i-MATH Winter School

DocCourse
Combinatorics and Geometry 2009
Discrete and Computational Geometry

Centre de Recerca Matemàtica
January to March 2009

Intensive Courses

Lectures by

Jiří Matoušek (Univerzita Karlova, Praha)

Günter M. Ziegler (Technische Universität Berlin)

Edited by

Marc Noy (Universitat Politècnica de Catalunya)

Julian Pfeifle (Universitat Politècnica de Catalunya)

Centre de Recerca Matemàtica
Campus de Bellaterra, Edifici C
08193 Bellaterra (Barcelona)

First edition: July 2010

ISBN:
Legal deposit:

Contents

Metric Embeddings

Presentation

These volumes contain the materials produced during the 2009 edition of the *DocCourse in Discrete and Computational Geometry*, celebrated in Barcelona at the Centre de Recerca Matemàtica from January to March 2009.

The first volume contains the lecture notes of the two intensive courses, delivered by the main speakers and partly transcribed by the participants. These notes provide a quick introduction to their respective subjects and open numerous possibilities for future research.

The course *Metric Embeddings*, by Jiří Matoušek, deals with low distortion embeddings of metric spaces into (preferably low-dimensional) normed spaces. One of the prominent technical tools in this area is the well-known *Johnson–Lindenstrauss Lemma*. The course discusses some important practical applications of this result, in particular to pseudorandom number generators, and to the new and burgeoning field of *compressed sensing*. Moreover, the author presents intrinsic lower bounds on the distortion of the embedding.

Günter M. Ziegler highlights some significant families of convex polytopes in his course *Convex Polytopes: Examples and Conjectures*. He presents numerous examples, constructions, and properties of these families, gathers together in one place many of the most important conjectures and open problems in the field, and asks several new questions. This makes these course notes into an excellent source of inspiration for future research in the field.

The second volume comprises extended summaries of the twelve invited lectures delivered by senior researchers in the field, and transcribed by the students. They cover a wide variety of topics and techniques situated at the forefront of current research. The topics range from convex polytopes and convex bodies, and point and line configurations, to relationships with algebraic geometry, number theory, graph theory and combinatorial optimization.

A distinguishing feature of the course consisted in the abundance of research problems that were proposed to the participating students. Having them team together and try to shed light on these open questions proved to be a very effective way of stimulating an exciting and relevant research experience. The consensus opinion among both the students and the researchers who posed the problems speaks to the success of this idea.

The third and last volume collects the research results obtained by the students during the period of the course, and documents the state of affairs at the end of the program. Many of the drafts collected here are now being actively improved and polished, and some of them will no doubt soon find their way as articles into the published literature.

The editors are most grateful to the participants Edward D. Kim and Vincent Pilaud for the many hours they devoted to the production of these volumes. The excellent quality of the text owes much to their tireless and extremely conscientious work. This edition of DocCourse has been supported by the Spanish project i-MATH and by the Centre de Recerca Matemàtica, to which we express our gratitude. We particularly wish to thank the director, Professor Joaquim Bruna, and the staff of the CRM for their support and their excellent job in organizing this edition of DocCourse.

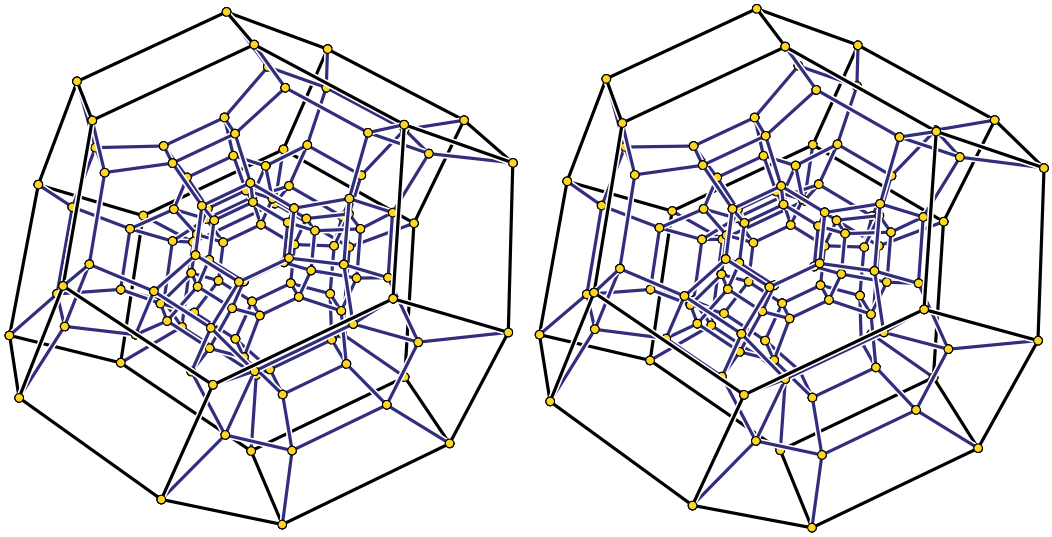
Marc Noy and Julian Pfeifle

Part I

Intensive Courses

Convex Polytopes: Examples and Conjectures

by Günter M. Ziegler



Lecture 1

Stacked Polytopes

1.1 Construction and counting

Definition 1.1.1 (Stacking onto a facet)**.** Let P be a d-polytope and F be a facet of P. The operation of *stacking onto* F consists of constructing the polytope $P' = P \cup (F \star p)$, where p is a point beyond the facet F but beneath all other facets of P, and $F \star p$ denotes the pyramid $\text{conv}(F \cup \{p\})$.

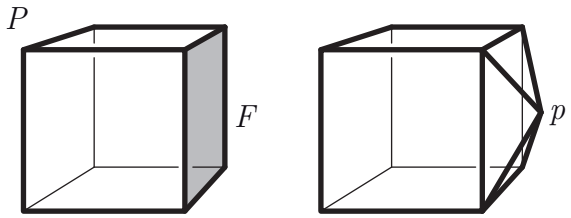


Figure 1.1: Stacking onto a facet of a polytope.

Observe that, during this operation, we destroy the facet F and create one new i-face for each $(i-1)$-face of F. In other words, the f-vector[1] of the resulting polytope P' is given by:

$$\begin{aligned} f_0(P') &= f_0(P) + 1, \\ f_i(P') &= f_i(P) + f_{i-1}(F) \quad \text{for all } 0 \leq i \leq d-2, \\ f_{d-1}(P') &= f_{d-1}(P) + f_{d-2}(F) - 1. \end{aligned}$$

Observe that the question of how to "find" p algorithmically highly depends on the presentation of the polytope P (i.e., on whether we know its vertex description or its facet description).

[1]The *f-vector* of a d-polytope P is the vector $f(P) = (f_0(P), \ldots, f_{d-1}(P))$, where $f_i(P)$ is the number of i-dimensional faces of P.

Definition 1.1.2 (Stacked polytopes)**.** A *stacked polytope* on $d+n$ vertices arises from a d-simplex by stacking $n-1$ times onto a facet ($n \geq 1$).

In other words, we obtain a (convex) tree of n d-simplices, and thus a stacked polytope is simplicial[2]. The f-vector of a stacked polytope on $d+n$ vertices is given by:

$$\begin{aligned} f_0 &= d+n, \\ f_i &= \binom{d}{i+1} + n\binom{d}{i} \quad \text{for all } 0 \leq i \leq d-2, \\ f_{d-1} &= 2 + n(d-1). \end{aligned}$$

Example 1.1.3. In dimension 2, every polytope is stacked (any triangulation of a convex polygon corresponds to a tree of triangles). In dimension 3, any cyclic polytope[3] is stacked, but the octahedron is not (since no vertex has degree 3, which should be the case of the last added vertex).

Lemma 1.1.4. *Let P be a d-polytope. The following assertions are equivalent:*

(i) *P is stacked;*

(ii) *P can be triangulated without new $(d-2)$-faces;*

(iii) *P is a tree of simplices.*

Furthermore, when $d \geq 3$, this triangulation without $(d-2)$-faces is unique.

Corollary 1.1.5. *The combinatorial type of a stacked polytope and its "shallow triangulation" are already determined by its graph (when $d \geq 3$).*

Observe that there exist many different combinatorial types of stacked polytopes (with same dimension and same number of vertices): The first example is given by the two stacked 3-polytopes with 7 vertices of Fig. 1.2.

Question 1.1.6. How many different (combinatorial types) of stacked d-polytopes with $d+n$ vertices are there?

---

[2]A polytope is *simplicial* if all its facets are simplices. A polytope is *simple* if its polar is simplicial, or equivalently, if all its vertex figures are simplices.

[3]The *cyclic polytope* with $n+1$ vertices in dimension d is the polytope

$$C_d(n+1) = \operatorname{conv}\{(i, i^2, \ldots, i^d)^T \mid i = 0, \ldots, n\}.$$

It is $\lfloor \frac{d}{2} \rfloor$-*neighborly*, meaning that any subset of $\lfloor \frac{d}{2} \rfloor$ of its vertices forms a face.
See Exercise 5.5.3 for a facet description and count.

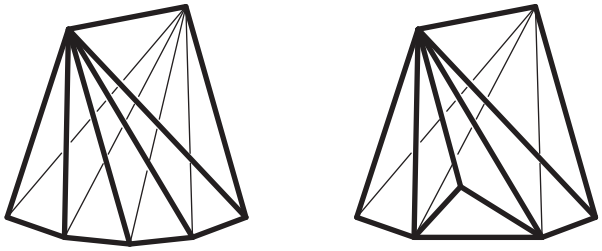


Figure 1.2: Two combinatorially different stacked 3-polytopes with 7 vertices.

As for many combinatorial objects, it is much easier to count *rooted stacked polytopes*, that is, stacked polytopes for which we have fixed one facet and labeled its vertices from 1 to d. Observe that when we fix a facet F of a stacked d-polytope (with $d \geq 3$), this automatically fixes a $(d+1)$-coloring of the vertices of P, and a $(d+1)$-coloring of the $(d-1)$-faces in the shallow tree of d-simplices.

Theorem 1.1.7. *There is a bijection between rooted stacked d-polytopes on $d+n$ vertices and plane d-ary trees with n internal nodes. In particular, the number of rooted stacked d-polytopes with $d+n$ vertices is given by the* Fuss–Catalan number*:*

$$S_{\text{rooted}}(d,n) = \frac{1}{(d-1)n+1}\binom{dn}{n}.$$

Corollary 1.1.8. *The number $S(d,n)$ of stacked d-polytopes with $n+d$ vertices is bounded by*

$$\frac{1}{d!(2+n(d-1))((d-1)n+1)}\binom{dn}{n} \leq S(d,n) \leq \frac{1}{(d-1)n+1}\binom{dn}{n}.$$

Question 1.1.9. How good is this bound?

1.2 The lower bound theorem

Stacked polytopes are important examples since they minimize the number of faces over all simplicial polytopes:

Theorem 1.2.1 (Lower Bound Theorem, Barnette 1971)**.** *A* simplicial *d-polytope P with m vertices has at least as many i-dimensional faces as any stacked d-polytope $\mathrm{St}_d(m-d)$ with m vertices (for all $0 \leq i \leq d-1$):*

$$f_i(P) \geq f_i(\mathrm{St}_d(m-d)).$$

Furthermore, equality holds for some $1 \le i \le d-1$ *if and only if* $d = 3$, *or* $d \ge 4$ *and* P *is stacked.*

Remarks 1.2.2 (On the proof(s) of the lower bound theorem).

(i) It suffices to prove it for $i = 1$ (McMullen–Perles–Walkup reduction), i.e., to prove the following statement: The number of edges of any simplicial d-polytope is a least that of a stacked d-polytope with the same number of vertices:

$$f_1 \ge df_0 - \binom{d+1}{2}.$$

(ii) It is related to Gromov rigidity: A simplicial complex X of dimension $d-1$ is *q-rigid* if and only if it is connected and if any set $A \subset V(X)$ that misses a facet of X hits at least $q|A|$ edges of X.

Let us mention the "opposite" theorem:

Theorem 1.2.3 (Upper Bound Theorem, McMullen 1970)**.** *A (simplicial) d-polytope P with m vertices has at most as many i-faces as the cyclic d-polytope $C_d(m)$ with m vertices (for all $0 \le i \le d-1$):*

$$f_i(P) \le f_i(C_d(m)).$$

Equality holds for some $\lfloor \frac{d}{2} \rfloor < i \le d-1$ if and only if P is neighborly[4].

1.3 Open problems

1.3.1 The generalized lower bound conjecture

Remember that we mentioned in Remark 1.2.2(i) that the lower bound theorem is equivalent to $f_1 \ge df_0 + \binom{d+1}{2}$, that is, in terms of h-numbers[5],

[4] A polytope is *k-neighborly* if any k of its vertices form a face (i.e., if its $(k-1)$-skeleton is combinatorially equivalent to that of the simplex with the same number of vertices). A d-polytope is said to be *neighborly* if it is $\lfloor \frac{d}{2} \rfloor$-neighborly.

[5] The *h-vector* of a polytope P is defined as $(h_0, \ldots, h_d)$, where

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \binom{d-i}{d-k} f_{i-1}.$$

For example, $h_0 = 1$, $h_1 = f_0 - d$, and $h_2 = f_1 - (d-1)f_0 + \frac{d(d-1)}{2}$. A nice way to express this relation is to collect the f- and h-numbers into the generating polynomials $F(x) = \sum_{i=0}^{d} f_{i-1} x^{d-i}$ and $H(x) = \sum_{i=0}^{d} h_i x^{d-i}$ (note the indexing!). Then the above relation simply reads $F(x) = H(x+1)$.

$h_1 \leq h_2$. McMullen and Walkup conjectured that not only this first inequality holds, but that

$$h_1 \leq h_2 \leq \ldots \leq h_{\lfloor \frac{d}{2} \rfloor}.$$

In terms of f-numbers, this yields the following formulation:

Conjecture 1.3.1 (McMullen–Walkup 1971)**.** *Let P be a simplicial d-polytope. Then, for any $0 \leq k \leq \lfloor \frac{d}{2} \rfloor + 1$,*

$$\sum_{j=-1}^{k} (-1)^{k-j} \binom{d-j}{d-k} f_j(P) \geq 0.$$

When $d \geq 4$, equality holds for some k $\iff$ P is k-stacked $\iff$ there is a triangulation without any interior $(d-k-1)$-face.

The first part of this conjecture is known, while the characterization remains open.

1.3.2 The non-simplicial case

The following theorem generalizes the lower bound theorem to non-simplicial d-polytopes:

Theorem 1.3.2 (Kalai 1988, Whiteley)**.** *The graph obtained by triangulating the 2-faces of a d-polytope is infinitesimally rigid*[6]*. In particular,*

$$f_1 + f_{02} - 3f_2 \geq df_0 - \binom{d+1}{2},$$

[6]Let $G = (V, E)$ be a graph, and $\phi\colon V \to \mathbb{R}^d$ be an embedding of its vertices into $\mathbb{R}^d$. A *motion* of G is a map $\psi\colon V \times [0,1] \to \mathbb{R}^d$ such that

(i) for all $v \in V$, $\psi(v,0) = \phi(v)$;

(ii) for all $v \in V$, the trajectory $t \mapsto \psi(v,t)$ of the vertex v is differentiable; and

(iii) for all $(v,w) \in E$, the distance between v and w is constant: for all $t \in [0,1]$,

$$\|\psi(v,t) - \psi(w,t)\| = \|\phi(v) - \phi(w)\|.$$

The graph G is *rigid* if any motion of G can be extended to an isometry of $\mathbb{R}^d$.

Looking at derivatives of motions leads to the notion of infinitesimal rigidity. An *infinitesimal motion* of G is a map $\tau\colon V \mapsto \mathbb{R}^d$ such that $\langle \phi(v) - \phi(w) \mid \tau(v) - \tau(w) \rangle = 0$ for all $(v,w) \in E$. The graph G is *infinitesimally rigid* if any infinitesimal motion corresponds to an isometry of $\mathbb{R}^d$. Observe that if G is rigid (or infinitesimally rigid), then

$$|E| \geq d|V| - \tbinom{d+1}{2}.$$

where f_{02} denotes the number of $(0,2)$-flags[7] *of P.*

Observe the following:

(i) If P is simplicial, then $f_{02} = 3f_2$, and we get back that $f_1 \geq df_0 + \binom{d+1}{2}$ (that is, the MPW-reduction of the lower bound theorem).

(ii) This is an equality when $d = 3$ (a triangulation of m vertices in the plane has exactly $3m - 6$ edges).

Question 1.3.3. Give a combinatorial proof (in other words, a correct proof for non-realizable spheres).

1.3.3 Universality

Even if some polytopes are not stacked, it is conjectured that stacked polytopes are sufficiently generic in the following sense:

Conjecture 1.3.4 (Kalai)**.** *Every d-polytope is a subpolytope*[8] *of a stacked polytope.*

Question 1.3.5. Is it true for $d = 3$? What about the octahedron? And the icosahedron?

1.3.4 Small coordinates

We have seen that there are many different combinatorial types of stacked polytopes in dimension 3. We would like to realize them by polytopes whose vertices have small coordinates:

Question 1.3.6. Can every combinatorial type of a stacked 3-polytope be realized with its vertices in $\{0, 1, \ldots, p(n)\}^3$ for some polynomial p?

1.3.5 Lower bound theorem for Delaunay polytopes

A *Delaunay polytope* is a polytope with all its vertices on a sphere.

Question 1.3.7. What is the minimal number of faces of a Delaunay polytope?

