CS271 Randomness & Computation **Fall 2022**

Lecture 17: October 20

Instructor: Alistair Sinclair

17.1 Geometric Embeddings

A *metric space*, denoted by (X, d), is a set X with a real valued function $d: X \times X \to \mathbb{R}$ (called a metric or a distance function) such that for every $x, y, z \in X$,

1. $d(x, y) \geq 0$ with equality if and only if $x = y$,
2. $d(x, y) = d(y, x)$, and
3. $d(x, y) \leq d(x, z) + d(y, z)$.

We will restrict our attention to the case where both $|X|$ and d are finite. The problem we are interested in is finding a way to map (X, d) to a "nicer" space (Y, d') by some mapping φ with small *distortion*, i.e., such that $\forall\ x, y \in X,\ d'(\varphi(x), \varphi(y)) \approx d(x, y)$. There are at least two motivations for doing this:

1. To solve a problem in a high dimensional space by first mapping it to a much lower dimension and then using an algorithm that is efficient in the lower dimensional space. (Many geometric algorithms have a running time that is exponential in the dimension.)
2. To solve a problem on a space with an arbitrary finite metric by first embedding it into a another space with an easier metric, such as the ℓ_p distance, and then solving the problem with this easier metric.

We now give a canonical example of a technique for each of these two cases.

17.1.1 Dimension reduction

We first consider the problem of mapping a Euclidean metric space to another Euclidean space of lower dimension. The fundamental theorem here is the so-called Johnson-Lindenstrauss Lemma.

Theorem 17.1 *([JL84]) For any set X of n points in $\mathbb{R}^d$ and for all $\varepsilon \in (0, 1)$, there exists a mapping $\varphi: \mathbb{R}^d \to \mathbb{R}^k$ where*

$$k = \left\lceil \frac{4 \ln n}{\varepsilon^2/2 - \varepsilon^3/3} \right\rceil \leq \left\lceil \frac{24}{\epsilon^2} \ln n \right\rceil$$

such that $\forall\ u, v \in X$,

$$(1 - \varepsilon) \|u - v\|_2^2 \ \leq\ \|\varphi(u) - \varphi(v)\|_2^2 \ \leq\ (1 + \varepsilon) \|u - v\|_2^2 .$$

NOTE: The value of the dimension k in the lemma is known to be optimal up to the constant factor [LN17].

Proof: There are several proofs of this result; we will follow [DG99]. The proof uses the probabilistic method, and the embedding is very simple: just project the points onto a random k-dimensional hyperplane. More precisely, for each point v in the original space, we let its image $\varphi(v)$ be $\sqrt{\frac{d}{k}}v'$, where v' is the projection of v onto the random k-dimensional hyperplane.

To analyze this embedding, we need to consider the distribution of the r.v.

$$\frac{\|\varphi(u)-\varphi(v)\|_2^2}{\|u-v\|_2^2},$$

over the choice of random projections φ. We may as well assume that $\|u-v\|_2^2 = 1$, i.e., $u-v$ is a unit vector. But note that the distribution of $\|\varphi(u)-\varphi(v)\|_2^2$ (the squared length of the projection of a fixed unit vector onto a random hyperplane) is the same as that of a *random* unit vector projected onto a *fixed* k-dimensional hyperplane. We can analyze this by picking a random point on the unit d-dimensional sphere and projecting it onto the hyperplane defined by the first k coordinate vectors. To generate a random unit vector on the sphere, we first generate a random vector $X = (X_1, \ldots, X_d)$, where the $X_i \sim N(0,1)$ are i.i.d. standard normal and then scale it to obtain the unit vector $Z = \frac{1}{\|X\|_2}(X_1, \ldots, X_d)$. Projecting Z onto the first k dimensions, we get the projected vector $Y = \frac{1}{\|X\|_2}(X_1, \ldots, X_k)$.

Our goal is to analyze the distribution of

$$L = \|Y\|_2^2 \;=\; \frac{X_1^2 + \ldots + X_k^2}{X_1^2 + \ldots + X_d^2}.$$

Note that, by symmetry, $\mu \equiv \mathrm{E}[L] = \frac{k}{d}$. (Let $\mu_i = \mathrm{E}[\frac{X_i^2}{X_1^2+\cdots+X_d^2}]$. Then all the μ_i are equal, and $\sum_i \mu_i = 1$.) This is why we scaled the projection by $\sqrt{\frac{d}{k}}$ above.

