CS 761 : Randomized Algorithms , Spring 2018 , Waterloo

Lecture 19 : Metric embedding

We give a short introduction to this topic and study two fundamental results.

One is the Johnson-Lindenstrauss theorem about dimension reduction.

Another is Bourgain's theorem of embedding into L_1-metric, which has surprising application to finding sparsest cuts.

Metric embedding

A metric space is a set of points V, with a distance function $d: V \times V \to \mathbb{R}_{\geq 0}$ that satisfies :

① $d(x,y) = 0$ iff $x = y$.

② $d(x,y) = d(y,x) \quad \forall x,y \in V.$ // symmetry

③ $d(x,y) + d(y,z) \geq d(x,z) \quad \forall x,y,z \in V$ // triangle inequality

It is called a pseudo-metric if the first property is replaced by $d(x,x) = 0 \quad \forall x \in V$.

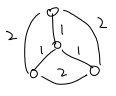
The following are some examples of distance functions :

- L_2-norm : $d(x,y) = \|x-y\|_2 = \sqrt{\sum_i (x_i - y_i)^2}$.
- L_1-norm : $d(x,y) = \|x-y\|_1 = \sum_i |x_i - y_i|$.
- L_∞-norm : $d(x,y) = \|x-y\|_\infty = \max_i |x_i - y_i|$.
- graph metric : each point is a vertex of a graph, and $d(x,y)$ is the shortest path distance from x to y.
- tree metric : a special case of graph metric where the underlying graph is a tree.
- edit distance : each point x is a string, and $d(x,y)$ is the edit distance between x and y.
- resistance distance : each point is a vertex of a graph, and $d(x,y)$ is the effective resistance between x and y.

Given a (general) metric space, sometimes we want to "embed" it into another (simpler) metric space so as to better understand the structure.

Ideally, the distances are preserved after the embedding, but it is not always possible.

For example, the metric in the picture cannot be realized in any Euclidean space.



The next best thing is an embedding with "low distortion".

Definition A mapping $f: X \to Y$, where X is a metric space with d_X and Y a metric space with d_Y, is called an α-embedding where $\alpha \geq 1$, if $\exists r > 0$ such that for all $a,b \in X$,

$$r \cdot d_X(a,b) \le d_Y(f(a),f(b)) \le \alpha \cdot r \cdot d_X(a,b).$$

The infimum of the number α is called the distortion of f.

Note that we can choose r to scale up / down $d_X(a,b)$ for all $a,b \in X$.

For example, we can set $r=1$ so that the metric Y does not "shrink" the distance of any pair,

or we can set $r=\frac{1}{\alpha}$ so that the metric Y does not "stretch" the distance of any pair.

Some basic results:

- Every metric can be isometrically embedded (i.e. no distortion, $\alpha=1$) into L_∞. (Exercise)
- Every tree metric can be isometrically embedded into L_1.
- L_2 can be isometrically embedded into L_1 (nontrivial, see [Matousek]).

Today we will try to show and mention:

- Every L_2-metric can be almost isometrically embedded into L_2 with $O(\log n)$ dimensions.
- Every metric can be embedded into L_2 with $O(\log n)$ distortion.
- Every metric can be embedded into a family of tree metrics with expected $O(\log n)$ distortion.

There are many applications of these theorem, and we will discuss some of them.

Dimension reduction

Given n points in the Euclidean space, we can always represent the vectors in n-dimensions.

In general, we cannot do better if no distortion is allowed.

Surprisingly, if we allow just a little distortion, then the number of dimensions can be significantly reduced.

Theorem (Johnson-Lindenstrauss Lemma) Given any $\varepsilon \in (0, \frac{1}{2})$ and any set of points $X=\{x_1, x_2, \ldots, x_n\}$,

there exists a map $A: X \to \mathbb{R}^k$ for $k = O(\frac{\log n}{\varepsilon^2})$ such that

$$1-\varepsilon \le \frac{\| A(x_i) - A(x_j) \|_2^2}{\| x_i - x_j \|_2^2} \le 1+\varepsilon.$$

Algorithm: The construction is very simple. It just projects the points in a random k-dimensional subspace.

Let d be the original dimension of the points.

