LSH-Preserving Functions and Their Applications

Flavio Chierichetti* Ravi Kumar†

Abstract

Locality sensitive hashing (LSH) is a key algorithmic tool that is widely used both in theory and practice. An important goal in the study of LSH is to understand which similarity functions admit an LSH, i.e., are LSHable. In this paper we focus on the class of transformations such that given any similarity that is LSHable, the transformed similarity will continue to be LSHable. We show a tight characterization of all such LSH-preserving transformations: they are precisely the probability generating functions, up to scaling.

As a concrete application of this result, we study which set similarity measures are LSHable. We obtain a complete characterization of similarity measures between two sets A and B that are ratios of two linear functions of $|A \cap B|, |A \triangle B|, |A \cup B|$: such a measure is LSHable if and only if its corresponding distance is a metric. This result generalizes the well-known LSH for the Jaccard set similarity, namely, the minwise-independent permutations, and obtains LSHs for many set similarity measures that are used in practice. Using our main result, we obtain a similar characterization for set similarities involving radicals.

*Work done in part while visiting Yahoo! Research. Supported in part by NSF grant CCF-0910940. Department of Computer Science, Cornell University, Ithaca, NY, 14853. Email: flavio@cs.cornell.edu

†Yahoo! Research, Sunnyvale, CA 94089. Email: ravikumar@yahoo-inc.com

1 Introduction

Locality sensitive hashing (LSH) is a key algorithmic tool that has profound connections to the theory of embeddings and has also found a myriad of applications in large-scale data processing, including data stream and sketching algorithms. The core idea behind LSH [16,17] is the following: construct a family of functions that hash objects into values such that objects that are similar will be hashed to the same value with high probability. The hash function family will be determined by the type of the objects and the notion of similarity between them.

A fundamental quest in this topic is to fully understand which similarity functions admit an LSH and to enlarge the scope of functions for which an LSH can be constructed. Charikar [11] offered a partial answer to the first problem: he showed that if a similarity function S admits an LSH, then $1 - S$ must be a metric.He also showed a close relationship between LSH and rounding algorithms and used this to obtain an LSH for cosine similarity in the Euclidean space and an LSH for the earth-mover distance between strings. In general, ℓ_p spaces are well understood from an LSH point of view: for example, Datar et al. [12] used p-stable distributions to obtain an LSH for p-norms in Euclidean spaces. See also the survey by Andoni and Indyk [3] and Andoni's thesis [2]. The condition that $1-S$ should be ℓ_1-embeddable has been essentially the only necessary condition for a similarity to have an LSH. In other words, very little is known in general on what makes it possible for a similarity to admit an LSH.

Jaccard coefficient — the ratio of the size of the intersection of two sets to the size of their union — has been widely studied in the context of LSH, primarily due to its applications in information retrieval and thanks to the work of Broder et al. [8–10]; minwise-independent permutations form an LSH for the Jaccard coefficient. In contrast, the large family of set similarities of which Jaccard coefficient is a very special case, has been studied far less from an LSH viewpoint. See Table 1 for a sample of such measures: they find extensive applications in areas like biology, ecology, software diagnosis, etc. [19].[1] In fact, there is a well-developed theory of similarity measures between two sets [15, 18], where the measure is syntactically a ratio of linear functions of the three set-theoretic measures $(\cap, \cup, \triangle)$ associated with two sets; we call these measures *rational set similarities*. The conditions under which a rational set similarity S yields a metric $1 - S$ are known (see [18]).

Main results. In this paper we study the following question: what is the class of transformations $f(\cdot)$ such that given any similarity S that admits an LSH, the transformed similarity $f(S)$ will continue to admit an LSH? In other words, we seek to study the class of *LSH-preserving* functions. Our main technical result is a surprisingly simple but complete characterization: a function is LSH-preserving if and only if it is a probability generating function (PGF), up to scaling. Our characterization yields conditions necessary for the existence of an LSH that go well beyond the current state of knowledge.

We observe that studying functions that preserve some properties of matrices is an idea that goes back at least to the classical work of Schoenberg [20, 21] and Assouad [5, 6], who studied functions that preserve ℓ_1 and ℓ_2 embeddability of distance matrices.

To show that scaled PGFs are LSH-preserving (Theorem 10), we develop simple tools for building new LSHs out of existing ones. These compositional tools allow us to obtain sufficient conditions under which functions of LSHable similarities are themselves LSHable. To show that LSH-preserving functions are scaled PGFs (Theorem 22), we exploit the negativity of a carefully chosen forward difference operator of a function that is not a scaled PGF and use linear duality in order to show the non-existence of an LSH.

[1] Many of them have their own pages in Wikipedia.

As a concrete application of this result, we study which of the set similarities have an LSH, i.e., are *LSHable*. We show that every rational set similarity S such that $1 - S$ is a metric is also LSHable (Corollary 28), thereby fully characterizing all rational set similarities in terms of their LSHability. The main technique is a delicate use of the compositional tools in order to construct the LSH; this algorithm can be viewed a novel generalization of the minwise-independent permutation for Jaccard similarity (Theorem 26). We also show how to approximate the sketches obtained from our LSH so that they can be used in practice (Section 5.2).

We then turn to similarities S such that $(1 - S)^\alpha$ is a metric for some $\alpha \in (0, 1]$. This case is much more involved and we obtain a similar result as before, but by showing that a particular function is a PGF, and hence LSHable, only for certain ranges of parameters (Theorem 31); showing this requires powerful analytic machinery (Section 4.2.1).

It is important to note that the ℓ_1-non-embeddability approach cannot be used to prove those non-LSHability results, because some of the non-LSHable root similarities are in fact ℓ_2-embeddable and therefore can be embedded in every ℓ_p metric, $1 \leq p \leq \infty$. To get a better understanding of the relationship between ℓ_p-embeddability and LSHability, we show that there exist similarities S admitting an LSH for which $1 - S$ is not ℓ_p-embeddable, for each $1 \lneq p \lneq \infty$. These results are discussed in Section 5.1.

2 Preliminaries

Let U be a (finite) universe of objects. A symmetric function $S : U^2 \to [0, 1]$ such that $S(A, A) = 1$ is called a *similarity*. We first define what it means for a similarity to admit a locality sensitive hash (LSH).

Definition 1 (LSHable similarity)**.** *An* LSH *for a similarity function S is a probability distribution over a set $\mathcal{H}$ of functions defined on U such that $\Pr_{h \in \mathcal{H}}[h(A) = h(B)] = S(A, B)$. A similarity S is* LSHable *if there is an LSH for S.*

A result of Charikar [11] provides a necessary condition for a similarity to be LSHable.

Lemma 2 (Charikar [11])**.** *If a similarity S is LSHable, then $1 - S$ is a metric.*

This condition is typically very useful for understanding whether a similarity is LSHable or not. (There is another necessary condition that subsumes Lemma 2 and which relates LSH to embeddability; we discuss it in Section 5.1.)

Next, we provide a tight characterization of functions on similarities that preserve LSHability.

3 LSH-preserving functions

We first define what it means for a univariate function to preserve the LSHability of a similarity.

Definition 3 (LSH-preserving function)**.** *A function $f : [0, 1) \to [0, 1]$ is said to be* LSH-preserving *if $f(S)$ is an LSHable similarity whenever S is an LSHable similarity.*

An alternate definition would be to let the domain of f be $[0, 1]$ but insist $f(1) = 1$; these definitions are equivalent.[2] We next recall the notion of a probability generating function.

[2]This holds because, in any LSHable similarity, each object must have similarity 1 with itself; that is, both $S(A, A) = 1$ and $(f(S))(A, A) = 1$ have to hold if S and $f(S)$ are LSHable similarities.

Definition 4 (Probability generating function (PGF))**.** *A function $f(x)$ is a* probability generating function (PGF) *if there is a probability distribution $(p_i)_{i=0}^{\infty}$ such that $f(x) = \sum_{i=0}^{\infty} p_i x^i$ for $x \in [0,1]$.*

PGFs are closed under convex combination and composition. We now state the main technical result of the paper, which will follow from Theorem 10 and Theorem 22.

Theorem 5. *A function $f : [0,1) \to [0,1]$ is LSH-preserving if and only there are a PGF $p(x)$ and a constant $\alpha \in [0,1]$ such that $f(x) = \alpha p(x)$.*

A theorem for multivariate functions is provided in Section 5.3.

3.1 Sufficiency

In this section we show that PGFs are LSH-preserving. In doing this, we develop simple ways of obtaining LSHs for new similarity functions using similarities that are known to be LSHable.

Observation 6. *The following similarities are LSHable: (i) $T_0(A,B) = 0$ for each $A \neq B$ and (ii) $T_1(A,B) = 1$ for each $A \neq B$.*

Proof. Both similarities can be easily sketched with a single hash function: for T_0 we use the identity function $h_0(A) = A$ and for T_1 we use the constant function $h_1(A) = 1$. □

Lemma 7. *If the similarities $S_0, S_1, \ldots$ are LSHable and if $(p_i)_{i=0}^{\infty}$ is a probability distribution, then the average similarity $T = \sum_{i=0}^{\infty} p_i S_i$ is also LSHable.*

Proof. To obtain a hash function h for the similarity T, we will first sample $X \in \mathbb{Z}^{\geq 0}$ according to $(p_i)_{i=0}^{\infty}$, and then a hash function h_X from the S_X's LSH. The hash function h is given by $h(A) = (X, h_X(A))$. Then,

$$\Pr_h [h(A) = h(B)] = \sum_{i=0}^{\infty} \left(p_i \Pr_{h_i} [h_i(A) = h_i(B)] \right) = \sum_{i=0}^{\infty} p_i S_i(A,B). \qquad \square$$

Lemma 8. *If the similarities S and T are LSHable, then the similarity $S \cdot T$ is also LSHable.*

Proof. Let h_S be a hash function chosen according to S's LSH and let h_T be a hash function independently chosen according to T's LSH. Consider the composite hash function $h = (h_S, h_T)$:

$$\begin{aligned} \Pr_h [h(A) = h(B)] &= \Pr_h [h_S(A) = h_S(B) \wedge h_T(A) = h_T(B)] \\ &= \Pr_{h_S} [h_S(A) = h_S(B)] \Pr_{h_T} [h_T(A) = h_T(B)] \quad = \quad S(A,B)T(A,B). \end{aligned} \qquad \square$$

Corollary 9. *If the similarity S is LSHable, then the similarity S^i, for $i \in \mathbb{Z}^{\geq 0}$, is also LSHable.*

Proof. If $i = 0$, let S^0 be the similarity T_1 of Observation 6. If $i \geq 1$, apply Lemma 8 to S^{i-1} and S. □

Theorem 10. *If $f(x)$ is a PGF and the similarity S is LSHable, then the similarity $f(S)$ is also LSHable. Furthermore, if $\alpha \in [0,1]$, then the similarity $\alpha f(S)$ is also LSHable.*

Proof. Let $f(x) = \sum_{i=0}^{\infty} p_i x^i$, where $\{p_i\}_{i=0}^{\infty}$ is a probability distribution over the non-negative integers. Since, $f(S) = \sum_{i=0}^{\infty} p_i S^i$ is the arithmetic mean of non-negative integer powers of S, we can apply Lemma 7 and Corollary 9 to get the first assertion. The second can be obtained by taking the $(\alpha, 1-\alpha)$-weighted average of similarity $f(S)$ and the similarity T_0 of Observation 6. □

In general, given a power series $p(x)$ with non-negative real coefficients and $p(1) < \infty$, it holds that $\frac{p(x)}{p(1)}$ is a PGF. Therefore, for each such $p(x)$ and for each LSHable similarity S, we have that $\frac{p(S)}{p(1)}$ is also LSHable; examples of such functions $p(x)$ are

$$\tan x,\ \sinh x,\ \cosh x,\ \frac{1}{\cos x} = \sec x,\ \frac{x}{\sin x} = x \csc x,\ \arcsin x$$

and

$$-W_0\left(-\frac{x}{e}\right),\ i \cdot \operatorname{erf}(-ix)$$

where W_0 is the main branch of the Lambert W function, erf is the Gauss error function, and $i = \sqrt{-1}$. There are other examples that involve the root, exponential and polylogarithm functions:

- $r_\alpha(S) = 1 - (1-S)^\alpha = \sum_{i=1}^{\infty} \left(\frac{(-1)^{i+1}\Gamma(\alpha+1)}{i!\,\Gamma(\alpha-i+1)} S^i \right)$, for each $\alpha \in (0,1)$ and, of course, $r_1(S) = S$,
- $e_\alpha(S) = \frac{1}{\alpha}\alpha^S = \sum_{i=0}^{\infty} \frac{(S \ln \alpha)^i}{\alpha i!}$, for each $\alpha > 1$,
- $\ell_\alpha(S) = \frac{1}{\zeta(\alpha)} \operatorname{Li}_\alpha(S) = \sum_{i=1}^{\infty} \left(\frac{S^i}{\zeta(\alpha)\, i^\alpha} \right)$, for each $\alpha > 1$ (where $\zeta(\alpha)$ denotes the Riemann Zeta function, $\zeta(\alpha) = \sum_{i=1}^{\infty} i^{-\alpha}$).

