of L, namely, $L + \tilde{\mathbf{x}}$.

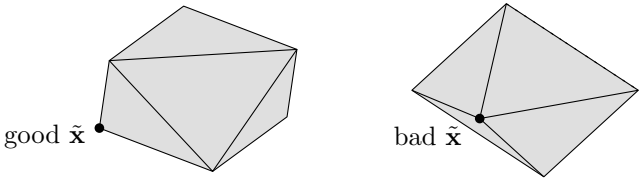
[10] Recently, alternative and even faster methods have been developed for computing a sparse solution of $A\mathbf{x} = \mathbf{b}$, under similar conditions on A, although they find the sparse solution only approximately.

When is $\tilde{\mathbf{x}}$ the unique point minimizing the ℓ_1 norm among all points of $L+\tilde{\mathbf{x}}$? Exactly when the affine subspace $L+\tilde{\mathbf{x}}$ just touches the ball B_1^n at $\tilde{\mathbf{x}}$; here is an illustration for $n=3$, $\dim L=1$, and $r=1$:



Good $\tilde{\mathbf{x}}$ Bad $\tilde{\mathbf{x}}$

Let π be the orthogonal projection of $\mathbb{R}^n$ on the orthogonal complement of L. Then $L+\tilde{\mathbf{x}}$ touches B_1^n only at $\tilde{\mathbf{x}}$ exactly if $\pi(\tilde{\mathbf{x}})$ has $\tilde{\mathbf{x}}$ as the only preimage. In particular, $\pi(\tilde{\mathbf{x}})$ has to lie on the boundary of the projected ℓ_1 ball.



Thus, BP-exactness of A can be re-phrased as follows: Every point $\tilde{\mathbf{x}}$ in each $(r-1)$ face of the unit ℓ_1 ball should project to the boundary of $\pi(B_1^n)$, and should have a unique preimage in the projection. (We note that this condition depends only on the kernel of A.)

In the case $n=3$, $r=1$, $\dim L=1$, it is clear from the above pictures that if the direction of L is chosen randomly, there is at least some positive probability of all vertices projecting to the boundary, in which case BP-exactness holds. The next theorem asserts that if the parameters are chosen appropriately and sufficiently large, then BP-exactness occurs with overwhelming probability. We won't need the just explained geometric interpretation in the proof.[11]

[11] The geometric interpretation also explains why, when searching for a sparse solution, it isn't a good idea to minimize the Euclidean norm (although this task is also computationally feasible). If L is a "generic" subspace of $\mathbb{R}^n$ and a translate of L touches the Euclidean ball at a single point, then this point of contact typically has all coordinates nonzero.

2.8.1 Theorem (BP-exactness of random matrices). *There are constants C and $c>0$ such that if n,k,r are integers with $1\le r\le n/C$ and $k\ge Cr\log\frac{n}{r}$ and if A is a random $k\times n$ matrix, with independent uniform ±1 entries (or, more generally, with independent entries as in Lemma 2.4.1—the general version of the random projection lemma), then A is BP-exact for sparsity r with probability at least $1-e^{-ck}$.*

It is known that the theorem is asymptotically optimal in the following sense: For $k=o(r\log\frac{n}{r})$, *no* $k\times n$ matrix at all can be BP-exact for sparsity r.

Let us say that a matrix A has the property of ***r*-restricted Euclidean *ε*-almost isometry**[12] if the corresponding linear mapping satisfies the condition of ε-almost isometry with respect to the ℓ_2 norm for every sparse $\mathbf{x}$; that is, if

$$(1-\varepsilon)\|\mathbf{x}\|_2\le\|A\mathbf{x}\|_2\le(1+\varepsilon)\|\mathbf{x}\|_2$$

for all r-sparse $\mathbf{x}\in\mathbb{R}^n$.

The next lemma is the main technical part of the proof of Theorem 2.8.1.

2.8.2 Lemma. *There is a constant $\varepsilon_0>0$ such that if a matrix A has the property of $3r$-restricted Euclidean ε_0-almost isometry, then it is BP-exact for sparsity r.*

Let us remark that practically the same proof also works for restricted ℓ_2/ℓ_1 almost isometry (instead of Euclidean), i.e., assuming $(1-\varepsilon)\|\mathbf{x}\|_2\le\|A\mathbf{x}\|_1\le(1+\varepsilon)\|\mathbf{x}\|_2$ for all $3r$-sparse $\mathbf{x}$.

Proof of Theorem 2.8.1 assuming Lemma 2.8.2. Let B be a matrix consisting of some $3r$ distinct columns of A. Proceeding as in the proof of Theorem 2.5.1 with minor modifications, we get that the linear mapping $\ell_2^{3r}\to\ell_2^k$ given by B (and appropriately scaled) fails to be an ε_0-almost isometry with probability at most $e^{-c_1\varepsilon_0^2k}$ for some positive constant c_1.

The number of possible choices of B is $\binom{n}{3r}\le\left(\frac{en}{3r}\right)^{3r}\le\left(\frac{n}{r}\right)^{3r}=e^{3r\ln(n/r)}$, using a well-known estimate of the binomial coefficient. Thus, A fails to have the $3r$-restricted ε_0-isometry property with probability at most $e^{3r\ln(n/r)}e^{-c_1\varepsilon_0^2k}\le e^{-ck}$ for r,k,n as in the theorem. □

[12] Sometimes abbreviated as 2-RIP (RIP = Restricted Isometry Property, the 2 referring to the ℓ_2-norm).

Proof of Lemma 2.8.2. Let us suppose that A has the property of $3r$-restricted Euclidean ε_0-almost isometry, and that $\tilde{\mathbf{x}}$ is an r-sparse solution of $A\mathbf{x} = \mathbf{b}$ for some $\mathbf{b}$.

For contradiction, we assume that $\tilde{\mathbf{x}}$ is not the unique minimum of (BP), and so there is another solution of $A\mathbf{x} = \mathbf{b}$ with smaller or equal ℓ_1 norm. We write this solution in the form $\tilde{\mathbf{x}} + \boldsymbol{\Delta}$; so

$$A\boldsymbol{\Delta} = \mathbf{0}, \quad \|\tilde{\mathbf{x}} + \boldsymbol{\Delta}\|_1 \le \|\tilde{\mathbf{x}}\|_1.$$

We want to reach a contradiction assuming $\boldsymbol{\Delta} \neq \mathbf{0}$.

Let us note that *if* A were an almost-isometry, then $\boldsymbol{\Delta} \neq \mathbf{0}$ would imply $A\boldsymbol{\Delta} \neq \mathbf{0}$ and we would have a contradiction immediately. Of course, we cannot expect the whole of A to be an almost-isometry—we have control only over small blocks of A.

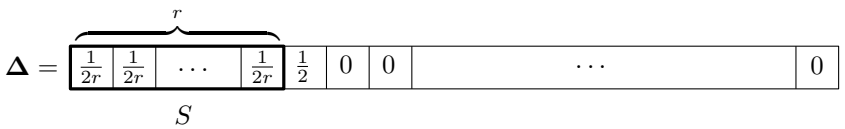
First we set $S := \{i : x_i \neq 0\}$ and we observe that at least half of the ℓ_1 norm of $\boldsymbol{\Delta}$ has to live on S; in symbols,

$$\|\boldsymbol{\Delta}_S\|_1 \ge \|\boldsymbol{\Delta}_{\overline{S}}\|_1,$$

where $\boldsymbol{\Delta}_S$ denotes the vector consisting of the components of $\boldsymbol{\Delta}$ indexed by S, and $\overline{S} = \{1, 2, \dots, n\} \setminus S$. Indeed, when $\boldsymbol{\Delta}$ is added to $\tilde{\mathbf{x}}$, its components outside S only *increase* the ℓ_1 norm, and since $\|\tilde{\mathbf{x}} + \boldsymbol{\Delta}\|_1 \le \|\tilde{\mathbf{x}}\|_1$, the components in S must at least compensate for this increase.

Since the restricted isometry property of A concerns the Euclidean norm, we will need to argue about the Euclidean norm of various pieces of $\boldsymbol{\Delta}$. For simpler notation, let us assume $\|\boldsymbol{\Delta}\|_1 = 1$ (as we will see, the argument is scale-invariant). Then, as we have just shown, $\|\boldsymbol{\Delta}_S\|_1 \ge \frac{1}{2}$ and thus $\|\boldsymbol{\Delta}_S\|_2 \ge \frac{1}{2\sqrt{r}}$ by the Cauchy–Schwarz inequality.

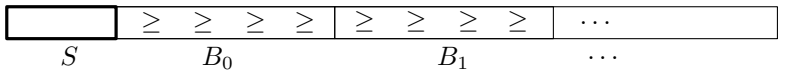
The first idea would be to use the restricted almost-isometry property to obtain $\|A_S\boldsymbol{\Delta}_S\|_2 \ge 0.9\frac{1}{2\sqrt{r}}$ (we use $\varepsilon_0 = 0.1$ for concreteness), and argue that the rest of the product, $A_{\overline{S}}\boldsymbol{\Delta}_{\overline{S}}$, is going to have smaller norm and thus $A\boldsymbol{\Delta} = A_S\boldsymbol{\Delta}_S + A_{\overline{S}}\boldsymbol{\Delta}_{\overline{S}}$ can't be $\mathbf{0}$. This doesn't quite work, because of the following "worst-case" scenario:



Here $\|\boldsymbol{\Delta}_{\overline{S}}\|_2$ is even much larger than $\|\boldsymbol{\Delta}_S\|_2$.

But this is not a problem: Since A has the $3r$-restricted almost-isometry property, as long as the bulk of the Euclidean norm is concentrated on at most $3r$ components, the argument will work.

So let $B_0 \subset \overline{S}$ consist of the indices of the $2r$ largest components of $\boldsymbol{\Delta}_{\overline{S}}$, B_1 are the indices of the next $2r$ largest components, and so on (the last block may be smaller).



We have $\|A_{S\cup B_0}\boldsymbol{\Delta}_{S\cup B_0}\|_2 \ge 0.9\|\boldsymbol{\Delta}_{S\cup B_0}\|_2 \ge 0.9\|\boldsymbol{\Delta}_S\|_2 \ge 0.9/2\sqrt{r} = 0.45/\sqrt{r}$. We want to show that

$$\sum_{j\ge 1} \|\boldsymbol{\Delta}_{B_j}\|_2 \le \frac{0.4}{\sqrt{r}}, \tag{2.3}$$

since then we can calculate, using restricted almost-isometry on $S \cup B_0$ and on each of $B_1, B_2, \dots$,

$$\|A\boldsymbol{\Delta}\|_2 \ge \|A_{S\cup B_0}\boldsymbol{\Delta}_{S\cup B_0}\|_2 - \sum_{j\ge 1} \|A_{B_j}\boldsymbol{\Delta}_{B_j}\|_2 \ge \frac{0.45}{\sqrt{r}} - 1.1\frac{0.4}{\sqrt{r}} > 0,$$

reaching the desired contradiction.

Proving (2.3) is a pure exercise in inequalities. We know that $\sum_{j\ge 0} \|\boldsymbol{\Delta}_{B_j}\|_1 = \|\boldsymbol{\Delta}_{\overline{S}}\|_1 \le \frac{1}{2}$. Moreover, by the choice of the blocks, the components belonging to B_j are no larger than the average of those in B_{j-1}, and thus

$$\|\boldsymbol{\Delta}_{B_j}\|_2 \le \left(2r \cdot \left(\frac{\|\boldsymbol{\Delta}_{B_{j-1}}\|_1}{2r}\right)^2\right)^{\frac{1}{2}} = \frac{\|\boldsymbol{\Delta}_{B_j}\|_1}{\sqrt{2r}}.$$

Summing over $j \ge 1$, we have

$$\sum_{j\ge 1} \|\boldsymbol{\Delta}_{B_j}\|_2 \le \frac{1}{\sqrt{2r}} \sum_{j\ge 0} \|\boldsymbol{\Delta}_{B_j}\|_1 \le \frac{1}{2\sqrt{2r}} < \frac{0.4}{\sqrt{r}},$$

which gives (2.3) and finishes the proof. □

2.9 Nearest neighbors in high dimensions

The topic of this section is only loosely related to the Johnson–Lindenstrauss lemma, but it can be regarded as an impressive instance of the "random projection" idea. It addresses an algorithmic problem very important in practice and highly interesting in theory: the *nearest neighbor problem*.

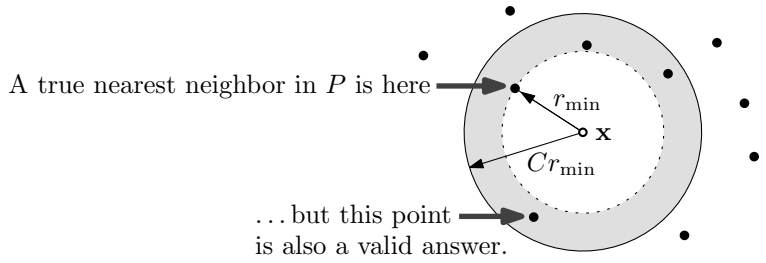
In this problem, we are given a set P of n points in some metric space; we will mostly consider the Euclidean space ℓ_2^k. Given a *query point* $\mathbf{x}$, belonging to the same metric space but typically not lying in P, we want to find a point $\mathbf{p} \in P$ that is the closest to $\mathbf{x}$.

A trivial solution is to compute the distance of $\mathbf{x}$ from every point of P and find the minimum. But since the number of queries is large, we would like to do better. To this end, we first *preprocess* the set P and store the results in a data structure, and then we use this data structure to answer the queries faster.

There are several efficient algorithms known for the case of $P \subset \ell_2^k$ with k small, say $k = 2, 3, 4$, but all of the known methods suffer from some kind of exponential dependence on the dimension: As k gets bigger, either the storage required for the data structure grows unrealistically large, or the improvement in the query time over the trivial solution becomes negligible (or both). This phenomenon is often called the *curse of dimensionality* and it is quite unpleasant, since in many applications the dimension is large, say from ten to many thousands.

Fortunately, it has been discovered that the curse of dimensionality can be broken if we are satisfied with approximate answers to the queries, rather than exact ones.

For a parameter $C \geq 1$, a *C-approximate algorithm* for the nearest neighbor problem is one that always returns a point $\mathbf{p} \in P$ at distance at most $Cr_{\min}$ from the query point $\mathbf{x}$, where $r_{\min}$ is the distance of $\mathbf{x}$ to a true nearest point.

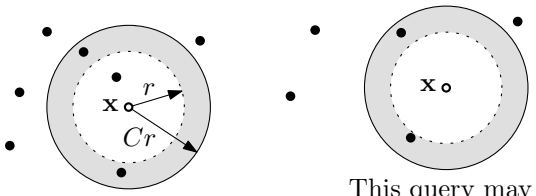


This is very much in the spirit of approximate metric embeddings, with C playing a role analogous to the distortion, and in some applications it even makes good sense. For example, if we search for a fingerprint left by John Doe, it seems reasonable to expect that John Doe's record

is going to be much closer to the query than any other record in the database.

Several C-approximate nearest-neighbor algorithms have been proposed. Here we present one of them, which is theoretically strong and elegant, and which also fares quite well in practical implementations (provided that some fine-tuning is applied, which is insignificant asymptotically but makes a big difference in real-world performance).

The r-near neighbor problem. Instead of considering the nearest neighbor problem directly, we describe an algorithm only for the following simpler problem, called the *C-approximate r-near neighbor problem*. We assume that together with the point set P, we are given a number $r > 0$ once and for all. Given a query point $\mathbf{x}$, the algorithm should return either a point $\mathbf{p} \in P$ at distance at most Cr to $\mathbf{x}$, or the answer NONE. However, NONE is a legal answer only if the distance of $\mathbf{x}$ to all points of P exceeds r.



This query must return a point in the larger disk.

This query may return NONE or a point in the larger disk.

It is known that an algorithm for the C-approximate *r-near* neighbor problem can be transformed to an algorithm for the $2C$-approximate *nearest* neighbor problem (where $2C$ can also be replaced by $(1+\eta)C$ for every fixed $\eta > 0$).

Such a transformation is very simple unless the ratio $\Delta := d_{\max}/d_{\min}$ is extremely large, where $d_{\max}$ and $d_{\min}$ denote the maximum and minimum distance among the points of P. The idea is to build several data structures for the C-approximate r-near neighbor problem for suitably chosen values of r, and to query all of them. For example, one can take $d_{\min}/2C$ as the smallest value of r and then keep doubling it until it exceeds $d_{\max}$. This incurs an extra factor of at most $O(\log \Delta)$ both in the storage and query time (here C is considered fixed). There are considerably more sophisticated methods avoiding the dependence on Δ, but these are mainly of theoretical interest, since in practice, we can almost always expect Δ to be reasonably small.

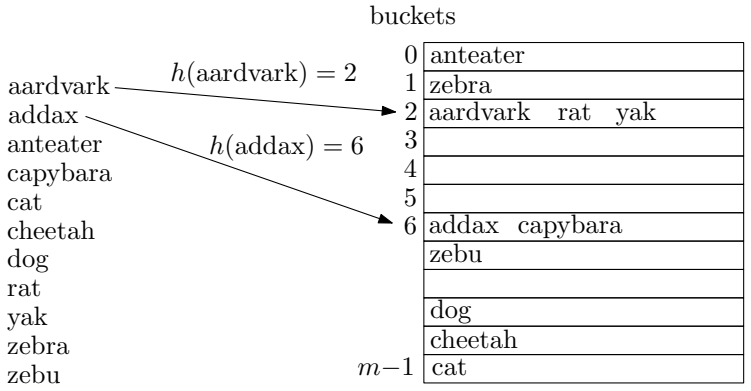
The probability of failure. The preprocessing phase of the algorithm is going to be randomized (while query answering is deterministic). Moreover, the algorithm is allowed to fail (i.e., say NONE even when it should really return a point) with some small probability δ. The failure probability is taken with respect to the internal random choices made during the preprocessing. That is, if we fix P and a query point $\mathbf{x}$, then the probability that the algorithm fails for that particular query $\mathbf{x}$ is at most δ.

In the algorithm given below, we will bound the failure probability by $\frac{3}{4}$. However, by building and querying t independent data structures instead of one, the failure probability can be reduced to $(\frac{3}{4})^t$, and thus we can make it as small as desired, while paying a reasonable price in storage and query time.

Hashing. The algorithm we will consider is based on *locality-sensitive hashing*. Before introducing that idea, we feel obliged to say a few words about hashing in general. This fundamental concept comes from data structures but it has also found many applications elsewhere.

Suppose that we want to store a dictionary of, say, 400,000 words in a computer in such a way that a given word can be looked up quickly. For solving this task by hashing, we allocate m *buckets* in the computer's memory, numbered 0 through $m-1$, where each bucket can store a list of words.[13] The value of m can be taken as a number somewhat larger than the number of words actually stored, say $m = 500{,}00$. Then we fix some *hash function* h, which is a function mapping every word to an integer in range from 0 to $m-1$, and we store each word w in the bucket number $h(w)$.

[13] An actual implementation of such buckets is not a completely trivial task. For readers interested in these matters, we sketch one of the possible solutions. We can represent the m buckets by an array of m pointers. The pointer corresponding to the ith bucket points to a linked list of the words belonging to that bucket. All of these linked lists are stored in a common auxiliary array. Many other schemes have been proposed, and the details of the implementation may make a large difference in practice.



Having stored all words of the dictionary in the appropriate buckets, it is very simple to look up an unknown word w: We search through the words stored in the bucket $h(w)$, and either we find w there, or we can claim that it doesn't occur in the dictionary at all.

If every word were assigned into a randomly chosen bucket, independent of the other words, then we would have less than one word per bucket on the average, and buckets containing more than ten words, say, would be very rare. Thus, in this hypothetical situation, the search for an unknown word is extremely fast in most cases.

However, it is not feasible to use a truly random hash function (how would one store it?). Reasonable hash functions resemble a random function in some respects, but they have a concise description and can be evaluated quickly.

Here is a concrete example. Let us represent words by 200-bit strings, say, which we then interpret as integers in the set $N = \{0, 1, \ldots, n-1\}$, where $n = 2^{200}$. We fix a prime $p \geq m$. The function $h\colon N \to \{0, 1, \ldots, m-1\}$ is defined using an integer parameter $a \in \{1, \ldots, p-1\}$, which we pick at random (and then keep fixed). For every word w, understood as an element of N, we set $h(w) := (aw \pmod p) \pmod m$. A theoretical analysis shows that for every fixed dictionary and a random, the words are very likely to be distributed quite uniformly among the buckets.

This was just one particular example of a good class of hash functions. The theory of hashing and hashing-based data structures is well developed, but we will not discuss it any further.

Locality-sensitive hashing. Hashing can be used for storing any kind of data items. In locality-sensitive hashing, we assume that the items are

points in a metric space, and we want that *two close points are more likely to be hashed to the same bucket than two faraway points*. This requirement goes somewhat against the spirit of hashing in general, since mapping the data items to the buckets in a "random-like" fashion is what allows a hash function to do its job properly. However, locality-sensitive hashing is used in a setting different from the one for the ordinary hashing.

We now introduce a formal definition. Let (X, d_X) be a metric space, whose points we want to hash, and let J be a set, whose elements we regard as indices of the buckets. In the algorithm below, we will have $J = \mathbb{Z}$ or $J = \mathbb{Z}^k$ for some integer k. We consider a family $\mathcal{H}$ of hash functions $h\colon X \to J$, together with a probability distribution on $\mathcal{H}$—in other words, we specify a way of choosing a random $h \in \mathcal{H}$.

Locality-sensitive hash family

A family $\mathcal{H}$ of hash functions from X to J is called **$(r, Cr, p_{\text{close}}, p_{\text{far}})$-sensitive**, where $r > 0$, $C \geq 1$, and $0 \leq p_{\text{far}} < p_{\text{close}} \leq 1$, if for every two points $x, y \in X$ we have

(i) If $d_X(x, y) \leq r$, then $\operatorname{Prob}[h(x) = h(y)] \geq p_{\text{close}}$.

(ii) If $d_X(x, y) > Cr$, then $\operatorname{Prob}[h(x) = h(y)] \leq p_{\text{far}}$.

In both cases, the probability is with respect to a random choice of $h \in \mathcal{H}$.

The algorithm. Now we present an algorithm for the C-approximate r-near neighbor in a metric space (X, d_X), assuming that a $(r, Cr, p_{\text{close}}, p_{\text{far}})$-sensitive hash family $\mathcal{H}$ is available for some constants $p_{\text{close}}, p_{\text{far}}$. A crucial quantity for the algorithm's performance is

$$\alpha := \frac{\ln p_{\text{close}}}{\ln p_{\text{far}}}$$

(note that both the numerator and the denominator are negative); we will achieve query time roughly $O(n^\alpha)$ and storage roughly $O(n^{1+\alpha})$.

First we need to "amplify" the gap between p_{close} and p_{far}. We define a new family $\mathcal{G}$, consisting of all t-tuples of hash functions from $\mathcal{H}$, where t is a parameter to be set later. That is,

$$\mathcal{G} = \Big\{ g = (h_1, \dots, h_t)\colon X \to J^t, h_1, \dots, h_t \in \mathcal{H} \Big\},$$

and for choosing $g \in \mathcal{G}$ at random, we pick $h_1, \dots, h_t \in \mathcal{H}$ randomly and independently. Then, clearly, $\mathcal{G}$ is $(r, Cr, p_{\text{close}}^t, p_{\text{far}}^t)$-sensitive.

In the preprocessing phase of the algorithm, we choose L random hash functions $g_1, \dots, g_L \in \mathcal{G}$, where L is another parameter to be determined in the future. For each $i = 1, 2, \dots, L$, we construct a hash table storing all elements of the point set P, where each $p \in P$ is stored in the bucket $g_i(p)$ of the ith hash table.

We note that the set J^t indexing the buckets in the hash tables may be very large or even infinite (in the instance of the algorithm for the Euclidean space we will have $J = \mathbb{Z}$). However, we can employ ordinary hashing, and further hash the bucket indices into a compact table of size $O(n)$. Thus, the total space occupied by the hash tables is $O(nL)$.[14]

To process a query x, we consider the points stored in the bucket $g_i(x)$ of the ith hash table, $i = 1, \dots, L$, one by one, for each of them we compute the distance to x, and we return the first point with distance at most Cr to x. If no such point is found in these buckets, we return the answer NONE. Moreover, if these L buckets together contain more than $3L$ points, we abort the search after examining $3L$ points unsuccessfully, and also return NONE.

Thus, for processing a query, we need at most kL evaluations of hash functions from the family $\mathcal{H}$ and at most $3L$ distance computations.

Here is a summary of the algorithm:

C-approximate r-near neighbor via locality-sensitive hashing

Preprocessing. For $i = 1, \dots, L$, choose $g_i = (h_{i,1}, \dots, h_{i,t}) \in \mathcal{G}$ at random and store the points of P in the ith hash table using the hash function g_i.

Query. Given x, search the points in the bucket $g_i(x)$ of the ith table, $i = 1, \dots, L$. Stop as soon as a point at distance at most Cr is found, or all points in the buckets have been exhausted, or $3L$ points have been searched.

Estimating the failure probability. The algorithm fails if it answers

[14] We need not store the values $g_i(p) \in J^t$ explicitly; we can immediately hash them further to the actual address in the compact hash table. The hash tables also don't store the points themselves, only indices. The theory of hashing guarantees that, when the compact hash tables are implemented suitably, the access to the points in the bucket $g_i(x)$ takes only $O(t)$ extra time; we refer to the literature about hashing for ways of doing this.

NONE even though there is a point $p_0 \in P$ with $d_X(x, p_0) \le r$. This may have two reasons:

(A) In none of the L hash tables, x was hashed to the same bucket as p_0.

(B) The L buckets searched by the algorithm contain more than $3L$ "far" points, i.e. points p with $d_X(x, p) > Cr$.

We want to set the parameters t and L so that the failure probability is bounded by a constant $\delta < 1$.

For points p_0 and x with $d_X(x, p_0) \le r$ we have $\mathrm{Prob}[g(x) = g(p_0)] \ge p_{\mathrm{close}}^t$, and so the probability of type (A) failure is at most $(1 - p_{\mathrm{close}}^t)^L$.

As for the type (B) failure, the probability that a far point q goes to the same bucket as x is at most p_{far}^t, and so the expected number of far points in the L searched buckets is no more than nLp_{far}^t. By Markov's inequality, the probability that we have more than $3L$ far points there is at most $np_{\mathrm{far}}^t/3$.

First we set t so that $np_{\mathrm{far}}^t/3 \le \frac{1}{3}$; this needs $t = (\ln n)/\ln(1/p_{\mathrm{far}})$ (ignoring integrality issues). Then, using $(1 - p_{\mathrm{close}}^t)^L \le e^{-p_{\mathrm{close}}^t L}$, we see that for $L = p_{\mathrm{close}}^{-t}$, the probability of type (A) failure is at most e^{-1}, and the total failure probability doesn't exceed $\frac{1}{3} + e^{-1} < \frac{3}{4}$. For the value of L this gives, again pretending that all numbers are integers,

$$L = p_{\mathrm{close}}^{-t} = e^{\ln(p_{\mathrm{close}})(\ln n)/\ln(p_{\mathrm{far}})} = n^{\alpha}$$

as claimed.

This finishes the analysis of the algorithm in the general setting. It remains to construct good locality-sensitive families of hash functions for metric spaces of interest.

Locality-sensitive hashing in Euclidean spaces. Now our metric space is ℓ_2^k. We may assume that $r = 1$, which saves us one parameter in the computations.

Let $\mathbf{Z} = (Z_1, \ldots, Z_k) \in \mathbb{R}^k$ be a normalized k-dimensional Gaussian random vector, and let $U \in [0, 1)$ be a uniformly distributed random variable (independent of $\mathbf{Z}$). Further let $w > 0$ be a real parameter. We define a family $\mathcal{H}_{\mathrm{Gauss}}$ of hash functions $\ell_2^k \to \mathbb{Z}$:

A random function $h \in \mathcal{H}_{\mathrm{Gauss}}$ is given by

$$h(\mathbf{x}) := \left\lfloor \frac{\langle \mathbf{Z}, \mathbf{x} \rangle}{w} + U \right\rfloor.$$

2.9.1 Lemma. *For every $C > 1$ and every $\varepsilon > 0$ there exists w such that the family $\mathcal{H}_{\mathrm{Gauss}}$ as above is $(1, C, p_{\mathrm{close}}, p_{\mathrm{far}})$-sensitive and $\alpha \le \frac{1}{C} + \varepsilon$, where $\alpha = \frac{\ln(1/p_{\mathrm{close}})}{\ln(1/p_{\mathrm{far}})}$ is as above.*

Proof. We will choose $w = w(C, \varepsilon)$ large, and then the probabilities p_{close} and p_{far} will be both close to 1. Thus, it is more convenient to estimate their complements.

