An optimal algorithm for the All–Nearest–Neighbors problem

Pravin M. Vaidya

Department of Computer Science
University of Illinois at Urbana–Champaign
Urbana, IL 61801

Abstract

Given a set V of n points in k–dimensional space, and an L_q–metric (Minkowski metric), the All–Nearest–Neighbors problem is defined as follows: For each point p in V, find all those points in $V-\{p\}$ that are closest to p under the distance metric L_q. We give an $O(n \log n)$ algorithm for the All–Nearest–Neighbors problem, for fixed dimension k and fixed metric L_q. Since there is an $\Omega(n \log n)$ lower bound, in the algebraic decision tree model of computation, on the time complexity of any algorithm that solves the All–Nearest–Neighbors problem (for $k=1$), the running time of our algorithm is optimal upto a constant.

1. Introduction

The All–Nearest–Neighbors problem is one of the very well studied proximity problems in computational geometry [2, 3, 4, 5, 7]. We are given a set V of n points in k–dimensional space and a distance metric L_q. Each point x is given as a k–tuple of real numbers $(x_1, x_2, ..., x_k)$. The L_q distance between a pair of points x, y, is given by $(\sum_i |x_i - y_i|^q)^{\frac{1}{q}}$ (Note that the L_∞ distance between x and y is given by $\max_i |x_i - y_i|$). The nearest (closest) neighbors of a point $p \in V$ are all those points in V that are closest to p under the distance metric L_q. The All–Nearest–Neighbors problem is defined as follows: For each point p in V, find the nearest (closest) neighbors of p. We assume that the dimension k and the distance metric L_q are fixed. We shall use distance for L_q distance, and $d(x, y)$ to denote the distance between x and y.

As far as the model of computation is concerned we assume that all arithmetic, comparison and memory access operations require constant time.

The simplest algorithm for the All–Nearest–Neighbors problem may be phrased as follows: For each point p in V, explicitly test if every point p' in $V-\{p\}$ is a closest neighbor of p. This algorithm runs in time $O(n^2)$. However, it is possible to obtain algorithms that require $o(n^2)$ time. In [2], Bentley utilises multidimensional divide and conquer to develop an $O(n (\log n)^{k-1})$ algorithm for the All–Nearest–Neighbors problem. An $O(n \log_2 \delta)$ algorithm is presented in [4], where δ is the ratio of the maximum to the minimum distance between a pair of points in V.

This research was supported by a fellowship from the Shell Foundation

We give an $O(n \log n)$ algorithm for the All–Nearest–Neighbors problem. In the algebraic decision tree model of computation, there is a lower bound of $\Omega(n \log n)$ on the time complexity of any algorithm that solves the All–Nearest–Neighbors problem for dimension $k=1$ [1, 5]. So the running time of our algorithm is optimal upto a constant.

A generalization of the All–Nearest–Neighbors problem is defined as follows. For a point $p \in V$, let $D(p)$ be the multiset of distances defined by $D(p) = \{ z : z = d(p, p'), p' \in V, z \neq 0 \}$, and let $d_1(p) \leq d_2(p) \leq .... \leq d_m(p)$ be the m smallest distances in $D(p)$. Then the m–Nearest–Neighbors of p are all those points in V whose distance from p is at most $d_m(p)$. The All–m–Nearest–Neighbors problem is as follows: For every point p in V, find the m–Nearest–Neighbors of p. A slight modification of our algorithm for the All–Nearest–Neighbors problem leads to an $O(m\, n \log n)$ algorithm for the All–m–Nearest–Neighbors problem.

We note that if the metric in the given problem is positive definite or semidefinite rather than one of the standard L_q metrics, the problem can be transformed in linear time to a problem with L_2 (euclidean) metric without increasing the dimension.

We define a box b to be the product $J_1 \times J_2 \times ..... \times J_k$ of k intervals (either closed, semi–closed or open), or equivalently, the set of those points $x = (x_1, x_2, ..., x_k)$ such that x_i lies in the interval J_i, for $i = 1, ..., k$. A box is a cubical box iff all the k intervals defining the box are of identical length. The centre $\alpha(b)$ of a box b is defined to be the point $(\alpha_1(b), ..., \alpha_k(b))$ where $\alpha_i(b)$ is the centre of the *ith* interval defining the box, for $i = 1, ..., k$. For a hyperplane $h = \{ x : \beta^T x = \gamma \}$, we define $L(h)$ to be the left open half–space $\{ x : \beta^T x < \gamma \}$, and $R(h)$ to be the right closed half–space $\{ x : \beta^T x \geq \gamma \}$.

2. An Overview

In the algorithm we maintain a collection B of disjoint closed cubical boxes which contain all the n points in the given set V. Each box has been shrunk as much as possible so that further shrinkage would either destroy the cubical shape of the box or push out of the box a point in V that was originally located in the box. With each box $b \in B$, there are associated two subets of B, *Neighbors*(b) and *Attractors*(b) such that

(1) For each point $p \in V$, every nearest neighbor of p is located in box b itself or in some box in *Neighbors*(b).

