Distance-Sensitive Hashing

Martin Aumüller
BARC and IT University of Copenhagen
maau@itu.dk

Tobias Christiani
BARC and IT University of Copenhagen
tobc@itu.dk

Rasmus Pagh
BARC and IT University of Copenhagen
pagh@itu.dk

Francesco Silvestri
University of Padova
silvestri@dei.unipd.it

ABSTRACT

Locality-sensitive hashing (LSH) is an important tool for managing high-dimensional noisy or uncertain data, for example in connection with data cleaning (similarity join) and noise-robust search (similarity search). However, for a number of problems the LSH framework is not known to yield good solutions, and instead ad hoc solutions have been designed for particular similarity and distance measures. For example, this is true for output-sensitive similarity search/join, and for indexes supporting *annulus queries* that aim to report a point close to a certain given distance from the query point.

In this paper we initiate the study of *distance-sensitive hashing* (DSH), a generalization of LSH that seeks a family of hash functions such that the probability of two points having the same hash value is a given function of the distance between them. More precisely, given a distance space (X, dist) and a "collision probability function" (CPF) $f : \mathbb{R} \to [0, 1]$ we seek a distribution over pairs of functions (h, g) such that for every pair of points $\mathbf{x}, \mathbf{y} \in X$ the collision probability is $\Pr[h(\mathbf{x}) = g(\mathbf{y})] = f(\text{dist}(\mathbf{x}, \mathbf{y}))$. Locality-sensitive hashing is the study of how fast a CPF can *decrease* as the distance grows. For many spaces, f can be made exponentially decreasing even if we restrict attention to the symmetric case where $g = h$. We show that the *asymmetry* achieved by having a pair of functions makes it possible to achieve CPFs that are, for example, increasing or unimodal, and show how this leads to principled solutions to problems not addressed by the LSH framework. This includes a novel application to privacy-preserving distance estimation. We believe that the DSH framework will find further applications in high-dimensional data management.

To put the running time bounds of the proposed constructions into perspective, we show lower bounds for the performance of DSH constructions with increasing and decreasing CPFs under angular distance. Essentially, this shows that our constructions are tight up to lower order terms. In particular, we extend existing LSH lower bounds, showing that they also hold in the asymmetric setting.

CCS CONCEPTS

• **Theory of computation → Randomness, geometry and discrete structures**; *Data structures design and analysis*; • **Information systems →** *Information retrieval*;

KEYWORDS

locality-sensitive hashing; similarity search; annulus query

1 INTRODUCTION

The growth of data from a variety of sources that need to be managed and analyzed has made it increasingly important to design data management systems with features that make them robust and tolerant towards noisy data. For example: different texts representing the same object (in data reconciliation), slightly different versions of a string (in plagiarism detection), or feature vectors whose similarity reflects the affinity of two objects (in recommender systems). In data management, such tasks are often addressed using the similarity join operator [49].

When data sets are *high-dimensional*, traditional algorithmic approaches often fail. Fortunately, there are general principles for handling high-dimensional data sets. One of the most successful approaches is the locality-sensitive hashing (LSH) framework by Indyk and Motwani [32], further developed in collaboration with Gionis [27] and Har-Peled [29]. LSH is a powerful framework for approximate nearest neighbor (ANN) search in high dimensions that achieves sublinear query time, and it has found many further applications. However, for a number of problems the LSH framework is not known to yield good solutions, for example output-sensitive similarity search, and indexes supporting *annulus queries* returning points at distance approximately r from a query point.

Motivating example. A classical application of similarity search is in recommender systems: Suppose you have shown interest in a particular item, for example a news article $\mathbf{x}$. The semantic meaning of a piece of text can be represented as a high-dimensional *feature vector*, for example computed using latent semantic indexing [25]. In order to recommend other news articles we might search the set P of article feature vectors for articles that are "close" to $\mathbf{x}$. But in general it is not clear that it is desirable to recommend the "closest" articles. Indeed, it might be desirable to recommend articles that are on the same topic but are not *too* aligned with $\mathbf{x}$, and may provide a different perspective [2].

Discussion. Unfortunately, existing LSH techniques do not allow us to search for points that are "close, but not too close". In a nutshell: LSH provides a sequence of hash functions $h_1, h_2, \ldots$ such that if $\mathbf{x}$ and $\mathbf{y}$ are close we have $h_i(\mathbf{x}) = h_i(\mathbf{y})$ for some i with constant probability, while if $\mathbf{x}$ and $\mathbf{y}$ are distant we have $h_i(\mathbf{x}) = h_i(\mathbf{y})$ only with small probability (typically $1/n$, where n upper bounds the number of distant points). As a special case, this paper discusses techniques that allow us to refine the first requirement: If $\mathbf{x}$ and $\mathbf{y}$ are "too close" we would like collisions to occur only with very small probability. At first sight this seems impossible because we will, by definition, have a collision when $\mathbf{x} = \mathbf{y}$. However, this objection is overcome by switching to an *asymmetric* setting where

we work with pairs of functions (h_i, g_i) and are concerned with collisions of the form $h_i(\mathbf{x}) = g_i(\mathbf{y})$.

More generally, we initiate the systematic study of the following question: In the asymmetric setting, what is the class of functions f for which it is possible to achieve $\Pr[h(\mathbf{x}) = g(\mathbf{y})] = f(\text{dist}(\mathbf{x}, \mathbf{y}))$, where the probability is over the choice of (h, g) and $\text{dist}(\mathbf{x}, \mathbf{y})$ is the distance between $\mathbf{x}$ and $\mathbf{y}$. We refer to such a function as a *collision probability function* (CPF). More formally:

Definition 1.1. A *distance-sensitive hashing* (DSH) scheme for the space (X, dist) is a distribution $\mathcal{D}$ over pairs of functions $h, g : X \to \mathbb{R}$ with collision probability function (CPF) $f : \mathbb{R} \to [0, 1]$ if for each pair $\mathbf{x}, \mathbf{y} \in X$ and $(h, g) \sim \mathcal{D}$ we have $\Pr[h(\mathbf{x}) = g(\mathbf{y})] = f(\text{dist}(\mathbf{x}, \mathbf{y}))$.

The theory of locality-sensitive hashing is the study of *decreasing* CPFs whose collision probability is high for neighboring points and low for far-away points.

1.1 Our contributions

We initiate the systematic study of distance-sensitive hashing (DSH), and in particular we:

- Show tight upper and lower bounds on the maximum possible growth rate of CPFs under *angular distance*. This extends upper and lower bound techniques for locality-sensitive hashing to the asymmetric setting.
- Provide several DSH constructions that exploit asymmetry between the functions g and h to achieve non-standard CPFs. For example, for Hamming distance we show how to achieve a CPF that equals any polynomial $\mathcal{P} : \{0, \ldots, d\} \to [0, 1]$ up to a scaling factor.
- Present several motivating applications of DSH: Hyperplane queries, annulus search, spherical range reporting, privacy-preserving distance estimation.

The lower bound for angular distance implies lower bounds for Hamming and Euclidean distance DSH. It also shows that existing (asymmetric) LSH constructions used to search for vectors close to a given hyperplane [52] are near-optimal.

On the upper bound side, our constructions show that asymmetric methods are significantly more expressive than standard, symmetric LSH constructions. Since asymmetric methods are often applicable in cases where symmetric methods are used, it seems relevant to re-assess whether such constructions can be improved, even in settings where a decreasing CPF is desired.

Though our DSH applications do not lead to quantitative running time improvements compared to existing, published ad-hoc solutions, we believe that studying the possibilities and limitations of the DSH framework will help unifying approaches to solving "distance sensitive" algorithmic problems. We now proceed with a more detailed description of our results.

1.1.1 Angular distance. We consider monotonically increasing CPFs for angular distance between vectors on the unit sphere $\mathbb{S}^{d-1}$. It will be convenient to express distances in terms of dot products $\langle \cdot, \cdot \rangle$ rather than by angles or Euclidean distances, with the understanding that there is a 1-1 correspondence between them.

The constructions rely on the idea of "negating the query point": leveraging state-of-the-art symmetric LSH schemes, we obtain DSH constructions with monotone CPFs by replacing the query point $\mathbf{q}$ by $-\mathbf{q}$ in a symmetric LSH construction. We initially apply this idea in section 2.1 to Cross-Polytope LSH [8] getting an efficient DSH with a monotonically decreasing CPF. We then show that a more flexible result follows with a variant of ideas used in the filter constructions from [10, 14, 21]. The filter based approach contains a parameter t that can be used for fine tuning the scheme. This parameter is exploited in the data structure solving the annulus query problem (see section 6.2). More specifically, the filter based approach gives the following result:

Theorem 1.2. *For every $t > 1$ there exists a distance-sensitive family $\mathcal{D}_-$ for $(\mathbb{S}^{d-1}, \langle \cdot, \cdot \rangle)$ with a monotonically decreasing CPF f such that for every $\alpha \in (-1, 1)$ with $|\alpha| < 1 - 1/t$ we have*

$$\ln(1/f(\alpha)) = \tfrac{1+\alpha}{1-\alpha} \tfrac{t^2}{2} + \Theta(\log t). \tag{1}$$

The complexity of sampling, storing, and evaluating $(h, g) \in \mathcal{D}_-$ is $O(dt^4 e^{t^2/2})$.

Note that in terms of the angle between vectors f is *increasing*, as desired. A corollary of Theorem 1.2 is that we can efficiently obtain a CPF f such that $\rho_- = \ln f(0)/\ln f(\alpha) \leq \frac{1-\alpha}{1+\alpha} + o_t(1)$. (We use $o_t(\cdot)$ to indicate that the function depends only on t.) The value ρ_- measures the gap between collision probabilities (the larger the gap, the smaller ρ_-), and is significant in search applications. The construction based on cross-polytope LSH has the same CPF as the construction stated in Theorem 1.2 with $t = \sqrt{2 \log d}$.

It turns out that the value ρ_- is optimal up to the lower-order term. To show this we consider vectors $\mathbf{x}, \mathbf{y} \in \{-1, +1\}^d$ that are either 0-correlated or α-correlated (i.e., $\Pr[\mathbf{x}_i = \mathbf{y}_i] = \frac{1+\alpha}{2}$ independently for each i). These are unit vectors up to a scaling factor $\sqrt{d}$, and for large d the dot product will be tightly concentrated around the correlation. Since correlation is invariant under linear transformation, we may without loss of generality consider $\mathbf{x}, \mathbf{y} \in \{0, 1\}^d$. In section 3 we show the following lower bound:

Theorem 1.3. *Let $\mathcal{D}$ be a distribution over pairs of functions $h, g : \{0, 1\}^d \to R$, and define $\hat{f} : [-1, 1] \to [0, 1]$ as $\hat{f}(\alpha) = \Pr[h(\mathbf{x}) = g(\mathbf{y})]$ where $\mathbf{x}, \mathbf{y}$ are randomly α-correlated and $(h, g) \sim \mathcal{D}$. Then for every $0 \leq \alpha < 1$ we have that $\hat{f}(\alpha) \geq \hat{f}(0)^{\frac{1+\alpha}{1-\alpha}}$.*

Considering $f(\alpha) = \lim_{d \to \infty} \hat{f}(\alpha)$, taking logarithms of the final inequality we get that $\rho_- = \ln f(0)/\ln f(\alpha) \geq \frac{1-\alpha}{1+\alpha}$, as desired. Theorem 1.3 extends standard (symmetric) LSH lower bounds [12, 37, 40] to an asymmetric setting in the following sense: If there exists a too powerful asymmetric LSH family $\mathcal{D}_+$, then by negating the query point $\mathbf{q}$ we obtain a family $\mathcal{D}_-$ with a monotone CPF that contradicts the statement in Theorem 1.3. Observe that this reasoning cannot be applied to standard LSH bounds since they do not handle the asymmetry we allow in our setting.

The proof of the theorem builds on the Reverse Small-Set Expansion Theorem [39] which, given subsets $A, B \subseteq \{0, 1\}^d$ and random α-correlated vectors $\mathbf{x}, \mathbf{y}$, lower bounds the probability of the event $[\mathbf{x} \in A \wedge \mathbf{y} \in B]$ as a function of α, $|A|$ and $|B|$. Through a sequence of inequalities we extend the lower bound for pairs of subsets of space to hold for distributions over pairs of functions that partition space, yielding a surprisingly powerful and simple lower bound.

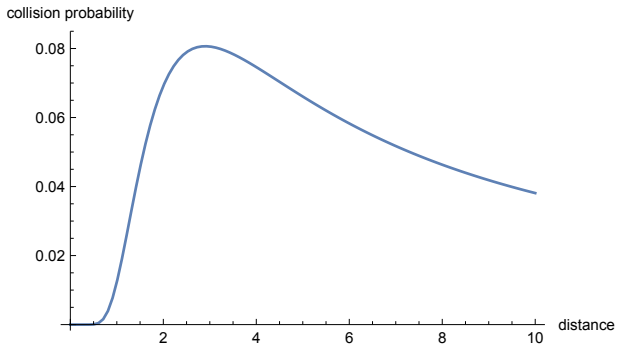

Figure 1: Collision probability function of (2) for $k = 3, w = 1$.

1.1.2 DSH constructions. In the paper we also presents several specific and general constructions for DSH families, in addition to the constructions on the unit sphere discussed above.

In section 4.1, we apply the negating trick to the well-known bit-sampling approach from [32] to get a simple construction of a DSH family with an increasing CPF for Hamming distance. The idea to obtain a DSH with increasing CPF is to use the bit-sampling approach, but negate the bit that is picked from the query point. We show that for small distances, the "collision probability gap" between points at distance r and r/c, which is measured by the value $\rho_- = \ln f(r)/\ln f(r/c)$, is worse than if we were to map bit strings onto the unit sphere and use the construction from Theorem 1.2. This can be considered somewhat surprising because the bit-sampling LSH family has an optimal collision probability gap for the collision probability of points at distance r and cr, measured by the value $\rho_+ = \ln f(r)/\ln f(cr)$, for small r [40].

The negating trick does not work for Euclidean space, since it is potentially unbounded. Nevertheless, we describe in section 4.2 a DSH construction with a ρ_--value asymptotically matching the performance of constructions on the unit sphere. The construction is based on an asymmetric version of the classical LSH family for Euclidean space by Datar et al. [23]: specifically, for parameters $w \in \mathbb{R}$ and $k \in \mathbb{N}$, we let

$$h\colon \mathbf{x} \mapsto \left\lfloor \frac{\langle \mathbf{a}, \mathbf{x}\rangle + b}{w} \right\rfloor, \qquad g\colon \mathbf{x} \mapsto \left\lfloor \frac{\langle \mathbf{a}, \mathbf{x}\rangle + b}{w} \right\rfloor + k, \tag{2}$$

where $b \in [0, w]$ is uniformly random and $\mathbf{a} \sim \mathcal{N}^d(0, 1)$ is a d-dimensional random Gaussian vector. We show that this method, for a suitable choice of parameters w and k, provides a near-optimal collision probability gap. This is surprising, since the classical construction in [23] is not optimal as an LSH for Euclidean space [7]. An example CPF for the above method is shown in Figure 1. As we can see from this figure, the collision probaiblity decreases rapidly on the left side of the maximum, but decreases more slowly on the right side. In particular, this construction has a CPF that is not monotone but *unimodal* (i.e., a distribution having a single local maximum).

