Locality-Preserving Hashing in Multidimensional Spaces

Piotr Indyk *
Department of Computer Science
Stanford University
indyk@cs.stanford.edu

Rajeev Motwani †
Department of Computer Science
Stanford University
rajeev@theory.stanford.edu

Prabhakar Raghavan
IBM Almaden Research Center
pragh@almaden.ibm.com

Santosh Vempala ‡
School of Computer Science
Carnegie-Mellon University
vempala@cs.cmu.edu

February 25, 1997

Abstract

We consider *locality-preserving hashing* — in which adjacent points in the domain are mapped to adjacent or nearly-adjacent points in the range — when the domain is a d-dimensional cube. This problem has applications to high-dimensional search and multimedia indexing. We show that simple and natural classes of hash functions are provably good for this problem. We complement this with lower bounds suggesting that our results are essentially the best possible.

1 Introduction

In a recent paper, Linial and Sasson [21] proved the following theorem about hash functions:

Theorem 1 *There exists a family $\mathcal{G}$ of functions from an integer line $[1, \ldots, U]$ to $[1, \ldots, R]$ and a constant C such that for any $S \subset [1, \ldots, U]$ with $|S| \leq C\sqrt{R}$:*

*Supported by NSF Award CCR-9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

†Supported by an Alfred P. Sloan Research Fellowship, an IBM Faculty Partnership Award, an ARO MURI Grant DAAH04-96-1-0007, and NSF Young Investigator Award CCR-9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

‡Supported in part by NSF National Young Investigator grant CCR-9357793.

- $\Pr_{f \in \mathcal{G}}(f_{|S} \textit{is one to one}) \geq \frac{1}{2}$
- *all* $f \in \mathcal{G}$ *are* non-expansive, *i.e., for any* $p, q \in U$ $d(f(p), f(q)) \leq d(p, q)$.

The family $\mathcal{G}$ *contains* $O(|U|)$ *functions, each of which is computable in* $O(1)$ *operations.*

Their result gives a family of hash functions with the surprising property that points close to each other in the domain are hashed to points close to each other in the range. A potential application of their result is to find near neighbors to points in one-dimensional space. Given a set of points on the line, one can hash these points so that we can search for the points closest to a query point as follows: we hash the query q to $h(q)$, then search the neighborhood of $h(q)$ in the hash table to retrieve, say, the nearest k points within some distance δ in the domain in time $\min\{k, \delta\}$. The advantage of the locality-preserving property is that it affords good paging performance: since the neighborhood of q (in the domain) is not scattered all over the range, the neighborhood of $h(q)$ in the hash table exhibits good locality of reference. This is counter to Knuth's suggestion (see [18], p. 540) that

> *In a virtual memory enviroment we probably ought to use tree search or digital tree search, instead of creating a large scatter table that requires bringing a new page nearly every time we hash a key.*

Of course, there are many other good ways of retrieving near neighbors in one dimension. However, efficient near-neighbor retrieval is considerably harder, and of growing importance, in higher dimensions. The main application comes from information retrieval: the process of retrieving text and multimedia documents matching a specified query. Other instances of near-neighbor search appear in algorithms for pattern recognition [6, 10], statistics and data analysis [26, 8], machine learning [5], data compression [12], data mining [13] and image analysis [20].

In the case of text retrieval, vector-space methods [3, 28] map each document into a point in high-dimensional space. Sometimes, statistical techniques such as principal components analysis [14], latent semantic indexing [7] or the Karhunen-Loéve/Hotelling transform [16, 22] are used to reduce the dimensionality of the vector space in which the documents are represented, but the number of dimensions could still be very large (say, even 200).

In image and multimedia retrieval, a common first step is to extract a set of numerically-valued features or parameters from the document. For instance, IBM's Query-by-image-content [11] and MIT's Photobook [27] extract image features such as color histograms (hues, intensities), shape descriptors, as well as quantities measuring texture. Once these features have been extracted, each image in the database may now be thought of as a point in this multidimensional feature space (one of the coordinates might, for the sake of a simplistic example, correspond to the overall intensity of red pixels, and so on). Dimension-reduction techniques can again be applied, in some cases getting down to under 25 dimensions.

Both for text and for images (as well as for other multimedia documents), a user-defined query is transformed to the same vector space where the documents are represented. The retrieval question now turns into one of finding near neighbors in this space. Most typically, the query one seeks to answer is of the form "give me the k nearest neighbors, but don't give me anything that's more than distance δ away." In practice k is usually a small number (such as 8), and the bound δ depends on the representation: the intent is to get (say) the nearest 8 images as long as there are at least 8 images "close by." (Users are typically not interested in seeing 8 images if many of them do not match the query very well.) Most current retrieval systems resort to brute-force linear seach, computing the distance from the query to all the points in the database. We know of no system embodying high-dimensional search that has sublinear worst-case query performance.

Generalizing the approach of Linial and Sasson to higher dimensions offers promise as a way of tackling this problem. In this paper we give such a generalized construction of locality-preserving hash functions in higher dimensions, together with negative results suggesting that our construction is theoretically the best possible. In addition to the theoretical value of these results, our construction will in principle afford fast retrieval and good paging behavior during search. Realistically, though, it will in practice only work for modest values of the dimension d (say, 10-20). Nevertheless it offers a simple approach for indexing problems (if the feature space after dimensionality reduction has moderate dimension) and in iterative computations for sparse finite-element relaxation methods. To put this in perspective, we survey some other approaches to near-neighbor finding.

Approaches for Near-Neighbor Search. Samet [29] surveys a variety of data structures used for this problem including variants of k-d trees, R-trees, and structures based on space-filling curves. While some of these structures perform reasonably well in 2 or 3 dimensions (and even admit probabilistic analyses for very simple probability distributions from which the points in the database are drawn), they all exhibit behavior that is poor in the worst case (and usually in typical cases as well).

