Tighter Lower Bounds for Nearest Neighbor Search and Related Problems in the Cell Probe Model

Omer Barkol* Yuval Rabani†

Abstract

We prove new lower bounds for nearest neighbor search in the Hamming cube. Our lower bounds are for randomized, two-sided error, algorithms in Yao's cell probe model. Our bounds are in the form of a tradeoff among the number of cells, the size of a cell, and the search time. For example, suppose we are searching among n points in the d dimensional cube, we use $\mathrm{poly}(n, d)$ cells, each containing $\mathrm{poly}(d, \log n)$ bits. We get a lower bound of $\Omega(d/\log n)$ on the search time, a significant improvement over the recent bound of $\Omega(\log d)$ of Borodin et al. This should be contrasted with the upper bound of $O(\log\log d)$ for approximate search (and $O(1)$ for a decision version of the problem; our lower bounds hold in that case). By previous results, the bounds for the cube imply similar bounds for nearest neighbor search in high dimensional Euclidean space, and for other geometric problems.

*Computer Science Department, Technion — IIT, Haifa 32000, Israel. Work partially supported by the Milton and Lillian Edwards Fellowship. Email: omerb@cs.technion.ac.il.

†Computer Science Department, Technion — IIT, Haifa 32000, Israel. This work was supported by grant number 386/99-1 of the Israel Science Foundation founded by the Israeli Academy of Sciences and Humanities, by the N. Haar and R. Zinn Research Fund, and by the Fund for the Promotion of Research at the Technion. Email: rabani@cs.technion.ac.il.

1 Introduction

Problem definition and motivation. This paper is concerned with *nearest neighbor search* (NNS), a fundamental problem in computational geometry, with applications to a variety of areas [15, 20, 17, 33, 16, 22, 14, 32, 19, 23, 34, 8, 21]. The problem is defined as follows: In some vector space endowed with a distance function (typically a d-dimensional Euclidean space), we are given a set of n points (called the database). Given any other point (called a query), we must find the closest point to it in the database. We have to pre-process the database efficiently and create a data structure that will support efficient search. More specifically, the trivial data structure storing the unprocessed list of points allows us to search spending $O(nd)$ arithmetic operations. A challenging goal is to design a similar sized data structure reducing the search time to $\mathrm{poly}(d, \log n)$ (or, in fact, to anything polynomial in d and sub-linear in n).

The problem (in Euclidean space) is a special case of point location in an arrangement of hyperplanes. As such, it has been studied extensively, especially in low dimension, where good solutions are known (see, for example [9]). However, the combinatorial complexity of arrangements grows exponentially with the dimension, rendering the problem seemingly intractable. Indeed, following a long list of contributions [18, 12, 36, 28, 1, 29], currently the best algorithms can find a nearest neighbor in time $\mathrm{poly}(d, \log n)$, but they need exponential ($n^{\Theta(d)}$) storage. On the other hand, there is little evidence in the form of concrete lower bounds to support the *curse of dimensionality* conjecture [13]; i.e., the belief that in high dimension the problem is indeed intractable (see below for more details).

Our results. We present here significant improvements over recently discovered lower bounds for nearest neighbor search [10]. Specifically, our main concern is nearest neighbor search in the d-dimensional Hamming cube. As previously observed [10], lower bounds for the cube imply lower bounds for geometric settings, such as ℓ_p^d ($\mathbb{R}^d$ with distances measured by the L^p norm) for all $1 \leq p < \infty$, as well as for related geometric problems. We prove lower bounds in Yao's cell probe model [35]. In this model, the database is pre-processed into s cells, each containing b bits. A search algorithm sequentially (and possibly adaptively) reads the contents of at most t cells to get the correct answer. In [10] it is proven that a randomized two-sided error cell probe algorithm that is restricted to use $\mathrm{poly}(n, d)$ cells of size $\mathrm{poly}(d, \log n)$ each, must probe at least $\Omega(\log d)$ cells. Here, we improve this bound to $\Omega(d/\log n)$. In fact, as in [10], we actually show tradeoffs among the three parameters s, b, and t, as follows.

Theorem 1. Assuming $d \in \omega(\log n) \cap n^{o(1)}$,[1] for any cell probe algorithm for NNS in the d-dimensional Hamming cube that uses s cells of size b each, and probes at most t cells, the following holds: either $s = 2^{\Omega(d/t)}$; or, $b = n^{\Omega(1)}/t$.

We note that similar bounds were shown in [10] for *deterministic* algorithms.

Our results are best contrasted with the bounds for approximate nearest neighbor search in the cube. In this version of the problem, the search algorithm is required to find a database point whose distance to the query is within a factor of $1+\varepsilon$ of the distance to a nearest neighbor, where $\varepsilon > 0$ is a predefined value. The best available (randomized) algorithm, following a long line of work [5, 13, 6, 26, 25, 27], uses (for an arbitrary constant ε, when stated in terms of the cell probe model) $\mathrm{poly}(n, d)$ cells of size $O(d)$ each, and searches probing $O(\log \log d)$ cells. This randomized upper bound nearly matches a recent deterministic lower bound of $\Omega(\log \log d / \log \log \log d)$ [11], which holds even for very poor approximation. It is worth noting that better upper bounds hold for approximate λ-neighbor (λN), a decision version of nearest neighbor search. In this problem, the search algorithm should answer "yes" iff there is a database point within distance at most λ from the query. In the approximate version, if the nearest neighbor is at distance in $(\lambda, (1+\varepsilon)\lambda)$, then the algorithm may answer either "yes" or "no," and otherwise it should behave as in the exact version. The above-mentioned upper bounds for approximate nearest neighbor search work by reduction to algorithms for approximate λ-neighbor that probe $O(1)$ cells. Our lower bounds (like those in [10]) are proven for λ-neighbor. Thus, we exhibit a very sharp contrast of $\Omega(d/\log n)$ versus $O(1)$ between the search time complexity of randomized, two-sided error, algorithms for exact λ-neighbor versus approximate λ-neighbor, respectively. We further note that by probabilistic arguments one can show the existence of a data structure with $\mathrm{poly}(n, d)$ cells, each with $\mathrm{poly}(d, \log n)$ bits, that allows us to find an approximate nearest neighbor deterministically in $\mathrm{poly}(d, \log n)$ probes [27, 24].

Unlike the randomized lower bound in [10], our results do not hold for exact partial match. In this problem queries may contain "don't care" bits (marked by $*$) that match both a zero and a one. A search should find a database point that precisely matches the query. There is an easy reduction from exact partial match to NNS (but not necessarily the other way around). We do, however, extend our results to partial match λ-neighbor. In this problem the search should find whether or not there is a database point within distance λ of a partial match query. Obviously, a lower bound for λN implies a lower bound for this problem, because queries may be void of don't cares. We show that our lower bounds hold even if the number of exposed bits k (i.e., bits $\neq *$) is fixed to any value in $\Omega(d)$. We get

[1] Notice that if d is outside this range, the problem in the cube becomes trivial.

Theorem 2. For every constant ρ, $0 < \rho \leq 1$, assuming $d \in \omega(\log n) \cap n^{o(1)}$ and $k = \rho d$, there exists λ such that for any cell probe algorithm for partial match λ-neighbor that can handle queries with exactly $d - k$ don't cares, which uses s cells of size b each, and probes at most t cells, the following holds: either $s = 2^{\Omega(d/t)}$; or, $b = n^{\Omega(1)}/t$.

The hidden constants in the lower bound decrease as k decreases. Indeed, notice that the problem with $k = \rho_1 d$ can be reduced to the problem with $k = \rho_2 d$, for $\rho_1 \leq \rho_2$ (see section 4.2).

Our methods. We prove our cell probe lower bounds via lower bounds in the asymmetric communication complexity model. In this model the input is split between two communicating parties, Alice and Bob. Alice gets the query, and Bob gets the database. Their goal is to compute a function of the entire input, the result of λN in our case. To do that, they may exchange bits. The complexity measure is the total number of bits communicated by each side. A protocol where Alice sends a bits and Bob sends b bits is called an $[a, b]$-protocol. In a randomized protocol, Alice and Bob have access to a source of random bits, which may affect the protocol.

The connection between the communication complexity model and the cell probe model is given by the following lemma due to Miltersen [30]:

Lemma 3 (Miltersen [30]). For any boolean 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 communication problem.

Thus, in order to prove lower bounds in the cell probe model, we exhibit lower bounds for the communication complexity of λN. For that, we appeal to the *richness technique* of [31]. It calls for showing that while a large fraction of the possible inputs produce a one value, every large submatrix of the communication matrix contains many zero values. As previously observed [10], the communication matrix for λN contains many large one-monochromatic sub-matrices, regardless of the value of λ. However, we show that for a judicious choice of λ, this is not the case for the complement function. The main idea underlying the proof is that if we take two query points that are about $d/2$ apart (in Hamming distance), then for a random database point the two distributions of the distances to the query points behave somewhat independently. (The precise bound, as well as the details of the richness technique, appear below in Section 2.) We note that our cell probe time lower bounds are asymptotically the best possible to derive using communication complexity. (Yet our communication complexity lower bounds could still be improved on the database side.)

