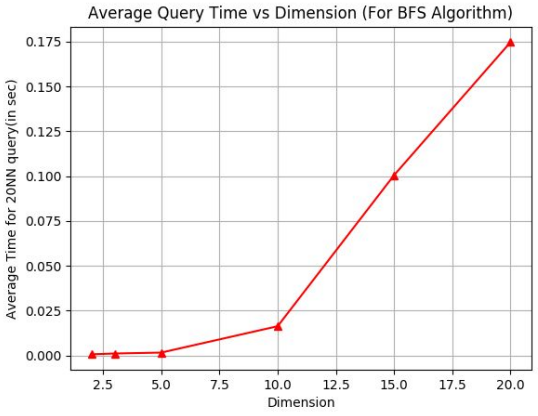
COL-362 : Assignment-3

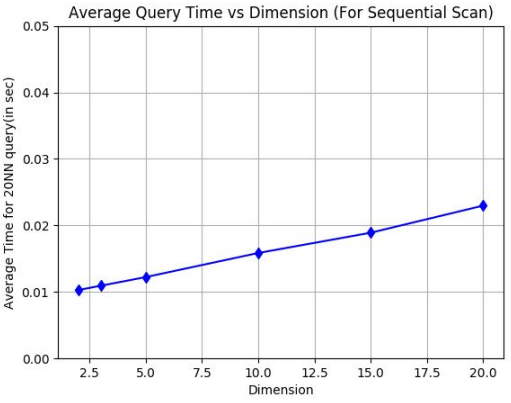
Project Report

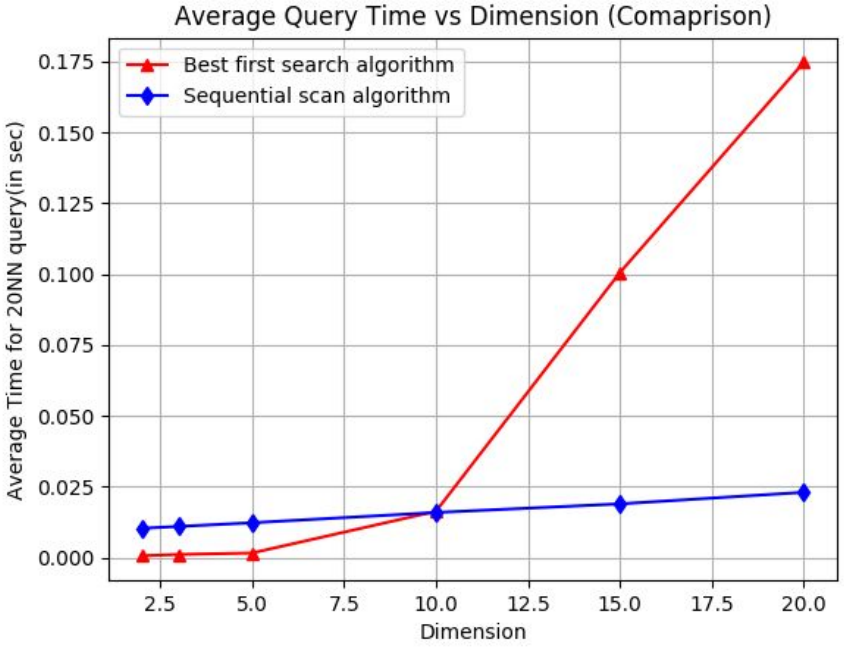
Submitted by:-

Ishu Jain – 2015EE30898
Harsh Malara - 2015EE30649
Yash Garg – 2015EE10691

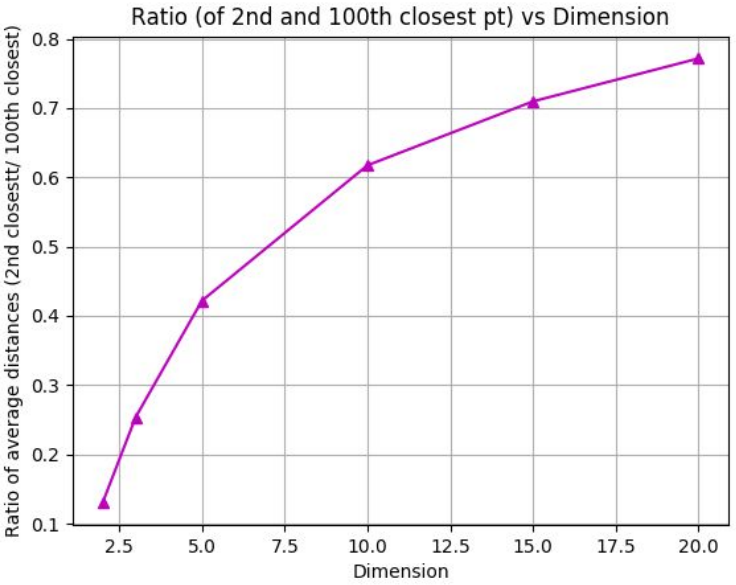
PART 2A)







PART 2B



PART 2C :

a)

In case of a KD Tree it is seen that as the dimensionality of the tree increases causes the working of the best first search(bfs algorithm) to degrade. This is due to the fact that 'PRUNING off' of branches decreases with increase in dimensions (That is because the sparseness of the tree increases). This causes the traversal for kNN query to almost every point in a tree. So the BEST FIRST algorithm gets similar order as SEQUENTIAL query ie O(nd*log(k)) .