The field of computational geometry has developed a rich theory for the study of proximity problems. Establishing upper bounds on the time required to answer a nearest-neighbor query in $\Re^d$ appears to have been first undertaken by Dobkin and Lipton [9]; they provided an algorithm with query time $O(2^d \log n)$ and pre-processing $O(n^{2^d})$. This was improved by Clarkson [4]: he gave an algorithm with query time $O(\exp(d) \cdot \log n)$ and pre-processing $O(n^{\lceil d/2 \rceil (1+\varepsilon)})$; here $\exp(d)$ denotes a function that grows at least as quickly as 2^d. The query time was later improved by Meiser [24] to $O(d^5 \log n)$ with pre-processing $O(n^{d+\epsilon})$. Recently, Kleinberg [17] has developed a scheme for *ap-*

proximate nearest neighbor problem that achieves query time $O(d^2 \log n)$ with preprocessing $n^{O(d)}$. There have been a number of other approaches and extensions (e.g. [31, 23, 25, 1, 2]). The best approaches from these studies are still impractical for the values of d encountered in the retrieval applications above.

Overview of Paper. In Section 3 we give a construction for locality-preserving hash functions in two dimensions; this is extended to higher dimensions in Section 4. For d-dimensions, our functions are $\sqrt{d}$-expansive under the l_2 norm, d-expansive under the l_1 norm, and non-expansive under the l_∞ norm. The constants in our guarantees (bucket size, collision probability) grow with the dimension, roughly as $O(c^d)$ where c is a fixed constant for l_2 norm and $O(d^d)$ for l_1 and l_∞. In Section 5 we present several negative results exploring the intrinsic limitations of locality-preserving hash functions. First, we establish a lower bound of $\Omega(1/R)$ on the collision probability for any family of non-expansive functions in $d \geq 2$ dimensions; this implies that obtaining low collision probability is essentially impossible for higher dimensions. Then we restrict ourselves to polynomial sized families of natural subclass of non-expansive hash functions. We show that even if we allow to store up to c elements in each bucket, no such family is able to hash more than roughly $O(R^{1-1/c})$ elements. Both results suggest that relaxation of the non-expansiveness constraint is essential in order to provide good bounds. We conclude in Section 6 with a number of issues which merit further study.

2 Preliminaries

The domain of the hash functions we consider in this paper is a d-dimensional cube $D = \{-U, \ldots, U\}^d$. The range of the functions will be a set of points in a d-dimensional cube I of side R. In this section we present our results for the case $d = 2$; in Section 4 we outline the generalization to higher dimensions. Although we consider only cubes (i.e., with equal-length sides) for the domain, the results can be extended to cuboids with unequal sides. Let $d(p, q)$ be the distance between any points p and q. We will use the notation that $p = (x_p, y_p)$ and $q = (x_q, y_q)$. We define

- $d_x(p, q) = |x_q - x_p|$ and $d_y(p, q) = |y_q - y_p|$,
- $d_r(p, q) = (d_x(p, q)^r + d_y(p, q)^r)^{\frac{1}{r}}$, for any $r \geq 1$.

These definitions extend naturally to $d > 2$ dimensions.

Definition 1 *A function $h : D \rightarrow I$ is* c-expansive *under a distance metric d if for any $p, q \in D$, $d(h(p), h(q)) \leq d(p, q) + c$. If $c = 0$, then h is said to be* non-expansive *under d.*

In our constructions we use the 1-dimensional family of hash functions $\mathcal{G}$ introduced by Linial and Sasson [21]. Their functions are obtained by "folding" the domain D along randomly-chosen turning points in such a way that any segment of D between two consecutive turning points has length $\Theta(R)$ (see [21] for a formal definition). Besides the properties described in the introduction, $\mathcal{G}$ has the property that $\Pr_{g \in \mathcal{G}}(g(x) = g(y)) = O(1/R)$ for any $x \neq y \in D$.

In the sequel, we often omit the subscript of $\Pr_{f \in \mathcal{F}}$ which denotes the probability space from which f is chosen – it should be assumed that f is chosen uniformly at random from the probability space implicit from the context.

3 Hashing in Two Dimensions

The main theorem we prove in this section asserts that we can achieve $O(1)$ bucket size using locality-preserving hash functions in the 2-dimensional setting. To this end, we describe a family of 1-expansive hash functions $\mathcal{H}_2 \subset \{h : D \to I\}$ based on applying random rotations to the 2-dimensional domain cube I.

Formally, we define $\mathcal{H}_2$ to be all functions h of the form $h(p) = t_2(g(r(p)))$, where for $p = (x_p, y_p)$,

- $t_c(p) = (\lfloor x_p/c \rfloor, \lfloor y_p/c \rfloor)$;
- $g(p) = (g_x(x_p), g_y(y_p))$, where $g_x, g_y \in \mathcal{G}$;
- To define $r(p)$ we need some extra notation. Let p_v be the column vector whose coordinates are the coordinates of p. Let $q_v = (x'_p, y'_p) = Mp_v$ where M is a 2×2 *rotation* matrix with column sums equal to 1. Then $r(p) = (\lfloor x'_p \rfloor, \lfloor y'_p \rfloor)$.

The function r rotates the domain cube, and then rounds off the rotated points to the nearest lattice points. It will suffice to restrict the entries of M to the discrete set of multiples of $1/R$ in the range $[0, 1]$. Hence, the size of the family $\mathcal{H}_2$ is equal to $|\mathcal{G}|R^2$.

Theorem 2 *There exist constants C and B such that for any domain size U and and range size R, the functions $h \in \mathcal{H}_2$ are 1-expansive under the d_1 distance metric and have the property that for any $p \in I$ and any $S \subset D$ with $|S| \leq CR$,*

$$\Pr(|h^{-1}(p) \cap S| \leq B) > \frac{1}{2}.$$

The proof of Theorem 2 is obtained using Lemma 1, Lemma 2 and Lemma 4 (proved below).

