Cell-probe lower bounds for the partial match problem

T.S. Jayram* Subhash Khot† Ravi Kumar* Yuval Rabani‡

Abstract

Given a database of n points in $\{0,1\}^d$, the *partial match problem* is: In response to a query x in $\{0,1,*\}^d$, find a database point y such that for every i whenever $x_i \neq *$, we have $x_i = y_i$. In this paper we show randomized lower bounds in the cell-probe model for this well-studied problem [Riv74, Knu73, Riv76, MNSW98, BOR99, CIP02].

Our lower bounds follow from a two-party asymmetric randomized communication complexity *near-optimal* lower bound for this problem, where we show that either Alice has to send $\Omega(d/\log n)$ bits or Bob has to send $\Omega(n^{1-o(1)})$ bits. When applied to the cell-probe model, it means that if the number of cells is restricted to be $\mathrm{poly}(n,d)$ where each cell is of size $\mathrm{poly}(\log n, d)$, then $\Omega(d/\log^2 n)$ probes are needed. This is an exponential improvement over the previously known lower bounds for this problem [MNSW98, BOR99] and suggests that an efficient algorithm for this problem is essentially hopeless!

Our lower bound also leads to new and improved lower bounds for related problems including a lower bound for the ℓ_∞ c-nearest neighbor problem for $c < 3$ and an improved communication complexity lower bound for the exact nearest neighbor problem.

*IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120. Email: {jayram, ravi}@almaden.ibm.com

†Department of Computer Science, Princeton University, Princeton, NJ 08544. Supported by Prof. Sanjeev Arora's David and Lucile Packard Fellowship, NSF Grant CCR-0098180, and an NSF ITR grant. Part of this work was done while the author was visiting the IBM Almaden Research Center and IBM T. J. Watson Research Center. Email: khot@cs.princeton.edu

‡Computer Science Department, Technion, Israel Institute of Technology, Technion City, Haifa 32000, Israel. Work at the Technion supported by BSF grant number 99-00217, by ISF grant number 386/99, by IST contract number 32007 (APPOL), and by the Fund for the Promotion of Research at the Technion. Part of this work was done while the author was visiting the IBM Almaden Research Center. Email: rabani@cs.technion.ac.il

1 Introduction

Given a database of points in a space and a query, the *nearest neighbor* problem is to find a database point that is "close" to the query. Here, the notion of closeness depends on the underlying space and also on whether a close, but not necessarily the closest data point suffices. The goal is to preprocess the database and create a static data structure that will support efficient search. The complexity measures are the size of the data structure (usually, a table) and the number of probes into the table.

The nearest neighbor problem is a fundamental problem in computational geometry with numerous applications and many efficient algorithms are known for low-dimensional Euclidean space. However, these algorithms become inefficient in terms of query time and/or space requirements as the dimension grows—this is the so-called *curse of dimensionality.* This prompted the quest for very efficient (and possibly randomized) algorithms for the approximate nearest neighbor problem in Euclidean/Hamming spaces and the last few years have seen tremendous progress made in this direction [Kle97, IM98, KOR00]; also supporting this line of study was the evidence for the curse of dimensionality for these spaces provided by the recent lower bounds of [BOR99, CCGL99, BR00]. However, for many nearest neighbor problems, there is still a huge gap between the upper and lower bounds; for many of these problems, the gap is exponential.

Our results. In this paper we consider an instantiation of the nearest neighbor problem—the *partial match* problem. This problem has been investigated for the last few decades (see, for example, [Riv74, Knu73]), but all solutions discovered so far have either essentially $\Omega(n)$ query time or $2^{\Omega(d)}$ space. This leads to the widely believed conjecture that this problem also suffers from the curse of dimensionality (see [BOR99, BR00, CIP02]). We give supporting evidence for this conjecture by proving essentially optimal communication complexity lower bounds for this problem, and thus dramatically improving the lower bounds in the cell-probe model.

Specifically, we show that in the cell-probe model, if the number of cells is restricted to be $\mathrm{poly}(n, d)$ where each cell is of size $\mathrm{poly}(\log n, d)$, then $\Omega(d/\log^2 n)$ probes are needed; this is an exponential improvement over the bounds in [MNSW98, BOR99]. Our lower bound follows from an asymmetric two-sided error randomized communication complexity lower bound for this problem, where we show that either Alice has to send $\Omega(d/\log n)$ bits or Bob has to send $\Omega(n^{1-o(1)})$ bits.

There are several consequences of this basic lower bound and the most notable ones include: (1) a near-optimal communication complexity lower bound and a cell-probe lower bound for the ℓ_∞ c-nearest neighbor problem for $c < 3$; this addresses the presumed hardness stated in [Ind01] and (2) a near-optimal communication complexity lower bound for the Hamming nearest neighbor problem; this strengthens the bound on Bob's communication obtained in [BR00].

