CS602

Assignment-1

Saksham Rathi

3rd year, B.Tech CSE

22B1003

CS 602 - Applied Algorithms: Assignment 1

Total Marks - 50

Instructions. Please try to be brief, clear, and technically precise. Use pseudocodes to describe the algorithms. To solve the problems, one may assume that the instances are in general position, unless stated otherwise. Novelty in the answer carries marks.

1. Given a point set P, we would like to perform a k-median clustering of it, where we are allowed to ignore m of the points. These m points are *outliers* which we would like to ignore since they represent irrelevant data. Unfortunately, we do not know the m outliers in advance. It is natural to conjecture that one can perform a local search for the optimal solution. Here one maintains a set of k centers and a set of m outliers. At every point in time the algorithm moves one of the centers or the outliers if it improves the solution.

Show that local search does not work for this problem; namely, the approximation factor is not a constant. [**15 Marks**]

2. Show that Sauer's lemma is tight. Specifically, provide a finite range space that has the number of ranges as claimed by the lemma. [**15 Marks**]

3. Given a set of n unit disks $\mathcal{S}$ in the plane, we are interested in the problem of computing the minimum length TSP that visits all the disks. Here, the tour has to intersect each disk somewhere.

1. Let P be the set of centers of the disks of $\mathcal{S}$. Show how to get a constant factor approximation for the case that $diam(P) \leq 1$. [**10 Marks**]
2. Extend the above algorithm to the general case. [**10 Marks**]

1. k-median clustering with m-outliers

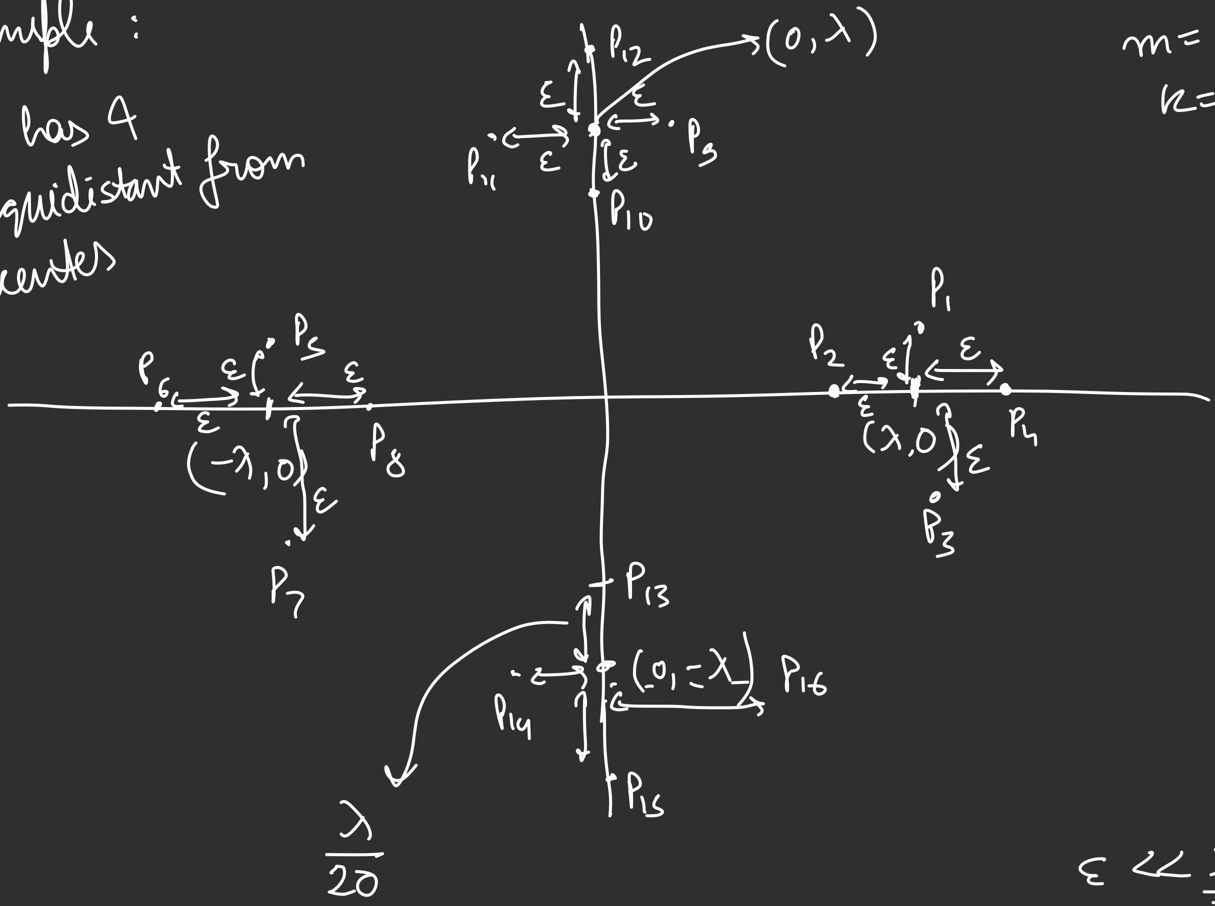
Now, to show that the approximation factor for local search is not constant, we will have to create a suitable example.