We now state an important application of Theorem 10 that involves the use of the geometric series; we will use it for building an LSH for rational set similarities.

Lemma 11. *If S is LSHable, then for each $w \geq 1$, $\frac{S}{S+w(1-S)}$ is also LSHable. Specifically, if $S(A,B) = \frac{f(A,B)}{f(A,B)+g(A,B)}$ is LSHable, then $\frac{f(A,B)}{f(A,B)+wg(A,B)}$ is also LSHable.*

Proof. Apply Theorem 10 to S and to the function $f(x) = \frac{x}{w-(w-1)x}$. It holds that $f(1) = 1$ and that f is equal to its power series expansion in $[0,1]$: $f(x) = \sum_{i=1}^{\infty} \left(w^{-1} \left(1 - w^{-1}\right)^{i-1} x^i \right)$. Since the power series has only non-negative coefficients and since they sum up to 1, f is a PGF and hence $f(S)$ is LSHable. Observe also that $f(S) = \frac{S}{w-(w-1)S} = \frac{S}{S+w(1-S)}$. The second part follows easily. □

Finally, Theorem 29 will provide a more general PGF than the one of Lemma 11; we will use it for creating an LSH for root similarities.

3.2 Necessity

We now show that if f is not a PGF, then f does not preserve LSHability. To prove this, we focus on a carefully chosen higher-order derivative of f and use it to construct a family of LSHable similarities. We then appeal to LP duality to show that f cannot be LSH-preserving when applied to these similarities. We first state some basic definitions and facts that we will use in the proof.

Definition 12 (Forward difference)**.** *The k*th forward difference *of function f at x with step h is*

$$\Delta_h^k(f,x) = \sum_{i=0}^{k} \left((-1)^{k-i} \binom{k}{i} f\left(x + ih\right) \right).$$

Definition 13 (Absolutely monotonic function)**.** *A function $f(x)$ is* absolutely monotonic *in $[a,b)$ if and only if $\Delta_h^k(f,x) \geq 0$ for each $x \in [a,b)$, for each $k \in \mathbb{Z}^{\geq 0}$, and for each $h > 0$ such that $x + kh < b$.*

We will make use of Bernstein's theorem on absolutely monotonic functions [7] (see [22, Chapter IV] for an English proof).

Theorem 14 (Bernstein's Theorem [7]). *If $f : \mathbb{R} \to \mathbb{R}$ is absolutely monotonic in $[a, b)$, then for each $x \in [a, b)$,*

$$f(x) = \sum_{i=0}^{\infty} \left(f_+^{(i)}(a) \frac{(x-a)^i}{i!} \right),$$

where $f_+^{(i)}(a) \geq 0$ is the ith right-hand derivative of f at a, which exists and is non-negative for each $i \geq 0$.

Finally, we need Farkas' Lemma from LP duality.

Lemma 15 (Farkas' Lemma). *The system $Ax = b$ has a solution $x \geq \mathbf{0}$ if and only if $yA \geq \mathbf{0} \Rightarrow yb \geq 0$.*

We begin the proof by defining a family of special similarities that will be used in the arguments.

Definition 16 (Intersection similarity). *Let $k, n \in \mathbb{Z}^+$ and $x, h \in \mathbb{R}$ be such that $0 \leq x < x+kh \leq 1$. The* Intersection similarity $H = H_{k,n,x,h} : V^2 \to [0, 1]$ *is given by*

$$V = V_{k,n} = \{X_{A,i} \mid A \subseteq [k] \wedge i \in [n]\},$$

and

$$H\left(X_{A,i}, X_{B,j}\right) = \begin{cases} x + |A \cap B|\, h & \text{if } A \neq B \vee i \neq j \\ 1 & \text{otherwise.} \end{cases}$$

Let k, n, x, h satisfy the conditions in Definition 16. We show that the intersection similarity is LSHable.

Lemma 17. *The intersection similarity $H = H_{k,n,x,h}$ is LSHable.*

Proof. We use $k + 2$ hash functions, one for each $j \in \{0, 1, \ldots, k + 1\}$. The hash function h_0^* will be a constant function, chosen with probability x. The hash function h_{k+1}^* will be the identity function, chosen with probability $1 - x - kh$. The hash function h_j^* for $1 \leq j \leq k$ is defined as

$$h_j^*(X_{A,i}) = \begin{cases} 1 & \text{if } j \in A \\ X_{A,i} & \text{otherwise,} \end{cases}$$

and will be chosen with probability h. An easy calculation shows this to be an LSH for the similarity $H = H_{k,n,x,h}$. □

We now prove an important structural property of forward differences for LSH-preserving functions.

Theorem 18. *If f is LSH-preserving then (i) $f(x) \in [0, 1]$ for each $x \in [0, 1)$ and (ii) for each $k \geq 1$, for each $0 \leq x < 1$, and for each $0 < h \leq \frac{1-x}{k}$, it holds that $\Delta_h^k(f, x) \geq 0$.*

Proof. (i) is trivial. Indeed, suppose there is an $x^* \in [0, 1)$ such that $f(x^*) < 0$ (resp., $f(x^*) > 1$). Then consider the similarity function S such that $S(A, B) = x^*$ for all $A \neq B$. Easily, S is LSHable by using the constant function with probability x^* and the identity function with probability $1-x^*$. On the other hand, $f(S)$ is not LSHable since $f(S(A, B)) < 0$ (resp., $f(S(A, B)) > 1$).

For (ii), we assume by contradiction that there exists k, x, h such that $0 > -\delta = \Delta_h^k(f, x)$. We consider the intersection similarity $H = H_{k,n,x,h}$ with $n = \left\lfloor (t+1) \cdot \frac{2^k}{\delta} \right\rfloor + 1$, and $|V| = n2^k$; t will be any positive integer. By Lemma 17, H is LSHable. Let the ℓ_1 distance between similarities F and $f(H)$ be defined as

$$\ell_1(F, f(H)) = \sum_{\{v,v'\} \in \binom{V}{2}} \left| F(v, v') - f(H(v, v')) \right|.$$

To complete the proof, we will show something stronger: no similarity F for which $\ell_1(F, f(H)) \leq n\, t\, 2^{k-1} = \Theta\left(t \cdot |V|\right) = \Theta\left(\delta \cdot 4^{-k} \cdot |V|^2\right)$ is LSHable.

Let $N = |\mathcal{H}|$ and let $\mathcal{H} = \{h_1, \ldots, h_N\}$ be an arbitrary enumeration of a maximal class of non-isomorphic hash functions of V. Consider the matrix M whose rows are indexed by the elements in $\binom{V}{2} \cup \{\star\}$ and columns are indexed by the elements in $\mathcal{H}$; clearly, M has $\binom{2^k n}{2} + 1$ rows and N columns. The $\star$ row of M will be all-ones, and the entry $M(\{v, v'\}, h_i) = 1$ if $h_i(v) = h_i(v')$ and 0 otherwise.

Given an arbitrary similarity vector s indexed by unordered pairs of elements in V, we augment s by adding a new element 1 indexed by $\star$ to obtain $s^\star$. Then, it is easy to see that the similarity vector s is LSHable if and only if there exists a non-negative solution p to the system $Mp = s^\star$. Indeed, any non-negative solution to the system gives a natural LSH for s (select h_i with probability $p(h_i)$) and conversely, from any LSH, one can get a valid solution p by setting $p(h_i)$ to be the probability that the LSH selects the hash function h_i.

We will set $s = F$ and use Farkas's Lemma (Lemma 15) to show that the linear system $Mp = F^\star$ does not admit a non-negative solution p. To be able to show this, we seek a vector π indexed by elements in $\binom{V}{2} \cup \{\star\}$ such that $\pi M \geq \mathbf{0}$ and $\pi F^\star < 0$. We show that the following (magical) choice works:

$$\pi(\{X_{A,i}, X_{B,j}\}) = (-1)^{|A|+|B|},$$

for each $\{X_{A,i}, X_{B,j}\} \in \binom{V}{2}$ and $\pi(\star) = n2^{k-1} = |V|/2$. Then, Lemma 19 and Lemma 20 together with Farkas's Lemma will complete the proof. □

Lemma 19. $\pi F^\star < 0$.

Proof. Let $f_i = f(x + hi)$ for $i = 0, \ldots, k$.

$$\begin{aligned}
\pi F^\star &= \sum_{\{X_{A,i}, X_{B,j}\} \in \binom{V}{2}} \left(\pi\left(\{X_{A,i}, X_{B,j}\}\right) F(X_{A,i}, X_{B,j})\right) + \pi(\star) \\
&\leq \sum_{\{X_{A,i}, X_{B,j}\} \in \binom{V}{2}} \left((-1)^{|A|+|B|} f_{|A \cap B|}\right) + \ell_1(F, f(H)) + \pi(\star) \\
&= n^2 \sum_{\substack{\{A,B\} \in \binom{2^{[k]}}{2} \\ A \neq B}} \left((-1)^{|A|+|B|} f_{|A \cap B|}\right) + \binom{n}{2} \sum_{A \in 2^{[k]}} \left((-1)^{2|A|} f_{|A|}\right) + \ell_1(F, f(H)) + \pi(\star) \\
&\leq n^2 \sum_{\substack{\{A,B\} \in \binom{2^{[k]}}{2} \\ A \neq B}} \left((-1)^{|A|+|B|} f_{|A \cap B|}\right) + \binom{n}{2} \sum_{A \in 2^{[k]}} f_{|A|} + \ell_1(F, f(H)) + \pi(\star).
\end{aligned}$$