In section 5, we extend our work by targeting the following natural question: let $\mathcal{P}(t)$ be a polynomial, does there exist a distance-sensitive hash family with CPF $f(t) = \mathcal{P}(t)$? We present two general approaches of constructing CPFs on the unit sphere and in Hamming space that cover a wide range of such polynomials. The result for the unit sphere easily follows from asymmetric embeddings previously used by Valiant [51] to solve the closest pair problem. For Hamming space, we propose an approach based on bit-sampling and polynomial factorization that obtains the desired CPF up to a scaling factor $\Delta \geq 1$ that depends on the roots of the polynomial.

1.1.3 Applications. We briefly describe here some applications of DSH constructions. They are all straightforward given a DSH, but give an idea of the versatility of the framework. We refer to section 6 for formal statements and details.

- **Hyperplane queries**. The problem of searching a set of unit vectors for a point (approximately) closest to a given hyperplane can be solved using a CPF f, parameterized by dot product, that peaks at $f(0)$. This approach was previously used in ad-hoc contructions, see [52].
- **Approximate annulus search**. The problem of searching for a point at approximately distance r from a query point can be solved using a CPF that peaks at r. Previous solutions to this problem used a different, two-stage filtering approach [41].
- **Approximate spherical range reporting**. Classical LSH data structures are inefficient when many near neighbors need to be found, since each neighbor may have a high collision probability and may be "found" many times. A "step function" CPF that is flat for small distances, and then rapidly decreases implies good output-sensitivity. Again, previous results addressing this problem were ad-hoc [4].
- **Privacy preserving distance estimation.** If two parties want to compute the distance between private vectors, limiting the leakage of other information, secure multi-party computations can be used (see e.g. [28]). Using "step function" CPFs we can transform this kind of question into a question about Hamming distance between vectors, for which much more efficient protocols exist [24, 26].

1.2 Related work

A recent book by Augsten and Böhlen [13] surveys algorithms for similarity join, with emphasis on data cleaning. Many of the commonly used algorithms are heuristics with weak theoretical guarantees, especially in high dimensions. Recently, however, a substantial literature has been devoted the theoretical study of similarity search and join in data management, e.g. [5, 30, 34, 54]. All of these papers address particular similarity or distance measures, and their results are not directly comparable to those obtained in this paper. In the following we review selected results from the LSH literature in more detail, referring to [6, 53] for comprehensive surveys.

For simplicity we consider only LSH constructions that are *isometric* in the sense that the probability of a hash collision depends only on the distance $\mathrm{dist}(\mathbf{x}, \mathbf{y})$. In other words, there exists a CPF $f\colon \mathbb{R} \to [0, 1]$ such that $\Pr[h(\mathbf{x}) = h(\mathbf{y})] = f(\mathrm{dist}(\mathbf{x}, \mathbf{y}))$. Almost all LSH constructions whose collision probability has been rigorously analyzed are isometric. Notable exceptions are recent *data dependent* LSH methods such as [11] where the LSH distributions, and thus the collision probabilities, depend on the structure of data.

ρ-values. Much attention has been given to optimal so-called ρ-values of locality-sensitive hash functions, where we consider *non-increasing* CPFs. Suppose we are interested in hash collisions when $\text{dist}(\mathbf{x}, \mathbf{y}) \leq r_1$ but want to avoid hash collisions when $\text{dist}(\mathbf{x}, \mathbf{y}) \geq r_2$, for some $r_2 > r_1$. The ρ-value of this setting (denoted in this paper with ρ_+) is the ratio of the logarithms of collision probabilities at distance r_1 and r_2, i.e., the real number in $[0, 1]$ such that $f(r_1) = f(r_2)^\rho$. The ρ-value determines the performance of LSH-based data structures for the (r_1, r_2)-*approximate near neighbor* problem, see [29]. In many spaces a good upper bound on ρ can be given in terms of the ratio $c = r_2/r_1$, but in general the smallest possible ρ can depend on r_1, r_2, $f(r_1)$, as well as the number of dimensions d. In our applications it will be natural to consider the "dual" ρ-value that measures the growth rate that can be achieved when distance increases.

LSHable functions. Charikar in [17] gave a necessary condition that all CPFs in the symmetric setting must fulfill, namely, $\text{dist}(\mathbf{x}, \mathbf{y}) = 1 - \Pr[h(\mathbf{x}) = h(\mathbf{y})]$ must be the distance measure of a metric, and more specifically this metric must be isometrically embeddable in ℓ_1. In the asymmetric setting this condition no longer holds since in general $\Pr[h(\mathbf{x}) = g(\mathbf{x})] < 1$.

Chierichetti et al. considered transformations that can be used to create new CPFs [18], and studied the decision problem to verify if there exists an LSH with given pairwise collision probabilities [19]. The transformations in [18, Lemma 7] are considered in a symmetric setting, but the same constructions applied in the asymmetric setting give the following result, proved for completeness in Appendix C.1:

Lemma 1.4. *Let $\{\mathcal{D}_i\}_{i=1}^n$ be a collection of n distance-sensitive families with CPFs $\{f_i\}_{i=1}^n$.*

(a) *There exists a distance-sensitive family $\mathcal{D}_{concat}$ with CPF $f(x) = \prod_{i=1}^n f_i(x)$.*
(b) *Given a probability distribution $\{p_i\}_{i=1}^n$ over $\{\mathcal{D}_i\}$, there exists a distance-sensitive family $\mathcal{D}^p$ with CPF $f(x) = \sum_{i=1}^n p_i f_i(x)$.*

An example application of Lemma 1.4 is shown in Figure 2. Interestingly, at least in the symmetric setting, the application of this lemma to a single CPF yields *all* transformations that are guaranteed to map a CPF to a CPF. Chierichetti et al. [20] recently extended the study of CPFs in the symmetric setting to allow *approximation*, i.e., allowing the collision probability to differ from a target function by a given approximation factor.

Asymmetric LSH. Motivated by applications in machine learning, Vijayanarasimhan et al. [52] presented asymmetric LSH methods for Euclidean space where the CPF is a decreasing function of the dot product $|\langle \mathbf{x}, \mathbf{y} \rangle|$. Shrivastava and Li [48] also explored how asymmetry can be used to achieve new CPFs (increasing), in settings where the inner product of vectors is used to measure closeness. Neyshabur and Srebro [38] extended this study by showing that the extra power obtained by asymmetry hinges on restrictions on the vector pairs for which we consider collisions: If vectors are not restricted to a bounded region of $\mathbb{R}^d$, no nontrivial CPF (as a function of inner product) is possible. On the other hand, if one vector is normalized (e.g. a query vector), the performance of known asymmetric LSH schemes can be matched with a symmetric method. But in the case where vectors are bounded but not normalized, asymmetric LSH is able to obtain CPFs that are impossible for symmetric LSH. Ahle et al. [5] showed further impossibility results for asymmetric LSH applied to inner products, and that symmetric LSH is possible in a bounded domain even without normalization if we just allow collision probability 1 when vectors coincide.

Indyk [31] showed how asymmetry can be used to enable new types of embeddings. More recently asymmetry has been used in the context of locality-sensitive *filters* [10, 21] and *maps* [22]. The idea is to map each point $\mathbf{x}$ to a pair of sets $(h(\mathbf{x}), g(\mathbf{x}))$ such that $\Pr[h(\mathbf{x}) \cap g(\mathbf{y}) \neq \emptyset]$ is constant if $\mathbf{x}$ and $\mathbf{y}$ are close, and very small if $\mathbf{x}$ and $\mathbf{y}$ are far from each other. This yields a nearest neighbor data structure that adds for each vector $\mathbf{x} \in P$ the elements of $h(\mathbf{x})$ to a hash table; a query for a vector $\mathbf{q}$ proceeds by looking up each key in $g(\mathbf{q})$ in the hash table. One can transform such methods into asymmetric LSH methods by using min-wise hashing [15, 16], see [21, Theorem 1.4].

Recommender systems. Returning to our motivating example, the topic of getting "interesting" recommendations using nearest neighbor methods is not new. Abbar et al. [1] built a nearest neighbor data structure on a *core-set* of P to guarantee diverse query results. However, this method effectively discards much of the data set, so may not be suitable in all settings. Indyk [31] and Pagh et al. [41] proposed data structures for finding the furthest neighbor in Euclidean space, leveraging random projections and using specialized data structures.

Privacy-preserving search. Privacy is an increasing concern in connection with data analytics. Proximity information is potentially sensitive, since it may be used to reveal the source of a data point. Ideally we would like information-theoretical privacy guarantees [3], but the standard technique of adding noise to data does not seem to work well for proximity problems, since adding noise merely shifts distances. Riazi et al. [45] considered answering nearest neighbor queries without leaking the actual distance. They showed that standard LSH approaches can compromise privacy under a "triangulation" attack, however this risk can be reduced by designing an LSH (symmetric) with a CPF that is "flat" in the region of interest of an attacker. However, only rather weak privacy guarantees were provided.

2 OPTIMAL ANGULAR DSH

This section describes DSH schemes with monotonically increasing and decreasing CPFs for the unit sphere that match the lower bounds shown in the following section 3. As LSH are monotone schemes too (i.e., CPF decreasing with distance or increasing with similarity), we refer to DSH schemes with the opposite monotonicity as *anti-LSH* (i.e., CPF increasing with distance or decreasing with similarity).

For notational simplicity, the CPF is expressed as function of the inner product of two points, however it holds for other angular similarity and distance measures: on the unit sphere, inner product is equivalent to the cosine similarity and there is a 1-1 correspondence with Euclidean distance and angular distance. Results on the unit sphere can be extended to ℓ_s-spaces for $0 < s \leq 2$ through Rahimi and Recht's [44] embedding version of Bochner's Theorem [46]

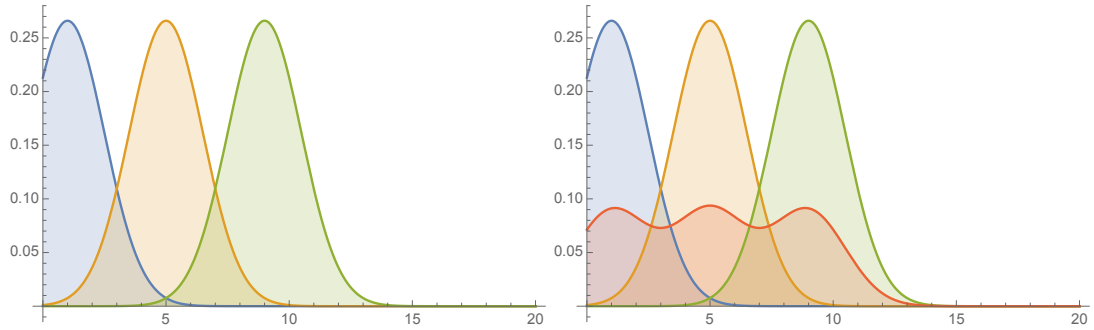

Figure 2: Composing several unimodal CPFs (left) to form a "step function" CPF (red curve on the right) using Lemma 1.4.

applied to the characteristic functions of s-stable distributions as used in [21].

We use the notation $\mathcal{D}_-$ to denote a family with CPF that is decreasing in the inner product between points, and $\mathcal{D}_+$ to denote a family with a CPF that is increasing in the inner product between points. The idea behind the following constructions is to take a standard (symmetric) locality-sensitive hash family $\mathcal{D}_+$ for the unit sphere with an increasing CPF, and transform it into a family $\mathcal{D}_-$ with a decreasing CPF by introducing asymmetry (i.e, by negating the query value).

2.1 Cross-polytope DSH schemes

Andoni et al. [8] described the following LSH family $\mathcal{CP}$ for the d-dimensional unit sphere $\mathbb{S}^{d-1}$: To sample a function $h \sim \mathcal{CP}$, sample a random matrix $A \in \mathbb{R}^{d\times d}$ in which each entry is independently sampled from a standard normal distribution $\mathcal{N}(0,1)$. To compute a hash value of a point $\mathbf{x}$, compute $\hat{\mathbf{x}} = A\mathbf{x}/\|A\mathbf{x}\| \in \mathbb{S}^{d-1}$ and map $\mathbf{x}$ to the closest point to $\hat{\mathbf{x}}$ among $\{\pm\mathbf{e_i}\}_{1\le i\le d}$, where $\mathbf{e_i}$ is the i-th standard basis vector of $\mathbb{R}^d$. Intuitively, a random hash function from $\mathcal{CP}$ applies a random rotation to a point and hashes it to its closest point on the cross-polytope.

To formally define a DSH from $\mathcal{CP}$, we sample $(h_+, g_+) \sim \mathcal{CP}_+$ by sampling $h \sim \mathcal{CP}$ and set $h_+ = g_+ = h$. In [8], Andoni et al. showed the following theorem, here reproduced in terms of inner product similarity.[1]

Theorem 2.1 (Theorem 1 in [8]). *Let f be the CPF of hash family $\mathcal{CP}_+$. Suppose that $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ such that $\langle \mathbf{x}, \mathbf{y}\rangle = \alpha$, where $\alpha \in (-1, 1)$. Then,*

$$\ln \frac{1}{f(\alpha)} = \frac{1-\alpha}{1+\alpha} \ln d + O_\alpha(\ln \ln d).$$

To obtain a DSH with a monotonically decreasing CPF (in the similarity), we define the family $\mathcal{CP}_-$ consisting of pairs (h_-, g_-) as follows: To sample a pair $(h_-, g_-) \sim \mathcal{CP}_-$, sample a function $h \sim \mathcal{CP}$. For each point $\mathbf{x} \in X$, set $h_-(\mathbf{x}) = h(\mathbf{x})$ and $g_-(\mathbf{x}) = h(-\mathbf{x})$. This means that we invert the query point before applying the hash function. Intuitively, we map the point to the point on the cross-polytope that is *furthest away* after applying the random rotation.