0272-5428/86/0000/0117$01.00 © 1986 IEEE

(2) $Attractors(b) = \{ b' : b \in Neighbors(b') \}$. Thus if $p \in b \cap V$ is the nearest neighbor of some point $p' \in V$ then p' must be located in box b itself or in some box in $Attractors(b)$.

In the beginning, B contains a single box which is a smallest cubical box containing all the points in V. At the end, every box $b \in B$ is degenerate and consists of a single point in V i.e. $|b| = |b \cap V| = 1$, and $Neighbors(b)$ is exactly the set of nearest neighbors of the unique point in $b \cap V$.

Let $d_{\max}(b)$ denote the greatest distance between a pair of points in b, and let $d_{\max}(b, b')$, $d_{\min}(b, b')$, respectively denote the maximum and minimum distance between a point in b and a point in b'. For each box $b \in B$, we maintain a parameter $Estimate(b)$ given by

$$Estimate(b) = \begin{cases} d_{\max}(b), & \text{if } |b \cap V| \geq 2 \\ \min\limits_{b' \in Neighbors(b)} \{ d_{\max}(b, b') \}, & \text{otherwise} \end{cases}$$

For each point $p \in b \cap V$, $Estimate(b)$ serves as an upper bound on the distance between p and a nearest neighbor of p. In order to eliminate unnecessary boxes from $Neighbors(b)$, we ensure that $Neighbors(b)$ satisfies the invariant

$$\forall\, b' \in Neighbors(b),\ d_{\min}(b, b') \leq Estimate(b)$$

The algorithm proceeds in stages. At the beginning of each stage we choose a box b in B that has the largest volume (size) among all the boxes in B for splitting. The chosen box b is split into 2^k cubical boxes $b_1, b_2, \ldots, b_{2^k}$, by k mutually orthogonal hyperplanes passing through its centre $\alpha(b)$, each hyperplane being perpendicular to one of the co-ordinate axes. Simultaneously, the set of points $b \cap V$ is split into the sets of points $b_1 \cap V, b_2 \cap V, \ldots, b_{2^k} \cap V$. Out of these 2^k boxes we discard those which do not contain a point in V, and each of the remaining boxes is shrunk as much as possible to obtain the set $successors(b)$. The shrinking process ensures that when b is split the set of data points $b \cap V$ is also split, or equivalently shrinking guarantees that there are at least two boxes in $successors(b)$. We then remove the split box b from B and add to B all the boxes in $successors(b)$. Finally, we create the estimates, and the neighbor and attractor sets, for the boxes in $successors(b)$; update the estimates and the neighbor sets for the boxes in $Attractors(b)$; and update the attractor sets for the required boxes in B. That concludes a stage. For each $b \in B$, at the end of a stage $Neighbors(b)$ satisfies the above described invariant.

Throughout the algorithm $Estimate(b)$ serves as an approximation to the distance between a point $p \in b \cap V$ and a nearest neighbor of p, and the volume (size) of b is a measure of the error in this aproximation. By splitting a box that has the largest volume, an estimate that has the largest error is refined.

3. The Algorithm

Let $h_i(b)$ be the hyperplane defined by $h_i(b) = \{ x : x_i = \alpha_i(b) \}$, where $\alpha(b) = (\alpha_1(b), \alpha_2(b), \ldots, \alpha_k(b))$ is the centre of box b. Let $L(h_i(b))$ be the left open halfspace $\{ x : x_i < \alpha_i(b) \}$ and $R(h_i(b))$ be the right closed halfspace $\{ x : x_i \geq \alpha_i(b) \}$. Let $Immediate\text{-}successors(b)$ be the set of boxes defined by

$$Immediate\text{-}successors(b) = \{ b' : b' = b \cap f_1 \cap \ldots \cap f_k, \text{ where } f_i = L(h_i(b)) \text{ or } f_i = R(h_i(b)),\ 1 \leq i \leq k \}.$$

Corresponding to a cubical box b, let $shrunk(b)$ be a cubical box such that

(1) If $|b \cap V| \leq 1$ then $shrunk(b) = b \cap V$.

(2) If $|b \cap V| \geq 2$ then (i) $shrunk(b) \subseteq b$, (ii) $shrunk(b) \cap V = b \cap V$, and (iii) the maximum L_∞ distance between a pair of points in $shrunk(b) \cap V$ equals the size of $shrunk(b)$.

We now give the algorithm for the All–Nearest–Neighbors problem.

Algorithm All–Nearest–Neighbors

1. $Neighbors(b_0) := \phi$, $Attractors(b_0) := \phi$, $B := \{b_0\}$,

   where b_0 is a smallest cubical box containing all the n points in V.