---

[7] Let $0 \leq i_1 < i_2 < \ldots < i_p \leq d$. An $(i_1, \ldots, i_p)$-*flag* of P is an increasing sequence $F_1 \subset F_2 \subset \ldots \subset F_p$ of faces of P of dimensions $i_1, \ldots, i_p$ respectively. We denote by $f_{i_1,\ldots,i_p}$ the number of $(i_1, \ldots, i_p)$-flags of P.

[8] A *subpolytope* of a polytope P is the convex hull of a subset of the vertices of P.

Lecture 2

2-simple 2-simplicial 4-polytopes

2.1 Face lattices of *d*-polytopes

We consider the *face lattice* L of a d-polytope P, that is, the set of all its faces, partially ordered by inclusion. This poset is in fact a graded lattice of length $d+1$.

Definition 2.1.1. The polytope P is *k-simplicial* if all its k-faces are simplices. Equivalently, for all $x \in L$ of rank at most $k+1$, the interval $[\hat{0}, x]$ is the Boolean lattice of rank k.

The polytope P is *h-simple* if every $(d-1-h)$-face is contained in exactly $h+1$ facets. Equivalently, for all x of corank at most $h+1$, the interval $[x, \hat{1}]$ is the Boolean lattice of rank k.

Remark 2.1.2. A d-simplicial d-polytope is a d-simplex. $(d-1)$-simplicial d-polytopes are exactly simplicial d-polytopes. Any polytope is 1-simplicial.

A d-simple d-polytope is a d-simplex. $(d-1)$-simple d-polytopes are exactly simple d-polytopes. Any polytope is 1-simple.

Proposition 2.1.3. *If $k+h > d$, then every k-simplicial h-simple d-polytope is a d-simplex. In particular, the first interesting example is that of* 2-simple 2-simplicial 4-polytopes (*2s2s 4-polytopes*).

Proposition 2.1.4. *Every 2s2s 4-polytope has a symmetric f-vector:*

$$f_0 = f_3 \quad \text{and} \quad f_1 = f_2.$$

The "classical" list of 2s2s 4-polytopes is really restricted: the simplex, the hypersimplex and its dual, and the 24-cell[1] are 2s2s 4-polytopes. The goal of the following section is to give a general method to construct 2s2s 4-polytopes, which provides an infinite family of such polytopes.

[1]The 24-*cell* is a regular 4-polytope each of whose 24 facets is an octahedron.

2.2 Constructing 2s2s 4-polytopes

Definition 2.2.1. Let P be a d-polytope. We call *deep vertex truncation* of P the polytope $DVT(P)$ obtained by cutting off all the vertices of P in such a way that, from each edge, exactly one point remains.

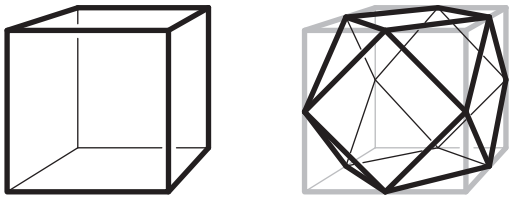


Figure 2.1: Deep vertex truncation of the 3-cube.

It is not always possible to realize the deep vertex truncation of a polytope. When it exists, $DVT(P)$ has one vertex for each edge of P, and two types of facets:

(i) the vertex figures of P, and

(ii) the deep vertex truncations of the facets of P.

Lemma 2.2.2. *If P is regular, then $DVT(P)$ can be constructed.*

Examples 2.2.3. We obtain interesting polytopes with $DVT(P)$ when P is:

1. the 4-simplex: $DVT(\triangle_4)$ is the hypersimplex —its facets are 5 tetrahedra (vertex figure of $\triangle_4$) and 5 octahedra ($DVT(\triangle_3)$);
2. the 4-dimensional cross-polytope: $DVT(C_4^*)$ is the 24-cell —it has 24 facets which are all octahedra (all the 8 vertex figures and the 16 DVT of facets of C_4^* give octahedra);
3. the 600-cell: we obtain a 4-polytope with 600 octahedral and 120 dodecahedral facets.

Theorem 2.2.4. *If P is a simplicial 4-polytope, and if $DVT(P)$ exists, then $DVT(P)$ is $2s2s$.*

Theorem 2.2.5. *Every combinatorial type of stacked d-polytope can be realized in such a way that the deep vertex truncation can be performed.*

Observe that if $\mathrm{St}_4(n)$ is a stacked 4-polytope with $4+n$ vertices, then the f-vector of the deep vertex truncation that we obtain is:

$$f(DVT(\mathrm{St}_4(n)) = (6+4n,\ 12+18n,\ 12+18n,\ 6+4n).$$

2.3 Open problems

2.3.1 Approximation

We would like to know whether there are enough 2s2s polytopes to approximate convex bodies:

Conjecture 2.3.1 (Shephard)**.** *Every convex body in* $\mathbb{R}^4$ *can be approximated by* 2*s*2*s* 4*-polytopes.*

2.3.2 Related polytopes

It remains open to find examples of the following generalizations of 2s2s polytopes:

Question 2.3.2. What about 2-cubical 2-cocubical 4-polytopes?

Question 2.3.3. What about ℓ-simple ℓ-simplicial polytopes, for $\ell \geq 3$?

For $\ell = 3$ and 4, only some sporadic examples are known:

1. half cubes are 3s3s (meaning the convex hull of every second vertex of the cube);
2. one non-trivial 4s4s polytope is known.

We have as yet no example (except for the simplex!) of a 5s5s polytope.

2.4 Selected exercises

Exercise 2.4.1. Show that any simple or simplicial d-polytope with $f_0 = f_{d-1}$ must be a simplex, or 2-dimensional.

Assume that P is a simplicial polytope with $f_0 = f_{d-1}$. Then any facet contains exactly d vertices, and thus $f_{0,d-1} = df_0 = df_{d-1} \leq f_{0,d-1}$, which implies that P is also simple. And only simplices and 2-dimensional polytopes are both simple and simplicial.

Exercise 2.4.2. What is the f-vector of a neighborly cubical 4-polytope[2]?

Let P be a neighborly cubical polytope. Since P has the graph of the n-cube (which is n-regular), we already know that $f_0 = 2^n$ and $f_1 = n2^{n-1}$. Moreover, all facets of P are cubes, which implies that $f_2 = 3f_3$. Using the Euler relation, we obtain that the f-vector of P is:

$$(2^n,\, n2^{n-1},\, 3(n-2)2^{n-2},\, (n-2)2^{n-2}).$$

[2]A *neighborly cubical d-polytope* is a d-polytope with the graph of the n-cube ($n \geq d$) and whose facets are $(d-1)$-cubes.

Exercise 2.4.3. Show that if a 4-polytope P is not simplicial, then its deep vertex truncation $DVT(P)$ cannot be 2-simplicial.

Assume that P is a 4-polytope such that $DVT(P)$ is 2-simplicial. Then the facets of $DVT(P)$ are simplicial. But $DVT(P)$ has two types of facets:

1. the vertex figures of P —this implies that the facets of P are simple;
2. the deep vertex truncations of the facets of P —this implies that the facets of P are simplicial.

Thus, the facets of P are both simple and simplicial, which ensures that they are simplices, and that P is simplicial.

Exercise 2.4.4. (a) Show that $f_{13} = f_{03} + 2f_2 - 2f_3$, and dually $f_{02} = f_{03} + 2f_1 - 2f_0$, hold for the flag vector of each 4-polytope.

(b) Derive from (a) that the inequality $2f_{03} \geq (f_1 + f_2) + 2(f_0 + f_3)$ is valid for all 4-polytopes, and that it is tight exactly for the 2-simple 2-simplicial 4-polytopes.

(a) Let P be a 4-polytope. For each facet F of P, we apply the Euler relation

$$f_2(F) - f_1(F) + f_0(F) - 2 = 0.$$

Summing these relations over all facets of P, we obtain

$$f_{23}(P) - f_{13}(P) + f_{03}(P) - 2f_3(P) = 0.$$

Since $f_{23}(P) = 2f_2(P)$ (each ridge is contained in exactly two facets), we obtain

$$f_{13}(P) = f_{03}(P) + 2f_2(P) - 2f_3(P).$$

The second relation is the same by duality.

(b) For any 2-face F of P, $f_0(F) \geq 3$ (with equality if and only if F is a triangle). Summing these inequalities over all 2-faces of P, we obtain that $f_{02}(P) \geq 3f_2(P)$, with equality if and only if P is 2-simplicial. Similarly, $f_{13}(P) \geq 3f_1(P)$ with equality if and only if P is 2-simple. Combining both, we have

$$f_{02}(P) + f_{13}(P) \geq 3f_2(P) + 3f_1(P),$$

with equality if and only if P is 2-simple 2-simplicial. Using the equalities of (a), this inequality can be transformed into:

$$2f_{03} \geq (f_1 + f_2) + 2(f_0 + f_3).$$

Lecture 3

Hypersimplices

3.1 0/1-polytopes

Definition 3.1.1. A 0/1-*polytope* in $\mathbb{R}^n$ is the convex hull of a subset of vertices of the 0/1-cube.

Similarly, one defines a ± 1-polytope to be the convex hull of a subset of the vertices of the ± 1-cube. These two families of polytopes are obviously affinely equivalent via the transformation $x \mapsto 2x - \mathbb{1}$.

Lemma 3.1.2. *Any hyperplane in $\mathbb{R}^n$ contains at most 2^{n-1} vertices of the 0/1-cube, with equality only for the hyperplanes defined by one of the equations*

$$x_i = 0, \quad x_i = 1, \quad x_i = x_j, \quad \text{or} \quad x_i = 1 - x_j$$

with $1 \leq i < j \leq n$.

Lemma 3.1.3. *Let n be an even integer. The number of vertices of the 0/1-cube on the hyperplane defined by the equation $x_1 + \cdots + x_n = \frac{n}{2}$ is the Catalan number*

$$\binom{n}{\frac{n}{2}} \sim \frac{2^n}{\sqrt{n}}.$$

We concentrate on four different properties of 0/1-polytopes:

3.1.1 "Many"

There are obviously $2^{2^n} - 1$ possible non-empty choices of subsets of $\{0,1\}^n$, but lots of them give equivalent polytopes. We are interested in the number of equivalence classes of 0/1-polytopes, for the following four notions of equivalence: we say that two 0/1-polytopes $P, Q \subset \mathbb{R}^n$ are

(i) 0/1-*equivalent* if there is an isometry of the 0/1-cube that moves P to Q;

(ii) *congruent* if there is an isometry of $\mathbb{R}^n$ that transforms P into Q;

(iii) *affinely equivalent* if there is an affine map that transforms P into Q;

(iv) *combinatorially equivalent* if P and Q have isomorphic face lattices.

Observe that

$$\text{0/1-equivalent} \Rightarrow \text{congruent} \Rightarrow \text{affinely equivalent} \Rightarrow \text{combinatorially equivalent},$$

but that all opposite implications are false (even if counterexamples are not completely obvious...).

The following table gives the number N_d of 0/1-equivalence classes of d-dimensional 0/1-polytopes for small d's:

d	0	1	2	3	4	5
N_d	1	1	2	12	349	1 226 525

Fig. 3.1 gives one representative for each of the 12 classes of 3-dimensional 0/1-polytopes.

Proposition 3.1.4 (Sarangarajan, Ziegler)**.** *There are more than* $2^{2^{d-2}}$ *different combinatorial types of* d*-dimensional* 0/1*-polytopes.*

Proof. The idea of the proof is to consider the 0/1-polytopes that contain all vertices of the bottom face of the cube, the vertices e_n and $\mathbb{1} = \sum e_i$, plus some other vertices, but that do not contain either $e_1 + e_n$ or $\mathbb{1} - e_1$. There are $2^{2^{n-1}-4}$ possible choices, and no more than $2^{n-1}(n-1)!$ can be 0/1-equivalent (since an isometry of the cube that transforms a polytope of our family into another one preserves the bottom face). This provides the desired bound. □

3.1.2 Volumes

The volume of a 0/1-polytope P is bounded by

$$\frac{1}{n!} \leq \operatorname{vol}(P) \leq 1.$$

The lower bound comes from the fact that any 0/1-polytope contains a 0/1-simplex S whose volume is

$$\operatorname{vol}(S) = \frac{1}{n!} \det V_S \geq \frac{1}{n!},$$

where V_S is the 0/1-matrix whose columns are the coordinates of the vertices of S.

Furthermore, these bounds are obviously tight:

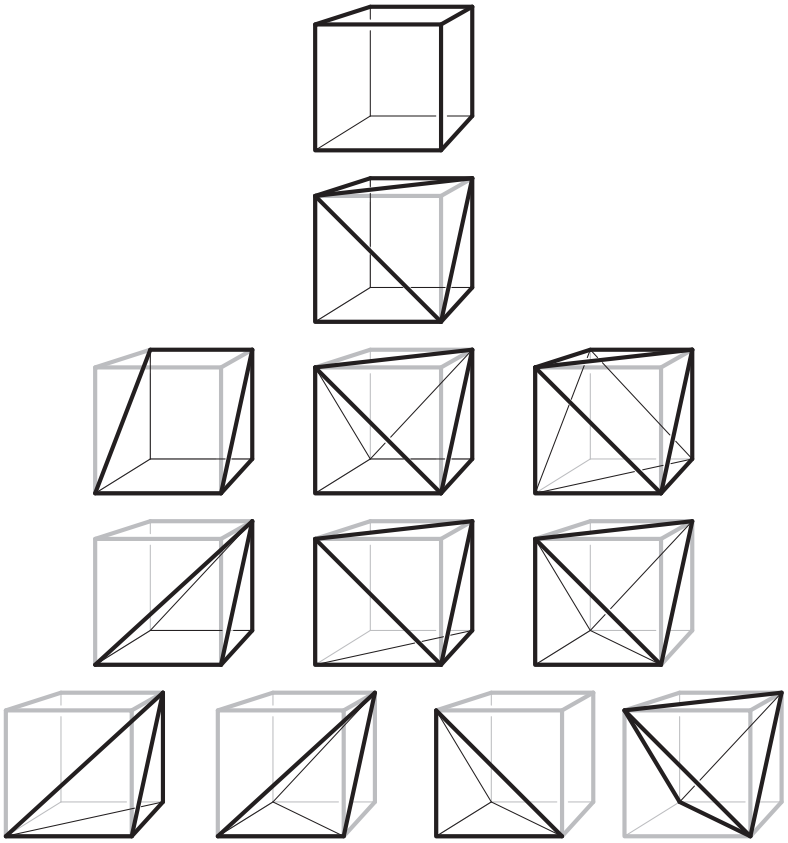


Figure 3.1: 0/1-equivalent classes of 0/1-polytopes of dimension 3.

(1) The 0/1-cube itself has volume 1.

(2) There exists a triangulation of the n-dimensional 0/1-cube into $n!$ simplices of volume $1/n!$ each. This triangulation can be described as follows:

   (i) either by the vertex description of its simplices —for each permutation σ of $\{1, \ldots, n\}$, we define the simplex

$$\triangle_\sigma = \operatorname{conv}\left\{0, e_{\sigma_1}, e_{\sigma_1} + e_{\sigma_2}, \ldots, \sum_{i=1}^{k} e_{\sigma_i}, \ldots, \sum_{i=1}^{n} e_{\sigma_i} = \mathbb{1}\right\}$$

   and we construct the triangulation T whose simplices are the simplices associated to all permutations,

   (ii) or by the cutting hyperplanes of the triangulation: the hyperplanes $x_i = x_j$ $(i \neq j)$ cut the 0/1-cube into the triangulation T.

Remarks 3.1.5. (i) This implies in particular that the maximum number of simplices of a triangulation of the 0/1-cube is $n!$. The minimum number of simplices in a triangulation of the 0/1-cube is not known.

(ii) For any 0/1-matrix A,

$$\det(A) = \frac{1}{2^n} \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & & & \\ \vdots & & 2A & \\ 0 & & & \end{pmatrix} = \frac{1}{2^n} \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ -1 & & & \\ \vdots & & 2A - J & \\ -1 & & & \end{pmatrix},$$

where $J = \mathbb{1}\mathbb{1}^T$ denotes the matrix with only 1 entries.

This last determinant can be bounded by the Hadamard bound (that says that the volume of a parallelepiped is at most the product of the length of its edges, with equality if and only if the edges are orthogonal). We obtain that

$$\det(A) \leq \frac{1}{2^n} \left(\sqrt{n+1}\right)^{n+1},$$

with equality for Hadamard matrices[1].

Consequently, the volume of any 0/1-simplex S is bounded by

$$\frac{1}{n!} \leq \mathrm{vol}(S) \leq \frac{(n+1)^{\frac{n+1}{2}}}{n! 2^n},$$

and equality is possible only for $n = 0$, $n = 1$ and $n \equiv -1 \bmod 4$.

3.1.3 Simplicial 0/1-polytopes

The maximum number of vertices of a d-dimensional 0/1-polytope is obviously 2^d (cube). But what about simplicial 0/1-polytopes?

Conjecture 3.1.6. *A d-dimensional simplicial 0/1-polytope has at most $2d$ vertices and 2^d facets, with equality only if P is a centrally symmetric cross-polytope.*

3.1.4 Number of facets

Theorem 3.1.7 (Bárány–Por, Fleiner–Kaibel–Rote)**.** *The number of facets $f(d)$ of a d-dimensional 0/1-polytope is bounded by*

$$\left(\frac{cd}{\log d}\right)^{\frac{d}{4}} \leq f(d) \leq 30(d-2)!$$

[1] A *Hadamard matrix* is a square ± 1-matrix with mutually orthogonal rows. The order of a Hadamard matrix must be 1, 2, or a multiple of 4.

for a certain constant $c \in \mathbb{R}$.