As observed in [31], proving stronger lower bounds in the cell probe model would imply non-linear lower bounds for Boolean branching programs. The converse is not necessarily true; and, to the best of our knowledge, recent breakthroughs in branching programs lower bounds [7, 2, 3] do not seem to apply directly to our problem.

Additional remarks. For a more comprehensive survey of the relevant literature, including previous lower bounds in algebraic and other concrete settings, as well as previous results on the cell probe model, see [10] and the references therein.

2 Preliminaries

Let C_d denote the d-dimensional binary cube $\{0,1\}^d$. For $p, q \in C_d$, let $H(p,q)$ denote the Hamming distance between p and q (i.e., the number of coordinates in which they differ).

Definition. Let $\lambda \in [0,d]$. Let $p, q \in C_d$. We say that q is a λ-*neighbor* of p (and vice-versa) iff $H(p,q) \leq \lambda$. Let $D \subseteq C_d$. We say that q is a λ-neighbor of D iff there exists $p \in D$ such that q is a λ-neighbor of p. For $q \in C_d$, we denote by $B_\lambda(q)$ the set of all λ-neighbors of q.

A two-party boolean (asymmetric) communication problem is specified by two input sets X and Y, and a boolean function $f : X \times Y \to \{0,1\}$. Informally, one party (Alice) gets an element $x \in X$, and the other party (Bob) get an element $y \in Y$. Their goal is to compute $f(x,y)$ by exchanging as few bits as possible according to a specified protocol. The *communication complexity* of the problem is the number of bits transmitted by each side. In a probabilistic protocol, the sides can use random bits to determine the protocol. For every input, the output is correct with a certain probability. A *two sided error* protocol returns the correct output with probability at least $2/3$. An $[a,b]$-*protocol* for the communication problem is a sequence of bit transmissions alternating between Alice and Bob, where the total number of bits Alice sends is at most a and the total number of bits Bob sends is at most b. It is convenient to specify a communication problem by its *communication matrix*. The rows of the matrix are labeled by the elements of X, and the columns are labeled by the elements of Y. An entry labeled (x,y) has value $f(x,y)$.

We are interested in the λ-neighbor problem (λN), where $X = C_d$, Y is the set of all n-tuples $(y^1, y^2, \ldots, y^n) \in C_d^n$,[2] and for $x \in X$, $y \in Y$, $f(x, y) = 1$ iff x is a λ-neighbor of y. We call an element of X a *query*, and an element of Y a *database*. Abusing notation, we denote the function f by λN. We denote the complement of λN by nλN (again, abusing notation, this is both a problem and a function). Notice that for two sided error protocols, lower bounding the communication complexity of a problem is equivalent to lower bounding the communication complexity of the complement problem. In order to derive asymptotic bounds, we consider an infinite sequence of such problems, for increasing values of n and $d = d(n)$. We assume that $d \in \omega(\log n) \cap n^{o(1)}$.

In order to derive our lower bounds, we use the following definition and lemma due to Miltersen et al. [31].

Definition. A communication problem $f : X \times Y \to \{0, 1\}$ is α-dense if

$$\frac{|\{(x, y) \in X \times Y; f(x, y) = 1\}|}{|X \times Y|} \geq \alpha.$$

The following lemma presents the richness technique of Miltersen et al. for two sided error protocols.

Lemma 4 (Miltersen et al. [31]). Let $\alpha, \beta > 0$. Let $f : X \times Y \to \{0, 1\}$ be an α-dense problem. If f has a randomized two sided error $[a, b]$-protocol, then the communication matrix for f contains a sub-matrix M of dimension at least

$$\frac{|X|}{2^{O(a)}} \times \frac{|Y|}{2^{O(a+b)}},$$

such that the fraction of zero entries in M is at most β. (The hidden constants depend only on α and β.) □

3 Lower Bounds for λ-Neighbor

The purpose of this section is to prove the following lower bound on the communication complexity of nλN. This, in turn, implies Theorem 1 giving time/space tradeoffs for nearest neighbor search.

[2] We allow multiple copies of the same point in order to simplify the analysis. Our results hold without substantial changes if the n points must be distinct.

In what follows we denote $\gamma = \frac{2}{\ln 2} \approx 2.885$, and put $\lambda = \frac{d}{2} - C\sqrt{d \log n}$, where $C \approx \frac{1}{\sqrt{\gamma}}$ is defined below.

Theorem 5. If there is a two sided error $[a, b]$-protocol for nλN; then, either $a = \Omega(d)$ or $b = \Omega(n^{\delta})$, where δ is any constant less than $\frac{1}{8}$.

The rest of this section is devoted to the proof of this theorem. The main idea of the proof is to show that every large set of queries contains large subsets of queries that are mutually far apart. For queries that are almost λ apart, nλN behaves "somewhat independently" on random databases. We begin with some properties of balls and intersections of balls in the cube.

Claim 6. Let $q \in C_d$. Then,

$$n^{-(\gamma+\nu)C^2} 2^d \leq |B_\lambda(q)| \leq n^{-(\gamma-\nu)C^2} 2^d,$$

$\nu = \nu(n)$ is monotonically decreasing in n, and moreover $\lim_{n\to\infty} \nu(n) = 0$.

For intuition, consider a uniform distribution over C_d. If p is a random point from this distribution, then the expected distance $H(p, q)$ is $\frac{d}{2}$. Hence, by the Chernoff bound (see, e.g., [4]),

$$\Pr[H(p,q) \leq \lambda] = \Pr\left[H(p,q) \leq \frac{d}{2} - C\sqrt{2\log n}\sqrt{\frac{d}{2}}\right] \leq e^{-(C\sqrt{\log n})^2} \leq n^{-C^2},$$

and therefore $|B_\lambda(q)| \leq n^{-C^2} 2^d$. Proving the tighter bounds stated in the claim requires estimating the binomial distribution directly (using Stirling's formula).

Proof. We first prove the lower bound.

$$|B_\lambda(q)| = \sum_{i=0}^{\lambda} \binom{d}{i} \geq \binom{d}{\lambda}$$

Recall that by Stirling's formula $\sqrt{2\pi k}(k/e)^k \leq k! \leq (1 + 1/4k)\sqrt{2\pi k}(k/e)^k$. Thus we have

$$\binom{d}{\lambda} = \frac{d!}{\lambda!\,(d-\lambda)!}$$

$$
\begin{aligned}
&\geq \frac{\sqrt{2\pi d}\left(\frac{d}{e}\right)^d}{\sqrt{2\pi\lambda}\left(1+\frac{1}{4\lambda}\right)\left(\frac{\lambda}{e}\right)^\lambda \cdot \sqrt{2\pi(d-\lambda)}\left(1+\frac{1}{4(d-\lambda)}\right)\left(\frac{d-\lambda}{e}\right)^{d-\lambda}} \\
&= \frac{2^{(d+1/2)\log d-(\lambda+1/2)\log\lambda-(d-\lambda+1/2)\log(d-\lambda)}}{\sqrt{2\pi}\left(1+\frac{1}{4\lambda}\right)\left(1+\frac{1}{4(d-\lambda)}\right)} \\
&\geq 2^{(d+1/2)\log d-(\lambda+1/2)\log\lambda-(d-\lambda+1/2)\log(d-\lambda)-4}. \qquad (1)
\end{aligned}
$$

We now explore the exponent. Because $\lambda = \frac{d}{2} - C\sqrt{d\log n}$,

$$
\begin{aligned}
\log\lambda &= \log\left(\frac{d}{2} - C\sqrt{d\log n}\right) \\
&= \log\left(\frac{d}{2}\left(1 - 2C\sqrt{\frac{\log n}{d}}\right)\right) \\
&= \log\left(\frac{d}{2}\right) + \log\left(1 - 2C\sqrt{\frac{\log n}{d}}\right).
\end{aligned}
$$

Using the Taylor expansion for $\ln(1-x)$ we have

$$
\ln\left(1 - 2C\sqrt{\frac{\log n}{d}}\right) = -2C\sqrt{\frac{\log n}{d}} - 2C^2\frac{\log n}{d} - o\left(\frac{\log n}{d}\right),
$$

where the last term follows from the fact that $d \in \omega(\log n)$.