One possible example design is (k+1) sets of points, where in each set, the points are closely to each other, and each set has size = m. The local search will not be able to get out of such a set. And then, we can have the intra-set distances different for different sets.

For example:

Each set has 4 points equidistant from some centers

$m = 4$
$k = 3$



$\underbrace{\varepsilon \ll \frac{\lambda}{20}}$
ε is very small.

Clearly, local search algorithm might get stuck in the cluster $(P_{13}, P_{14}, P_{15}, P_{16})$ and mark some other cluster points (eg: P_1, P_2, P_3, P_4) as outliers.
But this solution will be far from the optimal solution which involves $(P_{13}, P_{14}, P_{15}, P_{16})$ as outliers and the other 3 as $k=3$ clusters. And since the ratio of λ and ε can be varied, the approximation factor is not constant.

2. <u>Sauer's Lemma</u> : If (X, R) is a range space of VC-dimension δ with $|X| = n$, then $|R| \leq G_\delta(n)$

where $$G_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i}$$

<u>Proof of Tightness</u> : We basically need to construct an example where the equality holds.

Consider $Y = \mathbb{R}$ → set of all real numbers (basically the real line)

and R = set of all intervals.

Consider $X \subseteq Y$

where X = set of n points $= \{x_1, x_2, \ldots, x_n\}$

We already knew that the VC-dimension of the range space (real line, intervals) is of size two. [No subset of size = 3 can be shattered].

We need to find $|R|$. To realize empty set (ϕ), there must be an interval which does not intersect with any of these n points. Then, to get singleton points, there should be intervals which intersect with only one of the points. (n in number). Then to get two point intersection, we can choose intervals of the form $[x_i, x_j]$ $x_i \leq x_j$ $1 \leq i < j < n$ (Number of intervals of this form $= \frac{n(n-1)}{2}$)

$$|R| = 1 + n + \frac{n(n-1)}{2}$$

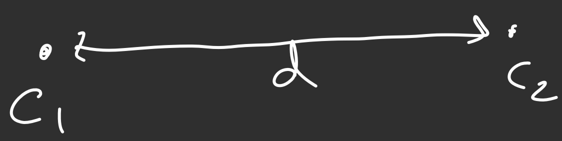
$$G_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} = 1 + n + \frac{n(n-1)}{2} = |R|$$
$(\delta = 2)$

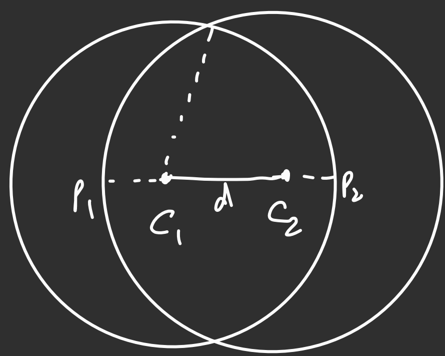
Hence, we have shown that Sauer's Lemma is tight.