In general, let $\mathbf{x}, \mathbf{y} \in \ell_2^k$ be points with distance s, where we consider s as a constant, and let

$$f(s) := \mathrm{Prob}[h(\mathbf{x}) \ne h(\mathbf{y})]$$

for h random. First we observe that for arbitrary real numbers a, b we have

$$\mathrm{Prob}\Big[\lfloor a + U \rfloor \ne \lfloor b + U \rfloor\Big] = \min(1, |a - b|).$$

Thus, for $\mathbf{Z}$ fixed to some $\mathbf{z}$, we have $\mathrm{Prob}[h(\mathbf{x}) \ne h(\mathbf{y}) \mid \mathbf{Z} = \mathbf{z}] = \min(1, |\frac{\langle \mathbf{z}, \mathbf{x} - \mathbf{y} \rangle}{w}|)$.

To compute $f(s)$, we need to average this over a random choice of $\mathbf{Z}$. By the 2-stability of the normal distribution, the difference $\langle \mathbf{Z}, \mathbf{x} \rangle - \langle \mathbf{Z}, \mathbf{y} \rangle = \langle \mathbf{Z}, \mathbf{x} - \mathbf{y} \rangle$ is distributed as sZ, where Z is (one-dimensional) standard normal. Hence

$$\begin{aligned} f(s) &= (2\pi)^{-1/2} \int_{-\infty}^{\infty} \min(1, |\frac{s}{w} t|) e^{-t^2/2} \, \mathrm{d}t \\ &= \mathrm{Prob}\big[\, |Z| \ge \tfrac{w}{s} \big] + \frac{s}{w} (2\pi)^{-1/2} \int_{-w/s}^{w/s} |t| e^{-t^2/2} \, \mathrm{d}t \\ &= \mathrm{Prob}\big[\, |Z| \ge \tfrac{w}{s} \big] + \frac{s}{w} \mathbf{E}[\,|Z|\,] - 2 \cdot \frac{s}{w} (2\pi)^{-1/2} \int_{w/s}^{\infty} t e^{-t^2/2} \, \mathrm{d}t. \end{aligned}$$

Both the first and third terms in the last expression decrease (faster than) exponentially as $w \to \infty$, and as we know from Section 2.5, $\mathbf{E}[\,|Z|\,] = \sqrt{2/\pi}$. Hence

$$f(s) = \sqrt{\tfrac{2}{\pi}} \cdot \frac{s}{w} + o(\tfrac{1}{w}).$$

Then, using $\ln(1 - t) = -t + o(t)$, we obtain

$$\alpha = \frac{\ln(1 - f(1))}{\ln(1 - f(C))} = \frac{(1 + o(1)) f(1)}{(1 + o(1)) f(C)} = \frac{1}{C} + o(1)$$

for C fixed and $w \to \infty$. Hence we can fix w so large that $\alpha \le \frac{1}{C} + \varepsilon$, which concludes the proof. □

With this lemma, we get that the general algorithm above can be used with α arbitrarily close to $1/C$. We need space $O(n^{1+\alpha})$ for storing the hash tables, $O(kn)$ space for the points themselves, and $O(ktL) = O(kn)$ space for representing the hash functions $g_1, \ldots, g_L$, so $O(kn + n^{1+\alpha})$ in total.

The cost of distance computation is $O(k)$, and evaluating a hash function g_i needs $O(tk) = O(k \log n)$ time. The total query time is $O(kn^{\alpha} \log n)$.

Remarks. If k is large, one may consider reducing the dimension by a random projection as in the Johnson–Lindenstrauss lemma. This increases the approximation factor C, but it allows us to replace k by $O(\log n)$. Of course, each query point has to be projected using the same matrix as for the original set P, and only then the nearest neighbor algorithm can be applied. Moreover, with some small probability, the random projection may distort the distance of the query point to the points of P by a large amount, and then the algorithm is likely to give a wrong answer.

In the proof of the last lemma, we have estimated the probabilities p_{close} and p_{far} asymptotically, using a large w. In practice, too a large w is not good, since then the parameter t in the algorithm also becomes large, which in turn makes the storage and query time worse. However, for a given C, one can find the value of w giving the best α numerically. It turns out that, first, the best w is not very large (below 5 for $C \le 5$, for example), and second, that the corresponding α is actually *strictly smaller* than $1/C$.

A more recent work even showed that the exponent α can be reduced further, arbitrarily close to $1/C^2$. We only sketch the idea.

The hash functions in $\mathcal{H}_{\text{Gauss}}$ described above can be geometrically interpreted as follows: To obtain $h(\mathbf{x})$, we randomly project $\mathbf{x}$ to $\mathbb{R}^1$, using independent $N(0,1)$ coefficients, then we translate the image by a random amount, and finally we round the result to the nearest integer. In the improved hashing strategy, we take a random Gaussian projection into $\mathbb{R}^d$ for a suitable $d = d(C)$, we translate it by a random vector, and finally, we round it to the nearest point in G, where G is a suitable discrete set in $\mathbb{R}^d$. Roughly speaking, we want G to be the set of centers of balls of a suitable radius w that form a "thin" covering of $\mathbb{R}^d$. We won't present any details here.

On the other hand, it is known that no locality-sensitive hashing family can achieve α below $0.462/C^2$.

The case of ℓ_1. Good locality-sensitive hash families are also known for several other important classes of metric spaces. Perhaps most significantly, for point sets in ℓ_1^k, the exponent α can be pushed as close to $1/C$ as desired (and, unlike in the Euclidean case, no further improvement is known).

The above analysis of the family $\mathcal{H}_{\text{Gauss}}$ relies mainly on the 2-stability of the normal distribution. Probably the conceptually simplest locality-sensitive hash family for ℓ_1^k uses a 1-stable distribution, namely, the **Cauchy distribution** with the density function

$$\frac{1}{\pi} \cdot \frac{1}{1+x^2}.$$

The 1-stability means that if $\mathbf{K} = (K_1, \ldots, K_k)$ is a vector of independent random variables with the Cauchy distribution, then for every $\mathbf{x} \in \mathbb{R}^k$, we have $\langle \mathbf{K}, \mathbf{x} \rangle \sim \|\mathbf{x}\|_1 K$, where K again has the Cauchy distribution. Defining a random hash function $h(\mathbf{x}) := \lfloor \frac{\langle \mathbf{K}, \mathbf{x} \rangle}{w} + U \rfloor$, one obtains a hash family $\mathcal{H}_{\text{Cauchy}}$, and an analysis similar to the proof of Lemma 2.9.1 shows that the resulting exponent α tends to $\frac{1}{C}$ as $w \to \infty$.

It is also interesting to note that while the Cauchy distribution is good enough for locality-sensitive hashing, it cannot be used for a "flattening lemma" in ℓ_1 (indeed, as we will see later, there is no flattening lemma possible for ℓ_1 even remotely comparable to the Johnson–Lindenstrauss lemma). The reason is that the Cauchy distribution is "heavy-tailed"; i.e., the probability of large deviations is quite significant, while for approximate preservation of many distances in the flattening lemma we would need a strong concentration.

2.10 Exercises

1. (Ball volume via the Gaussian distribution)

   (a) Calculate $I_n := \int_{\mathbb{R}^n} e^{-\|\mathbf{x}\|_2^2} \, \mathrm{d}\mathbf{x}$. (See Section 2.2.)

   (b)* Express I_n using $V_n = \mathrm{Vol}(B^n)$ and a suitable one-dimensional integral, where B^n denotes the n-dimensional Euclidean unit ball (consider the contribution to I_n of a very thin spherical shell). Compute the integral (set up a recurrence) and calculate V_n.

2.* Let $x, y \in S^{n-1}$ be two points chosen independently and uniformly at random. Estimate their expected (Euclidean) distance, assuming that n is large.

3.* Let $L \subseteq \mathbb{R}^n$ be a fixed k-dimensional linear subspace and let $\mathbf{x}$ be a random point of S^{n-1}. Estimate the expected distance of $\mathbf{x}$ from L, assuming that n is large.

4. (Lower bound for the flattening lemma)

(a) Consider the $n+1$ points $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n \in \mathbb{R}^n$ (where the $\mathbf{e}_i$ are the vectors of the standard orthonormal basis). Check that if these points with their Euclidean distances are $(1+\varepsilon)$-embedded into ℓ_2^k, then there exist unit vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^k$ with $|\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \le 100\varepsilon$ for all $i \ne j$ (the constant can be improved).

(b)** Let A be an $n \times n$ symmetric real matrix with $a_{ii} = 1$ for all i and $|a_{ij}| \le n^{-1/2}$ for all j, j, $i \ne j$. Prove that A has rank at least $\frac{n}{2}$.

(c)** Let A be an $n \times n$ real matrix of rank d, let k be a positive integer, and let B be the $n \times n$ matrix with $b_{ij} = a_{ij}^k$. Prove that the rank of B is at most $\binom{k+d}{k}$.

(d)* Using (a)–(c), prove that if the set as in (a) is $(1+\varepsilon)$-embedded into ℓ_2^k, where $100n^{-1/2} \le \varepsilon \le \frac{1}{2}$, then

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

5. Design a streaming algorithm that uses space $O(\log n)$ and solves the following problem. Let A be a set contining $n-1$ distinct numbers from $\{1, \ldots, n\}$. The algorithm reads a stream containing A in an arbitrary order and outputs the missing number $x \in \{1, \ldots, n\} \setminus A$.

6.* Extend the streaming algorithm for the ℓ_2 norm estimation from Section 2.6 to solve the *heavy hitters problem* in the following sense: After reading the stream, given an index $j \in \{1, 2, \ldots, n\}$, the algorithm outputs a number x_j^* such that with high probability, $x_j^* = x_j \pm \alpha \cdot \|\mathbf{x}\|_2$, where $\mathbf{x}$ is the current vector at the end of the stream. Here $\alpha > 0$ is a prescribed constant. A hint is given at the end of Section 2.6.

3

Lower bounds on the distortion

In this chapter we will consider lower bounds on the distortion, i.e., methods for showing that some metric space doesn't D-embed into another. We focus on embeddings of finite metric spaces into the spaces ℓ_p^k.

3.1 A volume argument and the Assouad dimension

The following problem is not hard, but solving it can help in getting used to the notion of distortion:

3.1.1 Problem. *Show that every embedding of the n-point equilateral space K_n (every two points have distance 1) into the Euclidean plane ℓ_2^2 has distortion at least $\Omega(\sqrt{n}\,)$.*

The usual solution is via a *volume argument*, very similar to the one for the existence of small δ-dense sets (Lemma 2.5.4). Assuming that $f\colon K_n \to \ell_2^2$ is non-contracting and doesn't expand any distance by more than D, we fix an arbitrary point $x_0 \in K_n$ and observe that the open $\frac{1}{2}$-balls centered at the images $f(x)$, $x \in K_n$, are all disjoint and contained in the ball of radius $D + \frac{1}{2}$ around $f(x_0)$. Comparing the areas gives $n(\frac{1}{2})^2 \le (D + \frac{1}{2})^2$, and $D = \Omega(\sqrt{n}\,)$ follows.

The same argument shows that K_n requires distortion $\Omega(n^{1/k})$ for embedding into ℓ_2^k (or any k-dimensional normed space, for that matter). For the equilateral space this is tight up to the multiplicative constant (right?), but there are worse n-point spaces, as we will see shortly. Before

getting to that, let us phrase the "volume argument" in a slightly different way, in which we encounter useful notions.

Dimensions defined by ball coverings. An important quantity for a metric space is, how many small balls are needed to cover a large ball.

> A metric space M is said to have **doubling dimension** at most k if, for every $r > 0$, every $2r$-ball in M can be covered by at most 2^k balls of radius r. A **doubling metric space** is one with a finite doubling dimension.

The notion of doubling dimension is good for rough calculations, where the precise value of the dimension doesn't matter. A slightly unaesthetic feature of it is that the doubling dimension of ℓ_2^k is hard to determine and certainly it doesn't equal k. Here is a more sophisticated notion:

> A metric space M is said to have **Assouad dimension** at most k if there exists $C > 0$ such that for every r, R with $0 < r < R$, every R-ball in M can be covered by at most $C(R/r)^k$ balls of radius r.

In other words, a metric space having Assouad dimension[1] at most k means that every R-ball has an r-dense subset of size $O((R/r)^k)$.

It is easily seen that ℓ_2^k has Assouad dimension k (one of the inequalities relies on Lemma 2.5.4, or rather, a variant of it for balls), and moreover, a metric space has a finite Assouad dimension iff it is doubling. On the other hand, Assouad dimension makes a good sense only for infinite spaces, because for a finite space the constant C swallows all information.

Now we re-phrase the argument for Problem 3.1.1 so that it avoids volume and works for embeddings of the equilateral space into any metric space M of Assouad dimension at most k. Indeed, suppose that K_n embeds into M with distortion at most D. Then the image forms, for some $r > 0$, an n-point r-separated set[2] S contained in a Dr-ball. This ball can be covered by $O((3D)^k)$ balls of radius $r/3$, each of them contains at most one point of S, and so $D = \Omega(n^{1/k})$.

[1] In the original source, Assouad called this the *metric dimension*, and this notion is also sometimes used in the literature (sometimes synonymously to the doubling dimension, though). But in graph theory the term "metric dimension" appears with a completely different meaning. Moreover, this Assouad dimension shouldn't be confused with the *Nagata–Assouad dimension*, which means something else.

[2] We recall that r-separated means that every two distinct points of the set have distance at least r.

Along similar lines, one can show that an embedding with a finite distortion cannot decrease the Assouad dimension.

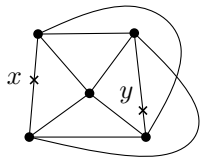
3.2 A topological argument and Lipschitz extensions

We have seen that the n-point equilateral space needs $\Omega(\sqrt{n}\,)$ distortion for embedding in the plane. Now we exhibit an n-point space requiring distortion $\Omega(n)$.

3.2.1 Proposition. *For all n there exists an n-point metric space (X, d_X) that doesn't cn-embed into the Euclidean plane ℓ_2^2, where $c > 0$ is a suitable constant.*

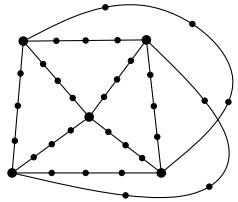
This result is asymptotically optimal, since it is known that every n-point metric space can be $O(n)$-embedded even in the real line (which we won't prove).

Proof. We begin by fixing a nonplanar graph G; for definiteness, let G be the complete graph K_5. Let $\overline{G}$ be the (infinite) metric space obtained from G by "filling" the edges; for each edge $e \in E(G)$, $\overline{G}$ contains a subspace s_e isometric to the interval $[0, 1]$, and these subspaces are glued together at the endpoints. The metric $d_{\overline{G}}$ on $\overline{G}$ is still the shortest-path metric. For example, for the points x and y in the picture,

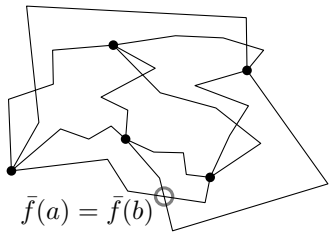


$d_{\overline{G}}(x, y) = 1.75$, assuming that x lies in the middle of its edge and y in a quarter (the drawing is a little misleading, since the intersection of the two edges drawn as arcs corresponds to two points lying quite far apart in $\overline{G}$).

Let us now choose X as a δ-dense subset of $\overline{G}$, with $\delta = O(1/n)$. For example, we may assume that $n = 10(m-1)+5$ for a natural number m, we put the vertices of the graph G into X, and on each of the 10 edges we choose $m-1$ equidistant points (so the spacing is $1/m$), as the drawing illustrates for $m = 4$:



We consider a mapping $f\colon X \to \ell_2^2$ with distortion D, and we want to bound D from below. We may assume that f is noncontracting and D-Lipschitz. We extend it to a mapping $\bar{f}\colon \overline{G} \to \ell_2^2$, by interpolating linearly on each of the short segments between two neighboring points in X. Then the image of $\bar{f}$ is a piecewise-linear drawing of the graph $G = K_5$, as in the next picture.



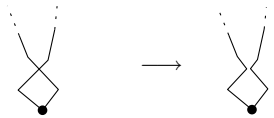
The core of the proof is the following claim.

3.2.2 Claim. *If $\bar{f}\colon \overline{G} \to \mathbb{R}^2$ is a piecewise-linear map as above, then there exist points $a, b \in \overline{G}$ whose distance in $\overline{G}$ is at least a positive constant, say $\frac{1}{2}$, and such that $\bar{f}(a) = \bar{f}(b)$.*

Proof of the claim. One possibility is to use the *Hanani–Tutte* theorem from graph theory, which asserts that *in every drawing of a nonplanar graph G in the plane, there are two non-adjacent edges (i.e. edges not sharing a vertex in G) that cross an odd number of times.* In particular, there exists a crossing of non-adjacent edges, and such a crossing yields the desired a and b, since non-adjacent edges have distance at least 1 in $\overline{G}$.

The Hanani–Tutte theorem is not an easy result, so we also sketch a more pedestrian proof, relying only on the nonplanarity of K_5. We proceed by contradiction: We assume that whenever $\bar{f}(a) = \bar{f}(b)$, we have $d_{\overline{G}}(a, b) < \frac{1}{2}$. This means

that in the corresponding piecewise-linear drawing of K_5, only edges sharing a vertex may cross. These (finitely many) crossings can be removed by transformations of the following kind:



Then we arrive at a planar drawing of K_5, which is a contradiction proving the claim.

We note that if we have an arbitrary (piecewise-linear) drawing of K_5 in which only adjacent edges may cross, it is not straightforward to remove the crossings by the above transformation. For example, in the following situation, trying to remove the circled crossing makes two non-adjacent edges cross:

However, such cases cannot occur in our setting, since by the assumption, if we follow an edge $e = \{u, v\}$ in the drawing from u to v, we first encounter all crossings with edges incident to u, and only then crossings with edges incident to v (this is where we use the assumption $d_{\overline{G}}(a, b) < \frac{1}{2}$).

Now the proof of Proposition 3.2.1 is finished quickly. Given $a, b \in \overline{G}$ as in the claim, we find $x, y \in X$ with $d_{\overline{G}}(x, a) \le \delta$ and $d_{\overline{G}}(y, b) \le \delta$, $\delta = O(1/n)$. Since f is D-Lipschitz, the extension $\bar{f}$ is D-Lipschitz as well (right?), and so

$$\begin{aligned} \|f(x) - f(y)\|_2 &\le \|\bar{f}(x) - \bar{f}(a)\|_2 + \|\bar{f}(y) - \bar{f}(b)\|_2 \\ &\le D(d_{\overline{G}}(x, a) + d_{\overline{G}}(y, b)) = O(D/n). \end{aligned}$$

On the other hand, since f is noncontracting, $\|f(x)-f(y)\|_2 \ge d_{\overline{G}}(x, y) \ge d_{\overline{G}}(a, b) - O(1/n) \ge \frac{1}{2} - O(1/n)$, and thus $D = \Omega(n)$ as claimed. □

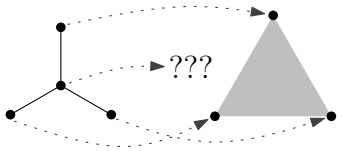
Lipschitz extendability. We will generalize the previous proof in order to exhibit, for every fixed k, n-point metric spaces badly embeddable in ℓ_2^k.

First we want to generalize the step where we extended the mapping f defined on X to a map $\bar{f}$ defined on the larger space $\overline{G}$; the important thing is to preserve the Lipschitz constant, or at least not to make it much worse.

Extendability of Lipschitz maps constitutes an extensive area with many results and techniques, and here we have a good occasion to mention two basic and generally useful results.

In a general setting, we consider two metric spaces (Y, d_Y) and (Z, d_Z) (the latter is usually taken as a normed space) and a subset $X \subseteq Y$, and we ask, what is the smallest C such that every 1-Lipschitz map $X \to Z$ can be extended to a C-Lipschitz map $Y \to Z$. (We mention in passing that the Johnson–Lindenstrauss lemma was discovered in connection with a problem of this type.[3])

We begin with a very simple example illustrating the nontriviality of the general problem. We consider the "tripod" graph with the shortest-path metric, and we map the leaves to the vertices of an equilateral triangle with side 2 in the Euclidean plane:



This is a 1-Lipschitz map, but there is no way of extending it to the central vertex while keeping it 1-Lipschitz.

However, if the target space has the ℓ_∞ norm, then all Lipschitz maps can be extended with no loss in the constant:

3.2.3 Proposition (Lipschitz extendability into ℓ_∞). *Let (Y, d_Y) be a metric space, let $X \subseteq Y$ be an arbitrary subset, and let $f\colon X \to \ell_\infty^k$ be a 1-Lipschitz map. Then there is a 1-Lipschitz map $\bar{f}\colon Y \to \ell_\infty^k$ extending f, i.e., with $\bar{f}(x) = f(x)$ for all $x \in X$.*

Proof. First we observe that it suffices to deal with maps into the real line, since a map into ℓ_∞^k is 1-Lipschitz if and only if each of its coordinates f_i is 1-Lipschitz.

[3] They proved that if Y is any metric space and $X \subseteq Y$ has n points, then every 1-Lipschitz $f\colon X \to \ell_2$ can be extended to an $O(\sqrt{\log n})$-Lipschitz $\bar{f}\colon Y \to \ell_2$. With the Johnson–Lindenstrauss lemma, Proposition 3.2.3 below, and the Kirszbraun theorem mentioned thereafter in our toolkit, the proof becomes a nice exercise.

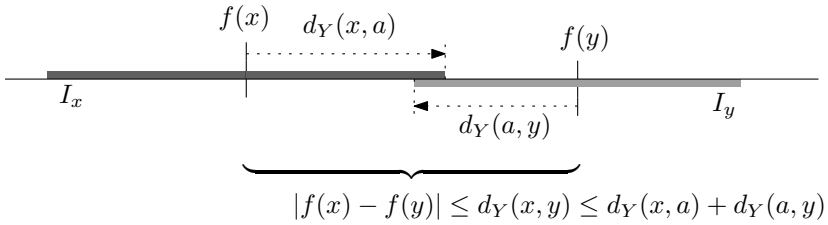
Next, we note that it is enough if we can extend the domain by a single point; in other words, to deal with the case $Y = X \cup \{a\}$. The general case then follows by Zorn's lemma.[4]

So the remaining task is, given a 1-Lipschitz (i.e. nonexpanding) $f\colon X \to \mathbb{R}$, find a suitable image $b \in \mathbb{R}$ for a so that the resulting extension $\bar{f}$ is 1-Lipschitz.

The 1-Lipschitz condition reads $|b - f(x)| \le d_Y(a,x)$ for all $x \in X$, which we rewrite to

$$b \in I_x := [f(x) - d_Y(a,x),\, f(x) + d_Y(a,x)].$$

A system of nonempty, closed and bounded intervals in the real line has a nonempty intersection if and only if every two of the intervals intersect (this, in addition to being easy to see, is the one-dimensional Helly theorem). So it suffices to check that $I_x \cap I_y \ne \emptyset$ for every $x, y \in X$, and this is immediate using the triangle inequality:



The proposition is proved. □

For the case where both the domain and the range are Euclidean (or Hilbert) spaces, one can use the following neat result:

3.2.4 Theorem (Kirszbraun's (or Kirszbraun–Valentine) theorem). *If Y is a Euclidean (or Hilbert) space, $X \subseteq Y$ is an arbitrary subset, and Z is also a Euclidean (or Hilbert) space, then every 1-Lipschitz map $X \to Z$ extends to a 1-Lipschitz map $Y \to Z$.*

The general scheme of the proof is the same as that for Proposition 3.2.3, but one needs a nontrivial lemma about intersections of balls in a Hilbert space, which we omit here.

[4]Readers who don't know Zorn's lemma or don't like it may want to consider only separable metric spaces, i.e., assume that there is a countable set $S = \{a_0, a_1, a_2 \ldots\} \subseteq Y$ whose closure is Y. We first extend on $X \cup S$ by induction, and then we extend to the closure in the obvious way, noticing that the Lipschitz constant is preserved.

With a little help from topology. Now we need a space analogous to the "K_5 with filled edges" $\overline{G}$ in Proposition 3.2.1. It should be nonembeddable in $\mathbb{R}^k$ (and moreover, every attempted embedding should fail by identifying two faraway points), and at the same time, it should be "low-dimensional", in the sense of possessing small δ-dense sets.

The best what topology has to offer here are k-dimensional spaces nonembeddable in $\mathbb{R}^{2m}$ (while every "sufficiently reasonable" m-dimensional space is known to embed in $\mathbb{R}^{2m+1}$). Using such spaces, we will prove the following analog of the claim in the proof of Proposition 3.2.1.

3.2.5 Lemma. *For every integer $m \ge 1$ there exists a metric space (Y, d_Y) with the following properties. For some constants $R, C_0, \beta > 0$ we have:*

(i) *(Bounded and "m-dimensional") The diameter of (Y, d_Y) is at most R, and for every $\delta > 0$ there exists a δ-dense subset $X \subseteq Y$ with $|X| \le C_0 \delta^{-m}$.*

(ii) *(Every continuous map in $\mathbb{R}^{2m}$ identifies two distant points) For every continuous map $f\colon Y \to \mathbb{R}^{2m}$, there exist points $a, b \in Y$ with $d_Y(a,b) \ge \beta$ and $f(a) = f(b)$.*

Assuming this lemma, the following generalization of Proposition 3.2.1 is easy.

3.2.6 Theorem. *For every $k \ge 1$ there exists $c_k > 0$ such that for every n, one can construct an n-point metric space requiring distortion at least $c_k n^{1/\lceil k/2 \rceil}$ for embedding in ℓ_2^k.*

Proof. It suffices to consider $k = 2m$. We fix (Y, d_Y) as in Lemma 3.2.5 and we choose an n-point δ-dense set $X \subseteq Y$, $\delta = O(1/n^{1/m})$.

Since we have prepared Lipschitz extendability for maps into ℓ_∞^k, it will be more convenient to work with ℓ_∞^k as the target space. Since the Euclidean and ℓ_∞ norm differ by a factor of at most $\sqrt{k}$ (i.e., $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \sqrt{k}\|\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \mathbb{R}^k$), this influences only the constant factor c_k.

So we consider a noncontracting and D-Lipschitz map $f\colon X \to \ell_\infty^k$, and we extend it to a D-Lipschitz map $\bar{f}$ defined on all of Y. By part (ii) of Lemma 3.2.5, there are points $a, b \in Y$ with distance at least β and such that $\bar{f}(a) = \bar{f}(b)$. The rest of the proof follows the argument for Proposition 3.2.1 almost literally and we omit it. □

Interestingly, the theorem is tight for $k \le 2$ (as we have already mentioned) and tight up to a logarithmic factor for all *even* k, since every n-point metric space can be embedded in ℓ_2^k with distortion $O(n^{2/k}(\log n)^{3/2})$.[5]

On the other hand, for $k = 3, 5, 7, \dots$ there is a more significant gap; e.g., for $k = 3$ the lower bound is $\Omega(n^{1/2})$, while the upper bound is only roughly $n^{2/3}$.

The Van Kampen–Flores complexes. First we describe a "standard" example of spaces for Lemma 3.2.5. For this we need the vocabulary of simplicial complexes (readers not familiar with it may skip this part or look up the relevant terms).

In the 1930s Van Kampen and Flores constructed, for every $m \ge 1$, an m-dimensional finite simplicial complex that doesn't embed in $\mathbb{R}^{2m}$, namely, the m-skeleton of the $(2m+2)$-dimensional simplex. For $m = 1$, this is the 1-skeleton of the 4-dimensional simplex, which is exactly the "K_5 with filled edges".

Metrically, we can consider the Van Kampen–Flores complex as a metric subspace Y of the regular $(2m+2)$-dimensional Euclidean simplex. Then property (i) in Lemma 3.2.5 is almost obvious, since the regular m-dimensional Euclidean simplex has δ-dense subsets of size $O(\delta^{-m})$ (arguing as in Lemma 2.5.4), and so has a finite union of such simplices.

The nonembeddability in part (ii) of Lemma 3.2.5 follows from a slightly stronger form of the Van Kampen–Flores result: For every continuous mapping $f\colon Y \to \mathbb{R}^{2m}$, there are points a, b belonging to *disjoint faces* of Y with $f(a) = f(b)$ (this is what the proofs actually give). For example, for the case $m = 1$, this asserts that in any drawing of K_5 in the plane, some two non-adjacent edges cross, which was the consequence of the Hanani–Tutte theorem we needed in Proposition 3.2.1.

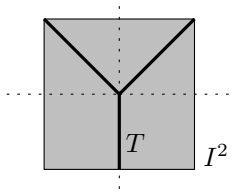
The proof of this result (for all m) is not exactly hard, but it requires some topological apparatus, and we won't treat it here.

A geometric proof from the Borsuk–Ulam theorem. We describe an alternative approach to Lemma 3.2.5, where all that is needed from topology is encapsulated in the following famous result.

[5] Sketch of proof: For embedding an n-point ℓ_2 metric in ℓ_2^k, one uses a random projection as in the Johnson–Lindenstrauss lemma—only the calculation is somewhat different; the resulting distortion is $O(n^{2/k}\sqrt{\log n})$. An arbitrary n-point metric space is first embedded in ℓ_2 with distortion at most $O(\log n)$, which relies on a theorem of Bourgain, to be discussed later.