Similarly,

$$
\log(d-\lambda) = \log\left(\frac{d}{2} + C\sqrt{d\log n}\right) = \log\left(\frac{d}{2}\right) + \log\left(1 + 2C\sqrt{\frac{\log n}{d}}\right),
$$

and

$$
\ln\left(1 + 2C\sqrt{\frac{\log n}{d}}\right) = 2C\sqrt{\frac{\log n}{d}} - 2C^2\frac{\log n}{d} + o\left(\frac{\log n}{d}\right).
$$

Assigning into the exponent in (1) we get

$$
\left(d+\frac{1}{2}\right)\log d -
$$

$$\left(\frac{d}{2} - C\sqrt{d\log n} + \frac{1}{2}\right)\left(\log\left(\frac{d}{2}\right) - \gamma C\sqrt{\frac{\log n}{d}} - \gamma C^2\frac{\log n}{d} - o\left(\frac{\log n}{d}\right)\right) -$$

$$\left(\frac{d}{2} + C\sqrt{d\log n} + \frac{1}{2}\right)\left(\log\left(\frac{d}{2}\right) + \gamma C\sqrt{\frac{\log n}{d}} - \gamma C^2\frac{\log n}{d} + o\left(\frac{\log n}{d}\right)\right) - 4 =$$

$$d\log d + \frac{1}{2}\log d - d\log\left(\frac{d}{2}\right) - \log\left(\frac{d}{2}\right) - \gamma C^2\log n + o(\log n) =$$

$$d - \gamma C^2\log n - \frac{1}{2}\log d + o(\log n).$$

Because $d \in n^{o(1)}$, there exists $\nu > 0$, where $\nu \to 0$ as $n \to \infty$, such that

$$\frac{1}{2}\log d - o(\log n) \leq \nu C^2\log n$$

Hence, the term in (1) is at least

$$2^{d-\gamma C^2\log n-(1/2)\log d+o(\log n)} \geq 2^{d-(\gamma+\nu)C^2\log n}.$$

Hence,

$$|B_\lambda(q)| \geq n^{-(\gamma+\nu)C^2}2^d.$$

On the other hand a similar argument gives the upper bound. In this case, we use

$$|B_\lambda(q)| = \sum_{i=0}^{\lambda}\binom{d}{i} \leq \lambda\binom{d}{\lambda}.$$

By the Stirling formula,

$$\begin{aligned}
\lambda\binom{d}{\lambda} &= \lambda\frac{d!}{\lambda!\,(d-\lambda)!} \\
&\leq \frac{d}{2}\frac{\left(1+\frac{1}{4d}\right)\sqrt{2\pi d}\left(\frac{d}{e}\right)^d}{\sqrt{2\pi\lambda}\left(\frac{\lambda}{e}\right)^\lambda\cdot\sqrt{2\pi(d-\lambda)}\left(\frac{d-\lambda}{e}\right)^{d-\lambda}} \\
&\leq 2^{\log(d/2)}2^{(d+1/2)\log d-(\lambda+1/2)\log\lambda-(d-\lambda+1/2)\log(d-\lambda)}.
\end{aligned}$$

Using similar arguments as for the lower bound, and the fact that we can set ν so that

$$\frac{1}{2}\log d + o(\log n) \leq \nu C^2\log n,$$

we get,

$$\lambda\binom{d}{\lambda} \leq 2^{d-(\gamma-\nu)C^2 \log n}.$$

Hence,

$$|B_\lambda(q)| \leq \lambda\binom{d}{\lambda} \leq n^{-(\gamma-\nu)C^2}2^d.$$

□

Lemma 7. Let $0 < \nu < \gamma$. For all sufficiently large n there exists C, $1/\sqrt{\gamma+\nu} \leq C \leq 1/\sqrt{\gamma-\nu}$, for which $2^d/n \leq |B_\lambda(q)| < 2^{d+1}/n$. (Recall that λ depends on C.)

Proof.

By Claim 6, for $C = \frac{1}{\sqrt{\gamma-\nu}}$,

$$n^{-(\gamma+\nu)/(\gamma-\nu)}2^d \leq |B_\lambda(q)| \leq n^{-1}2^d;$$

and for $C = \frac{1}{\sqrt{\gamma+\nu}}$,

$$n^{-1}2^d \leq |B_\lambda(q)| \leq n^{-(\gamma-\nu)/(\gamma+\nu)}2^d.$$

If we could claim for continuity of the size of $B_\lambda(q)$ we could claim that there is a C for which the size is exactly $n^{-1}2^d$. Instead, we show that by increasing the radius of a ball by one, the volume will not increase by more than twice, and from that the lemma follows.

We show that if we increase the radius by one to be $\frac{d}{2} - x$, with $x = o(d)$, the volume of the ball at most doubles, as

$$\binom{d}{\frac{d}{2}-x-2} + \binom{d}{\frac{d}{2}-x-1} \geq \binom{d}{\frac{d}{2}-x}.$$

To see this, notice that

$$\left(\frac{d}{2}-x-1\right)\left(\frac{d}{2}-x\right) + \left(\frac{d}{2}+x+2\right)\left(\frac{d}{2}-x\right) \geq \left(\frac{d}{2}+x+1\right)\left(\frac{d}{2}+x+2\right),$$

as

$$\frac{d^2}{2} - o(d^2) \geq \frac{d^2}{4} + o(d^2)$$

for sufficiently large d. Thus,

$$\frac{1}{\left(\frac{d}{2}+x+1\right)\left(\frac{d}{2}+x+2\right)}+\frac{1}{\left(\frac{d}{2}-x-1\right)\left(\frac{d}{2}+x+1\right)}\geq\frac{1}{\left(\frac{d}{2}-x\right)\left(\frac{d}{2}-x-1\right)},$$

which completes the proof. □

In what follows we set C to the value guaranteed by Lemma 7, thus setting the value of $\lambda = \frac{d}{2}-C\sqrt{d\log n}$. We denote the size of $B_\lambda(q)$ as $\xi\frac{2^d}{n}$, where $1\leq\xi<2$. Notice that, although C is not constant, for all n large enough, because $\nu\to 0$, we have $C\approx 1/\sqrt{\gamma}\approx 0.5887$.

Definition. Let $\epsilon>0$, and let $q^1,q^2\in C_d$. We say that q^1 and q^2 are *ϵ-close* iff $H(q^1,q^2)\leq\left(\frac{1}{2}-\sqrt{\epsilon}\right)d$. Otherwise, we say that q^1 and q^2 are *ϵ-far*.

Lemma 8. For every ϵ, $\frac{1}{36}>\epsilon>0$,[3] there exists $\delta>0$ such that the following holds for all n sufficiently large. If $q^1,q^2\in C_d$ are ϵ-far, then

$$\left|B_\lambda(q^1)\cap B_\lambda(q^2)\right|\leq\frac{\xi}{n^{1+\delta}}|C_d|.$$

Proof. Consider a uniform probability distribution over C_d, and let p be a random point from this distribution. We show that $\Pr[p\in B_\lambda(q^2)\mid p\in B_\lambda(q^1)]\leq n^{-\delta}$. As $\Pr[p\in B_\lambda(q^1)]=\frac{\xi}{n}$, the claim follows.

To see that, notice that choosing p uniformly at random in $B_\lambda(q^1)$ is equivalent to the following experiment: Choose a distance r, $0\leq r\leq\lambda$, with probability $\frac{\binom{d}{r}}{|B_\lambda(q^1)|}$. Then, for a given r, choose sequentially, uniformly, without replacement, a set of r coordinates $I=\{i_1,i_2,\ldots,i_r\}$. Finally, put $p_j=1-q^1_j$ for all $j\in I$, and $p_j=q^1_j$ otherwise.

Define p^t by $p^t_j=1-q^1_j$ for all $j\in\{i_1,i_2,\ldots,i_t\}$, and $p^t_j=q^1_j$ otherwise. (Thus $p^0=q^1$ and $p^r=p$.) For a fixed r, we define the following sequence of random variables:

$$X_t=E[H(p,q^2)|p^t].$$

As $X_t=E[X_{t+1}|p^t]$, the sequence is a martingale, in which $X_r=H(p,q^2)$. As for the value of X_0, from linearity of expectation

$$X_0=E[H(p,q^2)]=r\left(1-H(q^1,q^2)/d\right)+(d-r)H(q^1,q^2)/d=$$

[3]The upper bound $\frac{1}{36}$ is a somewhat arbitrary constant that can be improved.

$$\frac{d}{2} + \left(\frac{1}{2} - \frac{r}{d}\right)\left(2H(q^1, q^2) - d\right) > \frac{d}{2} - \left(\frac{1}{2} - \frac{r}{d}\right)\left(2d\sqrt{\epsilon}\right),$$

where the last inequality follows from the fact that
$H(q^1, q^2) > d/2 - d\sqrt{\epsilon}$.

We consider two cases, according to the value of r.