Let M be a $k \times d$ matrix, such that each entry of M is drawn from the normal $N(0,1)$ distribution

.2,

(Gaussian random variable with mean 0 and variance 1, with density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2}$.)

Define $A(x) = \frac{1}{\sqrt{k}} Mx$. This is efficiently computable.

Since A is a linear transformation ($A(x)+A(y) = A(x+y)$), the theorem can be reduced to the following.

Lemma If A is constructed by the above algorithm with $k = \Theta\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,

then $\Pr\left(1-\varepsilon \le \|A(x)\|_2^2 \le 1+\varepsilon\right) \ge 1-\delta$ for any unit vector $x \in \mathbb{R}^d$ and any $\varepsilon \in (0, \frac{1}{2})$.

First, we see how the lemma implies the theorem.

We set $\delta = \frac{1}{n^2}$ and thus $k = \Theta\left(\frac{\log n}{\varepsilon^2}\right)$.

For any pair $i \neq j$, the squared length of $x_i - x_j$ is maintained to within $1 \pm \varepsilon$ with probability $\ge 1 - \frac{1}{n^2}$.

By the union bound, the distances of all pairs are maintained to within $1 \pm \varepsilon$ with probability $\ge \frac{1}{2}$.

Henceforth, we focus on proving the lemma.

Proof idea: Consider the elementary unit vector $e_1 = (1, 0, \ldots, 0)$.

Then, Me_1 is just the first column of M, with independent and identical Gaussian values.

We are interested in the length of this column, which is the sum of squares of these Gaussians.

Note that $E\left[\sum_{i=1}^{k} M_{i,1}^2\right] = \sum_{i=1}^{k} E[M_{i,1}^2] = k$ as the variance of each M_{ij} is one, and so the expected length of Ae_1 is one as $E[\|Ae_1\|_2^2] = \frac{1}{k} E[\|Me_1\|_2^2] = 1$.

By setting k to be large enough, we expect that the length is highly concentrated around its expectation.

From our intuition of Chernoff bound, if we set $k = O(\frac{\log n}{\varepsilon^2})$, then the error probability is at most $2e^{-\mu\varepsilon^2/3} \le \frac{1}{n^2}$.

Proof The actual proof is similar to the above idea. There are two issues to handle.

① We cannot assume $x = e_1$, and we need to deal with any x.

② The standard Chernoff-Hoeffding bound cannot be directly applied, because the random variables are unbounded (although with a small tail).

The first issue can be taken care of by the nice properties of Gaussian random variables.

Consider an arbitrary entry y_j of the vector Mx for an arbitrary unit vector x.

Then $y_j = \sum_{i=1}^{d} M_{ji} x_i$ where M_{ji} is an $N(0,1)$ random variable.

So, y_j is a sum of Gaussian variable, and it is a well-known fact that y_j is an $N(0, \sum_{i=1}^{d} x_i^2)$

random variable. Since x is a unit vector, y_j is just an $N(0,1)$ random variable.

So, each of the k coordinates of Mx is just independent Gaussian.

By the same argument as in the proof idea, the expected length of $\frac{1}{\sqrt{k}}Mx$ is one.

The second issue requires some work, but we have all done it.

In Q5 of Hw1, we have computed the moment generating function of the sum of squares of independent Gaussians (i.e. $E[e^{tX^2}] = \frac{1}{\sqrt{1-2t}}$ for $t < \frac{1}{2}$ for $X \sim N(0,1)$) and use it to prove that $\Pr(\|Ax\|_2^2 \geq 1+\varepsilon) \leq e^{-k\varepsilon^2/8}$.

Similarly, we can bound the lower tail and get a similar result.

So, by setting $k = O(\frac{1}{\varepsilon^2}\ln(\frac{1}{\delta}))$, we have $\Pr(\left|\|Ax\|_2^2 - 1\right| > \varepsilon) \leq \delta$, proving the lemma. □

Remarks: The same result is true even when M is a random ± 1 matrix [Achioptas].

The proof is more difficult but the algorithm is much easier to implement.

Actually, if we use a random ± 1 matrix, this is very similar to what we did in L07 for estimating the L_2-norm in the data streaming model, expect we used 4-wise independence.

Dimension reduction is not possible for L_1-norm.

Applications: One immediate and important application is to do approximate near neighbor search.