Consider the first sum and let $I = A \cap B$ and $D = (A \cup B) \setminus (A \cap B)$. Observe that $D \neq \varnothing$. The number of unordered pairs $\{A, B\}$ of subsets of $[k]$ with intersection I and non-empty symmetric difference D is $2^{|D|-1}$. Using this, and writing the summation in terms of these, we obtain

$$\begin{aligned}
\pi F^\star &\leq n^2 \sum_{I \subset [k]} \sum_{\substack{D \subseteq [k]-I \\ D \neq \varnothing}} \left((-1)^{2|I|+|D|} 2^{|D|-1} f_{|I|} \right) + \binom{n}{2} \sum_{A \in 2^{[k]}} f_{|A|} + \ell_1(F, f(H)) + \pi(\star) \\
&= \frac{n^2}{2} \sum_{i=0}^{k-1} \left(\binom{k}{i} f_i \sum_{d=1}^{k-i} \left((-1)^d \binom{k-i}{d} 2^d \right) \right) + \binom{n}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i \right) + \ell_1(F, f(H)) + \pi(\star) \\
&= \frac{n^2}{2} \left(\sum_{i=0}^{k-1} \left(\binom{k}{i} f_i \sum_{d=0}^{k-i} \left((-1)^d \binom{k-i}{d} 2^d \right) \right) - \sum_{i=0}^{k-1} \left(\binom{k}{i} f_i \right) \right) + \binom{n}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i \right) + \ell_1(F, f(H)) + \pi(\star) \\
&= \frac{n^2}{2} \left(\sum_{i=0}^{k-1} \left(\binom{k}{i} f_i (-1)^{k-i} \right) - \sum_{i=0}^{k-1} \left(\binom{k}{i} f_i \right) \right) + \binom{n}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i \right) + \ell_1(F, f(H)) + \pi(\star)
\end{aligned}$$

where the last step follows from the binomial identity; continuing, we get

$$\begin{aligned}
\pi F^\star &\leq \frac{n^2}{2} \sum_{i=0}^{k-1} \left(\binom{k}{i} f_i (-1)^{k-i} \right) + \frac{n^2}{2} f_k - \frac{n}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i \right) + \ell_1(F, f(H)) + \pi(\star) \\
&= \frac{n^2}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i (-1)^{k-i} \right) - \frac{n}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i \right) + \ell_1(F, f(H)) + \pi(\star) \\
&= -\delta \frac{n^2}{2} - \frac{n}{2} \sum_{i=0}^{k} \left(\binom{k}{i} f_i \right) + \ell_1(F, f(H)) + \pi(\star) \\
&\leq -\delta \frac{n^2}{2} + \ell_1(F, f(H)) + \pi(\star) \quad \text{(since } f_i \geq 0 \text{ from (i))} \\
&\leq \frac{n}{2} \left(-n\delta + t2^k + 2^k \right) \quad \text{(since } \ell_1(F, f(H)) \leq nt2^{k-1} \text{, and } \pi(\star) = n2^{k-1}\text{)} \\
&< 0 \quad \text{(since } n > (t+1) \cdot 2^k \cdot \delta^{-1}\text{).}
\end{aligned}$$

□

Lemma 20. $\pi M \geq \mathbf{0}$.

Proof. It suffices to show for each $h \in \mathcal{H}$, $\pi M_h \geq 0$, where M_h is the column of M indexed by h. Let $C_1, \ldots, C_T$, $T = T(h)$, be the equivalence classes the hash function h induces on V; note that $\sum_{t=1}^{T} |C_t| = |V| = 2\pi(\star)$.

$$\begin{aligned}
\pi M_h &= \sum_{\substack{\{X_{A,i}, X_{B,j}\} \in \binom{V}{2} \\ h(X_{A,i}) = h(X_{B,j})}} \pi\left(\{X_{A,i}, X_{B,j}\} \right) + \pi(\star) = \sum_{\substack{\{X_{A,i}, X_{B,j}\} \in \binom{V}{2} \\ h(X_{A,i}) = h(X_{B,j})}} (-1)^{|A|+|B|} + \pi(\star) \\
&= \sum_{t=1}^{T} \sum_{\{X_{A,i}, X_{B,j}\} \in \binom{C_t}{2}} (-1)^{|A|+|B|} + \pi(\star) = \sum_{t=1}^{T} \left(\sum_{\substack{\{X_{A,i}, X_{B,j}\} \in \binom{C_t}{2} \\ |A|+|B| \text{ is even}}} 1 - \sum_{\substack{\{X_{A,i}, X_{B,j}\} \in \binom{C_t}{2} \\ |A|+|B| \text{ is odd}}} 1 \right) + \pi(\star).
\end{aligned}$$

Note that $|A| + |B|$ is even if and only if $|A|$ and $|B|$ have the same parity. Therefore the first (resp., second) inner sum is equal to the number of unordered pairs of objects in C_t whose sets

have cardinalities of the same (different) parities. For each $t = 1, \ldots, T$, we define e_t (resp., o_t) to be the number of objects $X_{A,i}$ in C_t such that A has even (resp., odd) cardinality. Rewriting the last expression in terms of these quantities,

$$\begin{aligned}\pi M_h &= \sum_{t=1}^{T} \left(\binom{e_t}{2} + \binom{o_t}{2} - e_t o_t \right) + \pi(\star) = \sum_{t=1}^{T} \frac{(e_t - o_t)^2 - (e_t + o_t)}{2} + \pi(\star) \\ &\geq \sum_{t=1}^{T} \frac{-(e_t + o_t)}{2} + \pi(\star) = -\frac{1}{2} \sum_{t=1}^{T} |C_t| + \pi(\star) = -\frac{1}{2}|V| + \pi(\star) = 0. \end{aligned}$$ □

We note that the previous proof implies the following corollary, which basically says that either $f : [0, 1) \to [0, 1]$ has all non-negative forward differences, or it will produce similarities, operating on universes V, that need to be modified by a total additive term of $\Omega(|V|^2)$ to become LSHable (we observe that $O(|V|^2)$ is enough for any similarity).

Corollary 21. *Let $f : [0, 1) \to [0, 1]$ be a function, and let k, x, h be such that $\Delta_h^k(f, x) = -\delta < 0$. Consider the intersection similarity $H_{k,n,x,h}$, with $n = \left\lfloor (t+1)\frac{2^k}{\delta} \right\rfloor + 1$, for an arbitrary positive integer t. $H_{k,n,x,h}$ operates on a universe set V of cardinality $|V| = n2^k$.*

Then, no similarity F for which $\ell_1 (F, f (H_{k,n,x,h})) \leq nt2^{k-1} = O\left(\delta 4^{-k} |V|^2\right)$ is LSHable.

Finally, we are ready to show that being a (possibly scaled-down) PGF is a necessary condition for a function to be LSH-preserving.

Theorem 22. *If f is LSH-preserving, then there exists a PGF $p(x)$ and a constant $\alpha \in [0, 1]$ such that $f(x) = \alpha p(x)$ for each $x \in [0, 1)$.*

Proof. Since f is LSH-preserving, Theorem 18(ii) shows that the requirements of Definition 13 are satisfied by f in the interval $[0, 1)$. Therefore, f is absolutely monotonic in $[0, 1)$ and Theorem 14 guarantees that $f(x)$, for each $x \in [0, 1)$, can be expressed as $f(x) = \sum_{i=0}^{\infty} p_i x^i$, with $p_i = \frac{f_+^{(i)}(0)}{i!} \geq 0$; this expression shows that f is continuous in $[0, 1)$. Therefore, if $\sum_{i=0}^{\infty} p_i > 1$, or if it diverges, then there must exist some $x < 1$ such that $f(x) > 1$. But this contradicts Theorem 18(i). Hence, it follows that $\sum_{i=0}^{\infty} p_i \leq 1$. Now, choosing $\alpha = \sum_{i=0}^{\infty} p_i$ and choosing $p(x) = \sum_{i=0}^{\infty} \left(\left(p_i \cdot \alpha^{-1} \right) \cdot x^i \right)$ (or, if $\alpha = 0$, choosing any PGF $p(x)$) finishes the proof. □

4 Application: Set similarities

We define a family of similarity notions on sets. Given $A, B \subseteq [n]$, we define $A \triangle B := (A \setminus B) \cup (B \setminus A)$ and $\overline{A \cup B} := [n] \setminus (A \cup B)$ to be respectively the symmetric difference and the complement of the union of A and B.

Definition 23 (Rational set similarity)**.** *Given $0 \leq x, y$ and $0 \leq z \leq z'$, with $\max(x, y, z') > 0$, we define the* rational set similarity $S_{x,y,z,z'}$ *between non-empty sets $A, B \subseteq U$ as*

$$S_{x,y,z,z'}(A, B) := \frac{x\,|A \cap B| + y\left|\overline{A \cup B}\right| + z\,|A \triangle B|}{x\,|A \cap B| + y\left|\overline{A \cup B}\right| + z'\,|A \triangle B|},$$

if it is defined and 1 otherwise.

Similarity coefficient S	x	y	z	z'	Is $1-S$ a metric?	Is S LSHable?	Is $(1-S)^\alpha$ a metric?	Is $1-(1-S)^\alpha$ LSHable?
Jaccard	1	0	0	1	yes	yes	yes	**yes**
Hamming	1	1	0	1	yes	yes	yes	**yes**
Anderberg	1	0	0	2	yes	**yes**	yes	**yes**
Rogers–Tanimoto	1	1	0	2	yes	**yes**	yes	**yes**
Sørensen–Dice	2	0	0	1	no	no	$\text{yes}_{(\alpha \leq 1/2)}$	**no**

Table 1: Some rational set similarities [15, 19]. Here, $\alpha \in (0, 1]$. Bold-faced entries are new results.

By definition, range $S \subseteq [0, 1]$ and and $S(A, A) = 1$. Many well-known and widely-used set similarity metrics are captured by Definition 23; Table 1 gives a few examples. Janssens [18] precisely characterized those rational set similarities that yield metrics.

Theorem 24 (Janssens [18])**.** *$1 - S_{x,y,z,z'}$ is a metric if and only if $z' \geq \max(x, y, z)$.*

For some of the rational similarities S, it turns out that only $\sqrt{1-S}$ is a metric (see Table 1). A general definition is as follows.

Definition 25 (Root similarity)**.** *Given $0 < \alpha \leq 1$, $x, y, z, z' \geq 0$ with $\max(x, y, z') > 0$, we define the* root similarity $S^{(\alpha)}_{x,y,z,z'}$ *between non-empty sets $A, B \subseteq U$ as*

$$S^{(\alpha)}_{x,y,z,z'}(A, B) := 1 - \left(1 - \frac{x\,|A \cap B| + y\,\left|\overline{A \cup B}\right| + z\,|A \triangle B|}{x\,|A \cap B| + y\,\left|\overline{A \cup B}\right| + z'\,|A \triangle B|}\right)^{\alpha},$$

if it is defined and 1 otherwise.

4.1 LSHability of rational set similarities

In this section we show that all rational set similarities that are metrics are LSHable. To this end, we show a sufficient condition for the LSHability of rational set similarities.

Theorem 26. *If $z' \geq \max(x, y, z) > 0$, then $S_{x,y,z,z'}$ is LSHable.*

Proof. We will first give an LSH for a slightly different similarity function.

Define $X = \{(\eta, \eta), (\eta, \perp), (\perp, \eta)\}$ to be the set of *action pairs* of a hash function on a set A. The pair (a, b) of actions will roughly mean the following: if an element is in the set, then apply action a, otherwise apply action b. Here, (i) η denotes the action of the hash function returning the identity of the element together with a bit denoting the presence/absence of the element in set A; this is similar in spirit to the shingle approach [10], but with the additional bit that is important in our case; (ii) $\perp$ denotes the postponement of the hash determination. Let $p_{a,b}$ be a probability distribution on the elements of X.

Let $G = [n] \times X$. Observe that $|G| = 3n$. Let S_G be the symmetric group on G. Our function $h = h_\pi$ will be determined by a random draw from $\pi \in S_G$. The permutation π will not be chosen uniformly at random. Instead, for each $i = 1, \ldots, |G|$, (i) an element $x = (k, a, b)$ will be chosen randomly in $G \setminus \{\pi(1), \ldots, \pi(i-1)\}$ with probability proportional to $p_{a,b}$ (if $p_{a,b} = 0$ for all the elements $(k, a, b) \in G \setminus \{\pi(1), \ldots, \pi(i-1)\}$, choose x arbitrarily), and (ii) $\pi(i)$ will be set to x.

The hash function $h(A) = h_\pi(A)$ will operate as in Figure 1. Observe that $h(A)$ is defined for each $A \subseteq [n]$.

We show the following.