Corollary 2.2. *Let f be the CPF of hash family $\mathcal{CP}_-$. Suppose that $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ such that $\langle \mathbf{x}, \mathbf{y}\rangle = \alpha$, where $\alpha \in (-1, 1)$. Then,*

$$\ln \frac{1}{f(\alpha)} = \frac{1+\alpha}{1-\alpha} \ln d + O_\alpha(\ln \ln d).$$

Proof. If $\langle \mathbf{x}, \mathbf{y}\rangle = \alpha$, then $\langle \mathbf{x}, -\mathbf{y}\rangle = -\alpha$. Since $h_-(\mathbf{x}) = g_-(\mathbf{y})$ corresponds to $h(\mathbf{x}) = h(-\mathbf{y})$, we can apply the result of Theorem 2.1 for similarity threshold $-\alpha$. □

2.2 Filter-based DSH schemes

Recent work on similarity search for the unit sphere [9, 11, 14, 21] has used variations of the following technique: Pick a sequence of random spherical caps[2] and hash a point $\mathbf{x} \in \mathbb{S}^{d-1}$ to the index of the first spherical cap in the sequence that contains $\mathbf{x}$. By allowing the spherical cap to have different sizes for queries and updates, it was shown how to obtain space-time tradeoffs for similarity search on the unit sphere that is optimal for random data [10, 21]. We obtain a family $\mathcal{D}_-$ with a decreasing CPF by taking a standard (symmetric) family with its sequence of spherical caps and introduce asymmetry by negating the query point. Intuitively, this means that for $(h, g) \sim \mathcal{D}_-$ we let h use the original sequence of spherical caps while g uses the spherical caps that are *diametrically opposite* to the ones used by h. We therefore get a collision $h(\mathbf{x}) = g(\mathbf{y})$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are contained in random diametrically opposite spherical caps. We proceed by describing the family $\mathcal{D}_+$ and the modification that gives us the family $\mathcal{D}_-$.

The family $\mathcal{D}_+$ takes as parameter a real number $t > 0$ and an integer m that we will later set as a function of t. We sample a pair of functions (h, g) from $\mathcal{D}_+$ by sampling m vectors $z_1, \ldots, z_m$ where $\mathbf{z}_i \sim \mathcal{N}^d(0,1)$. The functions h, g map a point $\mathbf{x} \in \mathbb{S}^{d-1}$ to the index i of the first projection $\mathbf{z}_i$ where $\langle \mathbf{z}_i, \mathbf{x}\rangle \ge t$. If no such projection is found, then we ensure that $h(\mathbf{x}) \ne g(\mathbf{x})$ by mapping them to different values. Formally, we set

$$h_+(\mathbf{x}) = \min(\{i \mid \langle \mathbf{z}_i, \mathbf{x}\rangle \ge t\} \cup \{m+1\}),$$
$$g_+(\mathbf{x}) = \min(\{i \mid \langle \mathbf{z}_i, \mathbf{x}\rangle \ge t\} \cup \{m+2\}).$$

[1] An inner product of $\alpha \in (-1, 1)$ between two vectors on the unit sphere corresponds to Euclidean distance $\tau = \sqrt{2(1-\alpha)}$.

[2] A spherical cap is a portion of a sphere cut off by a plane.

We use the idea of negating the query point to obtain a family $\mathcal{D}_-$ from $\mathcal{D}_+$ by setting:

$$g_-(\mathbf{x}) = g_+(-\mathbf{x}) = \min(\{i \mid \langle \mathbf{z}_i, \mathbf{x}\rangle \leq -t\} \cup \{m+2\}).$$

The analysis of CPF for $\mathcal{D}_-$ and $\mathcal{D}_+$ and the proof of Theorem 1.2 are provided in Appendix A.1.

3 LOWER BOUND FOR MONOTONE DSH

This section provides lower bounds on the CPFs of DSH families in d-dimensional Hamming space under the similarity measure $\mathrm{sim}_H(\mathbf{x},\mathbf{y}) = 1 - 2\|\mathbf{x}-\mathbf{y}\|_1/d$. These results extend to the unit sphere and Euclidean space through standard embeddings.

Our primary focus is to obtain the lower bound in Theorem 1.3, which holds for a CPF that is decreasing with the similarity. As with our upper bounds for the unit sphere, re-applying the same techniques also yields a lower bound for the case of an increasing CPF in the similarity. The proof combines the (reverse) small-set expansion theorem by O'Donnell [39] with techniques inspired by the LSH lower bound of Motwani et al. [37]. The main contribution here is to extend this lower bound for pairs of subsets of Hamming space to our object of interest: distributions over pairs of functions that partition space. We begin by introducing the required tools from [39].

Definition 3.1. For $-1 \leq \alpha \leq 1$ and $\mathbf{x},\mathbf{y} \in \{0,1\}^n$ we say that $(\mathbf{x},\mathbf{y})$ is randomly α-correlated if $\mathbf{x}$ is uniformly distributed over $\{0,1\}^n$ and each component of $\mathbf{y}$ is i.i.d. according to

$$\mathbf{y}_i = \begin{cases} \mathbf{x}_i & \text{with probability } \frac{1+\alpha}{2}, \\ 1-\mathbf{x}_i & \text{with probability } \frac{1-\alpha}{2}. \end{cases}$$

The reverse small-set expansion theorem lower bounds the probability that random α-correlated points $(\mathbf{x},\mathbf{y})$ end up in a pair of subsets A, B of the Hamming cube, as a function of the size of these subsets. In the following, for $A \subset \{0,1\}^d$ we refer to the quantity $|A|/2^d$ as the *volume* of A.

Theorem 3.2 (Rev. Small-Set Expansion [39]). *Let $0 \leq \alpha \leq 1$. Let $A, B \subseteq \{0,1\}^d$ have volumes $\exp(-a^2/2)$, $\exp(-b^2/2)$, respectively, where $a, b \geq 0$. Then we have that*

$$\Pr_{\substack{(\mathbf{x},\mathbf{y}) \\ \alpha\text{-corr.}}}[\mathbf{x} \in A, \mathbf{y} \in B] \geq \exp\left(-\frac{1}{2}\frac{a^2 + 2\alpha ab + b^2}{1-\alpha^2}\right).$$

We define a probabilistic version of the collision probability function that we will state results for. In Section 3.1 we will apply concentration bounds on the similarity between α-correlated pairs of points in order to make statements about the actual CPF. We will use R to denote the range of a family of functions which, without loss of generality, we can assume to be finite.

Definition 3.3 (Probabilistic CPF). Let $\mathcal{D}$ be a distribution over pairs $h, g\colon \{0,1\}^d \to R$. We define the probabilistic CPF $\hat{f}\colon [-1,1] \to [0,1]$ by

$$\hat{f}(\alpha) = \Pr_{\substack{(h,g)\sim\mathcal{D} \\ (\mathbf{x},\mathbf{y})\ \alpha\text{-corr.}}}[h(\mathbf{x}) = g(\mathbf{y})].$$

The proof of the lower bound will make use of the following technical inequality that follows from two applications of Jensen's inequality.

Lemma 3.4. *Let p, q denote discrete probability distributions, then for every $c \geq 1$ we have that*

$$\sum_i (p_i q_i)^c \geq \left(\sum_i p_i q_i\right)^{2c-1}$$

with reverse inequality for $c \leq 1$.

Proof. Assume $c \geq 1$. By Jensen's inequality, using the fact that $x \mapsto x^c$ and $x \mapsto x^{2-1/c}$ are convex we have that

$$\sum_i (p_i q_i)^c = \sum_i p_i (p_i^{1-1/c} q_i) \geq \left(\sum_i p_i^{2-1/c} q_i\right)^c \geq \left(\sum_i p_i q_i\right)^{2c-1}.$$

For $c \leq 1$ we have that $x \mapsto x^c$ and $x \mapsto x^{2-1/c}$ are concave and the inequality is reversed. □

We are now ready to state our main lemma that lower bounds $\hat{f}(\alpha)$ in terms of $\hat{f}(0)$. This immediately implies Theorem 1.3.

Lemma 3.5. *For every $0 \leq \alpha < 1$ and for every distribution $\mathcal{D}$ over pairs of functions $h, g\colon \{0,1\}^d \to R$, we have that $\hat{f}(\alpha) \geq \hat{f}(0)^{\frac{1+\alpha}{1-\alpha}}$.*

Proof. For a function $h\colon \{0,1\}^d \to R$ define its inverse image $h^{-1}\colon R \to 2^{\{0,1\}^d}$ by $h^{-1}(i) = \{\mathbf{x} \in \{0,1\}^d \mid h(\mathbf{x}) = i\}$. For a pair of functions $(h,g) \in \mathcal{D}$ and $i \in R$ we define $a_{h,i}, b_{g,i} \geq 0$ such that $|h^{-1}(i)|/2^d = \exp(-a_{h,i}^2/2)$ and $|g^{-1}(i)|/2^d = \exp(-b_{g,i}^2/2)$. For fixed (h,g) define $\hat{f}_{h,g}(\alpha) = \Pr_{(\mathbf{x},\mathbf{y})\ \alpha\text{-corr.}}[h(\mathbf{x}) = g(\mathbf{y})]$. We obtain a lower bound on $\hat{f}(\alpha)$ as follows:

$$\begin{aligned}
\hat{f}(\alpha) &= \mathop{\mathrm{E}}_{(h,g)\sim\mathcal{D}}\left[\sum_{i\in R} \Pr_{(\mathbf{x},\mathbf{y})\ \alpha\text{-corr.}}[h(\mathbf{x}) = g(\mathbf{y}) = i]\right] \\
&\overset{(1)}{\geq} \mathop{\mathrm{E}}_{(h,g)\sim\mathcal{D}}\left[\sum_{i\in R} \exp\left(-\frac{1}{2}\frac{a_{h,i}^2 + 2\alpha a_{h,i} b_{g,i} + b_{g,i}^2}{1-\alpha^2}\right)\right] \\
&\overset{(2)}{\geq} \mathop{\mathrm{E}}_{(h,g)\sim\mathcal{D}}\left[\sum_{i\in R} \exp\left(-\frac{1}{2}\frac{a_{h,i}^2 + b_{g,i}^2}{1-\alpha}\right)\right] \\
&\overset{(3)}{\geq} \mathop{\mathrm{E}}_{(h,g)\sim\mathcal{D}} \hat{f}_{h,g}(0)^{\frac{1+\alpha}{1-\alpha}} \overset{(4)}{\geq} \left(\mathop{\mathrm{E}}_{(h,g)\sim\mathcal{D}} \hat{f}_{h,g}(0)\right)^{\frac{1+\alpha}{1-\alpha}} \\
&= \hat{f}(0)^{\frac{1+\alpha}{1-\alpha}}.
\end{aligned}$$

Here, (1) is due to Theorem 3.2, (2) follows from the simple fact that $a^2 + \alpha(a^2+b^2) + b^2 \geq a^2 + 2\alpha ab + b^2$, (3) follows from Lemma 3.4 with $c = 1/(1-\alpha)$, and (4) follows from a standard application of Jensen's Inequality. □

3.1 Extending the lower bound

We will now use Lemma 3.5 to together with concentration inequalities to obtain a lower bound on the ρ_--value of DSH schemes with a CPF that is decreasing in $\mathrm{sim}_H(\mathbf{x},\mathbf{y})$. We introduce the following property:

Definition 3.6. Let $\mathcal{D}$ be a DSH family for (X, dist) with CPF f. We say that $\mathcal{D}$ is $(\alpha_-, \alpha_+, f_-, f_+)$-*decreasingly sensitive* (resp., $(\alpha_-, \alpha_+, f_-, f_+)$-*increasingly sensitive*) if it satisfies:

- For $\alpha \leq \alpha_-$, we have $f(\alpha) \geq f_-$ (respectively, $f(\alpha) \leq f_-$);
- For $\alpha \geq \alpha_+$, we have $f(\alpha) \leq f_+$ (respectively, $f(\alpha) \geq f_+$).

We observe that if dist is a distance (resp., similarity) measure, then a decreasingly (resp., increasingly) sensitive DSH scheme corresponds to a standard LSH.

Theorem 3.7. *Let $0 < \alpha_- < \alpha_+ < 1$ be constants. Then every $(\alpha_-, \alpha_+, f_-, f_+)$-decreasingly sensitive family $\mathcal{D}$ for $(\{0,1\}^d, \mathrm{sim}_H)$ must satisfy*

$$\rho_- = \frac{\log(1/f_-)}{\log(1/f_+)} \geq \frac{1-\alpha_+}{1+\alpha_+-2\alpha_-} - O(\sqrt{\log(1/f_+)/d}).$$

In the statement of Theorem 3.7 we may replace the properties from Definition 3.6 that hold for every $\alpha \leq \alpha_-$ and every $\alpha \geq \alpha_+$ with less restrictive versions that hold in an ε-interval around α_-, α_+ for some $\varepsilon = o_d(1)$. Furthermore, if we rewrite the bound in terms of relative Hamming distances δ and δ/c where δ, c are constants, we obtain a lower bound of $1/(2c-1) - o_d(1)$ — an expression that is familiar from known LSH lower bounds [12, 37].

We now prove Theorem 3.7 proving an analogous result for a (r, cr, p, q)-increasingly sensitive family under Hamming distance. Indeed, a $(\alpha_-, \alpha_+, f_-, f_+)$-decreasingly sensitive family for $(\{0,1\}^d, \mathrm{sim}_H)$ is a (r, cr, p, q)-increasingly sensitive family for the space $(\{0,1\}^d, \mathrm{dist}_H)$, with $\alpha_- = 1 - r/d$ and $\alpha_+ = 1 - cr/d$.

Theorem 3.8. *For every constant $\varepsilon > 0$, we have that every (r, cr, p, q)-increasingly sensitive family $\mathcal{A}$ for $\{0,1\}^d$ under Hamming distance with $r \leq (1-\varepsilon)d/2$ must satisfy*

$$\rho(\mathcal{A}) = \frac{\log 1/p}{\log 1/q} \geq \frac{1}{2c-1} - O(\sqrt{(c/r)\log(1/q)}).$$

Proof. Given $\mathcal{A}$ we define a distribution $\hat{\mathcal{A}}$ over pairs of functions $\hat{h}, \hat{g} \colon \{0,1\}^{\hat{d}} \to R$ where $\hat{d} \leq d$ remains to be determined. We sample a pair of functions $(\hat{h}, \hat{g})$ from $\hat{\mathcal{A}}$ by sampling (h, g) from $\mathcal{H}$ and setting $\hat{h}(\mathbf{x}) = h(\mathbf{x} \circ \mathbf{1})$ and similarly $\hat{g}(\mathbf{x}) = g(\mathbf{x} \circ \mathbf{1})$ where $\mathbf{1}$ denotes the $(d - \hat{d})$-dimensional all-ones vector. We will now turn to the process of relating p to $\hat{p} = \hat{f}(0)$ and q to $\hat{q} = \hat{f}(\alpha)$ for $\hat{\mathcal{A}}$.