A linear scan takes $\Theta(n^2)$ time, but only $O(n\log n)$ time after dimension reduction.

Note that it works for Euclidean distances only (e.g. not for L_1-distances).

Another application is approximate matrix multiplication.

Given two $n \times n$ matrices A and B, we do dimension reductions on the rows of A and the columns of B to $O(\log n)$ dimensions. So that the product can be done in $O(n^2 \log n)$ time, while each entry is approximately the same as the inner products are approximated with high prob.

Low distortion L_1-embedding

This is a fundamental result in metric embedding.

Theorem (Bourgain's theorem) Every finite metric (V,d) can be embedded into an L_1-metric (V,f) such that $\Omega(\frac{1}{\log n}) \leq \frac{\|f(u)-f(v)\|_1}{d(u,v)} \leq 1$.

Theorem (Bourgain's theorem) Every finite metric (V,d) can be embedded into an L_1-metric (V,f) such that $\Omega\left(\frac{1}{\log n}\right) \le \frac{\|f(u)-f(v)\|_1}{d_{uv}} \le 1$.

Algorithm We assume that $n = 2^\ell$ without loss of generality. Remember that $\ell = \log_2 n$.

- For $2 \le i \le \ell+1$, form S_i by picking each vertex of V independently with probability $\frac{1}{2^i}$.
- Define the i-th coordinate of point v to be $f_i(v) = \frac{1}{\ell} d(v,S_i)$, where $d(v,S_i)$ is the distance between v and its closest point in S_i, i.e. $d(v,S_i) = \min_{u \in S_i} d(v,u)$.

First, we see that the mapping does not stretch any edge, showing the second inequality in the theorem. Since L_1-distances are additive, it is enough to prove the following.

Lemma $|f_i(u) - f_i(v)| \le d(u,v)$ for all $u,v \in V$.

Proof Let S_i be the set and s_u and s_v be the closest vertices of S to u and v.

Assume $d(s_u,u) \ge d(s_v,v)$ without loss of generality.

Then $|f_i(u)-f_i(v)| = d(s_u,u) - d(s_v,v) \le d(s_v,u) - d(s_v,v) \le d(u,v)$ by triangle inequality. □

The main work is to show that an edge is not over-shrunk by too much.

It will depend on whether the "neighborhood" of a point is sampled.

Let $B(x,r) = \{ y \in V \mid d(x,y) \le r \}$ be the ball of radius r around x,

and $B^\circ(x,r) = \{ y \in V \mid d(x,y) < r \}$ be the open ball of radius r around x.

The basic mechanism to establish a lower bound on the expected contribution of the coordinate corresponding to a set S_i is the following.

Claim If for some $r_1 \ge r_2 > 0$ and constant c, $\Pr\left[\left(S_i \cap B(u,r_1) = \phi\right) \text{ and } \left(S_i \cap B(v,r_2) \ne \phi\right) \right] \ge c$, then the expected contribution of S_i is at least $c(r_1-r_2)/\ell$.

Proof If this event happens, then $d(u,S_i) > r_1$ and $d(v,S_i) \le r_2$, and thus $f_i(u) \ge \frac{r_1}{\ell}$ and $f_i(v) \le \frac{r_2}{\ell}$, and hence $|f_i(u)-f_i(v)| \ge (r_1-r_2)/\ell$, proving the claim. □

The probability of this event can be lower bounded if we know the sizes of the set.

Claim Let A and B be disjoint subsets of V, with $|A| < 2^t$ and $|B| \ge 2^{t-1}$ for some t, then $\Pr\left[(S \cap A = \phi) \text{ and } (S \cap B \ne \phi) \right] \ge \frac{1}{2}(1 - e^{-1/4})$ if each vertex is in S with prob $p = \frac{1}{2^{t+1}}$.

$\Pr\left[(S \cap A = \phi) \text{ and } (S \cap B \neq \phi) \right] \geq \frac{1}{2}(1 - e^{-1/4})$ if each vertex is in S with prob $p = \frac{1}{2^{t+1}}$.

Proof $\Pr(S \cap A = \phi) \geq 1 - p|A| \geq \frac{1}{2}$, where the first inequality is by the union bound.

$\Pr(S \cap B \neq \phi) = 1 - (1-p)^{|B|} \geq 1 - e^{-p|B|} \geq 1 - e^{-1/4}$.