The crucial fact is the following, which is a Chernoff-type large deviation bound tuned to the current geometric setting:

Claim 17.2 *For L and μ as defined above, we have*

(i) $\Pr[L \le (1-\varepsilon)\mu] \le \exp(\frac{-\varepsilon^2 k}{4})$

(ii) $\Pr[L \ge (1+\varepsilon)\mu] \le \exp(\frac{-k}{2}(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}))$

We will prove this Claim in a moment. First, we complete the proof of Theorem 17.1.

Note from the Claim that, for $k \ge \left\lceil \frac{4\ln n}{(\varepsilon^2/2)-(\varepsilon^3/3)} \right\rceil$, we have

$$\Pr[|L-\mu| \ge \varepsilon\mu] \;\le\; 2\exp(-2\ln n) \;=\; \frac{2}{n^2}.$$

Hence the probability that the distortion is "bad", i.e.,

$$\Pr[|L-\mu| > \varepsilon\mu \text{ for any pair } u,v] \;\le\; \frac{2}{n^2}\binom{n}{2} \;=\; \left(1-\frac{1}{n}\right).$$

Therefore, the embedding is good with probability $\ge \frac{1}{n}$. This certainly implies the existence of a good embedding. Moreover, if we repeatedly try random projections until we get a good embedding, the expected number of trials we will need will be only $O(n)$. ■

We now go back and prove Claim 17.2.

Proof: The proof is essentially a Chernoff bound, adapted to the current setting where the random variable in question is the ratio of the sum of independent Gaussians; thus we will follow the same steps as in the proof of our earlier Chernoff bounds. The main difference is that we will use the following distribution-dependent fact: If $X \sim N(0,1)$, then $\mathrm{E}\left[e^{(sX^2)}\right] = (1-2s)^{-\frac{1}{2}}$ for $-\infty < s < \frac{1}{2}$. The proof of this fact is left as a straightforward exercise (just ingegrate over the Normal density).

We now proceed to prove the lower tail bound (i):

$$\begin{aligned}
&\Pr[L \le (1-\varepsilon)\mu] \\
&= \Pr\left[\left(\frac{X_1^2+\ldots+X_k^2}{X_1^2+\ldots+X_d^2}\right) \le (1-\varepsilon)\frac{k}{d}\right] \\
&= \Pr\left[k(1-\varepsilon)(X_1^2+\ldots+X_d^2) - d(X_1^2+\ldots+X_k^2)) \ge 0\right] \\
&= \Pr\left[\exp\{t[k(1-\varepsilon)(X_1^2+\ldots+X_d^2) - d(X_1^2+\ldots+X_k^2)]\} \ge 1\right] \qquad \text{for any } t>0 \\
&\le \mathrm{E}\left[\exp\{t[k(1-\varepsilon)(X_1^2+\ldots+X_d^2) - d(X_1^2+\ldots+X_k^2)]\}\right] \qquad \text{[by Markov's inequality]} \\
&= \mathrm{E}\left[\exp\{tk(1-\varepsilon)X_1^2\}\right]^{(d-k)} \ \mathrm{E}\left[\exp\{t(k(1-\varepsilon)-d)X_1^2\}\right]^k \qquad \text{[since the } X_i \text{ are independent]} \\
&= (1-2tk(1-\varepsilon))^{-(d-k)/2} \ (1-2t(k(1-\varepsilon)-d))^{-k/2} \qquad \text{[by the fact stated above].}
\end{aligned}$$

Choosing t to maximize the above expression, we get after a little calculus

$$t = \frac{\varepsilon}{2(1-\varepsilon)(d-k(1-\varepsilon))}.$$

(Note that this value satisfies $t(k(1-\varepsilon)-d) < tk(1-\varepsilon) < \frac{1}{2}$, so we were justified in applying the fact in the final step above.) Plugging this value of t back into the expression for $\Pr[L \le (1-\varepsilon)\mu]$, we get

$$\begin{aligned}
&\Pr[L \le (1-\varepsilon)\mu] \\
&\le \left(1-\frac{k\varepsilon}{d-k(1-\varepsilon)}\right)^{-(d-k)/2}\left(1+\frac{\varepsilon}{1-\varepsilon}\right)^{-k/2} \\
&= \left(1+\frac{k\varepsilon}{d-k}\right)^{(d-k)/2}(1-\varepsilon)^{k/2} \\
&< \exp(\frac{k\varepsilon}{2}) \ (1-\varepsilon)^{k/2} \qquad \text{[since } (1+\tfrac{x}{y})^y < e^x\text{]} \\
&= \exp\left(\frac{k\varepsilon}{2}+\frac{k}{2}\ln(1-\varepsilon)\right) \\
&< \exp\left(-\frac{\varepsilon^2 k}{4}\right) \qquad \text{[by Taylor expansion: } \ln(1-\varepsilon) < (-\varepsilon-\tfrac{\varepsilon^2}{2})\text{]}.
\end{aligned}$$

