Lower bounds on Locality Sensitive Hashing

Rajeev Motwani*  
Stanford University  
Stanford, CA 94305-9045  
rajeev@cs.stanford.edu

Assaf Naor  
Microsoft Research  
Redmond WA, 98052-6399  
anor@microsoft.com

Rina Panigrahi†  
Stanford University  
Stanford, CA 94305-9045  
rinap@cs.stanford.edu

ABSTRACT

Given a metric space \( (X, d_X) \), \( c \geq 1, r > 0 \), and \( p, q \in [0, 1] \), a distribution over mappings \( \mathcal{H} : X \rightarrow \mathbb{N} \) is called a \((r, c r, p, q)\)-sensitive hash family if any two points in \( X \) at distance at most \( r \) are mapped by \( \mathcal{H} \) to the same value with probability at least \( p \), and any two points at distance greater than \( c r \) are mapped by \( \mathcal{H} \) to the same value with probability at most \( q \). This notion was introduced by Indyk and Motwani in 1998 as the basis for an efficient approximate nearest neighbor search algorithm, and has since been used extensively for this purpose. The performance of these algorithms is governed by the parameter \( \rho = \frac{\log(1/p)}{\log(1/q)} \), and constructing hash families with small \( \rho \) automatically yields improved nearest neighbor algorithms. Here we show that constructing hash families with small \( \rho \) achieves most matches the construction of Indyk and Motwani which achieves \( \rho \leq \frac{1}{2} \).

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Probabilistic algorithms;  
E.1 [Data Structures]: Miscellaneous

General Terms

Algorithms, Theory

Keywords

Nearest Neighbor Search, Locality Sensitive Hashing, Lower Bounds

1. INTRODUCTION

In this note we study the complexity of finding the nearest neighbor of a query point in certain high dimensional spaces

\[ d_X(x, y) \leq r \Leftrightarrow \Pr_{\mathcal{H}}[\mathcal{H}(x) = \mathcal{H}(y)] \geq p . \]

\[ d_X(x, y) > R \Leftrightarrow \Pr_{\mathcal{H}}[\mathcal{H}(x) = \mathcal{H}(y)] < q . \]

Given \( c \geq 1 \) and \( q \in (0, 1) \) we define \( \rho_X(c, q) \) to be the smallest constant \( \rho > 0 \) such that for every \( r > 0 \) there exists \( p \in (0, 1) \) and a \((r, c r, p, q)\)-sensitive hash family \( \mathcal{H} : X \rightarrow \mathbb{N} \) with \( \frac{\log(1/p)}{\log(1/q)} \leq \rho \). In other words

\[ \rho_X(c, q) = \sup_{r>0} \inf_{p>0} \frac{\log(1/p)}{\log(1/q)} . \]

\[ \exists (r, c r, p, q) - \text{sensitive hash family } \mathcal{H} : X \rightarrow \mathbb{N} \] . (1)

Of particular interest is the case \( X = \ell^d_s \), for some \( s > 0 \) and \( d \in \mathbb{N} \). Here, and in what follows, \( \ell^d_s \) denotes the space \( \mathbb{R}^d \) equipped with the \( \ell^s \) norm

\[ \|(x_1, \ldots, x_d)\|_s = (|x_1|^s + \cdots + |x_d|^s)^{1/s} \]
(this is only a quasi-norm when \(0 < s < 1\)). In this case we define
\[
\rho_s(c) = \sup_{0 < q < 1} \limsup_{d \to \infty} \rho_d(q, c).
\]

The importance of these parameters stems from the following application to approximate nearest neighbor search. It will be convenient to discuss it in the framework of the following decision version of the \(c\)-approximate nearest neighbor problem: Given a query point, find any point from the data set which is at distance at most \(cr\) from it, provided that there is a data point at distance at most \(r\) from the query point. This decision version is known as the \((r, cr)\)-near neighbor problem. It is well known that the reduction to the decision version adds only a logarithmic factor in the time and space complexity [6, 5]. The following theorem was proved in [6]; the exact formulation presented here is taken from [4].

**Theorem 1.2.** Let \((X, d_X)\) be a metric on a subset of \(\mathbb{R}^d\). Suppose that \((X, d_X)\) admits a \((r, cr, p, q)\)-sensitive hash family \(\mathcal{H}\), and write \(\rho = \frac{\log(1/p)}{\log(1/q)}\). Then for any \(n \geq \frac{1}{q}\) there exists a randomized algorithm for \((r, cr)\)-near neighbor on \(n\)-point subsets of \(X\) which uses \(O(dn + n^{1+p})\) space, with query time dominated by \(O(n^p)\) distance computations and \(O(n^p \log \log n)\) evaluations of hash functions from \(\mathcal{H}\).