Case 1:

$$\frac{d}{2} - 3C\sqrt{d \log n} \leq r \leq \frac{d}{2} - C\sqrt{d \log n} = \lambda.$$

(The constant 3 could be reduced to be close to $\sqrt{9/8}$, yielding a wider range for ϵ.) In this case,

$$X_0 \geq \frac{d}{2} - \left(\frac{1}{2} - \frac{r}{d}\right)\left(2d\sqrt{\epsilon}\right) \geq \frac{d}{2} - 6C\sqrt{\epsilon d \log n}.$$

We have

$$\begin{aligned}
\Pr[H(p, q^2) \leq \lambda] &= \Pr\left[X_r \leq \lambda\right] \\
&= \Pr\left[X_r \leq \frac{d}{2} - C\sqrt{d \log n}\right] \\
&= \Pr\left[X_r \leq \frac{d}{2} - 6C\sqrt{\epsilon d \log n} - SC\sqrt{d \log n}\right],
\end{aligned}$$

for some $0 < S < 1$ (recall that $\epsilon < \frac{1}{36}$). We need the following

Fact 9. The martingale $\{X_t\}$ satisfies the Lipschitz condition $|X_t - X_{t+1}| \leq 2$, for all $0 \leq t \leq r-1$.

Proof. Consider an arbitrary step t of the process of choosing coordinates. Let W_t denote the set of coordinates that have not been chosen by step t. Let $V_t \subseteq W_t$ be the subset where q^1 and q^2 agree, and let $U_t \subseteq W_t$ be the subset where q^1 and q^2 differ. Let $v_t = |V_t|$, and let $u_t = |U_t|$. Let P_t denote the probability of a coordinate in W_t to be chosen in step $t+1$. (I.e., $P_t = (r-t)/(d-t)$.) Finally, denote by X^v_{t+1} the value of X_{t+1} in case we choose in step $t+1$ out of V_t, and denote by X^u_{t+1} the value of X_{t+1} in case we choose in step $t+1$ out of U_t.

Notice that the expected distance after step t is

$$X_t = d - H(q^1, q^2) - v_t + u_t + P_t v_t - P_t u_t.$$

Also

$$X^v_{t+1} = d - H(q^1, q^2) - (v_t - 1) + u_t + P_{t+1}(v_t - 1) - P_{t+1} u_t,$$

and

$$X^u_{t+1} = d - H(q^1, q^2) - v_t + u_t - 1 + P_{t+1}v_t - P_{t+1}(u_t - 1).$$

Therefore

$$X^v_{t+1} - X^u_{t+1} = 2(1 - P_{t+1}) \leq 2. \tag{2}$$

Now, X_t is a convex combination of X^v_{t+1} and X^u_{t+1}, so $|X_t - X_{t+1}| \leq 2$. □

We proceed to get

$$\begin{aligned}
\Pr[H(p, q^2) \leq \lambda] &\leq \Pr\left[X_r \leq X_0 - SC\sqrt{d \log n}\right] \\
&= \Pr\left[X_r \leq X_0 - SC\sqrt{2 \log n}\sqrt{\frac{d}{2}}\right] \\
&\leq \Pr\left[X_r \leq X_0 - SC\sqrt{2 \log n}\sqrt{r}\right] \\
&= \Pr\left[\frac{1}{2}X_r \leq \frac{1}{2}X_0 - \frac{1}{2}SC\sqrt{2 \log n}\sqrt{r}\right] \\
&\leq e^{-\frac{1}{4}S^2C^2 \log n} = 2^{-\frac{1}{4}S^2C^2 \log e \log n} \\
&= n^{-\frac{1}{8}\gamma S^2C^2},
\end{aligned}$$

where the first inequality follows from $X_0 \geq d/2 - 6C\sqrt{\epsilon d \log n}$, the second inequality follows from $r < \frac{d}{2}$, and the third inequality follows from applying Azuma's Inequality to the martingale $\{\frac{1}{2}X_t\}$.

Case 2:

$$r < \frac{d}{2} - 3C\sqrt{d \log n}.$$

In this case we use Claim 6 to bound
$|B_r(q^1)| \leq n^{-(\gamma-\nu)9C^2}2^d$, and $|B_\lambda(q^1)| \geq n^{-(\gamma+\nu)C^2}2^d$. Therefore,

$$\Pr\left[r < \frac{d}{2} - 3C\sqrt{d \log n}\right] \leq \frac{|B_r(q^1)|}{|B_\lambda(q^1)|} < n^{-8\gamma C^2 + 10\nu C^2} \leq n^{-7},$$

for all sufficiently large n.

Summing up the two cases, we get

$$\Pr[H(p, q^2) \leq \lambda] \leq n^{-\frac{1}{8}\gamma S^2C^2} + n^{-7} \leq n^{-\delta},$$

for all sufficiently large n, and for every δ which is slightly smaller than $\gamma S^2 C^2/8$. □

Notice, as $\epsilon \to 0$, $S \to 1$, and as $n \to \infty$, $C \to \frac{1}{\sqrt{\gamma}}$. So, we get $\delta \approx \frac{1}{8}$. For sufficiently large n, we can set δ as a function of ϵ alone. In what follows $\delta = \delta(\epsilon)$ denotes the δ guaranteed by Lemma 8.

Consider the graph G_ϵ with node set C_d, and edges connecting pairs of points which are ϵ-close.

Definition. Let $I \subseteq C_d$. We say that I is *ϵ-apart* iff I is an independent set in G_ϵ (i.e., iff all the pairs $q^1, q^2 \in I$, $q^1 \neq q^2$, are ϵ-far).

Claim 10. Let $R \subseteq C_d$, $|R| \geq 2n^\delta |C_d| 2^{-\epsilon d}$. Then, assuming n is sufficiently large, there exists $I \subseteq R$, $|I| = 2n^\delta$, such that I is ϵ-apart.

Claim 11. For every $q \in C_d$, the degree of q in G_ϵ (i.e., the number of points which are ϵ-close to q) is strictly less than $|C_d| \cdot 2^{-\epsilon d}$.

Proof. By standard Chernoff bounds (see, e.g., [4]), the probability that a point q' chosen uniformly at random in C_d has $H(q, q') \leq \frac{d}{2} - d\sqrt{\epsilon}$ is at most $e^{-\epsilon d}$. Thus, for every $q \in C_d$, the degree of q in G_ϵ is strictly less than $\Delta = |C_d| \cdot 2^{-\epsilon d}$. Obviously, this is also true in the subgraph of G_ϵ induced by R. Therefore, R must contain an independent set of size at least

$$\frac{|R|}{\Delta} \geq 2n^\delta,$$

for n sufficiently large. (The greedy algorithm will output such a set.) □

We are now ready for

Lemma 12. Let $R \subseteq C_d$, $|R| = 4n^\delta |C_d| 2^{-\epsilon d}$. Then there exist disjoint ϵ-apart subsets $I_1, I_2, \ldots, I_z \subseteq R$ whose union contains at least half of the points in R, such that the cardinality of each subset is $2n^\delta$, and the number of subsets $z = 2^{n^{o(1)}}$.

Proof. Repeatedly apply Claim 10 to get an ϵ-apart subset of the remainder of R. This can be done as long as the remaining set has cardinality at least $2n^\delta |C_d| 2^{-\epsilon d}$. The number of subsets z is clearly upper bounded by $|R|/2n^\delta = 2|C_d|2^{-\epsilon d}$. As we assume that $d \in n^{o(1)}$, the bound on z follows. □

Lemma 13. Let $I \subseteq C_d$ be an ϵ-apart set, $|I| \geq n^\delta$. Let p be chosen uniformly at random in C_d. Then, $\Pr[p \text{ is a } \lambda\text{-neighbor of } I] \geq \frac{\xi}{2n^{1-\delta}}$.

Proof. Assume $|I| = n^\delta$. (Otherwise, take a subset of I of cardinality n^δ.) Using Lemmas 7 and 8, and the Bonferroni Inequalities,

$$\begin{aligned}\Pr[p \text{ is a } \lambda\text{-neighbor of } I] &\geq |I|\left(\frac{\xi}{n}\right) - \binom{|I|}{2}\frac{\xi}{n^{1+\delta}} \\ &= \frac{\xi}{n^{1-\delta}} - \frac{\xi n^\delta(n^\delta - 1)}{2n^{1+\delta}} \\ &\geq \frac{\xi}{2n^{1-\delta}}.\end{aligned}$$

□

We are now ready for bounding the communication complexity of nλN.

Lemma 14. nλN is α-dense for some constant α.

Proof. We show that in each row of the communication matrix of nλN at least a constant fraction of the entries are 1. Fix $q \in C_d$. Choose a database D uniformly at random in C_d^n. If $p \in C_d$ is chosen uniformly at random, then

$$\Pr[H(p,q) \leq \lambda] = \frac{\xi}{n}.$$

Thus,

$$\Pr[\text{n}\lambda\text{N}(q,D) = 1] = \Pr[\forall p \in D;\ H(p,q) > \lambda] = \left(1 - \frac{\xi}{n}\right)^n \geq e^{-\zeta}$$

for some constant $\zeta > \xi$. □

Claim 15. Let $I \subseteq C_d$, such that $|I| = 2n^\delta$ and I is ϵ-apart. Then the number of databases in C_d^n such that nλN$(q,D) = 1$ for less than a fraction of $\frac{1}{20}$ of the queries in I is at most $2^{nd - n^\delta/32}$.