Algorithm 1 The hash function $h(A) = h_\pi(A)$ will operate as above on set $A \subseteq [n]$.

if $A = \varnothing$ **and** $p_{\eta,\eta} = p_{\perp,\eta} = 0$ **then return** $h(A) = \varnothing$
if $A = [n]$ **and** $p_{\eta,\eta} = p_{\eta,\perp} = 0$ **then return** $h(A) = [n]$
for $i = 1, \ldots, |G|$ **do**
  $(k, a, b) = \pi(i)$
  if $k \in A$ **and** $a = \eta$ **then return** $h(A) = (\in, k)$
  if $k \notin A$ **and** $b = \eta$ **then return** $h(A) = (\notin, k)$
end for

Lemma 27. *For $A \neq B$, $A, B \subseteq [n]$,*

$$\Pr_h[h(A) = h(B)] = \frac{(p_{\eta,\eta} + p_{\eta,\perp})\,|A \cap B| + (p_{\eta,\eta} + p_{\perp,\eta})\,\left|\overline{A \cup B}\right|}{(p_{\eta,\eta} + p_{\eta,\perp})\,|A \cap B| + (p_{\eta,\eta} + p_{\perp,\eta})\,\left|\overline{A \cup B}\right| + |A \triangle B|} := S'(A, B).$$

Proof. Observe that the claim is trivial if either A or B equal $\varnothing$ and $p_{\eta,\eta} = p_{\perp,\eta} = 0$, or if either A or B equal $[n]$ and $p_{\eta,\eta} = p_{\eta,\perp} = 0$. We assume the contrary. Observe that the set R that spans the value of $h(A)$ and $h(B)$ is equal to

$$R = (A \cap B) \times \{(\eta,\eta), (\eta,\perp)\} \cup (\overline{A \cup B}) \times \{(\eta,\eta), (\perp,\eta)\} \cup (A \triangle B) \times \{(\eta,\eta), (\eta,\perp), (\perp,\eta)\}.$$

Consider the projection π' of the permutation π on the elements of R. Given A and B, the event $h(A) = h(B)$ is completely determined by π'. Furthermore, the algorithm we described to sample a permutation π on the ground set G, if used to sample π'' on R, would induce on π'' the same distribution we have on π'.

By Bayes' rule, the total mass of set R appears in the denominator of $\Pr_h[h(A) = h(B)]$. Since $h(A) = h(B)$ if and only if the first element of π' is in the set

$$(A \cap B) \times \{(\eta,\eta), (\eta,\perp)\} \cup (\overline{A \cup B}) \times \{(\eta,\eta), (\perp,\eta)\},$$

it follows that our expression of $\Pr_h[h(A) = h(B)]$ holds and there is an LSH for the similarity S'. □

We now turn to our original similarity function $S_{x,y,z,z'}$. There are two cases to consider.

(i) If $x \geq y$, then set $p_{\eta,\eta} = y/x$, $p_{\eta,\perp} = 1 - y/x$ and all other $p_{a,b}$'s to zero, obtaining an LSH for $S_{x,y,0,x}$. Applying Lemma 11, with $w = \frac{z'}{x}$, we get an LSH for $S_{x,y,0,z'}$.

(ii) If $y > x$, then set $p_{\eta,\eta} = x/y$, $p_{\perp,\eta} = 1 - x/y$ and all other $p_{a,b}$'s to zero, obtaining an LSH for $S_{x,y,0,y}$. We again apply Lemma 11, this time with $w = \frac{z'}{y}$, to get an LSH for $S_{x,y,0,z'}$.

Finally, we use Lemma 7 between the LSH for $S_{x,y,0,z'}$, with weight $1 - z/z'$, and the all-ones LSH (T_1 in Observation 6), with weight z/z', to obtain an LSH for $S_{x,y,z,z'}$. □

For Jaccard set similarity, it can be seen that the above scheme reduces to the minwise-independent permutations, the LSH for Jaccard. To summarize, from Theorem 26 and combining Lemma 2, Lemma 38, and Theorem 24, we obtain the following.

Corollary 28. *The following are equivalent: (i) $S_{x,y,z,z'}$ is LSHable, (ii) $1 - S_{x,y,z,z'}$ is a metric, (iii) $1 - S_{x,y,z,z'}$ can be isometrically embedded into ℓ_1 with unit scaling factor, and (iv) $z' \geq \max(x, y, z) > 0$.*

Section 5.2 discusses the size of the sketch, an important practical consideration.

4.2 LSHability of root similarities

We will first obtain the exact conditions under which the function $f_{\alpha,v,w}$ (defined in the following theorem) is a PGF. The proof of the following theorem can be found in Section 4.2.1.

Theorem 29. *Let $\alpha \in (0,1]$, $v, w \geq 0$, and let*

$$f_{\alpha,v,w}(x) = 1 - \left(1 - \left(\frac{v}{w} + \left(1 - \frac{v}{w}\right)\frac{x}{x + w(1-x)}\right)\right)^{\alpha}.$$

Then, $f_{\alpha,v,w}(x)$ is a PGF if and only if $w \geq \frac{\alpha+1}{2}$ and $v \leq w$.

We now employ Theorem 29 to obtain a sufficient condition for a root similarity to be LSHable.

Theorem 30. *The root similarity $S^{(\alpha)}_{x,y,z,z'}$ is LSHable if $\alpha \in (0,1]$, $z' \geq \frac{\alpha+1}{2}\max(x,y)$, and $0 \leq z \leq z'$.*

Proof. We start by using Lemma 27. If $x \geq y$, we set $p_{\eta,\eta} = y/x$, $p_{\eta,\perp} = 1 - y/x$ and $p_{\perp,\eta} = 0$, obtaining an LSH for $S^{(1)}_{x,y,0,x}$. If $y > x$, we set $p_{\eta,\eta} = x/y$, $p_{\perp,\eta} = 1 - x/y$ and $p_{\eta,\perp} = 0$, obtaining an LSH for $S^{(1)}_{x,y,0,y}$. If we let $M = \max(x,y)$, then we have an LSH for $S^{(1)}_{x,y,0,M}$. We now apply Theorem 29 to obtain an LSH for $S^{(\alpha)}_{x,y,z,z'} = f_{\alpha,z/M,z'/M}(S^{(1)}_{x,y,0,M})$. □

Conversely, we will show the following theorem in Section 4.2.2:

Theorem 31. *The root similarity $S^{(\alpha)}_{x,0,z,z'}$ is not LSHable if $z' < x\frac{\alpha+1}{2}$ or $z > z'$. Analogously, the root similarity $S^{(\alpha)}_{0,y,z,z'}$ is not LSHable if $z' < y\frac{\alpha+1}{2}$ or $z > z'$.*

We now state a result of Gower and Legendre [15, Theorem 12] on the Euclidean embeddability of a specific root similarity. (Gower and Legendre use the notation $\sqrt{1 - T_{z'}}$ to refer to the distance $1 - S^{(1/2)}_{1,0,0,z'}$).

Theorem 32 (Gower and Legendre [15])**.** *$1 - S^{(1/2)}_{1,0,0,z'}$ is isometrically embeddable into ℓ_2, for each $z' \geq \frac{1}{2}$.*

Using Theorem 30 and Theorem 31, we can obtain that the similarity $S^{(1/2)}_{1,0,0,z'}$ is LSHable iff $z' \geq \frac{3}{4}$.

$S^{(1/2)}_{1,0,0,z'}$, for $\frac{1}{2} \leq z' < \frac{3}{4}$, thus represents a family of similarity functions S such that $1 - S$ is embeddable[3] into ℓ_2, but S is not LSHable. As far as we know, this is the first family of similarities that was shown to have these two properties.

4.2.1 Proof of Theorem 29

We start by proving with the following lemma.

Lemma 33. *Let*

$$h_i = h_{i,\alpha,w}(x) = {}_2F_1\left(-i, -i+1; -i+1+\alpha; (1-x)(1-w)\right).$$

[3]We observe that ℓ_2-embeddability implies ℓ_p-embeddability for each $p \in [1,\infty]$.

Then, the ith derivative of the function $f_{\alpha,w}(x) = 1 - \left(1 - \frac{x}{x+w(1-x)}\right)^{\alpha}$, $i \geq 1$, *equals*

$$f^{(i)}_{\alpha,w}(x) = \frac{\alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}} \cdot h_{i,\alpha,w}(x).$$

Also,

$$h_{i+1,\alpha,w}(x) = h_{i,\alpha,w}(x) \cdot \left(1 - 2 \cdot \frac{i(1-w)(1-x)}{i-\alpha}\right) - h_{i-1,\alpha,w}(x) \cdot \frac{i(i-1)(1-w)(1-x)(w+x(1-w))}{(i-\alpha)(i-1-\alpha)}.$$

Proof. We prove the claim by induction. For $i = 1$, we have:

$$\begin{aligned}
\frac{d}{dx} f_{\alpha,w}(x) &= -\alpha \left(1 - \frac{x}{x+w(1-x)}\right)^{\alpha-1} \cdot \frac{d}{dx}\left(1 - \frac{x}{x+w(1-x)}\right) \\
&= \alpha \left(1 - \frac{x}{x+w(1-x)}\right)^{\alpha-1} \frac{(x+w(1-x)) - x \cdot (1-w)}{(x+w(1-x))^2} \\
&= \alpha \left(\frac{w(1-x)}{x+w(1-x)}\right)^{\alpha-1} \frac{w}{(x+w(1-x))^2} \\
&= \frac{\alpha w^{\alpha}}{(1-x)^{1-\alpha}(x+w(1-x))^{1+\alpha}} \\
&= \frac{\alpha w^{\alpha}}{(1-x)^{1-\alpha}(w+x(1-w))^{1+\alpha}} \cdot {}_2F_1\left(-1, 0; \alpha; (1-x)(1-w)\right) = f^{(1)}_{\alpha,w}(x),
\end{aligned}$$

where the last equality follows from

$$\begin{aligned}
{}_2F_1\left(-1, 0; \alpha; (1-x)(1-w)\right) &= \sum_{n=0}^{\infty} \frac{(n-1)^{\underline{n}} \cdot (n)^{\underline{n}}}{(n+\alpha)^{\underline{n}}} \cdot \frac{((1-x)(1-w))^n}{n!} \\
&= \sum_{n=0}^{0} \frac{(n-1)^{\underline{n}} \cdot (n)^{\underline{n}}}{(n+\alpha)^{\underline{n}}} \cdot \frac{((1-x)(1-w))^n}{n!} = 1,
\end{aligned}$$

the second to equality being implied by $(n-1)^{\underline{n}} = 0$ for each $n \geq 1$.

We have proved that the statement holds for $i = 1$; assuming it holds for $i \geq 1$, we now prove that it holds for $i+1$. Recall that

$$h_{i,\alpha,w}(x) = {}_2F_1\left(-i, -i+1; -i+1+\alpha; (1-x)(1-w)\right).$$

We have

$$\begin{aligned}
\frac{d^{i+1}}{dx^{i+1}} f_{\alpha,w}(x) &= \frac{d}{dx} f^{(i)}_{\alpha,w}(x) \\
&= \frac{d}{dx}\left(\frac{\alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}} \cdot h_{i,\alpha,w}(x)\right) \\
&= h_{i,\alpha,w}(x) \cdot \alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}} \cdot \left(\frac{d}{dx} \frac{1}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}}\right) + \\
&\quad \frac{\alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}} \cdot \left(\frac{d}{dx} h_{i,\alpha,w}(x)\right). \qquad (1)
\end{aligned}$$

We compute the former derivative in (1):

$$\begin{aligned}\frac{d}{dx}\frac{1}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}} &= -\left((1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}\right)^{-2}\cdot\frac{d}{dx}\left((1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}\right)\\ &= \frac{(1-x)^{i-1-\alpha}(w+x(1-w))^{i-1+\alpha}}{(1-x)^{2i-2\alpha}(w+x(1-w))^{2i+2\alpha}}\cdot\\ &\quad((i-\alpha)(w+x(1-w))-(i+\alpha)(1-w)(1-x))\\ &= \frac{(i-\alpha)(w+x(1-w))-(i+\alpha)(1-w)(1-x)}{(1-x)^{i+1-\alpha}(w+x(1-w))^{i+1+\alpha}}\end{aligned}$$

The latter derivative in (1) can be computed by using the equality $\frac{d}{dx}\,{}_2F_1(a,b;c;f(x)) = \frac{ab}{c}\,{}_2F_1(a+1,b+1;c+1;f(x))\frac{d}{dx}f(x)$ which holds for arbitrary hypergeometric functions.