Thus, obtaining bounds on \(\rho_{x}(c)\) is of great algorithmic interest. It is proved in [6] that \(\rho_1(c) \leq 1/c\), and for small values of \(c\), namely \(c \in [1, 10]\), is shown in [4] that this inequality is strict. We refer to [4] for numerical data on the best known estimates for \(\rho_1(c)\) for small \(c\). For \(c = 2\) a recent result of Andoni and Indyk [1] shows that \(\rho_2(c) \leq 1/c^2\), and for general \(s \in (0, 2]\) the best known bounds [4] are \(\rho_s(c) \leq \max\{1/c, 1/c^s\}\).

The main purpose of this note is to obtain lower bounds on \(\rho_1(c)\) and \(\rho_2(c)\) which nearly match the bounds obtained from the constructions in [6, 4, 1]. Our main result is:

**Theorem 1.3.** For every \(c, s \geq 1\),
\[
\rho_s(c) \geq \frac{2^{1/2} - 1}{2^{1/2} + 1} \geq \frac{c - 1}{c + 1} \cdot \frac{1}{c^s} \geq \frac{0.462}{c^s}.
\]

The second to last inequality in (2) follows from concavity of the function \(t \mapsto \frac{2^{1/2} - 1}{2^{1/2} + 1} t\) on \([0, \infty)\). Observe also that as \(c \to \infty\), \(\frac{1/c^s - 1}{1/c^s + 1} \sim \frac{1}{2c^s}\). It would be very interesting to determine \(\lim_{c \to \infty} \rho_1(c)\) exactly- due to Theorem 1.3 and the results of [6] we currently know that this number is in the interval \([1/2, 1]\).

2. **Proof of Theorem 1.2**

The basic idea in the proof of Theorem 1.3 is simple. Choose a random point \(x \in \{0, 1\}^d\) and consider the random subset \(A\) of the cube \(\{0, 1\}^d\) consisting of points \(u\) for which \(\mathcal{H}(u) = \mathcal{H}(x)\). The second condition in Definition 1.1 forces \(A\) to be small in expectation. But, when \(A\) is small we can bound from above the probability that after \(r\) steps, the random walk starting at a random point in \(A\) will end up in \(A\). We obtain this upper bound using a Fourier analytic argument, and in combination with the first condition in Definition 1.1 we deduce the desired bound on \(\rho_1(c)\).

Theorem 1.3 follows from the following result:

**Proposition 2.1.** Let \(\mathcal{H}\) be a \((r, R, p, q)\)-sensitive hash family on the Hamming cube \(\{0, 1\}^d, \|\cdot\|_1\). Assume that \(r\) is an odd integer and that \(R < \frac{d}{2}\). Then
\[
p \leq \left( q + e^{-\frac{1}{2}(\frac{d}{2} - R)^2} \right)^{2r/d - 1} e^{s(r, q + 1)}.
\]

Choosing \(R \approx \frac{d}{2} - \sqrt{\log d}\) and \(r \approx R/c\) in Proposition 2.1, and letting \(d \to \infty\), yields Theorem 1.3 in the case \(s = 1\). The case of general \(s \geq 1\) follows from the fact that for \(x, y \in \{0, 1\}^d, \|x - y\|_1 = \|x - y\|_{1/s}\).

**Remark 2.1.** Proposition 2.1 implies a non-trivial lower bound on \(\log(1/p)\) for any \((r, cr, p, q)\)-sensitive hash family on \(\{0, 1\}^d, \|\cdot\|_1\) even if \(q\) is allowed to depend on \(d\). Observe that with the definition given in (1), Theorem 1.3 implies such a lower bound only for constant \(q\). But, Proposition 2.1 is much stronger, and implies a bound which asymptotically coincides with the lower bound in 1.3 for every \(q \geq 2^{-o(d)}\).

The proof of Proposition 2.1 will be broken into a few lemmas.