Let $0 < \varepsilon_p < 1$ and set $\hat{d} = \lceil 2r/(1-\varepsilon_p) \rceil$. Then by applying standard Chernoff bounds we get

$$\Pr_{(\mathbf{x},\mathbf{y})\ 0\text{-corr.}}[\mathrm{dist}(\mathbf{x},\mathbf{y}) \leq r] \leq \exp\left(-\frac{\varepsilon_p^2}{1-\varepsilon_p}\frac{r}{2}\right).$$

For convenience, define $\delta_p = \exp\left(-\frac{\varepsilon_p^2}{1-\varepsilon_p}\frac{r}{2}\right)$. Then $\hat{p} \geq (1-\delta_p)p$.

In order to tie $\hat{q}$ to q we consider the probability of α-correlated points having distance greater than r/c. The expected Hamming distance of α-correlated $(\mathbf{x}, \mathbf{y})$ in $\hat{d}$ dimensions is $\hat{d}(1-\alpha)/2$. We would like to set α such that the probability of the distance exceeding r/c is small. Let X denote $\mathrm{dist}(\mathbf{x}, \mathbf{y})$, then the standard Chernoff bound states that:

$$\Pr[X \geq (1+\varepsilon)\mu] \leq e^{-\varepsilon^2\mu/3}.$$

For a parameter $0 < \varepsilon_q < 1$ we set α such that the following is satisfied:

$$(1+\varepsilon_q)\mu \geq (1+\varepsilon_q)\frac{2r}{1-\varepsilon_p}\frac{1-\alpha}{2} \geq r/c.$$

This results in a value of $\alpha = 1 - \frac{1-\varepsilon_p}{1+\varepsilon_q}\frac{1}{c}$ and we observe that

$$\delta_q \leq \exp(-\varepsilon_q^2\mu/3) \leq \exp\left(-\frac{\varepsilon_q^2}{1+\varepsilon_q}\frac{r}{3c}\right).$$

It follows that

$$\hat{q} \leq (1-\delta_q)q + \delta_q.$$

Let us summarize what we know so far:

$$\begin{aligned}
&\hat{p} \geq (1-\delta_p)p \\
&\hat{q} \leq (1-\delta_q)q + \delta_q \leq q(1+\delta_q/q) \\
&0 < \varepsilon_p, \varepsilon_q < 1 \\
&\delta_p \leq \exp\left(-\frac{\varepsilon_p^2}{1-\varepsilon_p}\frac{r}{2}\right), \quad \delta_q \leq \exp\left(-\frac{\varepsilon_q^2}{1+\varepsilon_q}\frac{r}{3c}\right) \\
&\alpha = 1 - \frac{1-\varepsilon_p}{1+\varepsilon_q}\frac{1}{c}, \quad \hat{q} \geq \hat{p}^{\frac{1+\alpha}{1-\alpha}}.
\end{aligned}$$

We assume that $0 < q < p < 1$ and furthermore, without loss of generality we can assume that $q \leq 1/e$ due to the powering technique (see Lemma 1.4(a)). In our derivations we also assume that $\delta_p \leq 1/2$ and $\delta_q \leq 1/(2e)$ such that $\delta_q/q \leq 1/2$. This will later be implicit in the statement of the result in big-O notation. From our assumptions and standard bounds on the natural logarithm we are able to derive the following:

$$\begin{aligned}
\frac{\ln(1/p)}{\ln(1/q)} &\geq \frac{\ln(1-\delta_p)\ln(1/\hat{p})}{\ln(1/q)} \geq \frac{\ln(1/\hat{p})}{\ln(1/q)} - 2\delta_p \\
&\geq \frac{\ln(1/\hat{p})}{\ln(1+\delta_q/q) + \ln(1/\hat{q})} - 2\delta_p \\
&\geq \frac{\ln(1/\hat{p})}{\ln(1/\hat{q})}\left(1 - \frac{\ln(1+\delta_q/q)}{\ln(1/\hat{q})}\right) - 2\delta_p \\
&\geq \frac{\ln(1/\hat{p})}{\ln(1/\hat{q})} - \frac{\ln(1+\delta_q/q)}{\ln(1/(1+\delta_q/q)q)} - 2\delta_p \qquad (3) \\
&\geq \frac{\ln(1/\hat{p})}{\ln(1/\hat{q})} - 2\delta_q/q - 2\delta_p.
\end{aligned}$$

In equation (3) we use the statement itself combined with our assumptions on p and q to deduce that

$$1 > \frac{\ln(1/p)}{\ln(1/q)} \geq \frac{\ln(1/\hat{p})}{\ln(1/\hat{q})}.$$

We proceed by lower bounding $\hat{\rho}$. Temporarily define $1 - \varepsilon' = \frac{1-\varepsilon_p}{1+\varepsilon_q}$ and observe that

$$\begin{aligned}
\frac{\ln(1/\hat{p})}{\ln(1/\hat{q})} &\geq \frac{1-\alpha}{1+\alpha} = \frac{(1-\varepsilon')/c}{2-(1-\varepsilon')/c} \\
&\geq \frac{1}{2c-1} - \frac{\varepsilon'}{(2c-1)^2} - \frac{\varepsilon'}{2c-1}.
\end{aligned}$$

We have that

$$\varepsilon' = 1 - \frac{1-\varepsilon_p}{1+\varepsilon_q} = \frac{1+\varepsilon_q-(1-\varepsilon_p)}{1+\varepsilon_q} \leq \varepsilon_q + \varepsilon_p,$$

and combining these bounds results in

$$\frac{\ln(1/p)}{\ln(1/q)} \geq \frac{1}{2c-1} - 2(\varepsilon_q + \varepsilon_p - \delta_q/q - \delta_p).$$

We can now set $\varepsilon_q = \varepsilon_p = K \cdot \sqrt{(c/r)\ln(1/q)}$ for some universal constant K to obtain Theorem 3.7. □

3.2 Lower bound for asymmetric LSH

We can re-apply the techniques behind Lemma 3.5 and Theorem 3.7 to state similar results in the other direction where for $\alpha_- < \alpha_+$ we are interested in upper bounding $f(\alpha_+)$ as a function of $f(\alpha_-)$. This is similar to the well-studied problem of constructing LSH lower bounds and our results match known LSH bounds [12, 37], indicating that the asymmetry afforded by $\mathcal{D}$ does not help us when we wish to construct similarity-sensitive families with monotonically increasing CPFs. Implicitly, this result already follows from the space-time tradeoff lower bounds for similarity search shown independently by Andoni et al. [10] and Christiani [21]. As with Lemma 3.5, the following theorem by O'Donnell [39] is the foundation of our lower bounds.

Theorem 3.9 (Gen. Small-Set Expansion). *Let $0 \le \alpha \le 1$. Let $A, B \subseteq \{0,1\}^d$ have volumes $\exp(-a^2/2)$, $\exp(-b^2/2)$ and assume $0 \le \alpha b \le a \le b$. Then,*

$$\Pr_{\substack{(\mathbf{x},\mathbf{y}) \\ \alpha\text{-corr.}}}[\mathbf{x} \in A, \mathbf{y} \in B] \ge \exp\left(-\frac{1}{2}\frac{a^2 - 2\alpha ab + b^2}{1-\alpha^2}\right).$$

Lemma 3.10. *For every $0 \le \alpha < 1$ and for every distribution $\mathcal{D}$ over pairs of functions $h, g \colon \{0,1\}^d \to R$, we have $\hat{f}(\alpha) \le \hat{f}(0)^{\frac{1-\alpha}{1+\alpha}}$.*

We are now ready to state the corresponding result for similarity-sensitive families.

Theorem 3.11. *Let $0 < \alpha_- < \alpha_+ < 1$ be constants. Then every $(\alpha_-, \alpha_+, f_-, f_+)$-sensitive family $\mathcal{D}$ for Hamming space $(\{0,1\}^d, \mathrm{sim}_H)$ must satisfy*

$$\frac{\log(1/f_+)}{\log(1/f_-)} \ge \frac{1-\alpha_+}{1+\alpha_+-2\alpha_-} - O(\sqrt{\log(1/f_-)/d}).$$

4 HAMMING AND EUCLIDEAN SPACE DSH

4.1 Anti-LSH construction in Hamming space

Bit-sampling [32] is one of the simplest LSH families for Hamming space, yet gives optimal ρ_+-values in terms of the approximation factor [40]. Its CPF is $f(t) = 1 - t$, where t is the relative Hamming distance. By using a function pair $(\mathbf{x} \mapsto x_i, \mathbf{x} \mapsto 1 - x_i)$ where $i \in \{1, \ldots, d\}$ is random, we get a simple asymmetric DSH family for Hamming space whose CPF $f(t) = t$ is monotonically increasing in the relative Hamming distance. We refer to this specific family as *anti bit-sampling*. For anti bit-sampling, we get that $\rho_- = \ln f(r)/\ln f(r/c) = \Omega(1/\ln c)$ as soon as the relative Hamming distance $r \in [0,1]$ is smaller than $1/e$. Perhaps surprisingly, anti bit-sampling is not optimal and a better result, with $\rho_- = O(1/c)$, follows by the anti-LSH schemes based on cross-polytope hashing and filters for the unit sphere in the following subsection. Similarly, the DSH construction for Euclidean space in section 4.2 gives a value of $O(1/c)$ for ρ_-.

4.2 A DSH construction in Euclidean space

A simple and elegant DSH family in Euclidean space is given by a natural extension of the LSH family introduced by Datar et al. [23], where we project a point onto a line and split this line up into buckets. Let k and w be two suitable parameters to be chosen below. Consider the family $\mathcal{R}_{k,w}$ of pairs of functions (h, g) defined in equation (2), indexed by a uniform real number $b \in [0, w]$ and a d-dimensional random Gaussian vector $\mathbf{a} \sim \mathcal{N}^d(0,1)$. We have the following result whose proof is provided in Appendix B:

Theorem 4.1. *Let r_- and r be two real values such that $0 < r_- < r$, and let $c = r/r_-$. Then there exists a constant $w = w(c)$ such that for each k the family $\mathcal{R}_{k,w}$ satisfies*

$$\rho_- = \frac{\ln(1/f(r))}{\ln(1/f(r_-))} = \frac{1}{c^2}\left(1 + O_k(1/k)\right).$$

The proof uses that for $\mathbf{a} \sim \mathcal{N}^d(0,1)$ the inner product $\langle \mathbf{a}, (\mathbf{x}-\mathbf{y})\rangle$ is distributed as $\mathcal{N}(0, \Delta)$ for two points $\mathbf{x}$ and $\mathbf{y}$ at distance Δ. For the hash values of $\mathbf{x}$ and $\mathbf{y}$ to collide, the inner product must roughly lie in the interval $[(k-1)w, (k+1)w]$. Because we are free to choose k and w, we can move the interval into the tail of the $\mathcal{N}(0, \Delta)$ distribution, where the target distances r and r/c have quadratic influence in the exponents.

5 GENERAL CONSTRUCTIONS

So far we have focused our attention on constructions with monotone CPFs, which just represent one kind of DSH schemes. It is natural to wonder if more advanced CPFs can be obtained. In this section, we provide some results in this direction by describing two constructions yielding a wide class of CPFs. We remark that these general constructions do not seem to provide any improved constructions in the monotone case.

Angular similarity functions. We say that $\mathrm{sim} \colon [-1,1] \to [0,1]$ is an *LSHable angular similarity function* if there exists an hash family $\mathcal{S}$ with collision probability function $\mathrm{sim}(\langle \mathbf{x}, \mathbf{y}\rangle)$ for each $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$. For example, the function $\mathrm{sim}(t) = 1 - \arccos(t)/\pi$ is LSHable using the *SimHash* construction of Charikar [17].

Valiant [51] described a pair of mappings $\varphi_1^{\mathcal{P}}, \varphi_2^{\mathcal{P}} \colon \mathbb{R}^d \to \mathbb{R}^D$, where $D = O(d^k)$, such that $\varphi_1^{\mathcal{P}}(\mathbf{x}) \cdot \varphi_2^{\mathcal{P}}(\mathbf{y}) = \mathcal{P}(\langle \mathbf{x}, \mathbf{y}\rangle)$, for any polynomial $\mathcal{P}(t) = \sum_{i=0}^k a_i t^i$. By leveraging this construction we get the following result (with proof provided in Appendix C.2).

Theorem 5.1. *Suppose that sim is an LSHable angular similarity function and that the polynomial $\mathcal{P}(t) = \sum_{i=0}^k a_i t^i$ satisfies $\sum_{i=0}^k |a_i| = 1$. Then there exists a distribution over pairs (h, g) of functions such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$, $\Pr[h(\mathbf{x}) = g(\mathbf{y})] = \mathrm{sim}(\mathcal{P}(\langle \mathbf{x}, \mathbf{y}\rangle))$.*

The computational cost of a naïve implementation of the proposed scheme may be prohibitive when d^k is large. However, by using the so-called *kernel approximation* methods [42], we can in near-linear time compute approximations $\hat{\varphi}_1^{\mathcal{P}}(\mathbf{x})$ and $\hat{\varphi}_2^{\mathcal{P}}(\mathbf{y})$ that satisfy $\hat{\varphi}_1^{\mathcal{P}}(\mathbf{x}) \cdot \hat{\varphi}_2^{\mathcal{P}}(\mathbf{y}) = \mathcal{P}(\langle \mathbf{x}, \mathbf{y}\rangle) \pm \varepsilon$ with high probability for a given approximation error $\varepsilon > 0$.

Hamming distance functions. It is natural to wonder which CPFs can be expressed as a function of the relative Hamming distance $d_h(\mathbf{x}, \mathbf{y})$. A first answer follows by using the anti bit-sampling approach from section 4 together with Lemma 1.4. This gives a scheme for matching any polynomial $\mathcal{P}(t) = \sum_{i=0}^k a_i t^i$ that satisfies $\sum_{i=0}^k a_i = 1$ and $a_i > 0$ for each i.

In this section, we provide another construction that matches, up to a scaling factor Δ, any polynomial $\mathcal{P}(t)$ having no roots with

a real part in $(0, 1)$. The scaling factor depends only on the roots of the polynomial. We claim that such a factor is unavoidable in the general case: indeed, without Δ, it would be possible to match the CPF $1 - t^2$ for Hamming space, which implies $\rho \leq 1/c^2$ in contradiction with the lower bound $1/c$ in [40]. (Nevertheless, it is an open question to assess how tight Δ is.) We have the following result that is proven in Appendix C.3:

Theorem 5.2. *Let $\mathcal{P}(t) = \sum_{i=0}^{k} a_i t^i$, Z be the multiset of roots of $\mathcal{P}(t)$, and $\psi \leq k$ be the number of roots with negative real part. Then there exists a DSH family with collision probability $\Pr(h(\mathbf{x}) = g(\mathbf{y})) = \mathcal{P}(d_h(\mathbf{x}, \mathbf{y}))/\Delta$ with $\Delta = a_k 2^{\psi} \prod_{z \in Z, |z|>1} |z_i|$.*

The construction exploits the factorization $\mathcal{P}(t) = a_k \prod_{z \in Z}(t - z)$ and consists of a combination of $|Z|$ variations of bit-sampling and anti bit-sampling. Although the proposed scheme may not reach the ρ-value given by the polynomial $\mathcal{P}(t)$, it can be used for estimating $\mathcal{P}(d_H(\mathbf{x}, \mathbf{y}))$ since the scaling factor is constant and only depends on the polynomial.

