Hilbert Exclusion: Improved Metric Search through Finite Isometric Embeddings

Richard Connor, Franco Alberto Cardillo, Lucia Vadicamo, Fausto Rabitti

Most research into similarity search in metric spaces relies upon the triangle inequality property. This property allows the space to be arranged according to relative distances to avoid searching some subspaces. We show that many common metric spaces, notably including those using Euclidean and Jensen-Shannon distances, also have a stronger property, sometimes called the four-point property: in essence, these spaces allow an isometric embedding of any four points in three-dimensional Euclidean space, as well as any three points in two-dimensional Euclidean space. In fact, we show that any space which is isometrically embeddable in Hilbert space has the stronger property. This property gives stronger geometric guarantees, and one in particular, which we name the Hilbert Exclusion property, allows any indexing mechanism which uses hyperplane partitioning to perform better. One outcome of this observation is that a number of state-of-the-art indexing mechanisms over high dimensional spaces can be easily refined to give a significant increase in performance; furthermore, the improvement given is greater in higher dimensions. This therefore leads to a significant improvement in the cost of metric search in these spaces.

CCS Concepts: •**Information systems** → **Data structures; Multidimensional range search; Proximity search; Database query processing; Information retrieval query processing; Retrieval models and ranking; Retrieval efficiency;** *Multimedia information systems;* •**Theory of computation** → **Random projections and metric embeddings;**

Additional Key Words and Phrases: Similarity search, Metric Space, Metric Indexing, Four-point property, Hilbert Embedding

ACM Reference Format:
Richard Connor, Franco Alberto Cardillo, Lucia Vadicamo, Fausto Rabitti, 2016. Hilbert Exclusion: Improved Metric Search through Finite Isometric Embeddings *ACM Trans. Inf. Syst.* V, N, Article XXXX (XXXX 2016), 28 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In the realm of similarity search, many metric indexing techniques are available. These rely on the metric properties of the distance function used, and in particular use the triangular inequality property in various ways to exclude parts of the space from a search for values similar to a given query.

Any proper metric space (U, d) is isometrically 3-embeddable in two dimensional Euclidean space (ℓ_2^2). That is, for any three objects within (U, d), there exists a function mapping those objects into ℓ_2^2 which preserves the distances between them. This is in fact a corollary of the metric properties of d.

Author's address: R. Connor, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, G1 1XH, United Kingdom; F.A. Cardillo, Institute for Computational Linguistics (ILC), National Research Council of Italy (CNR), Via Moruzzi 1, 56124 Pisa, Italy; L.Vadicamo, and F. Rabitti, Information Science and Technology Institute (ISTI), National Research Council of Italy (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy; email: richard.connor@strath.ac.uk, francoalberto.cardillo@ilc.cnr.it, {lucia.vadicamo, fausto.rabitti}@isti.cnr.it.

DOI: http://dx.doi.org/10.1145/0000000.0000000

In this paper we consider spaces with the stronger property of being isometrically 4-embeddable in three dimensional Euclidean space (ℓ_2^3). We show that these spaces include all those which have isometric embeddings in Hilbert space, notably including any space under Euclidean distance, as well as the proper metric forms of Jensen-Shannon, Triangular Discrimination and a novel form of Cosine distance.

Such spaces give stronger geometric properties. All metric indexing currently relies on one (or both) of two core principles: exclusion based on a bounding radius, or exclusion based on a hyperplane partition, both of which can be explained in terms of their 3-embeddabilty property. Using the stronger 4-embeddability, we show that a greater degree of exclusion is possible, and that this exclusion degrades more slowly as higher dimensions are considered.

Our main result is very simple. Consider any four points p_1, p_2, q and s in a metric space (U, d), where the intent is that q is a query, s is a solution to this query (i.e. $d(q, s) \leq t$ for some real value t), and p_1 and p_2 are points within the space which have been previously used to structure the data.

During the progress of a query evaluation, the distances $d(q, p_1)$ and $d(q, p_2)$ are evaluated. Assuming without loss of generality that $d(q, p_2) \leq d(q, p_1)$, then a well known property used during search is

$$\frac{d(q, p_1) - d(q, p_2)}{2} > t \Rightarrow d(s, p_1) > d(s, p_2)$$

Here, we show that for certain common classes of spaces

$$\frac{d(q, p_1)^2 - d(q, p_2)^2}{2\, d(p_1, p_2)} > t \Rightarrow d(s, p_1) > d(s, p_2)$$

Both properties can be used to avoid searching subspaces where all elements are known to be closer to p_1 than p_2. The second property however is strictly weaker, meaning that any indexing mechanism which uses the first can be made more efficient[1].

In the context of exact metric indexing, the best performing index for general purpose use is currently believed to be the distal SAT [Chávez et al. 2014; 2016], which uses a combination of pivot and hyperplane-based exclusion. For this structure, we show a significant performance increase for Euclidean and Jensen-Shannon spaces, especially in higher dimensions. This therefore gives, for these spaces, a new high performance benchmark for similarity search.