$$\frac{d}{dx}h_{i,\alpha,w}(x) = \frac{(-i)(-i+1)}{-i+1+\alpha}h_{i-1,\alpha,w}(x)\cdot\frac{d}{dx}\left((1-w)(1-x)\right) = (1-w)\cdot\frac{i(i-1)}{i-1-\alpha}h_{i-1,\alpha,w}(x).$$

Then, multiplying the two derivatives by their coefficients in (1), we get

$$\frac{d}{dx}\frac{\alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}} = \frac{\alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}}\cdot\left(\frac{i-\alpha}{1-x}-\frac{(1-w)(i+\alpha)}{w+x(1-w)}\right),$$

and,

$$\frac{\alpha w^{\alpha}(i-1-\alpha)^{\underline{i-1}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}}\cdot\left(\frac{d}{dx}h_{i,\alpha,w}(x)\right) = -\frac{\alpha w^{\alpha}(1-w)i(i-1)(i-2-\alpha)^{\underline{i-2}}}{(1-x)^{i-\alpha}(w+x(1-w))^{i+\alpha}}\cdot h_{i-1,\alpha,w}(x).$$

Since it is convenient we multiply both sides of (1) by the term $\frac{(1-x)^{i+1-\alpha}(w+x(1-w))^{i+1+\alpha}}{\alpha w^{\alpha}(i-\alpha)^{\underline{i}}}$, to get:

$$\begin{aligned}&\frac{(1-x)^{i+1-\alpha}(w+x(1-w))^{i+1+\alpha}}{\alpha w^{\alpha}(i-\alpha)^{\underline{i}}}\cdot\frac{d^{i+1}}{dx^{i+1}}f_{\alpha,w}(x)\\ &= h_{i,\alpha,w}(x)\cdot\frac{(1-x)(w+x(1-w))}{i-\alpha}\cdot\left(\frac{i-\alpha}{1-x}-\frac{(1-w)(i+\alpha)}{w+x(1-w)}\right)-\\ &\quad h_{i-1,\alpha,w}(x)\cdot(1-w)(1-x)(w+x(1-w))\frac{i(i-1)}{(i-\alpha)(i-1-\alpha)}\\ &= h_{i,\alpha,w}(x)\cdot\left((w+x(1-w))-(1-x)(1-w)\frac{i+\alpha}{i-\alpha}\right)-\\ &\quad h_{i-1,\alpha,w}(x)\cdot(1-w)(1-x)(w+x(1-w))\frac{i(i-1)}{(i-\alpha)(i-1-\alpha)}\\ &= h_{i,\alpha,w}(x)\cdot\left(1-(1-w)(1-x)-(1-x)(1-w)\frac{i+\alpha}{i-\alpha}\right)-\\ &\quad h_{i-1,\alpha,w}(x)\cdot(1-w)(1-x)(w+x(1-w))\frac{i(i-1)}{(i-\alpha)(i-1-\alpha)}\\ &= h_{i,\alpha,w}(x)\cdot\left(1-2(1-w)(1-x)\frac{i}{i-\alpha}\right)-\\ &\quad h_{i-1,\alpha,w}(x)\cdot(1-w)(1-x)(w+x(1-w))\frac{i(i-1)}{(i-\alpha)(i-1-\alpha)} = h_{i+1,\alpha,w}(x)\end{aligned}\tag{2}$$

where the last equality is implied by the hypergeometric contiguous relation which holds for any hypergeometric function (see [4]):
$\frac{c-(a+b+1)z}{c}\,{}_2F_1(a+1,b+1;c+1;z)+z(1-z)\frac{(a+1)(b+1)}{c(c+1)}\,{}_2F_1(a+2,b+2;c+2;z)=\,{}_2F_1(a,b;c;z)$.
In our case, we have $a=-i-1$, $b=-i$, $c=-i+\alpha$ and $z=(1-x)(1-w)$.

Equation (2) then implies:

$$\frac{d^{i+1}}{dx^{i+1}}f_{\alpha,w}(x)=\frac{\alpha w^{\alpha}(i-\alpha)^{\underline{i}}}{(1-x)^{i+1-\alpha}(w+x(1-w))^{i+1+\alpha}}\cdot h_{i+1,\alpha,w}(x),$$

and the induction is proved. □

The following theorem establishes the "core" of Theorem 29. At the end of this section, we will prove Theorem 29 as a corollary of Theorem 34.

Theorem 34. *The function* $f_{\alpha,w}(x)=1-\left(1-\frac{x}{x+w(1-x)}\right)^{\alpha}$ *is a PGF for* $\alpha\in(0,1)$, *and* $w\in\left[\frac{\alpha+1}{2},1\right)$. *Specifically,*

$$f_{\alpha,w}(x)=\sum_{i=1}^{\infty}\left(-\frac{P_i^{(-i-\alpha,-i+\alpha)}(1-2w)}{w^i}x^i\right),$$

where $P_n^{(a,b)}(x)$ *is the* n*th Jacobi polynomial with parameters* a,b.

Furthermore, if $w\in\left(0,\frac{\alpha+1}{2}\right)$, *the second derivative* $f^{(2)}_{\alpha,w}(x)$ *of* $f_{\alpha,w}(x)$ *exists, and is negative, for each* $x\in\left[0,\frac{1+\alpha-2w}{2-2w}\right)$. *Therefore* $f_{\alpha,w}(x)$ *is not a PGF if* $w\in\left(0,\frac{\alpha+1}{2}\right)$.

Proof. Observe that $f_{\alpha,w}(0)=0$ and that $f^{(i)}_{\alpha,w}(0)$ is defined for each $i\geq 1$. Therefore $f_{\alpha,w}(x)$ admits a Taylor expansion centered at zero:

$$f_{\alpha,w}(x)=\sum_{i=1}^{\infty}\left(\frac{f^{(i)}_{\alpha,w}(0)}{i!}x^i\right). \tag{3}$$

We will prove that for our choice of parameters, (a) for each $i\geq 1$, the ith Taylor expansion coefficient $c_{i,\alpha,w}=\frac{f^{(i)}_{\alpha,w}(0)}{i!}$ is non-negative, and (b) the Taylor error term goes to 0 as i increases.

Consider the relation between the hypergeometric terms of $f^{(i)}_{\alpha,w}(x)$, $f^{(i-1)}_{\alpha,w}(x)$, and $f^{(i-2)}_{\alpha,w}(x)$ derived in Lemma 33. Then,

Claim 35. *For each* $i\geq 1$, $\alpha\in(0,1)$, $w\geq\frac{\alpha+1}{2}$, *and* $x\in[0,1]$ *it holds that*

$$0\leq h_{i,\alpha,w}(x)\leq(w+x(1-w))^{i-1}\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}}.$$

Furthermore, if $w<\frac{\alpha+1}{2}$, *then* $h_{2,\alpha,w}(x)<0$ *for each* $x\in\left(0,\frac{1+\alpha-2w}{2-2w}\right)$.

Proof. For convenience, let us define $\rho=w+x(1-w)$. Observe that, given the assumptions in the lemma, it holds that $\frac{\alpha+1}{2}\leq w\leq\rho\leq 1$.

For $i=1$, we have that $h_{1,\alpha,w}(x)$ is the constant function $h_{1,\alpha,w}(x)=1$, and satisfies both inequalities, since the upper bound simplifies to 1.

For $i=2$ we have $h_{2,\alpha,w}(x)=1-\frac{2}{1-\alpha}(1-\rho)$. Since $1-\rho\leq\frac{1-\alpha}{2}$, we have that $h_{2,\alpha,w}\geq 0$. Furthermore, since $\frac{1}{1-\alpha}\geq 1$, to prove the upper bound on $h_{2,\alpha,w}$ it is sufficient to prove

$1 - \frac{2}{1-\alpha}(1-\rho) \leq \rho$; this is equivalent to $1 - \frac{2}{1-\alpha} \leq \rho\left(1 - \frac{2}{1-\alpha}\right)$, which, since $1 - \frac{2}{1-\alpha} < 0$, is equivalent to $\rho \leq 1$. The latter holds by our assumptions.

Therefore, the lemma holds for $i = 1, 2$. For $i \geq 3$, we use the recursive definition of $h_{i,\alpha,w}(x)$:

$$h_{i,\alpha,w}(x) = \left(1 - 2\frac{(i-1)(1-\rho)}{i-1-\alpha}\right) h_{i-1,\alpha,w}(x) + \rho(1-\rho)\frac{(i-1)(i-2)}{(i-1-\alpha)(i-2-\alpha)} h_{i-2,\alpha,w}.$$

The coefficient of $h_{i-1,\alpha,w}$ in the expression, $1 - 2\frac{(i-1)(1-\rho)}{i-1-\alpha}$, is non-negative for $i \geq 3$. Indeed, the expression is minimized when ρ is minimum, $\rho = \frac{\alpha+1}{2}$. In that case we have:

$$1 - 2\frac{i-1}{i-1-\alpha}\frac{1-\alpha}{2} = 1 - \frac{i-1-\alpha(i-1)}{i-1-\alpha} \geq 0,$$

where the inequality follows from $i \geq 3$.

The coefficient of $h_{i-2,\alpha,w}$ in the expression, $\rho(1-\rho)\frac{(i-1)(i-2)}{(i-1-\alpha)(i-2-\alpha)}$, is non-negative for $i \geq 3$, since all terms in the fraction are non-negative, and $0 \leq \rho \leq 1$.

We have then shown by induction that $h_{i,\alpha,w} \geq 0$ for each $i \geq 1$. We have also shown that the upper inequality in the lemma holds for $i = 1, 2$. We now prove by induction that the upper inequality also holds for each $i \geq 3$, concluding the proof. We again manipulate the recursive definition of $h_{i,\alpha,w}$, $i \geq 3$:

$$\begin{aligned}
h_{i,\alpha,w}(x) &= \left(1 - 2\frac{(i-1)(1-\rho)}{i-1-\alpha}\right) h_{i-1,\alpha,w}(x) + \rho(1-\rho)\frac{(i-1)(i-2)}{(i-1-\alpha)(i-2-\alpha)} h_{i-2,\alpha,w} \\
&\leq \left(1 - 2\frac{(i-1)(1-\rho)}{i-1-\alpha}\right) \rho^{i-2}\frac{(i-2)!}{(i-2-\alpha)^{\underline{i-2}}} + \rho(1-\rho)\frac{(i-1)(i-2)}{(i-1-\alpha)(i-2-\alpha)}\rho^{i-3}\frac{(i-3)!}{(i-3-\alpha)^{\underline{i-3}}} \\
&= \rho^{i-2} - 2(1-\rho)\rho^{i-2}\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}} + (1-\rho)\rho^{i-2}\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}} \\
&= \rho^{i-2} - (1-\rho)\rho^{i-2}\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}} \\
&= \rho^{i-2}\left(1 - \frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}}\right) + \rho^{i-1}\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}} \\
&\leq \rho^{i-1}\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}}.
\end{aligned}$$

where the last inequality follows from $\frac{(i-1)!}{(i-1-\alpha)^{\underline{i-1}}} \geq 1$.

To finish the proof of the lemma, we show that if $w \in \left(0, \frac{\alpha+1}{2}\right)$, then $h_{2,\alpha,w}(x) < 0$ for each $x \in \left(0, \frac{1+\alpha-2w}{2-2w}\right)$.