Finally, we observe that our scheme can be used to approximate any function $f(t)$ that can be represented with a Taylor series: indeed, it is sufficient to truncate the series to the term that gives the desired approximation, and then to apply our construction to the resulting truncated polynomial.

6 APPLICATIONS

6.1 Hyperplane queries and annulus search

Approximate annulus search is the problem of finding a point in the set P of data points with distance in an interval $[r_-, r_+]$ from a query point. On the unit sphere, hyperplane queries are a special case of annulus queries where we want to find a point with inner product close to 0 to a query point. This type of search has applications in machine learning (see [33, 36]).

The ad hoc solution for this problem [41] in Euclidean space works by first building an LSH data structure that aims to retrieve points at distance at most r while filtering out points at distance at least r_+. In each repetition and in each bucket of the hash table, one builds a data structure that is set up to filter away points at distance at least r_- while preserving points at distance at least r. The latter filtering can be thought of as applying an anti-LSH in each bucket. For $r_- = r/c$ and $r_+ = cr$ for a $c \geq 1$, the construction in [41] answers queries in time $\tilde{O}(dn^{\rho+1/c^2})$ and the data structure uses space $\tilde{O}(n^{1+\rho+1/c^2} + dn)$, where ρ describes the collision probability gap of the LSH that is used, disregarding logarithmic terms by using the $\tilde{O}(.)$ notation.

Having access to a DSH family with a CPF that peaks inside $[r_-, r_+]$ and is significantly smaller at the ends of the interval gives an LSH-like solution to this problem.

Theorem 6.1. *Suppose we have a set P of n points, an interval $[r_-, r_+]$, a distance $r \in [r_-, r_+]$, and assume we are given a DSH scheme with a CPF f that peaks inside $[r_-, r_+]$ and satisfies $f(r') \leq 1/n$ for all $r' \notin [r_-, r_+]$. Then there exists a data structure that, given a query $\mathbf{q}$ for which there exists $\mathbf{x} \in P$ with $\operatorname{dist}(\mathbf{q}, \mathbf{x}) = r$, returns $\mathbf{x}' \in P$ with $\operatorname{dist}(\mathbf{q}, \mathbf{x}') \in [r_-, r_+]$ with probability at least $1/2$. The data structure uses space $O(n^{1+\rho^*}/f(r) + dn)$ and has query time $O(dn^{\rho^*})$, where $\rho^* = \log(1/f(r))/\log n$.*

Proof. The data structure is a straightforward adaptation of the construction of a near neighbor data structure using LSH. Associate with each data point $\mathbf{x}$ and query point $\mathbf{y}$ the hash values $h(\mathbf{x})$ and $g(\mathbf{y})$, where (h, g) are independently sampled from the distance-sensitive family. Store all points $\mathbf{x} \in S$ according to $h(\mathbf{x})$ in a hash table. Let $\mathbf{y}$ be the query point and let $\mathbf{x}$ be a point at distance r. Compute $g(\mathbf{y})$ and retrieve all the points from S that have the same hash value. If a point within distance $[r_-, r_+]$ is among the points, output one such point. We expect $\max\{f(r_-)n, f(r_+)n\} \leq 1$ collisions with points at distance at most r_- or at least r_+. The probability of finding $\mathbf{x}$ is at least $f(r)$. Thus, $L = 1/f(r) \leq n^{\rho^*}$ repetitions suffice to retrieve $\mathbf{x}$ with constant probability $1/e$. If the algorithm retrieves more than $8L$ points, none of which is in the interval $[r_-, r_+]$, the algorithm terminates. By Markov's inequality, the probability that the algorithm retrieves $8L$ points, none of which is in the interval $[r_-, r_+]$, is at most $1/8$. □

We note that the assumption $f(r_+), f(r_-) \leq 1/n$ in the theorem is not critical: the standard technique of *powering* (see Lemma 1.4(a)) allows us to work with the CPF $f(x)^k$ for integer k, where k is the smallest integer such that $f(x)^k \leq 1/n$.

We observe that this data structure improves the trivial scanning solution when $\rho^* = \log(1/f(r))/\log n < 1$, that is when $f(r) > f(r_-)$ and $f(r) > f(r_+)$. This is satisfied by *unimodal* distance-sensitive hash families, that is when the CPF has a single maximum at t^* and is decreasing for both $t \leq t^*$ and $t \geq t^*$: as soon as t^* lies in the interval (r_-, r_+) we obtain a data structure with sublinear query time.

Obtaining a CPF that peaks inside of $[r_-, r_+]$ can be achieved by combining a standard LSH family $\mathcal{H}$ with a DSH family $\mathcal{A}$ that has an increasing CPF by means of powering each part. For example, when combining a bit-sampling and an anti bit-sampling family, concatenating k_1 bit-sampling and k_2 anti bit-sampling results in the CPF $f(t) = (1 - t)^{k_1} t^{k_2}$. Setting $k_1 = k_2(1 - t)/t$ results in f peaking at distance r. In general, for the value ρ^* in the statement of the theorem, this approach yields a bound of $\rho^* \leq \rho_+ + \rho_-$, where ρ_+ and ρ_- are the ρ-values of $\mathcal{H}$ and $\mathcal{A}$, respectively. This is essentially the same as the running time guarantee of the ad hoc data structure in [41] in Euclidean space.

In section 6.2, we show how to combine the two monotonic constructions from section 2.2 on the unit sphere. A key ingredient of the construction is that we do not need the powering approach described above but can combine a single LSH with a single anti-LSH function and set the thresholds of each part accordingly. With respect to hyperplane queries under inner product similarity, the resulting construction allows us to search a point set P of unit vectors for a vector approximately orthogonal to a query vector $\mathbf{q}$ in time $dn^{\rho^*+o(1)}$ for $\rho^* = \frac{1-\alpha^2}{1+\alpha^2}$, where we guarantee to return a vector $\mathbf{x}$ with $\langle \mathbf{x}, \mathbf{q} \rangle \in [-\alpha, \alpha]$ if an orthogonal vector exists. When applied to search for approximately orthogonal vectors, our technique improves ρ-values of previous techniques, but the improvement is not surprising in view of recent progress in angular LSH, see e.g. [8]. However, Theorem 6.4 in Section 6.2 supports searching a wide range of different annuli and not just the ones centered around vectors with zero correlation. As an additional result, we obtain a definition of an annulus in a space with bounded distances.

6.2 Annulus search on the unit sphere

We will construct a distance sensitive family $\mathcal{D}$ for solving the approximate annulus search problem. Let $\mathcal{D}_+$ be parameterized by t_+ and let $\mathcal{D}_-$ be parameterized by t_-. To sample a pair of functions (h, g) from $\mathcal{D}$ we independently sample a pair (h_+, g_+) from $\mathcal{D}_+$ and (h_-, g_-) from $\mathcal{D}_-$ and define (h, g) by $h(\mathbf{x}) = (h_+(\mathbf{x}), h_-(\mathbf{x}))$ and $g(\mathbf{x}) = (g_+(\mathbf{x}), g_-(\mathbf{x}))$.

Let $f(\alpha)$ denote the CPF of $\mathcal{D}$. We would like to be able to parameterize $\mathcal{D}$ such that $f(\alpha)$ is somewhat symmetric around a unique maximum value of α. It can be verified from the definition of $\mathcal{D}_+$ that $p_+(-1) = 0$ which implies that $f(-1) = f(1) = 0$. If we ignore lower order terms and define $\gamma > 0$ by $t_- = \gamma t_+$, then we can see that

$$\ln(1/f(\alpha)) \approx \frac{1-\alpha}{1+\alpha}\frac{t_+^2}{2} + \frac{1+\alpha}{1-\alpha}\frac{\gamma^2 t_+^2}{2}.$$

For simplicity, temporarily define $a(\alpha) = (1-\alpha)/(1+\alpha) > 0$. Given a fixed γ, the equation $a + \gamma^2/a$ is minimized (corresponding to approximately maximizing $f(\alpha)$) when setting $a = \gamma$. Let $\alpha_{\max} \in (-1, 1)$ and set $\gamma = a_{\max} = (1-\alpha_{\max})/(1+\alpha_{\max})$. To find values $\alpha_- < \alpha_{\max} < \alpha_+$ where $\ln(1/f(\alpha_-)) \approx \ln(1/f(\alpha_+))$ note that this condition holds for every $s > 1$ when we set $a_- = s a_{\max}$ and $a_+ = (1/s)a_{\max}$. We therefore parameterize $\mathcal{D}$ by $\alpha_{\max} \in (-1, 1)$ and $t > 0$ and set $t_+ = t$ and $t_- = (1-\alpha_{\max})/(1+\alpha_{\max})t_+$. By combining our bounds from Lemma A.5 with the above observations we are able to obtain the following theorem which immediately yields a solution to the approximate annulus search problem.

Theorem 6.2. *For every choice of $t > 0$ and constant $\alpha_{\max} \in (-1, 1)$ the family $\mathcal{D}$ satisfies the following: For every choice of constant $s > 1$ consider the interval $[\alpha_-, \alpha_+]$ defined to contain every α such that $\frac{1}{s}\frac{1-\alpha_{\max}}{1+\alpha_{\max}} \le \frac{1-\alpha}{1+\alpha} \le s\frac{1-\alpha_{\max}}{1+\alpha_{\max}}$, then*

- *For $\alpha \in [\alpha_-, \alpha_+]$ we have that*

$$f(\alpha) = \Omega\left((1/t^2)\exp\left(-(s+1/s)\frac{1-\alpha_{\max}}{1+\alpha_{\max}}\frac{t^2}{2}\right)\right).$$

- *For $\alpha \notin [\alpha_-, \alpha_+]$ we have that*

$$f(\alpha) = O\left((1/t^2)\exp\left(-(s+1/s)\frac{1-\alpha_{\max}}{1+\alpha_{\max}}\frac{t^2}{2}\right)\right).$$

The complexity of sampling, storing, and evaluating a pair of functions $(h, g) \in \mathcal{D}$ is $O(dt^4 e^{t^2/2})$.

See Figure 3 for a visual representation of the annulus for given parameters $\alpha_{\max}$ and s.

We define an approximate annulus search problem for similarity spaces and proceed by applying Theorem 6.2 to provide a solution for the unit sphere, resulting in Theorem 6.4.

Definition 6.3. Let $\beta_- < \alpha_- \le \alpha_+ < \beta_+$ be given real numbers. For a set P of n points in a similarity space (X, sim) a solution to the $((\alpha_-, \alpha_+), (\beta_-, \beta_+))$-annulus search problem is a data structure that supports a query operation that takes as input a point $\mathbf{x} \in X$ and if there exists a point $\mathbf{y} \in P$ such that $\alpha_- \le \mathrm{sim}(\mathbf{x}, \mathbf{y}) \le \alpha_+$ then it returns a point $\mathbf{y}' \in P$ such that $\beta_- \le \mathrm{sim}(\mathbf{x}, \mathbf{y}') \le \beta_+$.

Theorem 6.4. *For every choice of constants $-1 < \beta_- < \alpha_- < \alpha_+ < \beta_+ < 1$ such that $\frac{1-\alpha_-}{1+\alpha_-}\frac{1-\alpha_+}{1+\alpha_+} = \frac{1-\beta_-}{1+\beta_-}\frac{1-\beta_+}{1+\beta_+}$ we can solve the $((\alpha_+, \alpha_-), (\beta_+, \beta_-))$-annulus problem for $(\mathbb{S}^{d-1}, \langle\cdot,\cdot\rangle)$ with space usage $dn + n^{1+\rho+o(1)}$ words and query time $dn^{\rho+o(1)}$ where*

$$\rho = \frac{c_\alpha + 1/c_\alpha}{c_\beta + 1/c_\beta} \le \frac{2}{c + 1/c}$$

and we define $1 < c_\alpha < c_\beta$ by $c_\alpha = \sqrt{\frac{1-\alpha_-}{1+\alpha_-}/\frac{1-\alpha_+}{1+\alpha_+}}$, $c_\beta = \sqrt{\frac{1-\beta_-}{1+\beta_-}/\frac{1-\beta_+}{1+\beta_+}}$, and $c = c_\beta/c_\alpha$.

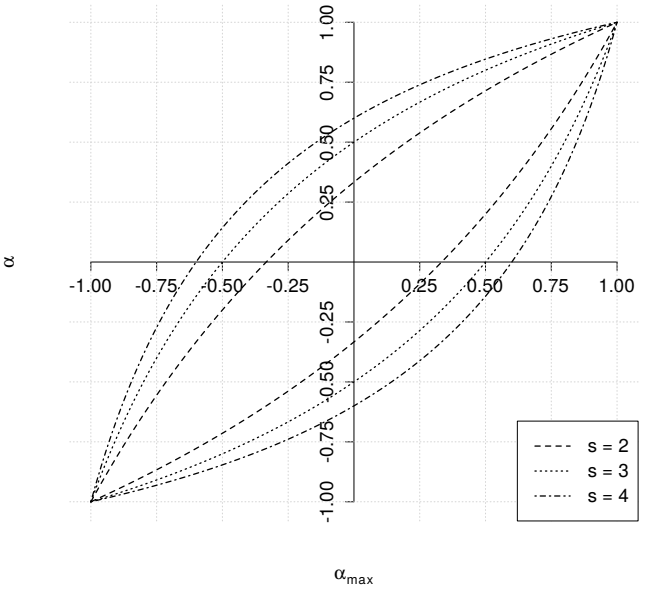

Figure 3: Annuli as defined in Theorem 6.2 for every value of $\alpha_{\max}$ and $s = 2, 3, 4$.

6.3 Spherical range reporting

Approximate spherical range reporting [4] aims to report all points in P within distance r from a query point. A common problem with LSH-based solutions for reporting all close points is that the CPF is monotonically decreasing starting with collision probability very close to 1 for points that are very close to the query point. On the other hand, many repetitions are necessary to find points at the target distance r. This means that the algorithm retrieves many duplicates for solving range reporting problems. The state-of-the-art data structure for range reporting queries [4] requires $O((1 + |S^*|)(n/|S^*|)^\rho)$, where S^* is the set of points at distance at most r_+.

CPFs that have a (roughly) fixed value in $[0, r]$ and then decrease rapidly to zero (so-called "step-function CPFs") yield data structures with good *output sensitivity*.

