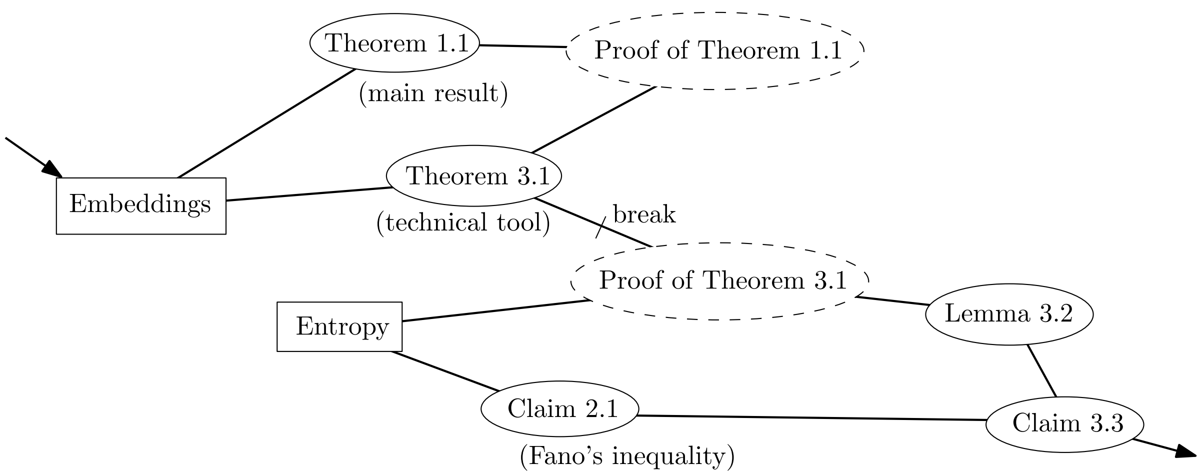
Entropy-based Bound on Dimension Reduction in L_1

Oded Regev

presented by Tomáš Gavenčiak



1 Embeddings

For convenience, let $[k]$ denote $\{1, \ldots, k\}$ and $U(S)$ an uniform distribution over S. All log's are base 2.

Let (X, d_X) and (Y, d_Y) be (possibly finite) metric spaces. We consider metrics of shortest distances in an (undirected) graph and L_1 metric in $\mathbb{R}^d$ (denoted l_1^d).

A mapping $f : X \to Y$ of metric spaces is called an *embedding* with *distortion* $C > 0$) if

$$Sd_X(x, y) \leq d_Y(f(x), f(y)) \leq CSd_X(x, y)$$

for some constant $S > 0$ (scaling factor).

2 Main result

Theorem 1.1 (Main result)

1. (large distortion) For every N, there is an N-point subset of L_1 such that for every $D > 1$, embedding it into l_1^d with distortion D requires $d \geq N^{\Omega(1/D^2)}$.

2. (small distortion) For every N, and every $\epsilon > 0$, there is an N-point subset of L_1 such that embedding it into l_1^d with distortion $1 + \epsilon$ requires $d \geq N^{1-O(1/\log(1/\epsilon))}$.

The main technical tool used in the proof is the following theorem:

Theorem 3.1. For any $k \geq 2$, $n \geq 1$ the following holds. Assume $f : [2k]^n \to \mathbb{R}^d$ and $\epsilon < 1/(k-1)$ satisfy:

1. For all $x_1, \ldots, x_n \in [2k]$, $\|f(x_1, \ldots, x_n)\|_1 \leq 1$

2. For all $l \in [n]$, $x_1, \ldots, x_{l-1} \in [2k]$, and $r \in [k-1]$,

$$\frac{1}{2k}\Bigg\| \sum_{b=1}^{r}(f(x_1, \ldots, x_{l-1}, b) + f(x_1, \ldots, x_{l-1}, b+k)) - \sum_{b=r+1}^{k}(f(x_1, \ldots, x_{l-1}, b) + f(x_1, \ldots, x_{l-1}, b+k))\Bigg\|_1 \geq 1 - \epsilon$$

where $f(x_1, \ldots, x_l)$ denotes the average of $f(x_1, \ldots, x_n)$ over $x_{l+1}, \ldots, x_n \in [2k]$.
Then $d \geq 2^{(\log k - \delta \log(k-1) - H(\delta))n-1} - 1/2$, where $\delta = (k-1)\epsilon/2 < 1/2$.

3 Embedded space

The theorem is applied on an L_1 metric space which is an ambedding of $G_{k,n}$, which can be L_1-embedded thanks to the following:

Theorem 4.1 [GNRS04]. Any (weighted) series-parallel graph can be embedded into L_1 with distortion at most 14. Moreover, the lengths of the edges can be preserved.

