Massachusetts Institute of Technology Michel X. Goemans
18.409: Topics in TCS: Embeddings of Finite Metric Spaces
November 6, 2006 Presentation and Notes by Alexandr Andoni

In this lecture, we will prove that embedding edit metric over $\{0,1\}^d$ into l_1 requires $\Omega(\log d)$ distortion, following the proof of [KR06]. We will start with the definitions.

The *edit metric* is a metric on $\{0,1\}^d$, where for two points $x, y \in \{0,1\}^d$, we define their edit distance, $ed(x,y)$, to be the minimum number of *edit operations* to transform one string into the other. The edit operations are character substitution, insertion, or deletion. For example strings $(10)^3 = 101010$ and $(01)^3 = 010101$ are at distance 2 (to obtain the second string from the first string, delete the first 1 and insert a 1 at the end). One can view the edit metric as the shortest path metric on the d-dimensional hypercube with some additional "shortcuts" (in addition to the hypercube edges, there is, for example, also an edge $((10)^3, (01)^3)$).

Definition 1. *We call $c_1(\{0,1\}^d, ed)$ to be the minimum distortion required to embed edit metric over $\{0,1\}^d$ into l_1. I.e., $c_1(\{0,1\}^d, ed)$ is the minimum D such that there exists a mapping $\phi : \{0,1\}^d \to l_1$ such that for any $x, y \in \{0,1\}^d$,*

$$ed(x,y) \le \|\phi(x) - \phi(y)\|_1 \le D \cdot ed(x,y)$$

In this lecture we prove the following theorem:

Theorem 1 ([KR06])**.** $c_1(\{0,1\}^d, ed) = \Omega(\log d)$.

For completeness, we mention that before [KR06], the previous lower bound was proven by Subhash Khot and Assaf Naor [KN05], who showed that $c_1(\{0,1\}^d, ed) = \Omega\left((\log d)^{1/2-o(1)}\right)$ using Fourier-analytic approach. The best upper bound on $c_1(\{0,1\}^d, ed)$ is[1] $2^{\tilde{O}(\sqrt{\log d})}$, proven by Rafail Ostrovsky and Yuval Rabani [OR05].

Open question 1. *Bridge the gap between $c_1(\{0,1\}^d, ed) \ge \Omega(\log d)$ and $c_1(\{0,1\}^d, ed) \le 2^{\tilde{O}(\sqrt{\log d})}$.*

1 Proof of the main theorem

As was mentioned earlier in this class, it is sufficient to exhibit two distributions τ and η on $\{0,1\}^d \times \{0,1\}^d$ such that

1. $$\sum_{x,y} \tau(x,y) \cdot ed(x,y) \le \alpha \sum_{x,y} \eta(x,y) \cdot ed(x,y)$$

2. for any boolean function $f : \{0,1\}^d \to \{0,1\}$, it holds that

$$\sum_{x,y} \tau(x,y) \cdot |f(x) - f(y)| > \beta \sum_{x,y} \eta(x,y) \cdot |f(x) - f(y)|$$

Then, $c_1(\{0,1\}^d, ed) \ge \beta/\alpha$.

We construct τ and η as the following probability distributions (i.e., $\sum \tau(x,y) = \sum \eta(x,y) = 1$):

[1] Notation $\tilde{O}(f(n))$ means $O(f(n) \cdot (\log f(n))^{O(1)})$.

- Distribution $\tau(x,y)$ (close pairs, or "edges"). Define the following *shift operation* $S : \{0,1\}^d \to \{0,1\}^d$: $S(x_1x_2 \ldots x_d) = x_d x_1 x_2 \ldots x_{d-1}$. Then let $E_S = \{(x, S(x)) \mid x \in \{0,1\}^d\}$, and τ_S is the uniform distribution over E_S. Also, let E_H be the set of edges in the hypercube: $E_H = \{(x,y) \mid \|x-y\|_1 = 1\}$. τ_H is the uniform distribution over E_H.

  Then $\tau(x,y) = \frac{\tau_S(x,y)+\tau_H(x,y)}{2}$.

- Distribution $\eta(x,y)$ (far pairs, or "diagonals") is defined to be simply uniform over all pairs (x,y).

We then prove the following two lemmas, which imply that $c_1(\{0,1\}^d, ed) = \Omega(\log d)$.

Lemma 2.

$$\mathbb{E}_\tau [ed(x,y)] \leq O\left(\frac{1}{d}\right) \cdot \mathbb{E}_\eta [ed(x,y)]$$

Lemma 3. *For any boolean function $f : \{0,1\}^d \to \{0,1\}$, we have that*

$$\mathbb{E}_\tau [|f(x) - f(y)|] > \Omega\left(\frac{\log d}{d}\right) \cdot \mathbb{E}_\eta [|f(x) - f(y)|]$$

The second lemma is the most technical part of the proof and is proven/discussed in the next section. We prove below the first lemma:

Proof of lemma 2. First we claim that $\mathbb{E}_\tau [ed(x,y)] \leq 2$. This results from the fact that for any $(x,y) \in E_S \cup E_H$, $ed(x,y) \leq 2$.