3.2.7 Theorem (Borsuk–Ulam theorem). *For every continuous map $f\colon S^n \to \mathbb{R}^n$, where $S^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \|\mathbf{x}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^{n+1}$, there exists a point $\mathbf{x} \in S^n$ with $f(\mathbf{x}) = f(-\mathbf{x})$.*

First we construct a weaker example than needed for Lemma 3.2.5; namely, with the target dimension $2m$ in part (ii) replaced by $2m-1$. Let $I^2 = [-1,1]^2$ be the unit square in the (Euclidean) plane, and let T be the "tripod" as in the picture:

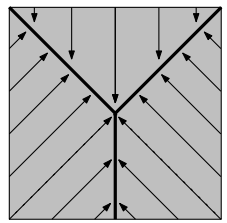


The weaker example is the space $K := T^m \subset \mathbb{R}^{2m}$, the Cartesian product of m copies of T. Since K is a union of finitely many (namely, 3^m) cubes, it is clear that it admits δ-dense subsets of size $O(\delta^{-m})$, so it remains to deal with the non-embeddability as in part (ii) of Lemma 3.2.5.

3.2.8 Lemma. *For every continuous map $f\colon T^m \to \mathbb{R}^{2m-1}$ there are points $\mathbf{a}, \mathbf{b} \in T^m$ with $f(\mathbf{a}) = f(\mathbf{b})$ and such that $\|\mathbf{a} - \mathbf{b}\|_2 \ge \beta$, where $\beta > 0$ is a universal constant.*

Proof. We plan to use the Borsuk–Ulam theorem but with S^n replaced with ∂I^{n+1}, the boundary of the $(n+1)$-dimensional cube. Using the central projection $S^n \to \partial I^{n+1}$, it is clear that also for every continuous map $g\colon \partial I^{n+1} \to \mathbb{R}^n$ there is an $\mathbf{x} \in \partial I^{n+1}$ with $g(\mathbf{x}) = g(-\mathbf{x})$.

To apply this result, we first we introduce a (piecewise-linear) mapping $\pi\colon I^2 \to T$ that "squashes" the square onto the tripod, as indicated in the picture:



Besides the continuity of π, we will use the following easily verified property: Whenever $\mathbf{x} \in \partial I^2$ is a point on the perimeter of the square, we have $\|\pi(\mathbf{x}) - \pi(-\mathbf{x})\|_2 \geq \beta$, for a suitable positive constant β.

Given a map $f\colon T^m \to \mathbb{R}^{2m-1}$ as in the lemma, we want to define a map $g\colon \partial I^{2m} \to \mathbb{R}^{2m-1}$. To this end, we represent a point $\mathbf{x} \in I^{2m}$ as an m-tuple $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in I^2$, and we note that $\mathbf{x} \in \partial I^{2m}$ exactly if $\mathbf{x}_i \in \partial I^2$ for at least one i (since both of these are equivalent to $\mathbf{x}$ having at least one ± 1 coordinate).

Now we set $g := f \circ \pi^m$; in other words, a point $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is first mapped to $\pi^m(\mathbf{x}) = (\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_m)) \in T^m$, and then to $f(\pi^m(\mathbf{x})) \in \mathbb{R}^{2m-1}$. The Borsuk–Ulam theorem gives us an $\mathbf{x} \in \partial I^{2m}$ with $g(\mathbf{x}) = g(-\mathbf{x})$; we claim that $\mathbf{a} := \pi^m(\mathbf{x})$ and $\mathbf{b} := \pi^m(-\mathbf{x})$ are the desired points. But this is obvious, since there is some i with $\mathbf{x}_i \in \partial I^2$, and then $\|\mathbf{a} - \mathbf{b}\|_2 \geq \|\pi(\mathbf{x}_i) - \pi(-\mathbf{x}_i)\|_2 \geq \beta$. □

The cone trick. Let us recall that for a set $B \subseteq \mathbb{R}^k$ and a point $\mathbf{p} \in \mathbb{R}^k$, the **cone** over B with apex $\mathbf{p}$ is defined as the union of all segments $\mathbf{xp}$ with $\mathbf{x} \in B$.

To finally prove Lemma 3.2.5, we define an m-dimensional subspace $Y \subset K = T^{m+1}$ as $Y := T^{m+1} \cap \partial I^{2m+2}$. In other words, Y consists of all points $(\mathbf{x}_1, \ldots, \mathbf{x}_{m+1}) \in T^{m+1}$ where at least one of the $\mathbf{x}_i$ is on the boundary of the square, i.e., is one of the "tips" of T. (A diligent reader may check that for $m = 1$, Y is homeomorphic to the complete bipartite graph $K_{3,3}$ with filled edges.)

It is easy to check that K is the cone over Y with the apex $\mathbf{0}$, and moreover, every $\mathbf{x} \in K \setminus \{\mathbf{0}\}$ is contained in a unique segment $\mathbf{y0}$ with $\mathbf{y} \in Y$. (This is because T is such a cone, and this property is preserved under Cartesian products.)

Let us consider a continuous map $f\colon Y \to \mathbb{R}^{2m}$. We define a map $\tilde{f}\colon K \to \mathbb{R}^{2m+1}$ in the natural way: A point $\mathbf{y}$ of the base Y is sent to $(f(\mathbf{y}), 0)$ (we append 0 as the last coordinate), the apex $\mathbf{0}$ is mapped to $(0, 0, \ldots, 0, 1)$, and for an arbitrary $\mathbf{x} \in K$ we interpolate linearly. Explicitly, we write $\mathbf{x} = t\mathbf{y} + (1-t)\mathbf{0} = t\mathbf{y}$ for $\mathbf{y} \in Y$, and we set $\tilde{f}(\mathbf{x}) := (tf(\mathbf{y}), 1-t)$ (this map is clearly continuous).

By Lemma 3.2.8, there are points $\tilde{\mathbf{a}}, \tilde{\mathbf{b}} \in K$ with $\tilde{f}(\tilde{\mathbf{a}}) = \tilde{f}(\tilde{\mathbf{b}})$ and $\|\tilde{\mathbf{a}} - \tilde{\mathbf{b}}\|_2 \geq \beta$. We have $\tilde{\mathbf{a}} = t_1\mathbf{a}$ and $\tilde{\mathbf{b}} = t_2\mathbf{b}$, and $\tilde{f}(\tilde{\mathbf{a}}) = \tilde{f}(\tilde{\mathbf{b}})$ means that $t_1 = t_2$ and $f(\mathbf{a}) = f(\mathbf{b})$ (note that neither t_1 nor t_2 can be 0, and so $\mathbf{a}, \mathbf{b}$ are determined uniquely). Then $\beta \leq \|\tilde{\mathbf{a}} - \tilde{\mathbf{b}}\|_2 = t_1\|\mathbf{a} - \mathbf{b}\|_2$, and Lemma 3.2.5 is proved. □

3.3 Distortion versus dimension: A counting argument

We know that every n-point metric space embeds isometrically in ℓ_∞^n. Can the dimension n be reduced substantially, especially if we allow for embeddings with some small distortion D?

The answer depends on the precise value of D in a surprising way, and it connects metric embeddings with a tantalizing problem in graph theory. Here we consider a lower bound, i.e., showing that the dimension can't be too small for a given D.

The counting argument for the lower bound goes roughly as follows. Suppose that all n-point metric spaces can be D-embedded in some k-dimensional normed space Z (where $Z = \ell_\infty^k$ is our main example, but the same argument works for every Z). We will exhibit a class $\mathcal{M}$ of many "essentially different" n-point metric spaces, and we will argue that Z doesn't allow for sufficiently many "essentially different" placements of n points corresponding to D-embeddings of all the spaces from $\mathcal{M}$.

First let us see a concrete instance of this argument, showing that it is not possible to embed all n-point metric spaces into a fixed normed space of a sublinear dimension with distortion below 3.

3.3.1 Proposition. *For every $D < 3$ there exists $c_D > 0$ such that if Z is a k-dimensional normed space in which all n-point metric spaces embed with distortion at most D, then $k \geq c_D n$.*

The assumption $D < 3$ turns out to be sharp: As we will see later, all n-point metric spaces can be 3-embedded in ℓ_∞^k with k only about $\sqrt{n}$.

Proof. Let us consider n even, and let $G = K_{n/2,n/2}$ be the complete bipartite graph on the vertex set $V := \{1, 2, \ldots, n\}$. Let $m = |E(G)| = (n/2)^2$ be the number of edges of G.

Let $\mathcal{H}$ be the set of all graphs H of the form (V, E') with $E' \subseteq E(G)$, i.e., subgraphs of G with the same vertex set as G. We have $|\mathcal{H}| = 2^m$.

For every $H \in \mathcal{H}$, let d_H denote the shortest-path metric of H. We define a new metric $\overline{d}_H$ by truncating d_H; namely, for $u, v \in V$ we set

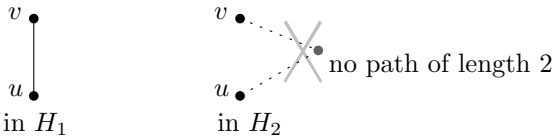
$$\overline{d}_H(u, v) := \min(d_H(u, v), 3).$$

It is easily seen that this is indeed a metric. (Strictly speaking, d_H need not always be a metric, since H may be disconnected and then some pairs of vertices have infinite distance, but the truncation takes care of this.)

These $\overline{d}_H$ define the "large" class $\mathcal{M}$ of metric spaces, namely, $\mathcal{M} := \{(V, \overline{d}_H) : H \in \mathcal{H}\}$. We note that every two spaces in $\mathcal{M}$ are "essentially different" in the following sense:

Claim. *For every two distinct $H_1, H_2 \in \mathcal{H}$ there exists $u, v \in V$ such that either $\overline{d}_{H_1}(u,v) = 1$ and $\overline{d}_{H_2}(u,v) = 3$, or $\overline{d}_{H_2}(u,v) = 1$ and $\overline{d}_{H_1}(u,v) = 3$.*

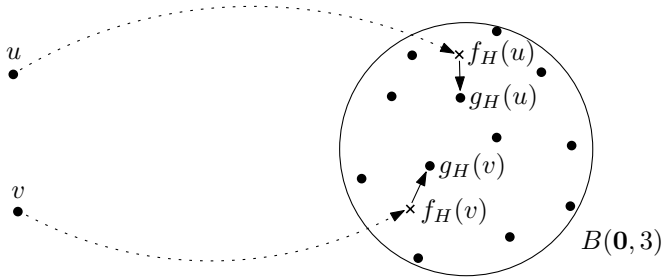
Proof of the claim. Since $E(H_1) \neq E(H_2)$, there is a edge $\{u,v\} \in E(G)$ belonging to one of H_1, H_2 but not to the other. Let us say that $\{u,v\} \in E(H_1)$. Then $\overline{d}_{H_1}(u,v) = 1$. Since $\{u,v\} \notin E(H_2)$, $\overline{d}_{H_2}(u,v)$ can't be 1. It can't be 2 either, since the path of length 2 connecting u and v in H_2 together with the edge $\{u,v\}$ would form a triangle in $G = K_{n/2,n/2}$, but G is bipartite and thus it has no triangles.



□

To prove Proposition 3.3.1, we suppose that for every $H \in \mathcal{H}$ there exists a D-embedding $f_H : (V, \overline{d}_H) \to Z$. By rescaling, we make sure that $\frac{1}{D}\,\overline{d}_H(u,v) \le \|f_H(u) - f_H(v)\|_Z \le \overline{d}_H(u,v)$ for all $u, v \in V$ (where $\|.\|_Z$ denotes the norm in Z). We may also assume that the image of f_H is contained in the ball $B_Z(\mathbf{0}, 3) = \{x \in Z : \|x\|_Z \le 3\}$.

We will now "discretize" the mappings f_H. Let us choose a small δ-dense set N in $B_Z(\mathbf{0}, 3)$, where δ will be fixed soon. As we know (Lemma 2.5.4), we may assume $|N| \le (\frac{4}{\delta})^k$. For every $H \in \mathcal{H}$, we define a new mapping $g_H : V \to N$ by letting $g_H(v)$ be the nearest point of N to $f_H(v)$ (ties resolved arbitrarily).



Next, we want to prove that two different subgraphs $H_1, H_2 \in \mathcal{H}$ give rise to different maps g_{H_1} and g_{H_2}. Let u, v be two vertices, as in the claim above, on which $\overline{d}_{H_1}$ and $\overline{d}_{H_2}$ differ; say $\overline{d}_{H_1}(u,v) = 1$ and $\overline{d}_{H_2}(u,v) = 3$.

We have, on the one hand,

$$\|g_{H_1}(u) - g_{H_1}(v)\|_Z \le \|f_{H_1}(u) - f_{H_1}(v)\|_Z + 2\delta \le \overline{d}_{H_1}(u,v) + 2\delta = 1 + 2\delta,$$

and on the other hand,

$$\|g_{H_2}(u) - g_{H_2}(v)\|_Z \ge \|f_{H_2}(u) - f_{H_2}(v)\|_Z - 2\delta \ge \frac{1}{D}\overline{d}_{H_2}(u,v) - 2\delta = \frac{3}{D} - 2\delta.$$

So for $\delta < \frac{1}{4}(\frac{3}{D} - 1)$, we arrive at $\|g_{H_1}(u) - g_{H_1}(v)\|_Z < \|g_{H_2}(u) - g_{H_2}(v)\|_Z$.

This shows that $g_{H_1} \neq g_{H_2}$. Hence the number of distinct mappings $g : V \to N$ cannot be smaller than the number of spaces in $\mathcal{M}$. This gives the inequality

$$|N|^n \ge |\mathcal{M}|.$$

Using $|N| \le (4/\delta)^k$, with $\delta > 0$ depending only on D, and $|\mathcal{M}| = 2^m$, $m = |E(G)| = (n/2)^2$, we obtain $k \ge c_D n$ as claimed. As a function of D, the quantity c_D is of order

$$\frac{1}{\log \frac{1}{3-D}}.$$

□

Graphs without short cycles. The important properties of the graph $G = K_{n/2,n/2}$ in the previous proof were that

- it has a large number of edges, and
- it contains no cycles of length 3.

If we start with a graph G containing no cycles of length at most ℓ, the same kind of calculations leads to the following result.

3.3.2 Proposition. *Let G be a graph with n vertices and m edges that contains no cycles of length at most ℓ. Then for every $D < \ell$ any normed space Z that admits a D-embedding of all n-point metric spaces, we have*

$$\dim Z \ge \frac{cm}{n}, \tag{3.1}$$

where $c = c(D, \ell) > 0$ depends only on D and ℓ and it can be taken as $1/\log_2(16\ell/(1 - \frac{D}{\ell}))$. □

But what is the largest number $m = m(\ell, n)$ of edges in a graph with n vertices and no cycle of length at most ℓ? This is the tantalizing graph-theoretic problem mentioned at the beginning of the section, and here is a short overview of what is known.

First, we note that it's easy to get rid of odd cycles. This is because every graph G has a bipartite subgraph H that contains at least half of the edges of G,[6] and so $m(2t+1, n) \geq \frac{1}{2} m(2t, n)$. Thus, neglecting a factor of at most 2, we can consider $m(\ell, n)$ only for odd integers ℓ.

Essentially the best known upper bound is

$$m(\ell, n) \leq n^{1+1/\lfloor \ell/2 \rfloor} + n. \tag{3.2}$$

We won't use it; we mention it only for completing the picture, and so we omit the proof, although it's quite simple.

This upper bound is known to be asymptotically tight in some cases, but only for several values of ℓ; namely, for $\ell = 3, 5, 7, 11$ (and thus also for $\ell = 2, 4, 6, 10$, but as we said above, it suffices to consider odd ℓ).

We have already considered the case $\ell = 3$; the graph witnessing $m(3, n) = \Omega(n^2)$ is very simple, namely, $K_{n/2, n/2}$. The construction for the next case $\ell = 5$, where $m(5, n)$ is of order $n^{3/2}$ is based on *finite projective planes*:[7] The appropriate graph has the points and the lines of a finite projective plane as vertices, and edges correspond to membership, i.e., each line is connected to all of its points. This is a bipartite graph, and it has no 4-cycle, because a 4-cycle would correspond to two distinct lines having two distinct points in common. If we start with a projective plane of order q, the resulting graph has $n = 2(q^2 + q + 1)$ vertices and $(q^2 + q + 1)(q + 1)$ edges, which is of order $n^{3/2}$.

The constructions for $\ell = 7$ and $\ell = 11$ are algebraic as well but more complicated. As an illustration, we present a construction for $\ell = 7$, but we won't verify that it actually works. The vertices are again points and lines, and edges correspond to membership, but this time we consider lines and points in the 4-dimensional projective space (over a finite field $\mathrm{GF}(q)$), and moreover, only the points and lines contained in the quadratic surface

$$Q = \{(x_0 : x_1 : x_2 : x_3 : x_4) : x_0^2 + x_1 x_2 + x_3 x_4 = 0\}$$

(here $(x_0 : \cdots : x_4)$ denotes a point of the projective 4-space, i.e., an equivalence class consisting of all 5-tuples $(\lambda x_0, \lambda x_1, \ldots, \lambda x_4)$, $\lambda \in \mathrm{GF}(q)$, $\lambda \neq 0$). There are no 4-cycles for the same reason as in the previous construction, and the absence of 6-cycles corresponds to the nonexistence of three lines spanning a triangle in the quadric surface.

Together with $\ell = 11$, we have exhausted all known cases where $m(\ell, n)$ is of order $n^{1+1/\lfloor \ell/2 \rfloor}$ as in the upper bound (3.2). For ℓ large, the best known constructions give about $n^{1+3/4\ell}$. A simple probabilistic construction (take a random graph with an appropriate edge probability and delete all edges in short cycles) gives a still weaker bound $m(\ell, n) \geq c_0 n^{1+1/(\ell-2)}$ for all odd ℓ, with some constant $c_0 > 0$ independent of ℓ.

Together with Proposition 3.3.2, the constructions for $\ell = 3, 5, 7, 11$ mentioned above yield the following: If Z is a normed space in which all n-point metric spaces embed with distortion at most D, then

- $\dim Z = \Omega(n)$ for $D < 3$,
- $\dim Z = \Omega(n^{1/2})$ for $D < 5$,
- $\dim Z = \Omega(n^{1/3})$ for $D < 7$, and
- $\dim Z = \Omega(n^{1/5})$ for $D < 11$

(here D is considered fixed and the implicit constant in $\Omega(.)$ depends on it).

We also obtain the following lower bound for embedding into Euclidean spaces, without any restriction on dimension.

3.3.3 Proposition (Lower bound on embedding in ℓ_2). *For all n, there exist n-point metric spaces whose embedding into ℓ_2 (i.e., into any Euclidean space) requires distortion at least $\Omega(\log n / \log\log n)$.*

Proof. If an n-point metric space is D-embedded into ℓ_2^n, then by the Johnson–Lindenstrauss lemma it can be $(2D)$-embedded into ℓ_2^k with $k \leq C \log n$ for some specific constant C. But a calculation using Proposition 3.3.2 shows that this dimension is too low unless the distortion is as large as claimed.

[6] To see this, divide the vertices of G into two classes A and B arbitrarily, and while there is a vertex in one of the classes having more neighbors in its class than in the other class, move such a vertex to the other class; the number of edges between A and B increases in each step. For another proof, assign each vertex randomly to A or B and check that the expected number of edges between A and B is $\frac{1}{2}|E(G)|$.

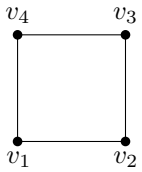
[7] We recall that a finite projective plane is a pair $(X, \mathcal{L})$, where X is a set, whose elements are called *points*, and $\mathcal{L}$ is a family of subsets of X, whose sets are called *lines*. Every two points $x, y \in X$ are contained in exactly one common line, every two lines intersect in exactly one point, and for some integer q, called the *order* of the projective plane, we have $|X| = |\mathcal{L}| = q^2 + q + 1$, and $|L| = q + 1$ for every $L \in \mathcal{L}$. For every prime power q, a projective plane of order q can be constructed algebraically from the q-element finite field.

In more detail, let us set $\ell = \lfloor 4D \rfloor$. Then, as was mentioned above, we have $m(\ell, n) \geq c_0 n^{1+1/(\ell-2)} \geq c_0 n^{1+1/4D}$, and (3.1) gives $k \geq c_{2D,\ell} m(\ell, n)/n = \Omega(n^{1/4D}/\log D)$. The resulting inequality $n^{1/4D}/\log D = O(\log n)$ then yields $D = \Omega(\log n / \log\log n)$. □

3.4 Nonembeddability of the ℓ_1 cube in ℓ_2

We will start proving lower bounds for the distortion using inequalities valid in the target space. In this section we demonstrate the approach in a simple case, concerning embeddings in Euclidean spaces. The method also applies to embeddings in other ℓ_p spaces and even in more general classes of normed spaces.

The 4-cycle. We begin with a rather small example, with only four points, where the metric is the graph metric of the 4-cycle:



3.4.1 Proposition. *The metric of the 4-cycle can be embedded in ℓ_2 with distortion $\sqrt{2}$, but not smaller.*

An embedding attaining distortion $\sqrt{2}$ is the obvious one, which goes into the Euclidean plane and is given by $v_1 \mapsto (0,0)$, $v_2 \mapsto (1,0)$, $v_3 \mapsto (1,1)$, and $v_4 \mapsto (0,1)$. (We note that if the image is considered with the ℓ_1 metric, rather than with the Euclidean one, then we have an *isometric* embedding of the four-cycle into ℓ_1^2.)

It remains to show that $\sqrt{2}$ is the smallest possible distortion. As we will see, any embedding that doesn't expand the length of the edges has to shorten at least one of the diagonals by at least $\sqrt{2}$.

Let us consider arbitrary four points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ in some Euclidean space; we think of $\mathbf{x}_i$ as the image of v_i under some mapping of the vertices of the 4-cycle. Let us call the pairs $\{\mathbf{x}_1, \mathbf{x}_2\}$, $\{\mathbf{x}_2, \mathbf{x}_3\}$, $\{\mathbf{x}_3, \mathbf{x}_4\}$, $\{\mathbf{x}_4, \mathbf{x}_1\}$ the *edges* of the considered 4-point configuration, while $\{\mathbf{x}_1, \mathbf{x}_3\}$ and $\{\mathbf{x}_2, \mathbf{x}_4\}$ are the *diagonals* (we note that the configuration $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ need not form a quadrilateral—the points need not lie in a common plane, for example).

The next lemma claims that for any 4-point Euclidean configuration, the sum of the squared lengths of the diagonals is never larger than the sum of the squared lengths of the edges:

3.4.2 Lemma (Short diagonals lemma). *For every choice of points $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ in a Euclidean space, we have*

$$\begin{aligned}\|\mathbf{x}_1 - \mathbf{x}_3\|_2^2 + \|\mathbf{x}_2 - \mathbf{x}_4\|_2^2 \quad \leq \quad & \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \|\mathbf{x}_2 - \mathbf{x}_3\|_2^2 \\ + \quad & \|\mathbf{x}_3 - \mathbf{x}_4\|_2^2 + \|\mathbf{x}_4 - \mathbf{x}_1\|_2^2.\end{aligned}$$

The lemma immediately implies Proposition 3.4.1, since for the metric of the 4-cycle, the sum of the squared lengths of the diagonals is 8, twice larger than the sum of the squared edges, and so any noncontracting embedding in a Euclidean space has to contract a diagonal by at least $\sqrt{2}$.

The use of *squared* lengths in the lemma is a key trick for proving Proposition 3.4.1, and it is an instance of a general rule of thumb, which we already met in the proof of Proposition 1.4.2: For dealing with ℓ_p metrics, it is usually easiest to work with pth powers of the distances.

First proof of Lemma 3.4.2. First we observe that it suffices to prove the lemma for points x_1, x_2, x_3, x_4 on the real line. Indeed, for the $\mathbf{x}_i$ in some $\mathbb{R}^k$ we can write the 1-dimensional inequality for each coordinate and then add these inequalities together.

If the x_i are real numbers, we calculate

$$\begin{aligned}&(x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^2 + (x_4 - x_1)^2 \\ &- (x_1 - x_3)^2 - (x_2 - x_4)^2 \\ &\qquad = \quad (x_1 - x_2 + x_3 - x_4)^2 \geq 0,\end{aligned} \tag{3.3}$$

and this is the desired inequality. □

Second proof of Lemma 3.4.2. Here is a more systematic approach using basic linear algebra. It is perhaps too a great hammer for this particular problem, but it will be useful for more complicated questions.

As in the first proof, it suffices to show that the quadratic form (3.3) is nonnegative for all $x_1, \ldots, x_4 \in \mathbb{R}$. The quadratic form can be rewritten in a matrix notation as $\mathbf{x}^T C \mathbf{x}$, where $\mathbf{x} = (x_1, x_2, x_3, x_4)$ (understood as

a column vector) and C is the symmetric matrix

$$\begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix}.$$

(Note that we first reduced the original problem, dealing with points in a Euclidean space, to a problem about points in $\mathbb{R}$, and then we represented a 4-tuple on these one-dimensional points as a 4-dimensional vector.)

So we need to show that C is positive semidefinite, and linear algebra offers several methods for that. For example, one can check that the eigenvalues are $0, 0, 0, 4$—all nonnegative. □

The cube. Let Q_m denote the space $\{0,1\}^m$ with the ℓ_1 metric. In other words, the points are all m-term sequences of 0s and 1s, and the distance of two such sequences is the number of places where they differ (this is also called the **Hamming distance**).

We can also regard Q_m as the vertex set $V := \{0,1\}^m$ the "graph-theoretical cube" with the shortest-path metric, where the edge set is

$$E := \{\{\mathbf{u}, \mathbf{v}\} : \mathbf{u}, \mathbf{v} \in \{0,1\}^m, \|\mathbf{u} - \mathbf{v}\|_1 = 1\}$$

Thus, for $m = 2$ we recover the 4-cycle.

3.4.3 Theorem. *Let $m \geq 2$ and $n = 2^m$. Then there is no D-embedding of the cube Q_m into ℓ_2 with $D < \sqrt{m} = \sqrt{\log_2 n}$. That is, the natural embedding, where we regard $\{0,1\}^m$ as a subspace of ℓ_2^m, is optimal.*

Historically, this is the first result showing that some metric spaces require an arbitrarily large distortion for embedding in ℓ_2. It also remains one of the simplest and nicest such examples, although the distortion is only $\sqrt{\log n}$, while order $\log n$ can be achieved with different n-point examples.

However, it is generally believed that the cube is the worst among all n-point ℓ_1 metrics. This hasn't been proved, but it is known that every n-point ℓ_1 metric embeds in ℓ_2 with distortion $O(\sqrt{\log n} \log\log n)$ (with a complicated proof beyond the scope of this text).

Proof of Theorem 3.4.3. We generalize the first proof above for the 4-cycle.

Let E be the edge set of Q_m, and let F be the set of the **long diagonals**, which are pairs of points at distance m. In other words,

$$F = \{\{\mathbf{u}, \overline{\mathbf{u}}\} : \mathbf{u} \in \{0,1\}^m\},$$

where $\overline{\mathbf{u}} = \mathbf{1} - \mathbf{u}$ (with $\mathbf{1} = (1, 1, \ldots, 1)$).

We have $|F| = 2^{m-1}$ and $|E| = m2^{m-1}$. Each edge has length 1, while the long diagonals have length m. Thus the sum of the squared lengths of all long diagonals equals $m^2 2^{m-1}$, while the sum of the squared lengths of the edges is $m2^{m-1}$, which is m-times smaller.

Next, let us consider an arbitrary mapping $f\colon \{0,1\}^m \to \ell_2$. We will show that

$$\sum_{\{\mathbf{u},\mathbf{v}\}\in F} \|f(\mathbf{u}) - f(\mathbf{v})\|_2^2 \leq \sum_{\{\mathbf{u},\mathbf{v}\}\in E} \|f(\mathbf{u}) - f(\mathbf{v})\|_2^2; \qquad (3.4)$$

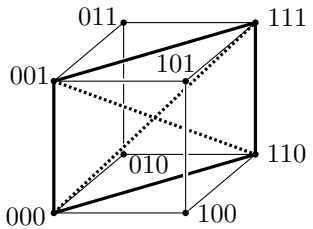
in words, for any configuration of 2^m points in a Euclidean space, the sum of squared lengths of the long diagonals is at most the sum of the squared lengths of the edges. This obviously implies the theorem, and it remains to prove (3.4).

We proceed by induction on m; the base case $m = 2$ is Lemma 3.4.2.

For $m > 2$, we divide the vertex set V into two parts V_0 and V_1, where V_0 are the vectors with the last component 0, i.e., of the form $\mathbf{u}0$, $\mathbf{u} \in \{0,1\}^{m-1}$. The set V_0 induces an $(m-1)$-dimensional subcube. Let E_0 be its edge set, let $F_0 = \{\{\mathbf{u}0, \overline{\mathbf{u}}0\} : \mathbf{u} \in \{0,1\}^{m-1}\}$ be the set of its long diagonals, and similarly for E_1 and F_1. Let $E_{01} = E \setminus (E_0 \cup E_1)$ be the edges of the m-dimensional cube connecting the two subcubes.