$G_{k,n}$ is defined as follows:

Let $G_{k,1}$ be a C_{2k} with edges labeled $1 \ldots 2k$, distinguished vertices *left* (between edges 1 and $2k$) and *right* (between edges k and $k+1$) and all edges oriented left-to-right.

Let $G_{k,n+1}$ be a $G_{k,n}$ with each edge with label l replaced with a copy of $G_{k,1}$ with the edge labels prefixed by l.

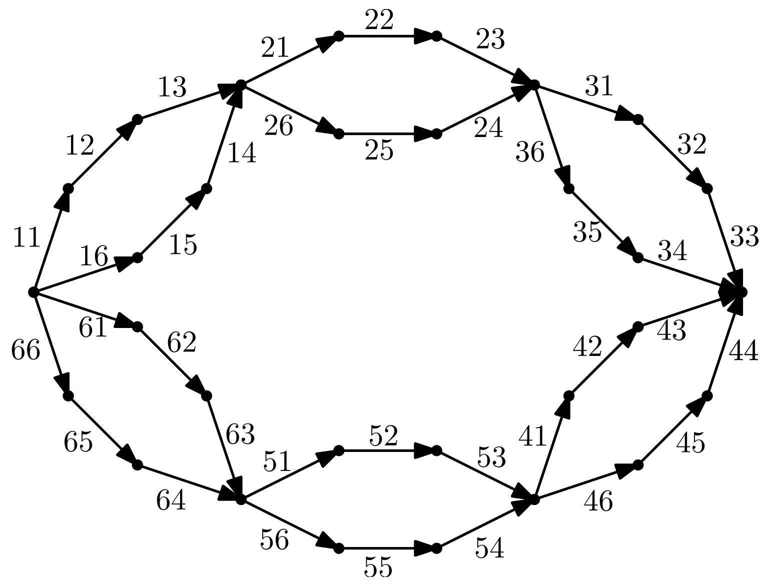


Figure 1: Graph $G_{3,2}$ (from the paper)

Let $F : G_{k,n} \to l_1^d$ be a non-expanding embedding. For $xy \in E(G_{k,n})$, define $f(xy) = F(x) - F(y)$.

Theorem 1.1 is proven by showing that f satisfies Theorem 3.1 and choosing apropriate k, n, ϵ.

4 Entropy

Entropy of a discrete random variable X is $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$. For convenience, let $H(p) = -p \log p - (1-p) \log(1-p)$ denote the entropy of a coin flip with probabilities p and $1-p$.

Conditional entropy $H(X|Y)$ is $\mathbb{E}H(X|Y=y)$ over y chosen according to Y. $H(X|Y) = H(XY) - H(Y)$.

Mutual information is defined as $I(X:Y) = H(X) + H(Y) - H(XY) = H(X) - H(X|Y)$, conditional mutual information $I(X:Y|Z)$ as $\mathbb{E}I(X:Y|Z=z)$ with z distributed as Z.

Data processing inequality: $I(f(X):Y) \leq I(X:Y)$.

Chain rule for entropy: $H(XY) = H(X) + H(Y|X)$.

Chain rule for mutual information: $I(XY:Z) = I(X:Z) + I(Y:Z|X)$.

Claim 2.1. **(Fano's Inequality)** Assume $X \sim U([k])$, Y arbitrary and that there is $f : Y \to X$ such that $P[f(Y) = X] = p \geq 1/2$. Then $I(X : Y) \geq \log k - (1-p)\log(k-1) - H(p)$.

5 Proof of main technical tool

The proof uses the following lemma and the lemma uses the claim below.

The lemma applies to any situation i T3.1 with fixed $X_1, \ldots, X_{l-1}$, $A = X_l$ and $B = \mathbb{E}_{X_{l+1},\ldots,X_n} M$

Lemma 3.2. Let A and B be two random variables such that A is uniformly distributed over $[2k]$ and for any $a \in [2k]$. Conditioned on $A = a$, B is distributed according to some probability distribution P_a on $[d]$.

Assume that for all $r \in [k-1]$,

$$\frac{1}{2k}\left\| \sum_{a=1}^{r}(P_a + P_{a+k}) - \sum_{a=r+1}^{k}(P_a + P_{a+k}) \right\|_1 \geq 1 - \epsilon$$

Then $I(A : B) \geq \log k - \delta \log(k-1) - H(\delta)$.

Claim 3.3. For any $p_1, \ldots, p_k \geq 0$,

$$\left(\sum_{i=1}^{k} p_i\right) - \max\{p_1, \ldots p_k\} \leq \frac{1}{2}\sum_{r=1}^{k-1}\left(\left(\sum_{i=1}^{k} p_i\right) - \left|\sum_{i=1}^{r} p_i - \sum_{i=r+1}^{k} p_i\right|\right).$$