Lemma 1 *Each $h \in \mathcal{H}_2$ is 1-expansive.*

Proof: It is easy to verify that each function r is 2-expansive, hence so is $g \circ r$. Application of t_2 reduces the expansion to 1. □

To bound the number of collisions, we look at two separate cases: pairs p, q which are close to each other (i.e., $d_\infty(p,q) < R$) and those that are far from each other (i.e., $d_\infty(p,q) \geq R$. The following fact takes care of the first case.

Lemma 2 *There exists a constant B_1 such that for each $p \in D$ and any $h \in \mathcal{H}_2$, the number of $q \in D$ such that $d_\infty(p,q) < R$ and $h(p) = h(q)$ is less than B_1.*

Proof: Follows from the construction of g, and the observation that for any $p \in D$, the number of $q \in D$ such that $r(p) = r(q)$ is bounded by a constant. □

Lemma 3 *Let $p, q \in D$ be such that $d_\infty(p,q) \geq R$, then there is constant A such that $r(p), r(q)$ differ in both coordinates with probability at least $1 - A/R$.*

Proof: Let

$$M = \begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \end{bmatrix},$$

such that $u_1 + u_2 = v_1 + v_2 = 1$. Let p, q be such that $d_\infty(p,q) = |x_p - x_q| \geq R$. Observe that

$$r(p)_{|x} - r(q)_{|x} = (Mp)_{|x} - (Mq)_{|x} = u_1(x_p - x_q) + v_1(y_p - y_q),$$
$$r(p)_{|y} - r(q)_{|y} = (Mp)_{|y} - (Mq)_{|y} = u_2(x_p - x_q) + v_2(y_p - y_q).$$

Then there is some constant A such that with probability at least $(1-A/R)$, both expressions above are at least 2 in magnitude. After the rounding step, the points will remain different on this coordinate. We conclude that with probability at least $1 - A/R$, both coordinates of $r(p)$ and $r(q)$ will be different. □

In order to prove the next lemma, we need the following fact which follows from the analysis due to Linial and Sasson [21].

Fact 1 *There is a constant C_1 such that for any $p, q \in D$ if p and q differ on k coordinates, then*

$$\Pr(t_2(g(p)) = t_2(g(q))) \leq \frac{C_1}{R^k}$$

Lemma 4 *There exists a constant C_2 such that for any $p, q \in D$, $d_\infty(p,q) \geq d$,*

$$\Pr(h(p) = h(q)) \leq \frac{C_2}{R^2}.$$

Proof: We proceed as follows:

$$\begin{aligned}
\Pr(h(p) = h(q)) &\leq \Pr(t_2(g(r(p))) = t_2(g(r(q)))) \\
&\leq \frac{A}{R} \cdot \frac{C_1}{R} + \\
&\quad \Pr(t_2(g(r(p))) = t_2(g(r(q))) \mid \\
&\quad r(p), r(q) \text{ differ in 2 coordinates}) \\
&\leq \frac{C_2}{R^2}.
\end{aligned}$$

$\square$

4 Hashing in Higher Dimensions

In this section we extend the locality preserving hash families to any dimension d. To this end we use more conventional version of rotation matrices, i.e., where the columns of the rotation matrix M are required to be orthonormal. This preserves d_2 distance between points, and allows us to obtain results for locality-preserving hashing under the d_2 distance metric. The results for other metrics then easily follow from the fact that any metric d_r for $r \geq 1$ can be approximated by d_2 with only $\sqrt{d}$-factor distortion. The constants in our guarantees (bucket size, collision probability) grow with the dimension, roughly as $O(c^d)$ (where c is a fixed constant) for a hash family that preserves d_2 distance and as $O(d^d)$ for hash families that preserve other distances.

We define $\mathcal{H}_d$ as the collection of all functions h of the form $h(p) = g(r(p))$, where for $p = (x_1, x_2, \ldots, x_d)$,

- $g(p) = (g_{x_1}(x_1), \ldots, g_{x_d}(x_d))$, where $g_{x_1}, \ldots, g_{x_d} \in \mathcal{G}$;
- We define $q_v = (x'_1, \ldots, x'_d) = Mp_v$ where M is a $d \times d$ rotation matrix with orthonormal columns. Then, $r(p) = (\lfloor x'_1 \rfloor, \ldots, \lfloor y'_1 \rfloor)$.

Theorem 3 *There exist constants C and B such that for any domension d, domain size U, and range size R, the functions $h \in \mathcal{H}_d$ are $\sqrt{d}$-expansive under the d_2 distance metric and have the property that for any $p \in I$ and any $S \subset D$ with $|S| \leq CR^{d/2}$,*

$$\Pr(|h^{-1}(p) \cap S| \leq B^d) > \frac{1}{2}.$$

First, we establish $\sqrt{d}$-expansiveness.

Lemma 5 *Each $h \in \mathcal{H}_d$ is $\sqrt{d}$-expansive.*

Proof: Let $h(p) = g(\lfloor M(p) \rfloor)$. As g and M are non-expansive, it suffices to note that for any $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ from $\Re^d$, $||p_i - q_i| - |\lfloor p_i \rfloor - \lfloor q_i \rfloor|| \leq 1$ for $i = 1, \ldots, d$ and hence rounding can change each coordinate of a difference vector between two points by at most 1. The result then follows from the triangle inequality. $\square$

We now concentrate on bounding the bucket size. To this end, the following fact will be useful.

Fact 2 *Let $\mathcal{B}_d(r)$ be the d-dimensional ball of radius r centered at the origin, and let $\mathcal{S}_d(r)$ be the bounding sphere of $\mathcal{B}_d(r)$. Then, letting $\Gamma(x)$ denote a gamma function,*

- *the volume* $|\mathcal{B}_d(r)|$ *of* $\mathcal{B}_d(r)$ *is equal to* $\dfrac{2r^d}{d}\dfrac{\pi^{d/2}}{\Gamma(d/2)}$,
- *the area* $|\mathcal{S}_d(r)|$ *of* $\mathcal{S}_d(r)$ *is* $\dfrac{2r^{d-1}\pi^{d/2}}{\Gamma(d/2)}$.

