

Analysis of Incomplete Data and an Intrinsic-Dimension Helly Theorem

Jie Gao

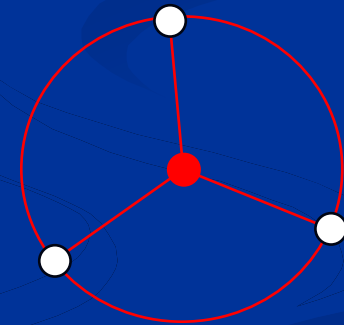
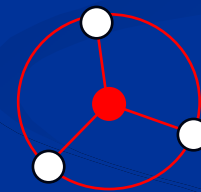
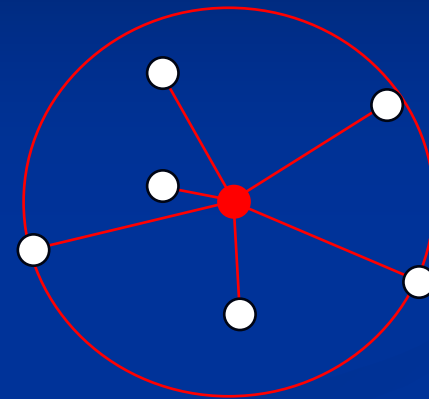
Stony Brook University

Joint work with Michael Langberg and Leonard Schulman @ Caltech

Clustering

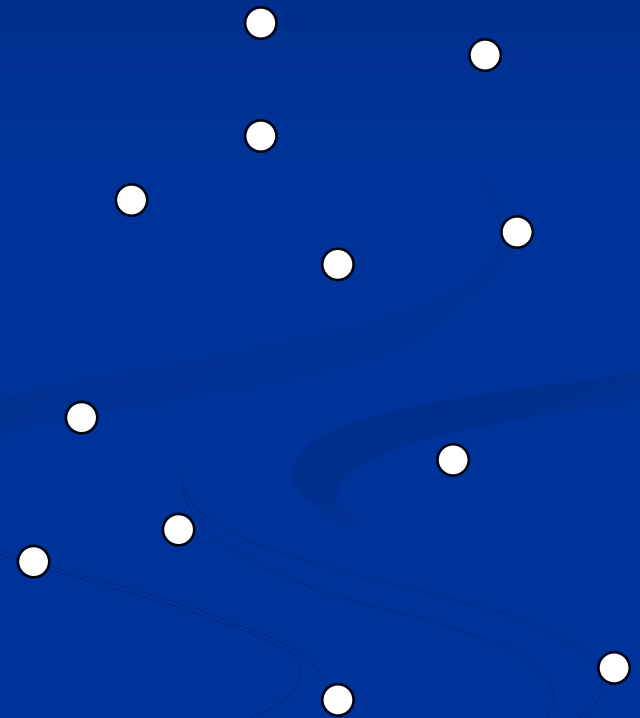
General framework:

- **Input:** data elements
 - Points in high dimension.
- **Objective:** extract pattern
 - **k**-median centers (min: $\Sigma dist$).
 - **k**-means centers (min: $\Sigma dist^2$).
 - **k**-center (min: *maximum dist*).
- **Extensively studied:**
 - Roughly speaking: NP-hard.
 - Exact and approximate algorithms studied.



This work: incomplete data

- Typically:
 - Data elements are points.
 - Represent **complete** data.
 - Answers to questionnaire.
 - Sensor readings.
- This talk - **incomplete** data:
 - **Incomplete** data:
 - Blank question in questionnaire.
 - Missing reading from sensor.