(ii) We can follow a similar procedure to prove that $\Pr[L \ge (1+\varepsilon)\mu] \le \exp(\frac{-k}{2}(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}))$; the only difference is that, in the end, we will be left with a term in $\ln(1+\varepsilon)$ instead of the $\ln(1-\varepsilon)$ term. We will then have to use the Taylor expansion $\ln(1+\varepsilon) < \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}$ to get the desired bound. ■

Remark: For a more efficient implementation of the above embedding, which replaces the above random projection with a more discrete one, see [A03].

17.1.2 Metric simplification: Embedding into ℓ_p metrics

We now turn to the case when we wish to map an arbitrary metric on n points into another metric space that is easier to handle. Typically, this target space will be the Euclidean space $\mathbb{R}^k$ (for some dimension k) equipped with an ℓ_p metric (e.g., the Euclidean metric ℓ_2 or the Manhattan metric ℓ_1). Here we will consider the question of embedding into ℓ_1, a case that has several algorithmic applications (see, e.g., [LLR95, AR95]).

We can think of any finite metric (X, d) as the shortest-path metric of a weighted undirected graph on vertex set X, in which $d(u, v)$ is the length of a shortest path from u to v. Of course, there is always the trivial representation on the complete graph, in which each edge (u, v) has weight $d(u, v)$. However, many metrics have a more compact representation, in which the graph is much sparser.

Intuitively, we would expect the embedding properties of the metric to depend on the structure of its graphical representation. For example, the complete graph on three vertices with unit edge weights can be embedded isometrically (i.e., with distortion 1, so that distances are preserved exactly) into $\mathbb{R}^3$ with coordinates $(\frac{1}{2}, 0, 0)$, $(0, \frac{1}{2}, 0)$, and $(0, 0, \frac{1}{2})$ under the ℓ_1 distance. Also, it is not hard to see that any weighted tree metric can be mapped isometrically into $\mathbb{R}^k$ with the ℓ_1 metric (exercise!). For an arbitrary graph, however, an isometric embedding into ℓ_1 need not exist; the simplest counterexample is the complete bipartite graph $K_{2,3}$ with unit edge weights, which requires distortion at least $\frac{4}{3}$. (Indeed, it can be shown that $K_{2,3}$ is the only obstacle to isometric embedding, in the sense that every metric supported on a graph G can be isometrically embedded into ℓ_1 if and only if G is $K_{2,3}$-free.) A major open problem is whether any metric supported on a *planar* graph can be embedded into ℓ_1 with *constant* distortion.

We now prove one of the most famous results in this area, due to Bourgain, which says that *any* finite metric on n points can be embedded into ℓ_1 with distortion $O(\log n)$. The proof can easily be adapted to the ℓ_2 case (exercise!).

Theorem 17.3 ([Bou85]) *Let (X, d) be a metric space with $|X| = n$. Then (X, d) can be embedded into $\mathbb{R}^k$ with the ℓ_1 metric with distortion at most $O(\log n)$ (and dimension $k = O(\log^2 n)$).*

Proof: We will define a mapping $\varphi : X \to \mathbb{R}^k$ such that

$$\frac{1}{c \log n} d(x, y) \leq \|\varphi(x) - \varphi(y)\|_1 \leq d(x, y), \tag{17.1}$$

for some constant c. We use a randomized construction.

Our embedding will be defined by $m = O(\log^2 n)$ random sets $A_i \subseteq X$, which will be picked in a manner to be described shortly. The embedding is then

$$\varphi(x) = \frac{1}{m} \left(d(x, A_1), \ldots, d(x, A_m) \right)$$

where $d(x, A_i) = \min_{y \in A_i} d(x, y)$. I.e., the coordinates of our embedding are just the distances of the given point from each of the subsets.

We will first show that this embedding satisfies the upper bound claimed above.