The rest of this article is structured as follows. Section 2 gives a general context of metric search and finite isometric embedding; after basic definitions, it goes on to show how the essential mechanisms of metric search can be explained in terms of finite embeddings. Section 3 briefly shows, in outline, why better performance can be expected from a space which is 4-embeddable in ℓ_2^3. Section 4 gives a formal definition of our new exclusion property for hyperplane partitioning, and proves its applicability to any space which is isometrically 4-embeddable in ℓ_2^3. Section 5 gives some background mathematics of Hilbert spaces, and shows the 4-embeddabilty property for three important metrics. Section 6 gives an analysis of the improvement, including relative performance measurements for some metric index implementations which use hyperplane partitioning. Section 7 shows how the new exclusion criterion degrades relatively less severely over higher dimensions than those currently used, and Section 8 summarises and outlines further possibilities.

[1]The distance $d(p_1, p_2)$ can be evaluated as the index is built, not during the query.

2. BACKGROUND AND RELATED WORK

2.1. Similarity Search and Metric Indexing

Our context of interest is in exact search in general metric spaces. That is, we are interested in searching a (large) finite set of objects S which is a subset of an infinite set U, where (U, d) is a metric space. The general requirement is to efficiently find members of S which are close to an arbitrary member of U, where the distance function d gives the only way by which any two objects may be compared. There are many important practical examples captured by this mathematical framework, see for example [Chávez and Navarro 2005; Zezula et al. 2006].

For (U, d) to be a metric space, the distance function $d : U \times U \to \mathbb{R}$ requires to satisfy

— Positivity: $\forall\, a, b \in U,\ d(a, b) \geq 0$ with equality if, and only if, $a = b$;
— Symmetry: $\forall\, a, b \in U,\ d(a, b) = d(b, a)$;
— Triangle inequality: $\forall\, a, b, c \in U,\ d(a, c) \leq d(a, b) + d(b, c)$.

Such spaces are typically searched with reference to a query object $q \in U$. A threshold search for some threshold t, based on a query $q \in U$, has the solution set $\{s \in S \text{ such that } d(q, s) \leq t\}$.

Typically S is too large to allow an exhaustive search. However such queries can often be performed efficiently by use of a *metric index*, one of a large family of data structures which make use of the triangle inequality property in order to arrange the set of objects S in such a way as to minimise the time required to retrieve the query result. Efficiency is primarily achieved by avoiding unnecessary distance calculations, although the efficient use of memory hierarchies is also extremely important. Both of these are optimised by structuring the set based on relative distances of objects from each other, so that triangle inequality can be used to determine subsets which do not need to be exhaustively checked. Such avoidance is normally referred to as *exclusion*.

For exact metric search, almost all indexing methods can be divided into those which at each exclusion possibility use a single "pivot" point to give radius-based exclusion, and those which use two reference points to give hyperplane-based exclusion. Many variants of each have been proposed, including many hybrids; [Chávez et al. 2001], [Zezula et al. 2006] give excellent surveys. In general the best choice seems to depend on the particular context of metric and data.

Here we focus particularly on mechanisms which use hyperplane-based exclusion. The simplest such index structure is the Generalised Hyperplane Tree [Uhlmann 1991]. Others include the Monotonic Bisector Tree [Noltemeier et al. 1992], the Metric Index [Novak et al. 2011], and the Spatial Approximation Tree [Navarro 2002]. This last has various derivatives, notably including the Dynamic SAT [Navarro and Reyes 2002] and the Distal SAT [Chávez et al. 2016], which includes a variant SAT_{out} which is believed to be, at time of writing, the most efficient known general-use indexing structure for performing exact search [Chávez et al. 2016]; therefore a significant improvement on this, as we show here, is a significant result.

2.2. Finite Isometric Embeddings

An isometric embedding of one metric space (V, d_v) in another (W, d_w) can be achieved when there exists a mapping function $f : V \to W$ such that $d_v(x, y) = d_w(f(x), f(y))$, for all $x, y \in V$. A finite isometric embedding occurs whenever this property is true for any finite selection of n points from V, in which case the terminology used is that V is isometrically n-embeddable in W.