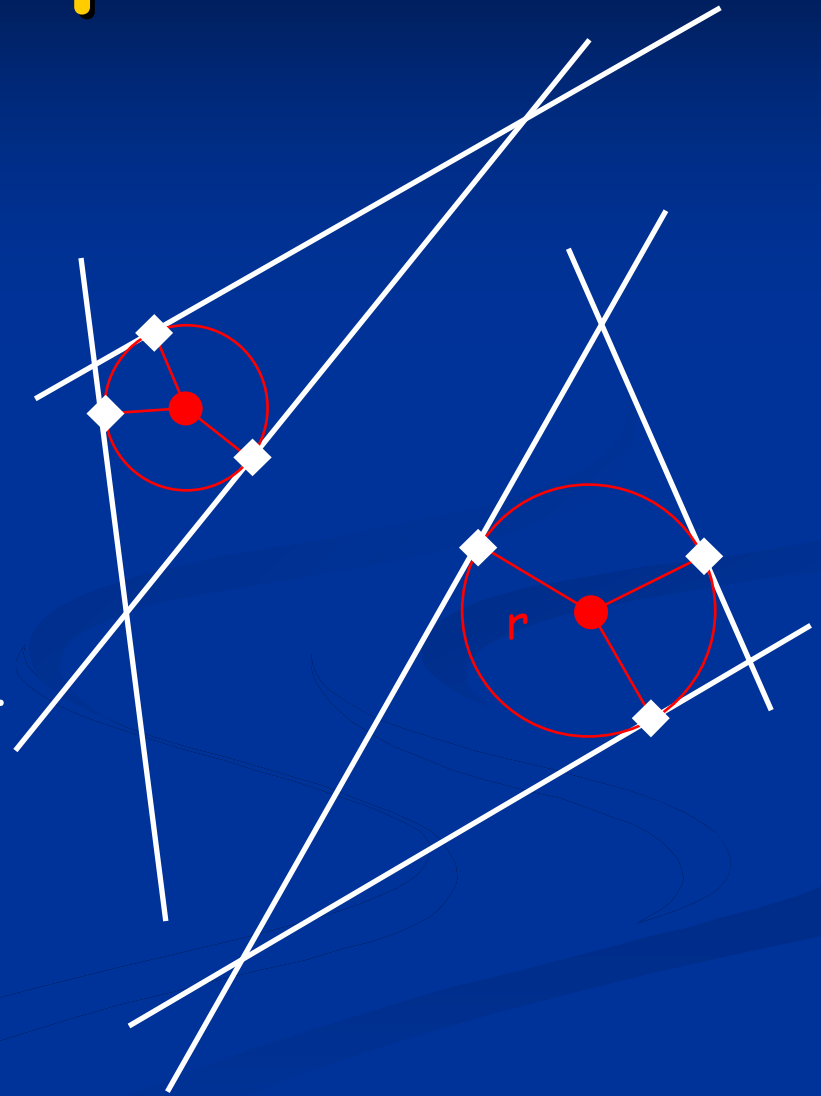
Model: incomplete data

- **Incomplete data:**
 - Blank question in questionnaire.
 - Missing reading from sensor.
- Model data elements as **lines** (or flats of higher dimension).
 - Missing answer: axis parallel line.
 - We consider general lines: correspond to correlations.



k-center on incomplete data

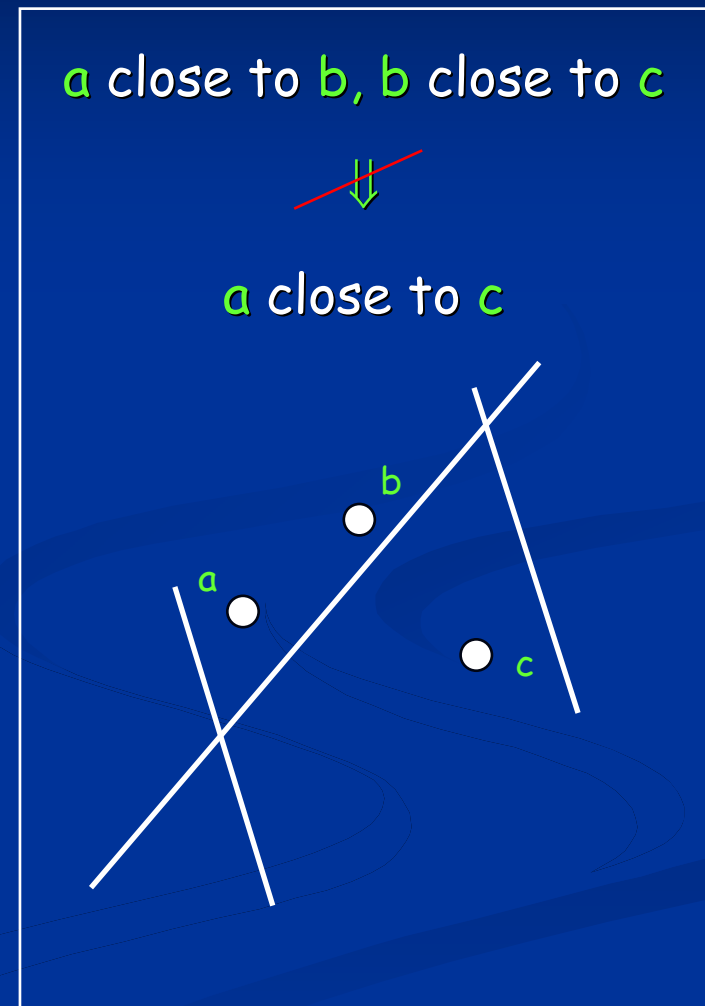
- Input:
 - n lines in \mathbb{R}^d .
 - Parameter k .
- Output:
 - k centers that minimize maximum dist. to lines.
 - k minimum radius balls that **cover** all lines.
- At least as hard as k-center for points, for general k .