Here we prove only the following upper bound, which is slightly weaker:

Proposition 3.1.8 (Bárány)**.** $f(d) \leq 2(d-1) + 2(d-1)!$

We need the following lemma:

Lemma 3.1.9. *For any* 0/1*-polytope* P, $f_{n-1}(P) \leq 2n + n!(1 - \mathrm{vol}(P))$.

Proof. The result is true for the 0/1-cube. Starting from our polytope P, we can insert one by one vertices of $\{0,1\}^n \setminus P$. At each step, we destroy some facets but we add pyramids over these facets. We derive the bound from the fact that every pyramid has volume at least $1/n!$. □

Proof of Proposition 3.1.8. Let P be a d-dimensional 0/1-polytope. Let $f_{d-1}^{\mathrm{lower}}(P)$, $f_{d-1}^{\mathrm{upper}}(P)$, and $f_{d-1}^{\mathrm{vert}}(P)$ denote the number of lower, upper and vertical facets respectively (according to the last coordinate), and let $\bar{P}$ denote the projection of P to the first $d-1$ coordinates. Then

$$f_{d-1}^{\mathrm{lower}}(P) \leq (d-1)!\,\mathrm{vol}(\bar{P}), \quad f_{d-1}^{\mathrm{upper}}(P) \leq (d-1)!\,\mathrm{vol}(\bar{P}),$$

and, according to Lemma 3.1.9,

$$f_{d-1}^{\mathrm{vert}}(P) \leq f_{d-2}(\bar{P}) \leq 2(d-1) + (d-1)!(1 - \mathrm{vol}(\bar{P})).$$

Summing up,

$$\begin{aligned} f_{d-1}(P) &= f_{d-1}^{\mathrm{lower}}(P) + f_{d-1}^{\mathrm{upper}}(P) + f_{d-1}^{\mathrm{vert}}(P) \\ &\leq 2(d-1) + (d-1)!(1 + \mathrm{vol}(\bar{P})) \\ &\leq 2(d-1) + 2(d-1)! \end{aligned}$$

□

3.2 Hypersimplices

Definition 3.2.1. For any integers $1 \leq k \leq n-1$, the *hypersimplex* $\triangle_{n-1}(k)$ is given by

$$\triangle_{n-1}(k) = \{x \in [0,1]^n \mid \langle \mathbb{1} \mid x \rangle = k\} = \mathrm{conv}\{x \in \{0,1\}^n \mid \langle \mathbb{1} \mid x \rangle = k\}.$$

Example 3.2.2. $\triangle_{n-1}(1)$ is the usual simplex $\triangle_{n-1} = \mathrm{conv}\{e_i \mid 1 \leq i \leq n\}$. $\triangle_{n-1}(n-1)$ is another embedding of the $(n-1)$-dimensional simplex.

Remarks 3.2.3. 1. Observe that $\triangle_{n-1}(k)$ is an $(n-1)$-dimensional polytope. Furthermore,

(i) from the facet description $\triangle_{n-1}(k) = \{x \in [0,1]^n \mid \langle \mathbb{1} \mid x\rangle = k\}$, it is easy to see that its number of facets is

$$f_{n-2}(\triangle_{n-1}(k)) = \begin{cases} n & \text{for } k \in \{1, n-1\}, \\ 2n & \text{otherwise.} \end{cases}$$

Let us insist on the fact that, when $2 \le k \le n-2$, there are two types of facets: we have n facets of type $\triangle_{n-2}(k)$ and n facets of type $\triangle_{n-2}(k-1)$;

(ii) from the vertex description

$$\triangle_{n-1}(k) = \operatorname{conv}\{x \in \{0,1\}^n \mid \langle \mathbb{1} \mid x\rangle = k\},$$

we get that it has $f_0 = \binom{n}{k}$ vertices. All vertices are symmetric.

2. $\triangle_{n-1}(k)$ and $\triangle_{n-1}(n-k)$ are affinely equivalent via $x \mapsto 1-x$. For this reason, if n is even, then $\triangle_{n-1}\left(\frac{n}{2}\right)$ is centrally symmetric[2].

3.2.1 f-vector

There is a correspondence between the faces of the n-dimensional 0/1-cube and the partitions of $[n] = \{1, \dots, n\}$ into three parts: the face associated to the partition $[n] = A \uplus B \uplus C$ is the $|C|$-face supported by the intersection of the hyperplanes of equations $x_a = 0$ $(a \in A)$ and $x_b = 1$ $(b \in B)$. Such a face will intersect the hyperplane of equation $\langle \mathbb{1} \mid x\rangle = k$ —and thus will contribute to a $(|C|-1)$-face of the hypersimplex $\triangle_{n-1}(k)$— if and only if $|A| < k$ and $|B| < n-k$. This yields the following formula for the number of $(i-1)$-faces of the hypersimplex $\triangle_{n-1}(k)$:

$$\begin{aligned} f_{i-1}(\triangle_{n-1}(k)) &= \left|\{[n] = A \uplus B \uplus C \,:\, |A| < k,\, |B| < n-k,\, |C| = i\}\right| \\ &= \sum_{\substack{0 \le s < k \\ k < s+i \le n}} \binom{n}{s}\binom{n-s}{i} \\ &= \sum_{\max(-1,\, k-i) < s < \min(k,\, n-i+1)} \frac{n!}{s!i!(n-s-i)!}. \end{aligned}$$

Since this formula is not completely explicit, it is interesting to consider the intermediate hypersimplex $\triangle_{n-1}\left(\frac{n}{2}\right)$ (for n large and even). It is easy to

[2] A polytope P is *centrally symmetric* if $P = -P$.

see that almost all faces of the hypercube (i.e., about 3^n) contribute to faces of this hypersimplex. This number 3^n is surprisingly low, according to the following conjecture:

Conjecture 3.2.4 (Kalai)**.** *Every centrally symmetric d-dimensional polytope has at least 3^d faces.*

In fact, an even stronger result was conjectured:

Conjecture 3.2.5 (Kalai)**.** *Every centrally symmetric polytope has an f-vector that is componentwise larger than the f-vector of a Hanner polytope*[3] *of the same dimension.*

This last conjecture was recently disproved using some tools that we will develop in the last lecture (see Section 5.4).

3.2.2 Volume

We want to compute the volume of the full-dimensional version of the hypersimplex:

$$\begin{aligned}\bar{\triangle}_{n-1}(k) &= \{x \in [0,1]^{n-1} \mid k-1 \leq \langle \mathbb{1} \mid x\rangle \leq k\} \\ &= \operatorname{conv}\{x \in \{0,1\}^{n-1} \mid \langle \mathbb{1} \mid x\rangle \in \{k-1,k\}\}.\end{aligned}$$

For this, we consider the *alcoved polytope*

$$C'_{n-1} = \{y \in \mathbb{R}^{n-1} \mid 0 \leq y_i - y_{i-1} \leq 1,\ \forall i \in [n-1]\}$$

(where $y_0 = 0$ by convention). This polytope is mapped to the standard cube $C_n = [0,1]^{n-1}$ via the map $\gamma\colon C'_{n-1} \to C_{n-1}$ given by

$$y = (y_1, \ldots, y_{n-1}) \longmapsto (y_1 - \lfloor y_1 \rfloor, \ldots, y_{n-1} - \lfloor y_{n-1} \rfloor).$$

We consider the triangulations of C'_{n-1} and C_{n-1} induced by the cutting hyperplanes $x_i = c$ and $x_i = x_j + c$ (where $1 \leq i < j \leq n-1$ and $c \in \mathbb{N}$). The mapping γ moves the simplices forming C'_{n-1} into the simplices forming C_{n-1}, and it turns out that the simplices that form the hypersimplex $\bar{\triangle}_{n-1}(k)$ in C_{n-1} come from simplices in C'_{n-1} corresponding to permutations with exactly k descents. Omitting details, we obtain:

[3]The class of *Hanner polytopes* is defined recursively as follows:

(i) any line segment is a Hanner polytope;

(ii) any polytope that can be written as the Cartesian product, or as the direct sum of two Hanner polytopes is a Hanner polytope.

For example, the Hanner 3-polytopes are the 3-cube and the octahedron. The Hanner 4-polytopes are the 4-cube, the 4-cross-polytope, the prism over an octahedron and the bipyramid over the 3-cube.

Theorem 3.2.6. *Let $A(n,k)$ denote the* Eulerian numbers*, that is, the number of permutations of $\{1,\dots,n\}$ with exactly k descents. Then the volume of the full-dimensional hypersimplex is given by*

$$\operatorname{vol}(\bar{\triangle}_{n-1}(k)) = \frac{A(n,k)}{(n-1)!}.$$

3.3 Selected exercises

Exercise 3.3.1. Classify the 0/1-polytopes of diameter $\sqrt{2}$.

Apart from the two particular cases of the square and the tetrahedron of Fig. 3.2, the only 0/1-polytopes of diameter $\sqrt{2}$ are the simplices of the form $\operatorname{conv}\{0, e_1, \dots, e_n\}$ (and their 0/1-equivalence class).

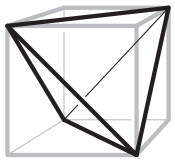


Figure 3.2: The only 0/1-polytope with diameter $\sqrt{2}$ and volume $\frac{1}{3}$.

Lecture 4

Associahedra

4.1 Four combinatorial structures

Let $a_1, \ldots, a_n$ denote n different letters.

4.1.1 Permutahedron

We consider the graph

- whose vertices are the permutations of these n letters,
- and whose edges are the pairs of *adjacent* permutations, that is, of permutations that differ by a transposition of two adjacent letters.

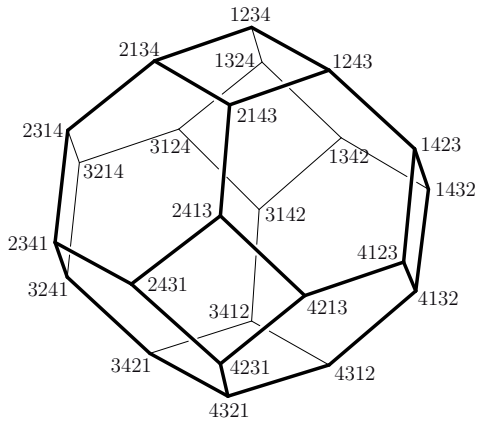


Figure 4.1: Permutahedron Π_3.

This graph has obviously $n!$ vertices and is regular[1] of degree $n-1$. It turns out that it is the graph of a simple $(n-1)$-polytope, called the *permutahedron* Π_{n-1} (see Fig. 4.1).

4.1.2 Associahedron

We consider the graph

- whose vertices are the bracketings[2] of our n letters,
- and whose edges are the pairs of bracketings of the form $m(no)$ and $(mn)o$ (where m, n and o are words forming valid bracketings).

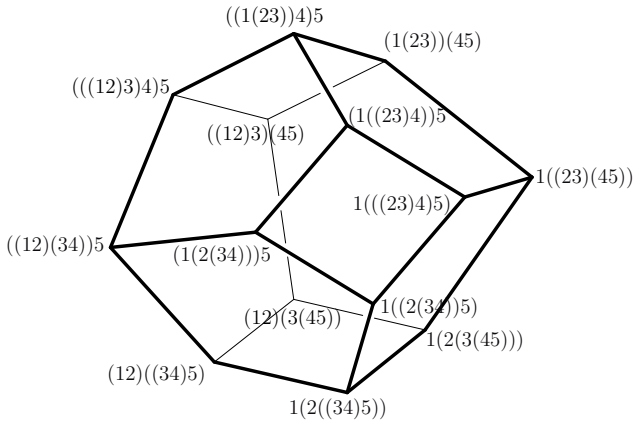


Figure 4.2: Associahedron A_3.

This graph has $C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}$ vertices and is regular of degree $n-2$. It turns out that it is the graph of a simple $(n-2)$-polytope, called the *associahedron* A_{n-1} (see Fig. 4.2).

4.1.3 Cyclohedron

We consider the graph

- whose vertices are the bracketings of cyclic permutations of our n letters,

[1] A graph is *regular* of degree d if all its vertices have degree d.

[2] A *bracketing* of n letters is a sequence of $2n-4$ parentheses ($n-2$ left and $n-2$ right) such that any of its prefixes contains more left parentheses than right parentheses. For example, 1(2(34)) and (1(23))4 are bracketings of 1234.

- and whose edges are (1) the pairs of vertices of the form $m(no)$ and $(mn)o$ (where m, n and o are words forming valid bracketings and where the letters are in the same order), and (2) pairs of vertices of the form mn and nm (where m and n are words forming valid bracketings).

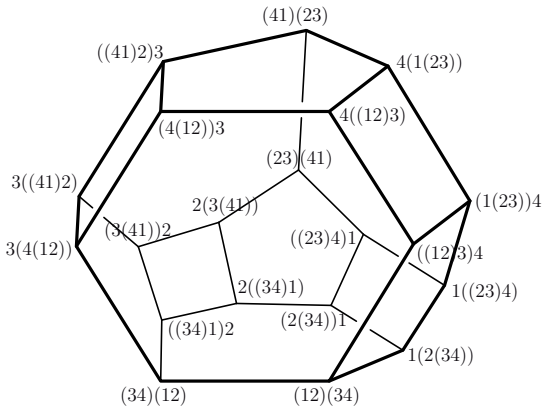


Figure 4.3: Cyclohedron Γ_3.

This graph has $nC_{n-1} = \binom{2n-2}{n-1}$ vertices and is regular of degree $n-1$ (thus, $n-2$ from rebracketing and 1 from cyclic permutation). It turns out that it is the graph of a simple $(n-1)$-polytope, called the *cyclohedron* Γ_{n-1} (see Fig. 4.3).

4.1.4 Permuto-associahedron

Finally, we consider the graph

- whose vertices are the bracketings of permutations of our n letters,
- and where a vertex is related with all vertices obtained either (1) by rebracketing without changing the order of the letters (for example, $1(2(34)) \to 1((23)4)$), or (2) by inverting two letters that are adjacent and bracketed together (for example, $1(2(34)) \to 1(2(43))$).

This graph has $n!C_{n-1} = (n-1)!\binom{2n-2}{n-1}$ vertices. Furthermore, all its vertices have degree at least $n-1$ and at most $n-2+\left\lfloor \frac{n}{2} \right\rfloor$. It turns out that it is the graph of a simple $(n-1)$-polytope, called the *permuto-associahedron* $K\Pi_{n-1}$ (see Fig. 4.4).

4.2 Polytopes from graphs

Lemma 4.2.1. *If G is the graph of a d-polytope, then*

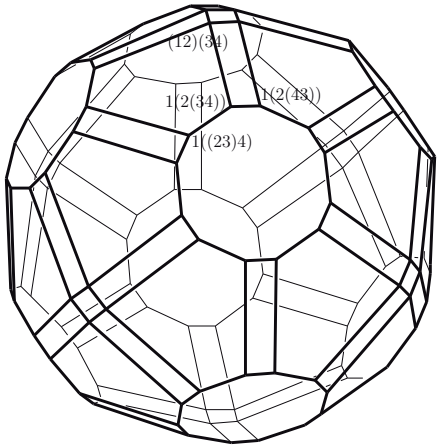


Figure 4.4: Permuto-associahedron $K\Pi_3$.

(i) *G is d-connected*[3] *(Balinski's Theorem);*

(ii) *for every vertex $v \in G$, there is a subdivision*[4] *of K_{d+1} contained in G which has v and d of its neighbors as its principal vertices;*

(iii) *if $d \leq 3$, then G is planar;*

(iv) *G is d-regular if and only if P is simple.*

In dimension 2, the graphs of polytopes are just the cycles. In dimension 3, Steinitz' Theorem characterizes graphs of polytopes:

Theorem 4.2.2 (Steinitz)**.** *A graph is the graph of a 3-polytope if and only if it is simple, planar and 3-connected.*

In higher dimension, however, we do not have characterizations of polytopal graphs. Small examples are enough to convince one that it is sometimes complicated to decide whether a given graph is or not the graph of a polytope:

Examples 4.2.3. 1. We consider the complete graph K_8 from which we remove a perfect matching. It has 8 vertices and constant degree 6. Thus, it is not planar, and if it is the graph of a d-polytope, then certainly $d \in \{4, 5, 6\}$. It turns out that it is the graph of the 4-dimensional cross-polytope and of the join of two squares (dimension 5). However, according to Lemma 4.2.1(ii), it is not realizable in dimension 6.

[3]A graph is *d-connected* if it is not possible to disconnect the graph by removing d vertices.

[4]A *subdivision* of a graph G is a graph obtained from G by replacing some of its edges by chains of edges. The principal vertices of this subdivision are the initial vertices of the graph G.

2. We consider the complete graph K_6 from which we remove the cycle C_6. It has 6 vertices and constant degree 3. Thus it can only be realized in dimension 3, and it is the graph of a prism.

3. We consider the complete graph K_7 from which we remove the cycle C_7. It is not planar (it contains a subdivision of $K_{3,3}$), but it does not contain any subdivision of K_5. By Lemma 4.2.1(ii), it is not the graph of a polytope.

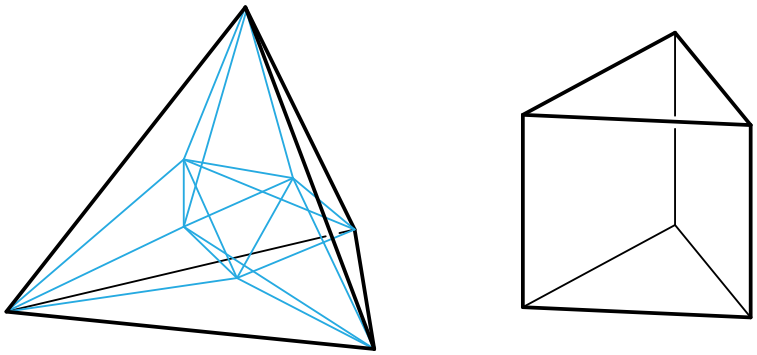


Figure 4.5: The Schlegel diagram of the 4-dimensional cross-polytope and a prism over a triangle (realizing the graphs of Examples 1 and 2 respectively).