Claim 17.4 $\forall\ x, y \in X,\ \|\varphi(x) - \varphi(y)\|_1 \leq d(x, y)$.

Proof: By definition of the ℓ_1 metric, we have

$$\|\varphi(x) - \varphi(y)\|_1 = \tfrac{1}{m} \sum_{i=1}^{m} |d(x, A_i) - d(y, A_i)|.$$

Hence it is sufficient to show that $|d(x, A_i) - d(y, A_i)| \leq d(x, y)$ for each i.

Given $x, y \in X$ and some A_i, assume w.l.o.g. that $d(x, A_i) \leq d(y, A_i)$. Then we need to prove that $d(y, A_i) \leq d(x, y) + d(x, A_i)$. Let z be the point in A_i closest to x, i.e., $d(x, A_i) = d(x, z)$. Then, by definition of $d(y, A_i)$, we have $d(y, A_i) \leq d(y, z)$. Now, applying the triangle inequality, we have $d(y, z) \leq d(x, z) + d(x, y) = d(x, A_i) + d(x, y)$, which is the desired result. ■

Note that this upper bound is rather trivial, and holds for any choice of the sets A_i (essentially because we scaled everything by $\frac{1}{m}$ to make it hold). We will now prove the much less trivial lower bound on $\|\varphi(x) - \varphi(y)\|_1$; here the choice of the A_i will play a central role.

Procedure for choosing sets A_i: For each $t \in \{1, 2, \ldots, \log n\}$, construct $r \log n$ random sets $\{A_i^t\}_{i=1}^{r \log n}$ as follows. For each i, include each element of X in A_i^t independently with probability 2^{-t}. Note that the expected size of each A_i^t is $\frac{n}{2^t}$, and there are $m = r \log^2 n$ sets in total.

Lemma 17.5 *For all pairs $x, y \in X$ and some constant c, we have $\|\varphi(x) - \varphi(y)\|_1 \geq \frac{1}{c \log n} d(x, y)$ with high probability.*

Proof: Fix some pair $x, y \in X$. Define the balls

$$\begin{aligned} B(x, \rho) &= \{z \in X : \ d(x, z) \leq \rho\} \\ B^o(x, \rho) &= \{z \in X : \ d(x, z) < \rho\}. \end{aligned}$$

Define the increasing sequence of radii $0 = \rho_0 < \rho_1 < \rho_2 < \cdots$ by

$$\rho_t = \min \left\{ \rho : \ B(x, \rho) \text{ and } B(y, \rho) \text{ both contain at least } 2^t \text{ points of } X \right\}.$$

We can construct this sequence by gradually increasing ρ until one of the sets has exactly 2^t points of X and the other has $\geq 2^t$ points. We continue this process while $\rho_t < \frac{1}{4} d(x, y)$. Let $t^* - 1$ be the last such t, and define $\rho_{t^*} = \frac{1}{4} \, d(x, y)$. Note that the balls $B(x, \rho_t)$ and $B(y, \rho_t)$ are disjoint for all t.

Now let's return to our random sets A_i^t. We say that A_i^t is "good" (for x, y) if it intersects the ball $B(y, \rho_{t-1})$ and does not intersect the ball $B^0(x, \rho_t)$. (Here we have assumed w.l.o.g. that the x-ball defines this radius; otherwise we can just interchange the roles of x and y. Furthermore, $B^0(x, \rho_t)$ is an open ball, so $|B^0(x, \rho_t)| < 2^t$.)

Here's why we call such a set "good": If A_i^t is good then its contribution to $\|\varphi(x) - \varphi(y)\|_1$ is $\geq \frac{1}{m}(\rho_t - \rho_{t-1})$ (since $d(x, A_i^t) \geq \rho_t$ and $d(y, A_i^t) \leq \rho_{t-1}$).

We will now show that, with high probability, enough of our sets are good so that we get a total contribution to $\|\varphi(x) - \varphi(y)\|_1$ of at least $\frac{1}{c \log n} d(x, y)$.