**Lemma 2.2.** Let \(\mathcal{H}\) be a \((r, R, p, q)\)-sensitive hash family on the Hamming cube \(\{0, 1\}^d, \|\cdot\|_1\), and fix \(x \in \{0, 1\}^d\). Then
\[
E[\mathcal{H}^{-1}(\mathcal{H}(x))] \leq \sum_{k=0}^{\lfloor R \rfloor} \binom{d}{k} + q \cdot \sum_{k=\lceil R \rceil + 1}^{d} \binom{d}{k}.
\]

**Proof.** We simply write
\[
E[\mathcal{H}^{-1}(\mathcal{H}(x))] = \sum_{u \in \{0, 1\}^d} \Pr[\mathcal{H}(u) = \mathcal{H}(x)]
\leq \left| \{u \in \{0, 1\}^d : \|u - x\|_1 \leq R \} \right| + q \cdot \left| \{ u \in \{0, 1\}^d : \|u - x\|_1 > R \} \right|
\leq \sum_{k=0}^{\lfloor R \rfloor} \binom{d}{k} + q \cdot \sum_{k=\lceil R \rceil + 1}^{d} \binom{d}{k},
\]
which is the required inequality.

**Corollary 2.3.** Assume that \(R < \frac{d}{2}\). Then, using the notation of Lemma 2.2, we have that
\[
E[\mathcal{H}^{-1}(\mathcal{H}(x))] \leq 2^d \left( q + e^{-\frac{1}{2}(\frac{d}{2} - R)^2} \right).
\]

**Proof.** This follows from Lemma 2.2 and the standard estimate \(\sum_{k \leq \frac{d}{2}} \binom{d}{k} \leq 2^d\).
$S \subseteq \{1, \ldots, d\}$, the Walsh function $W_S : \{0, 1\}^d \rightarrow \{-1, 1\}$ is defined by

$$W_S(u) = (-1)^{\sum_{j \in S} u_j}.$$  

For $f : \{0, 1\}^d \rightarrow \mathbb{R}$ we set

$$\hat{f}(S) = \frac{1}{2^d} \sum_{u \in \{0, 1\}^d} f(u)W_S(u),$$

so that $f$ can be decomposed as follows:

$$f = \sum_{S \subseteq \{1, \ldots, d\}} \hat{f}(S)W_S.$$

For every $f, g : \{0, 1\}^d \rightarrow \mathbb{R}$ we write

$$\langle f, g \rangle = \frac{1}{2^d} \sum_{u \in \{0, 1\}^d} f(u)g(u).$$

By Parseval’s identity,

$$\langle f, g \rangle = \sum_{S \subseteq \{1, \ldots, d\}} \hat{f}(S)\hat{g}(S).$$

For $\varepsilon \in [0, 1]$ the Bonami-Beckner operator $T_\varepsilon$ is defined as

$$T_\varepsilon f = \sum_{S \subseteq \{1, \ldots, d\}} \varepsilon^{\|S\|} \hat{f}(S)W_S.$$  

The Bonami-Beckner inequality \cite{3, 2} states that for every $f : \{0, 1\}^d \rightarrow \mathbb{R}$,

$$\sum_{S \subseteq \{1, \ldots, d\}} \varepsilon^{\|S\|} \hat{f}(S)^2 = \|T_\varepsilon f\|^2_2 = \frac{1}{2^d} \sum_{u \in \{0, 1\}^d} (T_\varepsilon f(u))^2 \leq \|f\|^2_2 \varepsilon^{\frac{d}{1+\varepsilon^2}}.$$  

Specializing to the indicator of $B \subseteq \{0, 1\}^d$ we get that

$$\sum_{S \subseteq \{1, \ldots, d\}} \varepsilon^{\|S\|} \hat{1}_B(S)^2 \leq \left( \frac{|B|}{2^d} \right)^{\frac{d}{1+\varepsilon^2}}. \tag{3}$$

Now, let $P$ be the transition matrix of the standard random walk on $\{0, 1\}^d$, i.e. $P_{uv} = 1/d$ if $u$ and $v$ differ in exactly one coordinate, $P_{uv} = 0$ otherwise. By a direct computation we have that for every $S \subseteq \{1, \ldots, d\}$,

$$PW_S = \left(1 - \frac{2|S|}{d}\right)W_S,$$

i.e. $W_S$ is an eigenvector of $P$ with eigenvalue $1 - \frac{2|S|}{d}$. The probability that the random walk starting form a random point in $B$ ends up in $B$ after $r$ steps equals

$$\Pr[Q_B \in B] = \frac{1}{|B|}\sum_{a,b \in B} (P^r)_{ab}$$

$$= \frac{2^d}{|B|} \sum_{S \subseteq \{1, \ldots, d\}} \hat{1}_B(S)^2 \left(1 - \frac{2|S|}{d}\right)^r$$

$$\leq \frac{2^d}{|B|} \sum_{S \subseteq \{1, \ldots, d\}} \hat{1}_B(S)^2 \left(1 - \frac{2|S|}{d}\right)^r,$$

where we used the fact that $r$ is odd (i.e. we dropped negative terms).