Theorem 6.5. *Suppose we have a set P of n points and two distances $r < r_+$. Assume we are given a DSH scheme with CPF f where $f(r') \le 1/n$ for all $r' \ge r_+$, and let $f_{\min} = \inf_{t\in[0,r]} f(t)$, $f_{\max} = \sup_{t\in[0,r]} f(t)$. Then there exists a data structure that, given a query $\mathbf{q}$, returns $S \subseteq \{\mathbf{x} \in P \mid \mathrm{dist}(\mathbf{q}, \mathbf{x}) \le r_+\}$ such that for each*

$\mathbf{x} \in P$ *with* $\text{dist}(\mathbf{q}, \mathbf{x}) \leq r$, $\Pr[\mathbf{x} \in S] > 1/2$. *The data structure uses space* $O(n^{1+\rho^*} + dn)$ *and the query has expected running time* $O(dn^{\rho^*} + d|S|f_{\max}/f_{\min})$, *where* $\rho^* = \log(1/f_{\min})/\log(1/f(r_+))$.

Proof. We assume that we build a standard LSH data structure as in the proof of Theorem 6.1 above. We use $1/f_{\min}$ repetitions such that each point within distance r is found with constant probability. Each repetition will contribute $O(1 + |S^*|f_{\max})$ points in expectation. Thus, the total cost will be $O((1 + |S^*|f_{\max})/f_{\min})$ from which the statement follows. □

In short, Theorem 6.5 provides a better analysis of the performance of a standard LSH data structure that takes into account the gap between $f_{\min}$ and $f_{\max}$. Again, the assumption $f(r_+) \leq 1/n$ in the theorem is not critical.

In particular, the theorem shows that if we have a constant bound on $f_{\max}/f_{\min}$ the output sensitivity is optimal in the sense that the time to report an additional close point is $O(d)$ which is the time it takes to verify its distance to the query point. Such step-function CPFs are implicit in the linear space extremes of the space-time tradeoff techniques for near neighbor search [10, 21]. To see why this is the case consider a randomized (list-of-points [10]) data structure that solves the approximate near neighbor problem using linear space. The data structure stores each point in exactly one bucket. During a query L different buckets are searched. The data structure has the property that near neighbors collide in at least one bucket with constant positive probability. We can now construct a family $\mathcal{D}$ where we sample $(h, g) \sim \mathcal{D}$ by sampling a random (data-independent) data structure. We set $h(\mathbf{x})$ equal to the index of the single bucket that $\mathbf{x}$ would be stored in during an update, and $g(\mathbf{x})$ is set to the index of a random bucket out of the L buckets that would be searched during a query for $\mathbf{x}$. If the data structure guarantees finding an r-near neighbor, then the probability of collision is $\Theta(1/L)$ for points $\mathbf{x}, \mathbf{y}$ with $\text{dist}(\mathbf{x}, \mathbf{y}) \leq r$. Even though this theoretical construction gives optimal output sensitivity, it is possible that a better value of ρ^* can be obtained by allowing a higher space usage.

6.4 Privacy-Preserving Distance Estimation

Consider a database consisting of private information, for example medical histories of patients, encoded as points in a space. We would like to be able to search the database for points that are similar to a given query point $\mathbf{q}$ without revealing sensitive information about the data points, except whether there exists a point within a given distance r from $\mathbf{q}$. This is nontrivial even in the case of a single point, so we focus on the distance estimation problem: Is the distance between points $\mathbf{q}$ and $\mathbf{x}$ at most r or not? Given $\mathbf{x}$ we would like to answer this while revealing as little information as possible about $\mathbf{x}$. Secure multi-party computations can be used for this (see e.g. [28]), but such protocols are not practical in general. In contrast, the intersection of two sets can be computed efficiently and in a private way that does not reveal anything about the items not in the intersection [24, 26, 43].

We are going to allow false positives and approximate false negatives as follows. For an approximation factor $c > 1$, and parameters $\varepsilon, \delta > 0$:

- If $\mathbf{q}$ and $\mathbf{x}$ have distance at most r, we say "Yes" with probability at least $1 - \varepsilon$.
- If $\mathbf{q}$ and $\mathbf{x}$ have distance at least cr, we say "No" with probability at least $1 - \delta$.

Our approach is to reduce this problem to private set intersection, as follows: For a parameter t to be chosen later, pick a DSH family $\mathcal{H}$ with a step-function CPF with collision probability $\Theta(1/t)$ at distances in $[0, r]$, and let $\rho > 0$ be the smallest constant such that the collision probability for distances larger than cr is $O(t^{-1/\rho})$. Without loss of generality we may assume that hash values are $O(\log t)$ bits (if not, hash them to this number of bits using universal hashing, increasing the collision probability only slightly). Now generate a sequence of $O(t \log(1/\varepsilon))$ hash functions pairs $(h_1, g_1), (h_2, g_2), \cdots \sim \mathcal{H}$, independently and consider the vectors $(h_1(\mathbf{x}), h_2(\mathbf{x}), \dots)$ and $(g_1(\mathbf{x}), g_2(\mathbf{x}), \dots)$. By definition the expected (component-wise) intersection size of the two vectors, i.e., the expected number of hash collisions, is $O(\log(1/\varepsilon))$. This means that the intersection of the two vectors reveals $O(\log(1/\varepsilon) \log t)$ bits of information about the vectors in expectation. Also, by adjusting constants, the probability that the intersection size is 0 when $\mathbf{q}$ and $\mathbf{x}$ are close is at most ε. Thus, saying "Yes" if and only if the intersection is nonempty satisfies the first condition. On the other hand, by a union bound the probability of a "Yes" when $\mathbf{q}$ and $\mathbf{x}$ have distance at least cr is $\delta = O(t \log(1/\varepsilon)/t^{1/\rho})$. Conversely, we need $t \approx (1/\delta)^{\rho/(1-\rho)}$ to have false positive probability δ.

We observe that our approach has a stronger privacy constraint than the result in [45]. Indeed, the step-function DSH preserves privacy even if the $\mathbf{q}$ and $\mathbf{x}$ points are very close (e.g., $\mathbf{q} = \mathbf{x}$), since the collision probability is almost equal in the range $[0, r]$. On the other hand, a standard LSH, as the one adopted in [45], has an high collision rate when the points are very close, revealing information on near points.

7 CONCLUSION

We have initiated the study of *distance-sensitive hashing*, an asymmetric class of hashing methods that considerably extend the capabilities of standard LSH. We proposed different constructions of such hash families and described some applications. Interestingly, DSHs provide an unique framework for capturing problems that have been separately studied, like nearest neighbor [32], finding orthogonal vectors [52], furthest point query [41], and privacy preserving search [45].

Though we settled some basic questions regarding what is possible using DSH, many questions remain. Ultimately, one would like for a given space a complete characterization of the CPFs that can be achieved, with emphasis on extremal properties. For example: For a CPF that has $f(x) = \Theta(\varepsilon)$ for $x \in [0, r]$, how small a value $\rho_+(c) = \log(f(r))/\log(f(cr))$ is possible outside of this range? Additionally, our solution to the annulus problem works by combining an LSH and an anti-LSH family to obtain a unimodal family: While we know lower bounds for both, it is not clear whether combining them yields optimal solutions for this problem. Finally, it is also of interest to consider other applications of DSH in privacy preserving search and in kernel density estimation (e.g. [35]).

ACKNOWLEDGMENTS

The authors thank Thomas D. Ahle for insightful conversations. The research leading to these results has received funding from the European Research Council under the European Union's 7th Framework Programme (FP7/2007-2013) / ERC grant agreement no. 614331. Silvestri has also been supported by project SID2017 of the University of Padova. BARC, Basic Algorithms Research Copenhagen, is supported by the VILLUM Foundation grant 16582.

A MONOTONE DSH CONSTRUCTIONS

A.1 Optimal monotone DSH for the unit sphere

We initially bound the CPF of the family $\mathcal{D}_+$, which translates to a bound for $\mathcal{D}_-$ through the following observation:

Lemma A.1. *Given $\mathcal{D}_+$ and $\mathcal{D}_-$ with identical parameters, we have that $f_+(\alpha) = f_-(-\alpha)$.*

Proof. A bivariate normally distributed variable with correlation α can be represented as a pair (X, Y) with $X = Z_1$ and $Y = \alpha Z_1 + \sqrt{1-\alpha^2}Z_2$ where Z_1, Z_2 are i.i.d. standard normal. By the symmetry of the standard normal distribution around zero it is straightforward to verify that $\Pr[Z_1 \geq t \wedge \alpha Z_1 + \sqrt{1-\alpha^2}Z_2 \geq t] = \Pr[Z_1 \geq t \wedge -\alpha Z_1 + \sqrt{1-\alpha^2}Z_2 \leq -t]$. □

The collision probability for $(h, g) \sim \mathcal{D}_+$ depends only on the inner product $\alpha = \langle \mathbf{x}, \mathbf{y} \rangle$ between the pair of points being evaluated and is given by

$$\Pr[h(\mathbf{x}) \leq m \vee g(\mathbf{y}) \leq m] \frac{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \wedge \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \vee \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}.$$

To see why this is the case first note that it is only possible that $h(\mathbf{x}) = g(\mathbf{y})$ in the event that $h(\mathbf{x}) \leq m \vee g(\mathbf{y}) \leq m$. Conditioned on this happening, consider the first i such that either $\langle \mathbf{z}_i, \mathbf{x} \rangle \geq t$ or $\langle \mathbf{z}_i, \mathbf{y} \rangle \geq t$. Now the probability of collision is given by $\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \wedge \langle \mathbf{z}, \mathbf{y} \rangle \geq t | \langle \mathbf{z}, \mathbf{x} \rangle \geq t \vee \langle \mathbf{z}, \mathbf{y} \rangle \geq t]$.

To bound the CPF of $\mathcal{D}_+$ and $\mathcal{D}_-$, we use the following tail bounds for the standard normal distribution and the tail bounds by Savage [47] for the bivariate standard normal distribution.

Lemma A.2 (Follows Szarek & Werner [50]). *Let Z be a standard normal random variable. Then, for every $t \geq 0$ we have that*

$$\frac{1}{\sqrt{2\pi}} \frac{1}{t+1} e^{-t^2/2} \leq \Pr[Z \geq t] \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2}.$$

Lemma A.3 (Savage [47]). *Let $\alpha \in (-1, 1)$ and let $Z_1, Z_2 \sim \mathcal{N}(0, 1)$. Define $X_1 = Z_1$ and $X_2 = \alpha Z_1 + \sqrt{1-\alpha^2}Z_2$. Then, for every $t > 0$ we have that*

$$\left(1 - \frac{(2-\alpha)(1+\alpha)}{1-\alpha} \frac{1}{t^2}\right) \frac{1}{2\pi t^2} \frac{(1+\alpha)^2}{\sqrt{1-\alpha^2}} \exp\left(-\frac{t^2}{1+\alpha}\right)$$
$$< \Pr[X_1 \geq t \wedge X_2 \geq t] < \frac{1}{2\pi t^2} \frac{(1+\alpha)^2}{\sqrt{1-\alpha^2}} \exp\left(-\frac{t^2}{1+\alpha}\right)$$

Corollary A.4. *By symmetry of the normal distribution the Lemma A.3 bounds apply to $\Pr[X_1 \geq t \wedge X_2 \leq -t]$ when we replace all occurrences of α with $-\alpha$.*

We are now ready to bound the CPF for $\mathcal{D}_+$.

Lemma A.5. *For every $t > 0$ and $\alpha \in (-1, 1)$ the family $\mathcal{D}_+$ satisfies*

$$f_+(\alpha) < \bar{f}_+(\alpha) := \frac{1}{\sqrt{2\pi}} \frac{t+1}{t^2} \frac{(1+\alpha)^2}{\sqrt{1-\alpha^2}} \exp\left(-\frac{1-\alpha}{1+\alpha} \frac{t^2}{2}\right),$$
$$f_+(\alpha) > \left(1 - \frac{(2-\alpha)(1+\alpha)}{1-\alpha} \frac{1}{t^2}\right) \frac{t}{t+1} \bar{f}_+(\alpha) - 2e^{-t^3}.$$

The complexity of sampling, storing, and evaluating a pair of functions $(h, g) \in \mathcal{D}_+$ is $O(dt^4 e^{t^2/2})$.

Proof. We proceed by deriving upper and lower bounds on the collision probability.

$$f_+(\alpha) \leq \frac{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \wedge \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t]}$$
$$\leq \frac{1}{\sqrt{2\pi}} \frac{t+1}{t^2} \frac{(1+\alpha)^2}{\sqrt{1-\alpha^2}} \exp\left(-\frac{1-\alpha}{1+\alpha} \frac{t^2}{2}\right).$$

We derive the lower bound in stages.

$$\Pr[h(\mathbf{x}) = g(\mathbf{y})]$$
$$\geq (1 - \Pr[h(\mathbf{x}) > m]) \frac{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \wedge \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \vee \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}$$
$$\geq \frac{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \wedge \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}{2 \Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t]} - \Pr[h(\mathbf{x}) > m].$$

The first part is lower bounded by

$$\frac{\Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t \wedge \langle \mathbf{z}, \mathbf{y} \rangle \geq t]}{2 \Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t]} \geq \left(1 - \frac{(2-\alpha)(1+\alpha)}{1-\alpha} \frac{1}{t^2}\right) \cdot$$
$$\frac{1}{2\sqrt{2\pi}} \frac{1}{t} \frac{(1+\alpha)^2}{\sqrt{1-\alpha^2}} \exp\left(-\frac{1-\alpha}{1+\alpha} \frac{t^2}{2}\right).$$

The probability of not being captured by a projection depends on the number of projections m. In order to make this probability negligible we can set $m = \lceil 2t^3/p' \rceil$ where p' denotes the lower bound from Lemma A.2.

$$\Pr[h(\mathbf{x}) > m] \leq (1 - \Pr[\langle \mathbf{z}, \mathbf{x} \rangle \geq t])^m \leq (1 - p')^{2t^3/p'} \leq e^{-2t^3}.$$

The bound on the complexity of sampling, storing, and evaluating a pair of functions $(h, g) \in \mathcal{D}_+$ follows from having $m = \lceil 2t^3/p' \rceil = O(t^4 e^{t^2/2})$ standard normal projections of length d to be sampled, stored, and evaluated. □

Combining the above ingredients we get the following results, which implies Theorem 1.2 by Lemma A.1.

Theorem A.6. *For every $t > 1$ there exists a distance-sensitive family $\mathcal{D}_+$ for $(\mathbb{S}^{d-1}, \langle \cdot, \cdot \rangle)$ with a CPF f such that for every $\alpha \in (-1, 1)$ satisfying $|\alpha| < 1 - 1/t$ we have that*

$$\ln(1/f(\alpha)) = \tfrac{1-\alpha}{1+\alpha} \tfrac{t^2}{2} + \Theta(\log t). \tag{4}$$

Furthermore, the CPF of $\mathcal{D}_+$ is monotonically increasing, and the complexity of sampling, storing, and evaluating $(h, g) \in \mathcal{D}_-$ is $O(dt^4 e^{t^2/2})$.

