Blockcourse on High-Dimensional Computational Geometry Spring 2013

Lecture 10: Johnson-Lindenstrauss L., Frechet&Hausdorff Metr.

Helmut Alt *Scribes: Rémy Thomasse, Udo Hoffmann*

1 Johnson-Lindenstrauss Lemma

Let's prove the lemma used in the proof of Johnson-Lindenstrauss flattening lemma.

Lemma 1. *For a random vector $x \in S^{n-1}$, let $f(x) = \sqrt{x_1^2 + \cdots + x_k^2}$.*
Then, there exist $= a(n,k)$ such that for all t,

$$\Pr[f(x) \geq a + t] \leq 2e^{-t^2 \frac{n}{2}}$$

and

$$\Pr[f(x) \leq a - t] \leq 2e^{-t^2 \frac{n}{2}}.$$

Moreover, there exist a constant c such that if $n \geq c$ and if $k \geq 10\ln(n)$, then $a \geq \frac{1}{2}\sqrt{\frac{k}{n}}$.

To proof this lemma, we need another lemma.

Lemma 2. *Levy's Lemma*
Let $f : S^{n-1} \to \mathbb{R}$ be a $1-$Lipschitz function (i.e $|f(x) - f(y)| \leq 1\|x - y\|_2 \ \forall x, y$).
Let $\mathrm{med}(f) := \sup\{t \in \mathbb{R} \mid \Pr(f \leq t) \leq \frac{1}{2}\}$. *Then,*

$$\Pr[f(x) \geq \mathrm{med}(f) + t] \leq 2e^{-t^2 \frac{n}{2}}$$

and

$$\Pr[f(x) \leq \mathrm{med}(f) - t] \leq 2e^{-t^2 \frac{n}{2}}.$$

Proof of the Lemma 1. The estimate on probability follows from Levy's Lemma with $a = \mathrm{med}(f)$.
For the lower bound for a, let's choose $x \in S^{n-1}$ uniformly at random.
Then, $1 = \mathbb{E}(\|x\|_2^2) = \sum_{i=1}^n \mathbb{E}(x_i^2)$. By the symmetry of the unit sphere, we get $\mathbb{E}(x_i^2) = \frac{1}{n}$ so $\mathbb{E}(f^2) = \frac{k}{n}$.
For $t \geq 0$,

$$\begin{aligned}\frac{k}{n} = \mathbb{E}(f^2) &\leq \Pr[f \leq a + t](a + t)^2 + \Pr[f > a + t]\underbrace{\max f^2(x)}_{\leq 1} \\ &\leq (a + t)^2 + 2e^{-t^2 \frac{n}{2}}\end{aligned}$$

Now let $t = \sqrt{\frac{k}{5n}}$. Since $k \geq 10\ln(n)$, $2e^{-t^2 \frac{n}{2}} = 2e^{-\ln(n)} = \frac{2}{n}$, so

$$\begin{aligned}&\frac{k}{n} \leq (a + t)^2 + \frac{2}{n} \\ &\Rightarrow k \leq n(a + t)^2 + 2 \\ &\Rightarrow \frac{1}{2}\sqrt{\frac{k}{n}} \leq \sqrt{\frac{k-2}{n}} - \sqrt{\frac{k}{5n}} \leq a.\end{aligned}$$

□

Theorem 3 (Bourgain's Theorem)**.** *Every $n-$points metric space (V,ρ) can be embedded into an Euclidian space ℓ_2 with distortion $O(\log n)$.*

Lemma 4. *Let $u,v \in V, u \neq v$. Then, there exist real numbers $\Delta_1,\dots,\Delta_q \geq 0$ with $\Delta_1+\dots+\Delta_q = \frac{1}{4}\rho(u,v)$, $q = \lfloor \log n \rfloor + 1$, such that for each $j = 1 \dots q$:*
If $A_j \subseteq V$ is drawn randomly, each element of V with probability 2^{-j}, then the probability p_j that $|\rho(u,A_j) - \rho(v,A_j)| \geq \Delta_j$ satisfies $p_j \geq \frac{1}{12}$.

Definition 5. *Line pseudometric*
A pseudometric has all the properties of a metric, except there may be $u,v, u \neq v$ but $\nu(u,v) = 0$.
A line pseudometric ν on V is a map $\nu : V \times V \to [0,+\infty)$ such that there exist a map $\varphi : V \to \mathbb{R}$ such that $\nu(u,v) = |\varphi(u) - \varphi(v)|$.

Lemma 6. *v Let (V,ρ) be a metric space, $\nu_1,\dots,\nu_N$ line pseudometrics on V with $\nu_i \leq \rho$ for $i = 1..N$, such that*

$$\sum_{i=1}^{N} \alpha_i \nu_i \geq \frac{1}{D}\rho$$

for some D and $\alpha_1,\dots,\alpha_n \in [0,1]$ summing up to 1.
Then, (V,ρ) can be $D-$embedded into ℓ_2^N.