Proof. Consider a database chosen uniformly at random in C_d^n. We think of the database as being chosen point by point. Let $x_1, x_2, \ldots, x_n$ be the (random) points of the database, in the order they are chosen. Let $D_i = \{x_{in^{1-\delta}+1}, \ldots, x_{(i+1)n^{1-\delta}}\}$, for $0 \leq i \leq n^\delta - 1$.

Let $M_0, M_1, M_2, \ldots, M_{n^\delta}$ be the following (random) subsets of I:

$$M_0 = \emptyset$$

$$M_{i+1} = \begin{cases} M_i \cup \{q\} & \text{if } \exists q \in I \setminus M_i,\ q \text{ is a } \lambda\text{-neighbor of } D_i; \\ M_i & \text{otherwise.} \end{cases}$$

(If more than one such q exists, pick one of them arbitrarily.) Denote $Z = |M_{n^\delta}|$. (Notice that this is a random variable.) Define a sequence of random variables $X_0, X_1, \ldots, X_{n^\delta}$ as follows.

$$X_i = E[Z \mid M_i].$$

As $X_i = E[X_{i+1} \mid M_i]$, the sequence is a martingale. Notice that $|X_i - X_{i-1}| \leq 1$ and that $X_{n^\delta} = Z$. Furthermore, for every i, $|I \setminus M_i| > n^\delta$. Therefore, by Lemma 13, the probability that a random database point is a λ-neighbor of $I \setminus M_i$ is at least $\frac{\xi}{2n^{1-\delta}}$. Hence the probability that none of the points of D_i are λ-neighbors of $I \setminus M_i$ is at most

$$\left(1 - \frac{\xi}{2n^{1-\delta}}\right)^{n^{1-\delta}} \leq e^{-\xi/2} \leq e^{-1/2}.$$

Therefore, by linearity of expectation, $X_0 \geq (1 - e^{-1/2})n^\delta$.

We use Azuma's Inequality to get

$$\Pr[X_{n^\delta} < 2n^\delta/20] \leq \Pr[X_{n^\delta} < (1 - e^{-1/2} - 1/4)n^\delta] \leq$$

$$\Pr[X_{n^\delta} < X_0 - (1/4)n^{\delta/2}n^{\delta/2}] \leq e^{-n^\delta/32}.$$

□

Now, we can show that there are no large nearly monochromatic sub-matrices in the communication matrix of nλN.

Lemma 16. For all sufficiently large n, in any $n^\delta 2^{(1-\epsilon)d+2} \times 2^{nd-n^\delta/33}$ sub-matrix of the communication matrix of nλN a fraction of at least $\frac{1}{80}$ of the entries are zeros.

Proof. Consider a sub-matrix $A \times B$ of the specified dimensions. Partition at least half of A into sets $I_1, I_2, \ldots, I_z$, as in Lemma 12. By Claim 15, for every I_j, $1 \leq j \leq z$, the number of databases $D \in B$ such that less than $\frac{1}{20}$ of the points in I_j are λ-neighbors of D is at most $2^{nd-n^\delta/32}$. Hence, the number of databases $D \in B$ such that there exists j for which less than $\frac{1}{20}$ of the points in I_j are λ-neighbors of D is at most

$$z \cdot 2^{nd-n^\delta/32} \leq 2^{nd-n^\delta/32+n^{o(1)}} \leq 2^{nd-n^\delta/33-1},$$

for n sufficiently large. Therefore, since at least half of the databases in B have at least $\frac{1}{20}$ of the points in any I_j as λ-neighbors (i.e., at least half of the databases in B have as λ-neighbors at least $\frac{1}{40}$ of the points in A), the fraction of zero entries in $A \times B$ is at least $\frac{1}{80}$. □

We are now ready to prove Theorem 5, using the Miltersen et al. richness technique.

Proof of Theorem 5: By Lemma 14, nλN is α-dense. By Lemma 16, every sub-matrix of size $n^{\delta}2^{(1-\epsilon)d+2} \times 2^{nd-n^{\delta}/33}$ has a fraction of at least $\beta = \frac{1}{80}$ of zero entries. Applying Lemma 4, we get that if there is an $[a, b]$-protocol for nλN, then, either

$$\frac{|C_d|}{2^{O(a)}} < |C_d| 2^{-\epsilon d+2} n^{\delta},$$

or

$$\frac{|C_d^n|}{2^{O(a+b)}} < |C_d^n| 2^{-n^{\delta}/33}.$$

The first inequality implies that $a = \Omega(\epsilon d - \delta \log n) = \Omega(d)$ (as $d = \omega(\log n)$). The second inequality implies that $a + b = \Omega(n^{\delta})$. Assuming that $a = o(d)$, this gives $b = \Omega(n^{\delta})$ (as $d = n^{o(1)}$). □

Proof of Theorem 1: Suppose there is a cell probe algorithm for NNS using s cells of size b each and at most t probes. In particular, this algorithm solves nλN. By Lemma 3 this implies a $[t\lceil \log s \rceil, tb]$-protocol for n$\lambda$N. By Theorem 5, either $t\lceil \log s \rceil = \Omega(d)$, or $tb = \Omega(n^{\delta})$. □

4 Lower bounds for Partial λ-Neighbor

In this section we discuss partially specified queries and prove Theorem 2. The problem we consider here generalizes the λ-neighbor problem in the cube. As in Section 3, the database consists of n points in C_d. The queries are taken from a set $Q_{d,k}$, defined as

$$Q_{d,k} = \left\{ q \in \{0, 1, *\}^d : |\{i : q_i \neq *\}| = k \right\}.$$

The character $*$ stands for "don't care", and it matches both a 0 and a 1. The entries of the query which are not $*$ are called the *exposed* bits of the query. Given $p \in C_d$ and $q \in Q_{d,k}$, we define $H'(p, q)$, the distance between q and p, as follows:

$$H'(p, q) = |\{i : q_i \neq * \ \wedge \ q_i \neq p_i\}|.$$

Partial λ-neighbor (PλN) is the problem of deciding for a query $q \in Q_{d,k}$ whether or not there is a database point at distance at most λ from q. We denote by nPλN the complement problem.

4.1 Lower Bounds for Communication Complexity of PλN

In this section we give lower bounds on the communication complexity of PλN which imply the trade-off lower bound for the cell probe model in Theorem 2. We give here a lower bound for the case in which the set of possible queries is $Q_{d,\rho d}$ (when $\rho \leq 1$ is a rational constant. For a given ρ we consider only d-s for which ρd is an integer.[4]

Recall from Section 3 that $\gamma = \frac{2}{\ln 2}$. Using $C \approx \sqrt{\frac{\rho}{\gamma}}$ to be defined later, we put $\lambda = \frac{\rho d}{2} - C\sqrt{d \log n}$. We define $B_\lambda(q) = \{p \in C_d : H'(p,q) \leq \lambda\}$. We prove the following theorem.

Theorem 17. If there is a two sided error $[a,b]$-protocol for nPλN; then, either $a = \Omega(d)$ or $b = \Omega(n^\delta)$, where δ is any constant less than $\frac{\rho}{8(2-\rho)}$.

The proof of this theorem is similar to the proof of the analogous Theorem 5 from Section 3. Hence, we show in detail only the arguments where there is a major difference between the two proofs.

The following claim is an easy modification to Claim 6.

Claim 18. Let $q \in Q_{d,\rho d}$. Then

$$n^{-(\gamma+\nu)C^2/\rho} 2^d \leq |B_\lambda(q)| \leq n^{-(\gamma-\nu)C^2/\rho} 2^d,$$

$\nu = \nu(n)$ is monotonically decreasing in n, and moreover $\lim_{n\to\infty} \nu(n) = 0$. □

Lemma 7 can be modified to give:

Lemma 19. Let $0 < \nu < \gamma$. For all sufficiently large n there exists C,

$$\sqrt{\frac{\rho}{\gamma+\nu}} \leq C \leq \sqrt{\frac{\rho}{\gamma-\nu}},$$

for which $2^d/n \leq |B_\lambda(q)| < 2^{d+1}/n$. □

We will set C to the value of guaranteed by the above lemma. This sets $\lambda = \frac{\rho d}{2} - C\sqrt{d \log n}$. We denote the size of $B_\lambda(q)$ as $\xi \frac{2^d}{n}$, where $1 \leq \xi < 2$. Notice that, although C is not a constant, for all n large enough, $C \approx \sqrt{\rho/\gamma}$.

[4]More generally, our argument can be extended to handle the case of queries taken from $Q_{d,g(d)}$ such that $\rho = \lim \frac{g(d)}{d} = \text{const}$. We omit the details.