2. Repeat Steps 2.1 through 2.6 until each box $\hat{b} \in B$ satisfies $|\hat{b}| = |\hat{b} \cap V| = 1$

   i.e. until each box $\hat{b} \in B$ is degenerate and consists of a single point in V.

   2.1. Choose for splitting, a box $b \in B$ that has the largest volume among all the

   boxes in B.

   2.2. Split the set of points $b \cap V$ into the sets of points $b_1 \cap V, b_2 \cap V, \ldots, b_{2^k} \cap V$,

   where $b_1, b_2, \ldots, b_{2^k}$, are the boxes in $Immediate\text{-}successors(b)$.

   2.3. $successors(b) := \{ b' : b' = shrunk(b''),\ b'' \in Immediate\text{-}successors(b),\ |b'' \cap V| \geq 1 \}$

   2.4. For each box $b' \in successors(b)$,
   create $Neighbors(b')$ and $Estimate(b')$.

   For each box $b' \in Attractors(b)$,
   update $Neighbors(b')$ and $Estimate(b')$.

   2.5. For each box $b' \in successors(b)$,
   create $Attractors(b')$.

   For each box $b' \in (B - successors(b))$,
   update $Attractors(b')$.

   2.6. $B := B - \{b\} \cup successors(b)$

end All–Nearest–Neighbors

Let B^* be the set of all the boxes generated during the execution of the algorithm. The boxes in B^* form a tree in which the boxes are the vertices, the children of b are the boxes in $successors(b)$, and the leaves are all the points in the set V. Every box in this tree has at least 2 and at most 2^k children. Also the total number of boxes in B^* is at most $2n$.

The running time of the algorithm may be divided into

(i) Time to select a largest box at each stage summed over all stages.

(ii) Total time required to split the set of data points $b \cap V$, and shrink box b, for all boxes $b \in B^*$.

(iii) Total time required to maintain *Estimate*(b), *Neighbors*(b), and *Attractors*(b) for all boxes $b \in B^*$.

We maintain a heap [6] for the boxes in B. This allows us to pick in constant time a box in B that has the largest volume. Since $|B^*| \leq 2n$ the total time for heap maintainence is $O(n \log n)$. Then (i) is $O(n \log n)$.

In Section 4 we show that (ii) is $O(n \log n)$. To split the set of data points $b \cap V$ efficiently we utilise k ordered lists *List*$_i(b)$, $1 \leq i \leq k$, the *ith* list containing the points in $b \cap V$ ordered on the *ith* co-ordinate. The list *List*$_i(b)$ is embedded in the leaves of a complete binary search tree $T_i(b)$. When b is split, from the k ordered lists for b we obtain similar lists for all the boxes in *successors*(b), In Section 4 we show how to split the set of data points in a box so that the total time to split $b \cap V$ for all boxes $b \in B^*$ is $O(n \log n)$. Once the k ordered lists corresponding to a box are available the box can be shrunk in constant time. So the total time for shrinking all the boxes in B^* is $O(n)$.

In Section 5 we show that (iii) is $O(n \log n)$. Since at each stage we split a box that has the largest volume, and force the neighbor set of each box $b \in B$ to satisfy the invariant $\forall\, b' \in$ *Neighbors*(b), $d_{\min}(b, b') \leq$ *Estimate*(b), we can bound the sizes of the attractor sets of all the boxes in B, and the sizes of the neighbor sets of non-degenerate boxes in B, by some constant dependent on k. This in turn enables us to obtain a bound of $O(n)$ on the total number of additions to (insertions into) the neighbor and attractor sets of all the boxes in B^*. Since the size of the attractor set of each box in B is bounded by a constant dependent on k, we get that the time for maintaining attractor sets is $O(n)$. For a box b, we implement *Neighbors*(b) by a data structure which allows insertions and deletions to be performed in $O(\log n)$ time, and allows access to a box b', with the greatest value of the parameter $d_{\min}(b, b')$, in $O(\log n)$ time. Then the total time, for maintaining neighbor sets of all the boxes in B^*, can be shown to be $O(n \log n)$. Finally, only addition of a box to *Neighbors*(b) can change the parameter *Estimate*(b), and the change due to the addition of a single box can be computed in constant time. So the total time for maintaining the estimates is $O(n)$. This gives a bound of $O(n \log n)$ on (iii).

4. Splitting the set of data points in a box

In this section we describe how to split the set of data points $b \cap V$ in a box b efficiently, so that the total time required to split $b \cap V$ for all the boxes generated during the algorithm is $O(n \log n)$.

