2004 Barbados Workshop on Computational Complexity

LECTURE NOTES ON FINITE METRIC SPACES AND LOW-DIMENSIONAL EMBEDDINGS

Title: Introduction	Lecture #: 1
Lecturer: Nati Linial	Scribe: Benoit Larose and Alexander Russell

1 Introduction

This lecture series will focus primarily on *finite metric spaces*.

Definition 1. *A* metric space *is a pair* (X, d) *where* X *is a non-empty set and* $d : X \times X \to \mathbb{R}^+$ *is a map satisfying, for all* $x, y, z \in X$*, the following axioms:*

1. $d(x, y) = d(y, x) \geq 0$*;*
2. $d(x, y) = 0$ *if and only if* $x = y$*;*
3. $d(x, y) + d(y, z) \geq d(x, z)$ *(the* triangle inequality*).*

Motivation. Examples of infinite metric spaces abound in mathematics. Our focus on finite metric spaces is motivated by their connection to several important families of algorithmic problems. Consider, for example, the following concrete example from bioinformatics: let X be the set of all known proteins (so that $|X| \approx 1.5 \times 10^6$). A *protein* may be viewed as a word of length between 40 and 5000 characters (about 400 on average) over an alphabet of size 20 (the *amino acids*). There are various interesting metrics one may impose on this set . For instance, consider "edit distance," i.e. the number of changes required to modify one word into another, useful in finding common evolutionary sources. (It should be pointed out that for some of these "metrics" the triangle inequality actually fails, but not in any fatal way.) There are many other relevant examples of large data sets, e.g., the files on the Internet. In biology, one often adopts a notion of *similarity*: given a *similarity function* $s(x, y)$ one defines a distance function d by

$$d(x, y) := s(x, x) + s(y, y) - 2s(x, y) \ .$$

One natural analytic framework for finite metric spaces is that of *clustering*: this is the process of decomposing a metric space in small "clouds" of points which are "close," the pieces being themselves "far" from one another (see Figure 1).

Remark. There is still a fairly wide gap between theory and practice in this area, and still much to be done in the field. Furthermore, there exist deep and beautiful mathematics related to finite metric spaces:

- The theory has very fruitful connections with the theory of Banach spaces.
- The study of finite metric spaces has given rise to interesting new algorithms (see Santosh Vempala's lectures).
- The theory is related to the highly successful *"geometrisation" of combinatorics* paradigm and, specifically, the study of *graph metrics*. (A *graph metric* is the metric induced on the vertices of a weighted graph G by assigning the distances between two vertices x and y to be the length of the shortest path between them.) See, for instance, the 1994 paper of Goemans and Williamson [14].

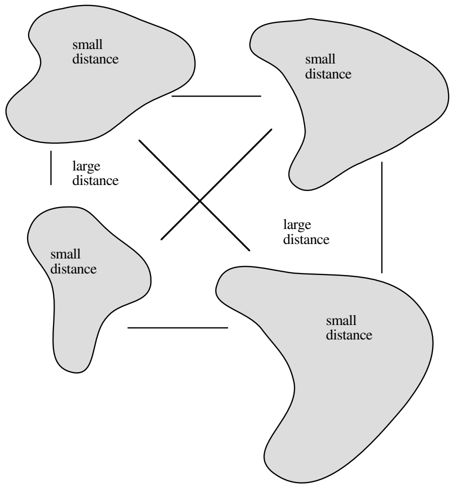

Figure 1: Clustering

2 Banach spaces; preparatory functional anaylsis

We shall now turn to Banach spaces, concentrating specifically on spaces of finite dimension.

Definition 2. *An n-*dimensional normed vector space *is a pair $(V, \| \ \|)$, where V is an n-dimensional $\mathbb{R}$-vector space and $\| \ \| : V \to \mathbb{R}$ is a* norm *map satisfying, for all $x, y \in V$ and all $\alpha \in \mathbb{R}$:*

1. $\|x\| \geq 0$*, and* $\|x\| = 0$ *if and only if* $x = 0$*;*
2. $\|\alpha x\| = |\alpha| \|x\|$*;*
3. $\|x + y\| \geq \|x\| + \|y\|$ *(the* triangle inequality*).*

Every normed space is a metric space, where the distance map is obtained by setting $d(x, y) = \|x - y\|$ for all $x, y \in X$. A *Banach space* is a *complete* normed space, i.e. a space in which every Cauchy sequence converges. (Note that Banach spaces need not have finite dimension, in general.) All finite-dimensional normed spaces are complete, and in fact all normed spaces considered in these lectures are Banach spaces.

Let us return to our example of proteins for a moment. In a typical instance, one is supplied with an enormous data set describing a finite metric space. As the set is both large and, potentially, poorly structured, carrying out basic operations on the associated metric space (e.g., clustering) is infeasible. One approach to this problem is to embed the metric space into a metric space with better geometric properties, ideally a low-dimensional Banach space. Specifically, one can ask when there exists a map Old version:

$$(X, d) \hookrightarrow (\mathbb{R}^n, \| \ \|)$$

which preserves the properties we wish to study.

Definition 3. *Let $(\mathbb{R}^n, \| \ \|)$ be a normed space. Define the* unit ball *of $(\mathbb{R}^n, \| \ \|)$ as the set of all vectors with norm no more than 1. Specifically,*

$$B_{\| \ \|} := \{x \in \mathbb{R}^n \mid \|x\| \leq 1\} \ .$$

The set $B = B_{\| \ \|}$ is a *centrally symmetric convex body of full dimension* in the space $(\mathbb{R}^n, \| \ \|)$; to be precise, it satisfies the following conditions:

1. if $x \in B$ then $-x \in B$ (*central symmetry*);
2. if $x, y \in B$ then $\lambda x + (1-\lambda)y \in B$ for all $0 \leq \lambda \leq 1$ (*convexity*);
3. B has non-empty interior, i.e. there exists some $x \in B$ with the following property: for all $y \in \mathbb{R}^n$, there exists an $\varepsilon > 0$ such that $x + \lambda y \in B$ for all $|\lambda| \leq \varepsilon$ (B is a *body*).

The ball $B_{\|\,\|}$ completely determines the geometry of the Banach space: given any $u \in \mathbb{R}^n$ with $\|u\| = k \neq 0$, let $v = \frac{1}{k}u$. Then

$$\|u\| = \frac{\|u\|_2}{\|v\|_2}$$

where $\|\,\|_2$ denotes the standard norm in $\mathbb{R}^n$.

Examples. The most important class of examples consists of the spaces ℓ_p^n. Let $1 \leq p \leq \infty$ and let n be a positive integer. The space ℓ_p^n (*n-dimensional* ℓ_p) is the pair $(\mathbb{R}^n, \|\,\|_p)$ where

$$\|(x_1, \ldots, x_n)\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and all $1 \leq p < \infty$; the norm $\|\,\|_\infty$ is given by

$$\|(x_1, \ldots, x_n)\|_\infty = \max_i |x_i|$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

The space ℓ_p is the infinite-dimensional analog: for $1 \leq p < \infty$ it is defined as the pair $(X, \|\,\|_p)$ where X consists of all (countably) infinite sequences of real numbers $(x_1, x_2, \ldots, x_n, \ldots)$ such that

$$\sum_{i=1}^\infty x_i^p < \infty \quad .$$

(the norm is defined in the obvious way); for $p = \infty$ the base set consists of all bounded sequences $(x_1, \ldots, x_k, \ldots)$ of real numbers and the norm is simply the supremum of the $|x_i|$.

Here are, in the plane, the unit balls of the three most important examples of these, namely the cases $p = 1, 2, \infty$:

The unit ball of ℓ_2. This is the familiar "round" ball ($\sum_{i=1}^n x_i^2 \leq 1$);

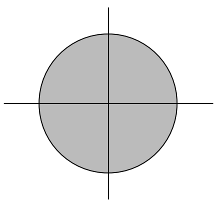

Figure 2: The ℓ_2-ball

The unit ball of ℓ_1. This is the octahedron ($\sum |x_i| \leq 1$), which is the convex hull spanned by the basis vectors $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$;

The unit ball of ℓ_∞. This is the convex set $|x_i| \leq 1$ spanned by the points $(x_1, \ldots, x_n)$ with $x_i \in \{-1, 1\}$.

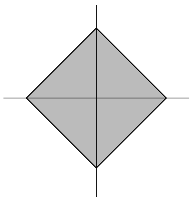

Figure 3: The ℓ_1-ball.

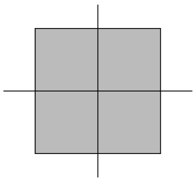

Figure 4: The ℓ_∞-ball

2.1 Elements of functional analysis

Definition 4. *Let $(X, \|\ \|)$ be a normed space. The* dual *of $(X, \|\ \|)$, denoted by $(X, \|\ \|)^*$ or X^* when the norm is understood, is the normed space of all linear functionals $f : X \to \mathbb{R}$ equipped with the norm*

$$\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|} \quad .$$

For instance, we have that $\ell_p^* = \ell_q$ when $1/p + 1/q = 1$ and hence that $(\ell_2^n)^* = \ell_2^n$ is its own dual space of functions. Observe that if $x_0 \in X$ then the map

$$x \mapsto \langle x, x_0 \rangle$$

is a linear functional on $\mathbb{R}^n$, and in fact every linear functional is of this form. Of course, an element $x \in X$ defines a natural linear functional E_x on X^*, given simply by evaluation (which is to say that $E_x(f) = f(x)$). This describes a *natural* (basis-independent) map $X \to X^{**}$ which, for finite-dimensional spaces, is an isomorphism. For infinite-dimensional spaces we have only $X^{**} \subseteq X$ in general. Finally note that ℓ_1^n and ℓ_∞^n are dual to one another.

The original motivation for functional analysis came from the study of differential equations: this leads to infinite-dimensional spaces, for instance the spaces L_p:

Definition 5. *Let (X, μ) be a (complete) measure space. $L_p(X, \mu)$ denotes the space of all measurable functions that satisfy*

$$\int |f|^p d\mu < \infty$$

with the norm

$$\left[\int |f|^p d\mu\right]^{1/p} \quad .$$

These are indeed generalisations of the previous examples, obtained by taking the measure space (X, μ) to be $\{1, 2, \dots, n\}$ with the uniform measure. The *Sobolev* spaces are smooth versions of the L_p-spaces.

3 Maps between metric spaces

The "ideal" maps between metric spaces are the *isometries*:

Definition 6. *Let (X, d) and (Y, ρ) be metric spaces. A map $\phi : X \to Y$ is an* isometry *if for all $x_1, x_2 \in X$ we have that $\rho(\phi(x_1), \phi(x_2)) = d(x_1, x_2)$.*

More generally, we shall relax this notion and focus on the following:

Definition 7. *Let (X, d) and (Y, ρ) be metric spaces. The* expansion *of a map $\psi : X \to Y$ is defined by*

$$\textit{expansion}(\psi) = \max_{x_1 \neq x_2 \in X} \frac{\rho(\psi(x_1), \psi(x_2))}{d(x_1, x_2)} \;\; ;$$

the contraction *of ψ is*

$$\textit{contraction}(\psi) = \max_{x_1 \neq x_2 \in X} \frac{d(x_1, x_2)}{\rho(\psi(x_1), \psi(x_2))};$$

finally, the distortion *of ψ is*

$$\textit{distortion}(\psi) = \textit{expansion}(\psi) \times \textit{contraction}(\psi) \;\; .$$

We say that a map $f : (X, d) \to (Y, \rho)$ has *distortion at most c* if there exists some $a > 0$ such that for all distinct $x_1, x_2 \in X$ we have

$$a \leq \frac{\rho(f(x_1), f(x_2))}{d(x_1, x_2)} \leq ca \;\; .$$

Remark. In cases where (at least) one of the two metric spaces involved is a vector space, a map f can naturally be *scaled* by a multiplicative constant $\alpha \in \mathbb{R}$. (If $f : X \to V$ and V is an $\mathbb{R}$-vector space, then $(\alpha f)(x)$ can be defined $f(x) = \alpha f(x)$; if $f : V \to X$, then (αf) can be defined $(\alpha f)(x) = f(\alpha x)$.) Observe that this family of maps contains examples with arbitrary expansion and contraction, but that the notion *distortion* is invariant under scaling.

General question. *Given a finite metric space (X, d), and a "nice" family of metric/normed spaces $\mathcal{F}$, find some $Y \in \mathcal{F}$ and a map $f : X \to Y$ that has small(est) distortion.*

This is related to the *Banach-Mazur distance*, a notion of distance between two normed spaces.

Definition 8. *Let $(X, \|\;\|_X)$ and $(Y, \|\;\|_Y)$ be two normed spaces. The* Banach-Mazur distance *between these spaces is the logarithm of the following quantity:*

$$d(X, Y) = \inf_{\substack{f : X \to Y \\ f \text{ linear, one-to-one}}} \|f\| \, \|f^{-1}\|$$

where

$$\|f\| = \sup_{x \neq 0} \frac{\|f(x)\|_Y}{\|x\|_X} \;\; .$$

This notion of distortion is a "bi-Lipshitz condition;" in our setting of finite metric spaces, distortion may be viewed as a non-linear version of the Banach-Mazur distance.

We begin now with our first embedding theorem.

Theorem 1 (Fréchet). *Every n-point metric space embeds isometrically in ℓ_∞^n (ℓ_∞ is "universal").*

Proof. Let $X = \{x_1, \ldots, x_n\}$ with metric d, and let A be the $n \times n$ matrix given by

$$a_{ij} = d(x_i, x_j)$$

for all $1 \leq i, j \leq n$. Let p_i denote the i-th row of A and consider the map from X into $\mathbb{R}^n$ given by the rule

$$x_i \longmapsto p_i \;\; .$$

This is an isometry: indeed, the ℓ_∞ distance between p_i and p_j satisfies

$$\|p_i - p_j\|_\infty = \max_{1 \leq k \leq n} |a_{ik} - a_{jk}| \geq |a_{ij} - a_{jj}| = d(x_i, x_j)$$

and for every $1 \leq k \leq n$ we have, by the triangle inequality,

$$|a_{ik} - a_{jk}| = |d(x_i, x_k) - d(x_k, x_j)| \leq d(x_i, x_k) + d(x_k, x_j) \leq d(x_i, x_j) \ .$$

$\square$

2004 Barbados Workshop on Computational Complexity

LECTURE NOTES ON FINITE METRIC SPACES AND LOW-DIMENSIONAL EMBEDDINGS

Title: Distortion versus dimension — Lecture #: 2

Lecturer: Nati Linial — Scribe: François Lemieux and Leonid Libkin

4 Some comments on the Banach-Mazur distance

Recall that if $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed spaces then their Banach-Mazur distance is defined as

$$d(X,Y) = \inf_{f \in \mathcal{L}} \|f\| \cdot \|f^{-1}\|$$

where $\mathcal{L}$ is the set of all linear functions $f : X \to Y$ and

$$\|f\| = \sup_{x \in X} \frac{\|f(x)\|_Y}{\|x\|_X} \quad .$$

Hence, if the Banach-Mazur distance between two normed spaces is d, then there is a (linear) d-embedding from one space into the other. The function $d(X,Y)$ itself is not a distance. To define a metric on normed spaces one has to use $\log d(X,Y)$ instead. The following theorem gives an upper bound on the Banach-Mazur distance between ℓ_2 and all other Banach spaces.

Theorem 2 (Löwner-John). *If X is an n-dimensional Banach space, then $d(X, \ell_2^n) = O(\sqrt{n})$.*

This result is tight for ℓ_1, i.e. $d(\ell_1^n, \ell_2^n) = \Theta(\sqrt{n})$. As a corollary we have that if X and Y are any two n-dimensional normed spaces then $d(X,Y) \le n$. This simply follows from the triangle inequality:

$$\log d(X,Y) \le \log d(X, \ell_2^n) + \log(Y, \ell_2^n) \quad .$$

There exists a bijection between Banach spaces and centrally symmetric convex bodies in $\mathbb{R}^n$. Given any norm over $\mathbb{R}^n$, the unit ball B^n is always symmetric and convex. Conversely, given any centrally symmetric convex body $B \subseteq \mathbb{R}^n$, one can define the norm

$$\|x\|_B = \min \{t > 0 \mid \frac{x}{t} \in B\}$$

(the triangle inequality then follows from the convexity of B).

Let B^n be the unit ball in ℓ_2^n. A (centrally symmetric) ellipsoid E is a convex body that can be defined as $f(B^n)$ where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear function. If X is the Banach space over $\mathbb{R}^n$ with norm $\| \, \|_E$ then f is an isometry from ℓ_2^n to X. It follows that Theorem 2 is a corollary of the next lemma which implies that any normed space X can be $\sqrt{n}$-embedded in a norm space Y whose unit ball forms an ellipsoid.

Lemma 1 (John [18]). *Given any symmetric convex body $B \subseteq \mathbb{R}^n$ there exists an ellipsoid E such that $E \subseteq B \subseteq \sqrt{n}E$.*

This result is tight for the cube $[-1,1]^n$. John's Lemma can be proved by taking an ellipsoid $E \subseteq B$ of maximal volume and then showing that $\sqrt{n}E$ contains B. Alternatively, one can start with an ellipsoid of minimal volume that contains B.

5 The theorem of Dvoretsky

In the last lecture, we have seen the Fréchet Theorem which states that every finite metric space embeds isometrically into ℓ_∞. What about embedding into the Euclidean space ℓ_2? We first consider the dual question: In what spaces can we embed ℓ_2?

Theorem 3 (Dvoretsky [10, 9]). *Let $(X, \| \|)$ be an n-dimensional Banach space with unit ball B and let $\epsilon > 0$ be given. Then, there exists a k-dimensional subspace L with $k = \Omega(\epsilon^2 \log n)$ such that $B \cap L$ is, up to $1 + \epsilon$, an ℓ_2-ball (i.e. B contains an ℓ_2-ball of radius r and is contained in an ℓ_2-ball of radius $r(1 + \epsilon)$).*

The full proof of this theorem will not be given here but Nati mentions that two major techniques that are used for proving results in this area are the *concentration of measure* and the use of clever *random projections*. Random projections will be discussed by Santosh in a next lecture. Concentration of measure can be introduced with the following *isoperimetric problem*. Let μ be a measure on some metric space. For any subset X, let X_ϵ be the ϵ-neighborhood of X, that is the set of points whose distance from X is bounded above by ϵ. Given $\epsilon > 0$ and $m > 0$, the problem is to find a subset X such that $\mu(X) = m$ and $\mu(X_\epsilon)$ is minimal.

