Data Science Through a Geometric Lens Fall 2019

Lecture 06 — October 16 & 21, 2019

Prof. Cyrus Rashtchian *Topics: Bourgain's Embedding*

Overview. In the past two lectures, we estimated the L2 norm of a data stream in two ways: the Alon-Matias-Szegedy (AMS) algorithm [2] and the Johnson-Lindenstrauss (JL) lemma [1].

In this lecture, we discuss an important and general metric embedding due to Bourgain [3]. The goal is to transform an arbitrary discrete metric space into Euclidean space by allowing some approximation (known as distortion) to the distances.

1 Metric Spaces and Embeddings

We first define a metric space then provide several examples. We consider discrete metric spaces $\mathcal{X}$ that consist of $n = |\mathcal{X}|$ points along with a distance function between pairs of points.

Definition 1. *A* metric space *is a pair $(\mathcal{X}, d_{\mathcal{X}})$, where $\mathcal{X}$ is a set and $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a nonnegative* distance function *on pairs in $\mathcal{X}$. The function $d_{\mathcal{X}}$ satisfies three properties:*

1. *$d_{\mathcal{X}}(x, y) = 0$ if and only if $x = y$,*
2. *$d_{\mathcal{X}}(x, y) = d_{\mathcal{X}}(y, x)$,*
3. *$d_{\mathcal{X}}(x, y) \leq d_{\mathcal{X}}(x, z) + d_{\mathcal{X}}(z, y)$ for all $x, y, z \in \mathcal{X}$.*

The third property (known as the *triangle inequality*) differentiates metric spaces from arbitrary notions of distance, and it implies much useful structure.

1.1 Examples

- **Hamming distance:** Binary strings, measuring # bits that differ, e.g., $d_{\mathrm{H}}(1011, 0111) = 2$.
- **Edit distance:** Arbitrary strings and measure the number of insertions, deletions, substitutions to transform one to the other, e.g., $d_{\mathrm{E}}(abcd, acd) = 1$ and $d_{\mathrm{E}}(abcd, abbb) = 2$.
- **Shortest path metric:** Take an arbitrary graph $G = (V, E)$, and define the distance between two vertices v and v' as $d_{\mathrm{G}}(v, v')$ = length of the shortest path between v and v'. This works for weighted graphs as well with positive weights.
- **Vector distances:** For two vectors $x, y \in \mathbb{R}^d$, you can look at the ℓ_p distance for of the difference: $\|x - y\|_p$ for any $p \geq 1$.
- **Matrix distances:** For two matrices $A, B \in \mathbb{R}^{n \times m}$, you can look at the ℓ_p norm of the singular values of $A - B$.
- **Trivial metric:** Take any set of points $\mathcal{X}$ and define $d(x, x) = 0$ and $d(x, y) = 1$ for $x \neq y$.

1.2 A very general embedding

How do we understand a new metric space $(\mathcal{X}, d_{\mathcal{X}})$? Well, we might try to compare it to a familiar metric space. For example, we could consider a mapping $F : \mathcal{X} \to \mathbb{R}^k$ into Euclidean space $\mathbb{R}^k$ equipped with the Euclidean norm $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_k^2}$.

How do we measure how good this mapping F is? One way to measure this is the distortion. This is a smallest number $C > 0$ such that

$$d(x, y) \leq \|F(x) - F(y)\|_2 \leq C \cdot d(x, y).$$

In this lecture, we prove the following result due to Jean Bourgain [3].

Theorem 2 (Bourgain, 1985)**.** *Any n-point metric space embeds into some Euclidean space $\mathbb{R}^k$ with distortion C that is bounded by $C = O(\log n)$. Moreover, we can achieve dimension $k = O(\log n)$.*

Today we will show $k = O(\log^2 n)$. But this is enough: we can get down to $O(\log n)$ dimensions by using Johnson-Lindenstrauss.

We start with a simpler but worse embedding due to Frechet.

2 Warm-up: Frechet's Embedding

We first demonstrate an embedding with distortion $D \leq \sqrt{n}$ and dimension $k = n$. Denote and index the n points in $\mathcal{X}$ as $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$.

Claim 3 (Frechet's Embedding)**.** *The embedding $f : \mathcal{X} \to \mathbb{R}^n$ defined by*

$$f(x) = (f_1(x), f_2(x), \ldots, f_n(x)), \qquad \text{where} \quad f_i(x) = d(x, x_i).$$

has distortion $\sqrt{n}$, that is, for every $x, y \in \mathcal{X}$ we have

$$d(x, y) \leq \|f(x) - f(y)\|_2 \leq \sqrt{n} \cdot d(x, y).$$

Proof. First, observe that every coordinate is 1-Lipschitz: for every $x, y \in \mathcal{X}$, we have

$$|f_i(x) - f_i(y)| = |d(x, x_i) - d(y, x_i)| \leq d(x, y),$$

where we have used the triangle inequality. From this we get that

$$\|f(x) - f(y)\|_2^2 = \sum_{i=1}^{n} (f_i(x) - f_i(y))^2 \leq n \cdot d(x, y)^2.$$

This implies that $\|f(x) - f(y)\|_2 \leq \sqrt{n} \cdot d(x, y)$ for every $x, y \in \mathcal{X}$.