We shall make the following definitions. Let T be a rooted tree such that each non-leaf vertex has at least two children. For a vertex v in T, we define $m(v)$, $v_{\max}$, $s(v)$, and $lsa(v)$. Let $m(v)$ be the number of leaves in the subtree rooted at v, and let the largest son of v, denoted by $v_{\max}$, be a child of v such that for any child v' of v, $m(v_{\max}) \geq m(v')$. For a non-leaf vertex v let $s(v) = m(v) - m(v_{\max})$, and for a leaf vertex v let $s(v) = 0$. We define the lowest smaller ancestor of v, denoted by $lsa(v)$, as follows. If each vertex on the path from the root to v, other than the root, is the largest son of its father then $lsa(v)$ is the root; otherwise $lsa(v)$ is the lowest vertex v' on the path from the root to v such that v' is not the largest son of its father.

In the case of the tree of boxes generated during the algorithm, for a box $b \in B^*$, $m(b)$ is identical to the number of points in $b \cap V$, $b_{\max}$ is a box in *successors*(b) that contains the largest number of points in V among all the boxes in *successors*(b), and $s(b)$ equals the number of points in $(b - b_{\max}) \cap V$.

In order to bound the total amount of work for splitting sets of data points in boxes we shall require a lemma about weightings on tree vertices.

Tree Weighting Lemma. Let T be a rooted tree with t vertices such that each non-leaf vertex has at least 2 and at most r children. Define the weight $w(v)$ of a non-leaf vertex v by $w(v) = s(v)(1 + \log_2 m(lsa(v)) - \log_2 s(v))$ and define the weight of a leaf to be 0. Then $\sum_{v \in T} w(v) \leq 4r\, t \log t$.

Proof. The lemma is proved by induction on the number of leaves in T. The base case is when T consists of a single leaf vertex v and then $w(v) = 0$. So let the number of leaves in T be at least 2, then $t \geq 2$. Let P be a sequence of vertices in T as follows. The first vertex in P is the root of T, and for $j > 1$, the jth vertex in P is the largest son of the $(j-1)st$ vertex in P. The sequence terminates in a leaf. Let Q be the set of all vertices v such that v is the son of some vertex in P and v is not the largest son of its father. For a vertex v, let $T(v)$ denote the subtree rooted at v. We inductively assume that

$$\forall\, v \in Q,\ \sum_{u \in T(v)} w(u) \leq 4r\, m(v) \log m(v)$$

We note that for each vertex v in P, $lsa(v) =$ root, and so

$$\forall\, v \in P,\ w(v) = s(v)(1 + \log_2 t - \log_2 s(v))$$

As a vertex v has at most r children we have $s(v) \leq (1 - 1/r)\, m(v)$, and so

$$\forall\, v \in P,\ w(v) \leq 4r\, s(v)(\log_2 t - \log_2 s(v))$$

Then

$$\sum_{v \in T} w(v) = \sum_{v \in P} w(v) + \sum_{v \in Q} \sum_{u \in T(v)} w(u)$$

$$\leq \sum_{v \in P} 4r\, s(v)(\log_2 t - \log_2 s(v)) + \sum_{v \in Q} 4r\, m(v) \log m(v)$$

Now since $\sum_{v \in P} s(v) \leq t$, and $\sum_{v \in Q} m(v) \log m(v) \leq \sum_{v \in P} s(v) \log s(v)$, we get that $\sum_{v \in T} w(v) \leq 4r\, t \log_2 t$. ■

A proper ancestor of a leaf in a rooted tree is a non–leaf *vertex on the path* from the root to the leaf. We shall also require an upper on the number of proper ancestors of a set of leaves in a complete binary tree, and such a bound is provided by the following lemma.

Ancestors Lemma. Let T be a complete binary tree of height $h(T)$ containing $2^{h(T)}-1$ vertices and let L be a set of leaves in T. Let $Ancestors(L, T)$ be the set of all the proper ancestors in T of the leaves in L. Then $|Ancestors(L, T)| \leq |L|(h(T)-\log|L|)-1$.

Proof. The bound holds for $h(T)=2$. Let us assume that the upper bound holds for all complete binary trees containing at most $2^{h(T)-1}-1$ vertices. Let T_1 and T_2 be the subtrees rooted at the two sons of the root of T, and let L_1, L_2, be the sets of of leaves in L that are in T_1, T_2 respectively. Then

$$|Ancestors(L_1, T_1)| \leq |L_1|(h(T)-1-\log|L_1|) - 1$$
$$|Ancestors(L_2, T_2)| \leq |L_2|(h(T)-1-\log|L_2|) - 1$$

and

$$|Ancestors(L, T)| = |Ancestors(L_1, T_1)| + |Ancestors(L_2, T_2)| + 1$$

After some algebraic mainpulation it is seen that

$$|Ancestors(L, T)| \leq |L|(h(T)-\log|L|) - 1 - |L_1|+|L_2|\log(\frac{|L_1|+|L_2|}{|L_2|}) - |L_2|+|L_1|\log(\frac{|L_1|+|L_2|}{|L_1|})$$