Related work. As mentioned earlier, the problem of partial match has been investigated for quite a while. The first non-trivial result for this problem was obtained in [Riv74, Riv76], who showed that for $d \leq 2\log n$, the "exhaustive storage" solution can somewhat be improved. Recently, [CIP02] presented two algorithms with the following tradeoffs: $n \cdot \exp(O(d \log^2 d \sqrt{c/\log n}))$ space and $O(n/2^c)$ query time for any c and nd^c space and $O(dn/c)$ query time for any $c \leq n$. If the number of cells is restricted to be at most $\mathrm{poly}(n, d)$ where each cell is of size $\mathrm{poly}(\log n, d)$, the previously known cell-probe lower bounds were $\Omega(\sqrt{\log d})$ [MNSW98] and $\Omega(\log d)$ [BOR99].

The nearest neighbor problem in d-dimensional Euclidean space or the d-dimensional Hamming cube is also a well-studied problem. Note that the exact version of nearest neighbor in the Hamming cube can be solved trivially by a single probe into a table of 2^d cells, each containing d bits. The best algorithms for the exact nearest neighbor problem in Euclidean space take $\mathrm{poly}(d, \log n)$ query time and need $n^{\Theta(d)}$ space (see, for instance, [Mei93]). For the approximate version of the problem, the best known algorithms are randomized and make $O(\log \log d)$ probes to a table of size $\mathrm{poly}(n, d)$ [IM98, KOR00]. The current best known cell-probe randomized lower bound for the exact nearest neighbor problem (in both the Hamming and the Euclidean cases) is $\Omega(d/\log n)$ for a table of size $\mathrm{poly}(n, d)$ [BOR99, BR00]; in a less general yet reasonable model, [BV02] showed a lower bound of $\Omega(d\sqrt{\log d/\log\log d})$. The current best known lower bound for the approximate nearest neighbor problem in the Hamming cube is $\Omega(\log\log d/\log\log\log d)$ [CCGL99]; however, this lower bound applies only to deterministic algorithms.

For the nearest neighbor problem under the ℓ_∞ norm, [Ind01] gave a 3-approximation algorithm that makes $O(d \log n)$ probes into a table of size $O(n^{1+\log d})$ and an $O(\log \log d)$-approximation algorithm that uses an essentially linear-sized table. This paper also shows that for factors below 3, this problem is as hard as the partial match problem. No lower bound better than $\Omega(\log d)$ was known before for the ℓ_∞ nearest neighbor problem.

Organization. In Section 2, we provide a brief description of the lower bound technique that we use. In Section 3, we will obtain a cell-probe lower bound for what we call the INTERSECT ALL problem. Finally, in Section 4, we use this lower bound to demonstrate near-optimal lower bounds for partial match, exact nearest neighbor, ℓ_∞ nearest neighbor, and other related problems.

2 Background

The cell-probe model. The *cell-probe model* was formulated by Yao [Yao81]. It is acknowledged as a general data structure model for proving lower bounds. In the basic cell-probe model, there are three parameters s, b, and t. Here, s denotes the *number of cells* in the data structure, b denotes the maximum *cell size* in terms of the number of bits in the cells, and t denotes the *number of probes* made into the data structure. An important technique to obtain lower bounds in the cell-probe model is via asymmetric communication complexity [Ajt88, Mil94]. For more details on the cell-probe model, see the survey by Miltersen [Mil99].

Asymmetric communication complexity. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. In the asymmetric communication complexity model, there are two parties Alice and Bob. Alice gets an input $X \in \mathcal{X}$ (to be thought of as the *query*) and Bob gets an input $Y \in \mathcal{Y}$ (to be thought of as the *database*). Their goal is to compute $f(X, Y)$ (to be thought of as finding the nearest neighbor of X). The complexity measure is the total number of bits communicated by each side—an $[a, b]$-*protocol* is one in which Alice sends a bits and Bob sends b bits. A randomized protocol is said to have two-sided error ϵ if it computes the function on every input correctly with probability at least $1 - \epsilon$.

Lower bounds on (randomized) asymmetric communication complexity lead to (randomized) lower bounds in the cell-probe model.

Lemma 1 (Miltersen [Mil94]). *For any function, if there is a (randomized) solution in the cell-probe model with parameters s, b, and t, then there is a (randomized) $[t\lceil \log s\rceil, tb]$-protocol for the corresponding communication problem.*

It therefore suffices to show lower bounds on the (randomized) asymmetric communication complexity. To accomplish this, we use the *richness* technique [MNSW98], which we describe below. This presentation is slightly different from the version present in [MNSW98], but is more convenient for us to work with.