The first observation to be made in this context is that any metric space is isometrically 3-embeddable in ℓ_2^2. This is apparent from the triangle inequality property of a proper metric, as illustrated in Figure 1. In fact the two properties are equivalent:

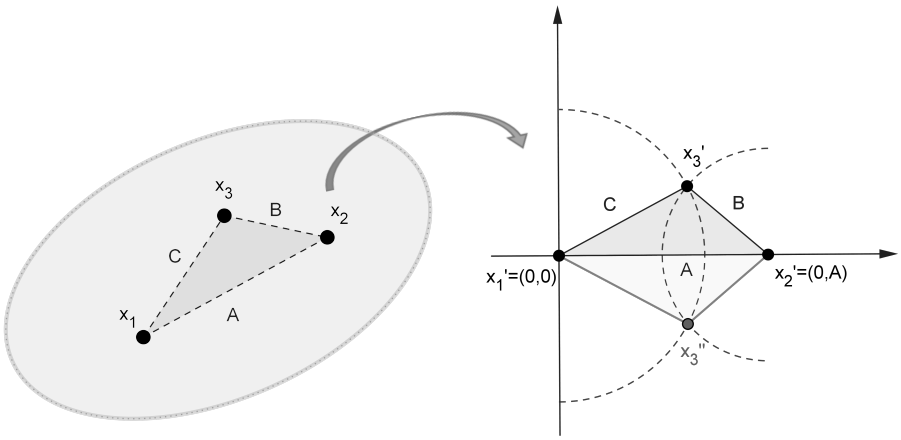


Fig. 1: For any three points x_1, x_2 and x_3 whose distances satisfy the triangle inequality property, a triangle can be constructed within 2D Euclidean space such that x'_1 is at the origin, x'_2 lies on the X-axis, and x'_3 is where the distances B and C intersect.

for any semi-metric space[2] (V, d_v) which is isometrically 3-embeddable in ℓ_2^2, triangle inequality also holds.

Much work was done on finite isometric embeddings in the 1930s, but it does not appear to have been a "hot topic" since then. Blumenthal [Blumenthal 1933] provides an excellent and concise summary of this work as it pertains to ours. He attributes our observation above, that any semi-metric space which is 3-embeddable in ℓ_2^2 is a metric space, to Menger. He uses the phrase *the four-point property* to mean a semi-metric space which is isometrically 4-embeddable in ℓ_2^3. Wilson [Wilson 1932] shows various properties of such spaces, and Blumenthal points out that results given by Wilson, when combined with work by Menger in [Menger 1931], generalise to show that some spaces have the *n-point property* (i.e. any n points can be isometrically embedded in ℓ_2^{n-1}.) This is in fact a more general result than our Lemma 5.2 which uses a more modern formulation for high dimensional Euclidean space.

The most important results in finite isometric embeddings from our perspective are given by Schoenberg and Blumenthal. [Schoenberg 1938] shows an initially surprising result that if a kernel function K has certain simple properties, then it can be used to construct a metric space which is isometrically embeddable in a Hilbert space. Blumenthal [Blumenthal 1953] shows that any space which is isometrically embeddable in a Hilbert space has the n-point property for every possible integer n. In combination these are extraordinarily strong from our perspective: for any kernel function K with the correct properties, we can construct a proper metric space with the four-point property. We expand on this observation in Section 5.

Although normally expressed in terms of the property of triangle inequality, the properties of a metric space that allow indexing can be equally well expressed in terms of the geometric guarantees afforded according to the 3-embeddability property in ℓ_2^2. To set the context, we briefly explain the two main indexing principles in terms of this property.

[2] a space where triangle inequality is not guaranteed

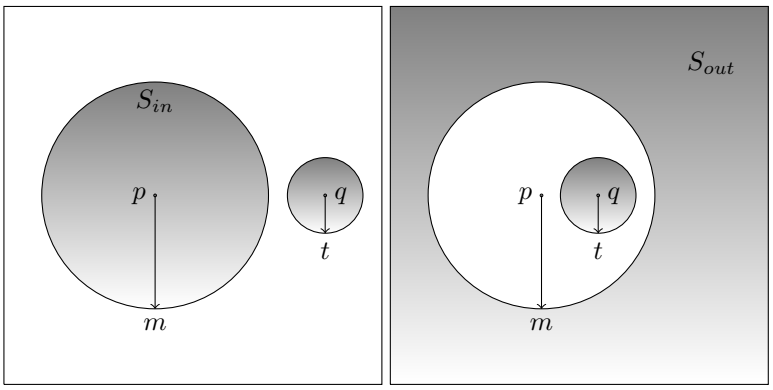


Fig. 2: Pivot-based exclusion illustrated by 3-embedding in ℓ_2^2. Objects in S_{in} are at most distance m from p, and objects in S_{out} are at least m from p. Given $d(q,p) > m+t$, S_{in} cannot contain a solution to the query. Similarily if $d(q,p) < m-t$, S_{out} cannot. Such diagrams should be treated with extreme care: for a general metric space, no more than three objects have a guarantee of isometric embedding within 2D Cartesian space. In these cases, it is necessary only to consider an embedding of the pivot, the query, and an arbitrary object within the solution space to see that the distance guarantee holds.