For the lower bound, consider any $y \in \mathcal{X}$, and any $x_j \in \mathcal{X}$. We directly see that

$$\|f(y) - f(x_j)\|_2^2 = \sum_{i=1}^{n} (f_i(y) - f_i(x_j))^2 \geq (f_j(y) - f_j(x_j))^2 = (d(y, x_j) - d(x_j, x_j))^2 = (d(y, x_j) - 0)^2$$

Therefore, $\|f(y) - f(x_j)\|_2 \geq d(y, x_j)$ as desired. □

3 Bourgain's Embedding

To improve the distortion (and prove the above theorem), we will use distances to *subsets* instead of to single points. For a subset $S \subseteq \mathcal{X}$ and any point $x \in \mathcal{X}$, we define

$$d(x, S) = \min_{s \in S} d(x, s),$$

as the distance between x and the closest point in S. Note that this map is also 1-Lipschitz. Indeed, the triangle inequality implies that

$$d(x, S) \leq d(y, S) + d(x, y).$$

Hence,

$$|d(x, S) - d(y, S)| \leq d(x, y) \tag{1}$$

for all $x, y \in \mathcal{X}$ and $S \subseteq \mathcal{X}$.

Fix a number $m = O(\log n)$ that we will choose shortly (we also assume $\log_2 n$ is an integer; otherwise, use $\lfloor \log_2 n \rfloor$ throughout, which doesn't change the resulting bounds).

Define sets $S_{tj} \subseteq \mathcal{X}$ for $t = 1, 2, \ldots, \log_2 n$ and $j = 1, 2, \ldots, m$. Each S_{tj} is an independent random subset of $\mathcal{X}$, where S_{tj} is formed by sampling every point in $\mathcal{X}$ independently with probability 2^{-t}.

The embedding $F : \mathcal{X} \to \mathbb{R}^k$ is defined using coordinates t, j, where we set $F(x)_{tj} = d(x, S_{tj})$. Note that $k = O(\log^2 n)$ because t, j are both $O(\log n)$.

3.1 Analysis

Using Eq. (1), we see that

$$\|F(x) - F(y)\| \leq \sqrt{m \log n} \cdot d(x, y), \tag{2}$$

for every $x, y \in \mathcal{X}$. This establishes the upper bound, showing the F doesn't increase the distances by too much.

We move on to the lower bound. To this end, define open and closed balls: for $r \geq 0$,

$$\begin{aligned} B(x, r) &= \{y \in \mathcal{X} \mid d(x, y) \leq r\} \\ B^o(x, r) &= \{y \in \mathcal{X} \mid d(x, y) < r\} \end{aligned}$$

Fir $x, y \in \mathcal{X}$ and $t \in [\log_2 n]$. Let r_t b the smallest radius such that

$$\max\{|B(x, r_t)|,\ |B(y, r_t)|\} \geq 2^t.$$

Let t^* be the smallest value of t such that $r_t \geq d(x, y)/4$ and reassign $r_{t^*} = d(x, y)/4$.

Observe that

$$\frac{d(x, y)}{4} = r_1 + (r_2 - r_1) + (r_3 - r_2) + \cdots + (r_{t^*} - r_{t^*-1}). \tag{3}$$

We will use the sets S_{tj} to get a contribution of $(r_t - r_{t-1})$ to the lower bound, and therefore Eq. (3) shows that we will get a contribution of $\Omega(d(x, y))$.

Consider some $t \in \{1, 2, \ldots, t^*\}$. For the sake of analysis, let $r_0 = 0$. Notice that, by the definition of r_t, we have that at least one of $|B(x, r_{t-1})| \geq 2^{t-1}$ or $|B(y, r_{t-1})| \geq 2^{t-1}$ will hold. Assume WLOG that this holds for x. It is also true that $|B^o(y, r_t)| < 2^t$. In other words, we have that

$$|B(x, r_{t-1})| \geq 2^{t-1} \qquad \text{and} \qquad |B^o(y, r_t)| < 2^t.$$

Let $S_t \subseteq \mathcal{X}$ be a random subset where every point is sampled independently with probability 2^{-t}. Consider the event

$$\mathcal{E}_t = \{S_t \cap B(x, r_{t-1}) \neq \emptyset \text{ and } S_t \cap B^o(y, r_t) = \emptyset\}.$$

Notice that

$$\mathcal{E}_t \text{ occurring implies that } |d(x, S_t) - d(y, S_t)| \geq r_t - r_{t-1}. \tag{4}$$

Intuition. The main idea is that we need to choose the different S_t sets with very different probabilities to handle different distance scales. For example, when $t = 1$, we choose roughly half the points in the space. This means that we are very likely to get a point close to both x and y, so $d(x, S_t)$ and $d(y, S_t)$ will both be small. At the other extreme, when we sample points in $S_{\log n}$ with probability $1/n$, we will be likely to be far from both x and y (assuming they are far apart themselves). We will glue together these different scales for the final embeddings (which is what (3) and the discussion afterwards is saying formally).