By induction, we have $\sum_{F_0} \leq \sum_{E_0}$ and $\sum_{F_1} \leq \sum_{E_1}$, where, e.g., $\sum_{F_0}$ is a shorthand for $\sum_{\{\mathbf{u},\mathbf{v}\}\in F_0} \|f(\mathbf{u}) - f(\mathbf{v})\|_2^2$.

For each $\mathbf{u} \in \{0,1\}^{m-1}$, we consider the quadrilateral with vertices $\mathbf{u}0$, $\overline{\mathbf{u}}0$, $\overline{\mathbf{u}}1$, $\mathbf{u}1$; for $\mathbf{u} = (0,0)$, it is indicated in the picture:



The sides of this quadrilateral are two edges of E_{01}, one diagonal from F_0 and one from F_1, and its diagonals are from F. Each edge from $E_{01} \cup F_0 \cup F_1 \cup F$ is contained in exactly one such quadrilateral.

Let us write the inequality of Lemma 3.4.2 for this quadrilateral and sum over all such quadrilaterals (their number is 2^{m-2}, since $\mathbf{u}$ and $\overline{\mathbf{u}}$ yield the same quadrilateral). This yields

$$\sum_F \le \sum_{E_{01}} + \sum_{F_0} + \sum_{F_1} \le \sum_{E_{01}} + \sum_{E_0} + \sum_{E_1} = \sum_E,$$

the last inequality relying on the inductive assumption for the two subcubes. The inequality (3.4) is proved, and so is Theorem 3.4.3. □

A sketch of another proof of (3.4). The inequality can also be proved in the spirit of the second proof of Lemma 3.4.2. This proof is perhaps best regarded in the context of harmonic analysis on the Boolean cube, which we will mention later.

But an interested reader can work the proof out right now, in the language of matrices. It suffices to establish positive semidefiniteness of an appropriate symmetric matrix C, whose rows and columns are indexed by $\{0,1\}^m$. It turns out that $C = (m-1)I_{2^m} - A + P$, where

- A is the usual **adjacency matrix** of the cube, i.e., $a_{\mathbf{uv}} = 1$ if $\{\mathbf{u},\mathbf{v}\} \in E$ and $a_{\mathbf{uv}} = 0$ otherwise;
- P is the adjacency matrix corresponding similarly to the edge set F; and
- I_{2^m} is the identity matrix.

Luckily, the eigenvectors of A are well known (and easily verified): They are exactly the vectors of the *Hadamard–Walsh* orthogonal system $(\mathbf{h_v} : \mathbf{v} \in \{0,1\}^m)$, where $(\mathbf{h_v})_\mathbf{u}$, i.e., the component of $\mathbf{h_v}$ indexed by $\mathbf{u} \in \{0,1\}^m$, equals $(-1)^{\langle \mathbf{u},\mathbf{v}\rangle}$. These $\mathbf{h_v}$ happen to be eigenvectors of C as well. So one can check that the eigenvalues are all nonnegative—we omit the computations.

The inequality (3.4) and "similar" ones are sometimes called **Poincaré inequalities** in the literature. The term Poincaré inequality is commonly used in the theory of Sobolev spaces (and it bounds the L_p norm of a differentiable function using the L_p norm of its gradient). The inequalities considered in the theory discussed here can be regarded, in a vague sense, as a discrete analog.

3.5 Nonembeddability of expanders in ℓ_2

In Proposition 3.3.3, we proved the *existence* of an n-point metric space requiring distortion $\Omega(\log n/\log\log n)$ for embedding in ℓ_2. Here we prove the slightly stronger, and asymptotically tight, lower bound of $\Omega(\log n)$ for an *explicit* example—by the method introduced in the previous section. As we'll see in the next section, the same lower bound actually holds for embeddings in ℓ_1 as well. The example is the vertex set of a constant-degree expander G (with the shortest-path metric).

Roughly speaking, expanders are graphs that are sparse but well connected. If a physical model of an expander is made with little balls representing vertices and thin strings representing edges, it is difficult to tear off any subset of vertices, and the more vertices we want to tear off, the larger effort that is needed.

For technical reasons, we will consider only regular expanders; we recall that a graph G is **regular** if the degrees of all vertices are the same, equal to some number r (then we also speak of an ***r*-regular** graph).

There are two definitions of constant-degree expanders, combinatorial and algebraic.

For the *combinatorial definition*, let G be a given graph with vertex set V and edge set E. Let us call a partition of V into two subsets S and $V \setminus S$, with both S and $V \setminus S$ nonempty, a **cut**, and let $\boldsymbol{E(S, V \setminus S)}$ stand for the set of all edges in G connecting a vertex of S to a vertex of $V \setminus S$. We define the **Cheeger constant** of G (also known as the **edge expansion** or **conductance** of G in the literature) as

$$h(G) := \min\left\{\frac{|E(S, V\setminus S)|}{|S|} : S \subseteq V, 1 \le |S| \le |V|/2\right\}.$$

Intuitively, if we want to cut off some vertices of G, but not more than half, and we pay one unit for cutting an edge, then $h(G)$ is the minimum price per vertex of the cut part.

> An infinite sequence $(G_1, G_2, \ldots)$ of r-regular graphs with $|V(G_i)| \to \infty$ as $i \to \infty$ is a family of **constant-degree expanders** if $h(G_i) \ge \beta$ for all i, where $\beta > 0$ is a constant independent of i.

(Since we want to bound $h(G)$ from below by some constant, but we don't want to specify which constant, it doesn't make much sense that a given single graph is an expander—indeed, *every* connected regular graph belongs to a family of constant-degree expanders.)

For the *algebraic definition* of expanders, we need to recall a few things about graph eigenvalues. We have already mentioned in passing the **adjacency matrix** $A = A_G$ of a graph G. For G with n vertices, A is an $n \times n$ matrix, with both rows and columns indexed by the vertices of G, given by

$$a_{uv} = \begin{cases} 1 & \text{if } u \neq v \text{ and } \{u,v\} \in E(G), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Since A is a symmetric real matrix, it has n real eigenvalues, which we write in a non-increasing order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We can also fix corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ (i.e., $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$) that form an orthogonal basis of $\mathbb{R}^n$ (all of this is basic linear algebra and can be found in many textbooks).

If G is r-regular, then we have $\lambda_1 = r$ and $\mathbf{v}_1 = \mathbf{1}$ (the vector of all 1s), as is easy to check. So the largest eigenvalue is kind of boring, but the *second* largest eigenvalue λ_2 is a key parameter of G. More precisely, the important quantity is the difference $r - \lambda_2$, which we denote by $\mathrm{gap}(G)$ and call the **eigenvalue gap** of G.

An infinite sequence $(G_1, G_2, \ldots)$ of r-regular graphs with $|V(G_i)| \to \infty$ as $i \to \infty$ is a family of **constant-degree expanders** if $\mathrm{gap}(G_i) \geq \gamma$ for all i, where $\gamma > 0$ is a constant independent of i.

It turns out that the combinatorial definition and the algebraic one yield the same notion. This is based on the following (nontrivial) quantitative result: For every graph G, we have

$$\frac{h(G)^2}{2r} \leq \mathrm{gap}(G) \leq 2h(G)$$

(proof omitted). Both of the inequalities are essentially tight; that is, there are graphs with $\mathrm{gap}(G) \approx h(G)$, as well as graphs with $\mathrm{gap}(G) \approx h(G)^2/r$.

We will need the *existence* of families of constant-degree expanders, but we won't prove it here. There is a reasonably straightforward probabilistic proof, as well as several explicit constructions. Some of the constructions are quite simple to state. For example, a family $(G_1, G_2, \ldots)$ of 8-regular expanders can be constructed as follows: G_m has vertex set $\mathbb{Z}_m \times \mathbb{Z}_m$ (where $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$ are the integers modulo m), and the neighbors of the vertex (x, y) are $(x + y, y)$, $(x - y, y)$, $(x, y + x)$, $(x, y - x)$, $(x + y + 1, y)$, $(x - y + 1, y)$, $(x, y + x + 1)$, and $(x, y - x + 1)$ (addition and subtraction modulo m, and G_i has loops and multiple edges). However, the proof that we indeed get a family of expanders is quite sophisticated, and the proofs for other constructions are of comparable difficulty or even much harder.

Nonembeddability in ℓ_2 via the eigenvalue gap. We'll use the algebraic definition to show that constant-degree expanders require $\Omega(\log n)$ distortion for embedding in ℓ_2.

The next lemma characterizes the eigenvalue gap of an arbitrary graph in terms of ℓ_2 embeddings of its vertex set.

3.5.1 Lemma. *Let $G = (V, E)$ be an arbitrary r-regular graph on n vertices, and let $F := \binom{V}{2}$ be the set of edges of the complete graph on V. Let α be the largest real number such that the inequality*

$$\alpha \cdot \sum_{\{u,v\} \in F} \|f(u) - f(v)\|_2^2 \leq \sum_{\{u,v\} \in E} \|f(u) - f(v)\|_2^2;$$

holds for all mappings $f : V \to \ell_2$. Then

$$\alpha = \frac{\mathrm{gap}(G)}{n}.$$

Proof. We follow the method of the second proof for the 4-cycle (Lemma 3.4.2). That is, we need to understand for what values of α is an appropriate $n \times n$ matrix C positive semidefinite. We calculate

$$C = rI_n - A - \alpha(n-1)I_n + \alpha(J_n - I_n) = (r - \alpha n)I_n - A + \alpha J_n,$$

where $A = A_G$ is the adjacency matrix of G, I_n is the identity matrix, and J_n is the all 1s matrix (thus, $J_n - I_n = A_{K_n}$).

We recall that $\mathbf{v}_1 = \mathbf{1}, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are mutually orthogonal eigenvectors of A, with eigenvalues $\lambda_1 = r \geq \lambda_2 \geq \cdots \geq \lambda_n$. Because of the very simple structure of the set F, the $\mathbf{v}_i$ are also eigenvectors of C. Indeed,

$$C\mathbf{1} = (r - \alpha n)\mathbf{1} - r\mathbf{1} + \alpha J_n \mathbf{1} = \mathbf{0},$$

so the eigenvalue of C belonging to $\mathbf{v}_1$ is 0. For $i \geq 2$,

$$C\mathbf{v}_i = (r - \alpha n)\mathbf{v}_i - \lambda_i \mathbf{v}_i - \alpha J_n \mathbf{v}_i = (r - \alpha n - \lambda_i)\mathbf{v}_i$$

since each $\mathbf{v}_i$, $i \geq 2$, is orthogonal to $\mathbf{1}$ and thus $J_n \mathbf{v}_i = \mathbf{0}$.

So the eigenvalues of C are 0 and $r - \alpha n - \lambda_i$, $i = 2, 3, \ldots, n$, and they are nonnegative exactly if $\alpha \leq (r - \lambda_2)/n = \mathrm{gap}(G)/n$. □

3.5.2 Lemma. *For every integer r there exists $c_r > 0$ such that the following holds. Let G be a graph of on n vertices of maximum degree r. Then for every vertex $u \in V(G)$ there are at least $\frac{n}{2}$ vertices with distance at least $c_r \log n$ from u.*

Proof. For any given vertex u, there are at most r vertices at distance 1 from u, at most $r(r-1)$ vertices at distance 2,..., at most $r(r-1)^{k-1}$ vertices at distance k. Let us choose k as the largest integer with $1 + r + r(r-1) + \cdots + r(r-1)^{k-1} \le \frac{n}{2}$; a simple calculation shows that $k \ge c_r \log n$. Then at least $\frac{n}{2}$ vertices have distance larger than k from u. □

Now we're ready for the main nonembeddability result in this section.

3.5.3 Theorem. *Let G be an r-regular graph on n vertices with $\mathrm{gap}(G) \ge \gamma > 0$. Then an embedding of the shortest-path metric of G in ℓ_2 requires distortion at least $c \log n$, for a suitable positive $c = c(r, \gamma)$. Using the existence of families of constant-degree expanders, we thus get that, for infinitely many n, there are n-point metric spaces requiring distortion $\Omega(\log n)$ for embedding in ℓ_2.*

Proof. For a graph $G = (V, E)$ as in the theorem and for every mapping $f: V \to \ell_2$ we have

$$\sum_{\{u,v\} \in F} \|f(u) - f(v)\|_2^2 \le \frac{n}{\gamma} \cdot \sum_{\{u,v\} \in E} \|f(u) - f(v)\|_2^2 \qquad (3.5)$$

by Lemma 3.5.1. On the other hand, letting d_G denote the shortest-path metric of G and using Lemma 3.5.2, we calculate that

$$\sum_{\{u,v\} \in F} d_G(u,v)^2 \ge \Omega(n \log^2 n) \cdot \sum_{\{u,v\} \in E} d_G(u,v)^2. \qquad (3.6)$$

The theorem follows by comparing these two inequalities. □

3.6 Nonembeddability of expanders in ℓ_1

In this section we'll show that the metric of constant-degree expanders requires distortion $\Omega(\log n)$ for embedding in ℓ_1 as well.

3.6.1 Theorem. *Let G be an r-regular graph on n vertices with $h(G) \ge \beta > 0$. Then an embedding of the shortest-path metric of G in ℓ_1 requires distortion at least $c \log n$, for a suitable positive $c = c(r, \beta)$. Using the existence of families of constant-degree expanders, we get that, for infinitely many n, there are n-point metric spaces requiring distortion $\Omega(\log n)$ for embedding in ℓ_1.*

This result is strictly stronger than the ℓ_2 lower bound proved in the previous section, since every ℓ_2 metric is also an ℓ_1 metric. However, we've presented the ℓ_2 result separately, since the proof is of independent interest.

We now offer two proofs of the ℓ_1 lower bound. The first one follows by an extension of the ℓ_2 proof and by an interesting isometric embedding. The second one is analogous to the ℓ_2 proof but it doesn't use it; rather, it is based on the combinatorial definition of expanders. Thus, if one wants to obtain both the ℓ_1 result and the ℓ_2 result quickly, the second proof in this section is probably the method of choice.

Metrics of negative type. For the first proof, we introduce a class of metrics, which has played a key role in several recent results on metric embeddings.

A metric d on a (finite) set V is called a **metric of negative type**, or alternatively, a **squared ℓ_2 metric**, if there exists a mapping $f: V \to \ell_2$ such that

$$d(u,v) = \|f(u) - f(v)\|_2^2$$

for all $u, v \in V$.

So a metric of negative type is a metric that can be represented by squared Euclidean distances of some points in some $\mathbb{R}^k$. There is a subtlety in this definition: If we take arbitrary points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ and define $d(\mathbf{x}_i, \mathbf{x}_j) := \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$, we need not obtain a metric of negative type, *because we need not obtain a metric at all.*[8] Indeed, for three distinct collinear points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ the triangle inequality for this d fails! So only rather special configurations of points in $\mathbb{R}^k$ give rise to squared ℓ_2

[8] Thus, the term "squared ℓ_2 metric" is potentially misleading, since one may be tempted to parse it as "squared (ℓ_2 metric)", as opposed to the correct "(squared ℓ_2) metric". Thus, the less explanatory but also less confusing traditional term "metric of negative type" seems preferable.

metrics. More precisely, the squared Euclidean distances in a set $X \subset \mathbb{R}^k$ form a metric exactly if no three points of X form a triangle with a (strictly) obtuse angle, as is not difficult to show.

In analogy to the metric cone $\mathcal{M}$ and the cone $\mathcal{L}_p$ introduced in Section 1.4, we let $\mathcal{N} \subset \mathbb{R}^N$, $N = \binom{n}{2}$, be the set of points representing metrics of negative type on an n-point set. By definition, we have $\mathcal{N} = \mathcal{L}_2 \cap \mathcal{M}$ (a metric of negative type is a square of a Euclidean metric *and* a metric), and so $\mathcal{N}$ is also a convex cone.

3.6.2 Lemma. *Every ℓ_1 metric is also a metric of negative type.*

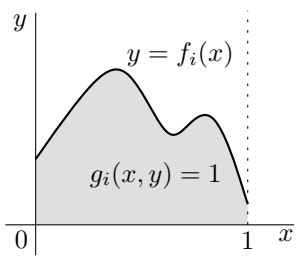
First proof. Since every ℓ_1 metric is a nonnegative linear combination of cut metrics (Proposition 1.4.1), and $\mathcal{N}$ is closed under nonnegative linear combinations, it suffices to show that every cut metric is of negative type. But this is obvious. □

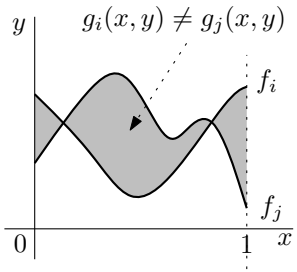
Second proof. We give a proof via function spaces, which is simple and natural. It could also be made "finite-dimensional" in the spirit of the discussion in Sections 1.5 and 2.5.

Given an ℓ_1 metric d on $\{1, 2, \dots, n\}$, let's represent it by *nonnegative* real functions $f_1, f_2, \dots, f_n \in L_1(0,1)$; that is, $d(i,j) = \|f_i - f_j\|_1$. Then we let $g_i\colon [0,1] \times \mathbb{R} \to \mathbb{R}$ be the characteristic function of the planar region between the x-axis and the graph of f_i; formally

$$g_i(x,y) := \begin{cases} 1 & \text{if } f_i(x) \in [0,y] \\ 0 & \text{otherwise.} \end{cases}$$

Then $\|g_i - g_j\|_2^2 = \int_0^1 \int_{-\infty}^{\infty} (g_i(x,y) - g_j(x,y))^2 \,\mathrm{d}y\,\mathrm{d}x$ is the area of the set $\{(x,y) \in [0,1] \times \mathbb{R} : g_i(x,y) \ne g_j(x,y)\}$, which in turn equals $\int_0^1 |f_i(x) - f_j(x)|\,\mathrm{d}x = \|f_i - f_j\|_1$:





So the ℓ_1 distances of the f_i equal the squared ℓ_2 distances of the g_i. Since $L_2([0,1] \times \mathbb{R})$ is a countable Hilbert space and thus isometric to ℓ_2, we indeed get a representation by a square of an ℓ_2 metric. □

First proof of Theorem 3.6.1. It suffices to observe that a minor modification of the proof in the previous section actually shows that constant-degree expanders also require $\Omega(\log n)$ distortion for embedding in any metric of negative type. Indeed, instead of the inequality (3.6) for the sum of *squares* of the expander metric, we derive from Lemma 3.5.2 the inequality

$$\sum_{\{u,v\} \in F} d_G(u,v) \ge \Omega(n \log n) \cdot \sum_{\{u,v\} \in E} d_G(u,v).$$

Comparing it with the inequality (3.5) for the squared Euclidean distances gives the claimed $\Omega(\log n)$ lower bound. This implies the same lower bound for embedding in ℓ_1 by Lemma 3.6.2. □

Nonembeddability in ℓ_1 via combinatorial expansion. Theorem 3.6.1 can also be proved directly, without a detour through the metrics of negative type.

In order to formulate an analog of Lemma 3.5.1, we introduce another parameter of the graph G, similar to the Cheeger constant $h(G)$. Namely, for a cut S in G we define the **density**

$$\phi(G,S) := \frac{|E(S, V \setminus S)|}{|S| \cdot |V \setminus S|}$$

(this is the ratio of the number of edges connecting S and $V \setminus S$ in G and in the complete graph on V), and $\boldsymbol{\phi(G)}$ is the smallest density of a cut in G. It's easy to see that $h(G) \le n\phi(G) \le 2h(G)$ for all G, and so $\phi(G)$ is essentially $h(G)$ with a different scaling.

3.6.3 Lemma. *Let $G = (V,E)$ be an arbitrary graph on n vertices, and let $F := \binom{V}{2}$ be the set of edges of the complete graph on V. Let β be the largest real number such that the inequality*

$$\beta \cdot \sum_{\{u,v\} \in F} \|f(u) - f(v)\|_1 \le \sum_{\{u,v\} \in E} \|f(u) - f(v)\|_1;$$

holds for all mappings $f\colon V \to \ell_1$. Then

$$\beta = \phi(G).$$

Assuming this lemma, the *second proof of Theorem 3.6.1* is completely analogous to the derivation of the ℓ_2 result, Theorem 3.5.3 from Lemma 3.5.1, and so we omit it. It remains to prove the lemma, and for this we can again offer two ways.

First proof of Lemma 3.6.3. As in the ℓ_2 case, we may assume that the values of f are real numbers. So we can write

$$\beta = \inf\{Q(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \text{ nonconstant}\}, \tag{3.7}$$

with

$$Q(\mathbf{x}) := \frac{\sum_{\{u,v\}\in E} |x_u - x_v|}{\sum_{\{u,v\}\in F} |x_u - x_v|}$$

(we call $\mathbf{x}$ nonconstant if there are u, v with $x_u \neq x_v$; the restriction to nonconstant $\mathbf{x}$ is needed to have the denominator nonzero).

We observe that the minimum of $Q(\mathbf{x})$ over all nonconstant $\mathbf{x} \in \{0,1\}^n$ is exactly $\phi(G)$ (right?). So it remains to show that the infimum in (3.7) is attained for $\mathbf{x} \in \{0,1\}^n$.

For a nonconstant $\mathbf{x} \in \mathbb{R}^n$, let $k(\mathbf{x}) \geq 2$ denote the number of distinct values attained by the components of $\mathbf{x}$. Given an $\mathbf{x}$ with $k(\mathbf{x}) \geq 3$, we'll find $\mathbf{x}' \in K$ with $k(\mathbf{x}') < k(\mathbf{x})$ and $Q(\mathbf{x}') \leq Q(\mathbf{x})$.

Suppose that $r < s < t$ are three consecutive values attained by the components of $\mathbf{x}$. We let s vary in the interval $[r,t]$ and keep everything else fixed. More precisely, for $s' \in [r,t]$, we define $\mathbf{x}'(s')$ by setting those components that equal s in $\mathbf{x}$ to s', and letting all the other components agree with $\mathbf{x}$.

The key observation is that both the numerator and the denominator of $Q(\mathbf{x}'(s'))$ are *linear* functions of s' (this is because the function $x \mapsto |a-x|$ is linear on each interval not containing a in the interior). Moreover, the denominator remains nonzero for all $s' \in [r,t]$. Then it's easily seen that $Q(\mathbf{x}'(s'))$ attains minimum for $s' = r$ or $s' = t$. The corresponding $\mathbf{x}'$ satisfies $k(\mathbf{x}') < k(\mathbf{x})$ because the value s has been eliminated.

Thus, the infimum in (3.7) can be taken only over the $\mathbf{x}$ with $k(\mathbf{x}) = 2$. By scaling, we can even restrict ourselves to nonconstant $\mathbf{x} \in \{0,1\}^n$, which concludes the proof. □

Second proof of Lemma 3.6.3. This proof is, in a way, simpler, but it needs more machinery. We use $\mathcal{L}_1$, the cone of ℓ_1 metrics introduced in Section 1.4, and the fact that every ℓ_1 metric is a nonnegative linear combination of cut metrics (Proposition 1.4.1).

We can re-formulate the definition of β as

$$\beta = \inf_{\mathbf{d}\in\mathcal{L}_1\setminus\{\mathbf{0}\}} R(\mathbf{d}), \tag{3.8}$$

where

$$R(\mathbf{d}) = \frac{\sum_{\{u,v\}\in E} d(u,v)}{\sum_{\{u,v\}\in F} d(u,v)}.$$

We claim that for every $\mathbf{d} \in \mathcal{L}_1$, there is a cut metric $\mathbf{d}^*$ with $R(\mathbf{d}^*) \leq R(\mathbf{d})$. Indeed, we know that $\mathbf{d}$ can be written as $\sum_{i=1}^k \lambda_i \mathbf{d}_i$, with the λ_i positive reals and the $\mathbf{d}_i$ cut metrics. Then $R(\mathbf{d}) \geq \min_i R(\mathbf{d}_i)$, using the inequality

$$\frac{a_1 + a_2 + \cdots + a_n}{b_1 + b_2 + \cdots + b_n} \geq \min\left\{\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_n}{b_n}\right\}$$

valid for all positive reals $a_1, \ldots, a_n, b_1, \ldots, b_n$ (a quick proof: if $a_i \geq \alpha b_i$ for all i, then $\sum_i a_i \geq \alpha \sum_i b_i$).

Thus, the infimum in (3.8) is attained for some cut metric, and it remains to observe (as in the first proof) that the minimum of R over (nonzero) cut metrics is exactly $\phi(G)$.

Essentially the same proof can also be expressed more geometrically. Let $\mathcal{P}$ be the convex polytope obtained as the intersection of $\mathcal{L}_1$ with the hyperplane $\{\mathbf{d} \in \mathbb{R}^N : \sum_{\{u,v\}\in F} d(u,v) = 1\}$ (it's easily seen that $\mathcal{P}$ is bounded). We have $\beta = \inf_{\mathcal{P}} R(\mathbf{d})$. Now R is a *linear* function on $\mathcal{P}$, and thus it attains its infimum at a vertex $\mathbf{d}_0$ of $\mathcal{P}$. The vertices are multiples of cut metrics, and we conclude as above. □

3.7 Computing the smallest distortion for embedding in ℓ_2

As a rule of thumb, the D-embeddability question, considered as a computational problem, is usually hard. For example, it is known that the question "Given an n-point metric space, does it 1-embed in ℓ_1?" is NP-hard, and there are several other results of this kind (although not all interesting cases have been settled).

The result of this section is an exception to this rule: When the target space is ℓ_2, a Euclidean space of unlimited dimension, then the minimum distortion D required for embedding a given n-point metric space can be computed in polynomial time (more precisely, it can be approximated to any desired accuracy).

3.7.1 Proposition. *The smallest distortion D required to embed a given finite metric space (V, d_V) in ℓ_2 can be approximated with any given accuracy $\varepsilon > 0$ in time polynomial in the number of bits needed to represent d_V and in $\log \frac{1}{\varepsilon}$.*

The proof has two ingredients, both of independent interest. The first one is a characterization of Euclidean metrics in terms of positive semidefinite matrices. Here it is convenient to index the points starting from 0, rather than by $\{1, \dots, n\}$ as usual.

3.7.2 Theorem. *Let $V = \{0, 1, \dots, n\}$, and let $\mathbf{z} = (z_{ij} : \{i, j\} \in \binom{V}{2})$ be a given vector of real numbers. Then $\mathbf{z} \in \mathcal{L}_2$ (in other words, there exists a Euclidean metric d_V on V such that $z_{ij} = d_V(i, j)^2$ for all i, j) if and only the $n \times n$ matrix G given by*

$$g_{ij} := \tfrac{1}{2}\left(z_{0i} + z_{0j} - z_{ij}\right), \quad i, j = 1, 2, \dots, n$$

(where $z_{ij} = z_{ji}$ for all i, j, and $z_{ii} = 0$ for all i), is positive semidefinite.

Proof. We need a standard linear-algebraic fact: An $n \times n$ matrix A is positive semidefinite if and only if it can be expressed as $A = B^T B$ for some $n \times n$ real matrix B.

First we check necessity of the condition in the theorem; that is, if $\mathbf{p}_0, \dots, \mathbf{p}_n \in \ell_2^n$ are given points and $z_{ij} := \|\mathbf{p}_i - \mathbf{p}_j\|_2^2$, then G is positive semidefinite. For this, we need the *cosine theorem*, which tells us that $\|\mathbf{a}-\mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - 2\langle \mathbf{a}, \mathbf{b}\rangle$ for every two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Thus, if we define $\mathbf{v}_i := \mathbf{p}_i - \mathbf{p}_0$, $i = 1, 2, \dots, n$, we get that $\langle \mathbf{v}_i, \mathbf{v}_j\rangle = \frac{1}{2}(\|\mathbf{v}_i\|^2 + \|\mathbf{v}_j\|^2 - \|\mathbf{v}_i - \mathbf{v}_j\|^2) = g_{ij}$. So G is the **Gram matrix** of the vectors $\mathbf{v}_i$, we can write $G = B^T B$ for the matrix B having $\mathbf{v}_i$ as the ith column, and hence G is positive semidefinite.

Conversely, let's assume that G as in the theorem is positive semidefinite; thus, it can be factored as $G = B^T B$ for some $n \times n$ matrix B. Then we let $\mathbf{p}_i \in \mathbb{R}^n$ be the ith column of B for $i = 1, 2, \dots, n$, while $\mathbf{p}_0 := \mathbf{0}$. Reversing the above calculation, we arrive at $\|\mathbf{p}_i - \mathbf{p}_j\|_2^2 = z_{ij}$, and the proof is finished. □

Proof of Proposition 3.7.1. Given a metric d_V on the set $V = \{0, 1, \dots, n\}$, we can use Theorem 3.7.2 to express the smallest D for

which (V, d_V) can be D-embedded in ℓ_2 as the minimum of the following optimization problem.

$$\begin{array}{lll} \text{Minimize} & D & \\ \text{subject to} & d_V(i,j)^2 \le z_{ij} \le D^2 d_V(i,j)^2 & \text{for all } \{i,j\} \in \binom{V}{2}, \\ & g_{ij} = \frac{1}{2}\left(z_{0i} + z_{0j} - z_{ij}\right) & i, j = 1, 2, \dots, n, \\ & \text{the matrix } G = (g_{ij})_{i,j=1}^n \text{ is positive semidefinite.} & \end{array}$$

The variables in this problem are D, z_{ij} for $\{i, j\} \in \binom{V}{2}$ (where z_{ij} refers to the same variable as z_{ji}, and z_{ii} is interpreted as 0), and g_{ij}, $i, j = 1, 2, \dots, n$.