The richness technique. The *function matrix* M_f is a 0–1 matrix whose rows are indexed by inputs $X \in \mathcal{X}$ to Alice and the columns are indexed by inputs $Y \in \mathcal{Y}$ to Bob and is such that $M_f(X,Y) \stackrel{\text{def}}{=} f(X,Y)$. Let $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ be probability distributions on $\mathcal{X}$ and $\mathcal{Y}$ respectively and let $\mu = \mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. For a rectangle R, we say that it has *weight* at least $\alpha \times \beta$ if the total weight of its rows when measured w.r.t. $\mu_{\mathcal{X}}$ is at least α and the total weight of its columns when measured w.r.t. $\mu_{\mathcal{Y}}$ is at least β.

We say that a matrix M has an *$[a,b]$-partition* if the matrix can be partitioned into monochromatic rectangles by a deterministic communication protocol where Alice uses a bits and Bob uses b bits. A column Y of a matrix M is called *α-good* if the weight of 1-entries in this column is at least α measured with respect to $\mu_{\mathcal{X}}$.

Definition 2 ((α,β)-richness). A function matrix M_f is *(α,β)-rich* if $\Pr_{Y \in \mu_{\mathcal{Y}}}[Y \text{is } \alpha\text{-good}] \geq \beta$.

The deterministic version of the richness lemma [MNSW98] states that if a function f has deterministic low communication protocol, then the function matrix has a large 1-monochromatic rectangle.

Lemma 3 (Deterministic richness lemma [MNSW98]). *If a matrix is (α,β)-rich and it has a $[a,b]$-partition into monochromatic rectangles $\{R_1, R_2, R_3, \ldots\}$, then there exists a 1-monochromatic rectangle R_{i_0} of weight at least $\frac{\alpha}{2^a} \times \frac{\beta}{2^{a+b}}$.*

The randomized version of richness lemma states that if a function f has a randomized two-sided error protocol with low communication complexity, then the function matrix has a large rectangle which is almost 1-monochromatic.

Lemma 4 (Randomized richness lemma [MNSW98]). *If a function f has a randomized δ-error $[a,b]$-protocol, the matrix M_f has at least a fraction c of 1's (w.r.t. the measure μ) and $c \geq 4\sqrt{\delta}$, then there exists a rectangle R_{i_0} of weight at least $\frac{c}{2^{a+2}} \times \frac{c}{2^{a+b+2}}$ such that at most $\sqrt{\delta}$-fraction of entries in R_{i_0} are 0.*

Note that Lemma 4 holds for two-sided error protocols. For one-sided error, meaning the protocol never errs on an input with $f(x,y) = 0$, the proof (see Appendix) shows that the rectangle R_{i_0} is in fact a 1-monochromatic rectangle for f.

3 A cell-probe lower bound

In this section we show randomized cell-probe lower bounds for the following problem. Our bounds follow from near-optimal communication complexity lower bounds.

Problem (Intersect all). Let the database $\mathcal{D}$ be a family of n sets $Y_1, Y_2, \ldots, Y_n$ where every $Y_i \subseteq [d]$. Given a query $X \subseteq [d]$, does X intersect Y_i for every $1 \leq i \leq n$?

An overview of the proof. In the cell-probe model, the aim is to build a concise data structure for Bob that will allow queries from Alice to be answered efficiently. By Lemma 1, a lower bound in the cell-probe model would follow from a lower bound in the asymmetric communication complexity setting. We prove the latter by a (rather straightforward) application of Lemma 4.

The function matrix for the Intersect all problem has rows indexed by Alice's inputs which are all the sets $X \subseteq [d]$ and columns indexed by Bob's inputs which are all tuples $\langle Y_1, Y_2, \ldots, Y_n \rangle$ where $Y_i \subseteq [d], \forall i$. A crucial step in our proof is to restrict Alice's input to only some specific sets, namely, sets $X_i \in \mathcal{X}$ where $\mathcal{X}$ is a *design*. All sets $X \in \mathcal{X}$ have the same size $d' = \Omega(d)$ and intersection of any two distinct sets $X, X' \in \mathcal{X}$ has size at most $\theta d'$ where $\theta > 0$ is a small constant. It is well-known that such designs with size $2^{\Omega(d)}$ exist and they have found several applications in complexity theory, most notably the Nisan-Wigderson's construction of a pseudorandom generator from hard functions [NW94].

Any two members X and X' of the design are almost disjoint. Therefore, when we pick a set Y (of some specific size) at random, the events $Y \cap X = \emptyset$ and $Y \cap X' = \emptyset$ are almost independent. This allows us to apply an inclusion-exclusion inequality to bound the probability that a randomly picked set Y intersects all the sets $X_1, X_2, \ldots, X_t$ for some $t \approx n$. Using this bound, we can immediately bound size of any 1-monochromatic rectangle in the matrix M_f.