Second, we claim that $\mathbb{E}_\eta [ed(x,y)] \geq \Omega(d)$. Fix any $x \in \{0,1\}^d$. Let's upper bound the number $N_{x,l}$ of strings y that satisfy $ed(x,y) \leq l$. Note that for any pair (x,y), we can assume that we perform first the deletions on x, then the insertions, then all the substitutions. Thus,

$$N_{x,l} \leq \binom{2d}{l} \cdot \binom{2d}{l} 2^l \cdot \binom{2d}{l} \leq 2^l \cdot \left(\frac{2de}{l}\right)^{3l}$$

For $l = d/100$, we get that

$$N_{x,d/100} \leq 2^{d/100} \cdot (200e)^{3d/100} \leq 2^{d/2}$$

Finally,

$$\mathbb{E}_\eta [ed(x,y)] = \mathbb{E}_x [\mathbb{E}_y [ed(x,y)]] \geq \mathbb{E}_x \left[(1 - N_{x,d/100} 2^{-d}) \cdot (d/100)\right] = \Omega(d)$$

□

2 Proof of lemma 3

To prove this lemma, we will use a deep theorem about boolean functions $f : \{0,1\}^d \to \{0,1\}$. The theorem is that of Kahn-Kalai-Linial [KKL88]. We will not prove the KKL theorem in this lecture. If you are interested in the proof this theorem, see [KKL88] for the original proof (using a Fourier-analytic approach), or, for example, [FSar] (and references therein) for alternative proofs (more combinatorial).

Consider a function $f : \{0,1\}^d \to \{0,1\}$. We define the *influence* of a variable as follows:

Definition 2. *For $i \in [d]$, call the* influence *of the i^{th} variable the quantity:*

$$Inf_i(f) = \Pr_{x \in \{0,1\}^d}[f(x) \neq f(x \oplus e_i)]$$

where e_i is the vector with 1 in the i^{th} position and 0 otherwise; $\oplus$ is the operation of coordinate-wise sum modulo 2.

Why "influence"? Imagine the following voting procedure. There are n players $x_1, x_2 \dots x_n$ with binary inputs (0 or 1), participating in a referendum. One can view the voting procedure as a function f from their inputs, $\{0,1\}^d$, to the outcome of the referendum, $\{0,1\}$. For example:

- In a democracy, the function is a *majority*: $f(x_1 \dots x_d) = 1$ iff $\sum_i x_i \geq d/2$ (assume d is odd, and ignore vote rigging). We call such function $f = \mathbf{Maj}$.
- The function could be a *dictatorship*, when $f(x_1 \dots x_d) = x_i$, i.e., exactly one person (i^{th}) establishes the outcome of the referendum.

Now, influence $Inf_i(f)$ is the probability that the i^{th} player has an influence on the result of the referendum after all the other players have fixed their value to random values. For example:

- In the majority, everybody has the same influence. $Inf_1(\mathbf{Maj})$ is precisely the probability that $\sum_{i=2}^d x_i = (d-1)/2$, which is roughly $\Theta(1/\sqrt{d})$. Thus $Inf_i(\mathbf{Maj}) = \Theta(1/\sqrt{d})$ for all $i \in [d]$.
- In a dictatorship $f(x) = x_i$, $Inf_i(f) = 1$ and $Inf_j(f) = 0$ for $j \neq i$.

KKL theorem roughly answers the following question: how small can be the largest influence, i.e., what is $\min_f \max_i Inf_i(f)$? Note that for a constant function $f(x) = 0$, all influences are zero, so the above question is trivial. But the question becomes much more non-trivial when we require the function f to be balanced ($Pr_x[f(x) = 0] = 1/2$). A relatively simple combinatorial arguments shows that $\sum_i Inf_i(f) \geq \Omega(1)$ for all balanced f (therefore, the max influence is at least $\Omega(1/d)$).

There is a function called *tribes function*, that obtains $Inf_i(f) = \Theta(\frac{\log d}{d})$ and is roughly balanced. The function is like this. Partition d players into t tribes, each of size $\log t$ (t satisfies $t \log t = d$). Note that $t \approx d/\log d$. Let the partition be $[d] = S_1 \cup S_2 \dots S_t$, where S_i is the i^{th} tribe. Then $f(x) = \vee_{i=1}^t \wedge_{j \in S_i} x_j$.

KKL theorem proves that the tribes function is essentially optimal:

Theorem 4 ([KKL88])**.** *Let $\mu = \Pr_x[f(x) = 1]$. There exists some constant $C > 1$ such that*

$$\max_i Inf_i(f) \geq C\mu(1-\mu)\frac{\log d}{d}$$

In our proof, we will need a slightly stronger theorem than the above one (although, it is possible to modify the proof in [KKL88] to obtain a similar stronger bound – see [KR06] for this).