2.3. Pivot-based indexing

This technique entails the selection of a *pivot* point $p \in S$, and the construction of one or more subsets of S based on a fixed distance m from p, e.g. S_{in} where $s \in S_{in} \Rightarrow d(p,s) \leq m$. For a query q, $d(q,p)$ is calculated; if this is greater than $m+t$, for a query threshold t, then no element of S within distance t of q can be within S_{in} and every element of S_{in} can therefore be excluded from the search. Similarly, S_{out} could be constructed such that $s \in S_{out} \Rightarrow d(p,s) > m$, in which case the elements of S_{out} can be excluded if $d(q,p) \leq m-t$.

The validity of the pivoting principle can be shown algebraically using the triangle inequality property of the metric, and many different mechanisms have been described using it [Chávez et al. 2001; Zezula et al. 2006]. They are often illustrated in the manner of Figure 2; using such illustrations relies upon isometric 3-embeddability within ℓ_2^2 of any metric space, but should also be treated with care whenever more than three objects are considered, as the distances among them cannot in general be preserved.

2.4. Partition-based indexing

In this type of indexing, two elements of S are chosen, and the rest of S is divided into two subsets according to which of these elements is closer. Formally:

$$
\begin{aligned}
&p_1, p_2 \in S \\
&\quad S_{p_1} = \{s \in S - \{p_1, p_2\}\,,\, d(s,p_1) < d(s,p_2)\} \\
&\quad S_{p_2} = \{s \in S - \{p_1, p_2\}\,,\, d(s,p_1) \geq d(s,p_2)\}
\end{aligned}
$$

To evaluate a query over q, the distances $d(q,p_1)$ and $d(q,p_2)$ are first calculated. If $|d(q,p_1) - d(q,p_2)| > 2t$, then the subset associated with the point further from q does not intersect with the solution set of the query and these values can be excluded from the search. Again, the exclusion condition is straightforward to derive algebraically from the triangle inequality property, but can also be shown in terms of 3-embeddability within ℓ_2^2.

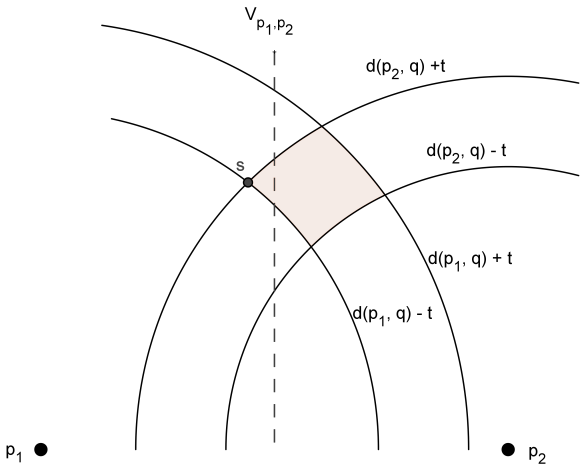


Fig. 3: The two pivot points and any solution to the query can be isometrically embedded in ℓ_2^2. The point q cannot be drawn in the same diagram. Given its distance from p_1 and p_2, any solution in the original metric space must lie in the region bounded by the four arcs shown in the ℓ_2^2 projection. If the point s lies to the right of V_{p_1,p_2} in ℓ_2^2, there is therefore no requirement to search to the left of the hyperplane in the original space. By symmetry, if $|d(q,p_1) - d(q,p_2)| > 2t$, then half of the search space can be excluded.

Figure 3 shows a graphical interpretation of this situation using the ℓ_2^2 embedding. The three points chosen for illustration, relying on the 3-embedding property, are p_1, p_2, and an arbitrary solution point to the query q. The two pivot points and any solution to the query can be isometrically embedded in ℓ_2^2. In general the point q may not be isometrically embedded in the same plane as the two pivot points p_1, p_2, and a solution s, and therefore cannot be drawn in the diagram.

The line V_{p_1,p_2} represents a boundary between S_{p_1} and S_{p_2} in the original space. If the whole of the region bounded by the four arcs lies to one side of this line, there is no requirement to search in the other part of the space. It can be seen from the diagram, if q is closest to p_2, that this occurs when $d(q,p_1)-t > d(q,p_2)+t$, i.e. $d(q,p_1)-d(q,p_2) > 2t$. This illustration alone in fact is not quite convincing; it must be further observed that, for any two 3-embeddings where two of the points are the same (in this case p_1 and p_2), then embedding functions can be chosen that map those two points to the same two points in ℓ_2^2 (e.g. see Figure 1) thus preserving the semantics of the line V_{p_1,p_2}.

3. PARTITION-BASED INDEXING WITH 4-EMBEDDING IN ℓ_2^3

We introduce the main result of this paper with simple observation that, for spaces that are isometrically 4-embeddable in ℓ_2^3, a tighter exclusion condition is possible for partitions.