Now we get to the details of the proof. We begin by defining a design.

Definition 5 (Design). A collection of sets $\{X_1, \ldots, X_m\}$ where $X_i \subseteq [d]$ is called a (d, m, α, β)-*design* if

1. $|X_i| = \alpha d \ \ \forall\, i$,
2. $|X_i \cap X_j| \leq \beta d \ \ \forall\, i \neq j$.

The following lemma shows the existence of designs (see, for instance, [NW94]). The Appendix contains a proof of this lemma.

Lemma 6. *For every $\theta > 0$, there exist $c_1, c_2 > 0$ such that for all sufficiently large integers d, there exists a $(d, 2^{c_2 d}, c_1, \theta c_1)$-design. In fact, one can take $c_1 = \theta/2$ and $c_2 = \theta^2/96$.*

Let $\mathcal{X}$ be the design guaranteed by the lemma, let

$$\mathcal{Y} \stackrel{\text{def}}{=} \{\langle Y_1, Y_2, \ldots, Y_n\rangle \mid Y_i \subseteq [d]\},$$

and let $d' \stackrel{\text{def}}{=} c_1 d$. $\mathcal{X}$ will be the set of inputs to Alice and $\mathcal{Y}$ will be the set of inputs for Bob. We will fix probability distributions $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ on the inputs to Alice and Bob respectively. $\mu_{\mathcal{X}}$ is the uniform distribution on $\mathcal{X}$. $\mu_{\mathcal{Y}}$ is obtained by picking a tuple $\langle Y_1, Y_2, \ldots, Y_n\rangle$ where the sets $Y_i \subseteq [d]$ are picked independently and each set Y_i is picked by taking every $i \in [d]$ with probability q, where $q \stackrel{\text{def}}{=} (\ln n)/d'$. Note that we have assumed $d = \omega(\ln n)$.

The function f is defined to be

$$f(X, \langle Y_1, \ldots, Y_n\rangle) \stackrel{\text{def}}{=} \bigwedge_{i=1}^{n} \textsc{disj}_d(X, Y_i),$$

where $\textsc{disj}_d(x, y)$ is the standard d-bit *disjointness* function $\textsc{disj}_d(x, y) \stackrel{\text{def}}{=} \vee_{i=1}^{d}(x_i \wedge y_i)$. If the bit strings X and Y_i denote the characteristic vectors, it follows that $f(X, \langle Y_1, \ldots, Y_n\rangle) = 1$ if and only if for all i, $X \cap Y_i \neq \emptyset$, i.e., f corresponds to the Intersect all problem.

The function matrix M_f has rows indexed by $\mathcal{X}$ and columns indexed by $\mathcal{Y}$. The entry of the matrix $M_f(X, \langle Y_1, \ldots, Y_n\rangle) = 1$ if and only if $X \cap Y_i \neq \emptyset \ \ \forall i$.

Lemma 7. *Each row of M_f has at least $1/(2e)$-fraction of 1's, where the fraction is measured w.r.t. the distribution $\mu_{\mathcal{Y}}$.*

Proof. Let X be a fixed set in $\mathcal{X}$ and $\langle Y_1, \dots, Y_n \rangle$ be a column picked at random according to $\mu_{\mathcal{Y}}$.

$$\Pr_{Y_i}[Y_i \cap X = \emptyset] = (1-q)^{|X|} = \left(1 - \frac{\ln n}{d'}\right)^{d'} \leq \exp(-\ln n) = \frac{1}{n}.$$

The fraction of 1's in the row corresponding to X is equal to

$$\Pr_{Y_1,\dots,Y_n}[Y_i \cap X \neq \emptyset \quad \forall i] \geq \left(1 - \frac{1}{n}\right)^n \geq \frac{1}{2e}.$$

□

Theorem 8. *M_f has no 1-monochromatic rectangle which has n rows and whose columns have total weight $2\exp(-\frac{1}{2}n^{1-3\theta})$.*

Proof. Fix any n rows $X_1, X_2, \dots, X_n$. We will show that when a column $\langle Y_1, Y_2, \dots, Y_n \rangle$ is picked at random with distribution $\mu_{\mathcal{Y}}$, then the probability that the column has 1's in all n rows is at most $\exp(-\frac{1}{2}n^{1-3\theta})$. Thus we need to upper bound the probability of the event

$$Y_j \cap X_i \neq \emptyset, \ \forall\ 1 \leq i, j \leq n.$$

Let us first compute the probability that a random set Y (where every element $i \in [d]$ is picked with probability q) intersects every X_i. Let E_i be the event that "$Y \cap X_i = \emptyset$". Clearly,

$$\Pr_Y[E_i] = (1-q)^{|X_i|} = \left(1 - \frac{\ln n}{d'}\right)^{d'} \geq \frac{1}{n^{1+\theta}},$$

and since for $i \neq j$, we have $|X_i \cap X_j| \leq \theta d'$, we have

$$\Pr_Y[E_i \wedge E_j] = (1-q)^{|X_i \cup X_j|} \leq \left(1 - \frac{\ln n}{d'}\right)^{(2-\theta)d'} \leq \frac{1}{n^{2-\theta}}.$$