Recall that $h_{2,\alpha,w}(x) = 1 - \frac{2}{1-\alpha}(1-\rho) = 1 - \frac{2}{1-\alpha}(1-w)(1-x)$. Suppose $x < \frac{1+\alpha-2w}{2-2w}$; then,

$$\begin{aligned}
h_{2,\alpha,w}(x) &= 1 - \frac{2}{1-\alpha}(1-w)(1-x) \\
&< 1 - \frac{2}{1-\alpha}(1-w)\frac{1-\alpha}{2-2w} = 1 - 1 = 0.
\end{aligned}$$

This concludes the proof of the claim. □

Since $f^{(i)}_{\alpha,w}(x)$ has the same sign of $h_{i,\alpha,w}(x)$, Claim 35 guarantees that

- the coefficients $c_{i,\alpha,w} = \frac{f^{(i)}_{\alpha,w}}{i!}$ of the Taylor series (3) are all non-negative, if $w \geq \frac{\alpha+1}{2}$; and
- the second derivative $f^{(2)}_{\alpha,w}(x)$ of $f_{\alpha,w}(x)$ is negative for $x \in \left(0, \frac{1+\alpha-2w}{2-2w}\right)$, if $0 < w < \frac{\alpha+1}{2}$.

We now use the Schlömilc remainder theorem to prove that the series converges to $f_{\alpha,w}(x)$ for all $x \in [0,1]$:

Claim 36. *The series (3) converges to $f_{\alpha,w}(x)$ for $x \in [0,1]$.*

Proof. We use the Schlömilch form of the Taylor error term for a function f expanded around a:

$$R_i(x) = \frac{f^{(i+1)}(\xi)}{i!p}(x-\xi)^{i-p+1}(x-a)^p,$$

that holds for any $p > 0$, for some $\xi \in (a, x)$.

We upper bound the value of the $(i+1)$stt derivative of f via Lemma 35; for any $x \in [0,1)$, it holds that

$$\begin{aligned} f^{(i+1)}(x) &= \frac{\alpha w^\alpha (i-\alpha)^{\underline{i}}}{(1-x)^{i+1-\alpha}(w+x(1-w))^{i+1+\alpha}} h_{i+1,\alpha,w}(x) \\ &\leq \frac{\alpha w^\alpha (i-\alpha)^{\underline{i}}}{(1-x)^{i+1-\alpha}(w+x(1-w))^{i+1+\alpha}} (w+x(1-w))^i \frac{i!}{(i-\alpha)^{\underline{i}}} \\ &= i! \frac{\alpha w^\alpha}{(1-x)^{i+1-\alpha}(w+x(1-w))^{1+\alpha}} \end{aligned}$$

The Taylor series (3) was expanded at $a = 0$, and we choose $p = \alpha$. This gives us:

$$\begin{aligned} R_i(x) &\leq \frac{w^\alpha}{(1-\xi)^{i+1-\alpha}(w+\xi(1-w))^{1+\alpha}} (x-\xi)^{i+1-\alpha} x^\alpha \\ &= \frac{w^\alpha}{(w+\xi(1-w))^{1+\alpha}} \left(\frac{x-\xi}{1-\xi}\right)^{i+1-\alpha} x^\alpha \\ &\leq w^{-1} x^{i+1-\alpha} x^\alpha \\ &= w^{-1} x^{i+1}. \end{aligned}$$

Thus, the error term tends to 0 for each $x \in [0,1)$. Therefore series (3) equals $f_{\alpha,w}(x)$, for each $x \in [0,1)$. Since $f_{\alpha,w}(x)$ is a continuous function in $[0,1]$, and since $f_{\alpha,w}(1) = 1$, one has that the sum of the coefficients of 3 cannot be smaller than 1 (for otherwise, there would exist some $x < 1$ such that $f_{\alpha,w}(x)$ would be larger than the series' value at x). Furthermore, the sum of the coefficients cannot be larger than 1, for otherwise, since the coefficients are all non-negative, there would exist some $x < 1$ where the sum would have value larger than 1. Therefore the sum of the coefficients equals 1, $\sum_{i=1}^{\infty} c_{i,\alpha,w} = 1$, and the series converges to 1 at $x = 1$. The lemma is proved. □

Finally, the $f_{\alpha,w}$ series expression in the Theorem statement can be obtained by using the well-known Jacobi polynomial/finite hypergeometric transformation: in general, for any positive integer n, it holds that

$$P_n^{(a,b)}(x) = \frac{(a+1)^{\overline{n}}}{n!} \, {}_2F_1\left(-n, n+a+b+1; a+1; \frac{1-x}{2}\right).$$

In the $c_{i,\alpha,w}$ expression, the hypergeometric function has parameters $n = i$, $a = -i+\alpha$, $b = -i-\alpha$, and $x = 2w-1$; i.e.,

$$\begin{aligned} c_{i,\alpha,w} &= \frac{f^{(i)}_{\alpha,w}(0)}{i!} \\ &= -\frac{(i-1-\alpha)^{\underline{i}}}{w^i i!}\, {}_2F_1\left(-i,-i+1;-i+1+\alpha;1-w\right) \\ &= -\frac{(-1)^i(-i+1+\alpha)^{\overline{i}}}{w^i i!}\, {}_2F_1\left(-i,-i+1;-i+1+\alpha;1-w\right) \\ &= -\frac{(-1)^i}{w^i} P_i^{(-i+\alpha,-i-\alpha)}(2w-1), \end{aligned}$$

Finally, by using the identity $P_n^{(a,b)}(x) = (-1)^n P_n^{(b,a)}(-x)$, we obtain

$$c_{i,\alpha,w} = -w^{-i} P_i^{(-i-\alpha,-i+\alpha)}(1-2w)$$

and

$$f_{\alpha,w}(x) = \sum_{i=1}^{\infty} \left(c_{i,\alpha,w} x^i\right) = \sum_{i=1}^{\infty} \left(-w^{-i} P_i^{(-i-\alpha,-i+\alpha)}(1-2w) x^i\right)$$

for all $x \in [0,1]$. This concludes the proof of the Theorem. □

After having established Theorem 34, we can finally provide a proof for Theorem 29:

Proof of Theorem 29. We first show that if $v = 0$, then $f_{\alpha,v,w}(x) = f_{\alpha,0,w}(x)$ is a PGF. Observe that if $w \geq 1$, then $g_w(x) = \frac{x}{x+w(1-x)}$ is a PGF. Furthermore, for $\alpha \in (0,1]$, $h_\alpha(x) = 1-(1-x)^\alpha$ is also a PGF. Therefore, $f_{\alpha,0,w}(x) = h_\alpha\left(g_w(x)\right)$ is a PGF.

If $w < 1$, then $\alpha < 1$ and Theorem 34 guarantees that $f_{\alpha,0,w}(x) = f_{\alpha,w}(x)$ is a PGF.

Now, suppose that $v \in (0,w]$. Then,

$$\left(1-\frac{v}{w}\right)^\alpha f_{\alpha,0,w}(x) + \left(1-\left(1-\frac{v}{w}\right)^\alpha\right)$$

is a PGF. We manipulate the latter expression:

$$\begin{aligned} \left(1-\frac{v}{w}\right)^\alpha f_{\alpha,0,w}(x) + \left(1-\left(1-\frac{v}{w}\right)^\alpha\right) &= \left(1-\frac{v}{w}\right)^\alpha \left(1-\left(1-\frac{x}{x+w(1-x)}\right)^\alpha\right) + \left(1-\left(1-\frac{v}{w}\right)^\alpha\right) \\ &= 1-\left(1-\frac{v}{w}\right)^\alpha \left(1-\frac{x}{x+w(1-x)}\right)^\alpha \\ &= 1-\left(1-\left(\frac{v}{w}+\left(1-\frac{v}{w}\right)\frac{x}{x+w(1-x)}\right)\right)^\alpha = f_{\alpha,v,w}(x), \end{aligned}$$

so $f_{\alpha,v,w}(x)$ is a PGF assuming only the conditions in the statement of the theorem.

To conclude, we note that the conditions $0 \leq v \leq w$ and $w \geq \frac{\alpha+1}{2}$ are in fact necessary. Indeed, $f_{\alpha,v,w}(0) = \left(1-\frac{v}{w}\right)^\alpha$ and hence if $v > w$, then the latter is negative and thus $f_{\alpha,v,w}$ would not be a PGF. Furthermore, Theorem 34 states that $w < \frac{\alpha+1}{2}$ implies that the second derivative of $f_{\alpha,0,w}(x)$ is negative for sufficiently small $x > 0$. Since $f_{\alpha,v,w}(x)$ is a convex combination of $f_{\alpha,0,w}(x)$ and a constant, we have that the second derivative of $f_{\alpha,v,w}(x)$ itself is negative for sufficiently small $x > 0$; therefore, $f_{\alpha,v,w}(x)$ would not be a PGF. □

4.2.2 Proof of Theorem 31

We first start with a technical statement about embedding the intersection similarity in the Jaccard similarity.

Lemma 37. *The intersection similarity $H = H_{k,n,0,h}$ can be embedded in the Jaccard Similarity on $n2^k \left\lceil \frac{1}{2h} \right\rceil$ dimensions in such a way that (i) pairs of elements with 0 (resp., 1) similarity will retain their values and (ii) all other similarities will be distorted by a factor of $1 + O(kh)$.*

Proof. Let $B = \left\lceil \frac{1}{2h} \right\rceil$. Pick the m-dimensional Jaccard similarity with $m = n2^k B + k$. The similarity $H_{k,n,x,h}$ contains an element $X_{S,i}$ for each pair S, i with $S \subseteq [k]$, $1 \leq i \leq n$. Consider any bijection $\beta : 2^{[k]} \to [2^k]$. The element $X_{S,i}$ will be embedded into the Jaccard similarity as the vector $Y_{S,i}$, with

$$Y_{S,i}(j) = \begin{cases} 1 & \text{if } (n\beta(S) + i - 1)\, B \leq j \leq (n\beta(S) + i)\, B - 1 \text{ or} \\ & \exists s \in S \text{ s.t. } n2^k B + s = j, \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $H(X_{S,i}, X_{T,j}) = 1$, if $S = T$ and $i = j$, and $H(X_{S,i}, X_{T,j}) = |S \cap T|\, h$ otherwise. Furthermore, the Jaccard similarity J between $Y_{S,i}$ and $Y_{T,j}$ will be equal to:

- $J(Y_{S,i}, Y_{T,j}) = 1$, if $S = T$ and $i = j$, and
- $J(Y_{S,i}, Y_{T,j}) = \frac{|S \cap T|}{2B + |S \cup T|}$, otherwise.

If $H(X_{S,i}, X_{T,j}) \in \{0, 1\}$, i.e., if $S = T$ and $i = j$ or if $S \cap T = \varnothing$ and either $i \neq j$ or $S \cup T \neq \varnothing$, we have $H(X_{S,i}, X_{T,j}) = J(Y_{S,i}, Y_{T,j})$. Otherwise, we can bound the distortion as

$$J(Y_{S,i}, Y_{T,j}) \leq \frac{|S \cap T|}{h^{-1}} = H(X_{S,i}, X_{T,j}),$$

and

$$J(Y_{S,i}, Y_{T,j}) \geq \frac{|S \cap T|}{h^{-1} + 2 + k} = H(X_{S,i}, X_{T,j}) \frac{1}{1 + 2h + kh} = H(X_{S,i}, X_{T,j})(1 - O(kh)).$$

□

Combining Lemma 37, Theorem 21, and Theorem 29 we prove the desired result.

Proof of Theorem 31. We consider the case when $x > 0$ and $y = 0$; the case $y > 0$ and $x = 0$ is analogous. Given the conditions in the statement, Theorem 29 guarantees that $f_{\alpha,z/x,z'/x}$ has a negative coefficient. Therefore, there exists $k \geq 1$ such that $\Delta_h^k(f_{\alpha,z/x,z'/x}, 0) = -\delta < 0$ for each sufficiently small $h > 0$. Observe that $S^{(\alpha)}_{x,0,z,z'} = f_{\alpha,z/x,z'/x}(S^{(1)}_{x,0,0,x})$, where $S^{(1)}_{x,0,0,x}$ is the Jaccard similarity.