The proof again proceeds in two stages. First, we bound the number of collisions caused by elements that are close to each other. Then, we bound the probability that two elements that are far apart are mapped to the same bucket.

Lemma 6 *There exists a constant* B_1 *such that for any* $p \in D$ *and any* $h \in \mathcal{H}_d$, *there are at most* B_1^d *choices of* $q \in D$ *such that* $d_2(p,q) < \sqrt{d}R$ *and* $h(p) = h(q)$.

Proof: The proof consists of two parts. First, we bound the number of points mapped to the same location by r. Then we bound the number of q's as above mapped to the same location as p by g.

For the first part, we need to bound the number of points from d-dimensional unit lattice contained in any (possibly rotated) unit cuboid $\mathcal{C}$. It suffices to bound the number of lattice cuboids intersecting $\mathcal{C}$, as each such cuboid contributes at most 2^d points. To this end, we bound the number of lattice cuboids intersecting a ball (of radius $\sqrt{d}$) circumscribed around $\mathcal{C}$, which again can be bounded by the number of points contained in a ball of radius $2\sqrt{d}$ around $\mathcal{C}$. As the volume of each cuboid is 1, it is sufficient to estimate the volume of a d-dimensional ball of radius $2\sqrt{d}$, which is

$$\frac{2(2\sqrt{d})^d}{d}\frac{\pi^{d/2}}{\Gamma(d/2)} \le \pi^{d/2}\frac{(2\sqrt{d})^d}{(d/2-2)!} \le B_1^d.$$

For the second part, notice that each g splits the domain into disjoint cuboids of side at least $\Omega(R)$ such that no two points from the same cuboid overlap. Hence it is sufficient to bound the number of such cuboids intersecting a d-dimensional ball of radius $\sqrt{d}R$. This can be done by an argument similar to the one above. □

Lemma 7 *Let* $p, q \in D$ *be such that* $d_2(p,q) \ge \sqrt{d}R$. *Then, there exists a constant* A *such that*

$$\Pr(r(p), r(q) \textit{ are equal in } k \textit{ coordinates}) \le \sqrt{d}\binom{d}{k}\left(\frac{A}{R}\right)^k.$$

Proof: Let $l = d_2(p,q)$ and $z = q - p$. It is sufficient to prove that for any $K \subset \{1,\ldots,n\}$, the probability that $R(p)$ and $r(q)$ are equal on K is at most $\sqrt{d}(A/R)^k$. Without loss of generality, we assume $K = \{1,\ldots,k\}$. To complete the proof of this lemma, we will employ the following two facts.

Fact 3 *Let o, p, q be any points in $\Re^d$. Futhermore, M_0 be a matrix of a a random rotation with origin at o, and similarly let M_q be a matrix of a random rotation with origin at q. Then, the distributions of $M_0(q-p)$ and $M_q(q-p)$ are identical.*

Fact 4 *Let M be a random orthonormal matrix. Then for any $x \in \Re^d$, Mx is a random unit vector.*

Proof: Let $M = (u_1, ..., u_d)$ be chosen thus: the first column, u_1, is a random unit vector, i.e., a vector chosen at random from the unit sphere. The ith column, u_i, is chosen at random from the set of unit vectors orthogonal to the first $i-1$ columns.

Consider $x = (1, 0, \ldots, 0) \in \Re^d$. Given the manner we selected the first column of M, the vector Mx is a random unit vector.

Now consider an arbitrary unit vector y. There exists an orthonormal matrix S such that $Sy = x$, in other words, $y = S^T x$. We wish to show that My is a random unit vector. But $My = (MS^T)x$. The orthonormal matrix S^T viewed as a function from the unit sphere to the unit sphere is 1-1 and space-preserving. (Note that for a discrete version of the lemma, e.g., if all coordinates are multiples of $1/R$, the 1-1 property suffices). Hence the matrix MS^T is a random orthonormal matrix, just like M. It follows that $(MS^T)x$ is uniformly distributed on the unit sphere, implying the desired result. □

Now we proceed as follows. From Fact 3 we may assume that the origin of rotation is located in q. By Fact 4 it is sufficient to consider only $z = (l, 0, \ldots, 0)$. Hence, $Mz = lu$, where $u = (u_1, \ldots, u_d)$ is the first column of M. It is now sufficient to show that

$$\begin{aligned} \Pr(|Mz_{|x_i}| \leq 1 \text{ for all } i \in K) &= \Pr(|lu_i| \leq 1 \text{ for all } i \in K) \\ &\leq \sqrt{d}\left(\frac{A}{R}\right)^k. \end{aligned}$$

Let C_d^i be the set $\Re^{i-1} \times [-1, 1] \times \Re^{d-i}$ and let $\phi = \sin^{-1}(1/l)$. The probability $\Pr(|u_i| \leq 1 \text{ for } i \in K)$ is equal to the area of the surface of

$$P_d^K(l) = \mathcal{B}_d(l) \cap \cap_{i \in K} C_d^i$$

divided by $|\mathcal{S}_d(l)|$. The area of the surface of $P_d^K(l)$ is

$$\begin{aligned} |P_d^K(l)| &= \int_{-\phi}^{+\phi} \cdots \int_{-\phi}^{+\phi} |\mathcal{S}_{d-k}(l \cos\phi_1 \ldots \cos\phi_k)| l^k d\phi_1 \ldots d\phi_k \\ &\leq \int_{-\phi}^{+\phi} \cdots \int_{-\phi}^{+\phi} |\mathcal{S}_{d-k}(l)| l^k d\phi_1 \ldots d\phi_k \\ &= \mathcal{S}_{d-k}(l) l^k \phi^k. \end{aligned}$$