Now we apply inclusion-exclusion bound. Let $t = n^{1-2\theta}$.

$$\begin{aligned}
\Pr_Y[\wedge_{i=1}^n (Y \cap X_i \neq \emptyset)] &\leq \Pr_Y[\wedge_{i=1}^t (Y \cap X_i \neq \emptyset)] \\
&= 1 - \Pr_Y[\vee_{i=1}^t (Y \cap X_i = \emptyset)] \\
&= 1 - \Pr_Y[\vee_{i=1}^t E_i] \\
&\leq 1 - \sum_{i=1}^t \Pr_Y[E_i] + \sum_{1 \leq i < j \leq t} \Pr_Y[E_i \wedge E_j] \\
&\leq 1 - t \cdot \frac{1}{n^{1+\theta}} + \frac{t^2}{2}\frac{1}{n^{2-\theta}} \\
&= 1 - \frac{1}{n^{3\theta}} + \frac{1}{2}\frac{1}{n^{3\theta}} \\
&= 1 - \frac{1}{2}\frac{1}{n^{3\theta}}.
\end{aligned}$$

Since $Y_j, 1 \leq j \leq n$ are picked independently, we have

$$\begin{aligned}
\Pr_{Y_1,\dots,Y_n}[\wedge_{j=1}^n \wedge_{i=1}^n Y_j \cap X_i \neq \emptyset] &\leq \left(1 - \frac{1}{2}\frac{1}{n^{3\theta}}\right)^n \\
&\leq \exp\left(-\frac{1}{2}n^{1-3\theta}\right).
\end{aligned}$$

□

Theorem 9. *Every rectangle in M_f which has at least n rows and whose columns have total weight at least $2\exp(-\frac{1}{2}n^{1-3\theta})$ contains at least $1/(2n)$-fraction of 0's.*

Proof. This follows by carefully looking at the proof of Theorem 8. What we have shown is, columns with weight at most $\exp(-\frac{1}{2}n^{1-3\theta})$ can have 1 in all the n rows. Therefore, for a rectangle with n rows and columns with total weight $2\exp(-\frac{1}{2}n^{1-3\theta})$, at least half the columns contain a zero in some row. This proves the claim. $\square$

Theorem 10. *For any $\epsilon > 0$ and for any two-sided error $[a, b]$-protocol for the* Intersect all *problem with error $1/3$, either $a = \Omega(10^{-5}\epsilon^2 d/\ln n)$ or $b = \Omega(n^{1-\epsilon})$.*

Proof. This follows easily from Theorem 9 and Lemma 4. Assume on the contrary that there exists a protocol with error 1/3 where Alice sends at most $10^{-5}\epsilon^2 d/\ln n$ bits and Bob sends at most $n^{1-\epsilon}$ bits. We first repeat the protocol $4\ln n$ times to get the error down to $1/n^4$. In this new protocol, Alice sends at most $10^{-5}4\epsilon^2 d$ bits and Bob sends at most $4n^{1-\epsilon}\ln n$ bits. We also note that by Lemma 7 the matrix M_f contains at least $1/(2e) \geq 1/8$ fraction of 1's. By Lemma 4, there exists a rectangle R with weight at least

$$\frac{1}{8 \cdot 2^{4\epsilon^2 d/10^5+2}} \times \frac{1}{8 \cdot 2^{4\epsilon^2 d/10^5+4n^{1-\epsilon}\ln n+2}}$$

that contains at most $1/n^2$ fraction of 0's. Let $\theta = \epsilon/4$ so that the matrix M_f has $2^{c_2 d} = 2^{\theta^2 d/96}$ rows. Therefore the number of rows of the rectangle R is

$$2^{c_2 d}\frac{1}{8 \cdot 2^{4\epsilon^2 d/10^5+2}} \geq 2^{\theta^2 d/96}\frac{1}{2^{\theta^2 d/200}} \geq 2^{\theta^2 d/200} \geq n.$$

The total weight of the columns of this rectangle is

$$\frac{1}{8 \cdot 2^{4\epsilon^2 d/10^5+4n^{1-\epsilon}\ln n+2}} \geq 2\exp\left(-\frac{1}{2}n^{1-3\theta}\right).$$