Definition. Let $q^1, q^2 \in Q_{d,\rho d}$. Then their *overlap*, denoted $h(q^1, q^2)$, is defined by

$$h(q^1, q^2) = |\{i : q_i^1 \neq * \ \wedge \ q_i^2 \neq *\}|.$$

The *distance* between q^1 and q^2, denoted $H'(q^1, q^2)$, is defined by

$$H'(q^1, q^2) = |\{i : q_i^1 \neq * \ \wedge \ q_i^2 \neq * \ \wedge \ q_i^1 \neq q_i^2\}|.$$

Definition. Let $\epsilon > 0$, and let $q^1, q^2 \in Q_{d,\rho d}$. We say that q^1 and q^2 are ϵ-*close* iff one of the following two conditions holds:

1. $h(q^1, q^2) \geq \rho^2 d + d\sqrt{2\epsilon}$; or,
2. $h(q^1, q^2) < \rho^2 d + d\sqrt{2\epsilon}$ and $H'(q^1, q^2) \leq \frac{h(q^1,q^2)}{2} - d\sqrt{2\epsilon}$.

Otherwise, we say that q^1 and q^2 are ϵ-*far*.

Lemma 20. For every ϵ, $\frac{\rho^2}{72} > \epsilon > 0$, there exists $\delta > 0$ such that the following holds for all n sufficiently large. If $q^1, q^2 \in Q_{d,\rho d}$ are ϵ-far, then

$$\left|B_\lambda(q^1) \cap B_\lambda(q^2)\right| \leq \frac{\xi}{n^{1+\delta}}|C_d|.$$

Proof. As q^1, q^2 are ϵ-far, we have $h(q^1, q^2) < \rho^2 d + d\sqrt{2\epsilon}$ and $H'(q^1, q^2) > \frac{1}{2}h(q^1, q^2) - d\sqrt{2\epsilon}$. Consider a uniform probability distribution over C_d, and let p be a random point from this distribution. We show that

$$\Pr[p \in B_\lambda(q^2) \mid p \in B_\lambda(q^1)] \leq n^{-\delta}.$$

As $\Pr[p \in B_\lambda(q^1)] = \frac{\xi}{n}$, the claim follows.

To see that, notice that choosing p uniformly at random in $B_\lambda(q^1)$ is equivalent to the following experiment: Let $I_{q^1} = \{i : \ q_i^1 = *\}$, and let $I_{q^1}^c$ be the indices of the remaining entries. Denote the elements of I_{q^1} by $i_1, i_2, \ldots, i_{(1-\rho)d}$. Choose a distance r, $0 \leq r \leq \lambda$, with probability $\frac{\binom{\rho d}{r}}{|B_\lambda(q^1)|}$. For the chosen r, choose sequentially, uniformly, without replacement, a set of r coordinates $I = \{i_{(1-\rho)d+1}, \ldots, i_{(1-\rho)d+r}\}$ from $I_{q^1}^c$. For every $j \in I_{q^1}$, choose p_j to be 0 or 1 uniformly and independently. For every other index j, put $p_j = 1 - q_j^1$ if $j \in I$ and $p_j = q_j^1$ otherwise. Define p^t by

$p_j^t = p_j$ for all $j \in \{i_1, i_2, \ldots, i_t\}$, and $p_j^t = q_j^1$ otherwise. (Thus $p^0 = q^1$ and $p^{(1-\rho)d+r} = p$.) Notice that p^t is not in C_d for any $t < (1-\rho)d$.

For a fixed r, we define the following sequence of random variables:

$$X_t = E[H'(p, q^2)|p^t].$$

As $X_t = E[X_{t+1}|p^t]$ the sequence is a martingale, in which $X_{(1-\rho)d+r} = H'(p, q^2)$. We now examine the value of X_0. Divide the exposed bits of q^2 into three parts (see figure 1): The first part (A) has no overlap with the exposed bits of q^1. Its size is $\rho d - h(q^1, q^2)$. The second part (B) is where q^2 is identical to q^1. Its size is $h(q^1, q^2) - H'(q^1, q^2)$. The third part (C) is where q^2 differs from q^1. Its size is $H'(q^1, q^2)$.

Figure 1: Overlap and Equivalence Segments

As $X_0 = E[H'(p, q^2)]$, we notice that the expected contribution of the coordinates in (A) to the distance is $\frac{1}{2}|A|$. Similarly, the expected contribution of (B) is $\frac{r}{\rho d}|B|$, and the expected contribution of (C) is $\left(1 - \frac{r}{\rho d}\right)|C|$. Hence, by the linearity of expectation,

$$\begin{aligned} X_0 &= E[H'(p, q^2)] = \frac{1}{2}\left(\rho d - h(q^1, q^2)\right) + \frac{r}{\rho d}\left(h(q^1, q^2) - H'(q^1, q^2)\right) \\ &+ \left(1 - \frac{r}{\rho d}\right) H'(q^1, q^2) \\ &= \frac{\rho d}{2} + \left(\frac{1}{2} - \frac{r}{\rho d}\right)\left(2H'(q^1, q^2) - h(q^1, q^2)\right) \\ &> \frac{\rho d}{2} + \left(\frac{1}{2} - \frac{r}{\rho d}\right)\left(-2d\sqrt{2\epsilon}\right), \end{aligned}$$

where the last inequality follows from the fact that $H'(q^1, q^2) > \frac{1}{2}h(q^1, q^2) - d\sqrt{2\epsilon}$.

We consider two cases, according to the value of r.

Case 1:

$$\frac{\rho d}{2} - 3C\sqrt{d\log n} \leq r \leq \frac{\rho d}{2} - C\sqrt{d\log n} = \lambda.$$

(The constant 3 could be reduced to be close to $\sqrt{\frac{16-7\rho}{16-8\rho}}$, yielding a wider range for ϵ.) In this case,

$$X_0 \geq \frac{\rho d}{2} + \left(\frac{1}{2} - \frac{r}{\rho d}\right)\left(-2d\sqrt{2\epsilon}\right) \geq \frac{\rho d}{2} - \frac{6C}{\rho}\sqrt{2\epsilon d\log n}. \tag{3}$$

Notice that

$$\begin{aligned}\Pr[H'(p,q^2) \leq \lambda] &= \Pr\left[X_{(1-\rho)d+r} \leq \lambda\right] \\ &= \Pr\left[X_{(1-\rho)d+r} \leq \frac{\rho d}{2} - \frac{6C}{\rho}\sqrt{2\epsilon d\log n} - SC\sqrt{d\log n}\right], \end{aligned} \tag{4}$$

for some $0 < S < 1$ (recall that $\epsilon < \frac{\rho^2}{72}$).

We need the following technical claim. Its proof appears at the end of Section 4.1.

Claim 21. The martingale $\{X_t\}$ satisfies, for all $0 \leq t \leq (1-\rho)d + r - 1$, the Lipschitz condition $|X_t - X_{t+1}| \leq 2$.

Using the above, we get

$$\begin{aligned}\Pr[H'(p,q^2) \leq \lambda] &\leq \Pr\left[X_{(1-\rho)d+r} \leq X_0 - SC\sqrt{d\log n}\right] \\ &= \Pr\left[X_{(1-\rho)d+r} \leq X_0 - SC\sqrt{\frac{\log n}{1-\rho/2}}\sqrt{(1-\frac{\rho}{2})d}\right] \\ &\leq \Pr\left[X_{(1-\rho)d+r} \leq X_0 - SC\sqrt{\frac{\log n}{1-\rho/2}}\sqrt{(1-\rho)d+r}\right] \\ &= \Pr\left[\frac{1}{2}X_{(1-\rho)d+r} \leq \frac{1}{2}X_0 - \frac{1}{2}SC\sqrt{\frac{\log n}{1-\rho/2}}\sqrt{(1-\rho)d+r}\right] \\ &\leq e^{-\frac{S^2C^2\log n}{8(1-\rho/2)}} = 2^{-\frac{S^2C^2\log e\log n}{8(1-\rho/2)}} = n^{-\frac{\gamma S^2C^2}{8(2-\rho)}}, \end{aligned}$$

where the first inequality follows from Equation 4 and Inequality 3, the second inequality follows from $r < \frac{\rho d}{2}$, and the third inequality follows from applying Azuma's Inequality to the martingale $\{\frac{1}{2}X_t\}$.

Case 2:

$$r < \frac{\rho d}{2} - 3C\sqrt{d \log n}.$$

In this case we use Claim 18 to bound

$$|B_r(q^1)| \leq n^{-(\gamma-\nu)9C^2/\rho} 2^d,$$

and

$$|B_\lambda(q^1)| \geq n^{-(\gamma+\nu)C^2/\rho} 2^d.$$

Therefore,

$$\Pr\left[r < \frac{\rho d}{2} - 3C\sqrt{d \log n}\right] < n^{-8\gamma C^2/\rho + 10\nu C^2/\rho} \leq n^{-7}$$

for all sufficiently large n.