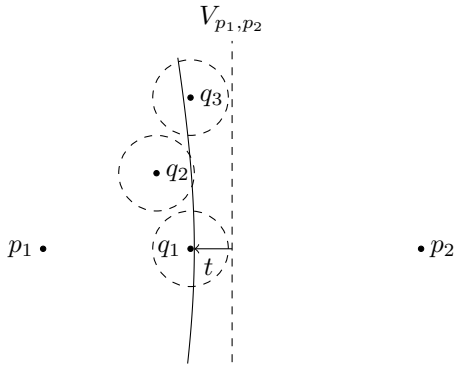


Fig. 4: Three queries, q_1, q_2 and q_3, each with threshold t, on the left side of the boundary V_{p_1,p_2}. Since $d(p_2, q_3) - d(p_1, q_3) < 2t$, S_{p_2} cannot be excluded from the search on q_3. ($p_1 = (-5, 0), p_2 = (5, 0), q_3 = (-1.1, 4), t = 1$). The hyperbola curve represents all possible points $x \in S_{p_1}$ such that $d(p_2, x) - d(p_1, x) = 2t$, i.e. the boundary of the exclusion condition.

Figure 4 shows an example taken from a metric space 3-embedded in ℓ_2^2, that is a standard metric space. Of the three queries, only q_1 and q_2 allow the partition on the far side of the hyperplane to be excluded, as for q_3 the exclusion condition is not met, even although the solution space appears geometrically separated from the right-hand side. This appearance however is an illusion, as in a general metric space it would require the isometric embedding of four points (p_1, p_2, q_3 and s, for any solution s) in the diagram.

In general, the boundary defined by the exclusion condition is given by the locus of points x such that $d(x, p_2) - d(x, p_1) = 2t$ which defines a hyperbola focussed at p_1 and p_2, with semi-major axis t. The minimum distance of this hyperbola from the line V_{p_1,p_2} is t, but this occurs only on the line passing through p_1 and p_2. When considering this diagram in two dimensions, the relative distances among p_1, p_2 and any individual q_i are significant, but as a general metric space guarantees only 3-embeddability, the circles drawn around the queries are meaningless with respect to the original space.

Consider now Figure 5, which shows the same situation but relying on a 4-embeddability in ℓ_2^3. Here the relative distances among any four points can be safely considered: in this case p_1, p_2, q, and any solution to q. The plane on which the diagram is drawn is that containing p_1, p_2 and q, and therefore the locus of any solution to q consists of a sphere, radius t, centred around q.

It is clear from this diagram, in comparison with Figure 3, that a more useful exclusion condition can be used: whenever the distance between q and V_{p_1,p_2} is greater than t, S_{p_1} does not require to be searched. Other than the single point on the line through p_1 and p_2 this distance is always strictly less than the nearest point on the corresponding hyperbola, and thus more exclusions are always possible.

Figure 6 gives an illustration of the two boundary conditions in ℓ_2^3. It can be seen that our new exclusion condition is weaker than the normal, hyperbolic, condition; in this sense weaker implies better, as it allows more queries to exclude the opposing semispace from further consideration. For discussion in the rest of the paper, we refer to the new exclusion condition as *Hilbert Exclusion*, and the former condition as *Hyperbolic Exclusion*. We proceed with a formal definition and proof of correctness of Hilbert Exclusion.

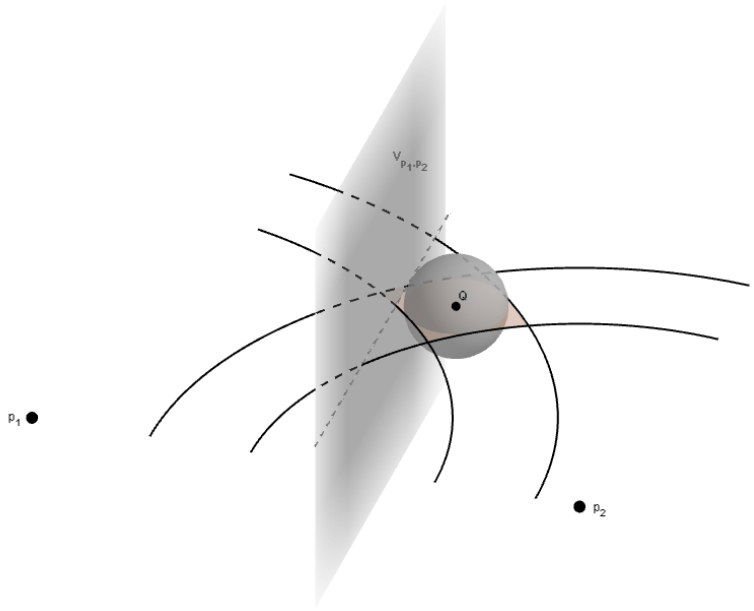


Fig. 5: Four points (p_1, p_2, q and s, s.t. $d(q,s) \leq t$) in ℓ_2^3. For fixed p_1, p_2 and q, any solution to the query lies within the sphere centred around q and cannot lie within S_{p_1}, even although $d(q,p_1) - d(q,p_2) < 2t$. Note that V_{p_1,p_2} in the figure now represents the hyperplane that divides the space into two subspaces: objects nearer to p_1 belonging to the left subspace and objects nearer to p_2 to the right.