This contradicts Theorem 9 and we are done. $\square$

Theorem 11. *For any $\epsilon > 0$ and for any one-sided error $[a, b]$-protocol for the* Intersect all *problem with error $1/3$, either $a = \Omega(10^{-5}\epsilon^2 d)$ or $b = \Omega(n^{1-\epsilon})$.*

Proof. This follows in a similar manner as the proof of Theorem 10. However, for protocols with one-sided error, Lemma 4 guarantees that the rectangle R is monochromatic. Thus we don't need to reduce the error probability and we save the $\ln n$ factor in communication. $\square$

Corollary 12. *For the* Intersect all *problem, any randomized algorithm in the cell-probe model that uses $\mathrm{poly}(n, d)$ cells with size $\mathrm{poly}(d, \log n)$ each, must make $\Omega(d/\log^2 n)$ probes. If the algorithm makes only one-sided error, it must make $\Omega(d/\log n)$ probes.*

Proof. Suppose there is a cell-probe algorithm with parameters s, b, t where $s = \mathrm{poly}(n, d)$ and $b = \mathrm{poly}(d, \log n)$. By Lemma 1, there is a communication protocol where Alice sends $t \log s$ bits and Bob sends tb bits and by Theorem 10, either $t \log s \geq \Omega(\epsilon^2 d/\log n)$ or $tb \geq n^{1-\epsilon}$. Therefore we must have the number of probes $t \geq \Omega(d/\log^2 n)$ $\square$

4 Consequences

We discuss the implications of the above lower bound for partial match, exact nearest neighbor, and other related problems.

Problem (Partial match). Given a database of n points in $\{0,1\}^d$, the *partial match problem* is: given a query x in $\{0,1,*\}^d$, is there a database point y such that for every i whenever $x_i \neq *$, we have $x_i = y_i$?

The partial match problem is equivalent to the Intersect all problem in the following way. Given an instance of Intersect all with database Y and query X, we obtain X' from X by replacing each 0 in X by a $*$ and each 1 by a 0. Now, it is easy to see that the partial match instance with respect to query X' and database Y is true if and only if the original instance of Intersect all is false. Therefore, using Lemma 1, Theorem 10, and the above reduction, we have

Theorem 13. *For any $\epsilon > 0$, any two-sided error randomized algorithm in the cell-probe model for the partial match problem that makes t probes, either uses $2^{\Omega(d/(t \log^2 n))}$ cells or uses cells of size $\Omega(n^{1-\epsilon}/t)$.*

In particular, if the number of cells is restricted to be $\text{poly}(n, d)$ where each cell is of size $\text{poly}(\log n, d)$, then the algorithm must make $\Omega(d/\log^2 n)$ probes.

Problem (Subset query). Let U be a universe and let $\mathcal{P}$ be a family of subsets of U. Given a query set $Q \subseteq U$, is there a $P \in \mathcal{P}$ such that $Q \subseteq P$?

The partial match problem is intimately related to the *subset query* problem. By virtue of the equivalence shown between these problems in [CIP02], we obtain a cell-probe lower bound similar to Theorem 13 for the subset query problem as well.

Interpreting the partial match problem in a geometric manner as stated in [BOR99], we obtain an improved cell-probe lower bound for the affine subspace problem: whether or not a query affine subspace contains at least one database point.

Problem (Nearest neighbor). Let $\text{ham}(\cdot,\cdot)$ denote the Hamming distance in $\{0,1\}^d$ and let $\mathcal{P}$ be a set of points in $\{0,1\}^d$. Given a query point $Q \in \{0,1\}^d$, find a point $P \in \mathcal{P}$ such that $\text{ham}(P, Q) = \min_{P' \in \mathcal{P}} \text{ham}(P', Q)$.

Using Theorem 10 and a reduction stated in [BOR99], we conclude that in the asymmetric communication problem for a decision version of the nearest neighbor problem, either Alice sends $\Omega(\epsilon^2 d/\ln n)$ bits or Bob sends $\Omega(n^{1-\epsilon})$ bits, for any $\epsilon > 0$. The best previous bound for this problem was that either Alice sends $\Omega(\epsilon d)$ bits or Bob sends $\Omega(n^\delta)$ bits, where $\delta < 1/8$ and ϵ depends on δ [BR00]. Using other reductions stated in [BOR99], our improved communication complexity lower bounds hold for decision versions of nearest neighbor in ℓ_p^d (d-dimensional real space under the ℓ_p norm), for all finite $p \geq 1$. Via easy reductions, the communication complexity lower bound of other geometric data structures problems is also improved. These include point location in an arrangement of hyperplanes, and a multi-dimensional generalization of the dictionary problem. For more details, see [BOR99].

