Lower bounds on Locality Sensitive Hashing

Rajeev Motwani Assaf Naor Rina Panigrahy

Abstract

Given a metric space (X, d_X), $c \geq 1$, $r > 0$, and $p, q \in [0, 1]$, a distribution over mappings $\mathscr{H} : X \to \mathbb{N}$ is called a (r, cr, p, q)-sensitive hash family if any two points in X at distance at most r are mapped by $\mathscr{H}$ to the same value with probability at least p, and any two points at distance greater than cr are mapped by $\mathscr{H}$ 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 ρ automatically yields improved nearest neighbor algorithms. Here we show that for $X = \ell_1$ it is impossible to achieve $\rho \leq \frac{1}{2c}$. This almost matches the construction of Indyk and Motwani which achieves $\rho \leq \frac{1}{c}$.

1 Introduction

In this note we study the complexity of finding the nearest neighbor of a query point in certain high dimensional spaces using *Locality Sensitive Hashing* (LSH). The nearest neighbor problem is formulated as follows: Given a database of n points in a metric space, preprocess it so that given a new query point it is possible to quickly find the point closest to it in the data set. This fundamental problem arises in numerous applications, including data mining, information retrieval, and image search, where distinctive features of the objects are represented as points in $\mathbb{R}^d$. There is a vast amount of literature on this topic, and we shall not attempt to discuss it here. We refer the interested reader to the papers [6, 5, 4, 7], and especially to the references therein, for background on the nearest neighbor problem.

While the exact nearest neighbor problem seems to suffer from the "curse of dimensionality", many efficient techniques have been devised for finding an approximate solution whose distance from the query point is at most c times its distance from the nearest neighbor. One of the most versatile and efficient methods for approximate nearest neighbor search is based on Locality Sensitive Hashing, as introduced by Indyk and Motwani in 1998 [6]. This method has been refined and improved in several papers- the most recent algorithm can be found in [4]. We also refer the reader to the LSH website, where more information on this algorithm can be found, including its implementation and code- all this can be found at `http://web.mit.edu/andoni/www/LSH/index.html`. The LSH approach to the approximate nearest neighbor problem is based on the following concept.

Definition 1.1. *Let (X, d_X) be a metric space, $r, R > 0$ and $p, q \in [0, 1]$. A distribution over mappings $\mathscr{H} : X \to \mathbb{N}$ is called a (r, R, p, q)-sensitive hash family if for any $x, y \in X$,*

- $d_X(x, y) \leq r \implies \Pr[\mathscr{H}(x) = \mathscr{H}(y)] \geq p$.
- $d_X(x, y) > R \implies \Pr[\mathscr{H}(x) = \mathscr{H}(y)] \leq q$.

Given $c \geq 1$ we define

$$\rho_X(c) = \sup_{r>0} \inf \left\{ \frac{\log(1/p)}{\log(1/q)} : \exists (r, cr, p, q) - \text{sensitive hash family } \mathscr{H} : X \to \mathbb{N} \right\} \quad . \tag{1}$$

Of particular interest is the case $X = \ell_s^d$, for some $s \geq 1$ and $d \in \mathbb{N}$. In this case we define

$$\rho_s(c) = \limsup_{d\to\infty} \rho_{\ell_s^d}(c) \quad .$$

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 element of 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 $\mathscr{H}$, 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, c) near neighbor on n-point subsets of X which uses $O\left(dn + n^{1+\rho}\right)$ space, with query time dominated by $O(n^\rho)$ distance computations and $O\left(n^\rho \log_{1/q} n\right)$ evaluations of hash functions from $\mathscr{H}$.*

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 was shown in [4] that this inequality is strict. We refer to [4] for numerical data on the best know estimates for $\rho_1(c)$ for small c. For $s = 2$ a recent result of Andoni and Indyk [1] shows that $\rho_2(c) \leq 1/c^2$, and for general $s \in [1, 2]$ 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 $c, s \geq 1$,*

$$\rho_s(c) \geq \frac{e^{\frac{1}{c^s}} - 1}{e^{\frac{1}{c^s}} + 1} \geq \frac{e-1}{e+1} \cdot \frac{1}{c^s} \geq \frac{0.462}{c^s} \quad . \tag{2}$$

The second to last inequality in (2) follows from concavity of the function $t \mapsto \frac{e^t-1}{e^t+1}$ on $[0, \infty)$. Observe also that as $c \to \infty$, $\frac{e^{1/c}-1}{e^{1/c}+1} \sim \frac{1}{2c}$. It would be very interesting to determine $\limsup_{c\to\infty} c \cdot \rho_1(c)$ 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 1.3

The basic idea in the proof of Theorem 1.3 is simple. We consider the random subset A of the cube $\{0, 1\}^d$ consisting of points u for which $\mathscr{H}(u) = \mathscr{H}(0)$. 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 $\mathscr{H}$ 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*

$$p \leq \left(q + e^{-\frac{1}{d}\left(\frac{d}{2}-R\right)^2}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} \quad .$$

Choosing $R \approx \frac{d}{2} - \sqrt{d \log d}$ 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 $x, y \in \{0,1\}^d$, $\|x-y\|_s = \|x-y\|_1^{1/s}$.