This optimization problem is an instance of a **semidefinite program**. A semidefinite program in general is the problem of minimizing or maximizing a given linear function of some k real variables over a set $S \subseteq \mathbb{R}^k$, called the set of **feasible solutions**. The set S is specified by a system of linear inequalities and equations for the variables *and* by the requirement that an $n \times n$ matrix X is (symmetric and) positive semidefinite, where each entry of X is one of the k variables. The **input size** of a semidefinite program is, speaking somewhat informally, the total number of bits needed to write down all the coefficients (in the optimized linear function and in the equations and inequalities).

Semidefinite programming, the second main ingredient in the proof of Proposition 3.7.1, is a research area concerned with properties of semidefinite programs and efficient algorithms for solving them, and it constitutes one of the most powerful tools in optimization. A key fact is that, roughly speaking, an optimal solution of a given semidefinite program can be approximated in polynomial time with any prescribed accuracy. Unfortunately, the last statement is not literally true; what is really known is the following:

3.7.3 Fact. *Suppose that a given semidefinite program has at least one feasible solution, and that every component in every feasible solution is bounded in absolute value by an explicitly given number R. Then, given $\varepsilon > 0$, it is possible to compute an optimal solution of the semidefinite program with accuracy ε in time polynomial in the input size and in $\log(R/\varepsilon)$.*

This fact can be proved by the *ellipsoid method*; conceptually it is simple, but there are many nontrivial technical details.

For our specific semidefinite program above, it is easy to come up with some apriori upper bound for the minimum distortion, which holds for

every possible input metric d_V. For example, $D \le n^2$ is easy to show, and Bourgain's theorem, to be discussed later, even shows $D \le C \log n$ for a suitable constant C. We can thus add an extra constraint $D \le n^2$, say, and then the set of all feasible solutions is clearly nonempty and bounded. Then Proposition 3.7.1 follows from Fact 3.7.3. □

3.8 "Universality" of the method with inequalities

In the previous sections, we have been proving lower bounds for the distortion of embeddings in ℓ_p by the approach with inequalities, which can in general be cast as follows. Given a metric d_V on the set $V = \{1, 2, \ldots, n\}$, we set up an inequality saying that the sum of the pth powers of the distances with some coefficients is never larger than the sum of the pth powers with some other coefficients. Then we show that this inequality is valid for all ℓ_p metrics, while for the original metric d_V it is violated at least with a multiplicative factor of D^p. That is, for some choice of nonnegative coefficients a_{uv} and b_{uv}, $\{u, v\} \in F := \binom{V}{2}$, we prove that, on the one hand, for every mapping $f: V \to \ell_p$ we have

$$\sum_{\{u,v\}\in F} a_{uv}\|f(u) - f(v)\|_p^p \le \sum_{\{u,v\}\in F} b_{uv}\|f(u) - f(v)\|_p^p, \tag{3.9}$$

and on the other hand, for the original metric d_V,

$$\sum_{\{u,v\}\in F} a_{uv} d_V(u, v)^p \ge D^p \sum_{\{u,v\}\in F} b_{uv} d_V(u, v)^p. \tag{3.10}$$

How strong is this method? As we will show next, using a simple argument about separation of convex sets, it is universal in the following sense.

3.8.1 Proposition. *Let (V, d_V) be a metric space on the set $\{1, 2, \ldots, n\}$, and let's suppose that it has no D-embedding in ℓ_p for some $p \in [1, \infty)$. Then there are nonnegative coefficients a_{uv} and b_{uv}, $\{u, v\} \in F$, such that (3.9) and (3.10) hold, and thus the method of inequalities always "proves" that (V, d_V) is not D-embeddable in ℓ_p.*

One should not get over-enthusiastic about this result: It tells us that the right coefficients always exist, but it doesn't tell us how to find them,

and moreover, even if we knew the coefficients, it's not clear in general how (3.9) can be verified for all ℓ_p metrics.

Proof of Proposition 3.8.1. We consider the following two convex sets in $\mathbb{R}^N$, $N = |F| = \binom{n}{2}$: the cone of pth powers of ℓ_p metrics

$$\mathcal{L}_p = \Big\{(\|f(u) - f(v)\|_p^p)_{\{u,v\}\in F} : f: V \to \ell_p\Big\},$$

and the set

$$\mathcal{K}_p = \mathcal{K}_p(d_V) = \big\{\mathbf{z} \in \mathbb{R}^N : d_V(u, v)^p \le z_{uv} \le D^p d_V(u, v)^p \text{ for all } \{u, v\} \in F\big\}.$$

This $\mathcal{K}_p$ includes all pth powers of metrics arising by noncontracting D-embeddings of (V, d_V). But not all elements of $\mathcal{K}_p$ are necessarily of this form, since the triangle inequality may be violated.

Both of $\mathcal{K}_p$ and $\mathcal{L}_p$ are convex (for $\mathcal{K}_p$ this is obvious, and for $\mathcal{L}_p$ we saw it in the proof of Proposition 1.4.2). The assumption that (V, d_V) has no D-embedding in ℓ_p means that $\mathcal{K}_p \cap \mathcal{L}_p = \emptyset$.

Therefore, $\mathcal{K}_p$ and $\mathcal{L}_p$ can be separated by a hyperplane; that is, there exist $\mathbf{c} \in \mathbb{R}^N$ and $b \in \mathbb{R}$ such that $\langle \mathbf{c}, \mathbf{z}\rangle \ge b$ for all $\mathbf{z} \in \mathcal{K}_p$, while $\langle \mathbf{c}, \mathbf{z}\rangle \le b$ for all $\mathbf{z} \in \mathcal{L}_p$.

Next, we check that we may assume $b = 0$. Since $\mathbf{0} \in \mathcal{L}_p$, we must have $b \ge 0$, and thus

$$\langle \mathbf{c}, \mathbf{z}\rangle \ge 0 \text{ for all } \mathbf{z} \in \mathcal{K}_p.$$

Since $\mathcal{L}_p$ is a cone (i.e., $\mathbf{z} \in \mathcal{L}_p$ implies $t\mathbf{z} \in \mathcal{L}_p$ for all $t \ge 0$), every $\mathbf{z} \in \mathcal{L}_p$ satisfies $\langle \mathbf{c}, \mathbf{z}\rangle \le \frac{1}{t} b$ for all $t > 0$, and therefore,

$$\langle \mathbf{c}, \mathbf{z}\rangle \le 0 \text{ for all } \mathbf{z} \in \mathcal{L}_p.$$

Now we define

$$a_{uv} := c_{uv}^+, \quad b_{uv} := c_{uv}^-,$$

where we use the notation $t^+ = \max(t, 0)$, $t^- = \max(-t, 0)$.

To check the inequality (3.10), we employ $\langle \mathbf{c}, \mathbf{z}\rangle \ge 0$ for the following $\mathbf{z} \in \mathcal{K}_p$:

$$z_{uv} = \begin{cases} d_V(u, v)^p & \text{if } c_{uv} \ge 0, \\ D^p d_V(u, v)^p & \text{if } c_{uv} < 0. \end{cases}$$

Then $\langle \mathbf{c}, \mathbf{z}\rangle \ge 0$ boils down to (3.10).

To verify (3.9) for a given $f: V \to \ell_p$, we simply use $\langle \mathbf{c}, \mathbf{z}\rangle \le 0$ for the $\mathbf{z} \in \mathcal{L}_p$ given by $x_{uv} = \|f(u) - f(v)\|_p^p$. □

3.9 Nonembeddability of the edit distance in ℓ_1

Here we present a nonembeddability result where a successful application of the method with inequalities relies on a sophisticated device—a result from *harmonic* (or *Fourier*) *analysis* on the discrete cube.

Generally speaking, harmonic analysis belongs among the most powerful tools in all mathematics, and in the last approximately twenty years it has also been used in theoretical computer science and in discrete mathematics, with great success.

There are several other applications of harmonic analysis in the theory of metric embeddings, besides the one presented here. The result we will cover is very clever but simple, it gives us an opportunity to encounter yet another very important metric, and the use of harmonic analysis in it is well encapsulated in a single (and famous) theorem.

The metric: edit distance. Edit distance is a way of quantifying the amount of difference between two strings of characters—for example, between two words, or two books, or two DNA sequences. It is also called the *Levenshtein distance*, after the author of a 1965 paper, on error-correcting codes, where it was introduced.

Let $\mathbf{u} = u_1u_2\dots u_n$ and $\mathbf{v} = v_1v_2\dots v_m$ be two strings. Their **edit distance**, denoted by $\mathrm{ed}(\mathbf{u},\mathbf{v})$, is the minimum number of **edit operations** required to transform $\mathbf{u}$ into $\mathbf{v}$, where an edit operation is the *insertion* of a character, the *deletion* of a character, and the *replacement* of one character by another.

For example, the strings BIGINIG and BEGINNING have edit distance 3: it is easy to find a sequence of three edit operations transforming one to the other, and slightly less easy to check that one or two operations won't do.

In many practical problems, we need to solve the *nearest neighbor problem* discussed in Section 2.9 with respect to the edit distance. That is, we have a large collection (database) of strings, and when a query string comes, we would like to find a nearest string in the database. Obvious applications include spelling check, detecting plagiarism, or matching fragments of DNA against known genomes.

A straightforward approach to this problem can be computationally very demanding. First, computing the edit distance of long strings is expensive: for strings of length a and b the best known exact algorithm runs in time roughly proportional to ab. Second, if we just try to match the query string against each of the strings in the database, we must compute the edit distance very many times.

Now consider how much easier things would be if we had a nice embedding of strings in ℓ_1. By this we mean an embedding of the metric space of all strings over the given alphabet, up to some suitable length, in ℓ_1, that

- has a small distortion,
- is quickly computable (i.e., given a string u, its image in ℓ_1 can be found reasonably fast), and
- the dimension of the target space is not too high.

Then we could use some (approximate) nearest neighbor algorithm in ℓ_1, such as the one presented in Section 2.9. We note that in at least some of the applications, it is reasonable to expect that one of the strings in the database is much closer to the query string than all others, and then approximate search is sufficient.

This program has been pursued with some success. An embedding has been found of the space of all strings of length n (over an alphabet of a constant size) in ℓ_1 with distortion at most

$$2^{O(\sqrt{(\log n)(\log\log n)})}.$$

This grows with n but more slowly than any fixed power n^ε, although much faster than $\log n$, say. (Note that, for a two-letter alphabet, the considered metric space has $N = 2^n$ points, and so Bourgain's theorem gives distortion only $O(n)$, which is fairly useless.) The embedding is also reasonable, although not entirely satisfactory, with respect to the computational efficiency and the dimension of the target space.[9]

However, the following theorem sets a limit on how small the distortion can be.

3.9.1 Theorem. *Let us consider the set $\{0,1\}^n$ of all strings of length n over the two-letter alphabet $\{0,1\}$, equipped with the edit distance* $\mathrm{ed}(.,.)$. *Then every embedding of the resulting metric space in ℓ_1 has distortion at least $\Omega(\log n)$.*

[9] A very interesting related result concerns the edit distance metric modified to allow *block operations* (i.e., swapping two arbitrarily large contiguous blocks as a single operation). The resulting block edit metric can be embedded into ℓ_1 with distortion $O(\log n \log^* n)$. Here $\log^* n$ is the number of times we need to iterate the (binary) logarithm function to reduce n to a number not exceeding 1.

Let us stress that the lower bound concerns embedding of *all* the 2^n strings simultaneously. A database of strings, say a dictionary, is probably going to contain only a small fraction of all possible strings, and so it might still be that such a smaller set of strings can be embedded with much smaller distortion. But that embedding can't be oblivious, i.e., it has to depend on the particular database.

Let us remark that, while the work on embedding of the edit distance in ℓ_1 certainly provided valuable insights, the current best result on fast approximation of edit distance use a different, although somewhat related, strategy. At the time of writing, the best, and brand new, randomized algorithm can approximate the edit distance of two strings of length n up to a factor of $(\log n)^{O(1/\varepsilon)}$ in time $O(n^{1+\varepsilon})$, where $\varepsilon > 0$ is a constant which can be chosen at will.

Boolean functions. In the proof of Theorem 3.9.1, we will need to consider properties of **Boolean functions on the Hamming cube**, i.e., functions $f\colon \{0,1\}^n \to \{0,1\}$.[10]

It may perhaps be useful for intuition to think about such a Boolean function f as a *voting scheme*. For example, let us imagine that a department of mathematics needs to take an important decision between two possible alternatives, labeled 0 and 1—namely, whether the coffee machine should be supplied with good but expensive coffee beans *or* with not so good but cheap ones. There are n members of the department, the ith member submits a vote $u_i \in \{0,1\}$, and the decision taken by the department is given by $f(u_1,\dots,u_n)$. Here are some examples.

- A familiar voting scheme is the **majority function**, denoted by Maj(.). Assuming, for simplicity, n odd, we have $\mathrm{Maj}(u_1,\dots,u_n) = 1$ if there are more 1s that 0s among $u_1,\dots,u_n$, and $\mathrm{Maj}(u_1,\dots,u_n) = 0$ otherwise.

- The function $\mathrm{Dict}_k(u_1,\dots,u_n) := u_k$ is the **dictatorship function**, where the decision is made solely by the kth member.

- The **parity function** $\mathrm{Parity}(u_1,\dots,u_n) := u_1 + u_2 + \cdots + u_n$ (addition modulo 2) is an example of a very erratic dependence of the decision on the votes (and it's hard to imagine a situation in which it would provide a reasonable voting scheme).

[10] In the literature dealing with harmonic analysis of Boolean functions, one often considers functions $\{-1,1\}^n \to \{-1,1\}$ instead of $\{0,1\}^n \to \{0,1\}$. This is only a notational change, which has nothing to do with the essence, but some formulas and statements come out simpler in this setting (some others may look less natural, though). We will stick to the 0/1 universe, mainly for compatibility with the rest of this text.

The *theory of social choice* is concerned, among others, with various properties of voting schemes and with designing "good" voting schemes.[11] (After all, while a good voting system in a country doesn't yet guarantee a good government, a bad voting system can bring a country to breakdown.)

Influences and the KKL theorem. Now we will introduce several parameters of a Boolean function. One of them is

$$\mu = \mu(f) := \frac{1}{2^n} \sum_{\mathbf{u}\in\{0,1\}^n} f(\mathbf{u}),$$

the arithmetic average of all values. (We stress that the numbers $f(\mathbf{u})$ on the right-hand side are added as real numbers—we consider the range $\{0,1\}$ of f as a subset of $\mathbb{R}$.)

A very useful point of view for interpreting μ and the other parameters mentioned next is to consider $\mathbf{u}$ to be chosen from $\{0,1\}^n$ *uniformly at random*. Then $f(\mathbf{u})$ becomes a random variable, and we can see that

$$\mu = \mathbf{E}[f],$$

the *expectation* of f. (This point of view is also often used in the theory of social choice; if one doesn't know much about the preferences of the

[11] Some of the results are rather pessimistic. For example, suppose that the society wants to rank *three* possible alternatives A, B, C from the least suitable to the best. To this end, three binary decisions are taken, using three possibly different voting schemes f, g, h. The first decision is whether A is better than B, with the two possible outcomes $A > B$ or $A < B$. The second decision is whether B is better than C, and the third one is whether A is better than C.

The *Concordet paradox* is the observation that if these three decisions are taken according to the majority, $f = g = h = \mathrm{Maj}$, it can happen that the outcome is *irrational* in the sense that the majority prefers A to B, B to C, and C to A (or some other cyclic arrangement)—even though every voter's decisions are rational (i.e., based on some consistent ranking).

A natural question is, can some voting schemes f, g, h guarantee that an such an irrational outcome never occurs? *Arrow's impossibility theorem* asserts that the *only* voting schemes with this property are $f = g = h = \mathrm{Dict}_k$ for some k, i.e., all decisions taken solely by a single (rational) dictator. (More precisely, for this result to hold, one needs to assume that f, g, h have the *unanimity property*, meaning that $f(0,0,\dots,0) = 0$ and $f(1,1,\dots,1) = 1$.)

A more recent result in this direction, obtained with the help of Fourier analysis, tells us that the Concordet paradox is, in a sense, robust: if the ranking of each voter is chosen at random, then for any voting schemes that are "sufficiently different" from the dictator function, the probability of an irrational outcome is bounded from below by a positive (and non-negligible) constant.

voters, it seems reasonable to study the behavior of a voting scheme for voters behaving randomly.) With this point of view in mind, we call f **unbiased** if $\mu = \frac{1}{2}$.

Knowing the expectation $\mathbf{E}[f] = \mu$, it is natural to ask about the *variance* $\mathrm{Var}[f] = \mathbf{E}\left[f^2\right] - \mathbf{E}[f]^2$, which measures the "how much non-constant" f is. It turns out that, since the values of f are only 0s and 1s, the variance is determined by the expectation, and we easily calculate that

$$\mathrm{Var}[f] = \mu(1-\mu).$$

Next, we introduce a quantity measuring how much the decision of the kth voter influences the outcome of the vote, provided that all others vote at random.

The **influence of the kth variable** for a function $f\colon \{0,1\}^n \to \{0,1\}$ is defined as

$$I_k(f) := \mathrm{Prob}[f(\mathbf{u}+\mathbf{e}_k) \neq f(\mathbf{u})] = 2^{-n} \sum_{\mathbf{u}\in\{0,1\}^n} |f(\mathbf{u}+\mathbf{e}_k) - f(\mathbf{u})|.$$

Here the probability is with respect to a random choice of $\mathbf{u}$, and $\mathbf{u}+\mathbf{e}_k$ simply means $\mathbf{u}$ with the kth coordinate flipped (the addition in $\mathbf{u}+\mathbf{e}_k$ is meant modulo 2, and $\mathbf{e}_k$ stands for the vector with 1 at the kth position and 0s elsewhere, as usual).

We also note that $f(\mathbf{u}+\mathbf{e}_k) - f(\mathbf{u})$ always equals 0, 1, or -1, and so the sum in the definition above indeed counts the number of $\mathbf{u}$'s where $f(\mathbf{u}+\mathbf{e}_k)$ differs from $f(\mathbf{u})$, i.e., where the decision of the kth voter is a "swing vote".

The **total influence**[a] of f is the sum of the influences of all variables:

$$I(f) := \sum_{k=1}^{n} I_k(f).$$

[a] This notion is apparently very natural, and it has several alternative names in the literature: the *energy*, the *average sensitivity*, the *normalized edge boundary*, etc.

Here are several examples, where the reader is invited to work out the calculations.

- For the dictator function we have, of course, $I_k(\mathrm{Dict}_k) = 1$ and $I_i(\mathrm{Dict}_k) = 0$ for $i \neq k$.

- $I_k(\mathrm{Parity}) = 1$ for all k.

- For majority, $I_k(\mathrm{Maj})$ is of order $n^{-1/2}$ (and independent of k, of course).

- We introduce yet another function, called the **tribes function**. In terms of a voting scheme, the n voters are partitioned into groups, referred to as tribes, each of size s, where s is a suitable parameter, to be determined later. (More precisely, if n is not divisible by s, then some tribes may have size s and some size $s-1$.) First a decision is taken in each tribe separately: the outcome in a tribe is 1 exactly if everyone in the tribe votes 1, and the final outcome is 1 if at least one of the tribes decides for 1.

  The defining logical formula is thus (for n divisible by s and $t := n/s$)

  $$\begin{aligned}\mathrm{Tribes}(u_1,\ldots,u_n) \quad = \quad & (u_1 \wedge u_2 \wedge \cdots \wedge u_s) \vee (u_{s+1} \wedge \cdots \wedge u_{2s}) \vee \cdots \\ & \vee (u_{(t-1)s+1} \wedge \cdots \wedge u_n).\end{aligned}$$

  The tribe size s is determined so that Tribes is approximately unbiased, i.e., so that $\mu(\mathrm{Tribes})$ is as close to $\frac{1}{2}$ as possible. Calculation shows that s is about $\log_2 n - \log_2 \log_2 n$, and then we obtain

  $$I_k(\mathrm{Tribes}) = \Theta\left(\frac{\log n}{n}\right).$$

A possible way of postulating that f should be "far" from a dictator function is to require that none of the influences $I_k(f)$ be too large. From this point of view, it is natural to ask, how small can the maximum influence be (for an unbiased function).

The somewhat surprising answer is given by the following fundamental theorem, which implies that the tribes function has the asymptotically smallest possible maximum influence.

3.9.2 Theorem (The KKL theorem; Kahn, Kalai, and Linial). *For every unbiased Boolean function f on $\{0,1\}^n$ we have*

$$\max_k I_k(f) = \Omega\left(\frac{\log n}{n}\right).$$

More generally, for an arbitrary, possibly biased, f we have

$$\max_k I_k(f) \geq c\mu(1-\mu)\frac{\log n}{n},$$

with a suitable constant $c > 0$.

This is a very important result: historically, it introduced Fourier-analytic methods into theoretical computer science, and it has a number of interesting applications.

As we will see in the course of the proof, the total influence $I(f)$ is bounded from below by $4\mu(1-\mu)$ (and this is a much easier fact than the KKL theorem). Thus, the average of the $I_k(f)$ is at least of order $\mu(1-\mu)\frac{1}{n}$. The point of the KKL theorem is in the extra $\log n$ factor for the *maximum* influence: for example, the $\log n$ lower bound for embedding the edit distance (Theorem 3.9.1) "comes from" exactly this $\log n$ in the KKL theorem.

The KKL theorem can usefully be viewed as a "local/global" result. Namely, the influence $I_k(f)$ measures the average "speed of local change" of f in the direction of the kth coordinate (and it can be regarded a discrete analog of a norm of the partial derivative $\partial f/\partial u_k$ for functions on $\mathbb{R}^n$), On the other hand, $\mu(1-\mu) = \operatorname{Var}[f]$ measures the "amount of change" of f globally, and the theorem gives a bound on the "global change" $\operatorname{Var}[f]$ in terms of the "local changes" $I_k(f)$. In this sense, the KKL theorem belongs to a huge family of local/global theorems in analysis with a similar philosophy.

We will present a proof of the KKL theorem in Appendix A. Actually, we will prove the following, slightly more general (and more technical) statement, which is what we need for the edit distance lower bound.

3.9.3 Theorem. *Let $f\colon \{0,1\}^n \to \{0,1\}$ be a Boolean function, and let $\delta := \max_k I_k(f)$. Then the total influence of f satisfies*

$$I(f) \geq c\mu(1-\mu)\log\frac{1}{\delta}$$

(assuming $\delta > 0$, of course, which is the same as assuming f nonconstant).

The lower bound for edit distance: proof of Theorem 3.9.1. Let $V := \{0,1\}^n$ be the point set of the considered metric space. The general approach is "as usual" in the method of inequalities: we compare the sum of distances over all pairs of points with the sum over suitably selected set of pairs, we get an inequality for the original space (V, ed) and an inequality in opposite direction in ℓ_1, and comparing them yields the distortion bound.

The proof relies only on the following three properties of the edit distance:

(P1) (Replacement is cheap) Strings at Hamming distance 1 also have edit distance 1; i.e., $\mathrm{ed}(\mathbf{u}, \mathbf{u}+\mathbf{e}_k) = 1$.

(P2) (Cyclic shift is cheap) We have $\mathrm{ed}(\mathbf{u}, \sigma(\mathbf{u})) \leq 2$, where $\sigma(\mathbf{u})$ denotes the *left cyclic shift* of $\mathbf{u}$, i.e., $\sigma(u_1u_2\cdots u_n) = u_2u_3\cdots u_nu_1$.

(P3) (Large typical distance) For every $\mathbf{u} \in V$, no more than half of the strings $\mathbf{v} \in V$ satisfy $\mathrm{ed}(\mathbf{u}, \mathbf{v}) \leq \alpha n$, where α is some positive constant.

The proof of (P3) is left as a slightly challenging exercise. One needs to estimate the number of different strings that can be obtained from $\mathbf{u}$ by a sequence of at most r edit operations. The main observation is that it is enough to consider suitable "canonical" sequences; for example, we may assume that all deletions precede all insertions.

Let $F := V \times V$ be the set of all pairs (note that we chose to work with *ordered* pairs, unlike in some of the previous sections—this has slight formal advantages here). Then (P3) implies that the average edit distance in F is $\Omega(n)$.

Next, we will choose a suitable collection of "selected pairs", which will reflect properties (P1) and (P2): first, the (directed) edges of the Hamming cube

$$E_{\mathrm{Hamm}} := \Big\{(\mathbf{u}, \mathbf{u}+\mathbf{e}_k) : \mathbf{u} \in V, k = 1, 2, \ldots, n\Big\},$$

and second, the *shift edges*

$$E_{\mathrm{shift}} := \Big\{(\mathbf{u}, \sigma(\mathbf{u})) : \mathbf{u} \in V\Big\}.$$

We have $|E_{\mathrm{Hamm}}| = n2^n$ and $|E_{\mathrm{shift}}| = 2^n$.

A fine point of the proof is that the shift edges need to be counted with weight n-times larger than the Hamming edges. Using (P3) for the

sum over F, and (P1) and (P2) for the sums over E_{Hamm} and E_{shift}, respectively, we obtain the inequality (with some constant $c_0 > 0$)

$$\frac{1}{2^{2n}} \sum_F \mathrm{ed}(\mathbf{u},\mathbf{v}) \geq c_0 \Big[\frac{1}{2^n} \sum_{E_{\mathrm{Hamm}}} \mathrm{ed}(\mathbf{u},\mathbf{v}) + \frac{n}{2^n} \sum_{E_{\mathrm{shift}}} \mathrm{ed}(\mathbf{u},\mathbf{v}) \Big].$$

To prove Theorem 3.9.1, it thus suffices to establish the counterpart in ℓ_1, i.e., the following "Poincaré inequality" for every mapping $f\colon V \to \ell_1$ and some constant C:

$$\frac{1}{2^{2n}} \sum_F \|f(\mathbf{u}) - f(\mathbf{v})\|_1 \;\leq\; \frac{C}{\log n} \Big[\frac{1}{2^n} \sum_{E_{\mathrm{Hamm}}} \|f(\mathbf{u}) - f(\mathbf{v})\|_1 + \frac{n}{2^n} \sum_{E_{\mathrm{shift}}} \|f(\mathbf{u}) - f(\mathbf{v})\|_1 \Big]. \quad (3.11)$$

Now by the standard consideration as in Section 3.6, using the fact that every ℓ_1 metric is a nonnegative linear combination of cut metrics, it suffices to prove a linear inequality such as this one for functions $V \to \{0,1\}$. So we are back to the realm of Boolean functions, the home of the KKL theorem.

For a Boolean f, two of the three sums in (3.11) can be interpreted using familiar parameters. First, $\sum_F |f(\mathbf{u}) - f(\mathbf{v})|$ counts the number of pairs $(\mathbf{u},\mathbf{v})$ with $f(\mathbf{u}) = 0$ and $f(\mathbf{v}) = 1$ or vice versa, and so we obtain

$$2^{-2n} \sum_F |f(\mathbf{u}) - f(\mathbf{v})| = 2\mu(1-\mu), \quad \mu = \mathbf{E}[f].$$

The average over the Hamming edges is identified as the total influence:

$$2^{-n} \sum_{E_{\mathrm{Hamm}}} |f(\mathbf{u}) - f(\mathbf{v})| = I(f).$$

Finally, since we don't recognize the sum over the shift edges as something familiar, we at least assign it a symbol, setting

$$\mathrm{SSh}(f) := 2^{-n} \sum_{E_{\mathrm{shift}}} |f(\mathbf{u}) - f(\mathbf{v})| = 2^{-n} \sum_{\mathbf{u}} |f(\mathbf{u}) - f(\sigma(\mathbf{u}))|.$$

With this new notation, we want to prove that

$$\mu(1-\mu) \leq O((\log n)^{-1})[\, I(f) + n \cdot \mathrm{SSh}(f)].$$

For contradiction, let us assume that this fails. Then both $I(f)$ and $\mathrm{SSh}(f)$ have to be small; quantitatively

(C1) $I(f) < c_1 \mu(1-\mu)\log n$, with c_1 a constant as small as we wish, and

(C2) $\mathrm{SSh}(f) < c_1 \mu(1-\mu)\frac{\log n}{n} \leq \frac{\log n}{n}$.

The inequality (C1) may hold for some functions f, since the smallest possible $I(f)$ is of order $\mu(1-\mu)$. But then the KKL theorem tells us that at least *some* of the influences $I_k(f)$ must be quite large. We will use (C2), the lower bound on $\mathrm{SSh}(f)$, to infer that *many* of the $I_k(f)$ must be large, and this will show that the total influence does violate (C1) after all—a contradiction. The tool for doing this is the next lemma.