Problem (ℓ_∞ c-nearest neighbor). Let ℓ_∞^d denote the d-dimensional real space under the ℓ_∞ norm and let $\mathcal{P}$ be a set of points in of ℓ_∞^d. Given a query point $Q \in \ell_\infty^d$, is there a $P \in \mathcal{P}$ such that $|P - Q|_\infty \leq c \cdot \min_{P' \in \mathcal{P}} |P' - Q|$?

For $c < 3$ this problem is as hard as the partial match problem [Ind01]. Therefore, a near-optimal communication complexity lower bound and a cell-probe lower bound follows for the c-nearest neighbor in ℓ_∞ norm, for $c < 3$.

Via an easy reduction from this problem, we obtain similar bounds for the following *range search* problem as well: The database is a set of points in d-dimensional real space and given a query which is a rectilinear range, does the query contain a database point. These bounds apply as well to the dual problem where the database is a set of ranges and the query is a point.

References

[Ajt88] M. Ajtai. A lower bound for finding predecessors in Yao's cell probe model. *Combinatorica*, 8:235–247, 1988.

[BOR99] A. Borodin, R. Ostrovsky, and Y. Rabani. Lower bounds for high dimensional nearest neighbor search and related problems. In *Proc. 31st Annual ACM Symposium on the Theory of Computing*, pages 312–321, 1999.

[BR00] O. Barkol and Y. Rabani. Tight bounds for nearest neighbor search and related problems in the cell probe model. In *Proc. 32nd Annual ACM Symposium on the Theory of Computing*, pages 388–396, 2000.

[BV02] P. Beame and E. Vee. Time-space tradeoffs, multiparty communication complexity, and nearest-neighbor problems. In *Proc. 34th Annual ACM Symposium on the Theory of Computing*, pages 688–697, 2002.

[CCGL99] A. Chakrabarti, B. Chazelle, B. Gum, and A. Lvov. A lower bound on the complexity of approximate nearest-neighbor searching on the Hamming cube. In *Proc. 31st Annual ACM Symposium on the Theory of Computing*, pages 305–311, 1999.

[Cha00] B. Chazelle. *The Discrepancy Method: Randomness and Complexity.* Cambridge University Press, 2000.

[CIP02] M. Charikar, P. Indyk, and R. Panigrahy. New algorithms for subset query, partial match, orthogonal range searching, and related problems. In *Proc. 29th International Colloquium on Algorithms, Logic, and Programming*, pages 451–462, 2002.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annual ACM Symposium on the Theory of Computing*, pages 604–613, 1998.

[Ind01] P. Indyk. On approximate nearest neighbors under ℓ_∞ norm. *J. of Computer and System Sciences*, 63:627–638, 2001.

[Kle97] J. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proc. 29th Annual ACM Symposium on the Theory of Computing*, pages 599–608, 1997.

[Knu73] D. Knuth. *The Art of Computer Programming: Sorting and Searching.* Addison-Wesley, 1973.

[KOR00] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. on Computing*, 30(2):451–474, 2000.

[Mei93] S. Meiser. Point location in arrangement of hyperplanes. *Information and Computation*, 106(2):286–303, 1993.

[Mil94] P. B. Miltersen. Lower bounds for union-split-find related problems on random access machines. In *Proc. 26th Annual ACM Symposium on the Theory of Computing*, pages 625–634, 1994.

[Mil99] P. B. Miltersen. Cell probe complexity - a survey. In *Pre-Conference Workshop on Advances in Data Structures at the 19th Conference on Foundations of Software Technology and Theoretical Computer Science*, 1999.

[MNSW98] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson. On data structures and asymmetric communication complexity. *J. of Computer and System Sciences*, 57(1):37–49, 1998.

[NW94] N. Nisan and A. Wigderson. Hardness vs. randomness. *J. of Computer and System Sciences*, 49(2), 1994.

[Riv74] R. Rivest. *Analysis of Associative Retrieval Algorithms.* PhD thesis, Stanford University, 1974.

[Riv76] R. Rivest. Partial match retrieval algorithms. *SIAM J. on Computing*, 5:19–50, 1976.

[Yao81] A. C. Yao. Should tables be sorted. *J. of the ACM*, 28(3):615–628, 1981.

A Proofs

Proof of Lemma 3.

Proof. We use induction on $a+b$. If $a+b=0$ then $a=0, b=0$, the matrix consists solely of 1's and there is only one rectangle which is the whole matrix. So the claim is clearly true in this case.

For the induction step, first consider the case when Bob sends the first bit. This splits the matrix M vertically into two parts, say M_1 and M_2. At least one of the two parts, say M_1, is $(\alpha, \beta/2)$-rich and it has $[a, b-1]$-partition. By induction hypothesis, one of the rectangles R_{i_0} in M_1 is 1-monochromatic and has weight at least $\frac{\alpha}{2^a} \times \frac{\beta/2}{2^{a+b-1}} = \frac{\alpha}{2^a} \times \frac{\beta}{2^{a+b}}$.