For any set A_i^t we have

$$\begin{aligned}
&\Pr[A_i^t \text{ is good for } x, y] \\
&= \Pr[A_i^t \text{ hits } B(y, \rho_{t-1}) \wedge A_i^t \text{ misses } B^0(x, \rho_t)] \\
&\geq \Pr[A_i^t \text{ hits } B(y, \rho_{t-1})] \times \Pr[A_i^t \text{ misses } B^0(x, \rho_t)] \\
&\quad \text{[these two events are not independent; however they are positively correlated!]} \\
&\geq \left(1 - \left(1 - 2^{-t}\right)^{2^{t-1}}\right) \times \left(\left(1 - 2^{-t}\right)^{2^t}\right) \\
&\quad \text{[there are } \geq 2^{t-1} \text{ points in } B(y, \rho_{t-1}) \text{ and } \leq 2^t \text{ points in } B^0(x, \rho_t); \text{ each point is in } A_i^t \text{ w.p. } 2^{-t}] \\
&\geq \left(1 - \frac{1}{\sqrt{e}}\right) \times \frac{1}{4} \\
&\quad \text{[for the second factor, note that it is increasing with } t, \text{ so the worst case is } t = 1] \\
&\geq \frac{1}{12}.
\end{aligned}$$

So with constant probability, each A_i^t is good for x, y. Since we are picking $r \log n$ such sets for each value of t, we have for each t

$$\mathrm{E}[\#\text{ of good sets for } x, y] \geq \frac{r \log n}{12} =: \mu.$$

Applying a Chernoff bound we have

$$\Pr[\#\text{ of good sets for } x, y \ \leq \tfrac{1}{2}\mu] \leq \exp(-\tfrac{\mu}{8}) = \exp(-\tfrac{r \log n}{96}) \leq \tfrac{1}{n^3},$$

if we choose $r = 288$. Taking a union bound over all $O(n^2)$ pairs x, y and all $O(\log n)$ values of t, we see that with probability $1 - O(\frac{n^2 \log n}{n^3}) = 1 - o(1)$ every pair x, y has at least $\frac{1}{2}\mu = \frac{r \log n}{24}$ good sets for every t.

We now assume the above condition holds, and compute a lower bound on $\|\varphi(x) - \varphi(y)\|_1$. Recalling that for a good set A_i^t we have $|d(x, A_i^t) - d(y, A_i^t)| \geq (\rho_t - \rho_{t-1})$, we get

$$\begin{aligned}
\|\varphi(x) - \varphi(y)\|_1 &= \frac{1}{m} \sum_{t=1}^{t^*} \sum_{i=1}^{r \log n} |d(x, A_i^t) - d(y, A_t^i)| \\
&\geq \frac{1}{m} \frac{r \log n}{24} \left((\rho_1 - \rho_0) + (\rho_2 - \rho_1) + \ldots + (\rho_{t^*} - \rho_{t^*-1})\right) \\
&= \frac{1}{m} \frac{r \log n}{24} (\rho_{t^*} - \rho_0) \\
&= \frac{1}{m} \frac{r \log n}{24} \frac{1}{4}\, d(x, y) \\
&= \frac{1}{96 \log n}\, d(x, y),
\end{aligned}$$

which is the bound we claimed in the lemma (with $c = 96$). In the next to last line we used the facts that $\rho_{t^*} = \frac{1}{4} d(x, y)$ and $\rho_0 = 0$ and in the last line, the fact that $m = r \log^2 n$. ■

We have thus completed our verification of both the upper and lower bounds in (17.1), which completes the proof of the theorem. Note that we have actually proved the existence of such an embedding via a randomized construction. Our proof further shows that a random embedding constructed as above will have the desired distortion with high probability. ■

References

[A03] D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of Computer & System Sciences* **66** (2003), pp. 671–687.

[AR95] Y. Aumann and Y. Rabani, "An $O(\log k)$ approximate min-cut max-flow theorem and approximation algorithm," *Proceedings of ACM/SIAM SODA* 1995.

[B85] J. Bourgain, "On Lipschitz embedding of finite metric spaces in Hilbert space," *Israel Journal of Mathematics* **52** (1985), pp. 46–52.

[DG99] S. Dasgupta and A. Gupta, "An elementary proof of the Johnson-Lindenstrauss Lemma," Technical Report TR-99-006, International Computer Science Institute, Berkeley, CA, 1999.

[JL84] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz maps into a Hilbert space," *Contemporary Mathematics*, 1984, 26:189–206.

[LN17] K.G. Larsen and J. Nelson, "Optimality of the Johnson-Lindenstrauss lemma," *Proceedings of IEEE FOCS*, 2017, pp. 633–638.

[LLR95] N. Linial, E. London and Y. Rabinovich, "The geometry of graphs and some of its algorithmic appplications," *Combinatorica* **15** (1995), pp. 215–245.