3.9.4 Lemma (Smoothness of the influences). *For every Boolean function f and every k, we have*

$$|I_{k+1}(f) - I_k(f)| \leq 2\,\mathrm{SSh}(f).$$

Assuming this lemma for a moment, we finish the proof of (3.11) as follows. Let us set $\delta := n^{-1/3}$. If the constant c_1 in (C1) is sufficiently small, we have $I(f) < c\mu(1-\mu)\log\frac{1}{\delta}$, and so by Theorem 3.9.3, there is some k with $I_k(f) > \delta$.

Now using Lemma 3.9.4 and (C2), we get that for all i with $|i| \leq n^{1/2}$, we have

$$I_{k+i}(f) \geq \delta - 2|i| \cdot \mathrm{SSh}(f) \geq n^{-1/3} - 2n^{1/2} \cdot \frac{\log n}{n} \geq \tfrac{1}{2} n^{-1/3}$$

for n sufficiently large. (Here $k+i$ is to be interpreted with wrapping around.) But then

$$I(f) \geq \sum_{-n^{1/2} \leq i \leq n^{1/2}} I_{k+i}(f) \geq n^{1/2} \cdot n^{-1/3} = n^{1/6},$$

which contradicts (C1) and finishes the proof of Theorem 3.9.1.

Proof of Lemma 3.9.4. This is a straightforward application of the triangle inequality.

$$\begin{aligned} I_k(f) &= 2^{-n} \sum_{\mathbf{u}} |f(\mathbf{u} + \mathbf{e}_k) - f(\mathbf{u})| \\ &= 2^{-n} \sum_{\mathbf{v}} |f(\sigma(\mathbf{v} + \mathbf{e}_{k+1})) - f(\sigma(\mathbf{v}))| \\ &\leq 2^{-n} \sum_{\mathbf{v}} \Big(|f(\sigma(\mathbf{v} + \mathbf{e}_{k+1})) - f(\mathbf{v} + \mathbf{e}_{k+1})| \\ &\qquad + |f(\mathbf{v} + \mathbf{e}_{k+1}) - f(\mathbf{v})| + |f(\mathbf{v}) - f(\sigma(\mathbf{v}))| \Big) \\ &= \mathrm{SSh}(f) + I_{k+1}(f) + \mathrm{SSh}(f). \end{aligned}$$

□

3.10 Impossibility of flattening in ℓ_1

Every n-point Euclidean metric space can be embedded in $\ell_2^{O(\log n)}$ with distortion close to 1 according to the Johnson–Lindenstrauss lemma, and this fact is extremely useful for dealing with Euclidean metrics.

We already know, by a counting argument, that no analogous statement holds for embedding metrics in ℓ_∞. For instance, there are n-point metrics that can't be embedded in ℓ_∞^{cn}, for a suitable constant $c > 0$, with distortion smaller than 2.9.

The following theorem excludes an analog of the Johnson–Lindenstrauss lemma for ℓ_1 metrics as well. Or rather, it shows that if there is any analog at all, it can be only quite weak.

3.10.1 Theorem. *For all sufficiently large n there exists an n-point ℓ_1 metric space M such that whenever M can be D-embedded in ℓ_1^k for some $D > 1$, we have $k \ge n^{0.02/D^2}$.*

Two particular cases are worth mentioning. First, for every fixed distortion D, the required dimension is at least a small but fixed power of n. Second, if we want dimension $O(\log n)$, the required distortion is at least $\Omega(\sqrt{\log n/\log\log n})$. Interestingly, the latter bound is almost tight: It is known that one can embed every n-point ℓ_1 metric in ℓ_2 with distortion $O(\sqrt{\log n}\log\log n)$ (this is a difficult result), then we can apply the Johnson–Lindenstrauss lemma to the image of this embedding, and finally embed $\ell_2^{O(\log n)}$ back in $\ell_1^{O(\log n)}$ with a negligible distortion.

The lower bound for the dimension for embeddings in ℓ_∞ was proved by counting—showing that there are more essentially different n-point spaces that essentially different n-point subsets of ℓ_∞^k. This kind of approach can't work for the ℓ_1 case, since it is known that if d is an ℓ_1 metric, then $\sqrt{d}$ is an ℓ_2 metric. Thus, if we had many ℓ_1 metrics on a given n-point set, every two differing by a factor of at least D on some pair of points, then there are the same number of Euclidean metrics on these points, every two differing by at least $\sqrt{D}$ on some pair—but we know that ℓ_2 metrics *can* be flattened.

Here is an outline of the forthcoming proof. We want to construct a space that embeds in ℓ_1 but needs a large distortion to embed in ℓ_1^k.

- We choose p a little larger than 1, namely, $p := 1 + \frac{1}{\ln k}$, and we observe that the "low-dimensional spaces" ℓ_1^k and ℓ_p^k are almost the same—the identity map is an $O(1)$-embedding (Lemma 3.10.2 below).
- Then we show that the "high-dimensional" spaces ℓ_1 and ℓ_p differ substantially. Namely, we exhibit a space X that embeds well in ℓ_1 (for technical reasons, we won't insist on an isometric embedding, but we'll be satisfied with distortion 2; see Lemma 3.10.3), but requires large distortion for embedding in ℓ_p (Lemma 3.10.4).

It follows that such an X doesn't embed well in ℓ_1^k, for if it did, it would also embed well in ℓ_p^k.

Let us remark that more recently a different, and in some respect simpler, proof was found by O. Regev [Entropy-based bounds on dimension reduction in L_1, arXiv:1108.1283; to appear in *Isr. J. Math.*].

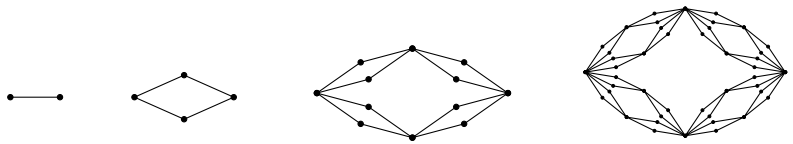
3.10.2 Lemma. *For $k > 1$ and $p := 1 + \frac{1}{\ln k}$, the identity map $\ell_1^k \to \ell_p^k$ has distortion at most* 3.

Proof. This is a very standard calculation with a slightly nonstandard choice of parameters. First, for $p_1 \le p_2$, we have $\|\mathbf{x}\|_{p_1} \ge \|\mathbf{x}\|_{p_2}$, and thus the identity map as in the lemma is nonexpanding. For the contraction Hölder's inequality yields

$$\begin{aligned}\|\mathbf{x}\|_1 &= \sum_{i=1}^k 1\cdot|x_i| \le k^{1-1/p}\|\mathbf{x}\|_p \\ &= e^{(\ln k)(p-1)/p}\|\mathbf{x}\|_p = e^{(\ln k)/(1+\ln k)}\|\mathbf{x}\|_p \le 3\|\mathbf{x}\|_p.\end{aligned}$$

□

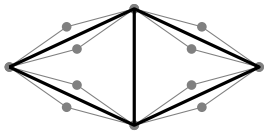
The recursive diamond graph. The space X in the above outline is a generally interesting example, which was invented for different purposes. It is given by the shortest-path metric on a graph G_m, where $G_0, G_1, G_2, \ldots$ is the following recursively constructed sequence:



G_0 G_1 G_2 G_3

Starting with G_0 a single edge, G_{i+1} is obtained from G_i by replacing each edge $\{u,v\}$ by a 4-cycle u,a,v,b, where a and b are new vertices. The pair $\{a,b\}$ is called the *anti-edge* corresponding to the edge $\{u,v\}$. Let us set $E_i := E(G_i)$, and let A_{i+1} be the set of the anti-edges corresponding to the edges of E_i, $i = 0,1,\ldots$.

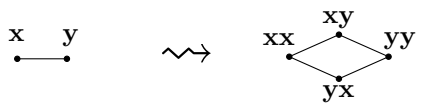
Since the vertex sets of the G_i form an increasing sequence, $V(G_0) \subset V(G_1) \subset \cdots$, we can regard $E_0, E_1, \ldots, E_m$ and $A_1, \ldots, A_m$ as sets of pairs of vertices of G_m. For example, the next picture shows E_1 and A_1 in G_2:



In G_m, the pairs in E_i and in A_{i+1} have distance 2^{m-i}.

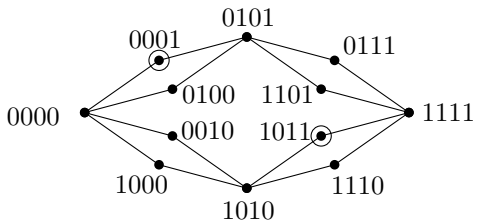
3.10.3 Lemma. *Every G_m embeds in ℓ_1 with distortion at most 2.*

Sketch of proof. The embedding is very simple to describe. Each vertex of G_m is assigned a point $\mathbf{x} \in \{0,1\}^{2^m}$. We start with assigning 0 and 1 to the two vertices of G_0, and when G_{i+1} is constructed from G_i, the embedding for G_{i+1} is obtained as follows:



(**xy** denotes the concatenation of **x** and **y**).

It is easily checked by induction that this embedding preserves the distance for all pairs in $E_0 \cup E_1 \cup \cdots$ and in $A_1 \cup A_2 \cup \cdots$ exactly. Consequently, the embedding is nonexpanding. However, some distances do get contracted; e.g., the two circled vertices in G_2



have distance 4 but their points distance only 2.

We thus need to argue that this contraction is never larger than 2. Given vertices u and v, we find a pair $\{u',v'\}$ in some E_i or A_i with u' close to u and v' close to v and we use the triangle inequality. This is the part which we leave to the reader. □

Let us mention that the embedding in the above lemma is not optimal—another embedding is known with distortion only $\frac{4}{3}$.

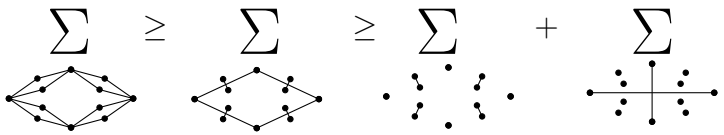
Finally, here is the promised nonembeddability in ℓ_p.

3.10.4 Lemma. *Any embedding of G_m in ℓ_p, $1 < p \le 2$, has distortion at least $\sqrt{1+(p-1)m}$.*

Proof. First we present the proof for the case $p = 2$, where it becomes an exercise for the method with inequalities we have seen for the Hamming cube and for expander graphs.

Let $E := E_m = E(G_m)$ and $F := E_0 \cup A_1 \cup A_2 \cup \cdots \cup A_m$. With d_{G_m} denoting the shortest-path metric of G_m, we have $\sum_E d_{G_m}(u,v)^2 = |E_m| = 4^m$ and $\sum_F d_{G_m}(u,v)^2 = 4^m + \sum_{i=1}^m |A_i| 4^{m-i+1} = 4^m + \sum_{i=1}^m 4^{i-1}4^{m-i+1} = (m+1)4^m$. So the ratio of the sums over F and over E is $m+1$.

Next, let us consider an arbitrary map $f\colon V(G_m) \to \ell_2$, and let $S_E := \sum_E \|f(u)-f(v)\|_2^2$. Applying the short-diagonals lemma to each of the small quadrilaterals in G_m, we get that $S_E \ge \sum_{A_m \cup E_{m-1}} \|f(u)-f(v)\|_2^2$. Next, we keep the sum over A_m and we bound the sum over E_{m-1} from below using the short-diagonals lemma, and so on, as in the picture:



In this way, we arrive at $\sum_F \|f(u)-f(v)\|_2^2 \le \sum_E \|f(u)-f(v)\|_2^2$, and so f has distortion at least $\sqrt{m+1}$.

For the case of an arbitrary $p \in (1,2]$ the calculation remains very similar, but we need the following result as a replacement for the Euclidean short-diagonals lemma.

3.10.5 Lemma (Short-diagonals lemma for ℓ_p). *For every four points $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4 \in \ell_p$ we have*

$$\|\mathbf{x}_1-\mathbf{x}_3\|_p^2 + (p-1)\|\mathbf{x}_2-\mathbf{x}_4\|_p^2 \le \|\mathbf{x}_1-\mathbf{x}_2\|_p^2 + \|\mathbf{x}_2-\mathbf{x}_3\|_p^2 + \|\mathbf{x}_3-\mathbf{x}_4\|_p^2 + \|\mathbf{x}_4-\mathbf{x}_1\|_p^2.$$

This lemma is a subtle statement, optimal in quite a strong sense, and we defer the proof to Appendix B. Here we just note that, unlike the inequalities used earlier, the norm doesn't appear with pth powers but rather with *squares*. Hence it is not enough to prove the lemma for the 1-dimensional case.

Given this short-diagonals lemma, we consider an arbitrary mapping $f: V(G_m) \to \ell_p$ and derive the inequality

$$\|f(s)-f(t)\|_p^2+(p-1)\sum_{A_1\cup A_2\cup\cdots\cup A_m}\|f(u)-f(v)\|_p^2 \le \sum_E \|f(u)-f(v)\|_p^2,$$

where s and t are the vertices of the single pair in E_0. We note that the left-hand side is a sum of squared distances over F but a *weighted* sum, where the pair in E_0 has weight 1 and the rest weight $p-1$. Comparing with the corresponding weighted sums for the distances in G_m, Lemma 3.10.4 follows. □

Proof of Theorem 3.10.1. We follow the outline. Let $f: V(G_m) \to \ell_1$ be a 2-embedding and let $X := f(V(G_m))$. Assuming that $(X, \|.\|_1)$ can be D-embedded in ℓ_1^k, we have the following chain of embeddings:

$$G_m \xrightarrow{2} X \xrightarrow{D} \ell_1^k \xrightarrow{3} \ell_p^k.$$

The composition of these embedding is a $6D$-embedding of G_m in ℓ_p, and so $6D \ge \sqrt{1+(p-1)m}$ with $p = 1+\frac{1}{\ln k}$. It remains to note that $n = |V(G_m)| \le 4^m$ for all $m \ge 1$. The theorem then follows by a direct calculation. □

3.11 Exercises

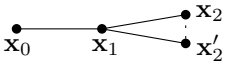
1. Show that there exists an n-point set $X \subset S^2$ such that every embedding of $(X, \|.\|_2)$ into $(\mathbb{R}^2, \|.\|_2)$ has distortion $\Omega(\sqrt{n})$. Use the Borsuk–Ulam theorem.

2.** Let P_n be the metric space $\{0, 1, \ldots, n\}$ with the metric inherited from $\mathbb{R}$ (in other words, a path of length n). Prove the following Ramsey-type result: For every $D > 1$ and every $\varepsilon > 0$ there exists an $n = n(D, \varepsilon)$ such that whenever $f: P_n \to (Z, d_Z)$ is a D-embedding of P_n into some metric space, there are $a < b < c$, $b = \frac{a+c}{2}$, such that f restricted to the subspace $\{a, b, c\}$ of P_n is a $(1+\varepsilon)$-embedding.

In other words, if a sufficiently long path is D-embedded, then it contains a scaled copy of a path of length 2 embeded with distortion close to 1.

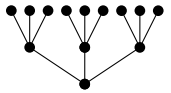
3. (Lower bound for embedding trees into ℓ_2.)

   (a)* Show that for every $\varepsilon > 0$ there exists $\delta > 0$ with the following property. Let $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2' \in \ell_2$ be points such that $\|\mathbf{x}_0 - \mathbf{x}_1\|_2, \|\mathbf{x}_1 - \mathbf{x}_2\|_2, \|\mathbf{x}_1 - \mathbf{x}_2'\|_2 \in [1, 1+\delta]$ and $\|\mathbf{x}_0 - \mathbf{x}_2\|_2, \|\mathbf{x}_0 - \mathbf{x}_2'\|_2 \in [2, 2+\delta]$ (so all the distances are almost like the graph distances in the following tree, except possibly for the one marked by a dotted line).



   Then $\|x_2 - x_2'\| \le \varepsilon$; that is, the remaining distance must be very short.

   (b)* Let $T_{k,m}$ denote the complete k-ary tree of height m; the following picture shows $T_{3,2}$:



   Show that for every r there exists k such that whenever the leaves of $T_{k,m}$ are colored by r colors, there is a subtree of $T_{k,m}$ isomorphic to $T_{2,m}$ with all leaves having the same color.

   (c)** Use (a), (b), and Exercise 2 to prove that for every $D > 1$ there exist m and k such that the tree $T_{k,m}$ considered as a metric space with the shortest-path metric cannot be D-embedded into ℓ_2.

4.* Show that for every r and ℓ there exists an r-regular graph $G(r, \ell)$ of girth at least ℓ (give an explicit inductive construction).

5. (a) Prove that if G is a graph whose average vertex degree is d, then G contains a subgraph with minimum vertex degree at least $d/2$.

   (b) Show that every graph G has a bipartite subgraph H that contains at least half the edges of G.

   (c)* Use (a) and (b) to prove that if $G = (V, E)$ is an n-vertex graph that contains no cycle of length ℓ or less, then $|E| = O(n^{1+1/\lfloor \ell/2 \rfloor})$.

6. (a) Show that the squared Euclidean distances between the points of a set $X \subset \mathbb{R}^k$ form a metric of negative type if and only if no triple in X forms an obtuse angle.

   (b) Show that every finite path (as a graph metric) is a metric of negative type, by giving an explicit embedding.

7.* Calculate the asymptotics of the influence $I_k(\mathrm{Maj})$ of the kth variable in the majority function of n variables, for $n \to \infty$.

8. Consider the Boolean function Tribes(.) of n variables, with tribes of size s.

   (a)* Estimate the value of $s = s(n)$ that makes the function as balanced as possible.

   (b) For that value of s, estimate $I_k(\mathrm{Tribes})$ asymptotically as $n \to \infty$.

9.* Given a string $\mathbf{u} \in \{0,1\}^n$, prove that at most half of the strings $\mathbf{v} \in \{0,1\}^n$ are at edit distance at most cn from $\mathbf{u}$, for a suitable constant $c > 0$.

10.** The *block edit distance* of two strings $\mathbf{u}, \mathbf{v} \in \{0,1\}^n$ is defined in a way similar to the edit distance, but in addition to the usual edit operations, also *block operations* are allowed: one can swap two arbitrarily large contiguous blocks as a single operation.

   The goal is to prove that that two randomly chosen strings $\mathbf{u}, \mathbf{v} \in \{0,1\}^n$ have block edit distance only $O(n/\log n)$ with probability close to 1.

   (a) Split both strings into blocks of length $k := c \log n$ for a small constant $c > 0$, and regard each block as a "supercharacter" in an alphabet of size 2^k. Let $n_{\mathbf{u}}(a)$ be the number of occurrences of a supercharacter a in $\mathbf{u}$, and similarly for $\mathbf{v}$. Prove that, with probability close to 1, $\sum_{a \in \{0,1\}^k} |n_{\mathbf{u}}(a) - n_{\mathbf{v}}(a)|$ is quite small compared to n.

   (b) Prove the main claim, i.e., that random $\mathbf{u}$ and $\mathbf{v}$ block edit distance only $O(n/\log n)$ with probability close to 1.

11. Verify that the embedding of the recursive diamond graphs G_m in ℓ_1, as described in the proof of Lemma 3.10.3, has distortion at most 2.

4

Constructing embeddings

In this chapter we present techniques for constructing low-distortion embeddings. The highlight is Bourgain's theorem, stating that every n-point metric space $O(\log n)$-embeds in ℓ_2.

Several embedding methods are known, with many variants and additional tricks. While the main ideas are reasonably simple, the strongest and most interesting embedding results are often rather complicated. So we'll illustrate some of the ideas only on toy examples.

4.1 Bounding the dimension for a given distortion

In this section we prove an upper bound almost matching the lower bounds from Section 3.3. We also obtain a weaker version of Bourgain's theorem, showing that every n-point metric space embeds in ℓ_2 with distortion $O(\log^2 n)$ (the tight bound is $O(\log n)$). The proof can also be taken as an introduction to the proof of Bourgain's theorem itself, since it exhibits the main ingredients in a simpler form.

Our concrete goal here is the following.

4.1.1 Theorem. *Let $D = 2q-1 \geq 3$ be an odd integer and let (V, d_V) be an n-point metric space. Then there is a D-embedding of V into ℓ_∞^k with*

$$k = O(qn^{1/q} \ln n).$$

For example, every n-point metric space can be 3-embedded in ℓ_∞^k with $k = O(\sqrt{n}\log n)$, while Proposition 3.3.1 tells us that there are spaces whose 3-embedding in ℓ_∞^k requires $k = \Omega(\sqrt{n}\,)$ (and for distortion

$D < 3$, the required dimension is $\Omega(n)$). Similarly matching results are also known for $D = 5, 7, 11$.

For $q = \lfloor \log n \rfloor$, the theorem implies that n-point metric spaces can be $O(\log n)$-embedded in ℓ_∞^k with $k = O(\log^2 n)$. Since the latter space $O(\log n)$-embeds in ℓ_2^k (by the identity mapping), we obtain the promised weaker version of Bourgain's theorem:

4.1.2 Corollary. *Every n-point metric space $O(\log^2 n)$-embeds in ℓ_2.* □

Fréchet embeddings. We begin with a somewhat informal introduction to the proof. To specify a mapping f of a metric space (V, d_V) into ℓ_p^k is the same as defining k functions $f_1, \dots, f_k : V \to \mathbb{R}$, the coordinates of the embedded points.

Let us call an embedding $f : V \to \ell_p^k$ a *Fréchet embedding* if each of the coordinates f_i is the distance from some set $A_i \subseteq V$; that is, if

$$f_i(v) = d_V(v, A_i) \text{ for all } v \in V$$

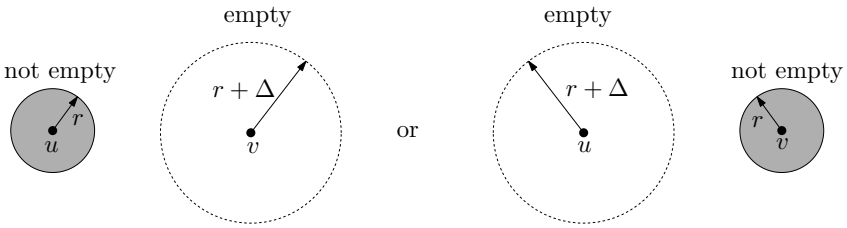
(more generally, $f_i(v)$ might also be $\alpha_i d_V(v, A_i)$ for some numeric factor α_i).

We saw a particularly simple Fréchet embedding in Section 1.5. There we had one coordinate for every point of V, and A_i consisted of the ith point of V. In the forthcoming proof, the A_i are going to be a suitable random subsets of V.

A pleasant property of a Fréchet embedding is that each f_i is automatically nonexpanding (since $|f_i(u) - f_i(v)| = |d_V(u, A_i) - d_V(v, A_i)| \leq d_V(u, v)$).

From now on, we focus on Fréchet embeddings in the target space ℓ_∞^k. Then, since each f_i is nonexpanding, f is nonexpanding as well. If f should have distortion at most D, then we need that for every pair $\{u, v\}$ of points of V, there is a coordinate $i = i(u, v)$ that "takes care" of the pair, i.e., such that $|f_i(u) - f_i(v)| \geq \frac{1}{D} d_V(u, v)$. So we "only" need to choose a suitable collection of the A_i that take care of all pairs $\{u, v\}$.

Let us consider two points $u, v \in V$. What are the sets A such that $|d_V(u, A) - d_V(v, A)| \geq \Delta$, for a given real $\Delta > 0$? For some $r \geq 0$, such an A must intersect the closed r-ball around u and avoid the open $(r+\Delta)$-ball around v, or conversely (with the roles of u and v interchanged):



If it so happens that the closed r-ball around u doesn't contain many fewer points of V than the open $(r+\Delta)$-ball around v, then a random A with a suitable density has a reasonable chance to satisfy $|d_V(u,A)-d_V(v,A)| \geq \Delta$.

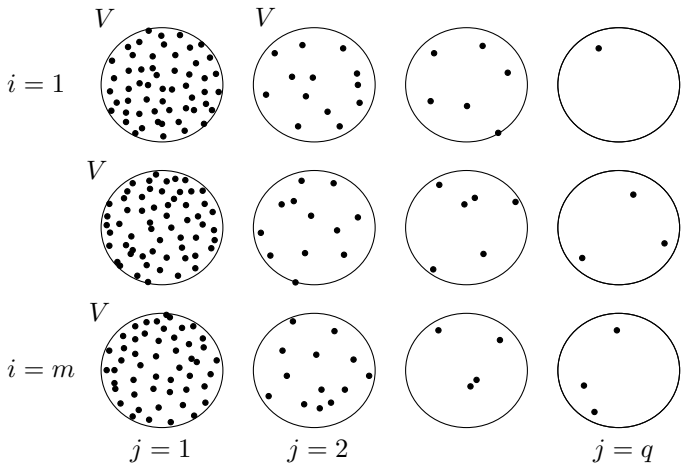
Generally we have no control over the distribution of points around u and around v, but by considering several suitable balls simultaneously, we will be able to find a good pair of balls. We also do not know the right density needed for the sample to work, but since we have many coordinates, we will be able to take samples of essentially all possible densities.

Now we can begin with the formal proof.

Proof of Theorem 4.1.1. We define an auxiliary parameter $p := n^{-1/q}$, and for $j = 1, 2, \ldots, q$, we introduce the probabilities $p_j := \min(\frac{1}{2}, p^j)$. Further, let $m := \lceil 24 n^{1/q} \ln n \rceil$, and let $k := mq$. We construct a Fréchet embedding $f\colon V \to \ell_\infty^k$.

It is convenient to divide the coordinates of ℓ_∞^k into q blocks by m coordinates. Thus, the coordinates, and the sets defining the Fréchet embedding, have double indices: $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, q$.

Each set $A_{ij} \subseteq V$ is a random sample with density p_j: Each point $v \in V$ has probability p_j of being included into A_{ij}, and these events are mutually independent. The choices of the A_{ij}, too, are independent for distinct indices i and j. Here is a schematic illustration of the sampling:



Then the Fréchet embedding $f\colon V \to \ell_\infty^k$ is given as above, by

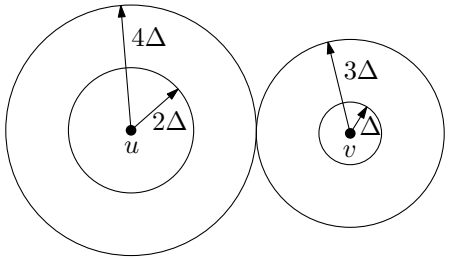
$$f_{ij}(v) = d_V(v, A_{ij}), \quad i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, q.$$

We claim that with a positive probability, f is a D-embedding. We have already noted that it is nonexpanding, and so it suffices to show that, with a positive probability, for every pair $\{u, v\}$ there is a good set A_{ij}, where we define A_{ij} to be *good for* $\{u, v\}$ if $|d_V(u, A_{ij}) - d_V(v, A_{ij})| \geq \frac{1}{D}\, d_V(u, v)$.

4.1.3 Lemma. *Let u, v be two distinct points of V. Then there exists an index $j \in \{1, 2, \ldots, q\}$ such that if the set A_{ij} is chosen randomly as above, then it is good for $\{u, v\}$ with probability at least $\frac{p}{12}$.*

First, assuming this lemma, we finish the proof of the theorem. Let us consider a fixed pair $\{u, v\}$ and select the appropriate index j as in the lemma. Then the probability that none of the sets $A_{1j}, A_{2j}, \ldots, A_{mj}$ is good for $\{u, v\}$ is at most $(1 - \frac{p}{12})^m \leq e^{-pm/12} \leq n^{-2}$. Since there are $\binom{n}{2} < n^2$ pairs $\{u, v\}$, the probability that we fail to choose a good set for any of the pairs is smaller than 1. □

Proof of Lemma 4.1.3. Let us set $\Delta := \frac{1}{D}\, d_V(u, v)$. Let $B_0 = \{u\}$, let B_1 be the (closed) Δ-ball around v, let B_2 be the (closed) 2Δ-ball around u,..., finishing with B_q, which is a $q\Delta$-ball around u (if q is even) or around v (if q is odd). The parameters are chosen so that the radii of B_{q-1} and B_q add to $d_V(u, v)$; that is, the last two balls just touch (recall that $D = 2q-1$):



We want to find balls B_t and B_{t+1} such that B_{t+1} doesn't have too many more points than B_t. More precisely, letting $n_t := |B_t|$ be the number of points in B_t, we want to select indices j and t such that

$$n_t \ge n^{(j-1)/q} \quad \text{and} \quad n_{t+1} \le n^{j/q}. \tag{4.1}$$

If there exists t with $n_t \ge n_{t+1}$, we can use that t (and the appropriate j). Otherwise, we have $n_0 = 1 < n_1 < \cdots < n_q$. We consider the q intervals $I_1, I_2, \ldots, I_q$, where $I_j = \bigl[n^{(j-1)/q}, n^{j/q}\bigr]$. By the pigeonhole principle, there exists t such that n_t and n_{t+1} lie in the same interval I_j, and then (4.1) holds for this j and t. In this way, we have selected the index j whose existence is claimed in the lemma, and the corresponding index t.

Let E_1 be the event "$A_{ij} \cap B_t \ne \emptyset$" and E_2 the event "$A_{ij} \cap B_{t+1}^\circ = \emptyset$", where B_{t+1}° denotes the interior of B_{t+1}. If both E_1 and E_2 occur, then A_{ij} is good for $\{u, v\}$.