The proof of the lemma follows from the observation that $\log(\frac{|L_1|+|L_2|}{|L_1|}) \leq \frac{|L_2|}{|L_1|}$ and $\log(\frac{|L_1|+|L_2|}{|L_2|}) \leq \frac{|L_1|}{|L_2|}$. ■

The boxes in B^* (i.e. the set of all the boxes generated during the algorithm) form a tree in which the vertices are the boxes in B^*, the children of every box b are the boxes in $successors(b)$, and the leaves are the points in V. In addition each non–leaf box in this tree has at least 2 and at most 2^k children, and the number of vertices in this tree is at most $2n$. Suppose we can split the set of data points $b \cap V$ in time proportional to $s(b)(1+\log_2 m(lsa(b))-\log_2 s(b))$. Then from the Tree Weighting Lemma it follows that the total time required to split $b \cap V$ for all $b \in B^*$ is $O(n \log n)$.

We now describe how to split $b \cap V$ in time proportional to $s(b)(1+\log_2 m(lsa(b))-\log_2 s(b))$. Corresponding to each box b we have k ordered lists $List_i(b)$, $1 \leq i \leq k$, the ith list containing the points in $b \cap V$ ordered on the ith co–ordinate. From an entry for a point p in each of these lists there are pointers to the entries for p in all the remaining $k-1$ lists. For each i, $1 \leq i \leq k$, $List_i(b)$ is embedded in the leaves of a corresponding complete binary search tree $T_i(b)$ of height $\lceil \log_2 m(lsa(b)) \rceil$. The leaves of $T_i(b)$ are the points in $lsa(b) \cap V$ ordered on the ith co–ordinate, however only the leaves that are points in $b \cap V$ are linked together to form the ordered list $List_i(b)$. In the process of spliting $b \cap V$, we split the k ordered lists $List_i(b)$, $i=1,\ldots,k$, and for each box obtain b' in $successors(b)$, we obtain the k ordered lists $List_i(b')$, $i=1,\ldots,k$. For each $b' \in (successors(b)-\{b_{\max}\})$, $List_i(b')$ will be embedded in the leaves of a complete binary search tree $T_i(b')$ of height $\lceil \log_2 m(b') \rceil$.

To split $b \cap V$ we first obtain the points in $(b-b_{\max}) \cap V$ in time proportional to $s(b)=|(b-b_{\max}) \cap V|$. As before let $h_i(b)$ be a hyperplane orthogonal to the ith co–ordinate axis and passing through the centre of b, and let $L(h_i(b))$ and $R(h_i(b))$ be the corresponding left open and right closed half spaces. Among the two boxes $b \cap L(h_i(b))$ and $b \cap R(h_i(b))$, let b_i be the one that contains the smaller number of points in V. We can obtain the set of points $b_i \cap V$ in time proportional to $|b_i \cap V|$ by searching $List_i(b)$ simultaneously from both ends and stopping the first time the hyperplane $h_i(b)$ is crossed. We have two cases depending on $|(\bigcup_{i=1}^{k} b_i) \cap V|$. If $|(\bigcup_{i=1}^{k} b_i) \cap V| < 2^{-k} m(b)$ then $(\bigcup_{i=1}^{k} b_i) \cap b_{\max} = \phi$ and $(b-b_{\max}) = \bigcup_{i=1}^{k} b_i$, and in this case we spend time proportional to $k|(\bigcup_{i=1}^{k} b_i) \cap V| = k\, s(b)$ in obtaining the set of points $(b-b_{\max}) \cap V$. On the other hand if $|(\bigcup_{i=1}^{k} b_i) \cap V| \geq 2^{-k} m(b)$ then we can afford spend $O(m(b))$ time in isolating the points in $(b-b_{\max}) \cap V$.

Once we have the set of points $(b-b_{\max}) \cap V$, we get the corresponding k sorted lists as follows. To obtain a list containing the set of points $(b-b_{\max}) \cap V$ sorted on the ith co–ordinate, we first label all the points in $List_i(b)$ that are located in $b-b_{\max}$. We then label all the vertices in $T_i(b)$ that are proper ancestors of the labelled points in $List_i(b)$. The labelling of the proper ancestors can be performed in time proportional to the number of proper ancestors plus the number of points in $(b-b_{\max}) \cap V$. Then from the Ancestors Lemma we can bound the time required for labelling by $O(s(b)(1+\log_2 m(lsa(b))-\log_2 s(b)))$. Next, we traverse the labelled vetices of $T_i(b)$ in order and thereby obtain the labelled points in $List_i(b)$ i.e. the points in $(b-b_{\max}) \cap V$ in sorted order on the ith co–ordinate. The in order traversal may be performed in time proportional to the number of labelled vertices. Finally, we unlabel all the labelled vertices.