As $\sin\phi = \frac{1}{l} \le \frac{1}{2}$, $\phi \le \frac{\pi}{6}$, so $\sin\phi \ge \frac{\phi}{2}$ and $\sin^{-1}(1/l) \le \frac{2}{l}$. Hence, $|P_d^K(l)| \le |\mathcal{S}_{d-k}(l)| \cdot 2^k$ and

$$\begin{aligned}
\Pr(|u_i| \le 1 \text{ for } i \in K) &\le 2^k \frac{|\mathcal{S}_{d-k}(l)|}{|\mathcal{S}_d(l)|} \\
&\le \frac{2^k}{\pi^{k/2} l^k} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-k}{2})} \\
&\le \frac{2^k}{\pi^{k/2}(R\sqrt{d})^k} d^{\lceil k/2 \rceil} \\
&\le \sqrt{d}\left(\frac{2}{R}\right)^k
\end{aligned}$$

□

Lemma 8 *There exists a constant C_2 such that for any $p, q \in D$, with $d_2(p,q) \ge \sqrt{d}R$,*

$$\Pr(h(p) = h(q)) \le \left(\frac{C_2}{R}\right)^d.$$

Proof: We recall Fact 1 and proceed as follows:

$$\begin{aligned}
\Pr(h(p) = h(q)) &\le \Pr(t_2(g(r(p))) = t_2(g(r(q)))) \\
&\le \sum_{k=0}^{d} \Pr(t_2(g(r(p))) = t_2(g(r(q))) \mid \\
&\qquad r(p), r(q) \text{ are equal in } k \text{ coordinates}) \\
&\qquad \times \Pr(r(p), r(q) \text{ are equal in } k \text{ coordinates}) \\
&\le \sum_{k=0}^{d} \sqrt{d}\binom{d}{k}\left(\frac{C_1}{R}\right)^{d-k}\left(\frac{A}{R}\right)^k \\
&\le \left(\frac{C_2}{R}\right)^d.
\end{aligned}$$

□

For any distance metric d_r where $r \ne 2$, we have the following.

Theorem 4 *There exist constants C and B such that for any dimension d, domain size U, and range size R, there exists a family $\mathcal{H}_d^r \subset \{h : D \to I\}$ of $d^{1/r}$-expansive functions (under the d_r distance metric) such that for any $p \in I$ and any $S \subset D$ with $|S| \le CR^{d/2}$,*

$$\Pr(|h^{-1}(p) \cap S| \le (Bd)^d) > \frac{1}{2}.$$

Proof: Take the family as in Theorem 4 and apply t_d to each coordinate. The proof follows from Theorem 4 and the fact that for any points p and q, and any choice of r, the distance metrics $d_2(p,q)$ and $d_r(p,q)$ differ by at most a factor $\sqrt{d}$. $\square$

Remark 1 *By a more detailed analysis we can establish that functions similar to the ones defined above are in fact $d^{1/r}-1$ expansive. This specifically implies existence of hash functions which are non-expansive in d_∞.*

Remark 2 *It may seem natural to try to improve the $O(d^d)$ bound for d_1 and d_∞ by applying a method similar to the one in the proof of Theorem 3. However, this method requires "random rotation" and it is not clear if a similar transformation exists for norms other than l_2. In fact, the notion of rotation in l_p for $p \neq 2$ is not well-defined. The reason is that the rotation is defined as a linear transformation described by a matrix with orthonormal columns. However, it is known that for R^d with l_p norm, where $p \neq 2$, no inner product exists (cf. [19], p. 133), and so orthonormality in such spaces is not well-defined.*

5 Negative Results

Can small bucket size be achieved with non-expansive hash functions? In this section we prove lower bounds that indicate the contrary. First we focus on lower bounding the collision probability for non-expansive functions. Then we show that in fact small bucket size is not possible for a sub-class of non-expansive functions which we call *multifoldings*. It is an open problem as to whether *any* non-expansive function can be viewed as a multifolding. We conjecture that this is indeed the case.

5.1 Lower Bounds on Collision Probability

Theorem 5 *For any family of non-expansive functions $\mathcal{H}$ under the d_1 distance metric, there exists a pair $x, y \in D$ such that* $\Pr(f(x) = f(y)) = \Omega(\frac{1}{R})$.

Proof: We prove the theorem for $d = 2$ ($d \geq 2$ can be handled similarily). Let $h \in \mathcal{H}$. We define *the c-crease set* of h as

$$C_c(h) = \{p \in D \mid \exists_{q \neq p} h(p) = h(q), d_1(p,q) \leq c\}.$$

The following lemma holds.

Lemma 9 *There exists a constant C such that for any $h : D \to I$, $\frac{|C_4(h)|}{|D|} \geq \frac{C}{R}$.*

Proof: Divide D into disjoint *slices*, i.e., rectangles of length $R+2$ and height 3. We show that each slice contains at least one point from $C_4(h)$. Without loss of generality, we examine only the set $[1.\ldots.R+2] \times [1.\ldots.3]$. Let $L = \{(i,2) \mid i = 2.\ldots.R+2\}$ and consider the set $h(L)$. If $h((2,2)) = h((3,2))$, there is nothing to prove. In the opposite case, either $d_x(h(2,2), h((3,2))) = 1$ or $d_y(h(2,2), h((3,2))) = 1$. Assume the first case (the second is symmetric). Then

- either $h(L)$ forms a line, i.e., $d_y(h(p), h(q)) = 0$ for every $p, q \in L$; or,
- there exists $i \geq 3$ such that $d_y(h(i,2), h(i+1,2)) = 1$.

In the first case clearly for every p, q, $d_1(h(p), h(q)) \leq R$ (as R is the width of I), so $|h(L)| \leq R$. As $|L| = R+1$, there are two points $p \neq q$ such that $h(p) = h(q)$ (assume $x_p < x_q$). Consider the sequence $S = h((x_p, 2))_{|x} \ldots h((x_q, 2))_{|x}$. Its first and last elements are equal and (due to the non-expansive property of h) the consecutive elements of S differ by at most 1. It follows that there exist two points $p', q' \in L$ such that $d_1(p', q') \leq 2$ and $h(p') = h(q')$.