Since A and B are disjoint, the two events are independent and the claim follows.

Now, we need to combine both the radius and the size of the set to prove the lower bound.

Let r_t be the smallest radius such that the ball around u and the ball around v each has at least 2^t vertices, i.e. $r_t = \min\{ r \geq 0 \mid |B(u,r)| \geq 2^t \text{ and } |B(v,r)| \geq 2^t \}$.

Clearly, $r_0 = 0$ and $r_\ell \geq d(u,v)$.

Let $\hat{t} = \max\{ t \mid r_t < d(u,v)/2 \}$, up to this t the two balls are still disjoint.

Let $c = \frac{1}{2}(1 - e^{-1/4})$ be the constant probability in one of the above claims.

Lemma The expected contribution of S_{t+1} to uv is at least $\frac{c}{\ell} \cdot (r_t - r_{t-1})$ for $1 \leq t \leq \hat{t}$.

The expected contribution of S_{t+1} to uv is at least $\frac{c}{\ell} \cdot \left(\frac{d(u,v)}{2} - r_{t-1} \right)$ for $t = \hat{t}+1$.

Proof By the definition of r_t, either $|B^\circ(u, r_t)| < 2^t$ or $|B^\circ(v, r_t)| < 2^t$, assume $|B^\circ(u, r_t)| < 2^t$.

Also, by definition, $|B(v, r_{t-1})| \geq 2^{t-1}$. The sets $B^\circ(u, r_t)$ and $B(v, r_{t-1})$ are disjoint.

By the above claims, with probability at least c, we have $S_{t+1} \cap B^\circ(u, r_t) = \phi$ and $S_{t+1} \cap B(v, r_{t-1}) \neq \phi$,

and this implies that the expected contribution of S_{t+1} is at least $\frac{c}{\ell}(r_t - r_{t-1})$.

By the definition of $\hat{t}$, we can assume $|B^\circ(u, d(u,v)/2)| < 2^{\hat{t}+1}$. And also $|B(v, r_{\hat{t}})| \geq 2^{\hat{t}}$.

The lemma follows by the same argument as above. □

Lemma The expected contribution of all sets $S_2, \ldots, S_{\ell+1}$ to uv is at least $\frac{c}{\ell} \frac{d(u,v)}{2}$.

Proof This follows by a telescoping sum: $\frac{c}{\ell}\left((r_1 - r_0) + (r_2 - r_1) + \ldots + \left(\frac{d(u,v)}{2} - r_{\hat{t}} \right) \right) = \frac{c}{\ell} \frac{d(u,v)}{2}$. □

Recall that $\ell = \log_2 n$, the above lemma implies that the expected total contribution to uv is $\Omega\left(\frac{1}{\log n}\right) \cdot d_{uv}$.

By repeating the whole procedure $O(\log n)$ times and taking the average, we can prove that the total contribution is close to its expected value with high probability by Chernoff bound.

Thus, we have an embedding into L_1 with distortion at most $O(\log n)$ using $O(\log^2 n)$ dimensions.

Tight example: The $O(\log n)$ distortion cannot be improved in general.

The shortest path metric of a constant degree expander graph requires $\Omega(\log n)$ distortion to

embed into L_1 (see [Matousek]).

Open question: Is it true that the shortest path metric of a planar graph can always be embedded into L_1 with constant distortion?

Sparsest cut

A surprising application of Bourgain's theorem is in designing a $O(\log n)$-approximation algorithm for the sparsest cut problem.

Given an undirected graph $G=(V,E)$, for each pair of vertices u,v, there is a capacity c_{uv} and a demand $dem(u,v)$ - the problem is to find $S \subseteq V$ that minimizes $\sum_{uv \in \delta(S)} c_{uv} / \sum_{uv \in \delta(S)} dem(u,v)$.

In words, the objective is to find a cut $S \subseteq V$ that minimizes the average cost to disconnecting a pair.

For example, minimizing the conductance of a d-regular graph can be reduced to the sparsest cut problem.

Linear programming relaxation

Let d_{uv} be a binary variable ($d_{uv} \in \{0,1\}$) to denote whether uv is in the cut.