Now consider the case when Alice sends the first bit. This splits the matrix M horizontally into two parts, say M_1 and M_2. At least one of the two parts, say M_1 is $(\alpha/2, \beta/2)$-rich and has a $[a-1, b]$-partition. Using the induction hypothesis, there is a 1-monochromatic rectangle R_{i_0} with weight at least $\frac{\alpha/2}{2^{a-1}} \times \frac{\beta/2}{2^{a-1+b}} = \frac{\alpha}{2^a} \times \frac{\beta}{2^{a+b}}$. □

Lemma 14. *If the total weight of* 1*-entries in a matrix* M *is* $\geq c$ *(w.r.t. the product measure* μ*) and the matrix has* $[a, b]$*-partition into monochromatic rectangles* $\{R_1, R_2, R_3, \dots\}$*, then there exists a* 1*-monochromatic rectangle* R_{i_0} *of weight at least* $\frac{c}{2^{a+1}} \times \frac{c}{2^{a+b+1}}$.

Proof. The hypothesis of the lemma implies that the matrix is $(c/2, c/2)$-rich. Apply Lemma 3. □

Proof of Lemma 4.

Proof. Fix the random coins of the protocol such that the deterministic function f' computed by the protocol satisfies

$$\mathrm{Pr}_{(x,y)\in\mu}[f(x,y) \neq f'(x,y)] \leq \delta.$$

The matrix $M(f')$ has at least $c-\delta$ fraction of 1's and it has $[a,b]$-partition into monochromatic rectangles $\{R_1, R_2, \dots\}$. Call a rectangle R_i *good* if

$$\mathrm{Pr}[f(x,y) \neq f'(x,y) | (x,y) \in R] \leq \sqrt{\delta}.$$

By an averaging argument,

$$\sum_{R_i \mid R_i \text{ is not good}} \mu(R_i) \leq \sqrt{\delta}.$$

Now consider the matrix M^* obtained from $M(f')$ where we convert every bad rectangle into a 0-monochromatic rectangle. The matrix M^* has at least $c - \delta - \sqrt{\delta} \geq c/2$ fraction of 1's and M^* also has the same rectangle partition. Applying Lemma 14, there is a 1-monochromatic rectangle R_{i_0} with weight at least $\frac{c}{2^{a+2}} \times \frac{c}{2^{a+b+2}}$. Since R_{i_0} is good, we have

$$\mathrm{Pr}[f(x,y) = 0 \mid (x,y) \in R_{i_0}] \leq \sqrt{\delta}.$$

□

Proof of Lemma 6.

Proof. We will prove Lemma 6 with $c_1 = \theta/2$ and $c_2 = \theta^2/96$. Thus we will construct a design of size $2^{\theta^2 d/96}$ where every set has size $\theta d/2$ and every pairwise intersection has size at most $\theta^2 d/2$. We will select sets $X_1, X_2, ...$ one by one in a greedy fashion. Having selected $X_1, X_2, \dots, X_r$, if there exists a set $X, |X| = \theta d/2$ such that it intersects every $X_i, 1 \leq i \leq r$ in at most $\theta^2 d/2$ points, we let $X_{r+1} = X$ and continue this process. Using a probabilistic argument, we show that as long as $r \leq 2^{\theta^2 d/96}$, there always exists such a set X.

Consider the process of picking a set X where each element of $[d]$ is selected with probability $3\theta/4$. We will use the following Chernoff bound (for example, see [Cha00, Lemma A.3]).

Lemma 15. *Suppose $Z_1, Z_2, \dots, Z_t$ are independent 0-1 random variables and $\mathrm{Pr}[Z_i = 1] = p$ for every i. Let $Z = \sum_{i=1}^{t} Z_i$. Then for $0 < a < 1/2$,*

$$\mathrm{Pr}[Z > (1+a)pt] \leq \exp(-pta^2/4)$$

$$\mathrm{Pr}[Z < (1-a)pt] \leq \exp(-pta^2/2)$$

Using this bound with $p = 3\theta/4$, we have

$$\mathrm{Pr}[|X| < \theta d/2] < \exp(-\theta d/24) \leq 1/2,$$

and

$$\mathrm{Pr}[|X \cap X_i| > \theta^2 d/2] < \exp(-\theta^2 d/96).$$

Taking a union bound, it follows that as along as $r < 2^{\theta^2 d/96}$, there exists a set X of size at least $\theta d/2$ that intersects every X_i in at most $\theta^2 d/2$ points. Taking any subset of X with size $\theta d/2$ completes the proof. □