In the following result, we only deal with regular graphs:

Lemma 4.2.4. *The graph G of a simple polytope, together with the* facial cycles *(that is, the cycles in G that correspond to 2-faces of P) determine the combinatorics of P.*

Theorem 4.2.5 (Blind–Mani, Kalai, Friedman)**.** *From the graph G of a simple polytope, the combinatorics of P can be reconstructed in polynomial time.*

4.3 Associahedra via fiber polytopes

Definition 4.3.1. Let $\pi\colon \mathbb{R}^d \to \mathbb{R}^e$ denote a projection, P be a d-polytope and Q be its image under the projection π. The *fiber polytope* $\Sigma(P \to Q)$ is the polytope defined by

$$\Sigma(P \to Q) = \left\{ \frac{1}{\mathrm{vol}(Q)} \int_Q \gamma(x)dx \ : \ \gamma \text{ section of } \pi \right\}$$

(where a section of π is a map $\gamma\colon \mathbb{R} \to \mathbb{R}^d$ such that $\pi \circ \gamma = \mathrm{Id}$).

Theorem 4.3.2. *The fiber polytope $\Sigma(P \to Q)$ is a $(d-e)$-dimensional polytope, included in the intersection of P with the preimage of the barycenter of Q. Its vertices correspond to "tight regular polyhedral strings in P", which project down to the finest regular subdivisions of Q by faces projected from P.*

Examples 4.3.3. 1. Here we fix $P = [0,1]^n$, $Q = [0,n]$, and $\pi\colon x \mapsto \langle \mathbb{1} \mid x\rangle$ (see Fig. 4.6). We obtain one vertex for each path from the source to the sink on the edges of P, that is, one vertex for each permutation. And we get one edge if two paths can be transformed one to the other by changing them along a facet of P, that is, if the two corresponding permutations differ by a transposition of two adjacent entries. Thus, this construction leads to the permutahedron (see Fig. 4.1).

2. The second interesting example is obtained with

$$P = \triangle_n = \operatorname{conv}\{0,\, e_1,\, e_1+e_2, \ldots, e_1+e_2+\cdots+e_n\},$$

$$Q = C_2(n+1) = \operatorname{conv}\left\{ \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}2\\4\end{pmatrix}, \begin{pmatrix}3\\9\end{pmatrix}, \ldots, \begin{pmatrix}n\\n^2\end{pmatrix} \right\},$$

$$\text{and} \quad \pi\colon \mathbb{R}^n \longrightarrow \mathbb{R}^2,\ e_i \longmapsto \begin{pmatrix}i\\i^2\end{pmatrix} - \begin{pmatrix}i-1\\(i-1)^2\end{pmatrix} = \begin{pmatrix}1\\2i-1\end{pmatrix}.$$

Then the fiber polytope $\Sigma(\triangle_n, C_2(n+1))$ is a realization of the $(n-2)$-dimensional associahedron (see Fig. 4.2).

4.4 Selected exercises

Exercise 4.4.1. Are the following graphs polytopal?

1. The *circulant graph* on vertex set $\mathbb{Z}_n$ with edges $(i, i+1)$ and $(i, i+2)$.
2. The *Petersen graph* (see Fig. 4.7).
3. The product of two Petersen graphs (see Fig. 4.7).

1. Let G_n denote the circulant graph on $\mathbb{Z}_n$ with edges $(i, i+1)$ and $(i, i+2)$. Observe first that, if $n \le 5$, then G_n is the complete graph on n vertices, and thus it is realized by the $(n-1)$-simplex. Furthermore, when $n = 2m$ is even, it is easy to see that the following 3-polytope realizes G_n (see Fig. 4.8):

$$P_{2m} = \operatorname{conv}\left(\left\{ \begin{pmatrix} \cos\left(\frac{2(2i-1)\pi}{m}\right) \\ \sin\left(\frac{2(2i-1)\pi}{m}\right) \\ 0 \end{pmatrix} \;:\; 1 \le i \le m \right\} \cup \left\{ \begin{pmatrix} \cos\left(\frac{4i\pi}{m}\right) \\ \sin\left(\frac{4i\pi}{m}\right) \\ 1 \end{pmatrix} \;:\; 1 \le i \le m \right\} \right).$$

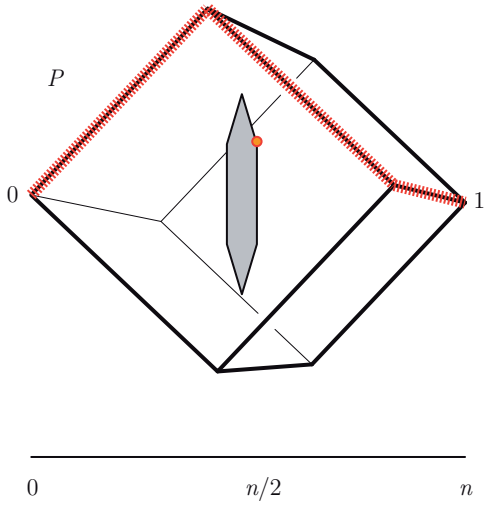


Figure 4.6: The permutahedron as the fiber polytope $\Sigma([0,1]^n, [0,1])$.

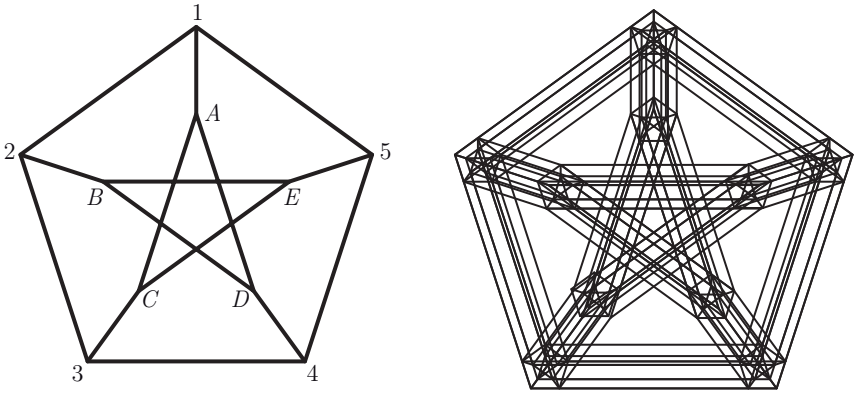


Figure 4.7: The Petersen graph and the cartesian product of two Petersen graphs.

We prove that there is no other realization of G_n. In particular, when n is odd and ≥ 7, the graph G_n is not polytopal. Observe first that G_n is regular of degree 4, and thus it cannot be realized in dimension different from 3 or 4.

It is easy to see that, if $n \geq 6$, the vertices $1, 2, 3, 4, 5$ are not principal vertices of a K_5 subdivision of G_n. Indeed, since 14, 25 and 15 are not in G_n, we need three paths of edges passing from $\{1, 2\}$ to $\{4, 5\}$. It is easy to see that either two of them pass through the vertex 6, or two of them pass through the edge $\{5, 7 \mod n\}$. Consequently, by Lemma 4.2.1(ii), G_n is not realizable in dimension 4 (except G_5).

Finally, when n is odd, G_n is not planar since it contains a subdivision of $K_{3,3}$ (see Fig. 4.8). Thus, it cannot be realized in dimension 3.

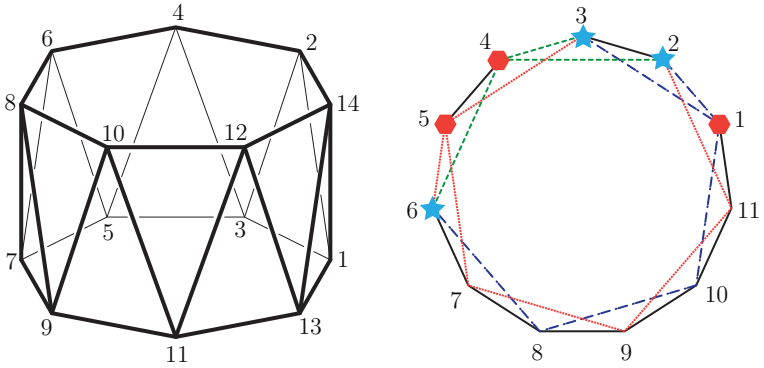


Figure 4.8: The polytope P_{14} realizing the circulant graph G_{14} in dimension 3, and a $K_{3,3}$ subdivision in the circulant graph G_{11}.

2. The Petersen graph is regular of degree 3 and non planar (there is a subdivision of $K_{3,3}$ with vertices $U = \{1, 2, 3\}$ and $V = \{B, D, E\}$). Thus, it is not the graph of a polytope.

Exercise 4.4.2. How many vertices, edges and facets has the $(n-1)$-dimensional permutahedron? How many vertices, edges and facets has the $(n-2)$-dimensional associahedron?

The $(n-1)$-dimensional permutahedron has one vertex for each of the $n!$ permutations of $\{1, \ldots, n\}$. It is regular of degree $n-1$, which implies that it has $n!(n-1)/2$ edges. Finally, its facets correspond to all the vertices of the cube that are neither the source 0 nor the sink $\mathbb{1}$ (in Fig. 4.6): the vertex set of a facet corresponds to all the paths on the edges of the cube that pass through a given vertex of the cube. Thus, the permutahedron has $2^n - 2$ facets.

As far as the $(n-2)$-dimensional associahedron is concerned, it has one vertex for each of the $\frac{1}{n}\binom{2n-2}{n-1}$ triangulations of the n-gon. It is regular of degree $n-2$, which implies that it has $\frac{n-2}{2n}\binom{2n-2}{n-1}$ edges. Finally, it has one facet for each of the $n(n-1)/2$ internal diagonals of the n-gon.

To sum up:

	vertices	edges	facets
permutahedron	$n!$	$\frac{n!(n-1)}{2}$	$2^n - 2$
associahedron	$\frac{1}{n}\binom{2n-2}{n-1}$	$\frac{n-2}{n}\binom{2n-2}{n-1}$	$\frac{n(n-1)}{2}$

Exercise 4.4.3. Compute the area, the barycenter, and the center of mass of the $(n+1)$-gon

$$C_2(n+1) = \operatorname{conv}\left\{ \binom{0}{0}, \binom{1}{1}, \binom{2}{4}, \binom{3}{9}, \dots, \binom{n}{n^2} \right\}.$$

Remember that, for any $1 \leq i \leq n$, the area of the triangle $\binom{0}{0}\binom{i}{i^2}\binom{j}{j^2}$ is given by the following determinant:

$$\operatorname{Area}\left(\binom{0}{0}\binom{i}{i^2}\binom{i+1}{(i+1)^2} \right) = \frac{1}{2} \det \begin{pmatrix} 1 & 1 & 1 \\ 0 & i & i+1 \\ 0 & i^2 & (i+1)^2 \end{pmatrix} = \frac{i(i+1)}{2}.$$

Consequently, we obtain

$$\operatorname{Area}(C_2(n+1)) = \sum_{i=1}^{n-1} \frac{i(i+1)}{2} = \frac{(n-1)n(n+1)}{6},$$

$$\operatorname{Bary}(C_2(n+1)) = \frac{1}{n+1} \sum_{i=0}^{n} \binom{i}{i^2} = \binom{n/2}{n(2n+1)/6},$$

$$\operatorname{CM}(C_2(n+1)) = \frac{1}{\operatorname{Area}(C_2(n+1))} \sum_{i=1}^{n-1} \frac{i(i+1)}{6} \binom{2i+1}{2i^2+2i+1} = \binom{n/2}{(6n^2+1)/15}.$$

Lecture 5

f-vector Shapes

Let P be a d-polytope and $f = f(P) = (f_0(P), \ldots, f_{d-1}(P))$ denote its f-vector (remember that $f_i = f_i(P)$ denotes the number of i-faces of P).

Definition 5.0.4. The *shape* of a d-polytope P is the function ϕ_P: "$[0,1] \to \mathbb{N}$" defined by

$$\phi_P(x) = f_{x(d-1)}(P).$$

We would like to answer the following question:

Question 5.0.5. What does ϕ_P typically look like?

Various conjectures have been made on the shapes of f-vectors of polytopes:

Conjecture 5.0.6 (Unimodality: Motzkin 1950, Welsh 1972)**.** *For any polytope P, ϕ_P is* unimodal*. In other words, there exists no i such that $f_{i-1} > f_i$ and $f_i < f_{i+1}$.*

Conjecture 5.0.7 (Partial Unimodality: Björner)**.** *For any polytope P, ϕ_P is increasing between 0 and $\frac{1}{4}$ and decreasing between $\frac{3}{4}$ and 1.*

Conjecture 5.0.8 (Minimal Entry: Bárány)**.** *For any polytope P and any $0 \leq i \leq d-1$,*

$$f_i \geq \min\{f_0, f_{d-1}\}.$$

5.1 Examples

5.1.1 Simplex

The f-vector of the $(d-1)$-simplex is given by

$$f_{i-1}(\triangle_{d-1}) = \binom{d}{i}.$$

Thus, $\phi_{\triangle_{d-1}}(x) = \binom{d}{x(d-1)} \sim \frac{\left(\frac{d}{e}\right)^d}{\left(\frac{xd}{e}\right)^{xd}\left(\frac{(1-x)d}{e}\right)^{(1-x)d}} = \left(\frac{1}{x^x(1-x)^{1-x}}\right)^d$.

We obtain a peak around $\frac{1}{2}$ which gets sharper when d grows (see Fig. 5.1(a)).

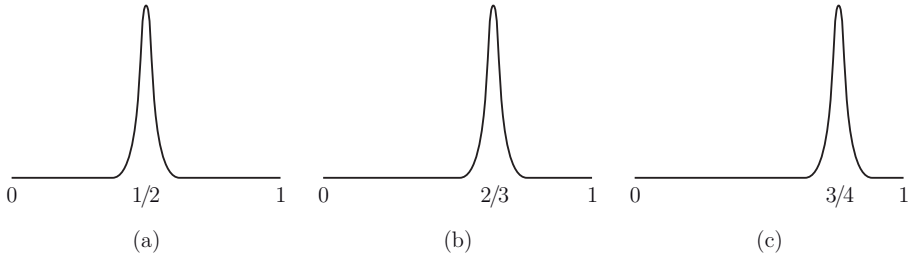


Figure 5.1: f-vector shape of (a) the simplex $\triangle_d$, (b) the cross-polytope C_d^*, and (c) the cyclic polytope $C_d(n)$.

5.1.2 Cross-polytope (and cube)

For the d-dimensional cross-polytope, the f-vector is given by

$$f_{i-1}(C_d^*) = \binom{d}{i} 2^i.$$

We obtain a peak around $\frac{2}{3}$ which gets sharper when d grows (see Fig. 5.1(b)).

These two examples already provide lots of different shapes: indeed, it is possible to "add" the two functions $\phi_{\triangle_d}$ and $\phi_{C_d^*}$ by recursively stacking onto facets of the cross-polytope. In particular, if the peaks are sharp enough, we obtain a non-unimodal function. Thus:

Corollary 5.1.1. *The "Unimodality Conjecture" is false, even for simplicial polytopes.*

5.1.3 Cyclic polytope

Let n, d denote two integers with n much larger than d. Then the f-vector of the d-dimensional cyclic polytope with n vertices $C_d(n)$ is approximately

$$f_{i-1}(C_d(n)) = \begin{cases} \binom{n}{i} & \text{if } i \leq \frac{d}{2} \\ \sim \binom{n}{\frac{d}{2}}\binom{\frac{d}{2}}{i-\frac{d}{2}} & \text{otherwise.} \end{cases}$$

This gives a peak around $\frac{3}{4}$ which gets sharper when d grows (see Fig. 5.1(c)). Observe the relation with the "Partial Unimodality Conjecture".

Remark 5.1.2. Using techniques similar to stacking on facets, one can prove that:

1. The "Unimodality Conjecture" is false for $d \geq 8$.
2. For simplicial polytopes, the "Unimodality Conjecture" is true for $d \leq 19$, but false for $d \geq 20$.

5.2 Simplicial polytopes

The following theorem, conjectured by McMullen and proved by Stanley using heavy machinery from algebraic geometry, yields a complete characterization of the f-vectors of simplicial d-polytopes:

Theorem 5.2.1 (The g-theorem: Stanley 1980, Billera–Lee 1980, Björner)**.** *A vector $f = (1, f_0, \ldots, f_{d-1}) \in \mathbb{N}^{d+1}$ is the f-vector of a d-dimensional simplicial polytope if and only if it can be written as $f = gM_d$, where $g = (g_0, \ldots, g_{\lfloor \frac{d}{2} \rfloor})$ is an M-sequence, and*

$$M_d = \left(\binom{d+1-j}{d+1-k} - \binom{j}{d+1-k} \right)_{\substack{0 \leq j \leq \lfloor \frac{d}{2} \rfloor \\ 0 \leq k \leq d}} \in \mathbb{N}^{(\lfloor \frac{d}{2} \rfloor + 1) \times (d+1)}.$$

Examples 5.2.2.

$$M_1 = \begin{pmatrix} 1 & 2 \end{pmatrix} \qquad \text{and } g = (1),$$

$$M_2 = \begin{pmatrix} 1 & 3 & 3 \\ 0 & 1 & 1 \end{pmatrix} \qquad \text{and } g = (1, g_1),$$

$$M_3 = \begin{pmatrix} 1 & 4 & 6 & 4 \\ 0 & 1 & 3 & 2 \end{pmatrix} \qquad \text{and } g = (1, g_1),$$

$$M_4 = \begin{pmatrix} 1 & 5 & 10 & 10 & 5 \\ 0 & 1 & 4 & 6 & 3 \\ 0 & 0 & 1 & 2 & 1 \end{pmatrix} \qquad \text{and } g = (1, g_1, g_2),$$

where $g_1 \geq 0$, $g_2 \geq 0$, and $g_2 \leq \binom{g_1+1}{2}$.