For example, consider the unit sphere S^{n-1} of the n-dimensional Euclidean space. What is the subset $X \subseteq S^{n-1}$ that minimizes $\mu(X_\epsilon)$? The answer is that we can always take X as a spherical cap. In particular, if $m = \mu(S^{n-1})/2$ then X can be some hemisphere of S_{n-1} and it can be shown that $\mu(X_\epsilon) \approx 1 - e^{-\frac{\epsilon^2 n}{2}}$ (see [26] for more details). Since the same result holds if we use the other hemisphere: in particular, this shows that for large n "most" points lie on a narrow band around the sphere S^{n-1}.

A form of the measure concentration that is used in the proof of Dvoretsky's Theorem concerns the *Lipschitz* functions. A function $f : S^{n-1} \to \mathbb{R}$ is 1-Lipschitz if $|f(x) - f(y)| \leq \|x - y\|_2$ for all $x, y \in S^{n-1}$. For instance, if a convex body B contains the unit ℓ_2-ball, then the function $f(x) = \|x\|_B$ is 1-Lipschitz. The proof of Dvoretsky's Theorem uses the following lemma which states that 1-Lipschitz functions are concentrated around their median where the median of f is defined as the largest m such that $\Pr[f(x) \leq m] \leq 1/2$.

Lemma 2 (Lévy [19]). *Let $f : S^{n-1} \to \mathbb{R}$ be 1-Lipschitz and let m be the median of f. For any $\epsilon > 0$ we have*

$$\Pr[|f - m| > \epsilon] < 2e^{-\frac{\epsilon^2 n}{2}}$$

From Dvoretsky's Theorem we know that ℓ_2^n sits inside every high dimensional normed space. However, it is not true that every metric space can be embedded in ℓ_2. For example, any undirected finite graph can be seen as a metric space where the distance between two nodes is given by the length of the shortest path from one node to the other. Consider now the graph of Figure 5, which cannot be isometrically embedded into ℓ^2. The best embedding that can

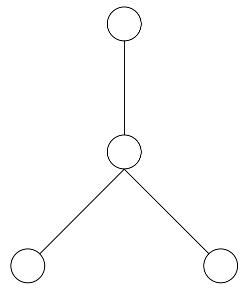

Figure 5: A metric space that cannot be isometrically embedded into ℓ^2

be achieved in ℓ_2^2 is the one appearing in the figure with angles of $120°$. The lowest possible distortion is $\sqrt{2}$ (this can be done in ℓ_2^3). As another example, for any $d \geq 2$ the ℓ_2 distortion of the d-cube is $\sqrt{d}$ (see [11]).

6 Distortion vs. dimension: Matoušek's theorem

The next theorem states that if many different n-point metric spaces embed with low distortion in a space X then the dimension of X must be high.

Theorem 4 (Matoušek [24]). *Let X be a normed space in which every n-point metric space embeds with a distortion d. If k is the dimension of X then we have:*

- *if $d < 3$ then $k \geq \Omega(n)$;*

- *if* $d < 5$ *then* $k \geq \Omega(n^{1/2})$*;*
- *if* $d < 7$ *then* $k \geq \Omega(n^{1/3})$.

It is believed that the above theorem could be generalized for all odd integers.

7 Proof: Graphs of high girth

To prove this, we use graph metrics, like the one shown above. We shall specifically focus on graphs of *high girth*. Recall that the *girth* of a graph G is the length of the shortest cycle in G.

Our goal is to prove the following.

Lemma 3. *Let G be a graph of girth l, with n vertices and m edges. Suppose X is a normed space into which the metric of any subgraph of G embeds with distortion $< D$, such that $D < l \leq 5D$* [1]*. Then*

$$\dim(X) \;\geq\; \frac{m}{n} \cdot \frac{1}{\log_2 \frac{16Dl}{l-D}}$$

Given this lemma we should like to find, for a given n, graphs which simultaneously achieve large m and l. These two requirements, of course, are contradictory, but one simply looks for large m for a given l.

Example 1. *To start with, take complete bipartite graph G with n vertices. The girth is 4, and so applying the lemma we see that if every n-point metric space (in particular, every metric space induced by a subgraph of G) embeds into X with distortion ≤ 3, then* $\dim(X) \geq \Omega(n)$ *– this is the first item in Matoušek's theorem.*

Example 2. *A more interesting example shows how the second line is proved. For that, we need the existence of graphs of girth ≥ 6 with $m = \Omega(n^{3/2})$ (just plug this into the equation in the Lemma and we get* $\dim(X) \geq \Omega(\sqrt{n})$*).*

To get this, consider $\mathbb{F}_q$ and an equivalence relation $\sim$ in $\mathbb{F}_q^3 - \{(0,0,0)\}$ defined by $(x,y,z) \sim (x',y',z')$ iff $x' = \lambda x, y' = \lambda y, z' = \lambda z$ for some $\lambda \in \mathbb{F}_q$. Consider the set U of equivalence classes (there are q^2+q+1 of them) and define a subgraph of $U \times U$ that contains an edge from (x,y,z) to (a,b,c) iff $ax+by+cz = 0$. This is a bipartite graph, and one can show (by linear algebra) that it has no 4-cycles—hence the girth is at least 6.

How do we prove the Lemma? There are two main ingredients:

- If G has high girth and $H_1 \neq H_2$ are two subgraphs of G, then their metrics must be "quite different".
- If X has low dimension, then it cannot admit, with low distortion, embedding of many distinct metrics.

For the first item, take $H_1 \neq H_2$, two subgraphs of G, whose girth is l. Suppose (u,v) is an edge in H_1 but not H_2. Then $d_{H_1}(u,v) = 1$ and $d_{H_2}(u,v) \geq l-1$ (otherwise the girth would be lower).

For the second one, more work is needed. Set $\beta = \frac{1}{4} \cdot (\frac{l}{D} - 1)$. Consider a *$\beta$-net* in $l \cdot \mathbb{B}_X$ (reminder: $\mathbb{B}_X$ is the unit ball). That is, a finite subset $S \subset l \cdot \mathbb{B}_X$ such that for every point $x \in l \cdot \mathbb{B}_X$ there is $y \in S$ with $\|x-y\|_X \leq \beta$. The reason we need this is that if we modify the embedding to go to elements of S as opposed to arbitrary elements of X, then this is an embedding into a discrete set, and the distortion does not change much.

We now find an upper bound on the the size of a β-net. We know what the number of different subgraphs of G is: 2^m. If we show that shifting the embedding to the β-net S doesn't change the distortion much, and put in these three elements together, we'll get the Lemma.

We now look at the size of the β-net. Let d stand for $\dim(X)$. Then we claim $|S| \leq (\frac{4l}{\beta})^d$. To see this, we construct the β-net greedily, adding a point at distance $> \beta$ from the current set S. This way $\frac{\beta}{2}$-balls are disjoint. When we ended up with a β-net of size N, we have

$$\mathrm{vol}(\mathbb{B}_X) \cdot (l+\beta)^d \;>\; N \cdot \mathrm{vol}(\mathbb{B}_X) \cdot \left(\frac{\beta}{2}\right)^d$$

[1] Nati makes the following remark: $l \leq 5D$ is not so important but $D < l$ is essential for this result.

which gives us a bound on N.

We finally connect this with embeddings of metrics of subgraphs of G. Let $H_1 \neq H_2$ be two subgraphs of G, with (u, v) an edge in H_1 but not H_2. Let $f_{H_1} : H_1 \to X$ and $f_{H_2} : H_2 \to X$ be non-expansive embeddings in X with distortion $< D$. We define g_{H_i} as the "rounding" of f_{H_i}, $i = 1, 2$: that is, $g_{H_i}(x)$ is the element of the β-net that is closest to $f_{H_i}(x)$.

We now calculate

$$\|g_{H_1}(u) - g_{H_1}(v)\|_X \leq \|f_{H_1}(u) - f_{H_1}(v)\|_X + 2\beta \leq 1 + 2\beta$$

and

$$\|g_{H_2}(u) - g_{H_2}(v)\|_X \geq \|f_{H_2}(u) - f_{H_2}(v)\|_X - 2\beta \geq \frac{l-1}{D} - 2\beta > 1 + 2\beta \ .$$

So the roundings are different for H_1 and H_2, which shows $2^m \leq |S|^n$. Thus,

$$2^m \ \leq \left(\frac{4l}{\beta}\right)^{dn} \ .$$

Putting in the definition of β, we see that $\frac{4l}{\beta} = \frac{16Dl}{l-D}$, and then taking $\log_2$ in the above inequality gives us the result of the lemma.

2004 Barbados Workshop on Computational Complexity

LECTURE NOTES ON FINITE METRIC SPACES AND LOW-DIMENSIONAL EMBEDDINGS

Title: Random projections and dimension reduction — Lecture #: 3

Lecturer: Santosh Vempala — Scribe: Pascal Tesson and Ricard Gavaldà

8 Dimension reduction

We have, thus far, focused on low-distortion embeddings of metric spaces into ℓ_p without worrying about the dimension of the target space. In this lecture we consider the challenge of embedding a finite metric space X (given as a submetric of the Euclidean space $\mathbb{R}^n$) into the Euclidean space $\mathbb{R}^k$ while trying to minimize both the distortion and the target dimension k. Throughout this lecture $\|\cdot\|$ denotes the euclidean norm $\|\cdot\|_2$.

For isometric (i.e., no distortion) embeddings, the following two claims give the complete picture:

Claim 1. *If $|X| = m$, then X embeds isometrically into $\mathbb{R}^{m-1}$.*

Indeed, if one of the points is moved to the origin, then X embeds isometrically in the $m - 1$ dimensional space spanned by the remaining points.

Claim 2. *For some sets X, dimension $m - 1$ is necessary to achieve isometry.*

For example, the metric of the complete graph K_m does not embed isometrically in less than $m - 1$ dimensions since there cannot be m equidistant points in $\mathbb{R}^k$ with $k < m - 1$.

We will look next at what happens when we tolerate some distortion in the embedding.

9 From dimension n to dimension 1

Let $\phi : u \mapsto u'$ be a map from $\mathbb{R}^n \to \mathbb{R}^k$ with $k < n$. Can ϕ be good at preserving distances? How different can $\|u - v\|$ and $\|u' - v'\|$ be? An equivalent question is to ask how good ϕ can be at preserving norms or, for that matter, the norm of a unit vector.

We start by studying the (instructive) case $k = 1$. Consider the map $\phi(u) = u \cdot r$ where r is a "random" vector in $\mathbb{R}^n$. By "random" we can mean a number of things; we consider the following two possibilities:

<u>Choice 1:</u> r is a unit vector chosen uniformly at random.
<u>Choice 2:</u> $r = (r_1, \ldots, r_n)$ and each r_i is picked independently from $N(0, 1)$.

Let us first consider Choice 1 and let $u' = \phi(u)$. If r is a uniformly chosen n-dimensional random unit vector then by symmetry we have $\mathrm{E}(r_i^2) = 1/n$. Thus

$$\mathrm{E}[\|u'\|^2] = \sum_{i=1}^{n} u_i^2 \,\mathrm{E}[r_i^2] = \frac{\|u\|^2}{n} \quad .$$

The norms are thus scaled down by an average factor of $\sqrt{n}$ but, moreover, we can show that $\|u'\|$ is unlikely to be much larger than its expected value. We want to estimate

$$\Pr\left[\|u'\|^2 > \frac{t^2\|u\|^2}{n}\right] = \Pr\left[\|u'\| > \frac{t\|u\|}{\sqrt{n}}\right] \quad .$$

We can assume w.l.o.g. that $\|u\| = 1$ so we are in fact trying to bound the probability that the inner product of a unit vector u and a random unit vector r is greater than $\epsilon = \frac{t}{\sqrt{n}}$. Geometrically speaking (see figure), this probability is the ratio of the outer surface of the two spherical caps to the area of the unit sphere B. In higher dimensions, as we

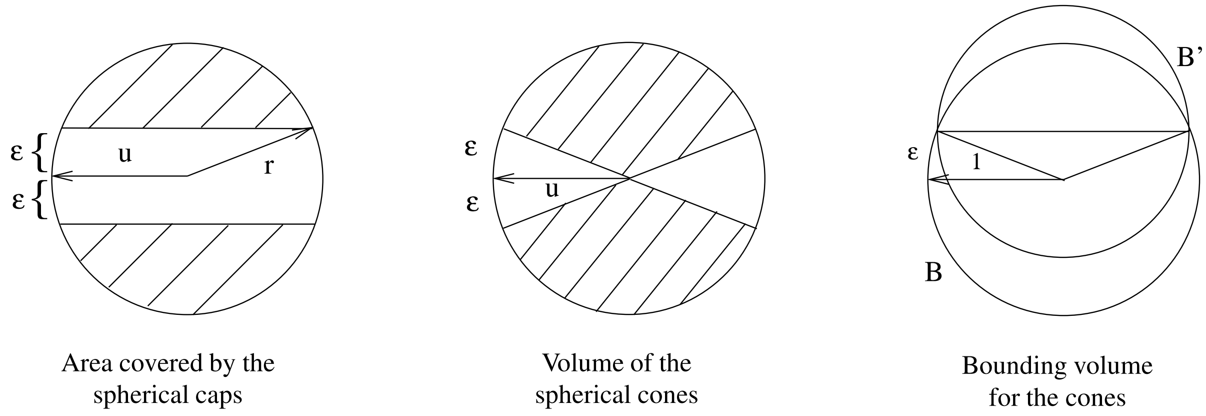

Figure 6: The metric structure of S^n

will now show, this ratio is quite small. We first observe that this ratio is the same as the ratio of the *volume* of the two corresponding spherical cones to the volume of the unit ball (one can, for instance, compare the two integrals). Each of these spherical cones is contained in a ball B' of radius $\sqrt{1-\epsilon^2}$ and so the ratio of the spherical cone's volume to the volume of the unit ball is bounded above by $\frac{\mathrm{Vol}(B')}{\mathrm{Vol}(B)} = (\sqrt{1-\epsilon^2})^n$.

Thus,

$$\Pr\left[\|u'\| > \frac{t\|u\|}{\sqrt{n}}\right] \le 2(1-\epsilon^2)^{n/2} \le 2e^{-\epsilon^2 n/2} \le 2e^{-t^2/2} \ .$$

If X is an m-point metric embedded in $\mathbb{R}^n$, then using the above inequality with $t = c\sqrt{\log m}$, we can find ϕ such that for all $u, v \in X$:

$$\frac{\|\phi(u) - \phi(v)\|}{\|u-v\|} \le c\sqrt{\log m} \ .$$

On the face of it, this is a surprisingly small expansion. But of course this does not mean that we obtain a map of low-distortion because almost all initial distances, both large and small, end up concentrating around their expected value after projection. In fact, one can show:

$$\Pr[\|\phi(u)\| < \frac{\|u\|}{t\sqrt{n}}] = O(1/t) \ .$$

If we rely instead on Choice 2 to pick a "random" r then, since the r_i's are independent $N(0,1)$, we have $\mathrm{E}(r_i^2) = 1$ and $\mathrm{E}(r_i r_j) = 0$ for any $i \neq j$. Thus:

$$\mathrm{E}\left[\|\phi(u)\|^2\right] = \mathrm{E}\left[\left(\sum u_i r_i\right)^2\right] = \sum \mathrm{E}\left[u_i^2 r_i^2\right] = \sum u_i^2 = \|u\|^2 \ .$$

The $\sqrt{n}$ scaling factor of the norms disappears but, as we will make clear in Section 11, it is still true that $\|u'\|$ is unlikely to be much larger than its expected value.

It is not surprising that projections into a 1-dimensional space will generally yield high distortion embeddings. If we start from a clique metric and project it on a line, some distances may be short but unavoidably some will be large and the resulting map will have distortion at least $m-1$. Thus, to look for low-distortion embeddings we have to project to higher dimensions.

10 Low-dimensional ℓ_1-embeddings

We are looking for low dimensional embeddings of ℓ_2-metrics into ℓ_2. In general, we could consider the same question for ℓ_p-embeddings. This question, however is easy to solve for ℓ_1.

Theorem 5. *Any ℓ_2-metric embeds isometrically into ℓ_1^∞.*

Proof. The idea is to embed the ℓ_2-metric into an infinite dimensional space where each dimension corresponds to a unit vector r of $\mathbb{R}^n$. The r-coordinate of $\phi(u)$ is then $u \cdot r$.

Clearly

$$d_2(u,v) = \|u-v\|_2 = \max_r |(u-v)\cdot r| = \max_r |u \cdot r - v \cdot r| = d_1(u,v) \ .$$

□

For finite ℓ_2-metrics, the above isometry is rather unsatisfactory since we are going from $m-1$ dimensions to infinite dimensions. But the construction can be fine-tuned.

Definition 9. *A* cut-metric *on the set X is a metric such that $X = S_1 \dot\cup S_2$ and any $d(x,y) = 0$ if x and y lie in the same S_i and $d(x,y) = 1$ otherwise.*

Lemma 4. *Every m-point ℓ_1-metric X can be realized as a convex combination of cut-metrics X_i.*

We should note that Santosh left this proof as an exercise. The following proof is provided by Ricard and Pascal for completeness.

Proof. We will show first that X can be realized as a convex combination of cut metrics, then argue that $\binom{m}{2}$ of them suffice.

Let $X = \{x_1, x_2, \ldots, x_m\} \subseteq R^n$. Since

$$\|x_i - x_j\|_1 = |x_i[1] - x_j[1]| + \cdots + |x_i[n] - x_j[n]| \ , \tag{1}$$

it suffices to show that we can express each value $|x_i[k] - x_j[k]|$ as a convex combination of cut metrics.

Fix k, $1 \le k \le n$, and assume w.l.o.g. that $x_1[k] \le x_2[k] \le \cdots \le x_m[k]$. Define cut-metrics $S_{k,1}, S_{k,2}, \ldots, S_{k,m}$ by $S_{k,\ell} = \{\ x_i \mid x_i[k] \le x_\ell[k]\ \}$. Define $\alpha_{k,\ell} = x_\ell[k] - x_{\ell-1}[k]$; note that $\alpha_{k,\ell} \ge 0$. It is then clear that for all i,j

$$|x_i[k] - x_j[k]| = \sum_\ell \alpha_{k,\ell} \cdot d_{S_{k,\ell}}(x_i, x_j) \tag{2}$$

where $d_{S_{k,\ell}}(u,v)$ is the cut metric induced by $S_{k,\ell}$.

