Algorithms for streaming data (mostly via embeddings)

Piotr Indyk
Streaming Data

- Problems defined over points $P=\{p_1, \ldots, p_n\}$
- The algorithm sees $p_1$, then $p_2$, then $p_3$, ...
- Key fact: it has limited storage
  - Can store only $s << n$ points
Example - diameter
Problems

• Diameter
• Minimum enclosing ball
Diameter in $l^d_{\infty}$

• Assume we measure distances according to the $l_{\infty}$ norm
• How can we compute it in the streaming model?
• Actually, how can we compute it efficiently?

\[
\max_{p,q \in P} \| p - q \|_{\infty} = \max_{p,q \in P} \max_{i=1,\ldots,d} |p_i - q_i| = \max_{i=1,\ldots,d} \max_{p,q \in P} |p_i - q_i| = \max_{i=1,\ldots,d} \left[ \max_{p \in P} p_i - \min_{p \in P} p_i \right]
\]
Diameter in $l_\infty$, streaming

- From the definition
  \[ \text{Diam}_\infty(P) = \max_{i=1 \ldots d} \left[ \max_{p \in P} p_i - \min_{p \in P} p_i \right] \]
- Can maintain max/min in constant space
- Total space = $O(d)$
- What about $l_1$?
Consider \( d=2 \)

- Define:
  \[
  f(x,y) = [x+y, x-y, -x+y, -x-y]
  \]
  - Since \( f \) linear, we have \( ||f(p)-f(q)|| = ||f(p-q)|| \)
  - \( ||(x,y)||_1 = |x|+|y| = \max[x+y, x-y, -x+y, -x-y] \)

- \( f \) is an **isometric** embedding of \( l^2_1 \) into \( l^4_\infty \)
The mapping $f$ is defined as

$$f(p) = (s_0 \cdot p, s_1 \cdot p, \ldots, s_{2^d-1} \cdot p)$$

where $s_i$ is the $i^{th}$ vector in $\{-1, 1\}^d$. Then

$$\|f(p) - f(q)\|_\infty = \|f(p - q)\|_\infty = \max_s |s \cdot (p - q)| = |p_1 - q_1| + \ldots + |p_d - q_d| = \|p - q\|_1$$

• $f$ is an **isometric** embedding of $l^d_1$ into $l^{2^d}_\infty$
Diameter in $l_1$

- Let $f : l_1^d \rightarrow l_\infty^{2^d}$ be an isometric embedding.
- We will maintain $\text{Diam}_\infty(f(P))$.
  - For each point $p$, we compute $f(p)$ and feed it to the previous algorithm.
  - Return the pair $p,q$ that maximizes $||f(p)-f(q)||_\infty$.
- This gives $O(2^d)$ space for $l_1^d$.
- What about $l_2$?
(1+\(\varepsilon\))-embedding of \(l^d_2\) into \(l^{d'}_\infty\)

- (1+\(\varepsilon\))-embedding:
  - No expansion
  - Contraction by at most 1+\(\varepsilon\)

- We will achieve \(d'\) equal to \(O(1/\varepsilon)^{(d-1)/2}\)
- Let’s start from \(d=2\)
Embedding of $l_2$ into $l_\infty$

• Again, use projections
  – Onto unit ($l_2$) vectors $v_1 \ldots v_k$
  – Requirement: vectors are “densely” spaced:
    for any $u$ there is $v_i$ such that
    \[ u^*v_i \geq ||u||_2 / (1+\varepsilon) \]
  – This implies $(1+\varepsilon)$ distortion
    • No expansion
    • Contraction by at most $1+\varepsilon$

• How big is $k$?
  – Can assume $||u||_2=1$

5/10/18
Lemma

• Consider two unit vectors $u$ and $v$, such that the $\angle(u, v) = \alpha$.

Then $u \cdot v \geq 1 - \Theta(\alpha^2)$

• Proof: $u \cdot v = \cos(\alpha) = 1 - \Theta(\alpha^2)$

• Therefore, suffices to use $2\pi/\epsilon^{1/2}$ vectors to get distortion $1 + \Theta(\epsilon)$

• The mapping $f(p) = (v_1 \cdot p, v_2 \cdot p, \ldots, v_k \cdot p)$ is a $(1 + \epsilon)$ distortion embedding of $l^2_2$ into $l^{\Theta(1/\epsilon^{1/2})}_\infty$. 