4. THE HILBERT EXCLUSION CONDITION

THEOREM 4.1. *Consider any three points $p_1, p_2, q \in \ell_2^3$ with $d(q,p_2) < d(q,p_1)$. Then the condition*

$$\frac{d(q,p_1)^2 - d(q,p_2)^2}{2\,d(p_1,p_2)} > t \tag{1}$$

implies that $d(s,p_2) < d(s,p_1)$ for all s s.t. $d(q,s) \leq t$.

PROOF.

It is sufficient to prove that the distance between the point q and the plane V_{p_1,p_2} is greater that t. In this case, $d(s,p_2) < d(s,p_1)$ for all s s.t. $d(q,s) \leq t$.

The equation of the plane V_{p_1,p_2} can be written as the scalar product $(p_2 - p_1) \cdot (x - \frac{(p_2+p_1)}{2}) = 0$, and so its distance from q is given by

$$\text{dist}(q, V_{p_1,p_2}) = \left| \left(q - \frac{(p_2+p_1)}{2} \right) \cdot \frac{(p_2-p_1)}{\|p_2-p_1\|_2} \right| = \frac{d(q,p_1)^2 - d(q,p_2)^2}{2\,d(p_1,p_2)}$$

Therefore if $\text{dist}(q, V_{p_1,p_2}) > t$, any point within distance t of q is closer to p_2 than to p_1 □

The practical application of this theorem is in search indexes which partition the search space. The exclusion condition

$$\frac{d(q,p_1)^2 - d(q,p_2)^2}{2\,d(p_1,p_2)} > t$$

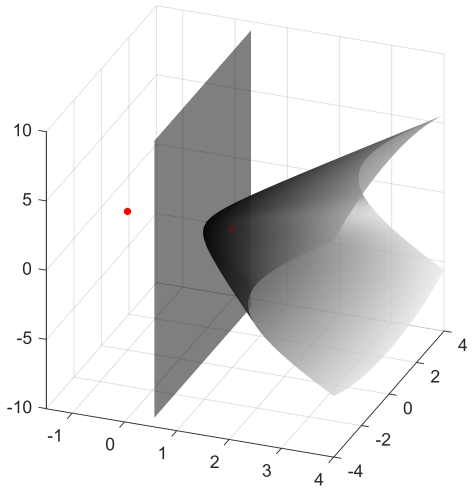


Fig. 6: This illustration shows the geometric principle behind the new exclusion condition, which can be applied to any metric space which is isometrically 4-embeddable in 3D Euclidean space. Here pivots are placed at $(-1, 0, 0)$ and $(1, 0, 0)$, the threshold selected is 0.5. The plane and hyperboloid shown in the 3D space represent the boundaries of the two exclusion conditions we now refer to as *Hilbert Exclusion* and *Hyperbolic Exclusion* respectively.

can be used in place of

$$\frac{d(q, p_1) - d(q, p_2)}{2} > t$$

in order to exclude any subspace which is known to be closer to p_1 than to p_2. The important point in our context is that the first condition is weaker than the second[3], and is therefore in general a more useful exclusion condition.

THEOREM 4.2. *For any metric space (U, d), and for any three points $p_1, p_2, q \in U$, the exclusion condition of Theorem 4.1 holds if (U, d) is isometrically 4-embeddable in ℓ_2^3.*

PROOF. Let (U, d) be a metric space isometrically 4-embeddable in ℓ_2^3. Let t be a real positive number and $p_1, p_2, q \in U$ be three points such that $d(q, p_2) < d(q, p_1)$ and

$$\frac{d(q, p_1)^2 - d(q, p_2)^2}{2\, d(p_1, p_2)} > t. \tag{2}$$

For any $s \in U$ such that $d(q, s) \leq t$ we want to prove that $d(s, p_2) < d(s, p_1)$. Since (U, d) is isometrically 4-embeddable in ℓ_2^3, there exists a function $f : (U, d) \to \ell_2^3$ which

[3]A simple proof is given in Appendix A.

preserves all the six distances:

$$\|f(p_1) - f(p_2)\|_2 = d(p_1, p_2) \tag{3}$$
$$\|f(q) - f(p_1)\|_2 = d(q, p_1) \tag{4}$$
$$\|f(q) - f(p_2)\|_2 = d(q, p_2) \tag{5}$$
$$\|f(s) - f(q)\|_2 = d(s, q) \leq t \tag{6}$$
$$\|f(s) - f(p_1)\|_2 = d(s, p_1) \tag{7}$$
$$\|f(s) - f(p_2)\|_2 = d(s, p_2). \tag{8}$$