5.3 Dimensions 3 and 4

5.3.1 Dimension 3

Let $(f_0, f_1, f_2) \in \mathbb{N}^3$ be an f-vector of a 3-polytope. Then $f_0 - f_1 + f_2 - 2 = 0$ (Euler relation) and $2f_1 = f_{12} \geq 3f_2$, which implies that $f_2 \leq 2f_0 - 4$.

Similarly, $f_0 \leq 2f_2 - 4$. Thus, apart from the simplex (whose f-vector is $(4, 6, 4)$), we obtain that the f-vector of a 3-polytope satisfies $f_0, f_2 > 4$ and

$$\frac{1}{2} \leq \frac{f_2 - 4}{f_0 - 4} \leq 2.$$

In fact, this inequality characterizes f-vectors of 3-polytopes:

Theorem 5.3.1 (Steinitz 1906)**.** *The f-vectors of 3-polytopes are exactly*

$$\{(4, 6, 4)\} \cup \left\{ (f_0, f_0 + f_2 - 2, f_2) \mid f_0, f_2 > 4, \frac{1}{2} \leq \frac{f_2 - 4}{f_0 - 4} \leq 2 \right\}.$$

We have represented in Fig. 5.2 the zone of possible f-vectors (where f_0 is represented on the horizontal axis, while f_2 is represented on the vertical axis).

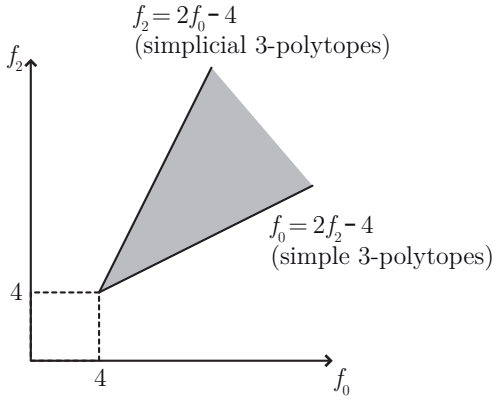


Figure 5.2: The cone of 3-dimensional f-vectors.

5.3.2 Dimension 4

Let $(f_0, f_1, f_2, f_3) \in \mathbb{N}^4$ be an f-vector of a 4-polytope. The following proposition states all currently known inequalities:

Proposition 5.3.2. *The f-numbers of a 4-polytope satisfy*

$$f_0 - f_1 + f_2 - f_3 = 0, \quad f_0 \geq 5, \quad f_3 \geq 5, \quad f_1 \geq 2f_0, \quad f_2 \geq 2f_3,$$

$$\textit{and} \quad 2f_1 + 2f_2 \geq 5f_0 + 5f_3 - 10.$$

Proof. We only prove the last inequality, using arguments similar to those in Exercise 2.4.4. Let P be a 4-polytope. For any facet F of P,

$$3f_0(F) \leq f_{01}(F) = 2f_1(F) \quad \text{and} \quad f_2(F) - f_1(F) + f_0(F) - 2 = 0.$$

Summing over all facets of P, we obtain

$$3f_{03}(P) \leq 2f_{13}(P) \quad \text{and} \quad f_{23}(P) - f_{13}(P) + f_{03}(P) - 2f_3(P) = 0.$$

Hence, $f_{13}(P) \leq 6f_2(P) - 6f_3(P)$, and, dually, $f_{02}(P) \leq 6f_1(P) - 6f_0(P)$. Now we use the generalized Lower Bound Theorem (Theorem 1.3.2):

$$4f_0(P) - 10 \leq f_1(P) + f_{02}(P) - 3f_2(P) \leq 7f_1(P) - 3f_2(P) - 6f_0(P),$$

$$\text{and, dually,} \quad 4f_3(P) - 10 \leq 7f_2(P) - 3f_1(P) - 6f_3(P).$$

Summing these two last inequalities, we obtain the desired bound. □

Corollary 5.3.3. *For any* 4*-polytope,*

$$\psi_1 \geq 1, \quad \psi_2 \geq 1, \quad \psi_1 - 1 \leq \psi_2 \leq \psi_1 + 1, \quad \textit{and} \quad \psi_1 + \psi_2 \geq \frac{5}{2},$$

$$\textit{where} \quad \psi_1 = \frac{f_1 - 10}{f_0 + f_3 - 10} \quad \textit{and} \quad \psi_2 = \frac{f_2 - 10}{f_0 + f_3 - 10}.$$

We have represented in Fig. 5.3 the polyhedron given by these inequalities (where ψ_1 is represented on the horizontal axis, while ψ_2 is represented on the vertical axis). It is not known exactly what integer points in this polyhedron correspond to 4-polytopes.

We finish by localizing certain 4-polytopes (and their duals!) in this polyhedron (see Fig. 5.3):

Examples 5.3.4. 1. Stacked polytopes:

$$f(\mathrm{St}_4(n)) = (4+n,\, 6+4n,\, 4+6n,\, 2+3n) \sim n(1,4,6,3).$$

Thus, in the (ψ_1, ψ_2)-coordinate system, we obtain the point $\begin{pmatrix} 1 \\ \frac{3}{2} \end{pmatrix}$.

2. Cyclic polytopes:

$$f(C_4(n)) = \left(n,\, \frac{n(n+1)}{2},\, n(n-3),\, \frac{n(n-3)}{2}\right) \sim \frac{n^2}{2}(0,1,2,1).$$

We obtain the point $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

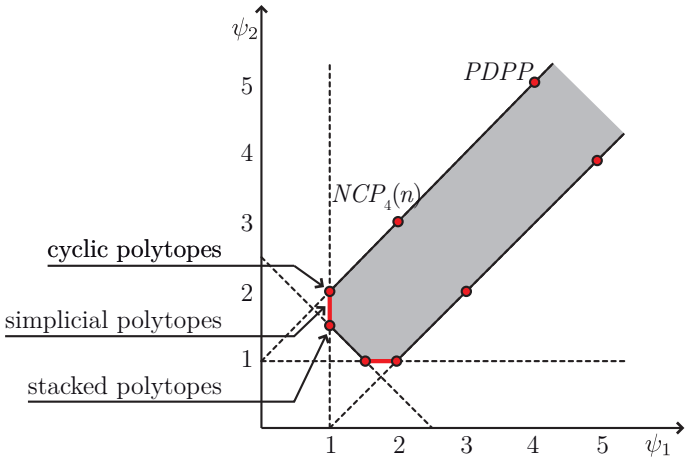


Figure 5.3: The possible zone of 4-dimensional f-vectors.

3. Neighborly cubical polytopes (see Exercise 2.4.2):

   $f(NCP_4(n)) = (2^n,\, n2^{n-1},\, 3(n-2)2^{n-2},\, (n-2)2^{n-2}) \sim n2^{n-2}(0,2,3,1).$

   We obtain the point $\begin{pmatrix}2\\3\end{pmatrix}$.

4. Projected deformed products of polygons give the point $\begin{pmatrix}4\\5\end{pmatrix}$.

5.4 Hansen polytopes

Let $G = (V, E)$ be a graph on $|V| = n$ vertices. Recall that

- an *independent set* of G is a subset I of V such that the induced subgraph has no edges;
- a *clique* of G is a subset C of V such that the induced subgraph is complete.

For any subset U of V, let χ^U denote the *characteristic function* of V, that is,

$$\chi^U : V \longrightarrow \{0,1\}, \quad v \longmapsto \begin{cases} 1 & \text{if } v \in U, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 5.4.1. The *stable set polytope* of G is the polytope defined by

$$\mathrm{Stab}(G) = \mathrm{conv}\{\chi^I \mid I \text{ independent set of } G\} \subset \mathbb{R}^n.$$

The *Hansen polytope* of G is the *twisted prism* of the stable set polytope of G:

$$\mathrm{Hans}(G) = \mathrm{conv}\{\{1\} \times \mathrm{Stab}(G),\ \{-1\} \times -\mathrm{Stab}(G)\} \subset \mathbb{R}^{n+1}.$$

Examples 5.4.2. The Hansen polytope of the complete graph K_2 on 2 vertices is an octahedron. The Hansen polytope of the independent graph $\bar{K}_2$ on 2 vertices is an affine cube.

For any stable set I of G and any clique C of G, it is clear that $|I \cap C| \leq 1$, i.e., that $0 \leq \langle \chi^I \mid \chi^C \rangle \leq 1$. This implies that

$$\mathrm{Stab}(G) \subseteq \big\{x \in \mathbb{R}^n \mid 0 \leq \langle x \mid \chi^C \rangle \leq 1 \text{ for all cliques } C \text{ of } G\big\},$$

and

$$\mathrm{Hans}(G) \subseteq \big\{(x_0, \bar{x}) \in \mathbb{R}^{n+1} \mid -1 \leq -x_0 + 2\langle \bar{x} \mid \chi^C \rangle \leq 1 \text{ for all cliques } C \text{ of } G\big\}.$$

Furthermore, Hansen proved that these inclusions are equalities if and only if the graph is perfect[1]. This yields a complete combinatorial description of Hansen polytopes of perfect graphs:

Remark 5.4.3. Let G be a perfect graph on n vertices. Then the Hansen polytope $\mathrm{Hans}(G)$ is a centrally-symmetric $(n+1)$-polytope with

(i) two vertices $\pm(1, \chi^I)$ for each independent set I of G, and

(ii) two facets $-1 \leq -x_0 + 2\langle \bar{x} \mid \chi^C \rangle$ and $-x_0 + 2\langle \bar{x} \mid \chi^C \rangle \leq 1$ for each clique C of G.

Furthermore, the vertex $(1, \chi^I)$ lies in the facet $-x_0 + 2\langle \bar{x} \mid \chi^C \rangle \leq 1 \iff \langle \chi^I \mid \chi^C \rangle = 1 \iff |I \cap C| = 1 \iff I \cap C \neq \emptyset$.

Observe also that the Hansen polytope $\mathrm{Hans}(G)$ is self-dual if and only if G is self-complementary.

Let us concentrate on the particular example of the Hansen polytope of the 4-path. Since this graph is perfect and self-complementary, its Hansen polytope is a centrally-symmetric and self-dual 5-polytope.

The following table compares its f-vector with the f-vectors of the Hanner polytopes (see the footnote on page 27). We do not explicitly write the

[1] A *perfect graph* is a graph such that each of its induced subgraphs has equal chromatic number and clique number. The Strong Perfect Graph Theorem affirms that a graph G is perfect if and only if neither G nor its complement contains an odd cycle of length at least 5.

f-vectors of the duals, but they are obtained by reversal:

P	$f_0(P)$	$f_1(P)$	$f_2(P)$	$f_3(P)$	$f_4(P)$	$f_0(P)+f_4(P)$
$\mathrm{Hans}(P_4)$	16	64	98	64	16	32
C_5^Δ	10	40	80	80	32	42
bip bip C_3	12	48	86	72	24	36
bip prism C_3^Δ	14	54	88	66	20	34
prism C_4^Δ	16	56	88	64	18	34

This disproves one of Kalai's conjectures already mentioned in Section 3:

Conjecture 5.4.4 (Kalai)**.** *Every centrally symmetric polytope has an f-vector that is componentwise larger than the f-vector of a Hanner polytope of the same dimension.*

Observe also that the total number of faces of the Hansen polytope of the 4-path,

$$1+\sum_{i=0}^{4} f_i(\mathrm{Hans}(P_4)) = 259,$$

is surprisingly low compared to the $3^5 = 243$ of the Hanner polytopes (and of the 3^d conjecture).

Similarly, the *bull graph* $B = (\{1,2,3,4,5\}, \{12,23,34,45,24\})$ gives an interesting example in dimension 6 (which in particular disproves Kalai's conjecture in dimension 6).

5.5 Selected exercises

Exercise 5.5.1. For $d = 3,4,5,\ldots$, construct a d-polytope with 12 vertices and 13 facets. How far do you get?

The following construction gives such a d-polytope (for any $3 \le d \le 10$): we start from a $(13-d)$-gon in dimension 2 ($f_0 = 13-d$ and $f_1 = 13-d$) and take a pyramid over it ($f_0 = 14-d$ and $f_2 = 14-d$). Then, we stack one of the triangular faces ($f_0 = 15-d$ and $f_2 = 16-d$). Finally, we take recursively $d-3$ pyramids over it and we obtain 12 vertices and 13 facets.

Exercise 5.5.2. Show that the f-vectors of d-polytopes are unimodal for $d \le 5$.

It is easy for $d \le 4$.

In dimension 5, we know that $2f_1 = f_{01} \ge 5f_0$, $2f_3 = f_{34} \ge 5f_4$ and $f_0 - f_1 + f_2 - f_3 + f_4 - 2 = 0$. If f is not unimodal, then $f_1 > f_2$ and $f_2 < f_3$.

This would imply that

$$\begin{aligned} 10 &= 5f_0 - 5f_1 + 5f_2 - 5f_3 + 5f_4 \\ &\leq 2f_1 - 5f_1 + 5f_2 - 5f_3 + 2f_3 \\ &= -3f_1 + 5f_2 - 3f_3 \leq -f_2, \end{aligned}$$

which is impossible.

Exercise 5.5.3. Derive an exact formula for the number of facets of the d-dimensional cyclic polytope.

Let us recall first the following description of facets of the cyclic polytope:

Proposition 5.5.4 (Gale's Evenness Criterion)**.** *A subset F of $[n]$ is a facet of the cyclic polytope $C_d(n)$ if and only if $|F| = d$ and all inner blocks*[2] *of F have even size. Furthermore, such a facet F is supported by the hyperplane*

$$H_F = \{z \mid \langle (\gamma_i(F))_{i\in[d]} \mid z\rangle = -\gamma_0(F)\},$$

where $\gamma_0(F), \ldots, \gamma_d(F)$ are defined as the coefficients

$$\Pi_F(t) = \prod_{i\in F}(t - t_i) = \sum_{i=0}^{d} \gamma_i(F)t^i.$$

Proof. Observe first the following:

(i) The Vandermonde determinant

$$\det \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \mu(x_0) & \mu(x_1) & \ldots & \mu(x_d) \end{pmatrix} = \prod_{0\leq i<j\leq d} (x_j - x_i)$$

ensures that any $d+1$ points on the d-dimensional moment curve are affinely independent, and, thus, that the cyclic polytope is simplicial.

(ii) For any $t \in \mathbb{R}$,

$$\langle (\gamma_i(F))_{i\in[d]} \mid \mu_d(t)\rangle + \gamma_0(F) = \sum_{i=0}^{d} \gamma_i(F)t^i = \Pi_F(t).$$

Let F be a subset of $[n]$ of size d. Then,

[2]The *blocks* of a subset F of $[n]$ are the maximal subsets of consecutive elements of F. The *initial* (resp. *final*) block is the block containing 1 (resp. n) —when it exists. Other blocks are called *inner* blocks. For example, $\{1, 2, 3, 6, 7, 9, 10, 11\} \subset [11]$ has 3 blocks: $\{1, 2, 3\}$ (initial), $\{6, 7\}$ (inner) and $\{9, 10, 11\}$ (final).

(i) for all $j \in F$, $\Pi_F(t_j) = 0$; thus, H_F is the affine hyperplane spanned by F;

(ii) for all $j \notin F$, the sign of $\Pi_F(t_j)$ is $(-1)^{|F \cap \{j+1,\dots,n\}|}$.

In particular, if F has an odd inner block $\{a, a+1, \dots, b\}$, then $\Pi_F(t_{a-1})$ and $\Pi_F(t_{b+1})$ have different signs, and F is not a facet. Reciprocally, if all inner blocks have even size, then the sign of all $\Pi_F(t_j)$ is $(-1)^\ell$, where ℓ is the size of the final block. □

From this criterion, it is easy to count the facets of the cyclic d-polytope, separating the cases when d is odd or even:

$$f_{2e-1}(C_{2e}(n)) = \binom{n-e}{e} + \binom{n-e-1}{e-1},$$

and

$$f_{2e}(C_{2e+1}(n)) = 2\binom{n-e-1}{e}.$$

Exercise 5.5.5. Compute the f-vector of the cyclic polytope $C_8(25)$. How bad is the approximation

$$f_{k-1}(C_d(n)) \sim h(d,k,n) = \binom{n}{\frac{d}{2}}\binom{\frac{d}{2}}{k-\frac{d}{2}} \quad \text{for even } d \text{ and } k > \tfrac{d}{2}?$$

The values of $f_k(C_8(25))$ and $h(8,k,25)$ are given in the following table:

k	0	1	2	3	4	5	6	7
$f_k(C_8(25))$	25	300	2 300	12 650	33 750	44 500	28 500	7 125
$h(8,k,25)$				12 650	50 600	75 900	50 600	12 650

Exercise 5.5.6. What is the f-vector of the product of ten 10-gons? Where is the peak?