5/10/18
Higher Dimensions

- For $d=2$ we get $d' = O(1/\varepsilon^{1/2})$
- For any $d$ we get $d' = O_d(1/\varepsilon)^{(d-1)/2}$
  - Can “cover” a unit sphere in $\mathbb{R}^d$ with $O_d(1/\alpha)^{d-1}$ vectors so that any $v$ has angle $<\alpha$ with at least one of the vectors
  - The remainder is the same

5/10/18
Covering vs Packing

• Assume we want to **cover** the sphere using disks of radius $\alpha$

• This can be achieved by **packing** , as many as possible, disks of radius $\alpha/2$

• How many disks can be pack ?
  – Each disk has volume $\Theta_d(\alpha/2)^{d-1}$ times smaller than the volume of the sphere
  – Inverse of that gives the packing/covering bound
Diameter in $l_2$

- Let $f: l_2^d \rightarrow l_\infty^{d'}$, $d' = O(1/\varepsilon)^{(d-1)/2}$, be a $(1+\varepsilon)$-distortion embedding
- Apply the same algorithm as before
- Space: $O(1/\varepsilon)^{(d-1)/2}$
Minimum Enclosing Ball

• Problem: given $P=\{p_1 \ldots p_n\}$, find center $o$ and radius $r>0$ such that
  – $P \subseteq B(o,r)$
  – $r$ is as small as possible

• Solve the problem in $l_\infty$

• Generalize to $l_1$ and $l_2$ via embeddings
MEB in $l_\infty$

- Let $C$ be the hyper-rectangle defined by max/min in every dimension.
- Easy to see that min radius ball $B(o,r)$ is a min size hypercube that contains $C$.
- Min radius = min hypercube side length/2.
- How to solve it in $l_2$?
MEB in $l_2$

• Let $f: l_2^d \rightarrow l_\infty^{d'}$ be an embedding as before:
  – No expansion
  – Contraction by at most $1+\varepsilon$

• Attempt I:
  – Maintain $\text{MEB}_\infty B'(o',r)$ of $f(p_1)\ldots f(p_n)$
  – Compute $o$ such that $f(o)=o'$
  – Report $o$
Problem

• There might be NO $o$ such that $f(o)=o'$

• The problem is that $f$ is into, not onto
The Correct Version

• Attempt II:
  – Maintain the min/max points $f(p_1)\ldots f(p_{2d'})$, two points per dimension
  – Compute MEB $B(o,r)$ of $p_1\ldots p_{2d'}$
  – Report $o$
Correctness

MEB radius for $P$

$= \text{Min } r \text{ s.t. } \exists o \ P \subseteq B(o,r) \ (\text{by definition})$

$\leq (1+\varepsilon) \text{Min } r' \text{ s.t. } \exists o \ f(P) \subseteq B(f(o),r')$

(contraction by at most $1+\varepsilon$)

$= (1+\varepsilon) \text{Min } r' \text{ s.t. } \exists o \ \{f(p_1)\ldots f(p_{2d'})\} \subseteq B(f(o),r')$

(a set of points $f(P)$ is contained in a hypercube iff the extreme points of $f(P)$ are contained in that hypercube)

$\leq (1+\varepsilon) \text{Min } r \text{ s.t. } \exists o \ \{p_1\ldots p_{2d'}\} \subseteq B(o,r)$

(no expansion of $f$)

$= (1+\varepsilon) \text{MEB radius for } \{p_1\ldots p_{2d'}\}$
Digression: Core Sets

• In the previous slide we use the fact that in $l_\infty$, for any set $P$ of points, there is a subset $P'$ of $P$, $|P|=2d'$, such that
  \[ \text{MEB}(P')=\text{MEB}(P) \]
• $P'$ is called a “core-set” for the MEB of $P$ in $l_\infty$
• For more on core-sets, see the web page by Sariel Har-Peled