Equations (3)-(6) together with equation (2) imply that points $\{f(p_1),\ f(p_2),\ f(q), f(s)\} \in \ell_2^3$ satisfy the exclusion condition of Theorem 1. Thus, $f(s)$ is closer to $f(p_2)$ than to $f(p_1)$, i.e., $\|f(s) - f(p_1)\|_2 > \|f(s) - f(p_2)\|_2$. This proves also that s is closer to p_2 than to p_1, in fact

$$d(s, p_1) = \|f(s) - f(p_1)\|_2 > \|f(s) - f(p_2)\|_2 = d(s, p_2).$$

□

Note that, for any solution s in U, a different mapping function f may be required, however the only importance of this function is that, for any four points, it exists: there is no requirement to identify it.

5. VECTOR SPACES ISOMETRICALLY 4-EMBEDDABLE IN ℓ_2^3

5.1. ℓ_2^n Space

Euclidean distance applied over many-dimensional data is probably the most common of metric searches. In these cases, we have an immediate result:

THEOREM 5.1. *Any n-dimensional Euclidean space (i.e. an ℓ_2^n space, for any n) is 4-embeddable in ℓ_2^3*

LEMMA 5.2. *In n dimensions, precisely one k-dimensional hyperplane passes through any $(k+1)$ points that do not lie in a $(k-1)$-dimensional hyperplane.*[4] *Moreover, a k-dimensional hyperplane can be regarded as a k-dimensional space in its own right. (See for example [Aleksandrov et al. 1999], Chapter 7.)*

PROOF. From Lemma 5.2, any ℓ_2^n space is $(k+1)$-embeddable in ℓ_2^k. Therefore any ℓ_2^n space is 4-embeddable in ℓ_2^3. □

COROLLARY 5.3. *The Hilbert Exclusion Condition is valid over Euclidean spaces of any dimension.*

However, we have a more general result: any metric space which has an isometric embedding in a Hilbert space is also 4-embeddable in ℓ_2^3. This includes Euclidean space of any dimension, but also includes other important spaces, notably any governed by the Jensen-Shannon distance.

5.2. Inner Product Spaces and Hilbert Spaces

The importance of Hilbert spaces is the generalisation of the notion of Euclidean space by extending the methods of vector algebra and calculus to spaces with any finite or infinite number of dimensions. A Hilbert space is an abstract vector space possessing the structure of an inner product that allows length and angle to be measured which gives certain geometric properties. These properties extend to abstract, non-geometric

[4] If the points are coplanar, an infinity of such hyperplanes exist; the important point for our purposes is only that at least one such hyperplane exists.

spaces which can be isometrically embedded in a Hilbert space. The key property of interest here is in 4-point isometric embedding in ℓ_2^3.

LEMMA 5.4 (SHOENBERG'S THEOREM [SCHOENBERG 1938; TOPSØE 2003]). *Let X be a nonempty set and $K : X \times X \to \mathbb{R}$ a mapping that satisfies the positivity and symmetric proprieties and such that, for all finite sets $(c_i)_{i\leq n}$ of real numbers and all finite sets $(x_i)_{i\leq n}$ of points in X, the implication*

$$\sum_{i=1}^{n} c_i = 0 \Rightarrow \sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \leq 0 \tag{9}$$

holds (i.e., K is conditionally negative semidefinite function). Then $(X, \sqrt{K})$ is a metric space which can be embedded isometrically as a subspace of a real Hilbert space.

The main importance from our perspective is that, given a metric space $(X, \sqrt{K})$, it is sufficient for K to be a conditionally negative semidefinite function in order to have isometric embeddability into a Hilbert Space.

LEMMA 5.5 (BLUMENTHAL LEMMA 53.1 [BLUMENTHAL 1953]).
A numerable semimetric space is isometrically embeddable in a Hilbert space if and only if it is isometrically n-embeddable in ℓ_2^{n-1} for every positive integer n.

LEMMA 5.6 (SCHOLTES PROPOSITION 1.3 [SCHOLTES 2013]). *Let $(X, \|\cdot\|)$ be a normed vector space. Then the following statements are equivalent:*

— *$(X, \|\cdot\|)$ is an inner product space, i.e., there exists an inner product $< \cdot, \cdot >$ on X which induces the norm: $\forall\, x \in X,\ \|x\| = \sqrt{< x, x >}$*
— *all subsets $\{u, v, w, x\} \subset X$ are isometrically embeddable in ℓ_2^3.*