Thus, using (3) we see that

$$\Pr[Q_B \in B] \leq \frac{2^d}{|B|} \sum_{S \subseteq \{1, \ldots, d\}} \hat{1}_B(S)^2 \cdot e^{-2r|S|/d}$$

$$= \left( \frac{|B|}{2^d} \right)^{\frac{d}{1+\varepsilon^2}} \cdot e^{-2r|S|/d},$$

completing the proof Lemma 2.4. \qed

**Proof of Proposition 2.1.** Assume that $r$ is an odd integer and $R < \frac{d}{2}$. For $x \in \{0, 1\}^d$ let $W_r(x) \in \{0, 1\}^d$ be the random point obtained by preforming a random walk for $r$ steps starting at $x$. Since $\|x - W_r(x)\|_1 \leq r$ we know that $\Pr[H(W_r(x)) = H(x)] = p$. Taking expectation with respect to the uniform probability measure on $\{0, 1\}^d$ we deduce that

$$p \leq \mathbb{E}_{x \in \{0, 1\}^d} \Pr[H(W_r(x)) = H(x)]$$

$$= \mathbb{E}_{x \in \{0, 1\}^d} \Pr[x \in \{0, 1\}^n : W_r(x) \in \mathcal{H}^{-1}(\mathcal{H}(x))]$$

$$= \mathbb{E}_{x \in \{0, 1\}^d} \sum_{k \in \mathbb{N}} \Pr[x \in \{0, 1\}^n : W_r(x) \in \mathcal{H}^{-1}(\mathcal{H}(x)) \wedge \mathcal{H}(x) = k]$$

$$= \mathbb{E}_{x \in \{0, 1\}^d} \sum_{k \in \mathbb{N}} \left( \frac{|\mathcal{H}^{-1}(k)|}{2^d} \right) \Pr[H^{-1}(k)]$$

$$\leq \mathbb{E}_{x \in \{0, 1\}^d} \sum_{k \in \mathbb{N}} \left( \frac{|\mathcal{H}^{-1}(k)|}{2^d} \right) \left( \frac{|\mathcal{H}^{-1}(k)|}{2^d} \right)^{\frac{2r/d}{d}}$$

$$= \mathbb{E}_{x \in \{0, 1\}^d} \left( \frac{|\mathcal{H}^{-1}(H(x))|}{2^d} \right)^{\frac{2r/d}{d}}$$

$$\leq \mathbb{E}_{x \in \{0, 1\}^d} \left( e^{-\frac{2r}{d}} \right)^{\frac{2r/d}{d}}$$

$$\leq \left( q + e^{-\frac{2r}{d}} \right)^{\frac{2r/d}{d}}$$

where in (4) we used Lemma 2.4, in (5) we used Jensen’s inequality, and in (6) we used Corollary 2.3. \qed
3. ACKNOWLEDGEMENTS
We are grateful to Piotr Indyk and Jirka Matoušek for helpful suggestions.

4. REFERENCES
[1] A. Andoni and P. Indyk. Faster algorithms for high dimensional nearest neighbor problems. Manuscript, 2005.
[2] W. Beckner. Inequalities in Fourier analysis. *Ann. of Math.* (2), 102(1):159–182, 1975.
[3] A. Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier (Grenoble)*, 20(fasc. 2):335–402 (1971), 1970.
[4] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SCG ’04: Proceedings of the Twentieth Annual Symposium on Computational Geometry (Brooklyn, NY, 2004)*, pages 253–262. ACM Press, New York, NY, 2004.
[5] S. Har-Peled. A replacement for Voronoi diagrams of near linear size. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*, pages 94–103. IEEE Computer Soc., Los Alamitos, CA, 2001.
[6] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC ’98: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (Dallas, TX, 1998)*, pages 604–613. ACM Press, New York, NY, 1998.
[7] R. Panigrahy. Entropy based nearest neighbor search in high dimensions. In *SODA ’06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm (Miami, FL, 2006)*, pages 1186–1195. ACM Press, New York, NY, 2006.