But there are two more factors that are responsible in causing the visible effect in the comparison graph i.e.

1.Overhead in traversal in tree –

as that is implemented by pointers and they have their overhead ,which is not prominent in the case of sequential query.

2. Also in case of sequential query only the L2 distance is computed for every point , but in case of BFS : MBR and L2 is computed for each node which further adds up in the time delay.

b)

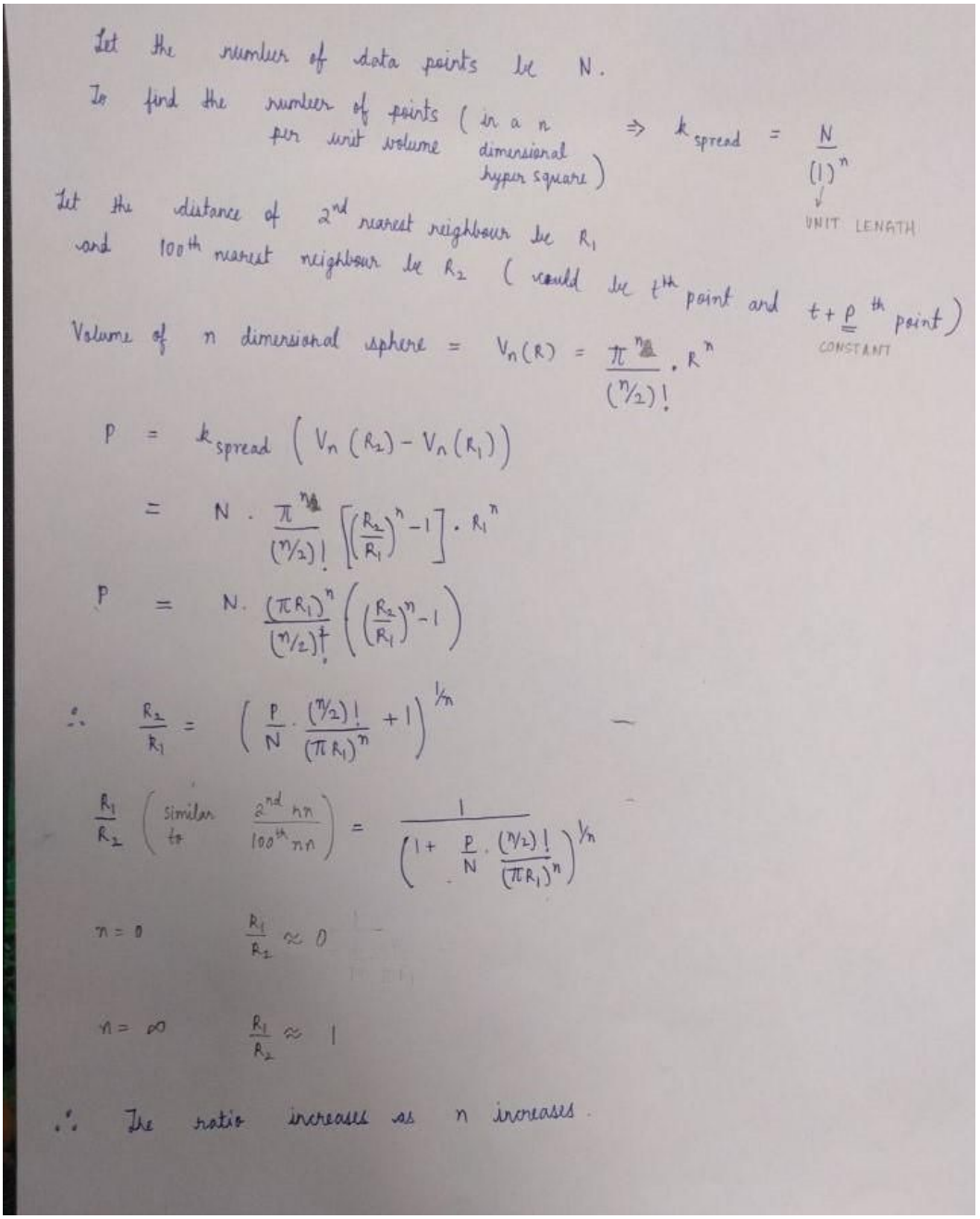
In the case of 2b, a variation of the ratio of average distance of 2^{nd} nearest point and 100^{th} nearest point has to be discussed. A 100NN query is done on structures of different dimensions and the results are reported in the graph.