3. (1) We have n unit circles

Consider a pair of circles:

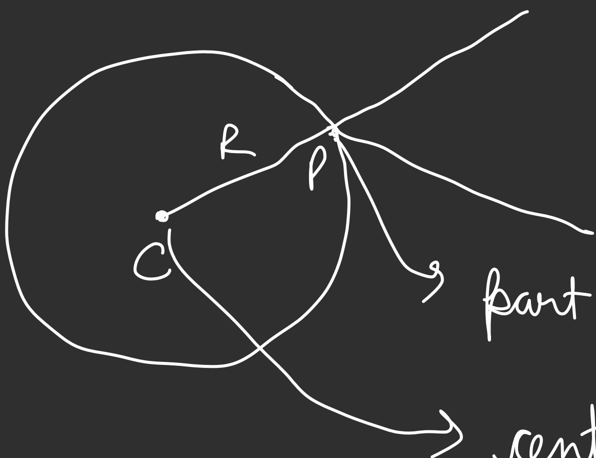
$d \le 1$ (diam ≤ 1 given)





If instead we calculate the optimal tour between the centres instead of circles:

$$2 \times OPT_{centres} \geq \text{Tour we calculate} \quad \text{(from class)}$$



part of the optimal tour

centre

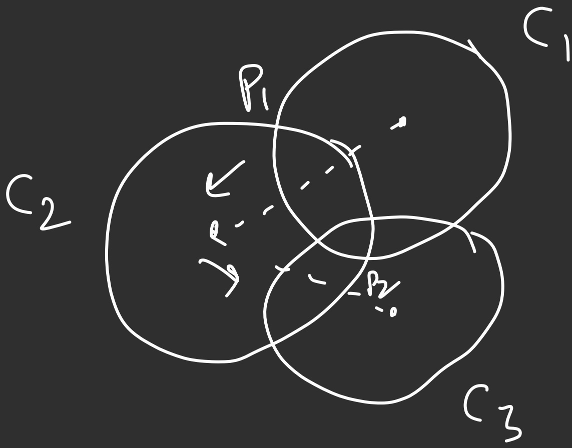
We are maximally adding 2R to the optimal tour for each circle

So,

$$opt_{centres} \le opt_{circles} + 2R \times \underset{n \text{ circles}}{n}$$

$$\text{Tour we calculate} \le \underbrace{2 \times OPT_{circles}}_{\text{best tour possible}} + \underbrace{4Rn}_{\text{we need to approximate this}}$$

If in the optimal tour of centres, we have the subsequence C_1, C_2, C_3 : (Arrows show direction of tour)

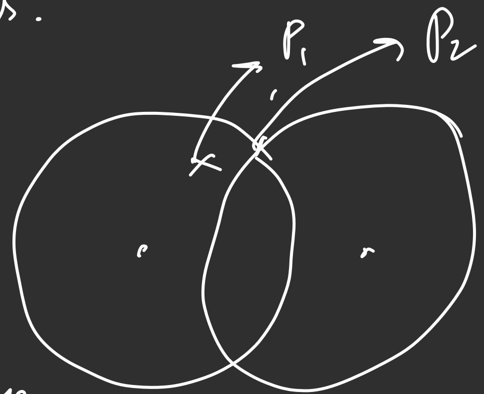


If instead of centres, we move our attention to the intersection points, then it might help.

For example replacing $C_1 \to C_2$ by $C_1 \to P_1$ or $C_1 \to P_2$ might help.

(2) In this part, we need to generalize from circles to disks and we have no constraint on diam (P).

It can be easily shown that the case of disks reduces to that of circles.



Moving the point from inside to outside, we can find another optimal tour of $\leq$ the previous length

diam (P) can be reduced by scaling the entire thing to make it ≤ 1

↳ we can use part (1) algo