A more careful analysis of the collision probabilities is required in order to combine the families $\mathcal{D}_-$ and $\mathcal{D}_+$ to form a unimodal family that can be used to solve the annulus search problem, see Theorem 6.1. These results are stated in Appendix 6.2.

B DSH FOR EUCLIDEAN SPACE

Proof of Theorem 4.1. For the sake of simplicity we assume $r = 1$ in the analysis (otherwise it is enough to scale down vectors accordingly). Let $\mathbf{x}$ and $\mathbf{y}$ be two points in $\mathbb{R}^d$ with distance Δ. We know that for $\mathbf{a} \sim \mathcal{N}^d(0,1)$ the inner product $\langle \mathbf{a}, (\mathbf{x}-\mathbf{y})\rangle$ is distributed as $\mathcal{N}(0,\Delta)$. A necessary but not sufficient condition to have a collision between $\mathbf{x}$ and $\mathbf{y}$ is that $\langle \mathbf{a}, (\mathbf{x}-\mathbf{y})\rangle$ lies in the interval $[(k-1)w, (k+1)w]$. Now, if $t := \langle \mathbf{a}, (\mathbf{x}-\mathbf{y})\rangle \in [(k-1)w, kw]$, then the random offset b must lie in an interval of length $t-(k-1)w$, putting $\langle \mathbf{a}, \mathbf{x}\rangle$ and $\langle \mathbf{a}, \mathbf{y}\rangle - (k-1)w$ into different buckets. For the interval $[kw, (k+1)w]$ similar observations show that b has to be chosen in an interval of length $(k+1)w - t$. Let $\phi(t) = 1/\sqrt{2\pi}e^{-t^2/2}$ be the density function of a standard normal random variable. Similarly to the calculations in [23], the collision probability at distance Δ can be calculated as follows:

$$
\begin{aligned}
f(\Delta) &= \Pr\left(\left\lfloor \frac{\langle \mathbf{a}, \mathbf{x}\rangle + b}{w} \right\rfloor - \left\lfloor \frac{\langle \mathbf{a}, \mathbf{y}\rangle + b}{w} \right\rfloor = k\right) \\
&= \int_{(k-1)w}^{kw} \frac{\phi(t/\Delta)}{\Delta}\left(\frac{t}{w} - (k-1)\right) dt \\
&\quad + \int_{kw}^{(k+1)w} \frac{\phi(t/\Delta)}{\Delta}\left(k+1-\frac{t}{w}\right) dt - \frac{\phi(kw/\Delta)}{\Delta} \\
&= \frac{1}{\sqrt{2\pi}\Delta}\left(\int_{(k-1)w}^{kw} e^{-\frac{t^2}{2\Delta^2}}\left(\frac{t}{w} - (k-1)\right) dt \right. \\
&\quad \left. + \int_{kw}^{(k+1)w} e^{-\frac{t^2}{2\Delta^2}}\left(k+1-\frac{t}{w}\right) dt - e^{-\frac{(kw)^2}{2\Delta^2}}\right).
\end{aligned}
$$

We now proceed to upper bound ρ^- by finding an upper bound on $f(1/c)$ and a lower bound on $f(1)$. Simple calculations give an upper bound of

$$
f(1/c) \leq \frac{2wc}{\sqrt{2\pi}} e^{-(c(k-1)w)^2/2}. \tag{5}
$$

For the lower bound, we only look at the interval $t \in [kw, (k+1/2)w]$ and obtain the bound:

$$
\begin{aligned}
f(1) &\geq \frac{1}{\sqrt{2\pi}} \int_{kw}^{(k+1/2)w} e^{-\frac{t^2}{2}}\left(k+1-\frac{t}{w}\right) dt \\
&\geq \frac{w}{4\sqrt{2\pi}} e^{-((k+1/2)w)^2/2}.
\end{aligned} \tag{6}
$$

Now we multiply the ratio of the logarithms of the right-hand sides of (5) and (6) with c^2 and proceed to show that this term is bounded by $1 + O(1/k)$, which shows the result. In the following, we set w such that $w \leq \sqrt{2\pi}/(2c)$. We compute:

$$
\begin{aligned}
&\frac{\ln\left(\frac{w}{4\sqrt{2\pi}} e^{-((k+1/2)w)^2/2}\right)}{\ln\left(\frac{2wc}{\sqrt{2\pi}} e^{-(c(k-1)w)^2/2}\right)} c^2 \\
&= \frac{-2\ln(\frac{w}{4\sqrt{2\pi}}) + ((k+1/2)w)^2}{-2\ln(\frac{2wc}{\sqrt{2\pi}})c^2 + ((k-1)w)^2} \leq \frac{-2\ln(\frac{w}{4\sqrt{2\pi}}) + ((k+1/2)w)^2}{((k-1)w)^2} \\
&= \frac{(k+1/2)^2}{(k-1)^2} + O(1/k^2) = 1 + O(1/k).
\end{aligned}
$$

□

C GENERAL CONSTRUCTIONS

C.1 Proof of Lemma 1.4

Proof. We consider the transformation in [18] in the asymmetric setting. Let $\mathbf{x}, \mathbf{y}$ be two arbitrary points from X. Part (a): Sample a pair (h_i, g_i) from $\mathcal{D}_i$ for each $i \in \{1, \ldots, n\}$ and set $h(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_n(\mathbf{x}))$ and $g(\mathbf{y}) = (g_1(\mathbf{y}), \ldots, g_n(\mathbf{y}))$. We observe that

$$
\Pr(h(\mathbf{x}) = g(\mathbf{y})) = \prod_{i=1}^{n} \Pr(h_i(\mathbf{x}) = g_i(\mathbf{y})) = \prod_{i=1}^{n} f_i(\mathrm{dist}(\mathbf{x}, \mathbf{y})).
$$

Part (b): Pick an integer $i \in \{1, \ldots, n\}$ according to $\{p_i\}$ at random. Then sample a pair (h_i, g_i) from $\mathcal{D}_i$. The hash function pair (h, g) is given by $(i, h_i(\mathbf{x}))$ and $(i, g_i(\mathbf{y}))$. We observe that

$$
\Pr(h(\mathbf{x}) = g(\mathbf{y})) = \sum_{i=1}^{n} p_i \Pr_{(h,g)\sim\mathcal{D}_i}(h(\mathbf{x}) = g(\mathbf{y})) = \sum_{i=1}^{n} p_i f_i(\mathrm{dist}(\mathbf{x}, \mathbf{y})).
$$

□

C.2 Angular similarity function

This section shows how to derive a distance sensitive scheme with collision probability $\mathrm{sim}(\mathcal{P}(\langle \mathbf{x}, \mathbf{y}\rangle))$, when $\sum_{i=0}^{k} |a_i| = 1$. Figure 4 gives some examples of functions that can be obtained from Theorem 5.1 using SimHash [17].

Proof of Theorem 5.1. Valiant [51] has shown how, for any real degree-k polynomial p, to construct a pair of mappings $\varphi_1^p, \varphi_2^p : \mathbb{R}^d \to \mathbb{R}^D$, where $D = O(d^k)$, such that $\varphi_1^p(\mathbf{x}) \cdot \varphi_2^p(\mathbf{y}) = \mathcal{P}(\langle \mathbf{x}, \mathbf{y}\rangle)$. For completeness we outline the argument here: First consider the monomial $\mathcal{P}(t) = a_k t^k$. For $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{x}^{(k)}$ denote the vector of dimension d^k indexed by vectors $\mathbf{i} = (i_1, \ldots, i_k) \in [d]^k$, where $\mathbf{x}_{\mathbf{i}}^{(k)} = \prod_{j=1}^{k} x_{i_j}$. It is easy to verify that $\langle \mathbf{x}^{(k)}, \mathbf{y}^{(k)}\rangle = (\langle \mathbf{x}, \mathbf{y}\rangle)^k$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. With this notation in place we can define $\varphi_1^p(\mathbf{x}) = \sqrt{|a_i|}\,\mathbf{x}^{(k)}$ and $\varphi_2^p(\mathbf{y}) = (a_i/\sqrt{|a_i|})\,\mathbf{y}^{(k)}$ which satisfy $\varphi_1^p(\mathbf{x}) \cdot \varphi_2^p(\mathbf{x}) = a_i(\langle \mathbf{x}, \mathbf{y},\rangle)^k$. The asymmetry of the mapping is essential to allow a negative coefficient a_k. To handle an arbitrary real polynomial $\mathcal{P}(t) = \sum_{i=0}^{k} a_i t^i$ we simply concatenate vectors corresponding to each monomial, obtaining a vector of dimension $\sum_{i=0}^{k} d^i = O(d^k)$.

Observe that $||\mathbf{x}^{(k)}||_2^2 = \langle \mathbf{x}^{(k)}, \mathbf{x}^{(k)}\rangle = (\langle \mathbf{x}, \mathbf{x},\rangle)^k = ||\mathbf{x}||_2^{2k}$. This means that for $||\mathbf{x}||_2^2 = 1$ we have $||\varphi_1^p(\mathbf{x})||^2 = \sum_{i=0}^{k} \sqrt{|a_i|}^2 = 1$, using the assumption $\sum_{i=0}^{k} |a_i| = 1$. Similarly, we have for $||\mathbf{y}||_2^2 = 1$ that $||\varphi_2^p(\mathbf{y})||^2 = \sum_{i=0}^{k} (a_i/\sqrt{|a_i|})^2 = \sum_{i=0}^{k} |a_i| = 1$. Thus, φ_1^p and φ_2^p map S^{d-1} to S^{D-1}.

Our family $\mathcal{F}$ samples a function s from the distribution $\mathcal{S}$ corresponding to sim and constructs the function pair (h, g) with $h(\mathbf{x}) = s(\varphi_1^p(\mathbf{x}))$, $g(\mathbf{y}) = s(\varphi_2^p(\mathbf{y}))$. Using the properties of the functions involved we have

$$
\Pr[h(\mathbf{x}) = g(\mathbf{y})] = \mathrm{sim}(\langle \varphi_1^p(\mathbf{x}), \varphi_2^p(\mathbf{y})\rangle) = \mathrm{sim}(\mathcal{P}(\langle \mathbf{x}, \mathbf{y}\rangle)).
$$

□

C.3 Hamming distance functions

Proof of Theorem 5.2. We initially assume that $a_0 \neq 0$ (i.e., 0 is not a root of $\mathcal{P}(t)$), and then remove this assumption at the end of the proof. We recall that a root of $\mathcal{P}(t)$ can appear with

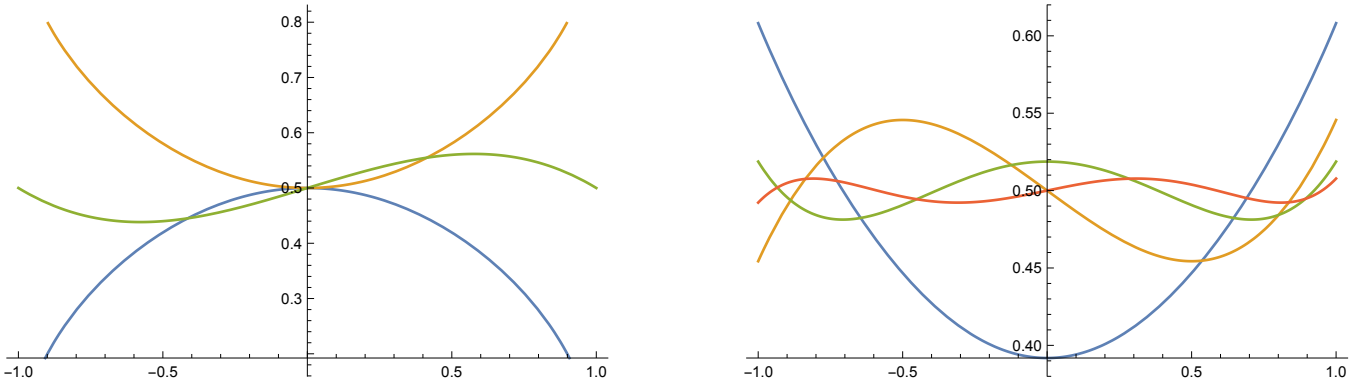

Figure 4: Examples of collision probability functions obtained using Theorem 5.1. The polynomials used are t^2, $-t^2$, $(-t^3+t^2-t)/3$ (left), and $(2t^2-1)/3$, $(4t^3-3t)/7$, $(8t^4-8t^2+1)/17$, $(16t^5-20t^3+5t)/41$ (right).

multiplicity larger than 1 and that, by the complex conjugate root theorem, if $z = a + bi$ is a complex root then so is its conjugate $z' = a - bi$. We let Z be the multiset containing the k roots of $\mathcal{P}(t)$, with Z_{r+} and Z_{r-} being the multiset of positive and negative real roots, respectively, and with Z_c being the multiset consisting of pairs of conjugate complex roots. By factoring $\mathcal{P}(t)$, we get:

$$\mathcal{P}(t) = a_k \prod_{z\in Z}(t-z) = |a_k| \prod_{z\in Z_{r+}}(z-t) \prod_{z\in Z_{r-}}(t+|z|)\cdot \prod_{z=a+bi\in Z_c}(t^2-2at+a^2+b^2), \tag{7}$$

where the last step follows since $a_k \prod_{z\in Z_{r+}}(z-t) = |a_k| \prod_{z\in Z_{r+}}(t-z) > 0$. Indeed, $\mathcal{P}(t)$ is positive in $(0,1)$ and the multiplicative terms associated with complex and negative real roots are positive in this range; this implies that the remaining terms are positive as well.

We need to introduce scaled and biased variations of bit-sampling or anti bit-sampling. Anti-bit sampling with scaling factor $\alpha \in [0,1]$ and bias $\beta \in [0,1]$ has the CPF $f(t) = \beta/2 + \alpha t/2$ and is given by randomly selecting one of following two schemes: (1) with probability $1/2$, the scheme is a standard hashing that maps data and query points to 0 with probability β, and otherwise to 0 and 1 respectively; (2) with probability $1/2$, the scheme is anti bit-sampling where the sampled bit is set to 0 with probability $1-\alpha$ on both data and query points, or kept unchanged otherwise. Similarly, bit-sampling with scaling factor $\alpha \in [0,1]$ has the CPF $f(t) = (1-\alpha t)$ and is given by using bit-sampling, where the sampled bit is set to 0 with probability $1-\alpha$ on both data and query points. (We do not need a biased version of bit-sampling.)