After the k ordered lists for the set $(b-b_{\max}) \cap V$ have been obtained, all the points in $(b-b_{\max}) \cap V$ are deleted from each of lists $List_i(b)$, $i=1,\ldots,k$. For each box b' in $successors(b)-\{b_{\max}\}$, we can obtain the k ordered lists $List_i(b')$ from the k ordered lists for $(b-b_{\max}) \cap V$, in $O(k|(b-b_{\max}) \cap V|) = O(k\, s(b))$ time. In addition, for each b' in $successors(b)-\{b_{\max}\}$, $List_i(b')$ may be embedded in a complete binary search tree $T_i(b')$ of height $\lceil \log_2 m(b') \rceil$ in $O(m(b'))$ time.

Thus the entire process of splitting $b \cap V$ may be accomplished in $O(k\, 2^k\, s(b)(1+\log_2 m(lsa(b))-\log_2 s(b)))$ time.

5. Maintaining the neighbor and attractor sets, and the estimates

In this section we describe how the neighbor and attractor sets, and the estimates, may be maintained $O(n \log n)$ time. We first observe that

(1) A box b' is added to or deleted from $Attractors(b)$ whenever b is added to or deleted from $Neighbors(b')$.

(2) $Estimate(b)$ changes only when a box is added to $Neighbors(b)$.

(3) When b is split, there can be additions only to the neighbor sets of boxes in $successors(b) \cup Attractors(b)$, only boxes in $successors(b)$ can possibly get added to the neighbor set of a box in $Attractors(b)$, and for all $b' \in successors(b)$, $Neighbors(b') \subseteq Neighbors(b) \cup successors(b)$.

The above observations lead to the following procedure for modifying the neighbor and attractor sets, and the estimates, during the stage when b is split.

Procedure Modify–sets–estimates

1. $\forall\, b' \in successors(b)$,
   $Neighbors(b') := Neighbors(b) \cup successors(b) - \{b'\}$,
   $Attractors(b') := Attractors(b) \cup successors(b) - \{b'\}$,
   If $|b' \cap V| \geq 2$ then $Estimate(b') := d_{\max}(b')$
   else
   $Estimate(b') := \min_{b'' \in Neighbors(b')} \{ d_{\max}(b', b'') \}$.
2. $\forall\, b' \in Attractors(b)$,
   $Neighbors(b') := Neighbors(b') \cup successors(b) - \{b\}$,
   $Estimate(b') := \min \{ Estimate(b'), \min_{b'' \in successors(b)} \{ d_{\max}(b', b'') \} \}$.
3. $\forall\, b' \in Neighbors(b)$,
   $Attractors(b') := Attractors(b') \cup successors(b) - \{b\}$.
4. $\forall\, b' \in successors(b) \cup Attractors(b)$,
   delete from $Neighbors(b')$ every box b'' which satisfies $d_{\min}(b', b'') > Estimate(b')$,
   and if b'' is deleted from $Neighbors(b')$
   then also delete b' from $Attractors(b'')$.

end Modify–sets–estimates

As before let B^* be the set of all the boxes generated during the execution of *Algorithm All–Nearest–Neighbors*. We shall bound the total number of boxes that are added to $Neighbors(b)$ and $Attractors(b)$ for all $b \in B^*$ during the entire execution of the All–Nearest–Neighbors algorithm. During the stage when b is split, there are at most

$$2\,|successors(b)|\,(|successors(b)| + |Neighbors(b)| + |Attractors(b)|)$$

additions to the neighbor and attractor sets. So to bound the total number of additions to the neighbor and attractor sets we must bound the size of $Neighbors(b)$ and $Attractors(b)$ when b is split. Such bounds are provided by the Packing Lemmas that follow.

We shall let b_L denote a box in the current set of boxes B such that b_L has the largest volume among all the boxes in B.

Packing Lemma 1. Let r be a positive integer. At the beginning of a stage, if $b \in B$ then the number of boxes b' in B such that $d_{\min}(b, b') \leq r\, d_{\max}(b_L)$ is at most $2^k(2rk+3)^k$.

Proof. Let $size(b')$ denote the length of each of the k intervals defining b'. Let

$$C(b) = \{ b' : b' \in B,\ d_{\min}(b, b') \leq r\, d_{\max}(b_L) \}$$

and

$$A(b) = \{ b' : b' \in B^*,\ successors(b') \cap C(b) \neq \phi \}$$

Since boxes are split in non–increasing order of size, $\forall\, b' \in A(b)$, $size(b') \geq size(b_L)$. So each box b' in $A(b)$ may be shrunk to a box b'' such that $d_{\min}(b, b'') = d_{\min}(b, b') \leq r\, d_{\max}(b_L)$, and $size(b'') = size(b_L)$. Let $\hat{A}(b)$ be the set of boxes obtained by shrinking all the boxes in $A(b)$ in this manner. We note that the L_q distance between two points in k–dimensional space is bounded by k times the L_∞ distance between the points. It then follows that each box $\hat{b}$ in $\hat{A}(b)$ must be contained in a box b^* such that $size(b^*) = (2rk+3)\, size(b_L)$, and b^* and b have the same centre. As the boxes in $\hat{A}(b)$ are disjoint we get $|\hat{A}(b)| \leq (2rk+3)^k$. Finally, as $|successors(b)| \leq 2^k$, we have

$$|C(b)| \leq 2^k |A(b)| \leq 2^k |\hat{A}(b)| \leq 2^k (2rk+3)^k. \blacksquare$$