Then the sparsest cut problem is written as $\min_{d} \frac{\sum_{uv} c_{uv} d_{uv}}{\sum_{uv} dem(u,v)\, d_{uv}}$, but this is NP-hard.

Note that the binary solution will satisfy the triangle inequality.

By relaxing the constraints $d_{uv} \in \{0,1\}$ to the constraints that d is a pseudo-metric, the minimization problem $\min_{d \text{ metric}} \frac{\sum_{uv} c_{uv} d_{uv}}{\sum_{uv} dem(u,v)\, d_{uv}}$ can be solved by linear programming in polytime.

L_1-metric

A crucial observation is that if the metric is an L_1-metric, then we know how to obtain an integral solution with the same objective value, i.e. $\min_{f} \frac{\sum_{uv} c_{uv} \|f(u)-f(v)\|_1}{\sum_{uv} dem(u,v) \|f(u)-f(v)\|_1}$ is exact.

First, since L_1-embedding is linear, we can assume that the optimal solution is of 1-dimension,

$$\min_{f: V \to \mathbb{R}} \frac{\sum_{uv} c_{uv} |f(u)-f(v)|}{\sum_{uv} dem(u,v) |f(u)-f(v)|}.$$

Then, using a random thresholding argument (i.e. assume $\min_v f(v)=0$, $\max_v f(v)=1$, choose a random threshold $t \in [0,1]$, consider the set $S_t = \{v \mid f(v) \geq t\}$, compute $E_t\left[\sum_{uv \in \delta(S_t)} c_{uv}\right]$ and $E_t\left[\sum_{uv \in \delta(S_t)} dem(u,v)\right]$), we can conclude that there exists t such that $\sum_{uv \in \delta(S_t)} c_{uv} / \sum_{uv \in \delta(S_t)} dem(u,v)$

has the same objective value as the L_1-minimization problem (see Trevisan's notes for details).

Approximation algorithm

Showing that the L_1-minimization problem is exact shows that L_1-minimization is NP-hard.

But we can use Bourgain's theorem to connect the LP relaxation and the L_1-minimization.

First, solve the LP relaxation for general metric d in polynomial time.

Then, use Bourgain's theorem to embed the metric d into an L_1-metric f with $O(\log n)$ distortion.

Note that this would imply that the objective value for f is at most $O(\log n)$ times that of d.

Now, we can use the random thresholding algorithm on f to produce an integral solution as good.

It is instructive to unfold the steps to see that the resulting algorithm is quite simple.

Remark: Solving the open problem positively will apply a constant factor approximation algorithm for the sparsest cut problem on planar graphs.

$\sqrt{\log n}$-approximation: There is a $O(\sqrt{\log n})$-approximation for the sparsest cut problem.

It is based on SDP, which enforces a stronger structure on the metric, which can be used to embed into L_1 with $O(\sqrt{\log n})$ distortion.

Tree embedding

There is another embedding result which has many applications in designing approximation algorithms.

While it is not true that an metric can be embedded into a tree metric with low distortion, it turns out that it is possible to construct a family of trees such that the average distortion is low. This is in a similar spirit of hashing where we need a hash family.

Theorem Given any metric (V, d) with $|V|=n$ and $\Delta = \max d(x,y) / \min d(x,y)$, there exists an efficiently sample-able distribution $\mathcal{T}$ over spanning trees of V s.t. for all $u,v \in V$:

① $d_T(u,v) \geq d(u,v)$ for all $T \in \mathcal{T}$.

② $E_{T \in \mathcal{T}}[d_T(u,v)] \leq O(\log n \log \Delta)\, d(u,v)$.

Note that $\log \Delta$ can be assumed to be $O(\log n)$, and so the expected distortion is at most $O(\log^2 n)$.

This result is improved to $O(\log n)$, which can be shown to be optimal.

It has many applications in approximation algorithms, for problems related to distances and cuts.

It can also be used to prove Bourgain's result, as tree metrics can be isometrically embed into L_1.

References

- Lecture notes in discrete geometry, chapter 15, by Matousek. (Highly recommended.)
- Approximation algorithms, by Vazirani, for sparsest cut and Bourgain's theorem.
- Lecture notes on "Leighton-Rao relaxation" by Trevisan for the random threshold argument.
- The design of approximation algorithms, by Williamson and Shmoys, for tree embedding.