We now assign to each multiplicative term of (7) a scaled and biased version of bit-sampling or anti bit-sampling as follows:

- z **is real and** $z < -1$**.** We assign to z an anti bit-sampling with bias 1 and scaling factor $1/|z| \le 1$: the CPF is $S_1(t,z) = (1/2 + t/(2|z|))$, and we have $(t+|z|) = 2|z|S_1(t,z)$.
- z **is real and** $-1 \le z < 0$**.** We assign to z an anti bit-sampling with bias $|z| \le 1$ and scaling factor 1: the CPF is $S_2(t,z) = |z|/2 + t/2$, and we have $(t+|z|) = 2S_2(t,z)$.
- z **is real and** $z \ge 1$**.** We assign to z a bit-sampling with scaling factor $1/z \le 1$: the CPF is $S_3(t,z) = (1-t/z)$, and we have $(t-z) = zS_3(t,z)$.
- (z,z') **are conjugate complex roots and** $\text{Real}(z) < -1$**.** Let $z = a+bi$ and $z' = a-bi$. The assigned scheme has CPF

$$S_4(t,z) = \left(\frac{b^2}{4(a^2+b^2)} + \frac{a^2}{a^2+b^2}\left(\frac{x}{2|a|}+\frac{1}{2}\right)^2\right)$$

and is obtained as follows: with probability $b^2/(a^2+b^2)$, the scheme maps data and query points to 0 and 0 with probability 1/4, or to 0 and 1 with probability 3/4; with probability $a^2/(a^2+b^2)$, the schemes consists of the concatenation of two anti bit-sampling with bias 1 and scaling factor $1/|a|$. Note that $t^2-2at+a^2+b^2 = 4(a^2+b^2)S_4(t,z)$.
- (z,z') **are conjugate complex roots and** $\text{Real}(z) \ge 1$**.** The scheme is similar to the previous one where we use two bit-sampling with scaling factor $1/a$ instead of the anti bit-sampling. The CPF is

$$S_5(t,z) = \left(\frac{b^2}{a^2+b^2} + \frac{a^2}{a^2+b^2}\left(1-\frac{x}{a}\right)^2\right),$$

and we get $t^2-2at+a^2+b^2 = (a^2+b^2)S_5(t,z)$.
- (z,z') **are conjugate complex roots,** $-1 \le \text{Real}(z) \le 0$**, and** $|z| = a^2+b^2 \ge 1$**.** We assign the following scheme with CPF

$$S_6(t,z) = \left(\frac{x^2}{4(a^2+b^2)} + \frac{|a|x}{2(a^2+b^2)} + \frac{1}{4}\right).$$

With probability 1/4 the scheme maps data and query points to 0; with probability 1/2, the scheme consists of anti bit-sampling with bias 0 and scaling factor $|a|/(a^2+b^2) \le 1$; with probability 1/4 the scheme consists of two anti bit-sampling with bias 0 and scaling factor $\sqrt{a^2+b^2}$ each. We have $t^2-2at+a^2+b^2 = 4(a^2+b^2)S_6(t,z)$.
- (z,z') **are conjugate complex roots,** $-1 \le \text{Real}(z) \le 0$**, and** $|z| = a^2+b^2 < 1$**.** We use the scheme of the previous point with different parameters, giving CPF

$$S_7(t,z) = \left(\frac{x^2}{4} + \frac{|a|x}{2} + \frac{a^2+b^2}{4}\right).$$

The scheme is the following: with probability 1/4, the scheme is a standard hashing scheme where data points are always mapped to 0 and where a query point is mapped to 0 with

probability $a^2 + b^2$ and to 1 with probability $1 - a^2 + b^2$; with probability $1/2$, the scheme consists of anti bit-sampling with bias 0 and scaling factor $|a| \leq 1$; with probability $1/4$, the scheme consists of two anti bit-sampling with bias 0 and scaling factor 1 each. We have $t^2 - 2at + a^2 + b^2 = 4S_7(t, z)$.

Consider the scheme obtained by concatenating the above ones for each real root and each pair of conjugate roots. Its CPF is $S(t) = \prod_{i=1}^{6} \prod_{z \in Z_i} S_i(t, z)$, where Z_i contains root with CPF S_i. Then, by letting ψ denote the number of roots with negative real part, we get from Equation 7:

$$\mathcal{P}(t) = \left(2^{\psi} |a_k| \prod_{z \in Z, |\text{Real}(z)|>1} |z| \right) S(t) = \Delta S(t).$$

Consider now $a_k = 0$ and let ℓ be the largest value such that $\mathcal{P}(t) = t^{\ell} \mathcal{P}'(x)$ with $\mathcal{P}'(0) \neq 0$. We get the claimed result by concatenating ℓ anti bit-sampling, which gives a CPF of x^{ℓ}, and the scheme for $\mathcal{P}'(t)$ obtained by the procedure described above. □

REFERENCES

[1] Sofiane Abbar, Sihem Amer-Yahia, Piotr Indyk, Sepideh Mahabadi, and Kasturi R. Varadarajan. 2013. Diverse near neighbor problem. In *Symposium on Computational Geometry (SoCG)*. 207–214.

[2] Serge Abiteboul, Marcelo Arenas, Pablo Barceló, Meghyn Bienvenu, Diego Calvanese, Claire David, Richard Hull, Eyke Hüllermeier, Benny Kimelfeld, Leonid Libkin, Wim Martens, Tova Milo, Filip Murlak, Frank Neven, Magdalena Ortiz, Thomas Schwentick, Julia Stoyanovich, Jianwen Su, Dan Suciu, Victor Vianu, and Ke Yi. 2017. Research Directions for Principles of Data Management (Abridged). *SIGMOD Rec.* 45, 4 (2017), 5–17.

[3] Dakshi Agrawal and Charu C. Aggarwal. 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. 20th Symposium on Principles of Database Systems (PODS)*. 247–255.

[4] Thomas D. Ahle, Martin Aumüller, and Rasmus Pagh. 2017. Parameter-free Locality Sensitive Hashing for Spherical Range Reporting. In *Proc. 28th Symposium on Discrete Algorithms (SODA)*. 239–256.

[5] Thomas D. Ahle, Rasmus Pagh, Ilya P. Razenshteyn, and Francesco Silvestri. 2016. On the Complexity of Inner Product Similarity Join. In *Proc. 35th ACM Symposium on Principles of Database Systems (PODS)*. 151–164.

[6] Piotr Indyk Alexandr Andoni. 2017. Nearest Neighbors in High-Dimensional Spaces. In *Handbook of Discrete and Computational Geometry, Third Edition*. Chapman and Hall/CRC, 1133–1153.

[7] Alexandr Andoni and Piotr Indyk. 2006. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. In *Proc. 47th Symposium on Foundations of Computer Science (FOCS)*. 459–468.

[8] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. 2015. Practical and Optimal LSH for Angular Distance. In *Proc. 28th Int. Conference on Neural Information Processing Systems (NIPS)*. 1225–1233.

[9] Alexandr Andoni, Piotr Indyk, Huy L. Nguyen, and Ilya Razenshteyn. 2014. Beyond Locality-Sensitive Hashing. In *Proc. 25th Symposium on Discrete Algorithms (SODA)*. 1018–1028.

[10] Alexandr Andoni, Thijs Laarhoven, Ilya P. Razenshteyn, and Erik Waingarten. 2017. Optimal Hashing-based Time-Space Trade-offs for Approximate Near Neighbors. In *Proc. 28th Symposium on Discrete Algorithms (SODA)*. 47–66.

[11] Alexandr Andoni and Ilya P. Razenshteyn. 2015. Optimal Data-Dependent Hashing for Approximate Near Neighbors. In *Proc. 47th Symposium on Theory of Computing (STOC)*. 793–801.

[12] Alexandr Andoni and Ilya P. Razenshteyn. 2016. Tight Lower Bounds for Data-Dependent Locality-Sensitive Hashing. In *Proc. 32nd Int. Symposium on Computational Geometry (SoCG)*. 9:1–9:11.

[13] Nikolaus Augsten and Michael H Böhlen. 2013. Similarity joins in relational database systems. *Synthesis Lectures on Data Management* 5, 5 (2013), 1–124.

[14] Anja Becker, Léo Ducas, Nicolas Gama, and Thijs Laarhoven. 2016. New directions in nearest neighbor searching with applications to lattice sieving. In *Proc. 27th Symposium on Discrete Algorithms (SODA)*. 10–24.

[15] Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proc. Compression and Complexity of Sequences*. 21–29.

[16] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer Networks and ISDN Systems* 29, 8-13 (1997), 1157–1166.

[17] Moses Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proc. 34th ACM Symposium on Theory of Computing (STOC)*. 380–388.

[18] Flavio Chierichetti and Ravi Kumar. 2015. LSH-preserving functions and their applications. *J. ACM* 62, 5 (2015), 33.

[19] Flavio Chierichetti, Ravi Kumar, and Mohammad Mahdian. 2014. The complexity of {LSH} feasibility. *Theoretical Computer Science* 530 (2014), 89 – 101.

[20] Flavio Chierichetti, Alessandro Panconesi, Ravi Kumar, and Erisa Terolli. 2017. The Distortion of Locality Sensitive Hashing. In *Proc. Conference on Innovations in Theoretical Computer Science (ITCS)*.

[21] Tobias Christiani. 2017. A Framework for Similarity Search with Space-Time Tradeoffs using Locality-Sensitive Filtering. In *Proc. 28th Symposium on Discrete Algorithms (SODA)*. 31–46.

[22] Tobias Christiani and Rasmus Pagh. 2017. Set Similarity Search Beyond Minhash. In *Proc. 49th Symposium on Theory of Computing (STOC)*.

[23] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. 2004. Locality-sensitive Hashing Scheme Based on P-stable Distributions. In *Proc. 20th Symposium on Computational Geometry (SoCG)*. 253–262.

[24] Emiliano De Cristofaro, Jihye Kim, and Gene Tsudik. 2010. Linear-Complexity Private Set Intersection Protocols Secure in Malicious Model. In *Asiacrypt*, Vol. 6477. 213–231.

[25] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391.

[26] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas. 2004. Efficient Private Matching and Set Intersection. In *Proc. Int. Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. 1–19.

[27] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity Search in High Dimensions via Hashing. In *Proc. 25th Int. Conference on Very Large Data Bases (VLDB)*. 518–529.

[28] Oded Goldreich. 2009. *Foundations of cryptography: volume 2, basic applications*. Cambridge University Press.

[29] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. 2012. Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality. *Theory of Computing* 8, 1 (2012), 321–350.

[30] Xiao Hu, Yufei Tao, and Ke Yi. 2017. Output-optimal Parallel Algorithms for Similarity Joins. In *Proc. Symposium on Principles of Database Systems (PODS)*. 79–90.

[31] Piotr Indyk. 2003. Better Algorithms for High-dimensional Proximity Problems via Asymmetric Embeddings. In *Proc. 14th Symposium on Discrete Algorithms (SODA)*. 539–545.

[32] Piotr Indyk and Rajeev Motwani. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proc. 30th ACM Symposium on the Theory of Computing (STOC)*. 604–613.

[33] Prateek Jain, Sudheendra Vijayanarasimhan, and Kristen Grauman. 2010. Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning. In *Proc. Conference on Neural Information Processing Systems (NIPS)*.

[34] Michael Kapralov. 2015. Smooth Tradeoffs between Insert and Query Complexity in Nearest Neighbor Search. In *Proc. 34th ACM Symposium on Principles of Database Systems, PODS 2015*. 329–342.

[35] C. G. Lambert, S. E. Harrington, C. R. Harvey, and A. Glodjo. 1999. Efficient On-Line Nonparametric Kernel Density Estimation. *Algorithmica* 25, 1 (1999), 37–57.

[36] Wei Liu, Jun Wang, Yadong Mu, Sanjiv Kumar, and Shih-Fu Chang. 2012. Compact Hyperplane Hashing with Bilinear Functions. In *Proc. Conference on Machine Learning (ICML)*.

[37] Rajeev Motwani, Assaf Naor, and Rina Panigrahy. 2007. Lower Bounds on Locality Sensitive Hashing. *SIAM J. Discrete Math.* 21, 4 (2007), 930–935.

[38] Behnam Neyshabur and Nathan Srebro. 2015. On Symmetric and Asymmetric LSHs for Inner Product Search. In *Proc. 32nd Conference on Machine Learning (ICML)*. 1926–1934.

[39] Ryan O'Donnell. 2014. *Analysis of Boolean Functions*. Cambridge University Press.

[40] Ryan O'Donnell, Yi Wu, and Yuan Zhou. 2014. Optimal lower bounds for locality-sensitive hashing (except when q is tiny). *ACM Transactions on Computation Theory (TOCT)* 6, 1 (2014), 5.

[41] Rasmus Pagh, Francesco Silvestri, Johan Sivertsen, and Matthew Skala. 2017. Approximate furthest neighbor with application to annulus query. *Information Systems* 64 (2017), 152–162.

[42] Ninh Pham and Rasmus Pagh. 2013. Fast and scalable polynomial kernels via explicit feature maps. In *Proc. 19th Int. Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 239–247.

[43] Benny Pinkas, Thomas Schneider, Christian Weinert, and Udi Wieder. 2017. Linear Size Circuit-based PSI via Two-Dimensional Cuckoo Hashing. (2017). Manuscript under submission.

[44] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Proc. 21st Conference on Neural Information Processing Systems (NIPS)*. 1177–1184.

[45] M. Sadegh Riazi, Beidi Chen, Anshumali Shrivastava, Dan S. Wallach, and Farinaz Koushanfar. 2016. Sub-Linear Privacy-Preserving Near-Neighbor Search with Untrusted Server on Large-Scale Datasets. (2016). ArXiv:1612.01835.

[46] Walter Rudin. 1990. *Fourier Analysis on Groups.* Wiley, New York.

[47] Richard I. Savage. 1962. Mill's Ratio for Multivariate Normal Distributions. *Jour. Res. NBS Math. Sci.* 66, 3 (1962), 93–96.

[48] Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS). In *Proc. 27th Conference on Neural Information Processing Systems (NIPS)*. 2321–2329.

[49] Yasin N Silva, Walid G Aref, and Mohamed H Ali. 2010. The similarity join database operator. In *Proc. Int. Conference on Data Engineering (ICDE)*. IEEE, 892–903.

[50] S. J. Szarek and E. Werner. 1999. A nonsymmetric correlation inequality for Gaussian measure. *Journal of Multivariate Analysis* 68, 2 (1999), 193–211.

[51] Gregory Valiant. 2015. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *J. ACM* 62, 2 (2015), 13.

[52] Sudheendra Vijayanarasimhan, Prateek Jain, and Kristen Grauman. 2014. Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 2 (2014), 276–288.

[53] J. Wang, H. T. Shen, J. Song, and J. Ji. 2014. Hashing for Similarity Search: A Survey. *CoRR* abs/1408.2927 (2014). http://arxiv.org/abs/1408.2927

[54] Haoyu Zhang and Qin Zhang. 2017. EmbedJoin: Efficient Edit Similarity Joins via Embeddings. In *Proc. Int. Conference on Knowledge Discovery and Data Mining (KDD)*. 585–594.