Packing Lemma 2. At the beginning of each stage, if $b \in B$ and $|b \cap V| \geq 2$ then $|Neighbors(b)| \leq c_1(k)$ where $c_1(k) = 2^k(2k+3)^k$.

Proof. Since $|b \cap V| \geq 2$, we have $Estimate(b) = d_{\max}(b) \leq d_{\max}(b_L)$, and hence $\forall\, b' \in Neighbors(b)$, $d_{\min}(b, b') \leq Estimate(b) \leq d_{\max}(b_L)$. So from Packing Lemma 1 we get $|Neighbors(b)| \leq 2^k(2k+3)^k$. $\blacksquare$

Packing Lemma 3. At the beginning of each stage, if $b \in B$ then $|Attractors(b)| \leq c_2(k)$ where $c_2(k) = 12^k + 2^k(8k+3)^k$.

Proof. Wlog let $\alpha(b)$, the centre of box b, be the origin. Let

$$A_F(b) = \{ b' : b' \in Attractors(b),\ d_{\min}(b, b') \geq 4 d_{\max}(b_L) \}$$

and let

$$A_S(b) = Attractors(b) - A_F(b).$$

From Packing Lemma 1,

$$|A_S(b)| \leq 2^k(8k+3)^k.$$

We first observe that there cannot be a pair of boxes b_1, b_2, in $A_F(b)$ such that the centres $\alpha(b) = 0$, $\alpha(b_1)$, $\alpha(b_2)$ are collinear. Assume that there exist such boxes b_1 and b_2 in $A_F(b)$, and let

$$d(0, \alpha(b_2)) = d(0, \alpha(b_1)) + d(\alpha(b_1), \alpha(b_2))$$

Then as

$$d_{\min}(b, b_2) + d_{\max}(b_L) \geq d(0, \alpha(b_2))$$

and

$$d_{\max}(b_1, b_2) \leq d(\alpha(b_1), \alpha(b_2)) + d_{\max}(b_L)$$

we get

$$d_{\min}(b, b_2) \geq d_{\max}(b_1, b_2) + d(0, \alpha(b_1)) - 2 d_{\max}(b_L)$$

$$> d_{\max}(b_1, b_2)$$

Then because of the invariant forced on the neighbor sets at each stage, we must have that $b \notin Neighbors(b_2)$ and $b_2 \notin Attractors(b)$ which is a contradiction.

There also cannot be a pair of boxes b_1, b_2 in $A_F(b)$ such that $d(\frac{\alpha(b_1)}{d(0,\alpha(b_1))}, \frac{\alpha(b_2)}{d(0,\alpha(b_2))}) \leq \frac{1}{3}$. Assume that there is such a pair of boxes b_1, b_2, in $A_F(b)$. Let $d(0,\alpha(b_1)) \leq d(0,\alpha(b_2))$ and $\lambda = \frac{d(0,\alpha(b_1))}{d(0,\alpha(b_2))}$. We have

$$d(\alpha(b_1), \lambda\alpha(b_2)) \leq \frac{1}{3} d(0,\alpha(b_1)) = \frac{1}{3} d(0, \lambda\alpha(b_2))$$

This gives

$$\begin{aligned} d(0,\alpha(b_2)) &= d(\alpha(b_1), \lambda\alpha(b_2)) + d(\lambda\alpha(b_2), \alpha(b_2)) \\ &\quad + d(0, \lambda\alpha(b_2)) - d(\alpha(b_1), \lambda\alpha(b_2)) \\ &\geq d(\alpha(b_1), \alpha(b_2)) + \frac{2}{3} d(0, \alpha(b_1)) \\ &\geq d(\alpha(b_1), \alpha(b_2)) + \frac{8}{3} d_{\max}(b_L) \end{aligned}$$

This would imply that $d_{\min}(b, b_2) > d_{\max}(b_1, b_2)$ and this cannot happen because of the invariant forced at each stage on the neighbor sets.