We have used at most mn cut metrics, and in any case there are no more than 2^{m-1}. In fact, there are only $\binom{m}{2}$ constraints to satisfy and thus one can further show that no more than $\binom{m}{2}$ cut metrics are required. □

Corollary 1. *If X is a finite ℓ_1-metric with $|X| = m$ then X is of dimension at most $\binom{m}{2}$.*

11 From n to k dimensions

We follow the same basic strategy and hope to construct a low-distortion embedding into $\mathbb{R}^k$ by projecting $\mathbb{R}^n$ to a randomly selected k-dimensional subspace. Again, there are many ways in which we could choose a "random" k-dimensional subspace. We can opt to pick a uniformly chosen orthonormal matrix R, i.e. a random k-dimensional basis and define $u' = \sqrt{\frac{n}{k}} R^T u$. (Note that the $\sqrt{\frac{n}{k}}$ is just a scaling factor chosen so that $\mathrm{E}[\|u'\|] = \|u\|$.)

For practical (computational) purposes, sampling the uniform distribution on orthonormal matrices might be tricky and the following alternative is more tempting: choose the matrix R with each entry $R_{ij} \sim N(0,1)$ independent. Then letting $u' = \frac{1}{\sqrt{k}} R^T u$ one can show $\mathrm{E}(\|u'\|) = \|u\|$ from the same computation used at the end of Section 9.

Let u be a unit vector in $\mathbb{R}^n$ and X be the random variable $\|u'\|^2 = \frac{1}{k}\sum_{j=1}^k (R_j \cdot u)^2$. If we define $X_j = R_j \cdot u$ then $X_j = R_j \cdot u = \sum_{i=1}^n R_{ij} u_i$ is normally distributed (because it is the sum of independent normals) and has mean 0. Its variance is the sum of variances $\mathrm{Var}[R_{ij} u_i]^2 = u_i^2$ and that sum is 1 since u is a unit vector.

Hence, X is the sum of squares of k i.i.d. $N(0,1)$ squared, i.e. X has the Chi-square distribution with k degrees of freedom. The expected value of X is 1 and the distribution is tightly concentrated around this value. Specifically, we want to show

Theorem 6. *Let R be $n \times k$ with $R_{ij} \sim N(0,1)$. If $u' = \frac{1}{\sqrt{k}} R^T u$ then for any $\epsilon > 0$*

$$\Pr[|\|u'\|^2 - \|u\|^2| > \epsilon \|u\|^2] \leq 2e^{-(\epsilon^2 - \epsilon^3)k/4} \quad .$$

Corollary 2 (Johnson-Lindenstrauss Flattening Lemma). *Let $S \subseteq \mathbb{R}^n$ with $|S| = m$, let R be as above and for any $0 < \epsilon < 1$ let $k = \frac{8\ln(2m)}{\epsilon^2 - \epsilon^3}$. Then with probability at least $1/2$, we have for all $u, v \in S$*

$$(1-\epsilon)\|u - v\| \leq \|u' - v'\| \leq (1+\epsilon)\|u - v\| \quad .$$

Proof of the Corollary. There are $\binom{m}{2}$ pairs of points u, v in S and by Theorem 6 the probability that there exists some pair whose distance is not well preserved is bounded by

$$\binom{m}{2} \cdot 2e^{-2\ln(2m)} < 1/2 \quad .$$

□

Note that in contrast to Dvoretsky's Theorem (see §5) the dimension of the space into which we obtain a low-distortion embedding depends on the number of points of the original metric rather than its dimension.

Proof of Theorem 6. For the purposes of the proof, let $X = k\|u'\|$ so that $\mathrm{E}(X) = k$. We need to show that

$$\Pr[X > (1+\epsilon)\,\mathrm{E}(X)] \leq e^{-(\epsilon^2-\epsilon^3)k/4} \quad .$$

Using Markov's inequality, we obtain

$$\Pr[X > (1+\epsilon)\,\mathrm{E}(X)] = \Pr[e^{\lambda X} > e^{\lambda(1+\epsilon)k}] \leq \frac{\mathrm{E}[e^{\lambda X}]}{e^{\lambda(1+\epsilon)k}}$$

for any λ.

Recall that $X = \sum X_j^2$ for i.i.d. $X_j's$, so

$$\begin{aligned}
\Pr[X > (1+\epsilon)\,\mathrm{E}(X)) &\leq \frac{\mathrm{E}[e^{\lambda \sum_j X_j^2}]}{e^{\lambda(1+\epsilon)k}} \\
&= \frac{\prod_j \mathrm{E}[e^{\lambda X_j^2}]}{e^{\lambda(1+\epsilon)k}} = \left[\frac{\mathrm{E}[e^{\lambda X_1^2}]}{e^{\lambda(1+\epsilon)}}\right]^k \quad .
\end{aligned}$$

The density function of a normal distribution $N(0,\sigma)$ is $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$ and thus, since $X_1 \sim N(0,1)$:

$$\begin{aligned}
\mathrm{E}(e^{\lambda X_1^2}) &= \int_{-\infty}^{\infty} e^{\lambda x^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2\lambda)\frac{x^2}{2}}\, dx \\
&= \frac{1}{\sqrt{1-2\lambda}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2\lambda}}{\sqrt{2\pi}} e^{-(1-2\lambda)\frac{x^2}{2}}\, dx \\
&= \frac{1}{\sqrt{1-2\lambda}}
\end{aligned}$$

since the last integral is the density function for $N(0, \sqrt{1-2\lambda})$.

The above equality holds for any $\lambda < 1/2$ and hence

$$\Pr[e^{\lambda X} > e^{\lambda(1+\epsilon)k}] \leq \left(\frac{e^{-2(1+\epsilon)\lambda}}{1-2\lambda}\right)^{k/2} \quad .$$

Elementary calculus shows that the right-hand expression is minimized for $\lambda = \frac{\epsilon}{2(1+\epsilon)}$ which yields

$$\Pr[X > (1+\epsilon)\,\mathrm{E}(X)] \leq ((1+\epsilon)e^{-\epsilon})^{k/2} \leq e^{-(\epsilon^2-\epsilon^3)^{k/4}} \quad .$$

where the last inequality holds since $1+\epsilon \leq e^{\epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3}}$. □

The Johnson-Lindenstrauss result shows that a low-dimensional embedding is obtained with high probability by picking a matrix whose entries are i.i.d. normal distributions, but for practical purposes one would like to obtain even simpler classes of random matrices R with the same property.

For instance, suppose each entry in R is chosen independently and uniformly from $\{-1, 1\}$. Then we can easily show that $\mathrm{E}(\|u'\|) = \|u\|$ and in fact $\|u'\|$ is also tightly concentrated around $\|u\|$. Indeed, as above, let $X_j = (R_j \cdot u)$ and $X = \|u'\|^2 = \sum_j (R_j \cdot u)^2 = \sum_j X_j^2$. Then, as before,

$$\Pr[X > (1+\epsilon)\,\mathrm{E}(X)) \leq \left[\frac{\mathrm{E}[e^{\lambda X_1^2}]}{e^{\lambda(1+\epsilon)k}}\right]^k \quad .$$

To bound $\mathrm{E}(e^{\lambda X_1^2})$, we first write the Taylor expansion

$$\mathrm{E}(e^{\lambda X_1^2}) = 1 + \lambda\,\mathrm{E}(X_1^2) + \frac{\lambda^2}{2}\,\mathrm{E}(X_1^4) + \ldots$$

and try to find some random variable Y such that $\mathrm{E}(Y^{2m}) \geq \mathrm{E}(X_1^{2m})$ for all m. For any such Y we clearly will have $\mathrm{E}(e^{\lambda X_1^2}) \leq \mathrm{E}(e^{\lambda Y^2})$.

In our case, $X_1 = \sum R_{i1} u_i$ where the R_{i1} are taken independently and uniformly from $\{-1, 1\}$. If we set $Y = R_i' u_i$ where the R_i' are i.i.d. $\sim N(0,1)$ then we have $\mathrm{E}[(R_{i1})^{2m}] = 1$ is smaller than $\mathrm{E}[(R_i')^{2m}] = \frac{(2m)!}{2^m\, m!}$ using the formula for moments of the normal distribution.

Expand $\mathrm{E}(X_1^{2n}) = \mathrm{E}\left[\left(\sum R_{i1} u_i\right)^{2n}\right]$ as a polynomial in the R_{j1}'s and u_{ij}'s. If a term of the polynomial contains an odd power of some R_{j1}, its expectation is 0 because the distribution of R_{j1} is centered around 0 and the other R_{i1}'s are independent of it. If a term contains only even powers of R_{j1}s, then its expectation is bounded above by that of the same term using R_{j1}'s. Consequently, the expectation of every term in this expansion is bounded above by the corresponding term in the expansion of $\mathrm{E}(Y^{2n})$ and so $\mathrm{E}(Y^{2n}) \geq \mathrm{E}(X_1^{2n})$.

We can now reuse the concentration result obtained in our proof of Theorem 6 to get:

$$\begin{aligned}
\Pr[X > (1+\epsilon)\,\mathrm{E}(X)) &\leq \left[\frac{\mathrm{E}[e^{\lambda X_1^2}]}{e^{\lambda(1+\epsilon)k}}\right]^k \\
&\leq \left[\frac{\mathrm{E}[e^{\lambda Y^2}]}{e^{\lambda(1+\epsilon)k}}\right]^k \\
&\leq e^{-(\epsilon^2-\epsilon^3)k/4} \quad .
\end{aligned}$$

Will this work for any distribution? In fact, not only the method, but the whole consideration fails. For instance, consider fat-tailed distributions such as

$$F_i = \begin{cases} -\sqrt{L} & \text{with probability } \frac{1}{2L}, \\ \sqrt{L} & \text{with probability } \frac{1}{2L}, \\ 0 & \text{with probability } 1 - \frac{1}{L}. \end{cases}$$

Then $\mathrm{E}(F_i^2) = 1$ but $\mathrm{E}(F_i^4) = L^3$, which is larger than the fourth moment of the normal distribution appearing in the proof. For such large L, the method fails because $\mathrm{E}[e^{LX^2}]$ is undefined, and in fact one can show that if R_{ij}s are picked according to F we may end up with high-distortion embeddings.

In general, we do not know which are the distributions such that $\|u'\|$ is strongly concentrated around its expected value.

Title: Girth of graphs — Lecture #: 4
Lecturer: Nati Linial — Scribe: Michal Koucký and Andreas Krebs

12 Girth and finite metric spaces

Summary. In previous lectures we have seen an application of graphs of large girth for proving a lower bound on embeddability of metric spaces into ℓ_2. In this lecture we will study this parameter closely.

Definition 10. *Let G be an undirected graph. The* girth of G, $\textsc{Girth}(G)$*, is the length of the shortest cycle in G.*

In previous lectures we have shown Matoušek's proof that uses metrics induced by graphs of large girth to prove lower bounds on the dimension of a host metric space. The idea of using graphs of large girth for proving lower bounds originates with Bourgain who used them to show that there are graphs that require large distortion when embedded into ℓ_2. There are many open questions regarding embeddability of these graphs.[2] Linial, London and Rabinovich make the following conjecture:

Conjecture 1 ([20]). *Let G be any graph with all vertices of degree ≥ 3. Then every embedding of G into ℓ_2 (regardless of dimension) has distortion $c_2(G) \geq \Omega(\textsc{Girth}(G))$.*

A stronger version of this conjecture asserts that the conclusion is true even for graphs with *average* degree at least $2+\delta$, for any constant $\delta > 0$. The requirement on the degree of vertices comes from the fact that one could replace every edge in a graph by a path thus increasing girth substantially without changing distortion. Linial, Magen and Naor prove the following weaker theorem.

Theorem 7 (Linial, Magen, and Naor [21]). *Let G be a d-regular graph, where $d \geq 3$. Then every embedding of G into ℓ_2 (regardless of dimension) has distortion $c_2(G) \geq \Omega(\sqrt{\textsc{Girth}(G)})$.*

It is not hard to show that the distortion for embeddings of graphs into ℓ_2 goes to infinity as their girth goes to infinity. It is open whether the same is true for embeddings into ℓ_1.

In the rest of the lecture we will focus on the following question: How large can the girth of a d-regular graph on n vertices be?

The following is known:

Theorem 8. *Let $g(n,d)$ denote the largest girth that a d-regular graph on n vertices can have. Then*

$$2 \cdot \frac{\log n}{\log(d-1)} + 2 \geq g(n,d) \geq \frac{4}{3} \cdot \frac{\log n}{\log(d-1)} \ .$$

The upper bound in this theorem is called Moore's bound. The lower bound follows from the construction of expanders given by Lubotzky, Phillips and Sarnak [22]. We will show the upper bound and also a somewhat weaker lower bound with the constant slightly below 1 instead of $4/3$. It is a challenge to present a simple proof for a lower bound with constant strictly larger than 1.

To prove the upper bound, consider any graph G of girth g on n vertices. Pick a vertex v in G and consider all vertices up to distance $\lfloor \frac{g-1}{2} \rfloor$ from v. Clearly, they all must be distinct because the girth of G is g. Hence, they form a tree, and $(d-1)^{\lfloor \frac{g-1}{2} \rfloor} \leq n$. The upper bound follows.

Before proving the lower bound we will digress to discuss Moore graphs.

[2] J. Matoušek maintains a list of open problems concerning embeddings at his web-page `http://kam.mff.cuni.cz/~matousek/`.

12.1 Moore graphs

From the argument for Moore's bound we know that a d-regular graph of girth g has at least $M(d,g)$ vertices, where $M(d,g)$ is given by:

$$M(d,g) = \begin{cases} 1 + d\sum_{r=0}^{(g-3)/2}(d-1)^r & \text{if } g \text{ is odd,} \\ 1 + d\sum_{r=0}^{(g-4)/2}(d-1)^r + (d-1)^{(g-2)/2} & \text{otherwise.} \end{cases}$$

A d-regular graph of girth g that has exactly $M(d,g)$ vertices is called a *Moore graph*. There is a striking similarity between Moore graphs and perfect codes: Moore graphs exist only for a small number of pairs (d,g). A cycle of length g is a Moore graph, for $g = 3$; every complete graph K_n is a Moore graph, for $g = 4$; every complete bipartite graph $K_{n,n}$ is a Moore graph; the remaining Moore graphs are described by the following theorem:

Theorem 9 ([16, 3, 8, 5, 13]). *For $d \geq 3$ and $g \geq 5$, d-regular Moore graphs of girth g exist only if:*

1. *$g = 5$ and $d = 3, 7$ or possibly 57, or*
2. *$g = 6, 8$ or 12.*

The 3-regular Moore graph of girth 5 is Petersen graph (Fig. 7) and the 7-regular Moore graph is Singleton-Hoffman graph.

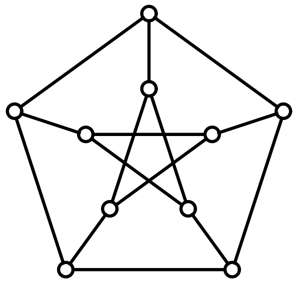

Figure 7: The Petersen graph

Proof. We prove the first part of Theorem 9. Let $g = 5$, $d \geq 3$ be an integer and let G be a d-regular Moore graph of girth g. Hence, G has $n = M(d,g) = d^2 + 1$ vertices. Let A be the adjacency matrix of G. Since the graph is d-regular on $d^2 + 1$ vertices with girth 5, the maximum distance between any two vertices is at most two. Thus, one can easily verify that $A^2 = (d-1)I + J - A$, where I is the identity matrix and J is the all one matrix. Clearly, $A\mathbf{1} = d\mathbf{1}$. Let $\lambda \neq d$ be another eigenvalue of A. Clearly, for any eigenvector u such that $Au = \lambda u$, we have that the inner product $\langle u, \mathbf{1}\rangle = \mathbf{0}$. Thus, $A^2u = (d-1)Iu + Ju - Au$, and $\lambda^2 u = (d-1)u + \lambda u$. This implies that $\lambda^2 = d - 1 - \lambda$ and hence all eigenvalues of A except d must be solutions of this quadratic equation. Call these eigenvalues λ_1 and λ_2, with multiplicities k and $n-k-1$, resp. (d has multiplicity one.) By consider the trace of A, we can conclude that $d + k\lambda_1 + (n-k-1)\lambda_2 = 0$. By solving for k we get that:

$$k = \frac{1}{2}\left(d^2 + \frac{2d - d^2}{\sqrt{1 + 4(d-1)}}\right) .$$

We observe that both

$$\frac{2d - d^2}{\sqrt{(1 + 4(d-1)}} \qquad \text{and} \qquad 256(d^2 - 2d)^2/(1 + 4(d-1))$$

must be integers, but that

$$256(d^2 - 2d)^2 = 256d^4 - 1024d^3 + 1024d^2 = (64d^3 - 208d^2 + 100d + 75)(4d - 3) + 225 .$$

Hence $1 + 4(d-1) = 4d - 3$ divides 225. Since d is an integer there are only the possibilities $1, 2, 3, 7, 12, 57$. By computing the corresponding values for k, we see that the only possible solutions for $d = 2, 3, 7, 57$. $\square$

A natural question is whether the Moore upper bound holds also for non-regular graph where d is replaced by the average degree. Alon, Hoory, and Linial [2] give a positive answer to this question.

12.2 Construction of graphs of large girth

In the rest of the lecture we will give an answer to the question whether one can construct d-regular graphs of an arbitrary girth.

The idea is to start with an arbitrary d-regular graph and apply random lifts to the graph to increase its girth.

Definition 11. *A graph H is a* cover *of the graph G if there exists a surjective graph homomorphism $\phi : H \to G$ (called* covering map*) that is one-to-one mapping of $N_H(v)$ to $N_G(\phi(v))$, for every vertex v of H. Here $N_H(v)$ is the neighborhood of the vertex v, that is $N_H(v) = \{v\} \cup \{v' \mid (v, v')$ is an edge in $H\}$.*

One can easily see that if H is connected then the covering map of G maps equal number of vertices of H to every vertex of G. This number is called the *covering number*. H is a *k-lift of G* if H is a cover of G of the covering number k. Notice that any lift of a d-regular graph is d-regular.

Our algorithm to increase the girth will be to start with a d-regular graph G and successively pick random 2-lifts of it.

Suppose that $\text{GIRTH}(G) = g$. Let N be the number of g-cycles (cycles of length g) in G, and let X be the number of g-cycles in a random 2-lift H of G. We observe that $\text{E}(X) \leq N$. (For every g-cycle of G, with probability $1/2$, two disjoint g-cycles of H are mapped to it and with the remaining probability, one $2g$-cycle of H is mapped to it. No cycle shorter than g can be mapped to a g-cycle. The observation follows.) If we remove the trivial 2-lift that has $2N$ g-cycles from the support of X we get $E(X) < N$. Hence there is always a 2-lift that reduces the number of g-cycles. Take this 2-lift and repeat lifting until no g-cycle in G is left.

Using the previous procedure again and again, we can construct d-regular graphs with arbitrarily high girth. Note, however, that the number of vertices appearing in these graph is enormous: the next step will be to reduce the number of vertices until we have $g + 1 \geq \frac{\log n}{\log(d-1)}$.