The non-empty faces of a product are exactly the products of non-empty faces of the factors. Thus, a k-face of $(C_{10})^{10}$ is obtained for any $0 \le i \le \lfloor \frac{k}{2} \rfloor$ by choosing:

(a) i of the complete 10-gons;

(b) one of the 10 edges in $k-2i$ of the ten 10-gons; and

(c) one of the 10 vertices in $10-k$ of the ten 10-gons.

Consequently, we obtain the f-vector

$$f_k((C_{10})^{10}) = \sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} \binom{10}{i} \binom{10-i}{k-2i} 10^{10-i}.$$

The peak is reached for $k = 6$:

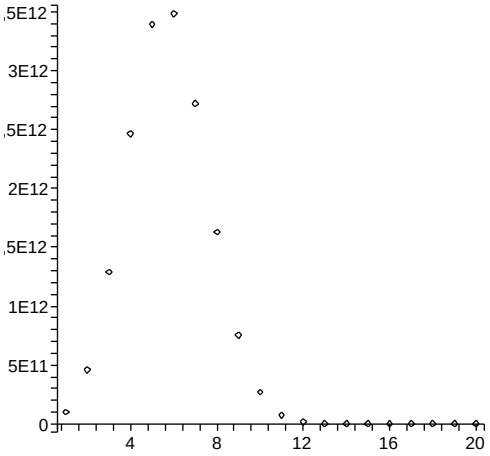
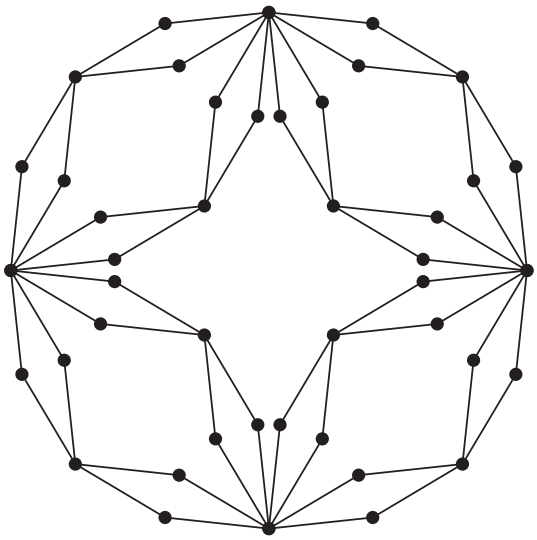


Figure 5.4: The f-vector shape of a product of ten 10-gons.

Metric Embeddings

by Jiří Matoušek



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

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

(M2) $\rho(x, x) = 0$,

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

(M4) $\rho(y, x) = \rho(x, y)$, and

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

If ρ 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).

1.1.1 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.

1.1.2 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 you in your finger.

My first encounter with metric embeddings occurred through bacteria in the late 1980s. There are enormous numbers 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 have picked creatures perhaps more familiar than bacterial species; the price to pay is that the numbers are completely artificial:

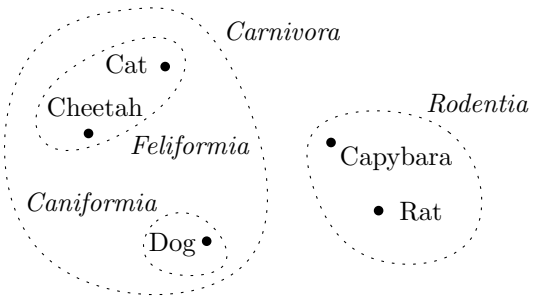
	Dog	Cat	Cheetah	Rat	Capybara
Dog	0				
Cat	0.50	0			
Cheetah	0.42	0.27	0		
Rat	0.69	0.69	0.65	0	
Capybara	0.72	0.61	0.59	0.29	0

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

It is hard to see any structure in this kind of 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.)

1.1.3 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

1.2.1 Impossibility of isometric embeddings

An exact representation of one metric space in another is formalized by the notion of isometric embedding. A mapping $f\colon (X,\rho) \to (Y,\sigma)$ of one metric space into another is called an *isometric embedding* or *isometry* if $\sigma(f(x), f(y)) = \rho(x,y)$ 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 is worth trying it *now*.

1.2.2 Approximate embeddings

For visualizing a metric space, we need not insist on representing distances exactly —often we do not 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, ρ) and (Y, σ) be metric spaces. An injective mapping $f\colon (X, \rho) \to (Y, \sigma)$ is called a *D-embedding*, where $D \geq 1$ is a real number, if there is a number $r > 0$ such that, for all $x, y \in X$,
>
> $$r \cdot \rho(x, y) \leq \sigma(f(x), f(y)) \leq Dr \cdot \rho(x, y).$$
>
> 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.

1.2.3 Lipschitz and bi-Lipschitz maps

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

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

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*.

1.2.4 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 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

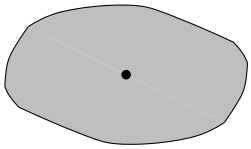
A metric can be defined on a completely arbitrary set, and it specifies distances for pairs of points. A norm is defined only on a vector space, and for each point it specifies its distance from the origin.

[2] A reminder of asymptotic notation: $f(n) = O(g(n))$ means that there are n_0 and C such that $f(n) \le Cg(n)$ for all $n \ge n_0$; $f(n) = o(g(n))$ 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 $f(n) = \Omega(g(n))$.

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.

1.3.1 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.

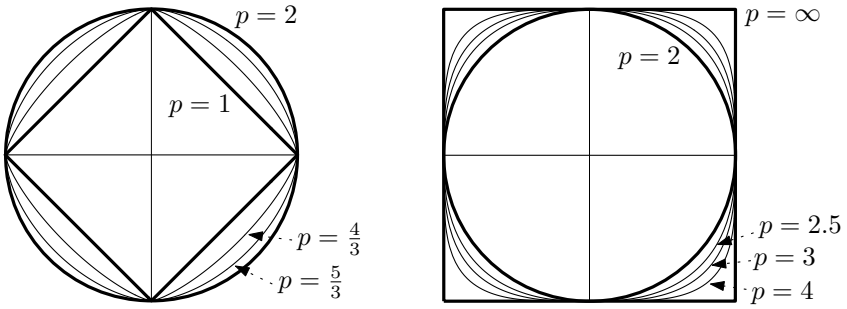
For a point $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ and for $p \in [1, \infty)$, the ℓ_p *norm* is defined as

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

We denote by ℓ_p^d the normed space $(\mathbb{R}^d, \|.\|_p)$.

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 $p \to \infty$.

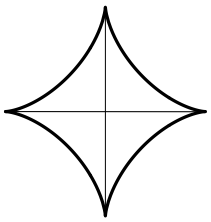
To gain some feeling about ℓ_p norms, let us look at their unit balls in the plane:



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 $p \geq 2$, one can see the unit ball expanding towards the square as $p \to \infty$. Sharp corners appear again for the ℓ_∞ norm.

1.3.2 The case $p < 1$

For $p \in (0, 1)$, the formula $\|\mathbf{x}\|_p = \big(|x_1|^p + \cdots + |x_d|^p\big)^{1/p}$ 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, $d_p(\mathbf{x}, \mathbf{y}) = |x_1 - y_1|^p + \cdots + |x_d - y_d|^p$ does define a metric on $\mathbb{R}^d$, 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 ρ on a finite set X is called an ℓ_p *metric* if there exists a natural number d and an isometric embedding of (X, ρ) into the space ℓ_p^d. For $p = 2$ we also speak of a *Euclidean metric*.

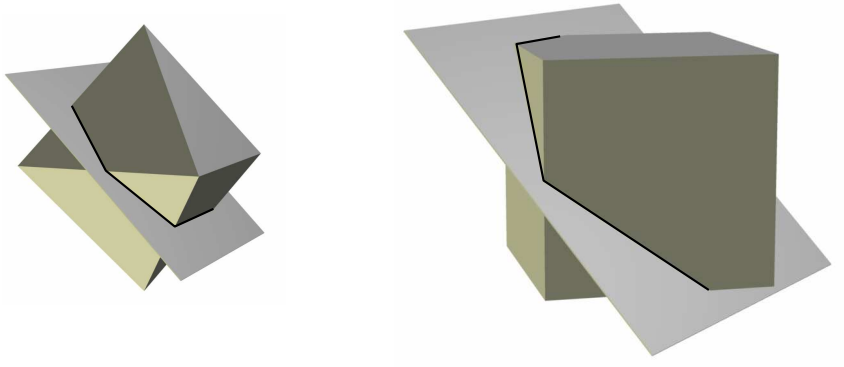
An ℓ_p *pseudometric* is defined similarly, but we consider isometric maps into ℓ_p^d that are not necessarily injective.

1.4.1 Dimension of isometric embeddings

This definition prompts a question: How high do we need to go with the dimension d in order to represent all possible ℓ_p metrics on n points?

For $p = 2$, the answer is easy: $d = n-1$ always suffices and it is sometimes necessary. Indeed, given any n points in $\mathbb{R}^d$, 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 of 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, and quite important, method is needed to show the following weaker bound on the dimension.

Proposition 1.4.1. *Every n-point space with an ℓ_p metric is isometrically embeddable in ℓ_p^N, where $N = \binom{n}{2}$.*

Proof. We first do the proof assuming $p = 1$, since this makes some things simpler, and then we point out modifications needed for the general case.

For notational convenience, let us assume that the point set of our space is $X = \{1, 2, \ldots, n\}$. The key trick is to think of a metric ρ on X as a point in a suitable space. Namely, ρ is specified by the N-component real vector $\boldsymbol{\rho} = (\rho(i,j) : 1 \le i < j \le n) \in \mathbb{R}^N$. We can now define, for example, the *metric cone* $\mathcal{M}_n \subset \mathbb{R}^N$ as

$$\mathcal{M}_n = \{\boldsymbol{\rho} \in \mathbb{R}^N : \rho \text{ is a pseudometric on } X\},$$

an interesting mathematical object.

For the purposes of this proof, we do not need the metric cone but rather the set

$$\mathcal{L}_1 = \{\boldsymbol{\rho} \in \mathbb{R}^N : \rho \text{ is an } \ell_1 \text{ pseudometric on } X\}.$$

We want to check that $\mathcal{L}_1$ is convex; more precisely, a convex cone, i.e., closed under linear combinations with nonnegative coefficients. Clearly, if $\mathbf{x} \in \mathcal{L}_1$, then $\lambda\mathbf{x} \in \mathcal{L}_1$ for all $\lambda \ge 0$, and so it suffices to verify that if $\mathbf{x}, \mathbf{y} \in \mathcal{L}_1$, then $\mathbf{x} + \mathbf{y} \in \mathcal{L}_1$. By definition, $\mathbf{x} \in \mathcal{L}_1$ means that there is a mapping $f\colon X \to \mathbb{R}^k$ such that $x_{i,j} = \|f(i) - f(j)\|_1$ (recall that the vectors in $\mathbb{R}^N$ are indexed by pairs (i,j), $i < j$). Similarly, for $\mathbf{y}$ we have a mapping $g\colon X \to \mathbb{R}^\ell$ with $y_{i,j} = \|g(i) - g(j)\|_1$. We define a new mapping $h\colon X \to \mathbb{R}^{k+\ell}$ by concatenating the coordinates of f and g; that is,

$$h(i) = (f(i)_1, \ldots, f(i)_k, g(i)_1, \ldots, g(i)_\ell) \in \mathbb{R}^{k+\ell}.$$

The point of $\mathcal{L}_1$ corresponding to h is $\mathbf{x} + \mathbf{y}$. Thus, $\mathcal{L}_1$ is a convex cone.

Next, we define a *line pseudometric* on X as a pseudometric that can be isometrically mapped into the real line $(\mathbb{R}^1, |.|)$. Line pseudometrics belong to $\mathcal{L}_1$. If we have an isometric embedding of a metric ρ in ℓ_1^k, we can express ρ as the sum of k line pseudometrics (each corresponding to one of the coordinates of the mapping). Thus, each ℓ_1 metric is a nonnegative linear combination of line pseudometrics.

Next, we want to use Carathéodory's theorem. The basic version of this theorem tells us that if $\mathbf{x}$ is in the convex hull of a set $S \subseteq \mathbb{R}^d$, then $\mathbf{x}$ is contained in the convex hull of some at most $d + 1$ points of S.

We need a version for convex cones: If a point $\mathbf{x} \in \mathbb{R}^d$ is a nonnegative linear combination of points in a set $S \subseteq \mathbb{R}^d$, then it can be expressed as a nonnegative linear combination of at most d of these points. In our situation, this shows that every ℓ_1 metric on X is a nonnegative linear combination of at most N line pseudometrics, and thus it embeds isometrically in ℓ_1^N. This concludes the proof for the case $p = 1$.

For $p > 1$, the proof is very similar but everything needs to be "raised to power p". Namely, we let $\mathcal{L}_p$ consist of pth powers of ℓ_p pseudometrics on X:

$$\mathcal{L}_p = \Big\{(\rho(i,j)^p)_{1 \le i < j \le n} : \rho \text{ is an } \ell_p \text{ pseudometric on } X\Big\}.$$

Then one verifies that $\mathcal{L}_p$ is a convex cone generated by pth powers of line pseudometrics, and uses Carathéodory's theorem —we omit the details. □

Corollary 1.4.2. *Let (X, ρ) be a finite metric space and suppose that for every $\varepsilon > 0$ there is some k such that (X, ρ) admits a $(1+\varepsilon)$-embedding in ℓ_p^k. Then (X, ρ) is an ℓ_p metric.*

Proof. Let $\Delta = \operatorname{diam}(X)$ be the largest distance in (X, ρ). For every $\varepsilon > 0$ there is a $(1+\varepsilon)$-embedding $f_\varepsilon \colon (X, \rho) \to \ell_p^N$, $N = \binom{|X|}{2}$, by Proposition 1.4.1.

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 $f \colon X \to \ell_p^N$ such that for every $\eta > 0$ there is some f_ε with $\|f(x) - f_\varepsilon(x)\|_p \le \eta$. Then f is the desired isometry. □

1.4.2 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 all infinite sequences $\mathbf{x} = (x_1, x_2, \ldots)$ of real numbers with $\|\mathbf{x}\|_p < \infty$, where $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p\right)^{1/p}$.

- The "big" $L_p = L_p(0,1)$, consisting of all measurable functions $f \colon [0,1] \to \mathbb{R}$ such that $\|f\|_p = \left(\int_0^1 |f(x)|^p \,\mathrm{d}x\right)^{1/p}$ 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 $p \neq 2$, 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 1.4.2. 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 $\|\mathbf{x}\|_p = \big(|x_1|^p + \cdots + |x_d|^p\big)^{1/p}$ it is probably not clear that 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 do not 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.

1.5.1 Dvoretzky's theorem and almost spherical slices

Part (ii) looks like something that should have a very direct and simple proof, but it does not.

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 1.4.2 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*[3] *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.

1.5.2 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^d isometrically into the infinite-dimensional space $L_1(S^{d-1})$. What is that? Similarly to $L_1 = L_1(0,1)$, the elements of $L_1(S^{d-1})$ are (equivalence classes of) measurable real functions, but the domain is the $(d-1)$-dimensional unit Euclidean sphere S^{d-1}. The distance of two functions f, g is $\|f-g\|_1 = \int_{S^{d-1}} |f(\mathbf{u}) - g(\mathbf{u})| \, \mathrm{d}\mathbf{u}$, where we integrate according to the uniform (rotation-invariant) measure on S^{d-1}, scaled so that the whole of

[3] The symmetry assumption can be dropped.

S^{d-1} has measure 1.

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

Let us fix some $\mathbf{v}_0 \in \ell_2^d$ with $\|\mathbf{v}_0\|_2 = 1$, and set

$$C = \|F(\mathbf{v}_0)\|_1 = \int_{S^{d-1}} |\langle \mathbf{v}_0, \mathbf{u}\rangle| \, \mathrm{d}\mathbf{u}.$$

By rotational symmetry, and this is the beauty of this proof, we have $\|F(\mathbf{v})\|_1 = C$ for every unit $\mathbf{v} \in \ell_2^d$, and hence in general $\|F(\mathbf{x})\|_1 = C\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \ell_2^d$. 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 $L_1(S^{d-1})$ are ℓ_1 metrics? With some handwaving we can argue like this: If we choose a "sufficiently uniformly distributed" finite set $A \subseteq S^{d-1}$, then the integral of every "reasonable" function f on S^{d-1}, such as our functions $f_{\mathbf{x}}$, over S^{d-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^d 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 1.4.1 and compactness allow us to conclude that every ℓ_2 metric is also an ℓ_1 metric.

1.5.3 The Fréchet embedding

We will prove that *every n-point metric space (X, ρ) 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 $f\colon X \to \ell_\infty^d$ means to define n functions $f_1, \ldots, f_n\colon X \to \mathbb{R}$, the coordinates of the embedded points. Here we set

$$f_i(x_j) = \rho(x_i, x_j).$$

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

1.5.4 Which p?

That is, if we have a collection of objects with a large number d of attributes (say 20 or more), such as a collection of bacterial strains in the motivating

example, how should we measure their distance? We assume that the considered problem does not suggest itself a particular distance function and that we can reasonably think of the attributes as coordinates of points in $\mathbb{R}^d$.

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 d is not very small compared to n, the number of points, and let us consider a random n-point set $X \subset \mathbb{R}^d$, 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.