Summing up the two cases, we get

$$\Pr[H'(p, q^2) \leq \lambda] \leq n^{-\frac{\gamma S^2 C^2}{8(2-\rho)}} + n^{-7} \leq n^{-\delta},$$

for all sufficiently large n, and for every δ which is slightly smaller than $\frac{\gamma S^2 C^2}{8(2-\rho)}$.

Notice that as $\epsilon \to 0$, $S \to 1$, and as $n \to \infty$, $C \to \sqrt{\frac{\rho}{\gamma}}$. So, we get $\delta \approx \frac{\rho}{8(2-\rho)}$. For sufficiently large n, we can set δ as a function of ϵ and ρ alone. In what follows $\delta = \delta(\epsilon, \rho)$ denotes the δ guaranteed by Lemma 20.

Consider the graph G'_ϵ with node set $Q_{d,\rho d}$, and edges connecting pairs of points which are ϵ-close.

Claim 22. For every $q \in Q_{d,\rho d}$, the degree of q in G'_ϵ (i.e., the number of points which are ϵ-close to q) is strictly less than $|Q_{d,\rho d}| \cdot 2^{-\epsilon d}$.

Proof. Choose $\tilde{q} \in Q_{d,\rho d}$ uniformly at random. This is equivalent to the following random experiment: Choose sequentially, uniformly, without replacement, a set of ρd coordinates $I = \{i_1, i_2, \ldots, i_{\rho d}\}$. For every coordinate $j \in I$ choose $\tilde{q}_j$ to be either 0 or 1 with equal probability. For every other coordinate j, set $\tilde{q}_j = *$. For every t, $0 \leq t \leq \rho d$, define $\tilde{q}^t$ as $\tilde{q}^t_j = \tilde{q}_j$ for all $j \in \{i_1, i_2, \ldots, i_t\}$, and $\tilde{q}^t_j = *$ otherwise. (Thus $\tilde{q}^{\rho d} = \tilde{q}$.)

We first analyze the probability that q and $\tilde{q}$ are ϵ-close because $h(q, \tilde{q}) \geq \rho^2 d + d\sqrt{2\epsilon}$. Define a sequence of random variables X_t as follows.

$$X_t = E[h(q, \tilde{q})|\tilde{q}^t].$$

As $X_t = E[X_{t+1}|\tilde{q}^t]$ this sequence is a martingale. Moreover, $X_0 = \rho^2 d$ and $X_{\rho d} = h(q,\tilde{q})$. Notice that $|X_t - X_{t-1}| \leq 1$ and therefore by Azuma's inequality:

$$\begin{aligned}\Pr\left[h(q,\tilde{q}) \geq \rho^2 d + d\sqrt{2\epsilon}\right] &= \Pr\left[X_{\rho d} \geq X_0 + \sqrt{\frac{2\epsilon d}{\rho}}\sqrt{\rho d}\right] \\ &\leq e^{\frac{-\epsilon d}{\rho}} \leq 2^{-4/3\epsilon d}.\end{aligned}$$

Now consider the case that q and $\tilde{q}$ are ϵ-close because $h(q,\tilde{q}) < \rho^2 d + d\sqrt{2\epsilon}$ and $H'(q,\tilde{q}) \leq \frac{h(q,\tilde{q})}{2} - d\sqrt{2\epsilon}$. In this case, we are only concerned with the values that are assigned to the $h(q,\tilde{q})$ places in $\tilde{q}$ that overlap those of q. We have $h(q,\tilde{q})$ independent trials with the expectation of $H'(q,\tilde{q})$ being $\frac{h(q,\tilde{q})}{2}$. Therefore, by standard Chernoff bounds we have

$$\Pr[H'(q,\tilde{q}) < \frac{h(q,\tilde{q})}{2} - d\sqrt{2\epsilon}] \leq e^{-\frac{2\epsilon d^2}{h(q,\tilde{q})}} \leq 2^{-2\epsilon d}.$$

Summing over the two cases we get that the probability that a random $\tilde{q}$ is ϵ-close to q is at most

$$2^{-2\epsilon d} + 2^{-4/3\epsilon d} < 2^{-\epsilon d},$$

for sufficiently large d.

We are now ready to prove Theorem 17, using the Miltersen et al. richness technique.

Proof of Theorem 17: Arguing as in Section 3, using Lemmas 19 and 20, and Claim 22, and assuming that $\rho > \epsilon$, one can show that for all sufficiently large n, in any $\binom{d}{\rho d} n^\delta 2^{(\rho-\epsilon)d+2} \times 2^{nd-n^\delta/33}$ sub-matrix of hte communication matrix for nPλN, a fraction of at least $\frac{1}{80}$ of the entries are zeros. Also, arguing as in Section 3, it is not difficult to show that nPλN is α-dense for some constant α. Applying Lemma 4, we get that if there is an $[a,b]$-protocol for nλN, then, either

$$\frac{|Q_{d,\rho d}|}{2^{O(a)}} < |Q_{d,\rho d}| 2^{-\epsilon d+2} n^\delta,$$

or

$$\frac{|C_d^n|}{2^{O(a+b)}} < |C_d^n| 2^{-n^\delta/33}.$$

This implies the theorem. □

Proof of Claim 21: We prove that $|X_t - X_{t+1}| \leq 2$ for the two different cases, considering the value of t. Consider an arbitrary step t of the process of choosing coordinates and their value. First

assume $t+1 \leq (1-\rho)d$. This means that the exposed bits of p^{t+1} are still equal to the exposed bits of q^1. Denote by w_t the number of coordinates that were assigned a value which differs from the matching coordinate of q^2, by step t. Then,

$$X_t = w_t + \frac{1}{2}((1-\rho)d - t) + \frac{r}{\rho d}\left(h(q^1,q^2) - H'(q^1,q^2)\right) + \left(1 - \frac{r}{\rho d}\right) H'(q^1,q^2),$$

and then either

$$\begin{aligned} X_{t+1} &= w_t + \frac{1}{2}((1-\rho)d - (t+1)) \\ &+ \frac{r}{\rho d}\left(h(q^1,q^2) - H'(q^1,q^2)\right) + \left(1 - \frac{r}{\rho d}\right) H'(q^1,q^2); \end{aligned}$$

or,

$$\begin{aligned} X_{t+1} &= w_t + 1 + \frac{1}{2}((1-\rho)d - (t+1)) \\ &+ \frac{r}{\rho d}\left(h(q^1,q^2) - H'(q^1,q^2)\right) + \left(1 - \frac{r}{\rho d}\right) H'(q^1,q^2). \end{aligned}$$

In both cases, $|X_t - X_{t+1}| = \frac{1}{2} \leq 2$.

Now assume $t \geq (1-\rho)d$, which means that $p^t \in C_d$. Denote by W_t the set of coordinates that have not been chosen by step t, by V_t the subset where q^1 and q^2 agree, and by U_t the subset where q^1 and q^2 differ. Let $v_t = |V_t|$, let $u_t = |U_t|$, and let $t' = t - (1-\rho)d$. Let P_t be the probability that a coordinate in W_t is chosen in step $t+1$ (i.e., $P_t = (r-t')/(\rho d - t')$). Denote by X^v_{t+1} the value of X_{t+1} in case we choose in step $t+1$ out of V_t, and by X^u_{t+1} the value of X_{t+1} in case we choose in step $t+1$ out of U_t. Finally, let w_t be the number of coordinates of I_{q^1} that were assigned a value that differs from the matching coordinate of q^2, by step t. (See figure 2.) Notice that the expected distance after step t is

$$X_t = w_t + h(q^1,q^2) - H'(q^1,q^2) - v_t + u_t + P_t v_t - P_t u_t.$$

Also

$$X^v_{t+1} = w_t + h(q^1,q^2) - H'(q^1,q^2) - (v_t - 1) + u_t + P_{t+1}(v_t - 1) - P_{t+1}u_t,$$

and

$$X^u_{t+1} = w_t + h(q^1,q^2) - H'(q^1,q^2) - v_t + u_t - 1 + P_{t+1}v_t - P_{t+1}(u_t - 1).$$

Figure 2: Notation of the Chosen Segments

Therefore

$$X^v_{t+1} - X^u_{t+1} = 2(1 - P_{t+1}) \leq 2. \tag{5}$$

As X_t is a convex combination of X^v_{t+1} and X^u_{t+1}, we have $|X_t - X_{t+1}| \leq 2$. □

4.2 Reductions for Different Rates of Exposed Bits

In this section we show that the partial λ-neighbor problem does not become less difficult as ρ increases.