Since $B_t \cap B_{t+1}^\circ = \emptyset$ the events E_1 and E_2 are independent. We estimate

$$\mathrm{Prob}[E_1] = 1 - \mathrm{Prob}[A_{ij} \cap B_t = \emptyset] = 1 - (1-p_j)^{n_t} \ge 1 - e^{-p_j n_t}.$$

Using (4.1), we have $p_j n_t \ge p_j n^{(j-1)/q} = p_j p^{-j+1} = \min(\frac{1}{2}, p^j) p^{-j+1} \ge \min(\frac{1}{2}, p)$. For $p \ge \frac{1}{2}$, we get $\mathrm{Prob}[E_1] \ge 1 - e^{-1/2} > \frac{1}{3} \ge \frac{p}{3}$, while for $p < \frac{1}{2}$, we have $\mathrm{Prob}[E_1] \ge 1 - e^{-p}$, and a bit of calculus verifies that the last expression is well above $\frac{p}{3}$ for all $p \in [0, \frac{1}{2})$.

Further,

$$\mathrm{Prob}[E_2] \ge (1-p_j)^{n_{t+1}} \ge (1-p_j)^{n^{j/q}} \ge (1-p_j)^{1/p_j} \ge \tfrac{1}{4}$$

(since $p_j \le \frac{1}{2}$). Thus $\mathrm{Prob}[E_1 \cap E_2] \ge \frac{p}{12}$, which proves the lemma. □

4.2 Bourgain's theorem

By a method similar to the one shown in the previous section, one can also prove a tight upper bound on Euclidean embeddings; the method was actually invented for this problem.

4.2.1 Theorem (Bourgain's theorem). *Every n-point metric space (V, d_V) can be embedded in ℓ_2 with distortion at most $O(\log n)$.*

The overall strategy of the embedding is similar to the embedding into ℓ_∞^k in the proof of Theorem 4.1.1. We again construct a Fréchet embedding: The coordinates in ℓ_2^k are given by distances to suitable random subsets. The situation is slightly more complicated than in the previous section, since for embedding into ℓ_∞^k, it was enough to exhibit one coordinate "taking care" of each pair, whereas for the Euclidean embedding, many of the coordinates will contribute significantly to every pair. Here is the appropriate analogue of Lemma 4.1.3.

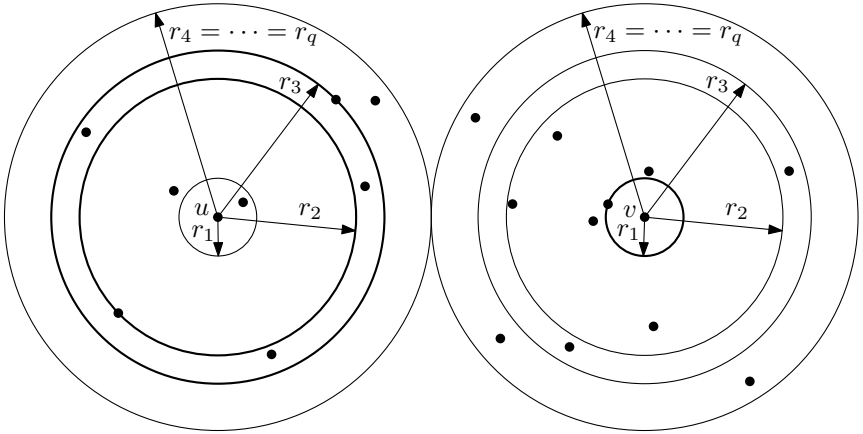
4.2.2 Lemma. *Let $u, v \in V$ be two distinct points. Then there exist real numbers $\Delta_1, \Delta_2, \ldots, \Delta_q \ge 0$ with $\Delta_1 + \cdots + \Delta_q = d_V(u,v)/2$, where $q := \lfloor \log_2 n \rfloor + 1$, and such that the following holds for each $j = 1, 2, \ldots, q$: If $A_j \subseteq V$ is a randomly chosen subset of V, with each point of V included in A_j independently with probability 2^{-j}, then the probability P_j of the event*

$$|d_V(u, A_j) - d_V(v, A_j)| \ge \Delta_j$$

satisfies $P_j \ge \frac{1}{12}$.

Proof. We fix u and v. As in the proof of Lemma 4.1.3, we will build a system of balls around u and around v. But now the construction will be driven by the *number of points* in the balls, and the radii will be set accordingly.

We think of two balls $B(u,r)$ and $B(v,r)$ of the same radius r, and we let r grow from $r_0 := 0$ to $d_V(u,v)/2$. During this growth, we record the moments $r_1, r_2, \ldots$, where r_j is the smallest r for which both $|B(u,r)| \ge 2^j$ and $|B(v,r)| \ge 2^j$. Here is an example:



The growth stops at the radius $d_V(u,v)/2$, where the balls just touch. For those j such that one or both of these touching balls have fewer than 2^j points, we set $r_j := d_V(u,v)/2$.

We are going to show that the claim of the lemma holds with $\Delta_j := r_j - r_{j-1}$.

If $\Delta_j = 0$, then the claim holds automatically, so we assume $\Delta_j > 0$, and thus $r_{j-1} < r_j$. Then both $|B(u, r_{j-1})| \ge 2^{j-1}$ and $|B(v, r_{j-1})| \ge 2^{j-1}$.

Let $A_j \subseteq V$ be a random sample with point probability 2^{-j}. By the definition of r_j, we have $|B^\circ(u, r_j)| < 2^j$ or $|B^\circ(v, r_j)| < 2^j$, where $B^\circ(x,r) = \{y \in V : d_V(x,y) < r\}$ denotes the open ball (this holds for $j = q$, too, because $|V| \le 2^q$).

We choose the notation u, v so that $|B^\circ(u, r_j)| < 2^j$. If A_j intersects $B(v, r_{j-1})$ and misses $B^\circ(u, r_j)$, then it has the desired property $|d_V(u, A_j) - d_V(v, A_j)| \ge \Delta_j$. As was already mentioned, we have $|B^\circ(u, r_j)| < 2^j$ and $|B(v, r_{j-1})| \ge 2^{j-1}$. Thus, we can estimate the probability that both $A_j \cap B(v, r_{j-1}) \ne \emptyset$ and $A_j \cap B^\circ(u, r_j) = \emptyset$ by exactly the same calculation as in the proof of Lemma 4.1.3 (with $p = \frac{1}{2}$), and we get that this probability is at least $\frac{1}{12}$. □

Proof of Theorem 4.2.1. We set $m := \lfloor C \log_2 n \rfloor$ for a sufficiently large constant C, $q = \lfloor \log_2 n \rfloor + 1$ is as in Lemma 4.2.2, and $k := mq$. For $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, q$ we sample the sets A_{ij} independently, where each point of V is included in A_{ij} independently with probability 2^{-j}. Then we define $f\colon V \to \ell_2^k$ as the Fréchet embedding with $f_{ij}(v) = d_V(v, A_{ij})$. So far this is almost exactly as in the proof of Theorem 4.1.1.

Since each f_{ij} is nonexpanding and there are $k = O(\log^2 n)$ coordinates, the mapping f doesn't expand any distance by more than $\sqrt{k} = O(\log n)$. It remains to show that, with a positive probability, f doesn't contract any distance by more than a constant factor.

Let us call an index i *good for u, v and j* if $|d_V(u, A_{ij}) - d_V(v, A_{ij})| \ge \Delta_j$, where $\Delta_1, \ldots, \Delta_q$ are the numbers as in Lemma 4.2.2 (depending on u and v).

4.2.3 Claim. *For the A_{ij} chosen randomly as above, the following holds with a positive probability:*

For every pair $\{u, v\}$ of points of V and for every $j = 1, 2, \ldots, q$, there are at least $\frac{m}{24}$ good indices i.

> **Proof of the claim.** For u, v and j fixed, the probability that a particular i is good is at least $\frac{1}{12}$. So, still with u, v, j fixed, the probability that there are fewer than $\frac{m}{24}$ good indices i is at most the probability that in m independent Bernoulli trials, each with success probability $\frac{1}{12}$, we get fewer than $\frac{m}{24}$ successes.
>
> This probability is at most e^{-cm} for a suitable positive constant $c > 0$. This follows from standard Chernoff-type estimates (see, e.g., Corollary A.1.14 in the Alon–Spencer book *The Probabilistic Method*), or it can also be calculated by elementary estimates of binomial coefficients.
>
> Thus, the probability that the condition in the claim fails for some u, v and j is at most $\binom{n}{2} q e^{-cm} \le O(n^2 \log n) e^{-cC \log n} < 1$ for C sufficiently large compared to c. The claim is proved. □

To finish the proof of Theorem 4.2.1, we fix a choice of the A_{ij} so that the condition in Claim 4.2.3 holds. Then, for the corresponding f and each pair $\{u, v\}$ we have

$$\begin{aligned} \|f(u) - f(v)\|_2^2 &= \sum_{j=1}^{q} \sum_{i=1}^{m} |f_{ij}(u) - f_{ij}(v)|^2 \\ &\ge \sum_{j=1}^{q} \sum_{i \text{ good for } u,v,j} \Delta_j^2 \ge \frac{m}{24} \sum_{j=1}^{q} \Delta_j^2 \\ &\ge \frac{m}{24} \cdot \frac{1}{q} \Big(\sum_{j=1}^{q} \Delta_j \Big)^2 = \frac{m}{24q} \cdot \frac{d_V(u,v)^2}{4} \end{aligned}$$

(the last inequality is Cauchy–Schwarz). Hence $\|f(u)-f(v)\|_2 = \Omega(d_V(u,v))$ as needed, and Theorem 4.2.1 is proved. □

Remark. Since every ℓ_2 metric is also an ℓ_p metric for every p, Theorem 4.2.1 immediately implies that every n-point metric space $O(\log n)$ embeds in ℓ_p.

However, it is nice to know that the same embedding f as in the above proof (satisfying the condition in Claim 4.2.3), regarded as an embedding in ℓ_p^k, also has distortion $O(\log n)$ (the implicit constant can be shown to behave as p^{-1} as $p \to \infty$). Indeed, only the final calculation needs to be modified, using Hölder's inequality in place of Cauchy–Schwarz.

An advantage of this "direct" embedding in ℓ_p is that the dimension k of the target space is only $O(\log^2 n)$. With additional ideas and more complicated proof, this has been improved to $k = O(\log n)$, still with $O(\log n)$ distortion (which is tight in the worst case).

4.3 Approximating the sparsest cut

We present one of the earliest and still most impressive algorithmic applications of low-distortion embeddings.

Let $G = (V, E)$ be a given graph. We would like to compute the quantity $\phi(G)$ as in Section 3.5, i.e., the minimum, over all sets $S \subseteq V$, $\emptyset \neq S \neq V$, of the density

$$\phi(G, S) := \frac{|E(S, V \setminus S)|}{|S| \cdot |V \setminus S|}.$$

That is, we want to assess, how good an expander G is.

The problem of computing $\phi(G)$ is known to be NP-hard, and assuming a famous and plausible-looking general conjecture, the *Unique Games Conjecture*, it is even hard to *approximate* $\phi(G)$ within any constant factor.[1]

As we'll explain, the tools we have covered, most notably Bourgain's theorem, yield a polynomial-time $O(\log n)$-approximation algorithm.

4.3.1 Proposition. *There is a randomized algorithm that, given a graph $G = (V, E)$ on n vertices, computes in (expected) polynomial time a set $S \subset V$ with $\phi(G, S) \le O(\log n) \cdot \phi(G)$.*

[1] This means that, assuming the Unique Games Conjecture, there are no constant C and polynomial-time algorithm $\mathcal{A}$ that computes, for every graph G, a number $\mathcal{A}(G)$ such that $\phi(G) \le \mathcal{A}(G) \le C\phi(G)$.

Proof. First we need Lemma 3.6.3, which we re-state as follows. Let $F := \binom{V}{2}$, $N := |F| = \binom{n}{2}$, and for a nonzero vector $\mathbf{z} = (z_{uv} : \{u,v\} \in F) \in \mathbb{R}^N$, let

$$R(\mathbf{z}) := \frac{\sum_{\{u,v\}\in E} z_{uv}}{\sum_{\{u,v\}\in F} z_{uv}}.$$

Then Lemma 3.6.3 shows that

$$\phi(G) = \min\bigl\{R(\mathbf{d}) : \mathbf{d} \in \mathcal{L}_1 \setminus \{\mathbf{0}\}\bigr\},$$

where, as usual, $\mathcal{L}_1 \subset \mathbb{R}^N$ is the cone of all ℓ_1 metrics on V.

Both of the proofs of Lemma 3.6.3 shown in Section 3.6 actually yield a *polynomial-time algorithm* that, given an ℓ_1 metric d on V represented by a mapping $f\colon V \to \ell_1^k$ (i.e., $d(u,v) = \|f(u) - f(v)\|_1$), finds an S such that $\phi(G, S) \le R(\mathbf{d})$. (Checking this may need revisiting the proofs and some thought.) Thus, finding a sparsest cut is equivalent to minimizing R over all nonzero ℓ_1 metrics (and so, in particular, the latter problem is also NP-hard).

The next step is the observation that we can efficiently minimize R over all nonzero (pseudo)metrics, ℓ_1 or not. To this end, we set up the following **linear program**[2]:

$$\begin{array}{lll} \text{Minimize} & \sum_{\{u,v\}\in E} z_{uv} & \\ \text{subject to} & \sum_{\{u,v\}\in F} z_{uv} = 1, & \\ & z_{uv} \ge 0 & \text{for all } \{u,v\} \in F, \\ & z_{uv} + z_{vw} \ge z_{uw} & \text{for all } u,v,w \in V \text{ distinct.} \end{array}$$

There are N variables z_{uv}, $\{u,v\} \in F$. Note the trick how we have got rid of the nonlinearity of R: Using homogeneity, we can assume that the denominator is fixed to 1 (this is the first constraint of the linear program), and we minimize the numerator. The remaining constraints express the nonnegativity of $\mathbf{z}$ and the triangle inequality, and thus they make sure that all feasible $\mathbf{z}$ are pseudometrics.

As is well known, the linear program can be solved in polynomial time, and thus we find a (pseudo)metric d_0 minimizing R. Since we minimized over a set including all ℓ_1 metrics, and $\phi(G)$ is the minimum over ℓ_1 metrics, we have $R(\mathbf{d}_0) \le \phi(G)$.

Now we D-embed the metric space (V, d_0) in ℓ_1^k for some k, with D as small as possible. Bourgain's theorem guarantees that there *exists*

[2] A linear program is the problem of minimizing a linear function of n real variables over a set $S \subseteq \mathbb{R}^n$ specified by a system of linear inequalities and equations.

an embedding with $D = O(\log n)$. The proof in Section 4.2 actually shows, first, that we can assume $k = O(\log^2 n)$, and second, that the embedding can be found by a randomized polynomial-time algorithm: We just choose the appropriate random subsets A_{ij}, and then we check whether the Fréchet embedding defined by them has a sufficiently low distortion—if not, we discard them and start from scratch.

So now we have an ℓ_1 metric d_1 on V, represented by an embedding in $\ell_1^{O(\log^2 n)}$, which differs from d_0 by distortion at most $D = O(\log n)$. It is easy to see that

$$R(\mathbf{d}_1) \le D \cdot R(\mathbf{d}_0) \le D\phi(G).$$

Finally, given d_1, we can find S with $\phi(G, S) \le R(\mathbf{d}_1)$, as was noted at the beginning of the proof. This is the set returned by the algorithm, which satisfies $\phi(G, S) \le O(\log n) \cdot \phi(G)$ as required. □

A better approximation algorithm. The above algorithm can be summarized as follows. We want to minimize R over the set $\mathcal{L}_1 \setminus \{\mathbf{0}\} \subset \mathbb{R}^N$ of all nonzero ℓ_1 metrics on V. Instead, we minimize it over the larger set $\mathcal{M}$ of all metrics, and the following properties of $\mathcal{M}$ are relevant:

(i) $\mathcal{M}$ contains $\mathcal{L}_1 \setminus \{\mathbf{0}\}$,

(ii) R can be minimized in polynomial time over $\mathcal{M}$, and

(iii) every element of $\mathcal{M}$ can be approximated by an element of $\mathcal{L}_1$ with distortion at most $O(\log n)$.

If we could find another subset of $\mathbb{R}^N$ with properties (i) and (ii) but with a better distortion guarantee in (iii), then we would obtain a correspondingly better approximation algorithm for the sparsest cut.

A suitable subset has indeed been found: the class $\mathcal{N}$ of all metrics of negative type on V. These are all metrics that can be represented as squares of Euclidean distances of points in ℓ_2; see Section 3.6.

First, by Lemma 3.6.2, we have $\mathcal{L}_1 \subseteq \mathcal{N}$, so (i) holds.

Second, the minimum of R over $\mathcal{N}$ can be computed by semidefinite programming. Indeed, to the linear program shown above, which expresses the minimization of R over $\mathcal{M}$, it suffices to add constraints expressing that the z_{uv} are squared Euclidean distances. This is done exactly as in the proof of Proposition 3.7.1.

Third, it is known that every n-point metric of negative type can be embedded in ℓ_2 (and thus also in ℓ_1) with distortion $O(\sqrt{\log n}\log\log n)$, and the proof provides a randomized polynomial-time algorithm.

This is a "logical" way to an improved approximation algorithm, but the historical development went differently. First came a breakthrough of Arora, Rao, and Vazirani, an $O(\sqrt{\log n}\,)$-approximation algorithm for the sparsest cut, which indeed begins by optimizing R over $\mathcal{N}$, but then it "rounds" the solution directly to a sparsest cut in the input graph, without constructing an embedding in ℓ_1 first.

Only later and with considerable additional effort it was understood that the geometric part of this algorithm's analysis also leads to low-distortion embeddings. Moreover, the distortion guarantee is slightly worse, by the $\log\log n$ factor, than the approximation guarantee of the algorithm.

However, there is a more general (and more important) algorithmic problem, the *sparsest cut for multicommodity flows*,[3] where the best known approximation algorithm is indeed obtained essentially according to the items (i)–(iii) above, using a low-distortion embedding of a metric of negative type in ℓ_1.

The embeddability of metrics of negative type in ℓ_1 and in ℓ_2 has been one of the most fascinating topics in metric embeddings in recent years (approximation algorithms providing a strong motivation), with some of the deepest and technically most demanding results. Although it was

[3] In the multicommodity flow problem, one can think of a graph $G = (V, E)$ whose vertices are cities and whose edges represent roads. Each edge (road) e has a nonnegative *capacity* $Cap(e)$ (trucks per day, say). There are k *demands*, such as that 56.7 trucks of DVD players per day should be shipped from Fukuoka to Tokyo, 123.4 trucks of fresh fish per day should be shipped from Kanazawa to Osaka, etc. Each demand is specified by an (unordered) pair $\{u, v\}$ of vertices and a nonnegative real number $Dem(u, v)$.

Assuming that the road capacities are not sufficient to satisfy all demands, one may ask (among others), what is the largest λ such that at least λ fraction of each demand can be satisfied, for all demands simultaneously. (To prevent confusion, let us stress that *this* algorithmic problem can be solved in polynomial time.)

If S is a subset of the cities such that the total capacity of all roads between S and $V \setminus S$ equals A and the sum of all demands with one city in S and the other outside S equals B, then $\frac{A}{B}$ is an upper bound for λ (and it's quite natural to call $\frac{A}{B}$ the *density* of the cut S).

For $k = 1$, a single-commodity flow, the well known max-flow/min-cut theorem asserts that there always exists a cut for which $\frac{A}{B}$ *equals* the maximum possible λ. For large k, there may be a gap; equality need not hold for any cut. But, using an argument slightly more complicated than that in the proof of Proposition 4.3.1, it can be shown that one can efficiently compute a cut S for which $\frac{A}{B}$ is within a multiplicative factor of $O(\log k)$ of the optimum λ (and in particular, that such a cut always exists).

Using the improved embedding of metrics of negative type in ℓ_2, the $O(\log k)$ factor in this result has been improved to $O(\sqrt{\log k}\log\log k)$. One also obtains a polynomial-time algorithm guaranteed to find a cut at most this factor away from optimal.

initially conjectured that all metrics of negative type might embed in ℓ_1 with distortion bounded by a universal constant, by now a lower bound of $\Omega((\log n)^c)$, for a small positive constant c, is known.

4.4 Exercises

1.* Refer to the proof of Theorem 4.2.1 (Bourgain's theorem). Show that the same mapping $f: V \to \mathbb{R}^k$ as given in the proof also provides an embedding of V into ℓ_p^k with $O(\log n)$ distortion, for every fixed $p \in [1, \infty)$. Describe only the modifications of the proof—you need not repeat parts that remain unchanged.

A

A Fourier-analytic proof of the KKL theorem

A.1 A quick introduction to the Fourier analysis on the Hamming cube

The reader has probably seen something from the "classical" Fourier analysis, where a "reasonable" function $f\colon [0,2\pi] \to \mathbb{R}$ is expressed by a series of sines and cosines. The Fourier analysis on the finite set $\{0,1\}^n$ instead of the interval $[0,2\pi]$ is analogous in some ways.[1] But its foundations are much simpler: there are no issues of convergence, only basic linear algebra in finite dimension.

So we consider the real vector space $\mathcal{F}$ of all functions $f\colon \{0,1\}^n \to \mathbb{R}$ (with pointwise addition and multiplication by scalars). It has dimension 2^n; the functions equal to 1 at one of the points and zero elsewhere form an obvious basis. The idea of the Fourier analysis is expressing functions in another basis.

So for every $\mathbf{a} \in \{0,1\}^n$, we define the function $\chi_{\mathbf{a}} \in \mathcal{F}$ by

$$\chi_{\mathbf{a}}(\mathbf{u}) := (-1)^{a_1u_1+\cdots+a_nu_n} = \prod_{i:a_i=1} (-1)^{u_i}.$$

The $\chi_{\mathbf{a}}$ are called the **characters**, and as we will soon check, they form a basis of $\mathcal{F}$.

[1] Both are instances of a general framework, where G is a locally compact Abelian topological group, and complex functions on G are expressed using the characters of G.

For $\mathbf{u},\mathbf{v} \in \{0,1\}^n$, $\mathbf{u}+\mathbf{v}$ stands for the componentwise sum of $\mathbf{a}$ and $\mathbf{b}$ modulo 2, the same notation as in Section 3.9. So we regard $\{0,1\}^n$ as the Abelian group $(\mathbb{Z}/2\mathbb{Z})^n$. We have $\chi_{\mathbf{a}}(\mathbf{u}+\mathbf{v}) = \chi_{\mathbf{a}}(\mathbf{u})\chi_{\mathbf{a}}(\mathbf{v})$ (in agreement with the general definition of a character of an Abelian group G as a homomorphism $G \to (\mathbb{C},\times)$), and also $\chi_{\mathbf{a}+\mathbf{b}}(\mathbf{u}) = \chi_{\mathbf{a}}(\mathbf{u})\chi_{\mathbf{b}}(\mathbf{u})$.

Next, we define a *scalar product* on $\mathcal{F}$ by

$$\langle f,g\rangle := \frac{1}{2^n} \sum_{\mathbf{u}\in\{0,1\}^n} f(\mathbf{u})g(\mathbf{u}).$$

It is easy to check that $(\chi_{\mathbf{a}} : \mathbf{a} \in \{0,1\}^n)$ form an *orthonormal system*, meaning that

$$\langle \chi_{\mathbf{a}}, \chi_{\mathbf{b}}\rangle = \begin{cases} 1 & \text{for } \mathbf{a}=\mathbf{b}, \\ 0 & \text{for } \mathbf{a}\neq\mathbf{b} \end{cases}$$

(this follows by an easy calculation, using the property $\chi_{\mathbf{a}}\chi_{\mathbf{b}} = \chi_{\mathbf{a}+\mathbf{b}}$ mentioned above, and we skip it). This implies that the $\chi_{\mathbf{a}}$ are linearly independent, and since there are 2^n of them, they constitute an orthonormal basis of $\mathcal{F}$.

The **Fourier coefficients** of a function $f \in \mathcal{F}$ are the coordinates of f in this basis. The coefficient of f corresponding to $\chi_{\mathbf{a}}$ is traditionally denoted by $\hat{f}(\mathbf{a})$.

A simple fact of linear algebra is that the coordinates with respect to an orthonormal basis can be computed using scalar products. In our case we thus have

$$\hat{f}(\mathbf{a}) = \langle \chi_{\mathbf{a}}, f\rangle.$$

Another easy general result of linear algebra tells us how the scalar product of two vectors is computed from their coordinates with respect to some orthonormal basis. In our case this rule reads

$$\langle f,g\rangle = \sum_{\mathbf{a}} \hat{f}(\mathbf{a})\hat{g}(\mathbf{a}) \tag{A.1}$$

(here and in the sequel, the summation is over $\{0,1\}^n$ unless specified otherwise). The most often used special case of this is with $f=g$, where it expresses the ℓ_2 norm of f using the Fourier coefficients; this is the **Parseval equality**:

$$\|f\|_2^2 := \langle f, f\rangle = \sum_{\mathbf{a}} \hat{f}(\mathbf{a})^2.$$

Here is another straightforward fact, whose proof is again omitted.

A.1.1 Fact (Fourier coefficients of a translation). *Let $\mathbf{w} \in \{0,1\}$ be a fixed vector, let $f \in \mathcal{F}$, and let $g \in \mathcal{F}$ be defined by $g(\mathbf{u}) := f(\mathbf{u}+\mathbf{w})$. Then*

$$\hat{g}(\mathbf{a}) = \chi_{\mathbf{a}}(\mathbf{w})\hat{f}(\mathbf{a}).$$

We will now express the influence of a Boolean function using its Fourier coefficients. So let $f\colon \{0,1\}^n \to \{0,1\}$, and let $\partial_k f$ stand for the function given by $\partial_k f(\mathbf{u}) := f(\mathbf{u}+\mathbf{e}_k) - f(\mathbf{u})$ (the notation should suggest a formal analogy with the partial derivative of a real function on $\mathbb{R}^n$). Since the values of $\partial_k f$ are in $\{0,-1,+1\}$, we have

$$I_k f = 2^{-n}\sum_{\mathbf{u}} |\partial_k f(\mathbf{u})| = 2^{-n}\sum_{\mathbf{u}} |\partial_k f(\mathbf{u})|^2 = \|\partial_k f\|_2^2,$$

and the Parseval equality thus gives

$$I_k(f) = \sum_{\mathbf{a}} \widehat{\partial_k f}(\mathbf{a})^2.$$

Using Fact A.1.1 with $\mathbf{w} = \mathbf{e}_k$, we easily calculate that

$$\widehat{\partial_k f}(\mathbf{a}) = \begin{cases} -2\hat{f}(\mathbf{a}) & \text{if } a_k = 1,\\ 0 & \text{if } a_k = 0.\end{cases}$$

We thus arrive at

$$I_k(f) = 4\sum_{\mathbf{a}:a_k=1} \hat{f}(a)^2, \tag{A.2}$$

and summing over k then yields

$$I(f) = \sum_{k=1}^{n} I_k(f) = 4\sum_{\mathbf{a}} |\mathbf{a}|\hat{f}(\mathbf{a})^2, \tag{A.3}$$

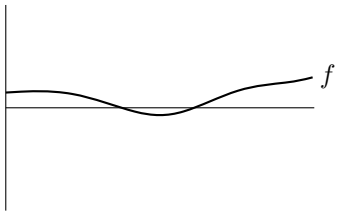
where $|\mathbf{a}|$ denotes the number of 1s in $\mathbf{a}$.

A.2 ℓ_p norms and a hypercontractive inequality

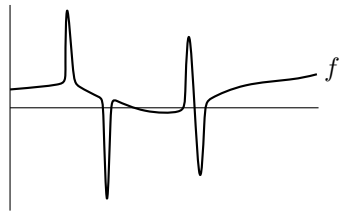
We have been using the ℓ_2 norm $\|f\|_2 = \langle f, f\rangle^{1/2}$ of functions on $\{0,1\}^n$; now we need to consider the ℓ_p norm

$$\|f\|_p = \left(2^{-n}\sum |f(\mathbf{u})|^p\right)^{1/p}.$$

One can say that for p small (close to 1), the ℓ_p norm measures mainly the "typical" behavior of f (in particular, $\|f\|_1 = \mathbf{E}[|f|]$), while for larger p, $\|f\|_p$ is more influenced by "spikes" of f. Here is a pictorial analog for functions $[0,1] \to \mathbb{R}$:



$\|f\|_p$ and $\|f\|_q$ not very different



$\|f\|_p$ much larger than $\|f\|_q$ for $p > q$

Let us define the **degree** of f as

$$\max\{|\mathbf{a}| : \hat{f}(\mathbf{a}) \neq 0\},$$

the maximum "level" of a nonzero Fourier coefficient.[2]

For a function $f \in \mathcal{F}$ and a threshold parameter $t \in [0,d)$, let us define

$$f_{\text{low}} := \sum_{\mathbf{a}:|\mathbf{a}|\le t} \hat{f}(\mathbf{a})\chi_{\mathbf{a}}$$

as the "low-degree part" of f; that is, f_{low} is obtained from f by truncating the Fourier expansion on level t.