Complete vs. Incomplete data

- Complete data (points in \mathbb{R}^d):
 - distances between points imply a **metric**.
 - **Locality** is crucial to success of clustering algorithms.
- Incomplete data (lines):
 - Distances between lines are **not** a metric.

Can one still argue locally?



Helly Theorem

Helly: In \mathbb{R}^d , if every set of $d+1$ convex sets intersect then all convex sets intersect.

- Suppose the lines can be covered by a radius r ball. We blow up each line to a cylinder with radius r .
- All the cylinders have a common intersection iff every $d+1$ cylinders have a common intersection.
- **Helly:** In \mathbb{R}^d , given n lines, if every set of $d+1$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius r .

Helly

In \mathbb{R}^2 : if every 3 lines intersect \Rightarrow all lines intersect.

- **Helly**: In \mathbb{R}^d , given n lines, if every set of $d+1$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius r .
- Not very useful as d can be very large (depends on n).
- **Goal**: remove the dependence on d in locality argument and apply to clustering.



Every three intersect



Every two intersect

Main idea

Can replace with any “convex set”

Helly: In \mathbb{R}^d , if every set of $d+1$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius r .

- Exploit the fact that we work on **lines (low-dim objects)** in **high dimensional (d-dim)** space.
- Intrinsic-dimension rather than dimension of the ambient space.



Our results

Helly: In \mathbb{R}^d , if every set of $d+1$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius r .

Independent of $d!$

- **Intrinsic Helly I:** In \mathbb{R}^d , given n lines, if every set of 3 lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $2r$. Relax quality
- **Intrinsic Helly II:** Given n lines, if every set of $\sim 1/\varepsilon^2$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $(1+\varepsilon)r$.

Theorem is constructive

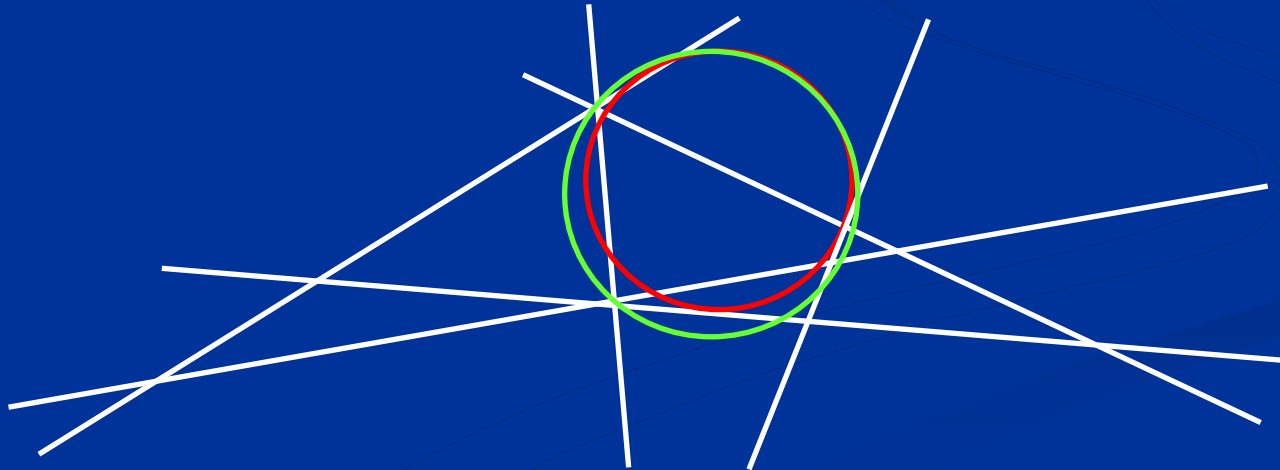
Intrinsic Helly I: Given n lines, if every set of 3 lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $2r$.