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

Lemma 2.2. *Let $\mathscr{H}$ be a (r, R, p, q)-sensitive hash family on the Hamming cube $(\{0,1\}^d, \|\cdot\|_1)$. Consider the random subset $S \subseteq \{0,1\}^d$ given by $A = \{u \in \{0,1\}^d : \mathscr{H}(u) = \mathscr{H}(0)\}$. Then*

$$\mathbb{E}|A| \leq \sum_{k=0}^{\lfloor R \rfloor} \binom{d}{k} + q \cdot \sum_{k=\lfloor R \rfloor+1}^{d} \binom{d}{k} \quad .$$

Proof. We simply write

$$\begin{aligned} \mathbb{E}|A| &= \sum_{u \in \{0,1\}^d} \Pr[\mathscr{H}(u) = \mathscr{H}(0)] \\ &\leq \left|\{u \in \{0,1\}^d : \|u\|_1 \leq R\}\right| + q \cdot \left|\{u \in \{0,1\}^d : \|u\|_1 > R\}\right| \\ &= \sum_{k=0}^{\lfloor R \rfloor} \binom{d}{k} + q \cdot \sum_{k=\lfloor R \rfloor+1}^{d} \binom{d}{k} \quad . \end{aligned}$$

□

Corollary 2.3. *Assume that $R < \frac{d}{2}$. Then, using the notation of Lemma 2.2, we have that*

$$\mathbb{E}|A| \leq 2^d \left(q + e^{-\frac{1}{d}\left(\frac{d}{2}-R\right)^2}\right) \quad .$$

Proof. This follows from Lemma 2.2 and the standard estimate $\sum_{k \leq \frac{d}{2}-a} \binom{d}{k} \leq 2^d \cdot e^{-\frac{a^2}{d}}$. □

Lemma 2.4 (Random walk lemma). *Let r be an odd integer. Given $\emptyset \neq B \subseteq \{0,1\}^d$, consider the random variable $Q_B \in \{0,1\}^d$ defined as follows: Choose a point $x \in B$ uniformly at random, and perform r-steps of the standard random walk on the Hamming cube starting from x. The point thus obtained will be denoted Q_B. Then*

$$\Pr[Q_B \in B] \leq \left(\frac{|B|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} \quad .$$

Proof. We begin by recalling some background and notation on Fourier analysis on the Hamming cube. Given $S \subseteq \{1, \ldots d\}$, the Walsh function $W_S : \{0,1\}^d \to \{-1,1\}$ is defined by

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

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

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

so that f can be decomposed as follows:

$$f = \sum_{S\subseteq\{1,\dots,d\}} \widehat{f}(S) W_S \quad .$$

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

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

By Parseval's identity,

$$\langle f, g\rangle = \sum_{S\subseteq\{1,\dots,d\}} \widehat{f}(S)\widehat{g}(S) \quad .$$

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

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

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

$$\sum_{S\subseteq\{1,\dots,d\}} \varepsilon^{2|S|}\widehat{f}(S)^2 = \|T_\varepsilon f\|_2^2 = \frac{1}{2^d} \sum_{u\in\{0,1\}^d} (T_\varepsilon f(u))^2 \le \|f\|_{1+\varepsilon^2}^2 = \left(\frac{1}{2^d} \sum_{u\in\{0,1\}^d} f(u)^{1+\varepsilon^2}\right)^{\frac{2}{1+\varepsilon^2}} \quad .$$

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

$$\sum_{S\subseteq\{1,\dots,d\}} \varepsilon^{2|S|}\widehat{\mathbf{1}_B}(S)^2 \le \left(\frac{|B|}{2^d}\right)^{\frac{2}{1+\varepsilon^2}} \quad . \tag{3}$$

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 $S \subseteq \{1,\dots,d\}$,

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

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

$$\begin{aligned}
\Pr[Q_B \in B] &= \frac{1}{|B|} \sum_{a,b\in B} (P^r)_{ab} \\
&= \frac{2^d}{|B|} \langle P^r \mathbf{1}_B, \mathbf{1}_B\rangle \\
&= \frac{2^d}{|B|} \sum_{S\subseteq\{1,\dots,d\}} \widehat{\mathbf{1}_B}(S)^2 \left(1 - \frac{2|S|}{d}\right)^r \\
&\le \frac{2^d}{|B|} \sum_{\substack{S\subseteq\{1,\dots,d\} \\ |S|\le d/2}} \widehat{\mathbf{1}_A}(S)^2 \left(1 - \frac{2|S|}{d}\right)^r \quad ,
\end{aligned}$$

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

Thus, using (3) we see that

$$\Pr[Q_B \in B] \le \frac{2^d}{|B|} \sum_{S \subseteq \{1,\dots,d\}} \widehat{\mathbf{1}_B}(S)^2 \cdot e^{-2r|S|/c} \le \frac{2^d}{|B|} \cdot \left(\frac{|B|}{2^d}\right)^{\frac{2}{1+e^{-2r/c}}} = \left(\frac{|B|}{2^d}\right)^{\frac{1-e^{-2r/c}}{1+e^{-2r/c}}} .$$

□

Proof of Proposition 2.1. Let A be as in Lemma 2.2. Assume that r is an odd integer and $R < \frac{d}{2}$. Since the Hamming distance from Q_A to A is at most r, we know that $\Pr[Q_A \in A] = \Pr[\mathscr{H}(Q_A) = \mathscr{H}(0)] \ge p$. On the other hand by Lemma 2.4, Corollary 2.3, and Jensen's inequality,

$$p \le \Pr[Q_A \in A] \le \mathbb{E}\left[\left(\frac{|A|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}}\right] \le \left(\frac{\mathbb{E}|A|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} \le \left(q + e^{-\frac{1}{d}\left(\frac{d}{2}-R\right)^2}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} .$$

□

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 *SoCG '04: Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pages 253–262, New York, NY, USA, 2004. ACM Press.

[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*, pages 604–613, New York, NY, USA, 1998. ACM Press.

[7] R. Panigrahy. Entropy based nearest neighbor search in high dimensions. In *SODA '06: Proceedings of the seventeenth annual ACM-SIAM Symposium on Discrete Algorithms*. To appear.