Let $\hat{A}_F(b) = \{ p : p = \frac{\alpha(b')}{d(0,\alpha(b'))}, b' \in A_F(b) \}$. Then $|\hat{A}_F(b)| = |A_F(b)|$, each point in $\hat{A}_F(b)$ is at a distance of 1 from the origin, and the distance between any two points in $\hat{A}_F(b)$ is at least 1/3. Around each point in $\hat{A}_F(b)$ draw a ball of radius 1/6. No two of these balls can intersect and the intersection of each such ball with the unit ball around the origin completely contains a ball of radius 1/12. So the number of points in $\hat{A}_F(b)$ and hence the number of boxes in $A_F(b)$ is at most 12^k. Thus

$$|Attractors(b)| = |A_S(b)| + |A_F(b)| \leq 2^k(8k+3)^k + 12^k. \blacksquare$$

From the packing lemmas and the procedure for maintaining the neighbor and attractor sets, it follows that during each stage there are a constant number of additions to neighbor and attractor sets. Since there are at most $2n$ stages it follows that $O(n)$ boxes are added to the neighbor and attractor sets during the entire execution of the All–Nearest–Neighbors algorithm. The size of the attractor set of any box in B never grows beyond a constant because at the beginning of a stage the size of the attractor set is bounded by a constant and during a stage there can be only a constant number additions to an attractor set. So the total time for maintaining attractor sets is $O(n)$. The time to maintain the estimates is also $O(n)$ because the change in $Estimate(b)$ due to the addition of a box to $Neighbors(b)$ can be computed in constant time. We implement $Neighbors(b)$ by a data structure which allows insertions and deletions in $O(\log n)$ time, and allows access to a box b' that has the largest value for the parameter $d_{\min}(b, b')$ in $O(\log n)$ time. A heap or a 2–3 tree suffices [6]. Then the total work to maintain the neighbors sets is $O(n \log n)$.

6. All–m–Nearest–Neighbors

To get an algorithm for the All–m–Nearest–Neighbors problem, the All–Nearest–Neighbors algorithm is modified as follows.

(1) Any box $b' \in Immediate-successors(b)$ such that $|b' \cap V| < m+1$ is immediately split into $|b' \cap V|$ boxes each containing exactly one point in V.

(2) Let $r_1(b) \leq r_2(b) \leq .... \leq r_m(b) \leq ....$ be a non–decreasing sequence of the distances in the multiset $\{ r : r = d_{\max}(b, b'), b' \in Neighbors(b) \}$. Then

$$Estimate(b) = \begin{cases} d_{\max}(b), \text{ if } |b \cap V| \geq m+1 \\ r_m(b), \text{ otherwise} \end{cases}$$

The sizes of the attractor sets of all the boxes in B can be bounded by $c_2(k)\,m$, and the sizes of the neighbor sets of non–degenerate boxes in B can be bounded by $c_1(k)$, where $c_1(k)$, $c_2(k)$ are constants dependent on k. During each stage there are $O(m^2)$ additions to the neighbor and attractor sets, and the number of stages is $O(n/m)$. Then the running time may be shown to be $O(m\, n \log n)$.

7. Conclusion

We have presented an $O(n \log n)$ algorithm for the All–Nearest–Neighbors problem. The running time of the algorithm is optimal upto a constant in the algebraic decision tree model of computation. If the metric in the given problem is positive definite or semidefinite rather than one the standard L_q metrics, the problem can be transformed in linear time to a problem with L_2 (euclidean) metric without increasing the dimension. A slight modification of the All–Nearest–Neighbors algorithm gives an $O(m\, n \log n)$ algorithm for the All–m–Nearest–Neighbors problem.

Acknowledgements

The author would like to thank Prof. Herbert Edelsbrunner for helpful discussions. The author is grateful to Prof. Alan Cline for pointing out that the algorithm works in the same asymptotic time complexity for any positive semi–definite metric.

References

1. M. Ben–Or, Lower bounds for algebraic computation trees, *Proc. 15th Annual ACM Symp. Theory Comput.* 1983, pp.80–86
2. J. L. Bentley, Multidimensional divide–and–conquer, *CACM*, Volume 23, 1980, Number 4, pp. 214–229.
3. J. L. Bentley, B. Weide, and A. C. Yao, Optimal expected–time algorithms for closest–point problems, *ACM Tran. Math. Software*, Volume 6, 1982, Number 4, pp. 563–579.
4. K. Clarkson, Fast Algorithms for the All–Nearest–Neighbors problem, *Proc. 24th Annual Symp. Found. Comp. Sc.*, 1983, pp.226–232.
5. F. P. Preparata, and M. I. Shamos, Computational Geometry: An introduction, 1985, Springer–Verlag New York Inc., New York.
6. E. M. Reingold, J. Nievergelt, and N. Deo, Combinatorial Algorithms: Theory and Practice, 1977, Prentice Hall Inc., Englewood Cliffs, New Jersey.
7. M. I. Shamos, Computational Geometry, Ph.D. Dissertation, Yale University, New Haven, Conn., 1978.