Intrinsic Helly II: Given n lines, if every set of $\sim 1/\epsilon^2$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $(1+\epsilon)r$.

- Given set of n lines with minimum enclosing ball of radius at least r , we find a subset of lines of size $\sim 1/\epsilon^2$ with minimum enclosing ball of radius at least $(1-\epsilon)r$ in time $\sim n \cdot \text{poly}(1/\epsilon)$.

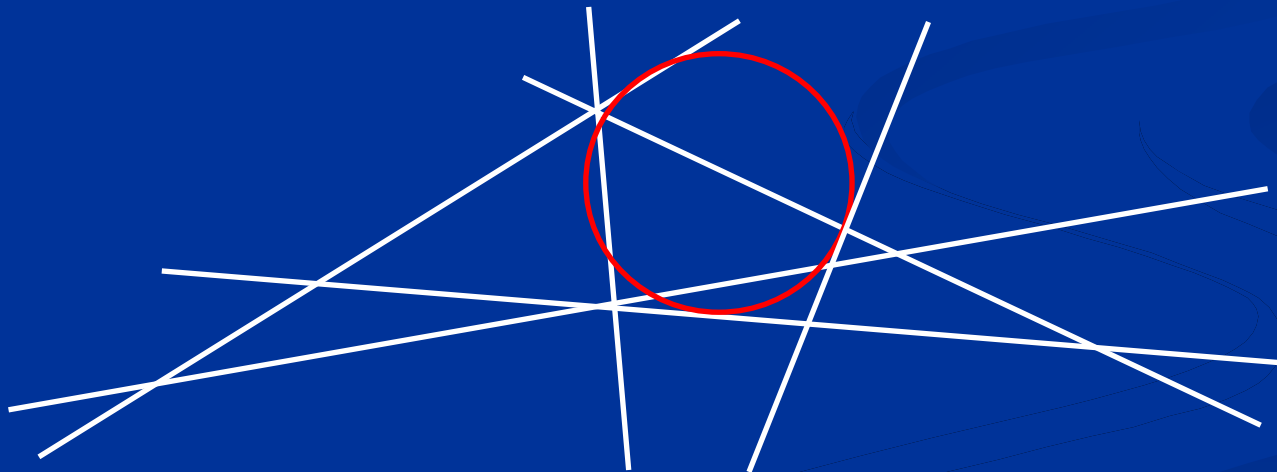
Application to clustering

- Minimum enclosing ball of lines (1-center)
 - **Input:** n lines.
 - **Output:** minimum radius ball covering all lines.
- Can be computed by convex programming in time $\sim n^{3/2}\text{poly}(d)$.
- **Using our algorithm:** can find approximation to radius of minimum enclosing ball within ratio $(1+\epsilon)$ in time $\sim n \cdot \text{poly}(1/\epsilon)$.



What next?