We start with a d-regular graph G with $\text{GIRTH}(G) = g$. Then we reduce the number of vertices, but not the girth, until $g + 1$ is at least diameter $\text{DIAM}(G)$ of G. Together with the fact that $\text{DIAM}(G) \geq \frac{\log n}{\log(d-1)}$ we shall achieve the desired result.

Suppose that $g < \text{DIAM}(G) - 1$. Then there are vertices u, v at distance $g + 1$.

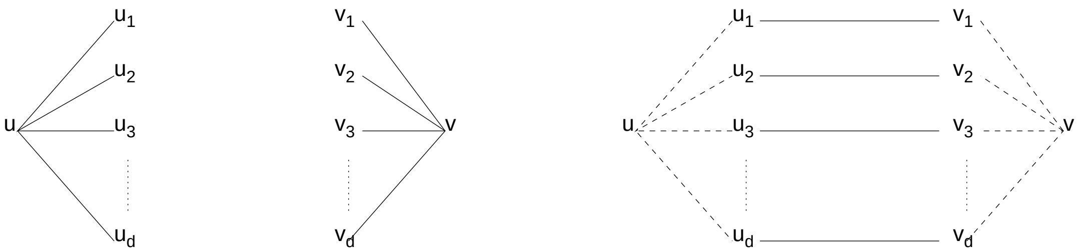

Figure 8: The vertex reduction step

We remove the vertices u, v and add edges (u_i, v_i) for all i. We can see that the modified graph is still d-regular. Since $d(u_i, v_i) \geq d(u, v) - 2 = g - 1$, we did not create any $(g-1)$-cycles by adding the edges (u_i, v_i). So the girth is still g and we repeat the procedure.

The conclusion is that we can construct graphs of arbitrary high girth so that the number of vertices is bounded by $g + 1 \geq \frac{\log n}{\log(d-1)}$.

13 Convex and semidefinite programming

There are no deep theorems in linear programming — just duality for *convex* programming, including semidefinite programming.

In general, a convex programming problem is phrased as follows. We are given a convex set P (which is often, but not necessarily, compact; for linear programming, P is a polytope) and a linear *objective function* f. We want to find the $x \in P$ that maximizes $f(x)$.

It's hard to find good books on this. Drawing pictures in the plane is too simple, and often gives the wrong idea — one should try to think about 3 or 4 dimensions at least!

A convex set P can be viewed in two useful ways:

1. As the convex hull of its *extremal points*, i.e., those points in P which are not convex combinations of two other points in P (this assumes compactness — for a non-compact set we consider its extremal rays, see below).

2. As the intersection of a collection of half-spaces, i.e., the conjunction of a set of linear inequalities.

Going from (1) to (2) and back again can be interesting and non-trivial.

The Duality Theorem can be stated, informally, as follows: any valid conclusion that can be derived from a system of linear equations and inequalities can be derived by taking admissible linear combinations of these (i.e., with nonnegative coefficients multiplying the inequalities).

To be more precise, we want the smallest possible C such that $x \in P$ implies $f(x) \leq C$. Since any such inequality can be derived by taking linear combinations of the inequalities that define P, this gives us another optimization problem: namely, optimize the coefficients in front of these inequalities to minimize C. Thus given P and f, we can construct another Q and g such that

$$\max_{x \in P} f(x) = \min_{y \in Q} g(y) = C \ .$$

Here Q is the set of admissible coefficients, and $g(y)$ is the upper bound on C given by combining P's inequalities with coefficients y. Note that

$$\forall x' \in P, y' \in Q \mid f(x') \leq C \leq g(y') \ .$$

For LP duality in particular, we have

$$\max_{\substack{x \geq 0 \\ Ax \leq b}} \langle x, c \rangle = \min_{\substack{y \geq 0 \\ yA \geq c}} \langle y, b \rangle = C \ .$$

The ellipsoid algorithm for linear programming can be generalized to a wide variety of convex programming problems. Consider the following two types of questions:

- *Membership*: given x, is $x \in P$?

- *Separation*: if $x \notin P$, give a separating hyperplane, i.e., a vector w and a number b s.t. $\langle w, x \rangle > b$ but $\langle w, x' \rangle < b$ for all $x' \in P$.

Then the ellipsoid algorithm works efficiently for any class of P for which we can answer these questions efficiently [GLS]. For linear programming these questions are trivial, since the number of inequalities is only polynomial. For semidefinite programming the number of inequalities is infinite, but luckily we can still answer them efficiently.

Semidefinite programming (SDP) is defined in the following way. P consists of the intersection of PSD_n with a finite collection of linear inequalities, where PSD_n is the cone of positive symmetric semidefinite $n \times n$ matrices, i.e., those whose eigenvectors are real and nonnegative. An important characterization of such matrices is the following:

$$A \in \mathrm{PSD}_d \text{ if and only if } A = BB^T \text{ for some nonsingular matrix } B.$$

In other words, A is a matrix of inner products $A_{i,j} = \langle B_i, B_j \rangle$ where B_i is the ith row of B.

14 The cut cone

Given a metric space consisting of n points in ℓ_1^∞, we can define a matrix of distances

$$d_{ij} = |x_i - x_j|_1 \quad .$$

Let $\mathcal{C}_n$ be the collection of matrices that can be realized in this way. $\mathcal{C}_n$ is a *convex cone*, i.e., nonnegative linear combinations of matrices in $\mathcal{C}_n$ are in $\mathcal{C}_n$. To see this, note that we can multiply d by any positive constant by rescaling the space, and given two metric spaces with distances d_1, d_2 in ℓ_1^∞ we can obtain a third metric space with distances $d = d_1 + d_2$ by concatenating coordinates.

$\mathcal{C}_n$ is not compact; however, if we like we can define a compact slice of it by setting $\sum_{i,j} d_{ij} = 1$. Then $\mathcal{C}_n$ consists of matrices in this slice multiplied by nonnegative constants.

Unfortunately, membership and separation for $\mathcal{C}_n$ are NP-complete. The problem is that $\mathcal{C}_n$ has an exponential number of facets, i.e., defining inequalities. One facet is given by the following inequality: given sets of points $\{x_i\}$ and $\{y_j\}$ we have

$$\sum_{i,j} d(x_i, y_j) \geq \sum_{i<j} d(x_i, x_j) + \sum_{i<j} d(y_i, y_j) \quad .$$

Taking five points x_1, x_2, y_1, y_2, y_3, this inequality shows that the graph metric of $K_{2,3}$ is not in ℓ_1.

However, while the set of facets is infeasible, we do know the set of extreme rays: namely, the *cut metrics*. Given a set S, the cut metric d_S associated with S is defined as

$$d_S(x, y) = \begin{cases} 0 & \text{if } x, y \in S \text{ or } x, y \notin S \ , \\ 1 & \text{if } x \in S \text{ and } y \notin S \text{ or vice versa.} \end{cases}$$

Such matrices have a simple block structure,

$$\left(\begin{array}{c|c} 0 & 1 \\ \hline 1 & 0 \end{array} \right) \quad .$$

Proposition 1. *Any matrix $d \in \mathcal{C}_n$ is a convex combination of cut metrics.*

Proof. Since d_{ij} is a sum of contributions from different dimensions, it suffices to show this for one-dimensional metrics. Order $x_1, \ldots, x_n$ in increasing order on the real line, and let $S_i = \{x_1, \ldots, x_i\}$ for $1 \leq i \leq n-1$. Then since for all $j \leq k$

$$|x_j - x_k| = \sum_{i=j}^{k-1} (x_{i+1} - x_i)$$

we have

$$d = \sum_{i=1}^{n-1} (x_{i+1} - x_i) d_{S_i} \quad .$$

□

Now, in any convex programming problem the optimum is always on an extreme ray. This means the ℓ_1^∞ metrics optimizing a linear objective function will be cut metrics—for instance, the solution to the dual of a Min Flow problem is a Max Cut. But which one? Does there exist a "computationally good" close approximation to $\mathcal{C}_n$, leading to $o(\log n)$-approximation algorithms?

One natural candidate is the intersection of "squared ℓ_2," i.e., where $d_{ij} = |x_i - x_j|_2^2$. These also form a cone, since by Pythagoras' theorem we can again add such matrices by concatenating coordinates. Since such things are not necessarily metrics, we take the intersection of this cone with MET, the cone of matrices d_{ij} that obey the triangle inequality. The resulting cone contains the cut metrics, and therefore all of $\mathcal{C}_n$.

By the way, here's a nice way to see that we only need $\binom{n}{2}$ dimensions to perfectly embed any ℓ_1 metric: the space of cut metrics has rank $\binom{n}{2}$.

15 Bourgain's theorem and network flows

Proposition 2 (Bourgain). *Given a finite metric space (X, d) and a set of pairs of points in it $\{(s_i, t_i) \mid i = 1, \ldots, k\}$, there exists a polynomial time algorithm to find a non-expansive embedding ϕ of X into $\ell_1^{O(n^2)}$, so that*

$$||\phi(s_i) - \phi(t_i)||_1 \ \geq \ \Omega\left(\frac{1}{\log k}\right) \cdot d(s_i, t_i)$$

for all $i = 1, \ldots, k$.

([LLR] showed that Bourgain's theorem is tight, e.g. for expander graphs.)

Here we are interested in a network represented by an undirected weighted graph $G = (V, E)$, where the weight function $C : E \to \mathbb{R}^+$ represents the capacity of edges. In general the network has multiple pairs of sources and sinks, where the ith pair is denoted by (s_i, t_i) and $i \in \{1, \ldots, k\}$ (assuming there are k pairs of source and sink). Additionally, each pair i has a *demand* to be satisfied which is denoted by D_i for $i = 1, \ldots, k$. The objective is to find a flow function that maximizes the fraction λ of each demand that can be concurrently satisfied without violating the capacity constraint of any edge. Observe that for $k = 1$ we get the standard version of 1-commodity flow and then the Ford-Fulkerson algorithm of augmenting paths provides an optimal solution. Also, we know that the max-flow in this case is equal to the min-cut of the graph G. In the general case of multi-commodity flow this is not true. Formally, for multi-commodity flow we define the following: Consider any $S \subset V$. Then the capacity of the cut induced by S (denoted by $\textsc{Cap}(S)$) is given by $\sum_{e \in cut(S)} C(e)$. The demand on the cut (denoted by $\textsc{Dem}(S)$) is given by $\sum_{i,j \text{ sep. } S} D_{ij}$. We would call the cost of a cut to be simply the ratio $\frac{\textsc{Cap}(S)}{\textsc{Dem}(S)}$. The minimum cost over all cuts in G is called the *min-cut*. Finding the min-cut turns out to be NP-hard. It is easy to see that the maximum concurrent flow λ is upper-bounded by the min-cut. There could, of course, be a gap between the two. Interestingly, Bourgain's embedding provides us with both an approximation algorithm for finding the min-cut and a bound for the gap between the max-flow and the min-cut for multi-commodity flow.

15.1 The connection to embeddings

There exists a LP formulation for the max flow problem and the optimal λ is just the maximum of this linear program. Using LP duality we can conclude that

$$\lambda = \min \frac{\sum_{(i,j) \in E} C_{i,j} \cdot d_{i,j}}{\sum_{j=1}^{k} D_j \cdot d_{s_j, t_j}}$$

where the minimum is over all graph metrics d on G. Now we make the following observation concerning the ℓ_1^m metric.

Proposition 3. *For any ℓ_1 embedding of the graph metric induced by the weighted graph G,*

$$\frac{\sum_{(i,j) \in E} C_{i,j} \cdot ||x_i - x_j||_1}{\sum_{j=1}^{k} D_j \cdot ||x_{s_j} - x_{t_j}||_1} \geq \textsc{MinCut}$$

Proof. Assuming we are in the ℓ_1^m metric (where m is the dimension of the space), we have

$$\frac{\sum_{i \neq j} C_{i,j} \cdot ||x_i - x_j||_1}{\sum_{j=1}^{k} D_j \cdot ||x_{s_j} - x_{t_j}||_1} = \frac{\sum_{r=1}^{m} \sum_{i \neq j} C_{i,j} \cdot |x_{i,r} - x_{j,r}|}{\sum_{r=1}^{m} \sum_{j=1}^{k} D_j \cdot |x_{s_{j,r}} - x_{t_{j,r}}|} \geq \min_{1 \leq r \leq m} \frac{\sum_{i \neq j} C_{i,j} \cdot |x_{i,r} - x_{j,r}|}{\sum_{j=1}^{k} D_j \cdot |x_{s_{j,r}} - x_{t_{j,r}}|}$$

Let the minimum ratio occur for the dimension indexed by l. It can also be proved using a variational argument[3] that the minimum is not increased by restricting the domain of values for all $x_{i,l}$ to $\{0, 1\}$. But then

$$\frac{\sum_{i \neq j} C_{i,j} \cdot |x_{i,l} - x_{j,l}|}{\sum_{j=1}^{k} D_j \cdot |x_{s_{j,l}} - x_{t_{j,l}}|} = \frac{\textsc{Cap}(S)}{\textsc{Dem}(S)}$$

where $S = \{i \mid x_{i,l} = 1\}$ is the cut inducing set. $\square$

The following steps now provide a deterministic polynomial time algorithm for obtaining an approximation of the min-cut in a graph representing a network with a demand for multiple commodities:

1. Solve the LP corresponding to the dual of max-flow. This gives a metric $d(i, j)$.
2. Use the algorithm from Proposition 2 to get a non-expansive embedding of the metric from the earlier step, where distortion is bounded by $\Omega(1/\log k)$ for demand pairs i.e. $||\psi(s_i, t_j)||_1 \geq C\frac{d(s_i,t_j)}{\log k}$, where ψ is the embedding.
3. For each co-ordinate r
   - Order the vertices according to the value of $\psi_r(i)$.
   - Consider all the $n - 1$ partitions along this order. Each partition gives a cut.
4. Among all these cuts, output the one with minimum cost.

Now we can apply a consequence of Bourgain's embedding as given in Proposition 2 and Proposition 3 to see that step 2 of the above algorithm gives

$$\textsc{MinCut} \leq \frac{\sum_{(i,j) \in E} C_{i,j} \cdot ||x_i - x_j||_1}{\sum_{j=1}^{k} D_j \cdot ||x_{s_j} - x_{t_j}||_1} = O(\log k)\lambda$$

where λ is the max-flow calculated by the LP in step 1 of the algorithm. This provides us the following theorem about the gap between the max-flow and min-cut in a muticommodity flow network:

Theorem 10. *In a k-commodity flow problem* $O(\log k)$ $\textsc{MaxFlow} \geq \textsc{MinCut} \geq \textsc{MaxFlow}$.

[3] See the paper by Linial, London and Rabinovich [20] for this argument.

16 Ramifications of Bourgain's Theorem

The focus of this lecture is on the proof and the consequences of Bourgain's theorem [6]:

Theorem 11 (Bourgain). *Every n-point metric space can be embedded in ℓ_1 (even in ℓ_2) with distortion $O(\log n)$.*

In a previous lecture, the k-COMMODITY MAX FLOW problem was defined. Recall that $\textsc{MaxFlow}(G)$ is the largest λ for which it is possible to simultaneously flow λD_i between the ith source and the ith sink for each i, where D_i is the demand associated with the ith source-sink pair. Then $\textsc{MinCut}(G)$ was defined as the minimum, over all nontrivial sets S of nodes, of $\frac{\textsc{Cap}(S,S^c)}{\textsc{Dem}(S,S^c)}$, where capacities are summed over all edges across the cut and demands are summed over all source-sink pairs separated by the cut. Bourgain's theorem can be used in surprising ways to show properties of graphs, for example [20]:

Theorem 12 (London, Linial, and Rabinovich). *For the k-COMMODITY MAX FLOW in a capacity network G,*

$$\textsc{MaxFlow}(G) \le \textsc{MinCut}(G) \le O(\log k) \cdot \textsc{MaxFlow}(G) \ .$$

The leftmost inequality in Theorem 12 is easy since a cut cannot support more total flow than the sum of the λD_i over all source-sink pairs separated by the cut. The rightmost inequality in Theorem 12 can be shown by Bourgain's theorem and we do not prove it here. We call the $O(\log k)$ factor in Theorem 12 *Bourgain's estimate*.

London, Linial, and Rabinovich [20] showed that the bound in Bourgain's Theorem is tight for ℓ_2 and ℓ_1 by exhibiting graphs that require $\Omega(\log n)$ distortion. It is possible, however, to get much better bounds for restricted classes of metrics. Note that in Bourgain's theorem, the ℓ_2 upper bound is stronger while in the tightness argument to be presented next the ℓ_1 lower bound is stronger, for the following reason. A theorem of Dvoretsky [9] shows that any n-dimensional normed space must contain an $d = \Omega(\varepsilon^2 \log n)$-dimensional subspace S into which we can embed ℓ_2^d with distortion $1+\epsilon$. Thus we can embed n points into ℓ_p by first using Bourgain to embed into ℓ_2^n and then embedding ℓ_2^n into the appropriate subspace of ℓ_p.

In the next section, will sketch a proof of the case of ℓ_1 in Bourgain's theorem. Before we do, we follow LLR and show that the bound in Theorem 11 is tight. We will do this by using expander graphs to set up a k-COMMODITY MAX FLOW in which the mincut is at least the Bourgain estimate times the maxflow. This will prove that Bourgain's theorem cannot be improved (since any improvement would improve the mincut upper bound in Theorem 12).

For a given graph, let $e(S,T)$ be the number of edges connecting S and T for a cut (S,T). We say that a graph is an ε-edge expander if $\forall S \subseteq V$ such that $|S| \le \frac{n}{2}$, we have $e(S,S^c) \ge \varepsilon|S|$.