Lecture 2

Dimension Reduction: Around the Johnson–Lindenstrauss Lemma

2.1 The lemma

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

Theorem 2.1.1. *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 $(1+\varepsilon)/(1-\varepsilon)$. 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 have not used the language of distortion, but rather a slightly different notion, which we turn into a general definition: Let us call a mapping $f\colon (X,\rho) \to (Y,\sigma)$ of metric spaces an *ε-almost isometry* if $(1-\varepsilon)\rho(x,y) \leq \sigma(f(x),f(y)) \leq (1+\varepsilon)\rho(x,y)$. For ε small, this is not very different from saying that f is a $(1+2\varepsilon)$-embedding (at

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

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 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:

Lemma 2.1.2 (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*

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

where $c > 0$ is a constant (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 *does not* say: It definitely does not claim that a random T is an ε-almost isometry —since, obviously, for $k < n$, a linear map $\mathbb{R}^n \to \mathbb{R}^k$ cannot 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

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

for every fixed $\mathbf{x}$. We apply this to the $\binom{n}{2}$ vectors $\mathbf{p}_i - \mathbf{p}_j$, $1 \le i < j \le n$, 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 $k \leq n$, while later we will 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.

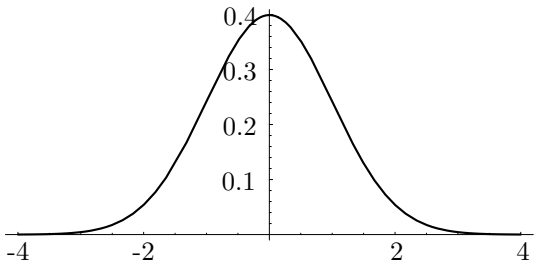
The standard normal (or Gaussian) distribution $N(0, 1)$ is well known, yet I first want to remind a beautiful computation related to it. The density of $N(0, 1)$ is proportional to $e^{-x^2/2}$, but what is the right normalizing constant? In other words, what is the value of the integral $I = \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x$? 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

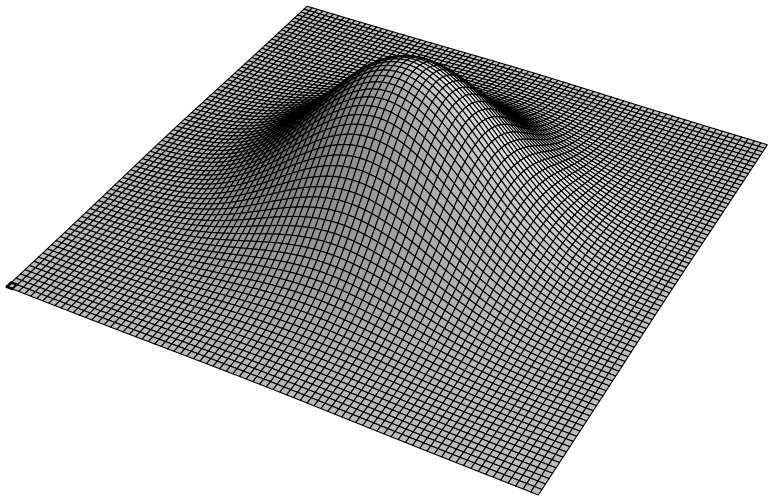
$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right) \\ &= \int_{\mathbb{R}^2} e^{-x^2/2} e^{-y^2/2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y \;=\; \int_0^{\infty}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

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

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]



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 $\mathrm{Var}\,[Z] = \mathbf{E}[(Z - \mathbf{E}[Z])^2] = \mathbf{E}[Z^2] = 1$ (this is the 1). The random variable aZ, $a \in \mathbb{R}$, has the normal distribution $N(0, a^2)$ with variance a^2.

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

2.2.1 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.

There are several ways of proving the 2-stability. The "right" one is perhaps through characteristic functions. One can also say that 2-stability has to be true because of the Central Limit Theorem. It is also possible to do the brute-force computation of the appropriate convolution. We offer a geometric way.

Let us first think of choosing the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathbb{R}^n$, and let $U = n^{-1/2}(Z_1 + Z_2 + \cdots + Z_n)$ be the orthogonal projection of $\mathbf{Z}$ on the diagonal line $x_1 = x_2 = \cdots = x_n$. By spherical symmetry, U has the same distribution as the orthogonal projection of $\mathbf{Z}$ on the x_1-axis, i.e., Z_1, and thus $Z_1 + \cdots + Z_n \sim \sqrt{n} Z_1$.

Now we want to prove that, with X, Y standard normal, $aX + bY$ is normally distributed. We may assume that a and b are positive integers (the general case follows by a continuity handwaving), and let $k = a^2$, $\ell = b^2$, and $n = k + \ell$. By the above $aX \sim Z_1 + \cdots + Z_k$ and $bY \sim Z_{k+1} + \cdots + Z_n$. Thus $aX + bY \sim Z_1 + \cdots + Z_n \sim \sqrt{n} Z_1$.

2.2.2 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 that X has a *subgaussian upper tail* if there exists a constant $a > 0$ such that, for all $\lambda > 0$,

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

We say that X has a *subgaussian upper tail up to* λ_0 if the previous bound holds for all $\lambda \leq \lambda_0$. 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 $Y = n^{-1/2}(X_1 + X_2 + \cdots + X_n)$ has a subgaussian tail (the normalization by $n^{-1/2}$ is chosen so that $\mathrm{Var}[Y] = 1$).

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

Lemma 2.2.1 (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. 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

Lemma 2.3.1 (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 $\mathbf{E}[Y_i^2] = \mathrm{Var}[Y_i] = 1$, and thus $\mathbf{E}[\|T(\mathbf{x})\|_2^2] = 1$. 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 is 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 does not 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.

Claim 2.3.2. *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

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

(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 (and 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: $\mathbf{E}\left[e^{u(1-Y^2)}\right] = e^u(1+2u)^{-1/2} = e^{u^2+O(u^3)}$. □

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

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

where $0 \le u \le u_0\sqrt{k}$, and similarly for $\mathbf{E}\left[e^{-uW}\right]$. 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,

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

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

$$\begin{aligned}\text{Prob}[1-\varepsilon \le \|T(\mathbf{x})\|_2 \le 1+\varepsilon] = \text{Prob}\left[(1-\varepsilon)^2 \le \|T(\mathbf{x})\|_2^2 \le (1+\varepsilon)^2\right] \ge \\ \text{Prob}\left[1-\varepsilon \le \|T(\mathbf{x})\|_2^2 \le 1+\varepsilon\right] = \text{Prob}\left[|W| \le \varepsilon\sqrt{k}\right] \ge 1 - 2e^{-c\varepsilon^2 k}.\end{aligned}$$

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.

Lemma 2.4.1 (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 $\mathbf{E}[R_{ij}] = 0$, $\text{Var}[R_{ij}] = 1$, and a uniform subgaussian tail. Then for every vector $\mathbf{x} \in \mathbb{R}^n$ we have

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

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 do not 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 does not 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:

Lemma 2.4.2. *If X is a random variable with $\mathbf{E}[X] = 0$ and $\operatorname{Var}[X] = \mathbf{E}[X^2] = 1$, 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, $F(t) = \operatorname{Prob}[X < t]$. 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:

Lemma 2.4.3. *Let $R_1, \ldots, R_n$ be independent random variables such that $\mathbf{E}[R_j] = 0$, $\mathrm{Var}[R_j] = 1$, and with 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.

Claim 2.4.4. *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*

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

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, $F(t) = \mathrm{Prob}[Y^2 < t]$. 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. Claim 2.4.4 is 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.

Theorem 2.5.1. *Given n and $\varepsilon \in (0,1)$, let $k \geq C\varepsilon^{-2}(\log \frac{1}{\varepsilon})n$ 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.

Lemma 2.5.2 (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*

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

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

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

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 do not 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.

Claim 2.5.3. *For Y standard normal we have*

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

with a suitable C and all $u \ge 0$ (note that we do not 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:

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

Thus $\mathbf{E}\left[e^{u(|Y|-\beta)}\right] \le e^{u^2/2}$. 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. □

2.5.1 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 the 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 *does not* 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* does not. 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 $\mathbf{E}[Y_i^2]$, 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 $C_p \geq c_p > 0$ (the best values are known) with

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

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.

2.5.2 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:

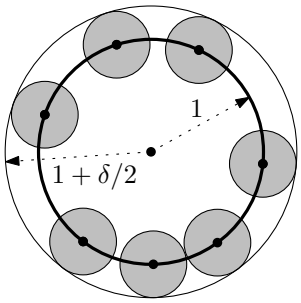
Lemma 2.5.4 (Small $\boldsymbol{\delta}$-dense sets in the sphere)**.** *For each $\delta \in (0, 1]$, there exists a δ-dense set $N \subseteq S^{n-1}$ with*

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

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, $\operatorname{vol}(B(0,2)) \geq |N| \operatorname{vol}(B(0,\frac{\delta}{2}))$, and since $\operatorname{vol}(B(0,r))$ in $\mathbb{R}^n$ is proportional to r^n, 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.

Lemma 2.5.5. *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 $1-\varepsilon \leq \|F(\mathbf{y})\|_1 \leq 1+\varepsilon$ for all $\mathbf{q} \in N$. Then T is a $2(\varepsilon+\delta)$-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 $\|\mathbf{x}-\mathbf{y}\| \leq \delta$, and the triangle inequality gives

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

But now we need to bound $\|T(\mathbf{x}-\mathbf{y})\|_1$ having a bound on $\|\mathbf{x}-\mathbf{y}\|_2$, and this is the same problem as bounding $\|T(\mathbf{x})\|_1$, only with a different vector.

The trick is to continue recursively, in a manner formally resembling the expansion of a real number into the continued fraction. We set $q_1 = \|\mathbf{x}-\mathbf{y}\|_2$, we define $\mathbf{x}_1 = \frac{1}{q_1}(\mathbf{x}-\mathbf{y}) \in S^{n-1}$, and we find an $\mathbf{y}_1 \in N$ with $\|\mathbf{x}_1 - \mathbf{y}_1\|_2 \le \delta$. Then

$$\begin{aligned}\|T(\mathbf{x})\|_1 &\le 1+\varepsilon+q_1\|T(\mathbf{x}_1)\|_1 \le 1+\varepsilon+\delta\|T(\mathbf{x}_1)\|_1\\ &\le 1+\varepsilon+\delta(\|T(\mathbf{y}_1)\|_1+\|T(\mathbf{x}_1-\mathbf{y}_1)\|_1)\\ &\le 1+\varepsilon\delta(1+\varepsilon)+\delta q_2\|T(\mathbf{x}_2)\|_1,\end{aligned}$$

with $q_2 = \|\mathbf{x}_1-\mathbf{y}_1\|_2$ and $\mathbf{x}_2 = \frac{1}{q_2}(\mathbf{x}_1-\mathbf{y}_1)$. Continuing in this manner we arrive at

$$\|T(\mathbf{x})\|_1 \le (1+\varepsilon)(1+\delta+\delta^2+\cdots) = \frac{1+\varepsilon}{1-\delta}.$$

(For perfectionists we remark that if it so happens and $\mathbf{x}_i = \mathbf{y}_i$ for some i, we can just stop the expansion at that i.)

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 $e^{-\Omega(\varepsilon^2 k)}$).

2.5.3 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 $c > 0$, similarly 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 cannot 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 has lost them but under a street lamp where there is 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, i.e., given i, testing whether the component x_i is exceptionally large compared to most others.

Thus, we consider the following problem, the ℓ_2 *norm estimation*: We are 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

[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 these 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 is at least $\sqrt{n}/10$.

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 *do not* 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.

2.6.1 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 cannot afford to store it.

2.6.2 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 it 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 cannot provide a theoretical guarantee for all possible vectors $\mathbf{x}$.

2.6.3 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 $\{0,1\}^{2\ell^2+\ell} \to \{0,1\}^{\ell 2^\ell}$.

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

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

$$(\sigma_0, a_1, b_1, \ldots, a_n, b_n),$$

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 $\mathrm{GF}(2^\ell)$.

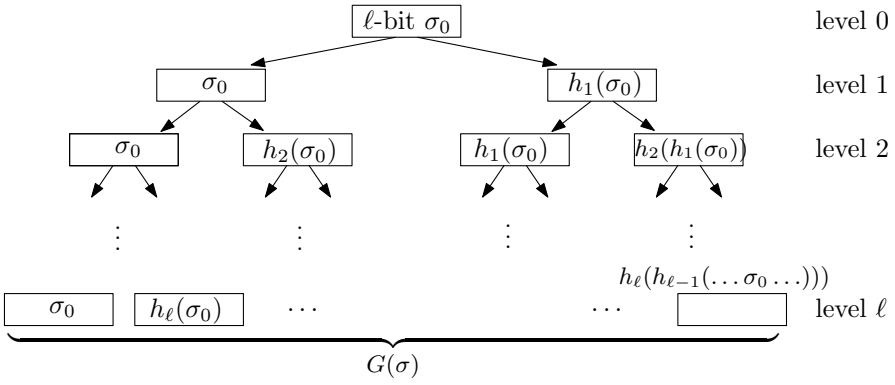
Each pair a_i, b_i defines a *hash function* $h_i\colon \mathrm{GF}(2^\ell) \to \mathrm{GF}(2^\ell)$ by $h_i(x) = a_i x + b_i$ (a small technical point is that we do not want a_i to be 0, but if the seed σ is random, we have $a_i = 0$ with probability $2^{-\ell}$, which will be negligible). Intuitively, 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 $a \in \mathrm{GF}(2^\ell) \setminus \{0\}$ and $b \in \mathrm{GF}(2^\ell)$ 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 $\mathrm{GF}(2)$. 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.

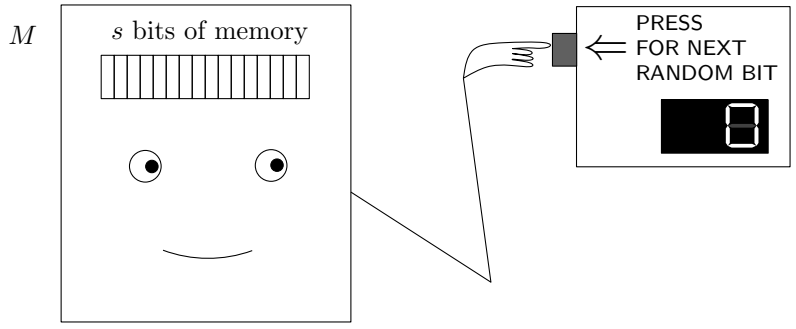
We construct a complete binary tree starting with a single node at level 0 with value σ_0. For a node at level i with value x, we construct two nodes at level $i+1$ with values x and $h_i(x)$. The string $G(\sigma)$ of length $\ell 2^\ell$ is the concatenation of the values of the leaves of the tree,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

2.6.4 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 machine M of the kind we are 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 $\omega = \omega_1\omega_2\omega_3\ldots$ 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 cannot 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 us 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.

Theorem 2.6.1 (Nisan's generator)**.** *If C in the above construction is a sufficiently large constant, then for all s and all machines M the probability distributions $p_{\text{truly}}(.)$ and $p_{\text{pseudo}}(.)$ are $2^{-\ell/10}$-close, which means that*

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

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 are ready to fix the random projection algorithm.

Theorem 2.6.2. *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 portion of $G(\sigma)$. 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 construct a hypothetical machine M and apply Theorem 2.6.1 to it.

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 $\|A\mathbf{x}\|_2^2$. If 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 pseudorandom bits from Nisan's generator, it finishes at GOOD with probability at least $1 - \delta - 2^{-\ell/10} \geq 1 - 2\delta$. 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 have 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.

2.6.5 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 $T\colon \ell_2^n \to \ell_1^{O(n)}$.

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.

2.7.1 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" of 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.

Lemma 2.7.1 (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 will 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 $\mathbf{E}[X^2] = 1$.

The upper bound $\mathbf{E}[\,|X|\,] \le \mathbf{E}[X^2] = 1$ 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 $\mathbf{E}[X^4]$ 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,

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

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 $\mathbf{E}[\epsilon_s^4]\, x_s^4 = x_s^4$ for some s, or $\mathbf{E}[\epsilon_s^2\epsilon_t^2]\, x_s^2 x_t^2 = x_s^2 x_t^2$ for some $s \neq t$. 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

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

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

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

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:

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

2.7.2 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 $\epsilon_1, \ldots, \epsilon_n$ with properties (i) and (ii) but on a probability space of size only $O(n^2)$.[9]

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

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

[9]For someone not familiar with t-wise independence, the first thing to realize is probably that 2-wise independence (every two of the variables 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 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) = A\mathbf{x}$. Then $\epsilon_1, \ldots, \epsilon_n$ 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 $\alpha_1, \ldots, \alpha_n$ 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:

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

here, e.g., $\alpha_1, \alpha_2, \ldots, \alpha_n$ represent m rows of A, since each α_i is interpreted as 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 $\alpha_i^1, \alpha_i^2, \ldots, \alpha_i^{2r}$ 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.

2.7.3 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 will 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 will not 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 (see, e.g., Ryan O'Donell's lecture notes at `www.cs.cmu.edu/~odonnell/boolean-analysis/lecture21.pdf`).

Theorem 2.7.4. *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 $\|\boldsymbol{\alpha}\|_2$.

Corollary 2.7.5. *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, and $\|\mathbf{x}\|_2 = 1$. Then $\mathbf{E}[\,|X'|\,] = \beta + O(b^{-1/2})$.*

Proof. We use the theorem with

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

□

The corollary as is provides 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:

Proposition 2.7.6. *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 do not 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

2.8.1 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 are not interested in all solutions —we are 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 $r = 0.08\,n$, i.e., about 8% of the transmitted data may be corrupted, we can take $n = 1.33\,m$, i.e., the encoding expands the message by 33%.

2.8.2 Compressed sensing

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 us 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 are 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 do not 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: Do not 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.)

2.8.3 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.