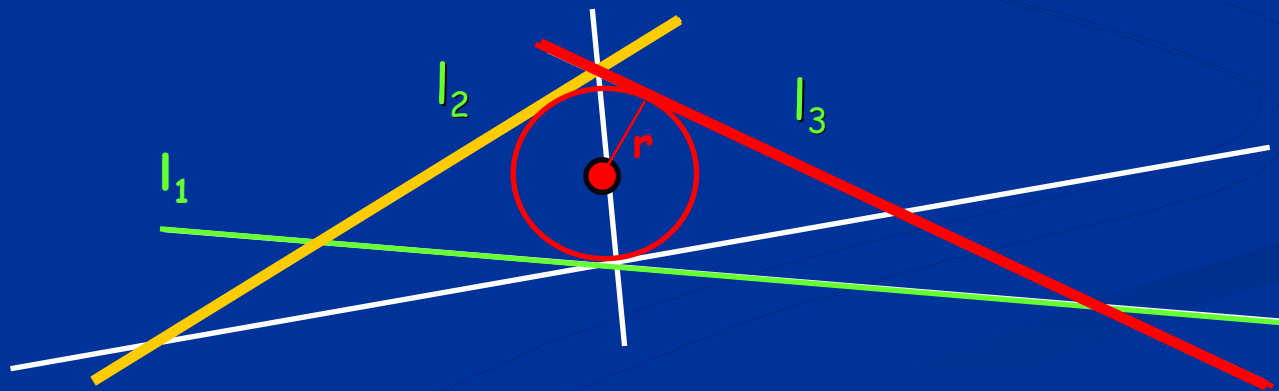
- Prove **Intrinsic Helly I** (**2** approximation).
- Sketch proof for **Intrinsic Helly II** (**$(1+\epsilon)$** app.).



2 approximation (Theorem I)

If every set of 3 lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $2r$.

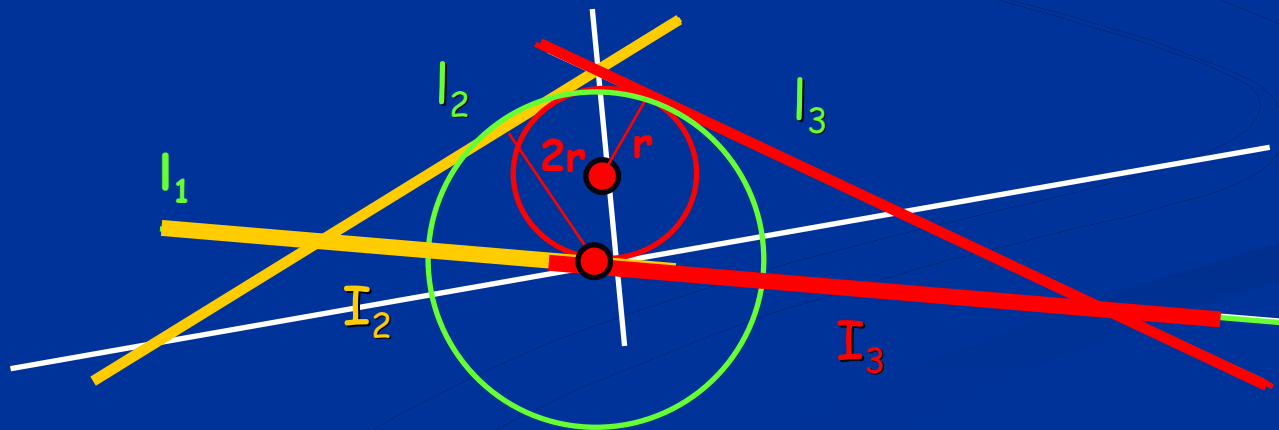
- **Claim:** We can find a ball of radius $2r$ that covers every line.
- Pick arbitrary line l_1 .
- Consider 2 lines l_2, l_3 , there is a ball centered at o with radius r that covers the three lines l_1, l_2, l_3 .



Proof

Helly: In \mathbb{R}^d , if every set of $d+1$ convex sets intersect then all convex sets intersect.