For an arbitrary graph G, we define a capacity network by assigning one commodity to each pair of vertices in G, setting the demand for each commodity to one and setting the capacity for each edge to one. Now pick a d-regular expander graph G, $d \ge 3$, and write λ for the maxflow of its associated capacity network. The following chain proves that Bourgain's theorem is optimal, where all the steps in the chain except the rightmost inequality are clear:

$$\textsc{MinCut}(G) = \min_{|S| \le n/2} \left\{ \frac{\textsc{Cap}(S,S^c)}{\textsc{Dem}(S,S^c)} \right\} \ge \frac{\varepsilon|S|}{|S|(n-|S|)} \ge \frac{\varepsilon}{n} \ge \lambda \cdot \Omega(\log n) \ .$$

To see the rightmost inequality, let E be the set of $O(n)$ edges of the d-regular graph. Now observe that in any regular graph of degree $d \ge 3$, the average distance between pairs of vertices is $\Omega(\log n)$. (This is because there are at most $d(d-1)^{t-1}$ nodes at distance t from any fixed node u; hence more than half of the nodes are at distance at least $\log_{d-1} n$ from u.) Thus we have:

$$
\begin{aligned}
|E| = \Theta(n) \quad &\geq \quad \sum_e \{\text{flow through } e\} \\
&= \quad \sum_{i,j} \sum_e \{\text{flow of commodity } (x_i, x_j) \text{ through } e\} \\
&\geq \quad \lambda \sum_{i,j} \text{dist}(x_i, x_j) \\
&\geq \quad \Omega(\lambda n^2 \log n) \Longrightarrow \lambda \cdot \log n = O\left(\frac{1}{n}\right)
\end{aligned}
$$

17 Bourgain's embedding

We now sketch a proof for the ℓ_1 case of Theorem 11. We will be showing a randomized construction based on Bourgain's construction that requires $O(\log^2 n)$ dimensions. Let $X = \{p_1, \ldots, p_n\}$ be the points in the metric being considered. Recall that in Fréchet's Theorem, we map each $x \in X$ to a vector with coordinates of the form $d(x, p_j)$. In the new embedding, the coordinates will correspond to the distances from x to *sets* of points.

For $A \subseteq X$ we define $d(x, A) = \min\{d(x, y) \mid y \in A\}$. So every coordinate in the Bourgain embedding is of the form $d(x, A)$. Note that $|d(x, A) - d(y, A)| \leq d(x, y)$ by the triangle inequality, so every such map from x to $d(x, A)$ is nonexpansive. (To see this, let $u, v \in A$ be points such that $d(x, A) = d(x, u)$ and $d(y, A) = d(y, v)$. Then $d(x, A) \leq d(x, v) \leq d(x, y) + d(y, v)$.)

Since we are dealing with nonexpansive maps, it remains to guarantee that distances do not shrink too much, in particular that $|d(x, A) - d(y, A)|$ is large on average over all sets A that we choose. We will show this for any two generic points x and y. The strategy is to choose a variety of different sets of different sizes to correspond to the coordinates. Our choice will be to select

$$
\begin{array}{ll}
\log n \text{ sets of density} & \frac{1}{2}, \\
\log n \text{ sets of density} & \frac{1}{4}, \\
\vdots & \vdots \\
\log n \text{ sets of density} & \frac{1}{2^j}.
\end{array}
$$

We will do this all the way up to some $j = O(\log n)$. To choose a set of density d we independently choose each x to be in the set with probability d.

Recall the birthday paradox:

> Let $A, B \subseteq [n]$ with $|A| = a$ and $|B| = b$ be chosen uniformly at random. If $ab \sim n$, then:
>
> $$\varepsilon < \Pr[A \cap B \neq \emptyset] < 1 - \varepsilon.$$

How does this help? Fix an x and a y.

We are happy when $|d(x, A) - d(y, A)|$ is large, i.e. a fixed fraction of $d(x, y)$ (we cannot hope for better as the map is nonexpansive). If we can find r and R such that $d(x, A) \leq r$ and $d(y, A) \geq R$ as in the picture, then obviously $|d(x, A) - d(y, A)| \geq |R - r|$. Let us define specific r, R for which $|R - r|$ is reasonably large with constant probability. In fact, $|R - r|$ individually for a specific A will not be large, but the sum over $O(\log n)$ sets A, one set A of each density, will be large.

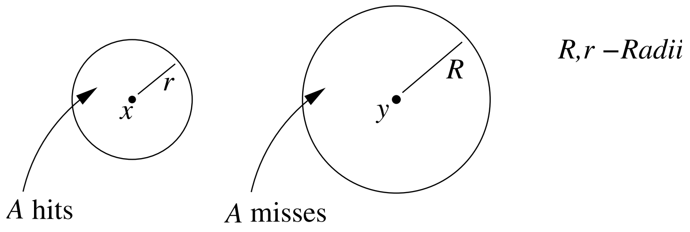

Figure 9: The distance-bounding event

Proposition 4. *Let $u, v \in X$ with $u \neq v$. Then there exist $\Delta_1, \ldots, \Delta_q \geq 0$ with $q \leq \lfloor \log_2 n \rfloor + 1$ and $\sum \Delta_i = d(u,v)/2$ such that $\forall i \in \{1, \ldots, q\}$ and $\forall A \subseteq X$ randomly chosen with density $\left(\frac{1}{2}\right)^i$,*

$$\Pr[|d(u,A) - d(v,A)| \geq \Delta_i] > \frac{1}{12} \ .$$

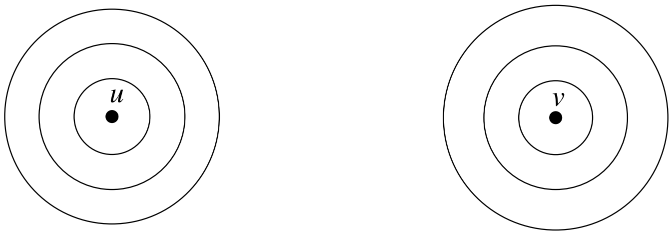

Figure 10: The concentric rings of Bourgain's theorem

Proof. Define $B(x,r) = \{y \mid d(x,y) \leq r\}$, $B^0 = \{y \mid d(x,y) < r\}$, and $r_i := \min_r \{|B(u,r)|, |B(v,r)| \geq 2^i\}$. In words, r_i is the smallest radius such that balls of this radius centered at u and v respectively encompass (at least) 2^i neighbours of u and 2^i neighbours of v. To get a better feel for the situation at level i, it is useful to think of a ball around u and a ball around v growing at an identical rate. The growth continues until both balls encompass 2^i points. So the first ball reaching 2^i keeps growing until the *trailer ball* reaches its target as well (both balls could reach their target simultaneously in which case the *trailer ball* is either one). In the following we distinguish between B and B^0 only to account for the case in which the trailer ball overshoots its target (owing to the fact that more than one point could lie precisely at the trailer ball's radius).

Let q be the largest i for which $r_{i-1} < d(u,v)/2$, and set r_q precisely to $d(u,v)/2$. Then $|B(v, r_q)| \leq 2^q$ and for all $i < q$, $B^0(u, r_i) \cap B(v, r_{i+1}) = \emptyset$.

The r_i thus produce $O(\log n)$ concentric rings around u and v, such that the ith ring of u has the same radius as the ith ring in v, as shown above. Now Δ_i for all i will correspond to the interval $|r_i - r_{i-1}|$. We will assume without loss of generality below that for each i, the trailer ball at level i is the ball centered at u. For $|d(u,A) - d(v,A)| \geq \Delta_i$ to hold, it suffices to show that A meets $B^0(v, r_{i-1})$ and misses $B^0(u, r_i)$.

We'd like A to miss $B^0(u, r_i)$ but A is selected with probability $1/2^i$ so this is likely to occur. (To see what is meant by "likely", note that A misses with probability

$$\left(1 - \frac{1}{2^i}\right)^{|B^0(u,r_i)|} > \left(1 - \frac{1}{2^i}\right)^{2^i}$$

and the latter is at least $\frac{1}{4}$ for $i > 0$.) We'd like A to meet $B(v, r_{i-1})$, a set of size $\geq 2^{i-1}$. By the birthday paradox, this is also likely to occur. (Specifically, this also occurs with constant probability. This is because the expected size of A is $\frac{n}{2^i}$ and $|A| \cdot |B(v, r_{i-1})|$ is $O(n)$ so the birthday paradox applies. In more detail, one would give up the constant probability that the size of A be less than its expected value, and would apply the birthday paradox to further reduce the complementary probability.) When both events occur we have $|d(u,A) - d(v,A)| \geq \Delta_i$ as required. Note that the sum of the Δ_i is r_q. This completes the proof of Proposition 4. $\square$

Now we apply Proposition 4 as a black box, as follows. For each i, $1 \leq i \leq q$, we repeat the probabilistic experiment m times (m to be determined). We are easy to please since it suffices that we be successful, namely that $|d(u,A) - d(v,A)| \geq \Delta_i$, once for each i. Now the failure probability at level i is at most $(\frac{11}{12})^m$, so the failure probability across all levels at most $q(\frac{11}{12})^m$. By choosing $m \in O(\log n)$ such that $(\frac{12}{11})^m > n^3$, we have

$$q\left(\frac{11}{12}\right)^m < \frac{1}{n^2}$$

since q is $O(\log n)$. This shows that for a specific pair of points u and v,

$$\Pr[d(u,v) \text{ shrinks to less than } d(u,v)/2] < \frac{1}{n^2} \quad .$$

Summing over all pairs (u,v),

$$\Pr[\text{some } d(u,v) \text{ shrinks to less than } d(u,v)/2] < \frac{1}{2} \quad .$$

Hence, with nonzero probability, none of the $d(u,v)$ will shrink to less than $d(u,v)/2$. This completes the ℓ_1 case of the proof of Bourgain's theorem. The ℓ_2 case goes through with a minor technical modification to the given embedding.

18 Goemans-Williamson algorithm for MAX CUT

The definition of the MAXCUT problem is given below.

INPUT: weighted graph $G = (V,E)$, $w : E \to \mathbb{R}^+$
OUTPUT: (S, S^c) such that $\Sigma_{i \in S; j \notin S} w_{i,j}$ is maximal

MAXCUT is NP-complete. We will show a polynomial time 1.13 randomized approximation algorithm by Goemans and Williamson [14] (there also exists a derandomized version of this algorithm [14, 23]). Our strategy will be to solve a relaxed version of the problem in polynomial time, and show that this solution approximates the optimal cut in the more restricted domain. A .94 hardness result was shown by Håstad [15].

The value of the maximum cut can be expressed as:

$$\text{MAXCUT}(G) = \max_{f: V \to \{-1,1\}} \sum_{i,j \in V} \frac{1 - f(i)f(j)}{2} w_{i,j} \quad .$$

Surprisingly, it is easier to do this in high dimensions. For a given weighted graph G, we define:

$$\text{V-MAXCUT}(G) = \max_{f: V \to S^n} \left\{ \sum_{i,j \in V} \frac{1 - \langle f(i), f(j) \rangle}{2} w_{i,j} \right\}$$

where S^n is the set of points on the unit sphere. Note that $\text{MAXCUT}(G) \leq \text{V-MAXCUT}(G)$. We will show that $\text{V-MAXCUT}(G) \leq 1.13 \cdot \text{MAXCUT}(G)$. So it remains to show that:

- V-MAXCUT can be computed in polynomial time.
- 1.13 MAXCUT $\geq$ V-MAXCUT.
- we can efficiently map the V-MAXCUT solution to the solution for MAXCUT.

First, we use PSD-programming to solve V-MAXCUT. Note that finding the mapping $f : V \to S^n$ that maximizes:

$$\sum \frac{1 - \langle f(i), f(j)\rangle}{2} w_{i,j}$$

is equivalent to finding the f that minimizes:

$$\sum \langle f(x), f(y)\rangle w_{i,j} \quad .$$

For any $f : V \to S^n$, let

$$F = \begin{pmatrix} ---f(1)--- \\ ---f(2)--- \\ \vdots \\ ---f(n)--- \end{pmatrix}, \qquad Z = FF^T \quad .$$

Note that by construction, Z is positive semi-definite with elements of the form $\langle f(i), f(j)\rangle$. We then solve the following PSD program:

$$\begin{array}{rl} \text{minimize} & \displaystyle\sum_{1 \le i,j \le n} w_{i,j} z_{i,j} \\ \text{s.t.} & (\forall i) z_{i,i} = 1;\ Z \in \text{PSD}_n \end{array}$$

From Z we can extract the optimal $f(1), \ldots, f(n)$. We can solve the PSD program in polynomial time using the ellipsoid method—this takes care of the first requirement.

Next we show how to map the V-MAXCUT solution to MAXCUT. Let $f(1), \ldots, f(n)$ be the vectors in an optimal solution for V-MAXCUT. We select a random $r \in S^n$ and assign each vertex according to the sign of $\langle r, f(i)\rangle$. This takes care of the third requirement.

As to the second requirement, we first want to show:

$$\mathrm{E}\left[\sum_{i \in S; j \notin S} w_{i,j}\right] \ge 0.878 \cdot \text{V-MAXCUT} \quad .$$

Now note that the probability that the unit vectors $f(i)$ and $f(j)$ are mapped to different sides of the cut by a random projection is $\left(\frac{\theta_{ij}}{\pi}\right)$ where θ_{ij} is the angle between $f(i)$ and $f(j)$. So the expected weight of the cut is

$$\sum_{ij} \left(\frac{\theta_{ij}}{\pi}\right) w_{i,j} \quad .$$

With some trigonometry we can show:

$$\forall_{\pi \ge \theta \ge 0}: \qquad \left(\frac{\theta}{\pi}\right) \ge 0.878 \cdot \left(\frac{1 - cos\theta}{2}\right) \quad .$$

So now:

$$\begin{aligned} \mathrm{E}\left[\sum_{i \in S; j \notin S} w_{i,j}\right] &= \sum_{ij} \left(\frac{\theta_{ij}}{\pi}\right) w_{i,j} \\ &\ge \sum_{i,j} 0.878 \cdot \left(\frac{1 - \cos\theta_{ij}}{2}\right) w_{i,j} \\ &= 0.878 \sum_{i,j} \left(\frac{1 - \langle f(i), f(j)\rangle}{2}\right) w_{i,j} \\ &= 0.878 \cdot \text{V-MAXCUT} \quad . \end{aligned}$$

Thus V-MAXCUT $\le 1.13 \cdot$ MAXCUT, as required. □

2004 Barbados Workshop on Computational Complexity	
LECTURE NOTES ON FINITE METRIC SPACES AND LOW-DIMENSIONAL EMBEDDINGS	
Title: Approximating Bandwidth	Lecture #: 7
Lecturer: Santosh Vempala	Scribe: Arkadev Chattopadhyay and Michal Koucký

19 Introduction to the bandwidth problem

Here we will see an application of the random projection technique. We will use it to obtain an approximate solution for the bandwidth minimization problem via rounding off a semi-definite programming relaxation of the problem.

The *bandwidth* of an undirected connected graph $G = (V, E)$ with respect to a given linear ordering of its vertices $\pi : V \to \{1, \ldots, n\}$ (π is a bijection) is defined to be the quantity $\max_{(i,j)\in E} |\pi(i) - \pi(j)|$. Bandwidth minimization is the problem of finding an ordering of the vertices of a given graph G that minimizes the bandwidth. The problem is known to be NP-hard for general graphs. In Figure 11, there are some examples of graphs and their bandwidths. It can be easily verified that the minimum bandwidth for the path on n vertices is 1, for the star is $n/2$, for the cycle is 2 and for the clique is $n - 1$.

The problem of bandwidth minimization originates in the past when Gaussian elimination was used for solving systems of linear equations. To minimize necessary resources for performing Gaussian elimination people tried to permute rows and columns of the matrix representation of the linear system so that non-zero entries were present only in narrow strip along the main diagonal. Such a matrix M of size $n \times n$ corresponds to a graph G on n vertices, which has an edge between vertices i and j iff $M_{i,j}$ is non-zero. Permuting rows and columns of M to get the most narrow strip along the main diagonal corresponds to the bandwidth minimization of G.

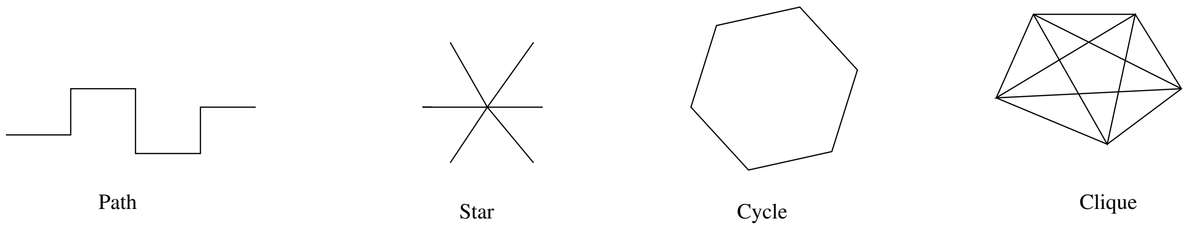

Figure 11: Bandwith examples

One can view the problem of bandwidth minimization as the following geometrical embedding problem: arrange the n vertices of the graph on the positive quadrant of a circle of radius n so that the vertices are equally spaced (at distance at least 1) and the Euclidean distance between any two points that form an edge in the graph is minimized. Projecting any such placement on either axes gives an optimal ordering of the vertices. (There is a correspondence between the bandwidth and the maximum euclidean length of any edge in the placement, since the points are equi-spaced on the circle).

We will generalize this idea to higher dimensions. We will place the points of the graph on an n-dimensional sphere of radius n so that certain criteria are optimized. Then we will project these points on a random line through the origin to obtain an approximate solution to the bandwidth minimization problem. To solve the problem efficiently we will try to describe the placement problem by a vector program which will then be solved by semi-definite programming.

There are two somewhat different approaches to what criteria should be optimized. One method is due to Blum et al. [1] and the other one due to Feige [12]. We begin with the former method. It will give us, with high probability, a linear ordering of the vertices with bandwidth bounded by $O(b\sqrt{\frac{n}{b}}\log n)$, where b is the optimal bandwidth. Notice that this corresponds roughly to a $\sqrt{n}$-approximation of the bandwidth. In the second part of the lecture we will consider Feige's method that uses a notion of *low volume distortion* to achieve a (better) poly-logarithmic approximation.

20 Spreading constraints method

We will formulate now the semi-definite program of Blum et al. Recall that when we place the vertices of the graph on the sphere we wish to enforce that they are "spread somewhat evenly" around the sphere. To express this as a semi-definite program, we make the following observation about evenly-space points on a line: *there are at most* $2k$ *points at distance at most* k *from any point*. This implies

$$\forall S \subseteq V,\ \forall i \in V,\ \sum_{j \in S} d(i,j) \ \geq\ 2(1+2+\ldots+|S|/2)\ \geq\ \frac{|S|^2}{4}\ .$$

This condition is called the *spreading* property and can be generalized to higher dimensions (see (7) of the vector program below). Metrics that have the spreading property are called *spreading metrics*. Using this idea, we formulate the following vector program with unkown vectors x_i's and a scalar b:

$$\begin{aligned}
&\min && b && (3)\\
\forall \{i,j\} \in E;\ ||x_i - x_j||_2 &\leq && b && (4)\\
\forall i,j;\ x_i \cdot x_j &\geq && 0 && (5)\\
\forall i;\ ||x_i||_2 &= && n && (6)\\
\forall i \in V, \forall S \subseteq V;\ \frac{1}{|S|}\sum_{j \in S} ||x_i - x_j||_2^2 &\geq && \frac{1}{6}\left(\frac{|S|}{2}+1\right)(|S|+1) && (7)
\end{aligned}$$

Note that $||v||_2^2$ can be expressed as the inner product of v with itself. Hence one can apply standard techniques to transform our vector program into an instance of semi-definite programming. This semi-definite program can be solved efficiently.