For the second case assume that i is the smallest index such that $d_y(h(i,2), h(i+1,2)) = 1$. Then either $d_x(h(i-1,2), h(i,2)) = 0$ (and there is nothing to prove) or $d_x(h(i-1,2), h(i,2)) = 1$. In the latter case, notice that there is a set T of 6 points in D different from $T' = \{(i-1,2), (i,2), (i+1,2)\}$ which are within distance 1 to some point from $T'' = \{(i-1,2), (i+1,2)\ \}$, while there are only 5 points in I different from $h(T')$ within distance 1 to some point from $h(T'')$. Hence at least two points from $T \cup T'$ are mapped to the same location. As for any $p, q \in T \cup T'$ $d_1(p,q) \leq 4$, the lemma follows. □

By a simple counting argument it follows from Lemma 9 that there exists $p \in U$ such that $\Pr(p \in C_4(h)) \geq C/R$. Consider the (at most) 24 points $q \neq p$ s.t. $d_1(p,q) \leq 4$. Clearly, for at least one of them $\Pr(h(p) = h(q)) \geq \frac{C}{24R}$. □

Theorem 6 *For any family of non-expansive functions $\mathcal{H}$ under the d_∞ distance metric, there exists a pair $x, y \in D$ such that* $\Pr(f(x) = f(y)) = \Omega(\frac{1}{R})$.

Proof: It is sufficient to prove Lemma 9 for d_∞ metric. To this end, define $L = \{(i,2) \mid i = 2.\ldots.R+2\}$ as before and consider the following cases:

Case 1: for all $p, q \in L$, $d_y(h(p), h(q)) = 0$, or for all $p, q \in L$, $d_x(h(p), h(q)) = 0$ – we proceed as in the proof of Lemma 9.

Case 2: $d_x(h(2,2), h(3,2)) = d_y(h(2,2), h(3,2)) = 1$. Then there exists a set T of 4 points from D different from $T' = \{(2,2), (3,2)\}$ within distance 1 to both points from T'. However, there are at most 2 points in I different from $h(T')$ within distance 1 to both points in $h(T')$, so at least two points from $T' \cup T$ collide.

Case 3: $d_x(h(2,2), h(3,2)) = 0$ and $d_y(h(2,2), h(3,2)) = 0$; we are done.

Case 4: $d_x(h(2,2),h(3,2)) = 1$ and $d_y(h(2,2),h(3,2)) = 0$. Let i be the smallest index such that $d_y(h(i,2),h(i+1,2)) = 1$ and let $(\Delta x, \Delta y) = h(i+1,2) - h(i,2)$. If $|\Delta x| = 1$, we proceed as in case 1. If $|\Delta x| = 0$, there exist a set T of 4 points from D different from $T' = \{(i,2),(i+1,2)\}$ within distance 1 to both points from T'. However, there are at most 3 points in I different from $h(T')$ within distance 1 to both points in $h(T')$, so at least two points from $T' \cup T$ collide.

Case 5: $d_x(h(2,2),h(3,2)) = 0$ and $d_y(h(2,2),h(3,2)) = 1$ – symmetric to Case 4.

□

Theorem 7 *There exists a set S of size $3(R+2)$ such that no non-expansive (under any distance metric) function $h : D \to I$ is one-to-one on S.*

Proof: Follows from the proof of Lemma 9. □

5.2 Lower Bounds on Multifoldings

We need some further notation. With any $s = (s_1, \ldots, s_d) \in \Re^d$ and $t \in \Re$ we associate a *hyperplane*

$$L_{(s,t)} = \{q \in \Re^d \mid s \cdot q = t\}.$$

In the sequel, we use both (s,t) and $L_{(s,t)}$ to denote a line; when convenient, we will think of the line as lying in D instead of $\Re^d$. We refer to s as the *slope* of $L_{(s,t)}$). For any line $L_{(s,t)}$ we also define its positive side

$$L^+_{(s,t)} = \{q \in \Re^d \mid s \cdot q \geq t\}$$

and negative side

$$L^-_{(s,t)} = \{q \in \Re^d \mid s \cdot q < t\}.$$

A *real folding along line L* is a function $r_L : \Re^d \to \Re^d$ such that

$$r_L(p) = \begin{cases} p & \text{if } p \in L^+ \\ \text{a reflection of } p \text{ about } L & \text{otherwise} \end{cases}$$

A *folding along L* is a function $f_L(p) = (\lfloor r_L(p)_{|x} \rfloor, \lfloor r_L(p)_{|y} \rfloor)$. A folding along any line crossing a point in $\mathcal{Z}^d$ with a slope from the set $B = \{-1,0,1\}^d - \{0,0\}$ is called *simple*. A *multifolding* is any composition of foldings and translations by vectors in $\mathcal{Z}^d$. We define a *real multifolding* analogously. A *simple multifolding* is any composition of simple foldings and translations.

The following theorem says that hash function families consisting of simple multifoldings have large bucket sizes even when hashing small sets.

Theorem 8 *For any c and d there exists c_1 such that for any R, and any family $\mathcal{H}$ of simple multifoldings $h : D \to I$, there exists a set S of $c_1 R^{1-\frac{1}{c}} \log |\mathcal{H}|$ points such that for any $h \in \mathcal{H}$ there exists $q \in I$ such that $|h^{-1}(q) \cap S| \geq c$.*

Proof: We prove the theorem for $d = 2$ ($d \geq 2$ can be handled similarily). We assume $U = kR$ for some constant k dependent on c. The following lemma will be helpful.