By Lemma 37, the intersection similarity $H_{k,n,0,h}$ can be embedded via ϕ into the Jaccard similarity $S^{(1)}_{x,0,0,x}$ with distortion $1 + O(kh)$ on non-integer similarities, and no distortion otherwise. Therefore $f_{\alpha,z/x,z'/x}(H_{k,n,0,h})$ can be embedded into $S^{(\alpha)}_{x,0,z,z'} = f_{\alpha,z/x,z'/x}(S^{(1)}_{x,0,0,x})$ with distortion $1 + O(kh)^\alpha$. Let V be the the universe on which $H_{k,n,0,h}$ operates. The ℓ_1 distance between the similarity $f_{\alpha,z/x,z'/x}(H_{k,n,0,h})$ and its embedding into $S^{(\alpha)}_{x,0,z,z'}$ is $O\left(|V|^2 (kh)^\alpha\right)$. By Theorem 21, no similarity F having ℓ_1 distance to $f_{\alpha,z/x,z'/x}(H_{k,n,0,h})$ less than $O\left(\delta 4^{-k} |V|^2\right)$ is LSHable. Since h can be chosen to be arbitrarily small, it follows that $S^{(\alpha)}_{x,0,z,z'}$ is not LSHable. □

5 Further properties of LSHable similarities

5.1 LSHability and embeddability

In this section we relate LSHability to embeddability. We have the following goals: proving an ℓ_1-embedding Lemma for similarities, showing a lower bound on the number of dimensions needed for embeddability, and obtaining an LSHable similarity that is not embeddable into the Hamming cube. We will also prove that no ℓ_p metric, for $1 < p < \infty$, can guarantee embeddability. (The ℓ_∞ metric, on the other hand, guarantees embeddability since any metric can be isometrically embedded into ℓ_∞.)

Lemma 38 is a minor correction of a result in [11], where the target metric space was the Hamming cube.

Lemma 38. *If a similarity S operating on a universe U is LSHable, then $1 - S$ is isometrically embeddable into $\ell_1^{O(|U|^2)}$.*

Before proving Lemma 38, we bound the maximum number of hash functions in the support of an LSH and the maximum size of their images.

Observation 39. *Suppose S is similarity operating on a universe U. If S admits an LSH, then it admits an LSH such that (i) every hash function that is assigned positive probability will have an image of cardinality upper bounded by $|U|$, (ii) the number of hash functions that are assigned positive probability is at most $\binom{|U|}{2} + 1$.*

Proof. The first claim is trivial: the cardinality of the image of a function cannot be larger than the cardinality of its domain.

Now, consider the system of linear equations for the similarity S. The system has one variable p_h for each hash function h, representing the probability assigned to that function by the LSH. Then, for each pair of distinct objects in $x, y \in U$, the system has the constraint

$$\sum_{\substack{h \\ h(x)=h(y)}} p_h = S(x, y).$$

Finally, the system contains the feasibility constraint $\sum_h p_h = 1$.

Observe that the system has $\binom{|U|}{2} + 1$ constraints. Therefore if there exists an LSH for S, there exists a solution to the linear system with at most $\binom{|U|}{2} + 1$ non-zero variables. This implies that the LSH can be supported by only $\binom{|U|}{2} + 1$ hash functions. □

We are now ready to prove the embedding lemma:

Proof of Lemma 38. By Observation 39 the similarity S admits an LSH using at most $\binom{|U|}{2} + 1$ hash functions each with codomain of size at most $|U|$. We allocate a block of $|U|$ coordinates to each hash function and hence the dimension of the space is $O(|U|^3)$.

Before describing the embedding let us fix arbitrarily, for each hash function h_i in the LSH of S, an injective function b_i from the codomain of h_i to the set of h_i's coordinates.

The embedding $\phi(x)$ or an element $x \in U$ is defined as follows: for each hash function h_i, $\phi(x)$ will have a single non-null coordinate in the set of h_i's coordinates. Specifically, the coordinate $b_i(h_i(x))$ will have value $\frac{1}{2} \cdot \Pr[h_i \text{ is chosen by the LSH}]$, and all the other coordinates in h_i's block will be zero.

Then, given two elements $x, y \in U$,

$$\ell_1\left(\phi\left(x\right), \phi\left(y\right)\right) = \sum_{\substack{h_i \\ h_i(x) \neq h_i(y)}} \left(2 \cdot \frac{1}{2} \cdot \Pr\left[h_i \text{ is chosen}\right]\right) = \Pr_h\left[h(x) \neq h(y)\right] = 1 - S(x, y).$$

Therefore there exists an embedding of $1 - S$ into $\ell_1^{O(|U|^3)}$. To get the embedding into $\ell_1^{O(|U|^2)}$ observe that S operates on the universe U, and that a set of n points embeddable into ℓ_1^m, for some $m \geq 1$, can also be embedded into $\ell_1^{\binom{n}{2}-1}$ [14]. □

We observe that the target metric of the embedding in the proof of Lemma 38 is a subset of ℓ_1 that closely resembles the Hamming cube: for a given sequence of non-negative weights $w_1, w_2, \ldots, \ldots, w_i, \ldots$, the embedding maps elements to vectors of the form $(b_1 \cdot w_1, b_2 \cdot w_2, \ldots, b_i \cdot w_i, \ldots)$ with $b_i \in \{0, 1\}$. (The Hamming cube would have had $w_i = 1$ for each i.)

It is then natural to try to use the Hamming cube as the target metric. Unfortunately, as we show next, there exists an LSHable similarity S such that $1 - S$ is not embeddable into the Hamming cube. (Still, for any LSHable similarity S one can get an embedding into the Hamming cube with distortion $1 + \epsilon$, for each $\epsilon > 0$; furthermore, if S has only rational values, then the embedding can be isometric, with a scaling factor.)

Observation 40. *There exists an LSHable similarity S such that $1 - S$ cannot be embedded into the n-dimensional Hamming cube, for every integer $n \geq 1$.*

Proof. We choose $S = S_{1,0,0,z'}$ operating on the ground set $\{0, 1, 2\}$, for an arbitrary irrational $z' > 1$; $S_{1,0,0,z'}$ is then LSHable by Corollary 28. Suppose by contradiction that $S_{1,0,0,z'}$ is isometrically embeddable into the Hamming cube with some scaling factor $c = c(z') > 0$. Consider the three sets $X = \{0\}, Y = \{0, 1\}$ and $Z = \{1, 2\}$; we can restrict the universe of S to be $U = \{X, Y, Z\}$.

Consider the distances $1 - S_{1,0,0,z'}(X, Y) = {}^{z'}/_{1+z'}$, $1 - S_{1,0,0,z'}(Y, Z) = {}^{2z'}/_{1+2z'}$. Since distances in the Hamming cube are natural numbers, it must be that the scaling factor c is commensurable to both ${}^{z'}/_{1+z'}$ and ${}^{2z'}/_{1+2z'}$, which must then be commensurable to each other, i.e., $\frac{{}^{2z'}/_{1+2z'}}{{}^{z'}/_{1+z'}} = \frac{2+2z'}{1+2z'}$ must equal $\frac{a}{b}$ for some integers a, b with $a > b$. This equality holds if and only if $z' = \frac{2b-a}{2a-2b}$. But this contradicts the irrationality of z'. It follows that $1 - S_{1,0,0,z'}$ is not isometrically embeddable into the Hamming cube. □

We now prove a lower bound on the number of dimensions needed for embedding $1 - S$ into ℓ_1, for an LSHable similarity S; Lemma 38 gave an upper bound of $O\left(|U|^2\right)$.

Observation 41. *There exists an LSHable similarity S operating on an arbitrarily large universe set U, such that $1 - S$ cannot be isometrically embedded in any ℓ_1 space with $o\left(\frac{|U|}{\log |U|}\right)$ dimensions.*

Proof. Alon and Pudlák [1] proved that embedding the equidistant metric on n points (i.e., the shortest path metric on the complete graph K_n) in ℓ_1 requires at least $c\frac{n}{\log n}$ dimensions, for some constant $c > 0$.

Consider the T_0 similarity of Observation 6. The associated metric $D = 1 - T_0$ assigns distance 1 to every pair of distinct points. Therefore if we set $S = T_0$ to operate on a universe U (of arbitrary cardinality), embedding $1 - S$ into ℓ_1 will require at least $c\frac{|U|}{\log |U|}$ dimensions. □

Finally, we show that there exists LSHable similarities S such that $1 - S$ is not embeddable into ℓ_p for each $1 < p < \infty$.

Observation 42. *There exists an LSHable similarity S such that $1 - S$ cannot be embedded into any ℓ_p metric, for each $1 < p < \infty$.*

Proof. We choose the similarity $S = S_{1,1,0,1}$ on the ground set $\{0, 1\}$. $S_{1,1,0,1}$ is LSHable by Corollary 28. If we define $x_0 = \varnothing$, $x_1 = \{0\}$, $x_2 = \{0, 1\}$, $x_3 = \{1\}$, we have

$$S(x_0, x_1) = S(x_1, x_2) = S(x_2, x_3) = S(x_3, x_0) = \frac{1}{2},$$

and

$$S(x_0, x_2) = S(x_1, x_3) = 0.$$

By $D(x, y) = 1 - S(x, y)$ we obtain

$$D(x_0, x_1) = D(x_1, x_2) = D(x_2, x_3) = D(x_3, x_0) = \frac{1}{2},$$

and

$$D(x_0, x_2) = D(x_1, x_3) = 1.$$

Consider a cycle graph G on four nodes, x_0, x_1, x_2, x_3, with edges of weight $1/2$. The shortest path metric on G coincides with D. Fichet [14, Lemma 8.6.1] (see also [13, Remark 3.2.5]) observed that the shortest path metric on G is not embeddable into ℓ_p (with any number of dimensions) for each $1 < p < \infty$. □

5.2 Sketch sizes

In this section we focus on the size of the sketches of LSHable similarities. This is an important consideration in practical applications, where it is imperative to have a sketch that is substantially more compact than the original object.

Recall the similarity T_0 of Observation 6 that assigned zero similarity to each pair of different objects. T_0, while admitting an LSH, requires $\lceil \log_2 N \rceil$ bits for a universe set U of $N = |U|$ objects. Indeed, in any LSH for T_0, there cannot exist a hash function with positive probability that maps two different objects to the same class: for otherwise, those two objects would end up having positive similarity in the LSH. Therefore the sketches have to have bit-size $\lceil \log_2 N \rceil$; this is also an upper bound by Observation 39. Contrast this against the sketch size for the Jaccard similarity. In that case, the set of objects was composed of all subsets of $[n]$, i.e., $N = 2^n$. Broder [8] showed that one can sketch an object using $\lceil \log_2 n \rceil = \lceil \log_2 \log_2 N \rceil$ bits. (More precisely, [8] shows that one can sketch all the subsets of $[n]$, with the exception of the empty set, with $\lceil \log_2 n \rceil$ bits — sketching the empty set might require an additional bit.) Therefore, there is an exponential gap in the number of bits required for sketches by the Jaccard similarity and the T_0 similarity.

In the case of set similarities, since $N = 2^n$, using a sketch of size $\lceil \log_2 N \rceil = n$ is undesirable. We are then motivated to weaken the requirement of LSHs, so to be able to use fewer bits in storing sketches. We introduce the following notion.

Definition 43 (Approximate LSH)**.** *An* ϵ-approximate LSH*, or* ϵ-aLSH *for a similarity function S over some set of objects is a set $\mathcal{H}$ of hash functions over the set of objects and a probability distribution over them, such that for any two objects A, B,*

$$(1 - \epsilon) \Pr_{h \in \mathcal{H}}[h(A) = h(B)] \leq S(A, B) \leq (1 + \epsilon) \Pr_{h \in \mathcal{H}}[h(A) = h(B)].$$

A similarity is ϵ-LSHable if it admits an ϵ-approximate LSH. Under the new definition, we show a theorem similar to Theorem 10.

Theorem 44. *If $f(x)$ is a PGF, and the similarity S is LSHable using sketches of size t, then the similarity $f(S)$ is ϵ-LSHable using sketches of size at most ct, for some constant $c = c(\epsilon, f)$ independent of S, if $\epsilon > 0$, and of unbounded size if $\epsilon = 0$.*

Proof. The case $\epsilon = 0$ is taken care of by Theorem 10. We assume $\epsilon > 0$.