(One needs to verify that for any set of vectors $x_1, \ldots, x_n$ and a real b we can efficiently determine whether constraints (4)-(7) are satisfied and, if not, which one is violated. This is the problem of showing that the *separation oracle* can be implemented efficiently). Although condition (7) implies exponentially many constraints one can check whether any of them is violated in polynomial time: for every vertex i, sort all other points j according to their distance from i, i.e., according $||x_i - x_j||_2^2$; once sorted, check (7) only for subsets S that are prefices of the sorted list. (4)-(6) can be checked in a straightforwad manner.)

Now we are ready to state our algorithm for computing the linear ordering:

Algorithm:

1. Solve the SDP relaxation as given in (4)-(7). Let vectors $x_1, \ldots, x_n$ be the solution.
2. Take a random unit vector ℓ and project $x_1, \ldots, x_n$ onto line l given by ℓ.
3. Output the order in which the points appear on l i.e., the order of increasing $x_i \cdot \ell$.

Note: Every edge $e = (i,j)$ in the graph corresponds to the vector $x_i - x_j$ (denoted by x_e and called the *edge vector*), where vectors x_i and x_j represent vertices i and j in the solution of the above vector program.

Theorem 13. *The algorithm above generates with high probability an ordering whose bandwidth is bounded by* $O(\sqrt{\frac{n}{b}}\log n)$ *times the optimum b.*

We will only give a sketch of the proof here, leaving out many technical details. The complete proof of this theorem can be found in [1].

Proof sketch. The crucial fact is that for any edge e in G, the corresponding edge vector x_e is almost orthogonal to a random unit vector ℓ with high probability. This means that the projected length of vector x_e is "small" with high probability. This becomes useful in the following way: we can show that the probability of the event that the projected length of x_e exceeds $W = 8b\sqrt{\log n}/\sqrt{n}$ is fairly small. We partition the random line ℓ into intervals of length W and thus every edge, with high probability spans at most two such intervals. We can now obtain our theorem by showing that with reasonable probability none of these intervals have more than $O(\sqrt{nb}\log n)$ points projected onto them.

The following helps to complete the argument: If X is a random variable that represents the number of points that get projected into a certain fixed interval of width W on ℓ, then $E[X^2] = O(W^2 n \log n)$. One can now apply Chebyshev's inequality to obtain

$$\frac{b}{4n} \geq \Pr\left[X > \sqrt{\frac{4n}{b}}\sqrt{W^2 n \log n}\right] = \Pr\left[X > 16\sqrt{bn}\log n\right] \tag{8}$$

We can also see that the middle n/b intervals of ℓ span a total width of $\frac{n}{b}W = 8\sqrt{n\log n}$. This implies that all vectors $x_1, \ldots, x_n$ fall into these intervals with very high probability. Using our bound from (8), we conclude that with probability close to $\frac{3}{4}$, each of these n/b intervals has only $O(\sqrt{nb}\log n)$ vertices, concluding the argument. $\square$

21 Low volume distortion method

Feige's method is based on the following observation. Consider two vertices of G connected by an edge together with some subset of other vertices from G. If the volume spanned by these vertices when embedded onto sphere is small relative to the length of the edge, then when projected on a random line many point from the subset will fall between the two end-points of the edge. This in turns translates into large bandwidth. Hence, Feige's approach is to embed the vertices of a graph on the sphere as to maximize volume spanned by any subset of the vertices.

We need the following notions to discuss this approach. For a subset $S \subseteq \mathbb{R}^n$, $\textsc{Evol}(S)$ denotes the *Euclidian volume* of the convex hull of S. For an arbitrary metric space (X, d) and any subset $S \subseteq X$, we define $\mathrm{MST}(S)$ to be any minimum spanning tree of S, i.e., a set of pairs $\{i,j\} \subseteq S$ such that $G(S, MST(S))$ is a connected graph and $\sum_{\{i,j\}\in \mathrm{MST}(S)} d(i,j)$ is minimum. Then the *tree volume of S* is

$$\textsc{Tvol}(S) = \prod_{\{i,j\}\in \mathrm{MST}(S)} d(i,j) \ .$$

Observe that although there may be many different minimum spanning trees of S, the tree volume of S is well-defined.

The following claim is usefull for the analysis:

Claim 3. *Let k be a positive integer and S be a subset of size k of metric space (X,d). Let ϕ be an embeding of S into $\mathbb{R}^n$. Then* $\textsc{Evol}(\phi(S)) \leq \frac{\textsc{Tvol}(S)}{(k-1)!}$.

Proof. By induction on size of S. If the size of S is less than 3 then the claim is trivial. Consider a set S of size at least 3. Assume that the claim is true for all sets of size less than $|S|$. Consider a $\mathrm{MST}(S)$ and let v be a leaf in $G(S, \mathrm{MST}(S))$ connected to some vertex u. Let $S' = S \setminus \{v\}$. Then $\mathrm{MST}(S) \setminus \{u, v\}$ is clearly $\mathrm{MST}(S')$. By induction hypothesis $\textsc{Evol}(\phi(S')) \leq \frac{\textsc{Tvol}(S')}{(k-2)!}$. Also $\textsc{Evol}(\phi(S)) \leq \textsc{Evol}(\phi(S')) \cdot d(u,v)/(k-1)$. Hence, $\textsc{Evol}(\phi(S)) \leq \frac{\textsc{Tvol}(S)}{(k-1)!}$.

For a metric space (X,d), an embedding ϕ of X into $\mathbb{R}^n$, and a positive integer k, we define the *volume distortion of ϕ*:

$$\eta(k,\phi) = \max_{S\subseteq X;\ |S|=k} \left(\frac{\textsc{Tvol}(S)}{(k-1)!\ EVOL(\phi(S))}\right)^{\frac{1}{k-1}} \ .$$

The key fact that Feige proves is a bound on the volume distortion:

Theorem 14 (Feige). *For any n-point metric space (X,d) and positive integer k, there exists a non-expanding embedding ϕ of X into Euclidian space such that $\eta(k,\phi) = O(\log n + \sqrt{k\log n \log k})$.*

The proof of this theorem gives a Bourgain-like embedding of required volume distortion of X into Euclidian space. In particular the embedding can be constructed as follows. Create a space (X', d'), where X' contains exactly $\log k$ copies $x^1, x^2, \ldots, x^{\log k}$ of every $x \in X$; the distance d' of two points in X' is the same as that of the original points in X (it is zero if the points are copies of the same point). The embedding ϕ has one coordinate ϕ_S for every subset S of X':

$$\phi_S(u) = \frac{d'(u,S)}{\sqrt{|S|\binom{n \log k}{|S|}}} \ ,$$

where $d'(u,S)$ is the distance of any copy of u in X' from S.

One can also construct a more efficient embedding by a randomized algorithn. This is discussed in the next lecture.

2004 Barbados Workshop on Computational Complexity

LECTURE NOTES ON FINITE METRIC SPACES AND LOW-DIMENSIONAL EMBEDDINGS

Title: Volume-Preserving Embeddings — Lecture #: 8

Lecturer: Santosh Vempala — Scribe: Mark Mercer and Mike Saks

22 Bandwidth approximation

If G is a graph, and τ maps $V(G)$ to a metric space (X, d), the *edge stretch* of τ, $\sigma(\tau)$, is the maximum of $d(\tau(v), \tau(w))$ over all edges (v, w). The *bandwidth of* G, $bw(G)$, is the minimum of $\sigma(\tau)$ over all bijections τ mapping $V(G)$ to $\{1, \ldots, n = |V(G)|\}$ with the natural metric. In the previous lecture we saw an $O(\sqrt{n}\log n)$ factor approximation algorithm for $bw(G)$. The basic approach for the approximation is given by:

Algorithm Template

1. Apply a suitable 1-1 mapping ρ from $V(G)$ to $\mathbb{R}^d$ for some d.
2. Choose a vector ℓ in $\mathbb{R}^d$ randomly according to a suitable distribution D and let $\pi_\ell : \mathbb{R}^d \longrightarrow \mathbb{R}$ be given by $\pi_\ell(x) = \langle x, \ell\rangle$. Apply π_ℓ to $\rho(V(G))$. Thus we have computed $\psi = \pi_\ell \circ \rho$.
3. Define τ to map $V(G)$ to $\{1, \ldots, |V(G)|\}$ according to the linear order given by ψ.

To specify an algorithm we need to choose ρ and D. The natural distribution for ℓ is the uniform distribution (i.e., given by Haar measure) from the unit sphere.

The choice of ρ in step (1) is the main ingredient. The $\sqrt{n}\log n$ approximation obtained in the previous lecture chose ρ to be the solution, ρ_{SDP}, of a particular semi-definite program (SDP). In this lecture we obtain a polylogarithmic approximation algorithm by a better choice of ρ. To motivate the new choice, let's summarize the previous analysis.

We want to bound $\sigma(\tau)$ from above. For a finite subset U of $\mathbb{R}$ and $L \geq 1$ the *L-crowding* $m_L(U)$ is the maximum number of points of U in any interval of length $1/L$. The stretch of τ is the maximum over $(v, w) \in E(G)$ of $|\{z \in V \mid \min\{\psi(v), \psi(w)\} < \psi(z) \leq \max\{\psi(v), \psi(w)\}\}|$ which, for any choice of L, is $O(\sigma(\psi)m_L(\psi(V))L)$.

To bound $\sigma(\psi)$ from above, we note that $\sigma(\psi) \leq \sigma(\rho)\sigma'$ where σ' is the stretch of π_ℓ restricted to the image of ρ. A routine calculation (essentially the Chernoff bound) shows that for any edge (v, w) and any $K > 1$, the probability (with respect to the choice of ℓ) that $\pi_\ell(\rho(v)) - \pi_\ell(\rho(w))$ exceeds $\|\rho(v) - \rho(w)\|\frac{K}{\sqrt{d}}$ is $e^{-O(K^2)}$. Choosing $K = O(\sqrt{\log n})$, we conclude that $\sigma' = O(\frac{\sqrt{\log n}}{\sqrt{d}})$ with high probability. Therefore $\sigma(\psi) = O(\sigma(\rho)\frac{\sqrt{\log n}}{\sqrt{d}})$ and the bandwidth of τ is thus $O(\sigma(\rho)m_{\sqrt{d}}(\psi(V))\sqrt{\log n})$.

In the previous lecture, we chose ρ to be the solution of a certain semi-definite program as follows. The standard metric on the integers satisfies a system of constraints called *spreading constraints*: For each $i \in \mathbb{Z}$, and each finite $S \subseteq \mathbb{Z}$,

$$\sum_{j\in S} d(i, j) \geq \lfloor(|S| + 1)/2\rfloor\lceil(|S| + 1)/2\rceil \ .$$

(Sometimes the right-hand side of the spreading constraints is replaced by some other quadratic function of $|S|$, which does not affect the relevant properties of such metrics).

The mapping ρ_{SDP} is that map from $V(G)$ to the (Euclidean) sphere of radius n in $\mathbb{R}^n$ that minimizes $\sigma(\rho_{\text{SDP}})$ subject to the condition that $(\rho_{\text{SDP}}(v) : v \in V(G))$ satisfies the spreading constraints with respect to the ℓ_2 metric. It was sketched in the last lecture how this can be formulated as an SDP and solved in polynomial time.

The optimal objective function value for this SDP is easily seen to be $O(bw(G))$: given an optimal ordering for bandwidth, we can get a feasible solution to the SDP with objective function value $O(bw(G))$ by placing the vertices in this order equally spaced along a semicircle.

Therefore the bandwidth of the final mapping is then $O(bw(G)m_{\sqrt{d}}(\psi(V))\sqrt{\log n})$, where $d = n$, and so the approximation ratio for the algorithm is $O(m_{\sqrt{n}}(\psi(V))\sqrt{\log n})$. In the previous lecture, it was shown that if ρ_{SDP} is used in step (1), then with probability close to 1 (over the choice of ℓ) , $m_{\sqrt{n}}(\psi(V))$ is $O(\sqrt{n \log n})$. This gave an approximation ratio of $O(\sqrt{n} \log n)$.

Here we will give a different choice of ρ which still satisfies $\sigma(\rho) = O(bw(G))$. However, the dimension d of the range of ρ is polylogarithmic in n and, most significantly, with high probability (over the choice of ℓ) $m_{\sqrt{d}}(\psi(V))$ is polylogarithmic in n. This will give a polylogarithmic approximation to bandwidth.

23 A sufficient condition for small crowding

We now consider the question: What conditions on a subset U of points on the radius d sphere of $\mathbb{R}^d$, ensure that for most ℓ, π_ℓ applied to U will have have small crowding? For a subset S of $\mathbb{R}^d$, the *spread* of the subset S, $\textsc{Spread}(S)$ is the random variable $\max_{v \in S} v \cdot \ell - \min_{v \in S} v \cdot \ell$.

Definition 12. *The Euclidean volume,* $\textsc{Evol}(S)$*, of a subset* S *of at most* $d+1$ *points lying on a sphere in* $\mathbb{R}^d$ *is the volume of the simplex spanned by the points.*

Lemma 5. *Let* S *be a finite set of points in* $\mathbb{R}^d$ *each having Euclidean length d. Let* ℓ *be a randomly chosen unit vector in* $\mathbb{R}^d$*. Then:*

$$\Pr[\textsc{Spread}(S) \leq \eta] \leq \frac{\eta\sqrt{d}}{(|S|-1)!\textsc{Evol}(S)} \quad .$$

The proof is omitted from these notes. From the lemma, we deduce:

Corollary 3. *Let* U *be a finite set of points in* $\mathbb{R}^d$ *each having Euclidean length d. Let* $k \leq |U|$ *be a positive integer and let* $\varepsilon > 0$*. Let* ℓ *be a randomly chosen unit vector in* $\mathbb{R}^d$*. Then*

$$\Pr\left[m_{\sqrt{d}}(\pi_\ell(U)) \leq 6 \left(\frac{1}{\varepsilon} \sum_S \frac{1}{|\textsc{Evol}(S)|} \right)^{1/k} \right] > 1 - \varepsilon \quad .$$

Proof. For every subset $S \subseteq U$ of size k, we define a random variable Y_S to be the indicator of the event that $\textsc{Spread}(S) \leq 1/\sqrt{d}$. Let $Y = \sum_{S:|S|=k} Y_S$ and $N = E[Y]$. Trivially $Y \leq \binom{U}{k} \leq (U/k)^k$. By Markov's inequality, $\Pr[Y \geq N/\varepsilon] \leq \varepsilon$, which implies that $\Pr[U \geq k(N/\varepsilon)^{1/k}] = \Pr[(U/k)^k \geq N/\epsilon] \leq \epsilon$. By the lemma,

$$\begin{aligned} N = E\left[\sum_{S:|S|=k} Y_S\right] = \sum_{S:|S|=k} E[Y_S] &= \sum \Pr[\textsc{Spread}(S) \geq 1/\sqrt{n}] \\ &\leq \sum \frac{1}{(k-1)!\textsc{Evol}(S)} \quad . \end{aligned}$$

The corollary now follows by observing that $k\left(\frac{1}{(k-1)!}\right)^{1/k} \leq 6$ for all integers k. ∎

24 Bounding $\left(\sum_{S:|S|=k} \frac{1}{\textbf{Evol}(S)}\right)^{1/k}$

By the above corollary, a sufficient condition for $\psi(V)$ to have crowding polylogarithmic in n (and thus yield a polylogarithmic approximation to bandwidth) would be that for some $k \geq \log n$, we have that

$$\sum_{S:|S|=k} \frac{1}{\textsc{Evol}(S)} \leq n^{O(1)}(\log n)^{O(k)} \quad ,$$

where the sum is over $S \subseteq \rho(V)$. Unfortunately, $\rho_{\mathrm{SDP}}(V)$ does not guarantee good enough upper bounds on $\sum_{S:|S|=k} \frac{1}{\textsc{Evol}(S)}$. So we now look for a mapping that does. A key observation is that ρ_{SDP} does satisfy an inequality that is similar to what we need, but with $\textsc{Evol}$(S) replaced by a quantity called *tree volume*:

Definition 13. *Let S be any finite subset of a metric space. The* volume *of a tree T on vertex set S is the product of the distances along all $|S|$-1 pairs of adjacent vertices. The* tree volume *of S, $\text{TVOL}(S)$ is the minimum volume of any tree on S (which is attained by any minimum spanning tree).*

The proof of the following lemma is omitted.

Lemma 6. *For any finite spreading metric M, and any $k \leq n$,*

$$\sum_{S \subseteq M || S|=k} \frac{1}{\text{TVOL}(S)} \leq |M|(\log |M|)^{k-1} \quad .$$

25 Converting tree volume to Euclidean volume

The set $X = X(G) = \{x_v \mid v \in V(G)\}$ obtained by solving the SDP satisfy the spreading constraints and therefore satisfies the conclusion of Lemma 6. Now for any set S in Euclidean space, it is easy to prove by induction on $|S|$ that $\text{EVOL}(S) \leq \text{TVOL}(S)/(|S|-1)!$. If we had equality for all S of size k for some $k \geq \log n$, then Lemma 6 would give a good enough bound on $\sum_{S \subseteq M||S|=k} \frac{1}{\text{EVOL}(S)}$ to imply a polylogarithmic approximation algorithm.

However, $\text{EVOL}(S)$ may be much smaller than $\text{TVOL}(S)/(|S|-1)!$, So our goal will be to find a nonexpansive embedding φ from X to a new Euclidean space in such a way that for each $S \subseteq X$ of size k, $\text{TVOL}(\phi(S))$ is "close" to $1/\text{EVOL}(\phi(S))(k-1)!$. We will call such a map *volume preserving*. Formally, we define the *k-volume distortion* $\eta(k, \varphi)$ of an embedding φ of X to be:

$$\eta(k, \varphi) = \max_{S \subseteq X, |S|=k} \left\{ \left(\frac{\text{TVOL}(S)}{(k-1)!\text{EVOL}(\varphi(S))} \right)^{\frac{1}{k-1}} \right\}$$

and we define $\eta(k)$ as

$$\eta(k) = \min_{\varphi}\{\eta(k, \varphi) \mid \varphi \text{ is nonexpansive}\} \quad .$$

In this section we will give an algorithm that for any metric space X, constructs a non-expansive embedding φ from X to Euclidean space of dimension $(\log n)^{O(1)}$ having k-volume distortion which is $(k \log n)^{O(1)}$.