We come to the main result of this section.

[2] We can formally express f as a (multilinear) polynomial. To this end, we write $\chi_{\mathbf{a}}(\mathbf{u}) = \prod_{i:a_i=1}(1-2u_i)$, which is a polynomial of degree $|\mathbf{a}|$ (note that this involves a "type cast": while we usually add the u_i modulo 2, here we regard them as real numbers). Then the degree of f is the degree of the corresponding polynomial.

Here we can also see the advantage of writing the Hamming cube as $\{-1,1\}^n$; then $\chi_{\mathbf{a}}(\mathbf{u})$ is simply the monomial $\prod_{i:a_i=-1} u_i$.

A.2.1 Proposition (A hypercontractive inequality). *There are constants C and $p < 2$ such that for every $f\colon \{0,1\}^n \to \mathbb{R}$ and every t we have*

$$\|f_{\mathrm{low}}\|_2 \le C^t \|f\|_p.$$

We will comment on the meaning of the word "hypercontractive" at the end of this section. Here we just observe that the ℓ_2 norm on the left-hand side is more sensitive to spikes than the ℓ_p norm on the right-hand side. So we can think of the inequality as a quantitative expression of the intuitive fact that removing high-level components of the Fourier expansion makes f smoother.

We will prove the proposition with $p = 4/3$ and $C = \sqrt{3}$, and the main step in the proof is the following lemma.

A.2.2 Lemma. *Let f be a function of degree at most t. Then*

$$\|f\|_4 \le \sqrt{3}^t \|f\|_2.$$

Proof. The exponents 2 and 4 are convenient, since they allow for a relatively simple inductive proof. We actually prove the fourth power of the required inequality, i.e.,

$$\|f\|_4^4 \le 9^t \|f\|_2^4$$

by induction on n.

In the inductive step, we want to get rid of the last variable u_n. We split the Fourier expansion of f into two parts, one with the characters that do not depend on u_n, and the other with those that do. For $\mathbf{u} \in \{0,1\}^n$, let $\overline{\mathbf{u}} := (u_1, \dots, u_{n-1})$, and for $\mathbf{b} \in \{0,1\}^{n-1}$, we write $\mathbf{b}0$ for $(b_1, b_2, \dots, b_{n-1}, 0)$ (and similarly for $\mathbf{b}1$). We have

$$\begin{aligned} f(\mathbf{u}) &= \sum_{\mathbf{a}\in\{0,1\}^n} \hat{f}(\mathbf{a})\chi_{\mathbf{a}}(\mathbf{u}) \\ &= \sum_{\mathbf{b}\in\{0,1\}^{n-1}} \hat{f}(\mathbf{b}0)\chi_{\mathbf{b}0}(\mathbf{u}) + \sum_{\mathbf{b}\in\{0,1\}^{n-1}} \hat{f}(\mathbf{b}1)\chi_{\mathbf{b}1}(\mathbf{u}) \\ &= \sum_{\mathbf{b}\in\{0,1\}^{n-1}} \hat{f}(\mathbf{b}0)\chi_{\mathbf{b}}(\overline{\mathbf{u}}) + (-1)^{u_n} \sum_{\mathbf{b}\in\{0,1\}^{n-1}} \hat{f}(\mathbf{b}1)\chi_{\mathbf{b}}(\overline{\mathbf{u}}) \\ &= g(\overline{\mathbf{u}}) + (-1)^{u_n} h(\overline{\mathbf{u}}). \end{aligned}$$

Here g is of degree at most t and h of degree at most $t-1$. Moreover, by the orthogonality of the characters we can see that

$$\|f\|_2^2 = \|g\|_2^2 + \|h\|_2^2.$$

We can begin the calculation for the inductive step.

$$\begin{aligned} \|f\|_4^4 &= 2^{-n} \sum_{\mathbf{u}} f(\mathbf{u})^4 \\ &= 2^{-n} \Bigg[\sum_{\mathbf{v}\in\{0,1\}^{n-1}} (g(\mathbf{v}) + h(\mathbf{v}))^4 + \sum_{\mathbf{v}\in\{0,1\}^{n-1}} (g(\mathbf{v}) - h(\mathbf{v}))^4 \Bigg]. \end{aligned}$$

We expand the fourth powers according to the Binomial Theorem; the terms with odd powers cancel out, while those with even powers appear twice, and we arrive at

$$\begin{aligned} &= 2 \cdot 2^{-n} \sum_{\mathbf{v}} \left(g(\mathbf{v})^4 + 6g(\mathbf{v})^2 h(\mathbf{v})^2 + h(\mathbf{v})^4 \right) \\ &= \|g\|_4^4 + 6\langle g^2, h^2 \rangle + \|h\|_4^4 \end{aligned}$$

(the norms and the scalar product are now for functions on $\{0,1\}^{n-1}$, one dimension less).

For the terms $\|g\|_4^4$ and $\|h\|_4^4$ we will simply use the inductive assumption. The only trick is with estimating the scalar product $\langle g^2, h^2 \rangle$: for that we use the Cauchy–Schwarz inequality $\langle \mathbf{x}, \mathbf{y} \rangle \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$, which in our case gives $\langle g^2, h^2 \rangle \le \|g^2\|_2 \|h^2\|_2 = \|g\|_4^2 \|h\|_4^2$. Only after that we apply induction, and we obtain

$$\begin{aligned} &\le \|g\|_4^4 + 6\|g\|_4^2 \|h\|_4^2 + \|h\|_4^4 \\ &\le 9^t \|g\|_2^4 + 6 \cdot 9^{t/2} \|g\|_4^2 \cdot 9^{(t-1)/2} \|h\|_4^2 + 9^{t-1} \|h\|_2^4 \\ &= 9^t \left(\|g\|_2^4 + 6 \cdot 9^{-1/2} \|g\|_4^2 \|h\|_4^2 + 9^{-1} \|h\|_2^4 \right) \\ &\le 9^t \left(\|g\|_2^2 + \|h\|_2^2 \right)^2 \\ &= 9^t \|f\|_2^4. \end{aligned}$$

The lemma is proved. (Well, wait a second... what is the basis of the induction?) □

Proof of Proposition A.2.1. If we apply the lemma just proved to f_{low}, we get

$$\|f_{\mathrm{low}}\|_4 \le \sqrt{3}^t \|f_{\mathrm{low}}\|_2, \tag{A.4}$$

but how do we relate this to the norm of f itself?

The first trick of the proof is to consider the scalar product $\langle f_{\text{low}}, f\rangle$ and express it using the Fourier coefficients:

$$\langle f_{\text{low}}, f\rangle = \sum_{\mathbf{a}} \widehat{f_{\text{low}}}(\mathbf{a})\hat{f}(\mathbf{a}) = \sum_{\mathbf{a}:|\mathbf{a}|\le t} \hat{f}(\mathbf{a})^2 = \|f_{\text{low}}\|_2^2$$

(Parseval).

The second and last trick is to write Hölder's inequality for $\langle f_{\text{low}}, f\rangle$, with exponents $p = 4$ and $q = \frac{4}{3}$, which gives

$$\|f_{\text{low}}\|_2^2 = \langle f_{\text{low}}, f\rangle \le \|f_{\text{low}}\|_4 \|f\|_{4/3}.$$

Now we bound $\|f_{\text{low}}\|_4$ using (A.4), divide the resulting inequality by $\|f_{\text{low}}\|_2$, and we arrive at the desired inequality $\|f_{\text{low}}\|_2 \le \sqrt{3}^t \|f\|_{4/3}$. □

On hypercontractive inequalities. Let $(Z, \|.\|)$ be a normed space. A 1-Lipschitz mapping $Z \to Z$ is often called *contractive*. This term is most often used for *linear* mappings $A\colon Z \to Z$, which in this context are referred to as *(linear) operators*. For a linear operator, contractivity means $\|Ax\| \le \|x\|$ for every $x \in Z$.

Now let us consider *two* different norms $\|.\|$ and $\|.\|'$ on Z, and assume that $\|x\| \le \|x\|'$ for all x. If the linear operator A even satisfies $\|Ax\|' \le \|x\|$, it is called **hypercontractive** (it is "more than contractive", since if x is small under the smaller norm, Ax is small even under the bigger norm).

For us, the relevant setting is $Z = \mathcal{F}$, the space of all functions $\{0,1\}^n \to \mathbb{R}$, and $\|.\| = \|.\|_p$, $\|.\|' = \|.\|_q$ with $p < q$. Note that we indeed have $\|f\|_p \le \|f\|_q$ for all $f \in F$. This holds for the functions on any probability space, and it follows from Hölder's inequality. (This should not be confused with the case of the ℓ_p norms on $\mathbb{R}^n$, where we have the *opposite* inequality $\|\mathbf{x}\|_p \ge \|\mathbf{x}\|_q$ for $p < q$.)

How does Proposition A.2.1 fit into this? We regard the "truncation" $f \mapsto f_{\text{low}}$ as an operator $L\colon F \to F$; it is linear because each Fourier coefficient $\hat{f}(\mathbf{a})$ depends linearly on f. The proposition then tells us that L is hypercontractive for some $p < 2$ and $q = 2$, well, almost, because there is the factor C^t.

This hypercontractive inequality is relatively easy to prove, and in a sense, it is tailored for the proof of the KKL theorem. The original, and usual, proof of the KKL theorem uses another hypercontractive inequality, proved by Bonami and independently by Gross (often also attributed to Beckner, who proved some generalizations later).

To state it, we introduce, for a real parameter $\rho \in [0,1]$, the **noise operator** $T_\rho\colon \mathcal{F} \to \mathcal{F}$. The simplest way of defining it is in terms of the Fourier expansion:

$$T_\rho f := \sum_{\mathbf{a}} \hat{f}(\mathbf{a})\rho^{|\mathbf{a}|}\chi_{\mathbf{a}};$$

that is, the higher Fourier coefficient, the more it is reduced by T_ρ. In particular, $T_1 f = f$ and $T_0 f$ is the constant function equal to $\hat{f}(\mathbf{0}) = \mathbf{E}[f]$.

To explain the name "noise operator", we need another definition. Let $p := (1-\rho)/2$, and let $\mathbf{x} \in \{0,1\}^n$ be a random vector, the noise, where each x_i is set to 1 with probability p and 0 with probability $1-p$, independent of all other x_j's. Then we have

$$T_\rho f(\mathbf{u}) := \mathbf{E}[f(\mathbf{u}+\mathbf{x})]\,.$$

In words, to evaluate the function $T_\rho f$ at some given $\mathbf{u}$, we first flip each coordinate of $\mathbf{u}$ with probability p, then we apply f to the resulting vector, and we take the expectation over the random flips. Thus, $T_\rho f(\mathbf{u})$ is a weighted average of the values of f, where (for $\rho < 1$) values at points closer to $\mathbf{u}$ are taken with larger weight. It is not too hard to verify that this definition is equivalent to the Fourier-analytic one above.

The hypercontractive inequality for T_ρ asserts that

$$\|T_\rho f\|_q \le \|f\|_p \tag{A.5}$$

for all $f \in \mathcal{F}$ if (and only if) $\rho^2 \le \frac{p-1}{q-1}$, $1 \le p \le q$. This, with $q = 2$, was used in the first proof of the KKL theorem. Essentially, one first derives Proposition A.2.1 from (A.5) and then proceeds as we will do in the next section.

In the usual proof of (A.5), one first proves the case $n = 1$, which is a laborious but essentially straightforward calculus problem. Then one derives the general case from the 1-dimensional one by a general inductive argument; this is often expressed by saying that the inequality (A.5) *tensors*. This is one of the most common approaches to proving multidimensional inequalities, such as various isoperimetric inequalities: one needs to find a version of the considered inequality that tensors, and then work out the one-dimensional version.

The KKL theorem, as well as (A.5), can also be derived from a **log-Sobolev inequality** for the Hamming cube, which asserts that

$$\sum_{\{\mathbf{u},\mathbf{v}\}\in E} |f(\mathbf{u}) - f(\mathbf{v})|^2 \ge \frac{1}{n}\operatorname{Ent}[f^2],$$

where E is the edge set of the Hamming cube, and where the **entropy** of a function g is defined as $\mathrm{Ent}[g] := \mathbf{E}[g\log(g/\mathbf{E}[g])]$. The log-Sobolev inequality is again proved for the 1-dimensional case and then tensored. This way of proving the KKL theorem is nicely presented in the survey P. Biswal: Hypercontractivity and its Applications, available from Biswal's home page. It is probably shorter than the one presented in the current chapter.

A.3 The KKL theorem

We begin with a quick Fourier-analytic proof of the following inequality, mentioned in Section 3.9, for the total influence of a Boolean function:

$$I(f) \geq 4\mathrm{Var}[f] = 4\mu(1-\mu).$$

(A quiz for the reader: this is an isoperimetric inequality in disguise, bounding from below the smallest number of edges connecting a subset A of vertices of the Hamming cube to the complement of A. Can you see why?)

Using the equality $I(f) = 4\sum_{\mathbf{a}} |\mathbf{a}|\hat{f}(\mathbf{a})^2$ derived earlier and the Parseval equality, we obtain

$$\tfrac{1}{4}\, I(f) \geq \sum_{\mathbf{a}:\mathbf{a}\neq\mathbf{0}} \hat{f}(\mathbf{a})^2 = \|f\|_2^2 - \hat{f}(\mathbf{0})^2 = \mathbf{E}\left[f^2\right] - \mathbf{E}[f]^2 = \mathrm{Var}[f]$$

(the penultimate equality is a little Fourier-analytic exercise).

The beginning of the proof of the KKL theorem is in a similar spirit. We recall that we actually want to prove a more general statement, Theorem 3.9.3, which reads as follows:

> For every $f\colon\{0,1\}^n \to \{0,1\}$, we have $I(f) \geq c\mu(1-\mu)\log\frac{1}{\delta}$, where $\delta := \max_k I_k(f)$.

(We leave the derivation of the KKL theorem itself from this to the reader.)

Let us write $W := \sum_{\mathbf{a}:\mathbf{a}\neq\mathbf{0}} \hat{f}(\mathbf{a})^2 = \mu(1-\mu)$ (the last equality is from the short proof above). We distinguish two cases, depending on whether f has more weight at "high" or "low" Fourier coefficients. We fix the threshold

$$t := \lfloor \tfrac{1}{2} c \log\tfrac{1}{\delta} \rfloor$$

to separate low from high.

Case 1: main weight at high coefficients. Here we assume

$$\sum_{\mathbf{a}:|\mathbf{a}|>t} \hat{f}(\mathbf{a})^2 \geq \tfrac{W}{2}.$$

Then we are done quickly:

$$\begin{aligned} I(f) &= 4\sum_{\mathbf{a}} |\mathbf{a}|\hat{f}(\mathbf{a})^2 \geq 4(t+1) \sum_{\mathbf{a}:|\mathbf{a}|>t} \hat{f}(\mathbf{a})^2 \\ &\geq 2(t+1)W \geq c\mu(1-\mu)\log\tfrac{1}{\delta} \end{aligned}$$

(here we see where the value of t comes from). Intuitively, lot of weight at high coefficients means that f varies quickly, and this implies large influences.

Case 2: main weight at low coefficients. This is the complement of Case 1, i.e., now $\sum_{\mathbf{a}:0<|\mathbf{a}|\leq t} \hat{f}(\mathbf{a})^2 > W/2$.

Here we use the assumption that $\delta = \max_k I_k(f)$ (so far we haven't needed it), and we show that $I(f)$ is even larger than claimed in the theorem.

For a while, we will work on an individual influence $I_k(f)$ for k fixed. Let $g := \partial_k f$; we will apply the (squared) hypercontractive inequality from Proposition A.2.1 to g. For simplicity, we use the specific numerical values $C = \sqrt{3}$ and $p = \frac{4}{3}$ obtained in the proof of the proposition. Thus

$$\|g_{\mathrm{low}}\|_2^2 \leq 3^t \|g\|_{4/3}^2 = 3^t \Big(2^{-n} \sum_{\mathbf{u}} |g(\mathbf{u})|^{4/3}\Big)^{2\cdot 3/4} = 3^t I_k(f)^{3/2},$$

the last equality holding since the values of g are in $\{0,-1,1\}$ and thus $I_k(f) = \|g\|_1 = \|g\|_p^p$ for all p. Roughly speaking, this tells us that if the influence $I_k(f)$ is small, say smaller than 3^{-2t}, then the contribution of the low Fourier coefficients to it is even considerably smaller.

Now $3^t = 3^{c\log(1/\delta)} \leq \delta^{-1/4}$, say, for c sufficiently small. We estimate $I_k(f)^{3/2} \leq I_k(f)\cdot\delta^{1/2}$, then we sum over k, and the total influence will appear on the right-hand side of the resulting inequality:

$$\sum_{k=1}^n \|(\partial_k f)_{\mathrm{low}}\|_2^2 \leq 3^t\delta^{1/2} \sum_{k=1}^n I_k(f) \leq \delta^{1/4} I(f).$$

In the same way as when we expressed $I(f)$ in terms of the Fourier coefficients (the passage from (A.2) to (A.3)), we get that the left-hand side equals

$$4 \sum_{\mathbf{a}:|\mathbf{a}|\leq t} |a|\hat{f}(\mathbf{a})^2 \geq 4 \sum_{\mathbf{a}:0<|a|\leq t} \hat{f}(\mathbf{a})^2 \geq 2W.$$

Thus

$$I(f) \geq 2W\delta^{-1/4} = 2\mu(1-\mu)\delta^{-1/4} \geq c\mu(1-\mu)\log\tfrac{1}{\delta}.$$

Theorem 3.9.3 is proved. □

A.4 Exercises

1. Prove Fact A.1.1.

B

Proof of the short-diagonals lemma for ℓ_p

We recall what is to be proved: for every four points $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4 \in \ell_p$ we have

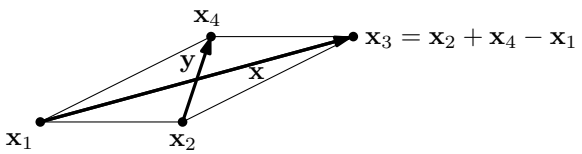
$$\begin{aligned}&\|\mathbf{x}_1-\mathbf{x}_3\|_p^2+(p-1)\|\mathbf{x}_2-\mathbf{x}_4\|_p^2\\ &\quad\le \|\mathbf{x}_1-\mathbf{x}_2\|_p^2+\|\mathbf{x}_2-\mathbf{x}_3\|_p^2+\|\mathbf{x}_3-\mathbf{x}_4\|_p^2+\|\mathbf{x}_4-\mathbf{x}_1\|_p^2. \qquad \text{(B.1)}\end{aligned}$$

First we will show that this result is an easy consequence of the following inequality:

$$\frac{\|\mathbf{x}+\mathbf{y}\|_p^2+\|\mathbf{x}-\mathbf{y}\|_p^2}{2}\ge \|\mathbf{x}\|_p^2+(p-1)\|\mathbf{y}\|_p^2,\quad 1<p<2, \qquad \text{(B.2)}$$

where $\mathbf{x},\mathbf{y}\in\mathbb{R}^k$ are arbitrary vectors. (The proof can also be extended for infinite-dimensional vectors in ℓ_p or functions in L_p, but some things come out slightly simpler in finite dimension.)

Deriving (B.1) from (B.2). For understanding this step, it is useful to note that (B.2) is equivalent to a special case of the short-diagonals lemma, namely, when $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$ are the vertices of a parallelogram:



In that case we have $\mathbf{x}_3=\mathbf{x}_2+\mathbf{x}_4-\mathbf{x}_1$, and writing (B.2) with $\mathbf{x}:=\mathbf{x}_2+\mathbf{x}_4-2\mathbf{x}_1$ and $\mathbf{y}:=\mathbf{x}_4-\mathbf{x}_2$ being the diagonals, we arrive at

$$\|\mathbf{x}_2+\mathbf{x}_4-2\mathbf{x}_1\|_p^2+(p-1)\|\mathbf{x}_4-\mathbf{x}_2\|_p^2\le 2\|\mathbf{x}_4-\mathbf{x}_1\|_p^2+2\|\mathbf{x}_2-\mathbf{x}_1\|_p^2. \qquad \text{(B.3)}$$

Now if $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$ are arbitrary, we use (B.2) for two parallelograms: The first one has vertices $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_2+\mathbf{x}_4-\mathbf{x}_1,\mathbf{x}_4$ as above, leading to (B.3), and the second parallelogram has vertices $\mathbf{x}_2+\mathbf{x}_4-\mathbf{x}_3,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$, leading to

$$\|\mathbf{x}_2+\mathbf{x}_4-2\mathbf{x}_3\|_p^2+(p-1)\|\mathbf{x}_4-\mathbf{x}_2\|_p^2\le 2\|\mathbf{x}_4-\mathbf{x}_3\|_p^2+2\|\mathbf{x}_2-\mathbf{x}_3\|_p^2. \qquad \text{(B.4)}$$

Taking the arithmetic average of (B.3) and (B.4) we almost get the inequality we want, *except* that we have $\frac12(\|\mathbf{x}_2+\mathbf{x}_4-2\mathbf{x}_1\|_p^2+\|\mathbf{x}_2+\mathbf{x}_4-2\mathbf{x}_3\|_p^2)$ instead of $\|\mathbf{x}_1-\mathbf{x}_3\|_p^2$ we'd like to have. It remains to see that the former expression is at least as large as the latter, and this follows by the convexity of the function $\mathbf{x}\mapsto\|\mathbf{x}\|_p^2$. Namely, we use $\frac12(\|\mathbf{a}\|_p^2+\|\mathbf{b}\|_p^2)\ge\|(\mathbf{a}+\mathbf{b})/2\|_p^2$ with $\mathbf{a}:=\mathbf{x}_2+\mathbf{x}_4-2\mathbf{x}_1$ and $\mathbf{b}:=2\mathbf{x}_3-\mathbf{x}_2-\mathbf{x}_4$. □

Proof of inequality (B.2). This exposition is based on a sketch given as the first proof of Proposition 3 in

> K. Ball, E. A. Carlen, and E. H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Invent. Math.* 115,1(1994) 463–482.

The second proof from that paper has been worked out by Assaf Naor; see `http://www.cims.nyu.edu/~naor/homepage/files/inequality.pdf`. I consider the first proof somewhat more conceptual and accessible for a non-expert.

First we pass to an inequality formally stronger than (B.2), with the same right-hand side:

$$\left(\frac{\|\mathbf{x}+\mathbf{y}\|_p^p+\|\mathbf{x}-\mathbf{y}\|_p^p}{2}\right)^{2/p}\ge\|\mathbf{x}\|_p^2+(p-1)\|\mathbf{y}\|_p^2. \qquad \text{(B.5)}$$

To see that the l.h.s. of (B.5) is never smaller than the l.h.s. of (B.2), we use the following well-known fact: The qth degree average $\left(\frac{a^q+b^q}{2}\right)^{1/q}$ is a nondecreasing function of q for a,b fixed. We apply this with $a=\|\mathbf{x}+\mathbf{y}\|_p^2$, $b=\|\mathbf{x}-\mathbf{y}\|_p^2$, $q=1$ and $q=p/2<1$, and we see that the new inequality indeed implies the old one. The computation with the new inequality is more manageable.

It is instructive to see what (B.5) asserts if the vectors $\mathbf{x}, \mathbf{y}$ are replaced by real numbers x, y. For simplicity, let us re-scale so that $x = 1$, and suppose that y is very small. Then the l.h.s. becomes $\left(\frac{(1+y)^p+(1-y)^p}{2}\right)^{2/p}$, and a Taylor expansion of this gives $(1 + p(p-1)y^2/2 + O(y^3))^{2/p} = 1 + (p-1)y^2 + O(y^3)$, while the r.h.s. equals $1 + (p-1)y^2$. So both sides agree up to the quadratic term, and in particular, we see that the coefficient $p-1$ in (B.5) cannot be improved.

The basic idea of the proof of (B.5) is this: With $\mathbf{x}$ and $\mathbf{y}$ fixed, we introduce an auxiliary real parameter $t \in [0,1]$, and we consider the functions $L(t)$ and $R(t)$ obtained by substituting $t\mathbf{y}$ for $\mathbf{y}$ in the left-hand and right-hand sides of (B.5), respectively. That is,

$$\begin{aligned} L(t) &:= \left(\frac{\|\mathbf{x}+t\mathbf{y}\|_p^p + \|\mathbf{x}-t\mathbf{y}\|_p^p}{2}\right)^{2/p} \\ R(t) &:= \|\mathbf{x}\|_p^2 + (p-1)t^2\|\mathbf{y}\|_p^2. \end{aligned}$$

Evidently $L(0) = R(0) = \|\mathbf{x}\|_p^2$. We would like to verify that the first derivatives $L'(t)$ and $R'(t)$ both vanish at $t = 0$ (this is easy), and that for the second derivatives we have $L''(t) \geq R''(t)$ for all $t \in [0,1]$, which will imply $L(1) \geq R(1)$ by double integration.

We have $R'(t) = 2(p-1)t\|\mathbf{y}\|_p^2$ (so $L(0) = 0$) and $R''(t) = 2(p-1)\|\mathbf{y}\|_p^2$.

For dealing with $L(t)$, it is convenient to write $f(t) := (\|\mathbf{x}+t\mathbf{y}\|_p^p + \|\mathbf{x}-t\mathbf{y}\|_p^p)/2$. Then

$$\begin{aligned} L'(t) &= \frac{2}{p} f(t)^{\frac{2}{p}-1} f'(t) \\ &= \frac{2}{p} f(t)^{\frac{2}{p}-1} \frac{p}{2} \sum_i \Big(|x_i + ty_i|^{p-1}\mathrm{sgn}(x_i+ty_i)y_i \\ &\qquad - |x_i - ty_i|^{p-1}\mathrm{sgn}(x_i - ty_i)y_i \Big) \end{aligned}$$

(we note that the function $z \mapsto |z|^p$ has a continuous first derivative, namely, $p|z|^{p-1}\mathrm{sgn}(z)$, provided that $p > 1$). The above formula for $L'(t)$ shows $L'(0) = 0$.

For the second derivative we have to be careful, since the graph of the function $z \mapsto |z|^{p-1}$ has a sharp corner at $z = 0$, and thus the function isn't differentiable at 0 for our range of p. We thus proceed with the calculation of $L''(t)$ only for t with $x_i \pm ty_i \neq 0$ for all i, which excludes finitely many values. Then

$$\begin{aligned} L''(t) &= \frac{2}{p}\left(\frac{2}{p} - 1\right) f(t)^{\frac{2}{p}-2} f'(t)^2 + \frac{2}{p} f(t)^{\frac{2}{p}-1} f''(t) \\ &\geq \frac{2}{p} f(t)^{\frac{2}{p}-1} f''(t) \\ &= f(t)^{\frac{2}{p}-1}(p-1)\left(\sum_i |x_i + ty_i|^{p-2} y_i^2 + \sum_i |x_i - ty_i|^{p-2} y_i^2 \right). \end{aligned}$$

Next, we would like to bound the sums in the last formula using $\|\mathbf{x}\|_p$ and $\|\mathbf{y}\|_p$. We use the so-called *reverse Hölder inequality*, which asserts, for nonnegative a_i's and strictly positive b_i's, $\sum_i a_i b_i \geq \left(\sum_i a_i^r\right)^{1/r}\left(\sum_i b_i^s\right)^{1/s}$, where $0 < r < 1$ and $\frac{1}{s} = 1 - \frac{1}{r} < 0$. This inequality is not hard to derive from the "usual" Hölder inequality $\sum_i a_i b_i \leq \|\mathbf{a}\|_p \|\mathbf{b}\|_q$, $1 < p < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$. In our case we use the reverse Hölder inequality with $r = p/2$, $s = p/(p-2)$, $a_i = y_i^2$, and $b_i = |x_i + ty_i|^{p-2}$ or $b_i = |x_i - ty_i|^{p-2}$, and we arrive at

$$L''(t) \geq (p-1) f(t)^{\frac{2}{p}-1} \|\mathbf{y}\|_p^2 \left(\|\mathbf{x}+t\mathbf{y}\|_p^{p-2} + \|\mathbf{x}-t\mathbf{y}\|_p^{p-2} \right)$$

Applying yet another inequality $\frac{a^\alpha + b^\alpha}{2} \geq \left(\frac{a+b}{2}\right)^\alpha$ with $a = \|\mathbf{x}+t\mathbf{y}\|_p^p$, $b = \|\mathbf{x}-t\mathbf{y}\|_p^p$, and $\alpha = (p-2)/p < 0$ (for $\alpha = -1$, for example, this is the inequality between the harmonic and arithmetic means), we get rid of the $f(t)$ term and finally obtain $L''(t) \geq 2(p-1)\|\mathbf{y}\|_p^2$.

We have thus proved $L''(t) \geq R''(t)$ for all but finitely many t. The function $L'(t) - R'(t)$ is continuous in $(0,1)$ and nondecreasing on each of the open intervals between the excluded values of t (by the Mean Value Theorem), and so $L'(t) \geq R'(t)$ for all t. The desired conclusion $L(1) \geq R(1)$ follows, again by the Mean Value Theorem. □