We can see as the dimensions increases the ratio value increases, the graph is similar to an exponential tending to a constant value. This effect can be attributed to the fact that as the number of dimensions increases (in case of a hyper-sphere) the amount of the total volume towards the sphere surface increases. This more volume towards the surface causes the distance between the 2^{nd} point and 100^{th} point to decrease thereby causing the value of the ratio to increase. This is a phenomenon usually encountered in machine learning literature as CURSE OF DIMENSIONALITY.

The more volume near the surface is an analogy of having more number of points near the surface (volume is proportional to number of data points as the number of points per unit volume is uniform).

It can be estimated that as the dimensions tend towards infinity the ratio tends to 1.

A little mathematical motivation is given below :-



Let the number of data points be N.

To find the number of points (per unit volume in a n dimensional hyper square) $\Rightarrow k_{spread} = \frac{N}{(1)^n}$ (1 → UNIT LENGTH)

Let the distance of 2nd nearest neighbour be R_1 and 100th nearest neighbour be R_2 (would be t^{th} point and $t+\underline{p}^{th}$ point) (p → CONSTANT)

Volume of n dimensional sphere $= V_n(R) = \frac{\pi^{n/2}}{(n/2)!} \cdot R^n$

$$P = k_{spread}\left(V_n(R_2) - V_n(R_1)\right)$$

$$= N \cdot \frac{\pi^{n/2}}{(n/2)!}\left[\left(\frac{R_2}{R_1}\right)^n - 1\right] \cdot R_1^n$$

$$P = N \cdot \frac{(\pi R_1)^n}{(n/2)!}\left(\left(\frac{R_2}{R_1}\right)^n - 1\right)$$

$$\therefore \frac{R_2}{R_1} = \left(\frac{P}{N} \cdot \frac{(n/2)!}{(\pi R_1)^n} + 1\right)^{1/n} \quad -$$

$$\frac{R_1}{R_2}\left(\text{similar to } \frac{2^{nd}\ nn}{100^{th}\ nn}\right) = \frac{1}{\left(1 + \frac{P}{N} \cdot \frac{(n/2)!}{(\pi R_1)^n}\right)^{1/n}}$$

$n = 0 \qquad \frac{R_1}{R_2} \approx 0$

$n = \infty \qquad \frac{R_1}{R_2} \approx 1$

$\therefore$ The ratio increases as n increases.