Armed with this mapping we then choose $k = O(\log n)$, and define the mapping ρ for the first step of the algorithm template to be the composition $\varphi \circ \rho_{\text{SDP}}$. The approximation factor of this algorithm is then

$$O(\sigma(\rho) m_{\sqrt{\log n}}(\psi) \sqrt{\log n}) \quad .$$

Since the map is non-expansive, we have $\sigma(\rho) = O(\sigma(\rho_{\text{SDP}})) = O(bw(G))$. Using the fact that ρ_{SDP} is a spreading metric, Lemma 6, the volume preservation property of φ, and Corollary 3, we conclude that the algorithm obtained has approximation factor $O((\log n)^{O(1)})$.

So it remains to give the algorithm for computing φ. Here is a general class of mappings from a finite metric space (X, d) in a finite dimensional Euclidean space. For any sequence $S_1, \ldots, S_N$ of subsets of X define, for each $i \in \{1, \ldots, N\}$, the map $\varphi_i : X \longrightarrow \mathbb{R}$ by $\varphi_i(x) = d(x, S_i)/\sqrt{N}$ (where $d(x, \emptyset)$ is defined to be the diameter of (X, d)), and define $\varphi : X \longrightarrow \mathbb{R}$ by $\varphi = (\varphi_1, \ldots, \varphi_N)$. It is easy to show that this mapping is non-expansive for any choice of $(S_1, \ldots, S_N)$. We now construct a probability distribution on subsets of X so that if each S_i is chosen independently according to this distribution, then with probability near 1, the resulting map has low k-volume distortion.

We will select N to be $(k \log n)^{O(1)}$ (as needed for the analysis below) and select each S_j independently according to the following probability distribution. For each $j \in \{0, 1, \ldots, \log n \log k\}$, define

$$p_j = 1 - \left(\frac{1}{k}\right)^{1/\left(1+\frac{1}{\log k}\right)^j} \quad .$$

Now each S_i is selected as follows: (1) Choose $j = j(i)$ uniformly from the set $\{0, 1, \ldots, \log n \log k\}$. (2) Independently, for each element $x \in X$, place x in S_i with probability $p(j)$.

Theorem 15. *(Feige) For S_i selected as above, with probability at least .99 the mapping φ has k-volume distortion $O\left(\log n + \sqrt{k \log k \log n}\right)$.*

Some notation: Henceforth we use η to be a parameter equal to the k-volume distortion claimed in the theorem. The *affine hull* of a finite subset $P \subseteq \mathbb{R}^N$, denoted Aff(P) is the set of points of the form $\sum_{x \in P} \lambda_x x$, where $\sum_x \lambda_x = 1$. This is the smallest affine subspace containing P. For a closed subset $Z \subseteq \mathbb{R}^N$ and $y \in \mathbb{R}^N$, we define $\|y - Z\|$ to be the minimum of $\|y - z\|$ over all $z \in Z$.

To prove the theorem, we show that for each subset S of X of size k, the event B_S that $\text{EVOL}(\phi(S)) \leq \frac{\text{TVOL}(S)}{(k-1)!(2\eta)^{k-1}}$ occurs with probability less than $(.01/\binom{n}{k})$. In the proof sketch below, we suppress the quantitative details of the upper bounds on low probability events, and say that probabilities are "sufficiently small" or "sufficiently close to 1" or that an event "holds almost certainly".

So fix $S \subseteq X$ of size k. We want to bound the ratio of $\text{EVOL}(\varphi(S))$ to $\text{TVOL}(S)$. Below, we will carefully select an ordering $u_1, \ldots, u_k$ of S. Having picked this ordering we define for each $i \in \{1, \ldots, k\}$:

- $U^i = \{u_1, \ldots, u_i\}$,
- A^i is the affine hull of $\varphi(U^i)$,
- $d_i = d(u_i, U^{i-1})$ (where $d(\cdot)$ is the distance function in X),
- $q_i = d_i/2$,
- $T_i = \{v \in X \mid d(u_i, v) < q_i\}$, and
- $t_i = |T_i|$.

We have:

$$\begin{aligned} \text{TVOL}(S) &\leq d_2 d_3 \cdots d_k = 2^{k-1} q_2 \cdots q_k \ , \\ \text{EVOL}(\varphi(S)) &= \left(\frac{1}{(k-1)!}\right) \prod_{i=2}^{k} \|\varphi(u_i) - A^{i-1}\| \ . \end{aligned}$$

The bound on $\text{TVOL}(S)$ is obtained by taking the product of the edges of the tree where u_i is joined to the closest point in U^{i-1}. The expression for $\text{EVOL}(\varphi(S))$ is proved by induction; for each $2 \leq i \leq k$, let C^i be the convex hull of U^i and VOL_i be the $(i-1)$-dimensional volume of C^i. Then, $\text{VOL}_2 = \|\varphi(u_2) - \varphi(u_1)\| = \|\varphi(u_2) - A^1\|$, and for $i \geq 3$, $\text{VOL}_i = \frac{1}{i}\|\varphi(u_i) - A^{i-1}\| \text{VOL}_{i-1}$.

Taking the ratio of the above expressions we get:

$$\frac{\text{TVOL}(S)}{(k-1)!\text{EVOL}(\varphi(S))} \leq \prod_{i=2}^{k} \frac{2q_i}{\|\varphi(u_i) - A^{i-1}\|} \ .$$

To show that the k-volume distortion of φ is almost certainly most η it therefore suffices to show:

Claim 4. *For each i*

$$\|\varphi(u_i) - A^{i-1}\| > 2q_i/\eta \tag{9}$$

holds almost certainly.

We show this provided that for the ordering $u_1, \ldots, u_k$ the parameters $t_2, \ldots, t_k$ (defined above) are in nonincreasing order. Such an ordering is obtained by choosing the points in reverse order: to choose u_i, assuming $u_{i+1}, \ldots, u_k$ have been selected, we take $U^i = S - \{u_{i+1}, \ldots, u_k\}$ and choose u_i from U^i so that

$$\left|\{v \in X \mid d(u_i, v) < d(U^i - u_i, u_i)/2\}\right|$$

is minimized. Henceforth we fix an ordering satisfying this property.

The nonincreasing order of the t_i easily implies:

For any $i < j$, the t_j closest points to u_i are disjoint from T_j.

To prove the claim, fix $i \in \{2, \ldots, k\}$. The selection of each set S_α, $1 \le \alpha \le N$, involves selecting the index $j(\alpha)$ and then, for each $v \in X$, independently deciding whether $v \in S_\alpha$ with probability $p_{j(\alpha)}$. Let us call $j(\alpha)$ the *type* of α. Let us say that index α is *useful* if for each $h < i$, $S_\alpha \cap (X - T_i)$ contains at least one of the t_i closest points to u_h. For each index α define S'_α to be S_α if α is not useful and $S_\alpha \cap (X - T_i)$ if α is useful. Let φ' be the map defined by the sets S'_α.

Proposition 5. *For each $h < i$, $\varphi'(u_h) = \varphi(u_h)$. Thus A^{i-1} is determined by the sets $\{S'_\alpha \mid 1 \le \alpha \le N\}$.*

Proof. Let $h < i$ and let α be an index. We need to show that $d(u_h, S'_\alpha) = d(u_h, S_\alpha)$. The claim is trivial for α that is not useful, so assume that α is useful. Then S'_α contains one of the t_i closest elements to u_h, and that element is closer to u_h then any element in T_i. The claim follows. $\square$

For any index α, the probability that α is useful given that it is type j is at least at least $(1-(1-p_j)^{t_i})^{i-1}$. Let j^* be the largest index such that $p_{j^*} \ge 1 - (1/k)^{1/t_i}$. (This is well defined since $p_0 = 1 - 1/k$.) For $j(\alpha) \le j^*$, we have $p_{j(\alpha)} \ge p_{j^*} \ge 1 - (1/k)^{1/t_i}$ and so the probability that α is useful is at least $1/e$. For each $j \le j(\alpha)$, the expected number of useful indices α with $j(\alpha) = j$ is then $N/2e \log n \log k$ and by a Chernoff bound argument we have:

Proposition 6. *With probability sufficiently close to 1 the following condition holds:*

(*) *for each $j \in \{0, \ldots, j^*\}$, there are at least $\frac{N}{2e \log n \log k}$ useful indices of type j.*

Now fix $j(\alpha)$ and S'_α for $\alpha \in \{1, \ldots, N\}$ and assume that these values satisfy (*). We show that, conditioned on these values, (9) almost certainly holds. As observed above, A^{i-1} is determined by the conditioning.

By the definition of q_i there is an index $c(i) \in \{1, \ldots, i-1\}$ such that $d(u_i, u_{c(i)}) = 2q_i$. Let

$$B = \{x \in A^{i-1} \mid \|x - u_{c(i)}\| \le 3q_i \quad .$$

Any point of A^{i-1} that is within $2q_i/\eta$ of u_i must be in B. We may choose a subset F of B of size $(O(1/\eta))^k$ such that every point of B is within q/η of a point of F. To verify (9) it suffices to show that for each point $x \in F$, the probability that $\|x - \varphi(u_i)\| < \frac{3q_i}{\eta}$ is suitably small. (Here "suitably small" means much smaller than $(k(n\eta)^k)^{-1}$, in order that the overall failure probabilty be small after adding over all points in F and over the k points in S and then over the $\binom{n}{k}$ choices for S.)

So fix $x \in F$. We look at $\|x - \varphi(u_i)\|^2$. The main thing to show is that the conditional expectation of this quantity is $\Omega((\frac{q_i}{\eta})^2)$; it can be shown by a Chernoff bound that the quantity is almost certainly at least half its expectation.

Let m^* be the largest integer such that $2^{m^*} \le t_i$. We consider two cases. For the first case, suppose there are at least 2^{m^*-1} points of X within distance $q_i/2$ of u_i. By definition of j^* and by the definition of p_j, we have

$$1 - (1/k)^{1/t_i} \le p_{j^*} \le 1 - (1/k)^{(1+1/\log k)/t_i} \quad .$$

For each index α of type j^*, $\varphi_\alpha(u_i) > q_i/\sqrt{N}$ with probability at least $(1-p_{j^*})^{t_i} \ge 1/2k$, and $\varphi_\alpha(u_i) \le q_i/2\sqrt{N}$ with probability at least

$$1 - (1-p_{j^*})^{2^{m^*-1}} \ge 1 - (1/k)^{2^{m^*-1}/t_i} \ge 1 - 1/k^{1/4} \quad .$$

This implies that $\|x_\alpha - \varphi_\alpha(u_i)\| \ge q_i/4\sqrt{N}$ with probability at least $\min\{\frac{1}{2k}, 1 - k{-}1/4\} = \Omega(1/k)$. Thus $E[\|x_\alpha - \varphi_\alpha(u_i)\|^2] = \Omega(\frac{q_i^2}{Nk^2})$. By condition (*), there are $\Omega(N/\log n \log k)$ indices α of type j^*, and summing over these indices we get:

$$E[\|x - \varphi(u_i)\|^2] = \Omega(\frac{q_i^2}{k^2 \log n \log k}) = \Omega(\frac{q_i^2}{\eta^2}) \quad .$$

For the second case, suppose there are fewer than 2^{m^*-1} points within distance $q_i/2$ of u_i. For each m between 0 and m^*-1, let r_m be the minimum radius such that there are at least 2^m points within distance r_i of u_i. We have $1 = r_0 < r_1 \le r_2 \le \ldots \le r_{m^*-1}$ and $r_{m^*-1} \ge q/2$. For h between 1 and m^*-1, there are $\Omega(\log k)$ integers j for which $2^h \le 1/p_j < 2^{h+1}$. For such a j consider a coordinate α of type j. The probability that $\varphi_\alpha(u_i) \ge r_h/\sqrt{N}$ is at least $(1-p_j)^{2^h} \ge (1-2^{-h})^{2^h} \ge 1/4$. The probability that $\varphi_\alpha(u_i) \le r_{h-1}\sqrt{N}$ is at least

$1-(1-p_j)^{2^{h-1}} \geq 1-(1-2^{-(h+1)})^{2^{h-1}} \geq 1-1/e^{1/4}$. Thus $E[\|x_\alpha - \varphi_\alpha(u_i)\|^2] = \Omega(\frac{(r_h - r_{h-1})^2}{N})$. There are $\Omega(\log k)$ indices j for which $2^{h-1} \leq 1/p_j < 2^h$, and by (*), for each such j, there are $\Omega(N/\log n \log k)$ coordinates of type j, these coordinates contribute $\Omega(\frac{(r_h - r_{h-1})^2}{\log n})$ to the expectation. Summing for $1 \leq h \leq m^* - 1$, bounding a sum of squares by the square of the sum over the number of terms (which is less than $\log n$) we get:

$$E[\|x - \varphi(u_i)\|^2] = \Omega\left(\frac{(r_{m^*-1} - r_0)^2}{\eta^2}\right) = \Omega\left(\frac{q_i^2}{\eta^2}\right).$$

Thus $\|x - \varphi(u_i)\|^2$ has expectation as claimed. As mentioned earlier, an argument based on the Chernoff bound shows that almost certainly this quantity is at least half its expectation, and this is what is needed to complete the proof that the given map has k-volume distortion $O((k \log n)^{O(1)})$.

2004 Barbados Workshop on Computational Complexity	
LECTURE NOTES ON FINITE METRIC SPACES AND LOW-DIMENSIONAL EMBEDDINGS	
Title: Approximating Arbitrary Metrics by Tree Metrics	Lecture #: 9
Lecturer: Nati Linial	Scribe: Anna Gál and Andreas Krebs

26 Tree metrics

Let $T = (V, E, w)$ be a weighted tree, then d_T is the metric of the shortest path, e.g. $d_T(u, v) = \sum_{e \in P} w(e)$ where P is the set of edges of the shortest path from u to v.

Definition 14. *A n-point metric (X, d) is called a tree metric, if there exists a tree $T = (V, E, w)$ such that $X \subseteq V$ and $d(u, v) = d_T(u, v)$ for all $u, v \in X$.*

Example 3. *Every tree T with its shortest path metric d_T is a tree metric.*

Example 4. *Let (X, d) be a n-point metric, where every two points have distance 1. In order to see that this is also a tree metric we choose $T = (X \cup R, \{(R, x) | x \in X\}, w : e \mapsto 1/2)$, so the shortest path between two points $u \neq v$ is always $u \to R \to v$.*

26.1 Embedding tree metrics in ℓ_1

Definition 15. *We say that (X, d) is a cut metric if there is a set $S \subseteq X$ and a positive value α so that:*

$$d(u,v) = \begin{cases} 0 & \text{if both } u, v \text{ belong to } S \text{ or both } u, v \text{ do not belong to } S, \\ \alpha & \text{if one of } u, v \text{ belongs to } S \text{ and the other does not.} \end{cases}$$

Lemma 7. *Every cut metric embeds isometrically into ℓ_1.*

Proof. We map all points in S to the vector (0) and all other points to the vector (α). □

Lemma 8. *The sum of ℓ_1 metrics is an ℓ_1 metric.*

Proof. Let (T_i, ρ_i) be a collection of ℓ_1 metrics, then we can assume that each metric is embedded in a subspace of the normed space ℓ_1. The direct sum of these vector spaces is again an ℓ_1 space, with the distance function $\sum \rho_i$. □

Lemma 9. *Every tree metric embeds isometrically into ℓ_1.*

Proof. We will show that every tree metric is the sum of cut metrics and then use the fact that we can embed a cut metric into ℓ_1.

Let (X, d) be any tree metric. For every tree metric there is a weighted tree $T = (V, E, w)$ with $X \subseteq T$ and the same distances on X. For every edge $e \in E$ we let d_e be the cut metric of the edge e. That is two points on the same side of the edge have distance 0 and two points on different sides have distance $w(e)$.

$$d_e(u,v) = \begin{cases} 0 & \text{if } u, v \text{ are on the same side of } e, \\ w(e) & \text{if } u, v \text{ are on different sides of } e. \end{cases}$$

But then we have for two points $u, v \in X$ that

$$d(u, v) = d_T(u, v) = \sum_{e \in P} w(e) = \sum_{e \in P} d_e(u, v) = \sum_{e \in E} d_e(u, v) \ ,$$

where P is the set of edges of the shortest path from u to v.

So the original metric is the sum of cut metrics, but all cut metrics are ℓ_1 metrics and the sum of ℓ_1 metrics is again an ℓ_1 metric. □

We could try to embed tree metrics into ℓ_2 instead of ℓ_1. This is not possible with distortion 1: the metric of $K_{1,3}$ (the 4-vertex tree with a root and three leaves) has no isometric embedding into ℓ_2. However, by a result of Matousek [25], every n-vertex tree embeds into ℓ_2 with distortion $O(\sqrt{\log\log n})$. Bourgain [7] on the other hand showed that this result is tight for complete binary trees.

26.2 Approximation of general metrics

In the following we will look at the general case. By a probabilistic method we will approximate an arbitrary metric by a distribution of tree metrics. This approach has many algorithmic applications for problems that have a good solution for tree metrics.

Definition 16. *A metric (X, d_1) dominates (X, d_2) if for all u, v, we have $d_1(u, v) \geq d_2(u, v)$.*

In general, the idea of just trying to approximate a given metric by tree metrics that dominate it is too naive. For example, for a graph metric, taking a spanning tree will dominate it, but possibly with very large distortion: points at distance 1 in the graph may have distance n in the spanning tree. The idea will be to find a distribution of metrics that dominate the original metric and, at the same time, prevent distances from expanding much (on average).

We start with a metric (X, d) and try to find a family $\{(T_j, \rho_j) \mid j \in J\}$ of tree metrics such that they dominate X, and there exist λ_j, α such that:

1. $\alpha \geq 1, \sum_j \lambda_j = 1$, and
2. $\forall u, v : \alpha d(u, v) \geq \sum_j \lambda_j \rho_j(u, v)$.

In algorithmic applications, this can be used as follows: select one of the metrics $\{(T_j, \rho_j)\}$ (with probability λ_j), and solve the problem using this metric. If the problem at hand can be efficiently approximated for tree metrics to within β, this gives $O(\alpha\beta)$ randomized approximation for the original problem.

The following theorem ensures that we can always approximate a given metric with $\alpha = O(\log n)$.

Theorem 16 (Fakcharoenphol, Rao, and Talwar [17]). *For every n-point metric (X, d) there is a collection of tree metrics $\{(T_j, \rho_j) \mid j \in J\}$, and coefficients $\lambda_j \geq 0$, where $\sum_{j\in J} \lambda_j = 1$, such that*

1. *$\forall j \in J, \rho_j \geq d$, and*
2. *$\sum_{j\in J} \lambda_j \rho_j \leq O(d \log n)$.*

This theorem improves a result of Bartal [4] who proved a similar statement with a bound of $O(\log n \log\log n)$.