2.8.4 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:

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

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

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

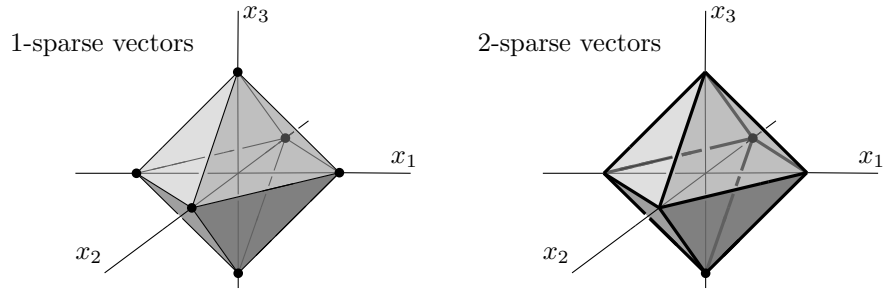
Such linear programs can be solved quite efficiently.[10]

2.8.5 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

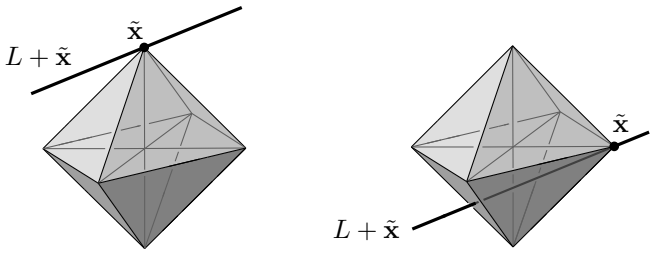
[10] Recently, alternative and even faster methods have been developed for computing a sparse solution of $A\mathbf{x} = \mathbf{b}$, under similar conditions on A, although they find the sparse solution only approximately.

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 $r = 1, 2$.

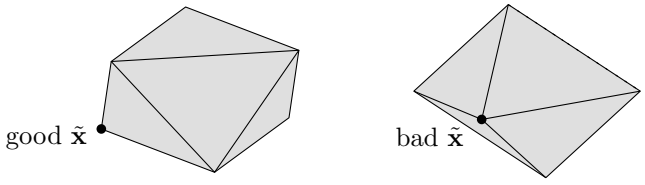


Let A be a $k \times n$ matrix of rank k and let $L = \operatorname{Ker} A$; then $\dim L = n-k = m$. A given r-sparse vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$ satisfies the linear system $A\mathbf{x} = \mathbf{b}_{\tilde{\mathbf{x}}}$, where $\mathbf{b}_{\tilde{\mathbf{x}}} = A\tilde{\mathbf{x}}$,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

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 $r = 1$:



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 will not need the just explained geometric interpretation in the proof.[11]

Theorem 2.8.1 (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 the general version of the random projection lemma —Lemma 2.4.1), 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.

Lemma 2.8.2. *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

[11] The geometric interpretation also explains why, when searching for a sparse solution, it is not 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.

[12] Sometimes abbreviated as 2-RIP.

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 $e^{-c_1\varepsilon_0^2 k}$ 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 $e^{3r\ln(n/r)}e^{-c_1\varepsilon_0^2 k} \le e^{-ck}$ for r, k, n as in the theorem. □

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.

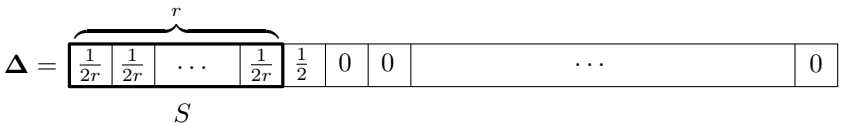
First we set $S = \{i : \Delta_i \neq 0\}$ and we observe that at least half of the ℓ_1 norm of $\boldsymbol{\Delta}$ has to live on S; in symbols,

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

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.

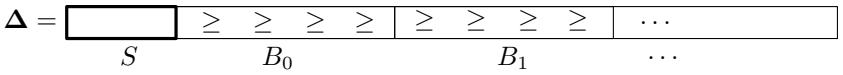
The first idea would be to use the restricted almost-isometry property to obtain $\|A_S\boldsymbol{\Delta}_S\|_2 \ge 0.9\frac{1}{2\sqrt{r}}$ (we use $\varepsilon_0 = 0.1$ for concreteness), and argue that the rest of the product, $A_{\overline{S}}\boldsymbol{\Delta}_{\overline{S}}$, is going to have smaller norm and thus $A\boldsymbol{\Delta} = A_S\boldsymbol{\Delta}_S + A_{\overline{S}}\boldsymbol{\Delta}_{\overline{S}}$ cannot be $\mathbf{0}$. This does not 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 $\|A_{S\cup B_0}\boldsymbol{\Delta}_{S\cup B_0}\|_2 \geq 0.9\|\boldsymbol{\Delta}_{S\cup B_0}\|_2 \geq 0.9\|\boldsymbol{\Delta}_S\|_2 \geq 0.9/2\sqrt{r} = 0.45/\sqrt{r}$. We want to show that

$$\sum_{j\geq 1} \|\boldsymbol{\Delta}_{B_j}\|_2 \leq \frac{0.4}{\sqrt{r}}, \tag{2.2}$$

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

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

reaching the desired contradiction.

Proving (2.2) is an exercise in inequalities. We know that $\sum_{j\geq 0} \|\boldsymbol{\Delta}_{B_j}\|_1 = \|\boldsymbol{\Delta}_{\overline{S}}\|_1 \leq \frac{1}{2}$. 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

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

Summing over $j \geq 1$, we have

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

which gives (2.2) and finishes the proof. □

Lecture 3

Lower Bounds

Some parts, not included here, were taught more or less according to Chapter 15 in the book *Lectures on Discrete Geometry* by J. Matoušek (lower bounds by a counting argument, lower bounds for the ℓ_1 cube and for expander graphs using inequalities, and Bourgain's embedding).

3.1 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 cannot 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.

Theorem 3.1.1. *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^d for some $D > 1$, we have $d \geq n^{0.02/D^2}$.*

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_∞^d. This kind of approach cannot work for the ℓ_1 case, since it is known that if ρ is an ℓ_1 metric, then $\sqrt{\rho}$ 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^d.

- We choose p a little larger than 1, namely, $p = 1 + \frac{1}{\ln d}$, and we observe that the "low-dimensional spaces" ℓ_1^d and ℓ_p^d are almost the same —the identity map is an $O(1)$-embedding (Lemma 3.1.2 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 will not insist on an isometric embedding, but we will be satisfied with distortion 2; see Lemma 3.1.3), but requires large distortion for embedding in ℓ_p (Lemma 3.1.4).

It follows that such an X does not embed well in ℓ_1^d, for if it did, it would also embed well in ℓ_p^d.

Lemma 3.1.2. *For $d > 1$ and $p = 1 + \frac{1}{\ln d}$, the identity map $\ell_1^d \to \ell_p^d$ has distortion at most* 3.

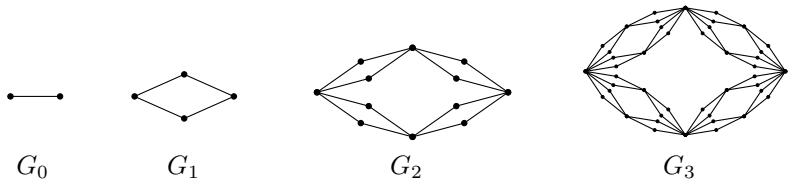
Proof. This is a very standard calculation with a slightly nonstandard choice of parameters. First, for $p_1 \le p_2$, we have $\|\mathbf{x}\|_{p_1} \ge \|\mathbf{x}\|_{p_2}$, and thus the identity map as in the lemma is nonexpanding. For the contraction Hölder's inequality and the standard estimate $1 + x \le e^x$ yield

$$\begin{aligned}\|\mathbf{x}\|_1 &= \sum_{i=1}^{d} 1 \cdot |x_i| \le d^{1-1/p}\|\mathbf{x}\|_p \\ &\le e^{(\ln d)(p-1)/p}\|\mathbf{x}\|_p = e^{(\ln d)/(1+\ln d)}\|\mathbf{x}\|_p \le 3\|\mathbf{x}\|_p.\end{aligned}$$

□

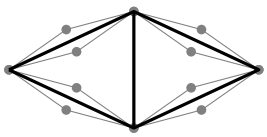
3.1.1 The recursive diamond graph

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_k, where $G_0, G_1, G_2, \dots$ is the following recursively constructed sequence:



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, $i = 0, 1, \dots$

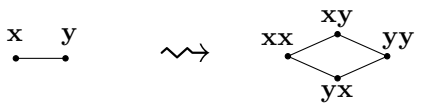
Since the vertex sets of the G_i form an increasing sequence, $V(G_0) \subset V(G_1) \subset \cdots$, we can regard $E_0, E_1, \dots, E_k$ and $A_1, \dots, A_k$ as sets of pairs of vertices of G_k. For example, the next picture shows E_1 and A_1 in G_2:



In G_k, the pairs in E_i and in A_{i+1} have distance 2^{k-i}.

Lemma 3.1.3. *Every G_k embeds in ℓ_1 with distortion at most 2.*

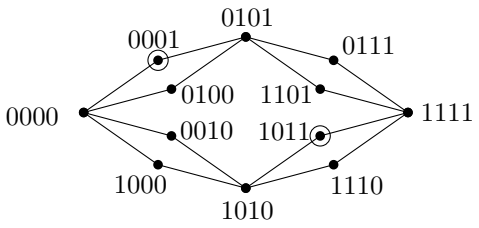
Sketch of proof. The embedding is very simple to describe. Each vertex of G_k is assigned a point $\mathbf{x} \in \{0,1\}^{2^k}$. We start with assigning 0 and 1 to the two vertices of G_0, and when G_{i+1} is constructed from G_i, the embedding for G_{i+1} is obtained as follows:



(**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, since another embedding is known with distortion only $\frac{4}{3}$.

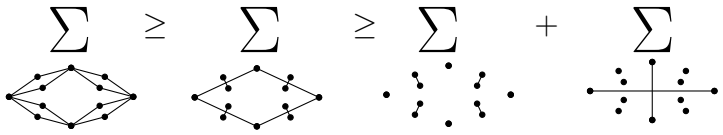
Finally, here is the promised nonembeddability in ℓ_p.

Lemma 3.1.4. *Any embedding of G_k in ℓ_p, $1 < p \le 2$, has distortion at least $\sqrt{1+(p-1)k}$.*

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 $E = E_k = E(G_k)$ and $F = E_0 \cup A_1 \cup A_2 \cup \cdots \cup A_k$. With ρ denoting the shortest-path metric of G_k, we have $\sum_E \rho(u,v)^2 = |E_k| = 4^k$ and $\sum_F \rho(u,v)^2 = 4^k + \sum_{i=1}^k |A_i| 4^{k-i+1} = 4^k + \sum_{i=1}^k 4^{i-1}4^{k-i+1} = (k+1)4^k$. So the ratio of the sums over F and over E is $k+1$.

Next, let us consider an arbitrary map $f\colon V(G_k) \to \ell_2$, 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_k, we get that $S_E \ge \sum_{A_k \cup E_{k-1}} \|f(u) - f(v)\|_2^2$. Next, we keep the sum over A_k and we bound the sum over E_{k-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 $\sqrt{k+1}$.

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.

Lemma 3.1.5 (Short-diagonals Lemma for $\boldsymbol{\ell_p}$)**.** *For every four points* $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4 \in \ell_p$ *we have*

$$\begin{aligned}&\|\mathbf{x}_1 - \mathbf{x}_3\|_p^2 + (p-1)\|\mathbf{x}_2 - \mathbf{x}_4\|_p^2 \\ &\le \|\mathbf{x}_1 - \mathbf{x}_2\|_p^2 + \|\mathbf{x}_2 - \mathbf{x}_3\|_p^2 + \|\mathbf{x}_3 - \mathbf{x}_4\|_p^2 + \|\mathbf{x}_4 - \mathbf{x}_1\|_p^2.\end{aligned}$$

This lemma is a subtle statement, optimal in quite a strong sense, and we prove it in the next section. Here we just note that, unlike the inequalities used earlier, the norm does not 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 $f\colon V(G_k) \to \ell_p$ and derive the inequality

$$\|f(s) - f(t)\|_p^2 + (p-1) \sum_{A_1 \cup A_2 \cup \cdots \cup A_k} \|f(u) - f(v)\|_p^2 \le \sum_E \|f(u) - f(v)\|_p^2,$$

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_k, Lemma 3.1.4 follows. □

Proof of Theorem 3.1.1. We follow the outline. Let $f\colon V(G_k) \to \ell_1$ be a 2-embedding and let $X = f(V(G_k))$. Assuming that $(X, \|.\|_1)$ can be D-embedded in ℓ_1^d, we have the following chain of embeddings:

$$G_k \xrightarrow{2} X \xrightarrow{D} \ell_1^d \xrightarrow{3} \ell_p^d.$$

The composition of these embedding is a $6D$-embedding of G_k in ℓ_p, and so $6D \ge \sqrt{1 + (p-1)k}$ with $p = 1 + \frac{1}{\ln d}$. It remains to note that $n = |V(G_k)| \le 4^k$ for all $k \ge 1$. The theorem then follows by a direct calculation. □

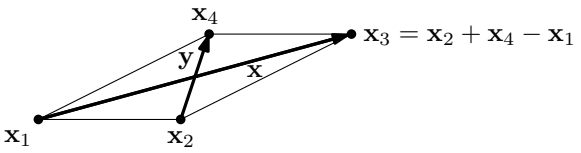
3.2 Proof of the short-diagonals lemma for $\boldsymbol{\ell_p}$

Lemma 3.1.5 is an easy consequence of the following inequality:

$$\frac{\|\mathbf{x} + \mathbf{y}\|_p^2 + \|\mathbf{x} - \mathbf{y}\|_p^2}{2} \ge \|\mathbf{x}\|_p^2 + (p-1)\|\mathbf{y}\|_p^2, \quad 1 < p < 2, \tag{3.1}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ 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.)

Proof of Lemma 3.1.5 assuming (3.1). For understanding this step, it is useful to note that (3.1) 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 (3.1) 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

$$\|\mathbf{x}_2 + \mathbf{x}_4 - 2\mathbf{x}_1\|_p^2 + (p-1)\|\mathbf{x}_4 - \mathbf{x}_2\|_p^2 \le 2\|\mathbf{x}_4 - \mathbf{x}_1\|_p^2 + 2\|\mathbf{x}_2 - \mathbf{x}_1\|_p^2. \quad (3.2)$$

Now if $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ are arbitrary, we use (3.1) 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 (3.2), 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

$$\|\mathbf{x}_2 + \mathbf{x}_4 - 2\mathbf{x}_3\|_p^2 + (p-1)\|\mathbf{x}_4 - \mathbf{x}_2\|_p^2 \le 2\|\mathbf{x}_4 - \mathbf{x}_3\|_p^2 + 2\|\mathbf{x}_2 - \mathbf{x}_3\|_p^2. \quad (3.3)$$

Taking the arithmetic average of (3.2) and (3.3) we almost get the inequality we want, *except* that we have $\frac{1}{2}(\|\mathbf{x}_2 + \mathbf{x}_4 - 2\mathbf{x}_1\|_p^2 + \|\mathbf{x}_2 + \mathbf{x}_4 - 2\mathbf{x}_3\|_p^2)$ instead of $\|\mathbf{x}_1 - \mathbf{x}_3\|_p^2$ as we would 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 $\frac{1}{2}(\|\mathbf{a}\|_p^2 + \|\mathbf{b}\|_p^2) \ge \|(\mathbf{a}+\mathbf{b})/2\|_p^2$ 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 (3.1). 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 `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 (3.1), with the same right-hand side:

$$\left(\frac{\|\mathbf{x}+\mathbf{y}\|_p^p + \|\mathbf{x}-\mathbf{y}\|_p^p}{2}\right)^{2/p} \ge \|\mathbf{x}\|_p^2 + (p-1)\|\mathbf{y}\|_p^2. \quad (3.4)$$

To see that the left-hand side of (3.4) is never smaller than the left-hand side of (3.1), 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 (3.4) 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 left-hand side 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 right-hand side 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 (3.4) cannot be improved.

The basic idea of the proof of (3.4) 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 (3.4), 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 $L(t)$, write $f(t) = (\|\mathbf{x}+t\mathbf{y}\|_p^p + \|\mathbf{x}-t\mathbf{y}\|_p^p)/2$. Then

$$\begin{aligned} L'(t) &= \frac{2}{p} f(t)^{\frac{2}{p}-1} f'(t) \\ &= \frac{2}{p} f(t)^{\frac{2}{p}-1} \frac{p}{2} \sum_i \Big(|x_i + ty_i|^{p-1} \operatorname{sgn}(x_i + ty_i) y_i \\ &\qquad - |x_i - ty_i|^{p-1} \operatorname{sgn}(x_i - ty_i) y_i \Big) \end{aligned}$$

(we note that the function $z \mapsto |z|^p$ has a continuous first derivative, namely $p|z|^{p-1}\operatorname{sgn}(z)$, provided that $p > 1$). The above formula for $L'(t)$ shows that $L'(0) = 0$.

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 is not 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

$$L''(t) \geq (p-1) f(t)^{\frac{2}{p}-1} \|\mathbf{y}\|_p^2 \left(\|\mathbf{x} + t\mathbf{y}\|_p^p + \|\mathbf{x} - t\mathbf{y}\|_p^p\right) = 2(p-1)\|\mathbf{y}\|_p^2.$$

We have thus proved that $L''(t) \geq R''(t)$ 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. □

Index