Lemma 23. Let $0 < \rho \leq \rho' \leq 1$, $\rho' \leq 3\rho$, and let $C' > 0$. Then, for all sufficiently large n, for all $d \in \omega(\log n)$, there exist λ, $n' > n$, and $\lambda' = 2\rho' d - 2C'\sqrt{d \log n'}$, and there exist efficiently computable functions $\phi_1 : Q_{d,\rho d} \to Q_{4d,\rho' 4d}$ and $\phi_2 : C^n_d \to C^{n'}_{4d}$, such that for every $q \in Q_{d,\rho d}$ and $D \in C^n_d$, q is a λ-neighbor of D iff $\phi_1(q)$ is a λ'-neighbor of $\phi_2(D)$.

Proof. Put $n' = \lceil n^{4\rho/\rho'} \rceil$.[5] Let $C = C'\sqrt{\frac{\rho}{\rho'}}$, and let $\lambda = \frac{\rho d}{2} - C\sqrt{d \log n}$.

Given $q \in Q_{d,\rho d}$ and $D \in C^n_d$ we define the functions $q' = \phi_1(q)$ and $D' = \phi_2(D)$ as follows. Let $I = \{i_1, i_2, \ldots, i_t\}$ be the first t coordinates of q which are $*$, where $t = (\rho' - \rho)d$. For

[5]For the sake of simplicity, we will omit the ceiling notation from the rest of the proof. The reader can verify easily that this does not affect the validity of the argument.

$i \in \{1, 5, 9, \ldots, 4d-3\}$ define,

$$q'_i q'_{i+1} q'_{i+2} q'_{i+3} = \begin{cases} q_{\lceil i/4 \rceil} q_{\lceil i/4 \rceil} q_{\lceil i/4 \rceil} q_{\lceil i/4 \rceil} & q_{\lceil i/2 \rceil} \neq * ; \\ 0011 & \left(q_{\lceil i/4 \rceil} = *\right) \wedge \left(\lceil i/4 \rceil \in I\right) ; \\ **** & \text{otherwise.} \end{cases}$$

As for ϕ_2, first apply the transformation ϕ_1 to each of the n points of the database D. Then, add the $n' - n$ first (in lexicographic order) points with the property that for all $0 \leq i \leq 2(d-1)$ either $p_{2i+1}p_{2i+2} = 10$ or $p_{2i+1}p_{2i+2} = 01$.[6] Denote this subset of D' by D''.

We now claim that q is a λ-neighbor of D iff $q' = \phi_1(q)$ is a λ'-neighbor of $D' = \phi_2(D)$. First notice that

$$\forall p \in D'' \ : \ H(p, q') = 2\rho' d > 2\rho' d - 2C'\sqrt{d \log n'} = \lambda'.$$

I.e., q' is a λ'-neighbor of D' only if it is a λ'-neighbor of $D' \setminus D''$.

As for $p \in D' \setminus D''$, notice that the coordinates in I contribute to the distance exactly $2t$. Furthermore, the coordinates where $q \neq *$ contribute to the distance a factor of four times their original contribution. Hence, for $p \in D$,

$$H(\phi_1(p), \phi_1(q)) = 2t + 4H(p, q).$$

Therefore, $H(p,q) \leq \frac{\rho d}{2} - C\sqrt{d \log n} = \lambda$ iff $2(\rho' - \rho)d + 4H(p,q) \leq 2\rho' d - 4C\sqrt{d \log n}$ iff $2t + 4H(p,q) \leq 2\rho' d - C\sqrt{\frac{\rho'}{\rho}}\sqrt{16\frac{\rho}{\rho'} d \log n}$ iff $H(\phi_1(p), \phi_1(q)) \leq \frac{\rho' 4d}{2} - C'\sqrt{4d \log n'} = \lambda'$. (Recall that $n' = n^{4\rho/\rho'}$, so $\log n' = 4\frac{\rho}{\rho'} \log n$.)

Corollary 24. Let $0 < \rho \leq \rho' \leq 1$. For every $C' > 0$, for all sufficiently large n, for all $d \in \omega(\log n)$, there exist λ, $n' = \text{poly}(n)$, $d' = \text{poly}(d)$, and $\lambda' = 2\rho' d - 2C'\sqrt{d \log n'}$, and an efficient reduction from the partial λ-neighbor problem with the parameters ρ, n, and d, to the partial λ'-neighbor problem with parameters ρ', n', and d'.

Proof. This follows from lemma 23, which claims the same for $\rho' \leq 3\rho$. If this is not the case, perform the reduction shown in Lemma 23 a constant number of times.

Notice that we can take C' to be the constant stipulated by Lemma 19 for the parameters ρ', n', and d'.

[6] Notice that there are 2^{2d} such vectors, because as $d \in \omega(\log n')$, then $n < 2^{2d}$. Furthermore, $\rho/\rho' \geq 1/3$ and thus $n' > n$.

References

[1] P.K. Agarwal and J. Matoušek. Ray shooting and parametric search. In *Proc. of 24th STOC*, pp. 517–526, 1992.

[2] M. Ajtai. Determinism versus non-determinism for linear-time RAMs. In *Proc. of 31st STOC*, pp. 632–641, 1999.

[3] M. Ajtai. A non-linear time lower bound for Boolean branching programs. Preprint, April 1999.

[4] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley, 1992.

[5] S. Arya and D. Mount. Approximate nearest neighbor searching. In *Proc. of 4th SODA*, pp. 271–280, 1993.

[6] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. In *Proc. of 5th SODA*, pp. 573-582, 1994.

[7] P. Beame, M. Saks, and J.S. Thathachar. Time-space tradeoffs for branching programs. In *Proc. of 39th FOCS*, pp. 254–263, 1998.

[8] J.S. Beis and D.G. Lowe. Shape indexing using approximate nearest-neighbor search in high-dimensional spaces. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, pp. 1000–1006, 1997.

[9] M. de Berg, M. van Kreveld, M. Overmars, O. Schwarzkopf. Computational Geometry, Algorithms and Applications. Springer, 1997.

[10] A. Borodin , R. Ostrovsky, and Y. Rabani. Lower bounds for high dimensional nearest neighbor search and related problems. In *Proc. of 31th STOC*, pp. 312–321, 1999.

[11] A. Chakrabarti, B. Chazelle, B. Gum, A. Lvov. A good neighbor is hard to find. In *Proc. of 31st STOC*, pp. 305–311, 1999.

[12] K. Clarkson. A randomized algorithm for closest-point queries. *SIAM J. Comput.*, 17:830–847, 1988.

[13] K. Clarkson. An algorithm for approximate closest-point queries. In *Proc. of 10th SCG*, pp. 160–164, 1994.

[14] S. Cost and S. Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10:57–67, 1993.

[15] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE-IT*, 13:21–27, 1967.

[16] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. Info. Sci.*, 41(6):391–407, 1990.

[17] L. Devroye and T.J. Wagner. Nearest neighbor methods in discrimination. In *Handbook of Statistics*, Vol. 2, P.R. Krishnaiah and L.N. Kanal eds. North Holland, 1982.

[18] D. Dobkin and R. Lipton. Multidimensional search problems. *SIAM J. Comput.*, 5:181–186, 1976.

[19] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P.Yanker. Query by image and video content: the QBIC system. *IEEE Computer*, 28:23–32, 1995.

[20] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[21] R. Fagin. Fuzzy queries in multimedia database systems. In *Proc. of PODS*, 1998.

[22] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic, 1991.

[23] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. In *1st International Conf. on Knowledge Discovery and Data Mining*, 1995.

[24] P. Indyk. Dimensionality reduction techniques for proximity problems. Manuscript, August 1999. To appear in *SODA 2000*.

[25] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. of 30th STOC*, pp. 604–613, 1998.

[26] J. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proc. of 29th STOC*, pp. 599–608, 1997.

[27] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. In *Proc. of 30th STOC*, pp. 614–623, 1998.

[28] J. Matoušek. Reporting points in halfspaces. In *Proc. of 32nd FOCS*, pp. 207–215, 1991.

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

[30] P.B. Miltersen. Lower bounds for union-split-find related problems on random access machines. In *Proc. of 26th STOC*, pp. 625–634, 1994.

[31] P.B. Miltersen, N. Nisan, S. Safra, and A. Widgerson. On data-structures and asymmetric communication complexity. In *Proc. of 27th STOC*, pp. 103–111, 1995.

[32] A. Pentland, R.W. Picard, and S. Sclaroff. Photobook: tools for content-based manipulation of image databases. In *Proc. SPIE Conf. on Storage and Retrieval of Image and Video Databases II*, 1994.

[33] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[34] A.W.M. Smeulders and R. Jain (eds). *Proc. 1st Workshop on Image Databases and Multi-Media Search*, 1996.

[35] A.C. Yao. Should tables be sorted? *J. Ass. Comp. Mach.*, 28:615–628, 1981.

[36] A.C. Yao and F.F. Yao. A general approach to d-dimension geometric queries. In *Proc. of 17th STOC*, pp. 163–168, 1985.