Lemma 10 *Let $h : X \to Y$ be a function such that for at least a fraction a_1 of $x \in X$ we have $|h^{-1}(h(x))| \geq c$, for some constants a_1 and c. Then there exists c_1 such that for any $\epsilon > 0$ for a random $S \subset X$ of cardinality $c_1 |X|^{1-\frac{1}{c}} \log \frac{1}{\epsilon}$*

$$\Pr(\textit{for some } y \in Y, |h^{-1}(y) \cap S| \geq c) > 1 - \epsilon$$

Proof: Let $m = |X|$. Let $A_1, \ldots, A_l$ be a family of disjoint c-subsets of A, such that $|h(A_i)| = 1$ holds for each $i = 1, \ldots, l$. Clearly, we can ensure $l \geq a_1 |X|/2c$. Now assume that we sample t elements of X independently and uniformly at random *without replacement* (clearly, sampling with replacement can only improve our bounds). It is now sufficient to show that if $t \geq c_1 m^{1-\frac{1}{c}}$, then there is a constant probability of having bucket size at least c (the $\log 1/\epsilon$ factor follows then from the fact that we can repeat perform sampling several times). To this end, first notice that there is a constant probability that at least $t' = a_1 t/3$ sampled elements belong to $\cup A_i$, so assume we sample t' elements from this set. For any A_i, if at least c sampled elements belong to A_i, there is a constant probability that they hit all elements of A_i. To complete the proof, notice that sampling $\Theta(l^{1-\frac{1}{c}})$ elements from the set of size l with a constant probability results in some element being sampled at least c times (this is a generalized version of the well-known *birthday problem* for $c > 2$). $\square$

For any slope s we define $M_{s,\Delta} = \{L_{(s,i\cdot\Delta)} \mid i \in \mathcal{Z}\}$. We define the set S to be a random subset of a set

$$X_R = D \cap (\cup_{s \in B} M_{s,R})$$

In other words, X_R consists of subintervals of equally spaced horizontal/vertical/diagonal lines, such that the consecutive lines of the same slope are within distance $O(R)$ from each other. We show that for any multifolding h the set X_R satisfies the condition for the set X of Lemma 10. To this end, we need a following

Lemma 11 *There exist lines T_s for $s \in B$ (called* thresholds*) and rectangle shaped regions $A_s \subset D$ for $s \in B$ of side bU, for some constant $b > 0$, such that:*

- *$(0,0) \in T_s^+$ for all $s \in B$,*
- *$A_s \subset T_s^-$ for all $s \in B$,*

- $A_s \subset T^+_{s'}$ *for all* $s, s' \in B$ *such that* $s \neq s'$,
- $f_{T_s}(A_s) \subset T^+_{-s}$, *for any* $s \in B$ *(i.e.,* A_s *after mapping along* T_s *cannot exceed* T_{-s}*),*
- *for any* $s \in B$*, the distance between* T_s *and* T_{-s} *is at least* $\frac{U}{2}$.

Proof: Refer to Figure 1. □

Figure 1: D is depicted as a square, thresholds are dotted.

Without loss of generality, we may assume that h is exclusively composed of foldings. Let $h = h_l \circ h_2 \cdots \circ h_1$. Also let L_i be the line defining folding h_i and s_i be a slope of L_i. We may assume that $(0,0) \in L^+_i$ and that for any $i, j = 1, \ldots, l$ such that $s_i = s_j$ we have $L^+_j \subset L^+_i$ (in other words, lines of the same slope are monotonic). We say that a folding h_i is *large*, if the distance between L_i and the last line L_j of the same slope as L_i is greater than aU, for some constant $b > a > 0$ to be defined later. Otherwise, we call the folding *small*.

Lemma 12 *For any* $t = 1, \ldots, l$*, if no folding* h_i *from the set* $h_1, \ldots, h_t$ *"exceeds" its threshold* T_{s_i} *(i.e., it is not the case that* $L^+_i \subset T^+_{s_i}$*), then* $h_1(\ldots h_t(A_s)) = A_s$ *for any* $s \in B$ *(i.e., sets* A_s *remain unchanged by* $h_1, \ldots, h_t$*).*

The lower bound strategy may be now described as follows. We proceed in phases. In the first phase we apply the consecutive foldings $h_1, h_2, \ldots$. We stop the phase at h_t if one of the following happens:

Case 1: h_t is large; or,

Case 2: h_t exceeds its threshold and no h_i for $i \leq t$ is large.

Clearly, one of these two cases has to happen, as some h_t has to exceed its threshold (recall that D is mapped to a square of side R). We proceed as follows. Let $s = s_t$.

Case 1: In this case we know that both width and length of the rectangular shaped region $A = A_s \cap L_t^-$ is more than $aU/2$. This region is mapped by h_t to a rectangular region $A' \subset L_t^+$ of the same size. As a result of this mapping, the intervals from $M \cap A$ belonging to lines of slope s are mapped to the corresponding intervals in $M \cap A'$. We can now restrict our attention only to some square region $D' \subset A'$, i.e., apply the above adversary strategy assuming $D := D'$, starting the next phase of the adversary strategy. Clearly, if this recursion step is (at some moment t') applied more than $4c$ times, then for some slope and for any point p from intervals of this slope belonging to the final square D'' we have $|(h_t \circ \cdots \circ h_1)^{-1}(p)| \geq c$. The constants k and b can be easily chosen such that the side of D'' is still $> 2R$. Hence the number of points p as above is at least R.

Case 2: Let t be the folding which exceeds T_s and for simplicity assume the first phase. As we know that all foldings (including the ones along the slope s) were small, it implies that all intervals from A_s (of length at least bU) were mapped to the corresponding intervals of length at most aU. By the pigeonhole principle, there have to be at least $(1 - ac/b)U$ points of those intervals mapped to locations to which more than c elements were mapped. By choosing b large enough with respect to a, we can ensure that at least R points are mapped to buckets of size at most c.

The above analysis shows that for any $h \in \mathcal{H}$ the fraction of points from X_R mapped to buckets of size at least c is constant. We can apply Lemma 10 with $\epsilon = 1/2|\mathcal{H}|$ to show that a random subset S of X_R of cardinality $c_1 R^{1-1/c}$ contains such a point with probability ϵ. Hence the probability that such a point from S exists for each $h \in \mathcal{H}$ is at least $1/2$, which proves that a set S with such a property exists. □

Remark 3 *The assumption that all slopes of foldings belong to B is not crucial. A slight modification of the above proof shows that the theorem holds for any set of slopes, provided the absolute difference of angles between any two slopes is constant. Also, if $c = 2$, the theorem holds without this assumption. Finally, a set of size $O(R)$ can be constructed for any constant number of slopes, for any value of c.*

Definition 2 *A t-multifolding is any function $h : \mathcal{Z}^d \to \mathcal{Z}^d$ such that there exists a real multifolding f such that for any $p, q \in \mathcal{Z}^d$, $d(h(p), h(q)) \leq d(f(p), f(q)) + t$.*

Remark 4 *Theorem 8 holds when "multifoldings" are replaced by "t-multifoldings" (the constant c_1 depends then on both c and t). The proof needs only minor modifications – lines are made "thicker."*

We conclude this section with the following conjecture.