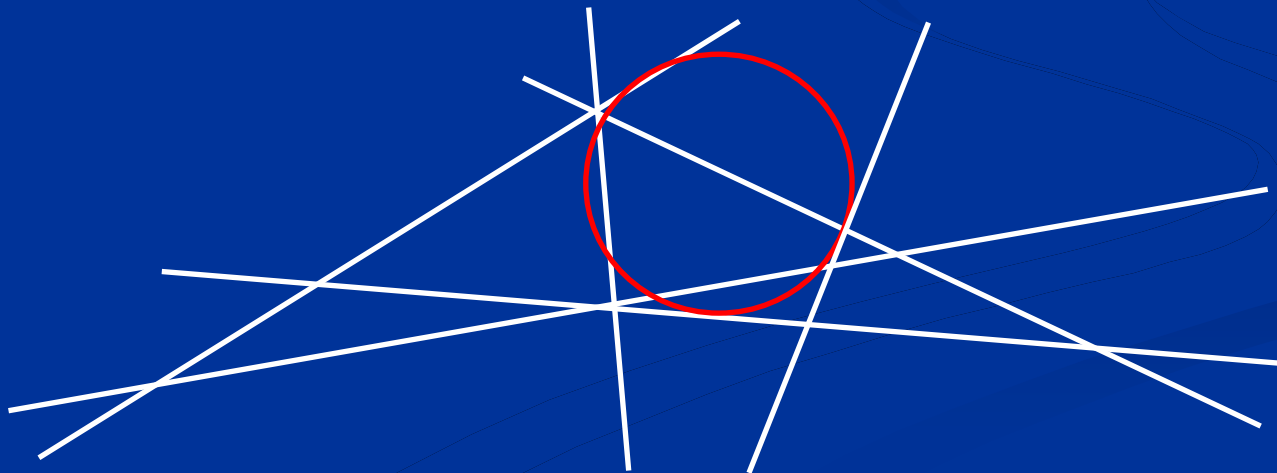
- For each of the 2 lines l_2, l_3 , define interval I_2, I_3 on l_1 consisting of points within distance $2r$ from l_2, l_3 , respectively.
- **Claim:** the two intervals I_2, I_3 intersect.
- In fact, any two intervals I_i, I_j for line l_i, l_j , intersect.
- By Helly theorem, all $n-1$ intervals intersect \Rightarrow take the ball centered at a point in the common intersection. It cover all the lines with radius $\leq 2r$. QED



$(1+\epsilon)$ approx. (Theorem II)

If every set of $\sim 1/\epsilon^2$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $(1+\epsilon)r$.

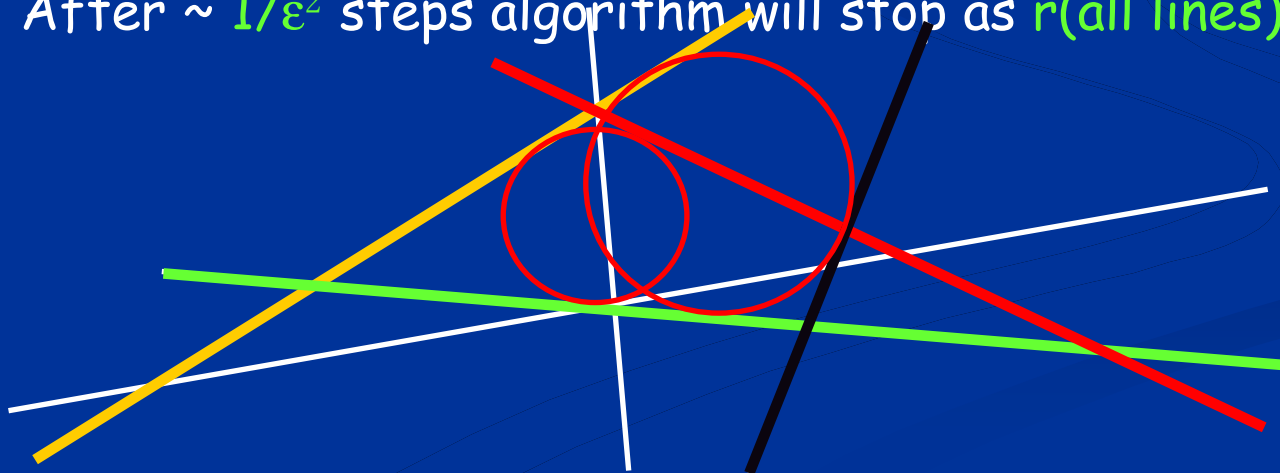
- **Input:** n lines (optimal ball that covers has radius 1).
- **Output:** $\sim 1/\epsilon^2$ lines that have covering radius $\geq (1-\epsilon)$.