Proof of the lemma 6. Let $\rho_i : V \to \mathbb{R}$ a map inducing ν_i.
Define $f : V \to \ell_2^N$ by $f(v)_i = \sqrt{\alpha_i}\rho_i(v)$.
Then,

$$\begin{aligned}
\|f(u) - f(v)\|^2 &= \sum_{i=1}^{N} (f(u)_i - f(v)_i)^2 \\
&= \sum_{i=1}^{N} [\sqrt{\alpha_i}(\varphi_i(u) - \varphi_i(v)]^2 \\
&= \sum_{i=1}^{N} \alpha_i \nu_i(u,v)^2 \leq \rho(u,v)^2
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|f(u) - f(v)\| &= \left(\sum_{i=1}^{N} \alpha_i \nu_i(u,v)^2 \right)^{\frac{1}{2}} \\
&= \underbrace{\left(\sum_{i=1}^{N} \alpha_i \right)^{\frac{1}{2}}}_{=1} \left(\sum_{i=1}^{N} \alpha_i \nu_i(u,v)^2 \right)^{\frac{1}{2}} \\
&= \|(\sqrt{\alpha_1},\dots,\sqrt{\alpha_N})\| \|(\sqrt{\alpha_1}\nu_1(u,v),\dots,\sqrt{\alpha_N}\nu_N(u,v))\| \\
&\geq \sum_{i=1}^{N} \alpha_i \nu_i(u,v) \text{ (by Cauchy-Schwartz Inequality)} \\
&\leq \frac{1}{D}\rho(u,v)
\end{aligned}$$

□

Proof of Bourgain's Theorem. Let $\Pi_j(A) = \Pr[$ ramdom subset A_j as in the lemma 4$= A]$ for $A \subseteq V$. Then, $\sum_{A \subseteq V} \Pi_j(A)\nu_A(u,v)$ is the expected value of $|\rho(u, A_j) - \rho(v, A_j)|$.
By the lemma 4 and Markov Inequality,

$$\sum_{A \subseteq V} \Pi_j(A)\nu_A(u,v) \leq \frac{1}{12}\Delta_j.$$

So, by summing,

$$\sum_{A \subseteq V} \left[\sum_{j=1}^{q} \Pi_j(A)\right] \nu_A(u,v) \geq \frac{1}{12}\sum_{j=1}^{q}\Delta_j = \frac{1}{48}\rho(u,v).$$

Let $\alpha_A = \frac{1}{q}\sum_{j=1}^{q}\Pi_j(A)$. Then $\sum_{A\subseteq V}\alpha_A = 1$ since $\sum_{A\subseteq V}\Pi_j(A) = 1$ for all j. So,

$$\sum_{A \subseteq V} \alpha_A\nu_A(u,v) \geq \frac{1}{48q}\rho(u,v)$$

for all u, v. By lemma 6, V can be embedded into ℓ_2 with distortion $48q = 48(\lfloor \log n \rfloor + 1)$, since ν_A's are line pseudometrics. □

Remark: In the proof of the theorem, V is embedded into ℓ_2^N where $N = 2^{|V|}$ (all subsets of V), but since n points are embedded, they lie in a $(n-1)-$dimensional subspace, so we can embed V into ℓ_2^{n-1}. It can be reduce to $\ell_2^{O(\log n)}$ by Johnson-Lindenstrauss Lemma.

2 Nearest neighbour search with respect to Hausdorff- or Frechet-distance

In this section we consider pointsets that forms a pattern. We have a database of patterns and want to query the nearest neighbour for a pattern q.

2.1 Hausdorff-distance

(Indyk, Farach-Colton 1999) Consider the Hausdorff distance d_H over finite subsets of size s (point patterns). There exists a $(1+\epsilon)$-embedding into l_∞^D where $D \approx O(\frac{s^2}{\epsilon^d})$. In l_∞^D we can approximate the nearest neighbour search.

For $d = 2$ and $d = 3$ there exist algorithms with the following properties.

1. A constant factor approximation with query time $O(s^2 \log_n)$ and space $n^{O(\log s)}$.
2. A $\log\log s$ approximation with query time $O(s^2 \log_n)$ and space $s^2 n^{1+\delta}$.

For similar patterns in a different position we can introduce the metric

$$d_H^T(A,B) = \min_{t \in T} d_H(t(A), B)$$

, where T are rigid motions, i.e. under translations and rotations. Under translations the query time and space requirements of the algorithms above does not change. If we also allow rotations the query time and space requirements are multiplied by s.

For an arbitrary underlying metric space M we can introduce embed d_H in l_∞^D with $D \approx s^2 m^\delta$ where $m = |M|$. This way, an efficient approximation algorithm for the nearest neighbour in M gives an efficient approximation algorithm for the NN problem for the Hausdorff distance over M.

2.2 Frechet-distance

For an underlying metric space (X, ρ) let X^* be the set of finite sequences over X. For $q = (p_1, \ldots, p_k)$ and $q = (q_1, \ldots, q_l)$ from X^* we define the discrete Frechet d_F distance as

$$d_F(p,q) = \min_s \min_{f,g} \max_t \rho(p_{f(t)}, q_{g(t)}),$$

where $f : \{1, \ldots, s\} \to \{1, \ldots, k\}$ and $g : \{1, \ldots, s\} \to \{1, \ldots, l\}$ are monotonically increasing and $g(1) = f(1) = 1$, $f(s) = k$, $g(s) = l$, and $f(i) - f(i-1) \leq 1$. This can be seen as comparing a discrete point set on two curves as approximation of the Frechet distance of the curves.

For this problem product metrics: for metric spaces $(M_1, \rho_1) \times \cdots \times (M_k, \rho_k)$ the metric space $M = M_1 \times \cdots \times M_k$ and $\rho(p,q) := \max_i \rho_i(p_i, q_i)$ is defined. Then we can approximate the NN in each M_i with query time $Q(n)$ and space $S(n)$:

1. Construct a $c(\log \log n)$-approximate NN with query time $O(Q(n) \log n + k \log n)$ and space $O(kS(n)n^{1+\delta})$ for the product metric.

2. The nearest neighbour for discrete Frechet metric can be reduced to several Frechet metrics over shorter sequences.

The result will be a $O((\log m + \log \log n)^\delta)$-approximate NN where m is the maximal length of a sequence. The space requirement is $O(1)^{1+\delta} m$ and the query time $(m + \log n)^{1+O(\delta)}$.