Let $f(x) = \sum_{i=0}^{\infty} p_i x^i$, where $\{p_i\}_{i=0}^{\infty}$ is a probability distribution. Given ϵ, let $k \geq 0$ be the smallest integer such that $\sum_{i=k}^{\infty} p_i \leq \epsilon$. Let $\{p_i^*\}_{i=0}^{k}$ be the following probability distribution:

$$p_i^* = \begin{cases} p_i & \text{if } 0 \leq i < k \\ \sum_{j=k}^{\infty} p_j & \text{if } i = k \\ 0 & \text{if } i > k \end{cases}$$

We let $f^*(x) = \sum_{i=0}^{k} p_i^* x^i$ and apply Lemma 7 and Corollary 9 to get an LSH for $f^*(S)$. Observe that in the application of Lemma 7 and Corollary 9, the total number of the original similarity S sketches that will make up the new similarity $f^*(S)$ sketches will be $\sum_{i=0}^{k} i \leq k^2$. Therefore the sketches of $f^*(S)$ will have size at most $O(k^2 t)$ and k only depends on f and ϵ.

Now, it is easy to see that $f^*(x) \geq f(x)$.

$$f^*(x) = \sum_{i=0}^{k} (p_i^* \cdot x^i) = \sum_{i=0}^{k-1} (p_i^* \cdot x^i) + (p_k^* \cdot x^k) = \sum_{i=0}^{k-1} (p_i \cdot x^i) + \sum_{i=k}^{\infty} (p_i \cdot x^k) \geq \sum_{i=0}^{\infty} (p_i \cdot x^i) = f(x),$$

where the inequality follows from $x \leq 1$. As the final step, we give the reverse inequality:

$$\begin{aligned} f^*(x) &= \sum_{i=0}^{k} (p_i^* \cdot x^i) = \sum_{i=0}^{k-1} (p_i^* \cdot x^i) + (p_k^* \cdot x^k) \leq \sum_{i=0}^{k-1} (p_i \cdot x^i) + \epsilon \cdot x^k \\ &\leq \sum_{i=0}^{k-1} (p_i \cdot x^i) + \epsilon \cdot x^k \cdot \frac{\sum_{i=0}^{k-1} p_i}{1-\epsilon} \leq \sum_{i=0}^{k-1} (p_i \cdot x^i) + \epsilon \cdot \frac{\sum_{i=0}^{k-1} (p_i \cdot x^i)}{1-\epsilon} \\ &\leq \frac{1}{1-\epsilon} \cdot \sum_{i=0}^{k-1} (p_i \cdot x^i) \leq \frac{1}{1-\epsilon} \cdot \sum_{i=0}^{\infty} (p_i \cdot x^i) = \frac{1}{1-\epsilon} f(x), \end{aligned}$$

where the second inequality is justified by $\sum_{i=0}^{k-1} p_i \geq 1 - \epsilon$.

Since $f(S)$ satisfies $(1-\epsilon) f^*(S) \leq f(s) \leq f^*(S)$, we have that the LSH for $f^*(S)$ is an ϵ-LSH for $f(S)$. □

By replacing Theorem 10 with Theorem 44 in Lemma 11, we can obtain a geometric series ϵ-aLSH function. If we then use that geometric aLSH function in Theorem 26, we get an ϵ-aLSH using only logarithmically (that is, $O(\log n) = O(\log \log N)$) many bits per sketch for any rational set similarity $S_{x,y,z,z'}$ that is LSHable. Therefore, the LSH we propose for rational set similarities can be used in practical applications with little loss over the number of bits per sketch used by the Jaccard similarity sketch approach of [8].

5.3 Multivariate LSH-preserving functions

We show that a multivariate power series is LSH-preserving if and only if it is a (possibly scaled down) multivariate PGF.

Lemma 45. *Let $k \geq 1$ and let*

$$f(x_1, \ldots, x_k) = \sum_{i_1,\ldots,i_k \geq 0} \left(p_{i_1,\ldots,i_k} \prod_{j=1}^{k} x_j^{i_j} \right).$$

Then f is LSH-preserving if and only if (i) $p_{i_1,\ldots,i_k} \geq 0$ for $i_1, \ldots, i_k \geq 0$ and (ii) $\sum_{i_1,\ldots,i_k \geq 0} p_{i_1,\ldots,i_k} \leq 1$.

Proof. If (i) and (ii) hold then we can prove that f is LSH-preserving using the same approach of Theorem 10; that is, Lemma 8 and Corollary 9 imply that $\prod_{j=1}^{k} S_j^{i_j}$ is LSHable, for non-negative integers i_j and LSHable similarities S_j, $j = 1, \ldots, k$. Then, using Lemma 7, we obtain that the similarity $f(S_1, \ldots, S_k)$ is LSHable provided that similarities S_j, $j = 1, \ldots, k$, are LSHable.

Observe that if f is LSH-preserving, then for any sequence of non-negative integers $e_1, \ldots, e_k$, the univariate function $g(x) = g_{e_1,\ldots,e_k}(x) = f\left(x^{e_1}, \ldots, x^{e_k}\right)$ must be LSH-preserving.

We prove (i) by contradiction: we assume that there exist indices $i_1^*, \ldots, i_k^* \geq 0$ such that $p_{i_1^*,\ldots,i_k^*} < 0$, and we prove that for some e_j's, g is not LSH-preserving.

Let $b = 1 + \sum_{j=1}^{k} i_j^*$. We set $e_j = b^k + b^{j-1}$ for $j = 1, \ldots, k$. Then,

$$\begin{aligned} g(x) = f\left(x^{e_1}, \ldots, x^{e_k}\right) &= \sum_{i_1,\ldots,i_k \geq 0} \left(p_{i_1,\ldots,i_k} \prod_{j=1}^{k} x^{e_j i_j} \right) \\ &= \sum_{i_1,\ldots,i_k \geq 0} \left(p_{i_1,\ldots,i_k} x^{\sum_{j=1}^{k}(e_j i_j)} \right) = \sum_{n \geq 0} \left(x^n \sum_{\substack{i_1,\ldots,i_k \geq 0 \\ \sum_{j=1}^{k}(e_j i_j) = n}} p_{i_1,\ldots,i_k} \right). \end{aligned} \tag{4}$$

Let $n^* = \sum_{j=1}^{k} \left(e_j i_j^* \right)$. Then

$$n^* = b^k \sum_{j=1}^{k} i_j^* + \sum_{j=1}^{k} (b^{j-1} i_j^*) = b^k(b-1) + \sum_{j=1}^{k} (b^{j-1} i_j^*) \leq b^k(b-1) + b^{k-1}(b-1) = b^{k+1} - b^{k-1}.$$

Observe that we also have $n^* = b^k(b-1) + \sum_{j=1}^{k} (b^{j-1} i_j^*) \geq b^{k+1} - b^k$.

Consider the coefficient c_{n^*} of x^{n^*} in (4):

$$c_{n^*} = \sum_{\substack{i_1,\ldots,i_k \geq 0 \\ \sum_{j=1}^{k}(e_j i_j) = n^*}} p_{i_1,\ldots,i_k}.$$

We claim that $i_1^*, \ldots, i_k^*$ is the only sequence of indices $i_1, \ldots, i_k$ that satisfy the condition in the sum. This implies that $c_{n^*} = p_{i_1^*,\ldots,i_k^*}$ and therefore the coefficient c_{n^*} of x^{n^*} is negative. Theorem 22 implies that g cannot be LSH-preserving, yielding a contradiction.

Suppose that $p_{i_1,\ldots,i_k}$ is part of the c'_{n^*} sum. Then, if we let

$$n' = \sum_{j=1}^{k} (e_j i_j) = b^k \sum_{j=1}^{k} i_j + \sum_{j=1}^{k} (b^{j-1} i_j),$$

we have that n' must equal n^*, $n' = n^*$. Observe that if $\sum_{j=1}^k i_j > \sum_{j=1}^k i_j^* = b-1$, then $n' \geq b^{k+1} > b^{k+1} - b^{k-1} \geq n^*$; also, if $\sum_{j=1}^k i_j < \sum_{j=1}^k i_j^* = b-1$, then $n' \leq b^k(b-2) + b^{k-1}(b-2) = b^{k+1} - b^k - 2b^{k-1} < b^{k+1} - b^k \leq n^*$. That is, in either case, $n' \neq n^*$.

We then assume that $\sum_{j=1}^k i_j = \sum_{j=1}^k i_j^* = b-1$. Since

$$n' = \sum_{j=1}^{k} (e_j i_j) = b^k(b-1) + \sum_{j=1}^{k} (b^{j-1} i_j),$$

and

$$n^* = \sum_{j=1}^{k} (e_j i_j^*) = b^k(b-1) + \sum_{j=1}^{k} (b^{j-1} i_j^*),$$

we have that $n' = n^*$ if and only if $\sum_{j=1}^k (b^{j-1} i_j) = \sum_{j=1}^k (b^{j-1} i_j^*)$. By $\sum_{\ell_1}^k i_\ell = b-1$, we get $i_j \leq b-1$ for each $j = 1, \ldots, k$. Therefore, $\sum_{j=1}^k (b^{j-1} i_j) = \sum_{j=1}^k (b^{j-1} i_j^*)$ holds if and only if $i_j = i_j^*$ for each $j = 1, \ldots, k$.

We have proved that (i) holds. Proving (ii) is easy: just observe that if

$$\sum_{i_1, \ldots, i_k \geq 0} p_{i_1, \ldots, i_k} > 1,$$

then by the continuity of f, there exists some $x < 1$ such that $f(x, \ldots, x) > 1$. If S is an LSHable similarity assigning value x to any two objects, we have that $f(S, \ldots, S) > 1$ and therefore f is not LSH-preserving. □

Acknowledgments

We thank the anonymous referees for suggestions.

References

[1] N. Alon and P. Pudlák. Equilateral sets in ℓ_p^n. *Geometric And Functional Analysis*, 13:467–482, 2003.

[2] A. Andoni. *Nearest Neighbor Search: the Old, the New, and the Impossible.* PhD thesis, MIT, 2009.

[3] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *CACM*, 51(1):117–122, 2008.

[4] G. E. Andrews, R. Askey, and R. Roy. *Special Functions.* Cambridge University Press, Cambridge, 1999.

[5] P. Assouad. Produit tensoriel, distances extrémales et réalisation de covariances. *Comptés Rendus de l'Académie des Sciences de Paris*, 288, 1979.

[6] P. Assouad. Plongements isométriques dans l^1: aspect analytique. *Séminaire Choquet (Initiation á l'Analyse)*, 14, 1980.

[7] S. Bernstein. Sur la définition et les propriétés des fonctions analytiques d'une variable réelle. *Mathematische Annalen*, 75:449–468, 1914.

[8] A. Broder. On the resemblance and containment of documents. In *Proc. SEQUENCES*, pages 21–29, 1997.

[9] A. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proc. 6th WWW*, pages 391–404, 1997.

[10] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *JCSS*, 60(3):630–659, 2000.

[11] M. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. 34th STOC*, pages 380–388, 2002.

[12] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proc. 20th SoCG*, pages 253–262, 2004.

[13] M. M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer-Verlag, 1997.

[14] B. Fichet. Dimensionality problems in ℓ_1-norm representations. *Classification and Dissimilarity Analysis*, 1994.

[15] J. C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48, 1986.

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

[17] P. Indyk, R. Motwani, P. Raghavan, and S. Vempala. Locality-preserving hashing in multidimensional spaces. In *Proc. 29th STOC*, pages 618–625, 1997.

[18] S. Janssens. *Bell Inequalities in Cardinality-Based Similarity Measurement*. PhD thesis, Universiteit Gent, 2006.

[19] L. Naish, H. J. Lee, and K. Ramamohanarao. A model for spectra-based software diagnosis. *ACM TOSEM*, 2011.

[20] I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39, 1938.

[21] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44, 1938.

[22] D. V. Widder. *The Laplace Transform*. Princeton University Press, 1946.