Algorithm

If every set of $\sim 1/\varepsilon^2$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $(1+\varepsilon)r$.

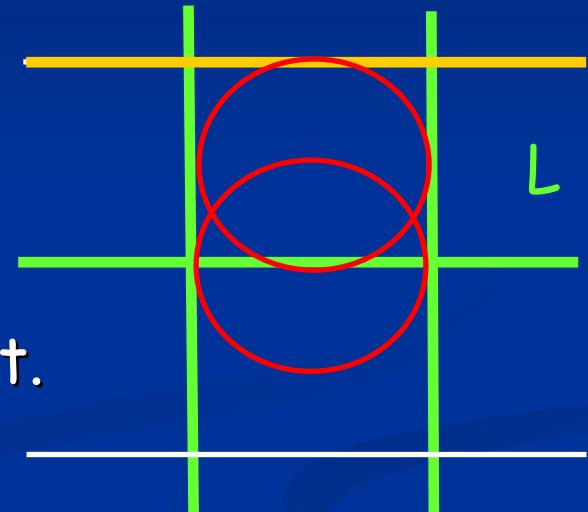
- We wish to follow the alg for points [BHI02].
- Start with 3 lines which are 2 approximation, denote set by L .
- These lines have minimum cover radius of at least $1/2$: $r(L) \geq 1/2$.
- Find additional line l s.t. $r(L+l) \gg r(L)$.
- **Claim:**
 - If $r(L) < 1-\varepsilon$, find a line l for which $r(L+l) \geq r(L)(1+\varepsilon^2)$.
 - After $\sim 1/\varepsilon^2$ steps algorithm will stop as $r(\text{all lines})=1$.



Algorithm

If every set of $1/\varepsilon^2$ lines can be covered by a ball of radius r then all lines can be covered by a ball of radius $(1+\varepsilon)r$.

- **Claim:** If $r(L) < 1-\varepsilon$, always exists line l for which $r(L+l) \geq r(L)(1+\varepsilon^2)$.
- Actually claim is **not** true:
- **Can prove** existence of **two** lines s.t. $r(L+l_1+l_2) \geq r(L)(1+\varepsilon^2)$.
 - Technically involved
- **Algorithm:** in each iteration find **2** lines that will increase the radius substantially. After $\sim 1/\varepsilon^2$ steps we are done. QED



Extensions to Δ -dim flats

- **Intrinsic Helly I:** In \mathbb{R}^d , given n Δ -dim flats, if every set of $\Delta+2$ flats is covered by a ball of radius r then all flats can be covered by a ball of radius $2r$.
Tight!
- **Intrinsic Helly II:** Given n Δ -dim flats, if every set of $\sim \Delta^4/\varepsilon^2$ flats can be covered by a ball of radius r then all flats can be covered by a ball of radius $(1+\varepsilon)r$.

Connection to LP-type problems

Intrinsic Helly I: In \mathbb{R}^d , given n Δ -dim flats, if every set of $\Delta+2$ flats can be covered by a ball of radius r then all flats can be covered by a ball of radius $2r$.

- Minimum intersecting ball is an LP-type problem with combinatorial dimension $d+1$.
- With 2-approx. we can restrict the center to be in one of the Δ -dim flats \rightarrow LP-type problem has combinatorial dimension $\Delta+1$.
- Nina Amenta [1996] showed LP-type problems with combinatorial dimension c have Helly number $\leq c+1$.
- This is another proof of **Intrinsic Helly I**.

Concluding remarks

- Initiated study of clustering on incomplete data.
- Presented **Intrinsic dimension Helly theorems + applications** to $(1+\epsilon)$ -approximate 1-center clustering.
- Existence of small representative subset (core-set).
- The 1-center result also implies preliminary results on $(1+\epsilon)$ -approximate **k-center** problem on lines with running time $\sim (n/\epsilon^{d+1})^k$.

Facility location for mobile clients

- Input:
 - n clients moving along lines in \mathbb{R}^d .
 - Parameter k .
- Output:
 - k gas stations such that each client has a nearby gas station.
- At least as hard as k -center for points, for general k .