Conjecture 1 *Every non-expansive hash function $h : D \to I$ is a multifolding.*

6 Summary and Further directions

In this paper we have given a locality-preserving hashing scheme with for set size $\Theta(R^{d/2})$, while showing a bound of $O(R^{1-\epsilon})$ for non-expansive hashing schemes based on multifoldings. These bounds complement each other (modulo our conjecture that multifoldings are comprehensive). We also give a non-expansive hash family with collision probability $1/R$ (for d_∞), while showing a matching lower bound for collision probability for non-expansive hashing in d_∞.

We now mention some extensions and further directions. If the dilation is allowed to be multiplicative (not additive), then one can modify the family described above by simulating each bucket storing E elements by E buckets, each storing at most one element. In this way we can obtain *perfect* family of hash functions with constant multiplicative and $\sqrt{d}$ additive expansion terms. Can one get small bucket size with $O(R)$ elements (rather than $O(R^{1/2})$)? It is quite easy (by adopting the hashing scheme of [21]) to store $O(R)$ elements in $O(R)$ buckets such that each element can be stored in at most $O(\log\log R + \log d)$ buckets and each bucket has $O(1)^d \cdot O(\log\log R)$ size. Can it be achieved with a constant depending only on d? Finally, an important open problem is to improve the upper bound on the bucket size, especially the $O(d^d)$ bound for d_1.

The other questions concern lower bounds. Can we prove the folding theorem for arbitrary slopes for $c > 2$? Can we prove that any non-expansive (in the Manhattan or Euclidean metric) function can be represented by simple folding functions (or general foldings)? Can we show that exponential bucket size is inevitable for hash functions with small additive dilation ?

Acknowledgments

The first author would like to thank Moses Charikar and Suresh Venkatasubramanian for helpful discussions.

References

[1] P.K. Agarwal and J. Matousek, "Ray shooting and parametric search," *Proc. 24th STOC* (1992), 517 – 526.

[2] S. Arya, D. M. Mount, N.S. Netanyahu, R. Silverman, A. Wu, "An optimal algorithm for approximate nearest neighbor searching," *Proc. 5th SODA* (1994), pp. 573 – 582.

[3] C. Buckley, A. Singhal, M. Mitra, and G. Salton, New Retrieval Approaches Using SMART: TREC 4. *Proc. Fourth Text Retrieval Conference*, National Institute of Standards and Technology, 1995.

[4] K. Clarkson, "A randomized algorithm for closest-point queries," *SIAM J. Computing*, 17 (1988), pp. 830–847.

[5] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Machine Learning*, 10 (1993), pp. 57–67.

[6] T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, 13 (1967), pp. 21–27.

[7] S. Deerwester, S. T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, 41 (1990), pp. 391–407.

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

[9] D. Dobkin and R. Lipton, "Multidimensional search problems," *SIAM J. Computing*, 5 (1976), pp. 181–186.

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

[11] C. Faloutsos, R. Barber, M. Flickner, W. Niblack, D. Petkovic and W. Equitz, "Efficient and effective querying by image content", *Journal of Intelligent Information Systems*, 3 (1994), pp. 231–262.

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

[13] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *First International Conference on Knowledge Discovery and Data Mining*, 1995.

[14] H. Hotelling, "Analysis of a complex of statistical variables into principal components", *Journal of educational psychology*, 27 (1933), pp. 417–441.

[15] *IEEE Computer Special Issue on Content-based Image Retrieval Systems*, 28 (1995).

[16] K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae*, Ser. A137, 1947.

[17] J. Kleinberg, "Two algorithms for nearest-neighbor search in high dimensions", these proceedings.

[18] D. Knuth, *The Art of Computer Programming, vol. 3, Sorting and Searching*, Addison Wesley, 1973.

[19] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, 1989.

[20] V. Koivune and S. Kassam, "Nearest neighbor filters for multivariate data," *IEEE Workshop on Nonlinear Signal and Image Processing*, 1995.

[21] N. Linial and O. Sasson, Non-Expansive Hashing, In *Proc. 28th STOC* (1996), pp. 509–517.

[22] M. Loéve. Fonctions aleastoires de second ordere. *Processus Stochastiques et mouvement Brownian.* Hermann, Paris, 1948.

[23] J. Matousek, "Reporting points in halfspaces," *Proc. 32nd FOCS*, 1991.

[24] S. Meiser, "Point location in arrangements of hyperplanes," *Information and Computation* (1993), 106, 2, pp. 286-303.

[25] K. Mulmuley, "Randomized multi-dimensional search trees: further results in dynamic sampling," *Proc. 32nd FOCS*, 1991.

[26] Panel on Discriminant Analysis and Clustering, National Research Council, *Discriminant Analysis and Clustering*, National Academy Press, 1988.

[27] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: tools for content-based manipulation of image databases", In *Proc. SPIE Conference on Storage and Retrieval of Image and Video Databases II*, 2185, 1994.

[28] G. Salton, *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1989.

[29] H. Samet, *The Design and Analysis of Spatial Data Structures*, Addison-Wesley, Reading, MA, 1989.

[30] A.W.M. Smeulders and R. Jain, editors, *Image Databases and Multi-media Search*, Proceedings of the First International Workshop, IDB-MMS '96, Amsterdam. Amsterdam University Press, 1996.

[31] A.C. Yao and F.F. Yao, "A general approach to d-dimensional geometric queries," *Proc. 17th STOC* (1985), pp. 163 – 168.