We first understand the probabilities of these $\mathcal{E}_t$ events. A key fact will be that because $r_t \leq d(x, y)/4$, we have that the balls $B(x, r_{t-1})$ and $B^o(y, r_t)$ are disjoint. Then, intuitively, $\mathcal{E}_t$ will hold with constant probability because (i) with constant probability S_t will contain at least one point in $B(x, r_{t-1})$, and (ii) also with constant probability, S_t will miss all the points in $B^o(y, r_t)$. This makes sense because in *expectation* we will see $1/2$ of a point from $B(x, r_{t-1})$, and also less than one point from $B^o(y, r_t)$. So there's a chance we see some in one, and none in the other.

Each set S_t might be bad by itself (so we don't get the contribution for (3) like we want). So we will repeat the sampling $m = O(\log n)$ times, to get sets $S_{t1}, \ldots, S_{tm}$. But these will be independent events, so we will just focus on one such trial, and we call it S_t for brevity.

Then, we show that if enough of the events hold at the same time (i.e., $\Omega(m)$ of the m trials, for each t value), the embedding will be good overall. This will complete the proof (by setting $m = O(\log n)$ and taking a union bound overall pairs in the metric space).

Now back to the proof, understanding the probability of $\mathcal{E}_t$.

Claim 4. $\Pr[\mathcal{E}_t] \geq \frac{1}{12}$.

Proof. Observe that $r_t \leq d(x, y)/4$, hence $B(x, r_t)$ and $B(y, r_t)$ are disjoint.

In particular, the two events composing $\mathcal{E}_t$ are independent, and it suffices to lower bound their probabilities separately. First, note that

$$\begin{aligned}
\Pr\left[S_t \cap B(x, r_{t-1}) \neq \emptyset\right] &\geq 1 - \Pr\left[S_t \cap B(x, r_{t-1}) = \emptyset\right] \\
&\geq 1 - (1 - 2^{-t})^{|B(x, r_{t-1})|} \\
&\geq 1 - (1 - 2^{-t})^{2^{t-1}} \\
&\geq 1 - \frac{1}{\sqrt{e}} \geq \frac{1}{3},
\end{aligned}$$

where we have used the fact that $(1 - 1/k)^k \leq 1/e$ for $k > 1$. Next, calculate

$$\Pr[S_t \cap B^o(y, r_t) = \emptyset] = (1 - 2^{-t})^{|B^o(y, r_t)|} \geq \frac{1}{4},$$

where we have used that $(1 - 1/k)^k \geq 1/4$ for $k \geq 2$. In conclusion, by independence, both events occur together with probability at least $\frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$ as desired. □

Now, let $\mathcal{E}_{tj}$ be the event corresponding to Eq. (4) but for the set S_{tj}.

Corollary 5. *If $\Omega(m)$ of the events $\mathcal{E}_{tj}$ for $j \in [m]$ occur, then*

$$\|F(x) - F(y)\|_2^2 \geq \Omega(m) \cdot (r_t - r_{t-1})^2.$$

In fact, we have something stronger

$$\Omega(m) \text{ of the events } \{\mathcal{E}_{tj} \mid j \in [m]\} \text{ occur for every } t \in [\log_2 n]. \tag{5}$$

We can make this concrete by saying that at least $m/24$ of the events hold (for each t) with very high probability. This follows from a Chernoff bound.

Then, since the contributions come from disjoint sets of coordinates,

$$\begin{aligned}
\|F(x) - F(y)\|_2^2 &\geq \Omega(m) \cdot \sum_{t=1}^{t^*} (r_t - r_{t-1})^2 \\
&\geq \Omega\left(\frac{m}{t^*}\right) \cdot \left(\sum_{t=1}^{t^*} (r_t - r_{t-1})\right)^2 \\
&\geq \Omega\left(\frac{m}{t^*}\right) \cdot d(x,y)^2 \\
&\geq \Omega\left(\frac{m}{\log n}\right) \cdot d(x,y)^2.
\end{aligned}$$

The second inequality is Cauchy-Schwarz, and the third is from Eq. (3). We could plug in $m/24$ instead of using the Ω notation.

Combining this with Eq. (2), our map has distortion $O(\log n)$ as long as we choose m large enough so that Eq. (5) holds with probability, say, $1 - 1/n^3$. That's because we can then take a union bound over all possible pairs $x, y \in \mathcal{X}$. But since each event $\mathcal{E}_{tj}$ occurs with probability at least $1/12$, a simple Chernoff bound shows that choosing some $m = O(\log n)$ suffices.

References

[1] William Johnson, Joram Lindenstrauss Extensions of Lipschitz mappings into a Hilbert space *Contemporary Mathematics*, 189–206, 1984.

[2] Noga Alon, Yossi Matias, Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[3] Jean Bourgain. On Lipschitz Embedding of Finite Metric Spaces in Hilbert Space. *Israel J. Math.*, 52(1-2):46–52, 1985