Proof. We make the following technical assumptions about the original metric (X, d): Without loss of generality, we assume that

$$\min_{x\neq y} d(x, y) > 1 ,$$

and that

$$\max d(x, y) = \Delta = 2^\delta .$$

For every $j \in J$, the leaves of T_j will consist of the points of the set X. We obtain the tree metrics (T_j, ρ_j) based on a hierarchical partitioning of X into clusters, as follows.

We construct a sequence of decompositions $D_0, D_1, \ldots, D_\delta$ of X, such that $D_\delta = X$ and for each i, D_i is a refinement of D_{i+1}. Each cluster will have a center associated with it: the center is an element of X that does not necessarily belong to the cluster. In the decomposition D_i the radius of each cluster is at most 2^i; thus the diameter of each cluster is at most 2^{i+1}. In D_0 the clusters are singletons since $\min_{x\neq y} d(x, y) > 1$.

Such a sequence $D_0, D_1, \ldots, D_\delta$ defines a rooted tree as follows. The root is labeled by $D_\delta = X$ and nodes at level i of the tree are labeled by the clusters in D_i. For a given node at level $i + 1$ labeled by some cluster S in D_{i+1} its children will be labeled by clusters in D_i that are subsets of S. (The leaves are labeled by singletons.)

The distances in a tree T obtained this way are defined so that $\rho_T(u, v)$ is the sum of the radii of the clusters labeling the nodes along the path from u to v in the tree T. We have then for every u, v:

$$\rho_T(u, v) \geq d(u, v) \ .$$

Thus, the metric (T, ρ) dominates the metric (X, d).

We will use a probabilistic process to define the hierarchical decomposition. The process of selecting the clusters is as follows: Pick a random permutation π of the elements of X. Also pick a random $\beta \in [1, 2]$ according to the probability density function $p(x) = \frac{1}{x \ln 2}$. To define the clusters for a given level i, every point u associates itself with the earliest (according to the ordering π) vertex w, which is at distance $d(u, w) \leq \beta 2^{i-1}$ in the metric (X, d). The points within a given cluster S of D_{i+1} that chose the same w at level i will form a subcluster of S in D_i; w will be the center of this subcluster.

This probabilistic process can be used to define a collection of tree metrics $\{(T_j, \rho_j) \mid j \in J\}$ such that each (T_j, ρ_j) dominates (X, d) and each (T_j, ρ_j) is obtained with some probability λ_j.

It remains to prove the second inequality of the theorem. We will show that for any u, v,

$$\mathrm{E}(\rho_T(u, v)) \leq O(\log n \cdot d(u, v)) \ .$$

We can also define $d_T(u, v)$ as the radius of the least common ancestor of u and v; this makes at most a constant factor difference. We will use this definition for the rest of the proof.

We say that $w \in X$ *separates* u and v at stage i if, before stage i, u and v always selected the same center but, at stage i, one of them (say u) selected w and the other selected a different point.

We denote by lca(u, v) the level of the least common ancestor of u and v in the tree. We have

$$\mathrm{E}(\rho_T(u, v)) \leq \sum_i 2^i Pr[\mathrm{lca}(u, v) = i] \leq \sum_{i,w} 2^i \Pr[w \text{ separates } u, v \text{ at level } i] \ .$$

To estimate the later, order the vertices $w \in X$ according to increasing distances from the pair u, v, that is, $d(w_1, \{u, v\}) \leq d(w_2, \{u, v\}) \leq \ldots$. Let w_s be the s-th vertex in this ordering.

Then w_s separates u and v at level i only if $d(w_s, u) \leq \beta 2^{i-1} < d(w_s, v)$ and w_s is the first among $w_1, \ldots, w_s$ in the ordering π.

Thus, for a given w_s,

$$\sum_i 2^i \Pr[w_s \text{ separates } u, v \text{ at level } i] \leq \sum_i 2^i \frac{1}{s} \Pr[d(w_s, u) \leq \beta 2^{i-1} < d(w_s, v)] \leq \frac{1}{s} \int_{d(w_s,u)}^{d(w_s,v)} \frac{1}{x \ln 2} 8x dx \ .$$

We used that $2^i \leq 8\beta 2^{i-1}$ and that $x = \beta 2^{i-1}$ is selected from $[2^{i-1}, 2^i]$ with density $\frac{1}{x \ln 2}$. Since x cancels out in the last expression, we have

$$\frac{1}{s} \int_{d(w_s,u)}^{d(w_s,v)} \frac{1}{x \ln 2} 8x dx = \frac{1}{s} O(d(w_s, v) - d(w_s, u)) = \frac{1}{s} O(d(u, v)) \ .$$

This gives

$$\mathrm{E}(\rho_T(u, v)) \leq \sum_{s=1}^{n} \frac{1}{s} O(d(u, v)) = O(\log n \cdot d(u, v)) \ .$$

$\square$

27 Embeddings in ℓ_2 and open questions

Given a finite metric space $X = (X, d)$, let $c_2(X, d)$ be the minimum distortion $f : X \hookrightarrow \ell_2$. The following theorem asserts that we can compute c_2 in polynomial time using semidefinite programming (SDP). SDP is a subcase of quadratic programming that have polynomial time algorithm.

Theorem 17. *Given a finite metric space (X, d) and a parameter ϵ, in time poly$(|X|, 1/\epsilon)$ we can find an embedding with distortion that is at most $(1 + \epsilon)c_2(X, d)$.*

Proof. Let $n = |X|$, and assume $X = \{1, ..., n\}$. Let $f : X \hookrightarrow \ell_2$ be an embedding. We may assme f maps X to vectors in $\mathbb{R}^n$, where $f(i) = x_i$. Let the $n \times n$ matrix $X = (x_1, x_2, \ldots, x_n)$. The Gram Matrix $Z = X^T X$ is positive semidefinite, and Z_{ij} is the inner product of x_i and x_j. It is then clear that $c_2(X, d)$ is the square root of the optimal solution of the following semidefinite program:

$$\begin{aligned} \min c \ & \text{ such that} \\ Z \ & \text{ is positive semidefinite,} \\ \forall i, j, \ & d_{ij}^2 \le z_{ii} + z_{jj} - 2z_{ij} \le c \cdot d_{ij}^2 \end{aligned}$$

In polynomial (in the input size and error) time we can approximate the optimal solution arbitrarily close and find X such that $X^T X = Z$ in the (nearly) optimal solution. $\square$

The semidefinite formulation of the problem has another nice consequence. By duality principle of semidefinite programming, we get

Theorem 18. $c_2(X, d) = \max_Q \sqrt{\frac{\sum_{q_{ij}>0} q_{ij} d_{ij}^2}{\sum_{q_{ij}<0} |q_{ij}| d_{ij}^2}}$, *where the maximum is over all the positive semidefinite matrices Q such that $Q\mathbf{1} = \mathbf{0}$ (i.e. the sum of every row is 0).*

Any Q gives a lower bound on the distortion. If you choose the best Q, then you get the best distortion.

Example 5. *Let $X = \{0, 1\}^r$ be the r dimensional cube with Hamming distance d.*

What is an obvious upper bound on $c_2(X, d)$? The simplest thing that comes to mind is the identity mapping. Two points with Hamming distance j have ℓ_2 distance $\sqrt{j}$. So the distortion of the identity mapping is $\sqrt{r}$. Therefore $c_2 \le \sqrt{r}$.

To prove a lower bound, find a positive semidefinite matrix Q satisfying $Q\mathbf{1} = \mathbf{0}$. Let

$$Q = -A + P + (r - 1)I \ ,$$

where all the matrices are indexed by the 2^r points (viewed as subsets of $[r]$) on the r-cube. A is the adjacency matrix of the binary vectors; P is the antipodal matrix that is 0 everywhere except 1 at $(X, \overline{X})$. It is easy to check $Q\mathbf{1} = \mathbf{0}$.

The eigenvalues and eigenvectors of A are well known. The eigenvectors of A are χ_X, where χ_X is the vector indexed by the 2^r subsets with $\chi_{XY} = (-1)^{|X \cap Y|}$. The corresponding eigenvalue is $r - 2|X|$ for each χ_X. These eigenvectors are also eigenvectors of P and I. It is clear they are eigenvectors of Q, and all the eigenvalues are non-negative. Hence, Q is positive semidefinite.

So, using this Q with Theorem 18, an easy calculation shows that $c_2 \geq \sqrt{(r^2 \cdot 2^r)/(r \cdot 2^r)} = \sqrt{r}$. Thus we proved $c_2 = \sqrt{r}$ for the r-cube. This proves that the identity mapping is the best embedding of the cube. Observe that this also implies that the identity is the best embedding of the square in the plane, as claimed earlier.

Open question: What is $\max c_2(X, d)$, where $(X, d) \in \ell_1$ and $|X| = n$? Is the maximum $\Theta(\sqrt{\log n})$, as in the case of the cube?

Theorem 19 (Linial and Magen). *Fix any positive d and δ. Let $G = (V, E)$ be a d-regular graph of order n with the second largest eigenvalue $\lambda_2 \leq d - \delta$ (i.e., G is an expander), then any embedding of G into squared-ℓ_2 has distortion $\Omega(\log n)$.*

Recall that this contains ℓ_1. Every ℓ_1 metric is in squared-ℓ_2. In the theorem, the unspecified coefficient depends on d and δ.

Proof sketch. As an easy exercise, one can show there exists a pairing of the vertices such that x is paired with y then the $\text{dist}_G(x, y) \geq c \log n$ for some constant c.

The construction of Q is an analogy to the previous example. Let P be the (symmetric) permutation matrix that gives this pairing. Again A is the adjacency matrix of the graph. Consider $Q = -A + dI + \frac{\delta}{2}(-I + P)$. The sum of the rows is zero and the matrix can easily be seen to be positive semidefinite by spectral assumptions.

By applying the semidefinite programming duality principle, we get the desired result. □

Open question: Find new classes of metrics with large $c_2, c_1, c_p, \ldots$

Trees can be embedded into ℓ_2 with relatively low distortion. For every tree metrix T with n points, $c_2(T) \leq O(\sqrt{\log \log n})$. The bound is tight as achieved by complete binary trees. The following result shows that the metric spaces of weighted planar graphs have good embeddings into ℓ_2.

Theorem 20 (Rao). *Any planar metric with n points embeds into ℓ_2 with only distortion $O(\sqrt{\log n})$.*

This is shown to be tight by Newman and Rabinovich who use the following recursive construction displayed in Figure 12.

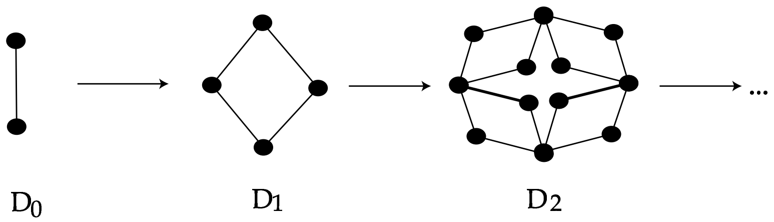

Figure 12: Newman and Rabinovich's diamond graphs

We can't embed this with better than $\sqrt{\log n}$ distortion.

One might think about what is special about planarity.

Rao's theorem works for any minor closed class. It used a structure theorem of any non-trivial minor closed family of graphs by Klein, Plotkin, and Rao.

Open question: Which graph metrics are not isometrically embedded into a planar metric?

One can show that the metric of the unweighted graph K_5 can be embedded into a planar metric. But the metric of unweighted $K_{3,3}$ is not.

28 Embeddings in ℓ_1

We know much less about embeddings into ℓ_1. Trees embedded isometrically into ℓ_1. For planar graphs we have the following open problem.

Conjecture 2. *(Rabinovich and Linial) There exist c such that every planar metric embeds into ℓ_1 with distortion at most c. Even more, the same should hold for any nontrivial minor closed family.*

Work by Newman, Rabinovich, Gupta, and Sinclair proved a special case of this general conjecture.

Recall: Johnson and Lindenstrauss show that for all $x_1, ..., x_n \in \mathbb{R}^n$, there exists $y_1, ..., y_n \in \mathbb{R}^{o(\frac{\log n}{\epsilon^2})}$ such that

$$1 + \epsilon \geq \frac{||y_i - y_j||}{||x_i - x_j||} \geq 1 \quad .$$

Is this also true in ℓ_1? The answer is no. There is no such dimension reduction in ℓ_1. This is shown by the Newman-Rabinovich counterexample shown in figure 12. It is planar and embeds in ℓ_1 with constant distortion. Brinkman and Charikar showed that this doesn't have a dimension reduction.

Theorem 21. *The diamond D_n embeds in ℓ_1^k with distortion D implies $k > \Omega(n^{1/D})$.*

Lee and Naor offered a shorter proof (1.5 pages). They looked at $\ell_{1+\epsilon}$ ($\epsilon \to 0$).

Open question: Is there a dimension reduction in other ℓ_p? ($p \neq 1, 2, \infty$)

Bartal, Linial, Mendel, Naor asked the following Ramsey-type question.

Open problem: Given $n, \alpha > 1$, find the largest $m = m(n, \alpha)$ such that every n point space contains a subset of m points that it embeds into ℓ_2 with distortion at most α.

The following question was asked by Bourgain, Figiel, Milman. Asked again in an online context by Karloff, Rabani, and Ravi. For all $\alpha > 1$, $m(n, \alpha)$ is $\Omega(\log n)$. Is this tight for $\alpha < 1 + \epsilon_0$, for some ϵ_0? Bartal, Linial, Mendel, and Noar have the following results.

Theorem 22 (Bartal, Linial, Mendel, and Noar). *With notation as above:*

1. *This is tight up to $\alpha < 2$.*
2. *For $\alpha > 2$, $m(n, \alpha) \geq n^{c(\alpha)}$ with $c(\alpha) > 0$ and $c(\alpha) \to 1$ as $\alpha \to \infty$.*

29 Approximate nearest neighbor problem

Consider a mapping $f : (x, d) \hookrightarrow (Y, \rho)$ where (Y, p) is some nice space. The mapping f is called *monotone* if for any points x_1, x_2, x_3, x_4 in X,

$$d(x_1, x_2) > d(x_3, x_4) \Rightarrow \rho(f(x_1), f(x_2)) > \rho(f(x_3), f(x_4)) \quad .$$

An easy fact is that every n-point metric can be monotonically mapped into ℓ_2^n. Consider the $n - 1$-dimensional simplex with n points, we can perturb the points so that $||x_1 - x_j|| = 1 + \epsilon_{ij}$ where the order of ϵ_{ij} is the same as d_{ij}.

The metric is really permutations over the $\binom{n}{2}$ numbers. Using the Milnor-Thom theorem from Algebraic Geometry, it is not hard to show that almost every permutation requires $\Omega(n)$ dimensions to get a monotone embedding.

Open problem: Find concrete instances where $\Omega(n)$ dimensions are required in any monotone embedding.

Note, by Bilu and Linial we can find examples of n point metrics which cannot be mapped monotonically into any nomred space of dimension $o(\log n)$. Surely this can be improved a lot.

References

[1] R. Ravi A. Blum, G. Konjevod and S. Vempala. Semi-definite relaxations for minimum bandwidth and other vertex ordering problems. *Theoretical Computer Science*, 235:25–42, 2000.

[2] N. Alon, S. Hoory, and N. Linial. The moore bound for irregular graphs. *Graphs Combin.*, 18:53–57, 2002.

[3] E. Bannai and T. Ito. On moore graphs. *Journal of Faculty of Sciences, University Tokyo*, pages 191–208, 1973.

[4] Y. Bartal. On approximating arbitrary metrics by tree metrics. *STOC98*, pages 161–168, 1998.

[5] N. L. Biggs. *Algebraic Graph Theory*. Cambridge University Press, Cambridge, England, 2nd edition, 1993.

[6] J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel J. Math.*, 52:46–52, 1985.

[7] J. Bourgain. The metrical interpretation of superreflexivity in banach spaces. *Israel J. Math.*, 56:222–230, 1986.

[8] R. M. Damerell. On moore graphs. In *Proc. Cambridge Philos. Soc.*, volume 74, pages 227–236, 1973.

[9] A. Dvoretsky. Some results on convex bodies and banach spaces. In *Proc. Int. Symp. Linear Spaces, LNCS 1295*, pages 223–226, 1959.

[10] A. Dvoretsky. A theorem on convex bodies and application to banach spaces. In *Proc. Natl. Acad. Sci. USA*, pages 123–160, 1961.

[11] P. Enflo. On the nonexistence of uniform homeomorphisms between l_p-spaces. *Ark. Mat.*, 8:103–105, 1969.

[12] U. Feige. Approximating the bandwidth via volume respecting embeddings. *JCSS*, 60(3):510–539, 2000.

[13] W. Feit and G. Higman. The non-existence of certain generalized polygons. *Journal of Algebra*, 1:114–131, 1964.

[14] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems. *J. ACM*, 42:1115–1145, 1995.

[15] J. Håstad. Some optimal inapproximability results. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pages 1–10, 1997.

[16] A. J. Hoffman and R. R. Singleton. On moore graphs of diameter 2 and 3. *IBM J. Res. Develop.*, 4:497–504, 1960.

[17] K. Talwar J. Fakcharoenphol, S. Rao. A tight bound on approximating arbitrary metrics by tree metrics. *STOC03*, pages 448–455, 2003.

[18] F. John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays, presented to R. Courant on his 60th birthday January 8, 1948*, pages 187–204. Interscience Publishers, Inc., New York, N.Y., 1948. Reprinted in J. Moser (editor): Fritz John, Collected Papers, Volume 2, Birkhauser, Boston, Massachusetts, 1985, pages 543-560.

[19] F. Lévy. *Problèmes concrets d'analyse fonctionnelle*. Gauthier Villars, Paris, 1951.

[20] N. Linial, E. London, and Yu. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15:215–245, 1995.

[21] N. Linial, A. Magen, and A. Naor. Girth and euclidean distortion. *GAFA*, 12:380–394, 2002.

[22] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[23] S. Mahajan and H. Ramesh. Correctly derandomizing goemans and williamson's max-cut algorithm. Manuscript, 1995.

[24] J. Matoušek. On the distortion required for embedding finite metric spaces into normed spaces. *Isral J. Math.*, 93:333–344, 1996.

[25] J. Matoušek. On embedding trees into uniformly convex banach spaces. *Israel J. Math.*, 114:221–237, 1999.

[26] J. Matoušek. *Lectures on Discrete Geometry*. Number 212 in Graduate Texts in Mathematics. Springer, 2002.