Theorem 5 ([Tal94])**.** *Let $\mu = \Pr_x[f(x) = 1]$. There exists some constant $C > 1$ such that*

$$\sum_i \frac{Inf_i(f)}{\log(e/Inf_i(f))} \geq C\mu(1-\mu)$$

Note that theorem 5 implies theorem 4.

Ok, we can finally prove lemma 3.

Proof of lemma 3. Suppose, wlog, $\mu = \Pr_x[f(x) = 1] \leq 1/2$ (otherwise, invert the function). Note that $\mathbb{E}_\eta\left[|f(x) - f(y)|\right] = \mu(1-\mu) \leq \mu$.

Assume, for contradiction, that, for any small $c > 0$, $\mathbb{E}_\tau\left[|f(x) - f(y)|\right] \leq \frac{c \log d}{d} \mathbb{E}_\eta\left[|f(x) - f(y)|\right] \leq \frac{c\mu \log d}{d}$.

Then,

$$\sum_i Inf_i(f) = d\mathbb{E}_{\tau_H}\left[|f(x) - f(y)|\right] \leq 2d\mathbb{E}_\tau\left[|f(x) - f(y)|\right] < 2c\mu \log d.$$

By theorem 5, there exists some j_0 such that $Inf_{j_0}(f) \geq d^{-1/8}$ (otherwise, $\sum Inf_i / \log(e/Inf_i) < O(\sum Inf_i / \log d) = O(\mu)$).

We will prove that there are at least $d^{1/4}$ other variables with big influences, concluding that $\sum Inf_i(f) = \Omega(d^{1/8})$, a contradiction.

First, note that

$$Pr_x[f(x) \neq f(S(x))] = \mathbb{E}_{\tau_S}\left[|f(x) - f(S(x))|\right] \leq 2\mathbb{E}_\tau\left[|f(x) - f(y)|\right] \leq \frac{2c\mu \log d}{d}$$

For any $k \in \{1, \ldots d^{1/4}\}$, we have that (define $Inf_z(f) = Inf_{z-n}(f)$ if $z > n$):

$$Inf_{j_0+k}(f) = \Pr_x[f(x) \neq f(x \oplus e_{j_0+k})] = \mathbb{E}_x\left[|f(x) - f(x \oplus e_{j_0+k})\right] \leq$$

$$\mathbb{E}_x\left[|f(x) - f(S(x))|\right] + \mathbb{E}_x\left[|f(S(x)) - f(S(x \oplus e_{j_0+k}))\right] + \mathbb{E}_x\left[|f(S(x \oplus e_{j_0+k})) - f(x \oplus e_{j_0+k})|\right] =$$

$$\mathbb{E}_x\left[|f(S(x)) - f(S(x) \oplus e_{j_0+k+1})\right] + 2\mathbb{E}_x\left[|f(x) - f(S(x))\right] = Inf_{j_0+k+1} + 2\Pr[f(x) \neq f(S(x))]$$

Therefore, $Inf_{j_0+k+1} \geq Inf_{j_0+k} - 2\frac{2c\mu \log d}{d}$. Since $Inf_{j_0} \geq n^{-1/8}$, we have that $Inf_{j_0+k} \geq d^{-1/8} - \frac{ck\mu \log d}{d} \geq d^{-1/8}/2$ for $k \in [d^{1/4}]$.

In total, we have that $\sum_i Inf_i \geq d^{-1/8}/2 \cdot d^{1/4} = d^{1/8} > c\mu \log d$, a contradiction. ☐

Remark 1. *Lemma 3 is tight for the following function f. Fix some $k = \Theta(\log d)$. Then $f(x) = 1$ iff x contains 0^k as a substring (allowing the string to wrap-around in x). Then, the function has balance $\mu \in [1/10, 9/10]$, and $\mathbb{E}_{\tau_H}\left[|f(x) - f(y)|\right] \leq O(k/d)$, whereas $\mathbb{E}_{\tau_S}\left[|f(x) - f(y)|\right] = 0$.*

References

[FSar] Dvir Falik and Alex Samorodnitsky. Edge-isoperimetric inequalities and influences. In *Combinatorics, Probability, and Computing*, to appear.

[KKL88] J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *Proceedings of the Symposium on Foundations of Computer Science*, pages 68–80, 1988.

[KN05] Subhash Khot and Assaf Naor. Nonembeddability theorems via fourier analysis. In *FOCS '05: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 101–112, Washington, DC, USA, 2005. IEEE Computer Society.

[KR06] Robert Krauthgamer and Yuval Rabani. Improved lower bounds for embeddings into l_1. In *SODA'06: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pages 1010–1017, New York, NY, USA, 2006. ACM Press.

[OR05] Rafail Ostrovsky and Yuval Rabani. Low distortion embeddings for edit distance. In *Proceedings of the Symposium on Theory of Computing*, 2005.

[Tal94] Michel Talagrand. On Russo's approximate 0-1 law. *Ann. Probab.*, 22(3):1576–1587, 1994.