By definition, any Hilbert space is a normed vector space which is also an inner product space. From the above lemmata, we can observe that for any semimetric, negative semidefinite kernel function K over $\mathbb{R}^n$, then $(\mathbb{R}^n, \sqrt{K})$ is a proper metric space which can be searched using our new exclusion rule. The fact that the resulting metric space is a subspace of Hilbert space is not strictly necessary for this purpose, although it gives other potentially valuable geometric properties as well. In fact, the Hilbert embeddability guarantees the n-point property for all n, while just the 4-point property is required for our new exclusion rule. It is worth noting that in [Blumenthal 1953] a weaker version of the Schoenberg's theorem is used to characterise any metric space which has the 4-point property:

LEMMA 5.7 ([BLUMENTHAL 1953]). *A metric space (X, d) is isometrically 4-embeddable in ℓ_2^3 if and only if for all set $\{c_1, c_2, c_3, c_4\}$ of real numbers and all finite sets $\{x_1, x_2, x_3, x_4\}$ of points in X, the implication*

$$\sum_{i=1}^{4} c_i = 0 \Rightarrow \sum_{i,j=1}^{4} c_i c_j d(x_i, x_j)^2 \leq 0 \tag{10}$$

holds.

5.3. Jensen-Shannon Distance

LEMMA 5.8 (TOPSØE [FUGLEDE AND TOPSØE 2004]). *For an appropriate definition of Jensen-Shannon divergence (JSD), the space $(M_+^1(A), \sqrt{JSD})$ is isometrically isomorphic to a subset in Hilbert Space.*

The term Jensen-Shannon divergence is used variously with slightly different meanings; to avoid ambiguity, we define it here as

$$JSD(v, w) = 1 - \tfrac{1}{2} \sum_i (h(v_i) + h(w_i) - h(v_i + w_i))$$

where

$$h(x) = -x \log_2 x$$

which formulation, explained in [Connor et al. 2013], is consistent with other authors and neatly bounds the range into [0,1].

Here, the set $M_+^1(A)$ is the set of probability distributions, which we can safely interpret as a set of positive numeric vectors $\{v\} \in \mathbb{R}^n$ for some n where $\sum_i^n v_i = 1$ (although the original definition extends to continuous spaces as well.) Topsøe uses Schoenberg's conjecture to prove this property by showing that JSD is itself a negative semidefinite mapping with the semi-metric properties. Although it has already been proved by more than one author that Jensen-Shannon distance (with the meaning of $\sqrt{JSD}$ in Topsøe's notation) is a proper metric ([Endres and Schindelin 2003],[Österreicher and Vajda 2003]) this proof of Hilbert space embedding gives that as a rather more elegant side-effect.

THEOREM 5.9. *The space $(M_+^1(A), \sqrt{JSD})$ is isometrically 4-embeddable in ℓ_2^3, and can therefore use Hilbert Exclusion with hyperplane partitioning.*

This is now a direct consequence of Lemmata 5.6 and 5.8.

5.4. Triangular Distance

To establish the generality of our results, we give one more example of a proper metric which is also Hilbert space embeddable and can therefore be indexed using Hilbert Exclusion.

The function

$$k(v, w) = \sum_i \frac{(v_i - w_i)^2}{v_i + w_i}$$

(where $v, w \in \mathbb{R}^n, \sum_i v_i = \sum_i w_i = 1$) has been identified and named in [Topsøe 2000] as *Triangular Discrimination*. Although rarely used in pratice, it is of significant interest as it has relatively tight upper and lower bounds over the much more expensive Jensen-Shannon distance [Topsøe 2000]. k is a semi-metric, so if it is negative semidefinite then $\sqrt{k}$ is a Hilbert-embeddable proper metric.

As k is a summation it is sufficient to prove that

$$f(x, y) = \frac{(x - y)^2}{x + y}$$

is conditionally negative semidefinite. Recalling the definition of negative semidefinite (Equation 10) we require

$$\sum_{i,j} \frac{(x_i - x_j)^2}{x_i + x_j} c_i c_j \leq 0$$

for any finite set of real numbers $(c_i)_{i \leq m}$ such that $\sum_i c_i = 0$ and for any finite set $(x_i)_{i \leq n}$ of points in X.

Observing that $(x_i - x_j)^2 = (x_i + x_j)^2 - 4x_i x_j$ we obtain

$$\sum_{i,j}^{m} c_i c_j \frac{(x_i - x_j)^2}{x_i + x_j} = \sum_{i,j}^{m} c_i c_j x_i + \sum_{i,j}^{m} c_i c_j x_j - 4 \sum_{i,j}^{m} c_i c_j \frac{x_i x_j}{x_i + x_j}$$
$$= -4 \sum_{i,j}^{m} c_i c_j \frac{x_i x_j}{x_i + x_j}$$

as the first two terms sum to zero. Thus it is sufficient to prove that

$$\sum_{i,j}^{m} c_i c_j \frac{x_i x_j}{x_i + x_j} \geq 0$$

As the index i, j such $x_i = 0$ or $x_j = 0$ do not contribute to the summation, we can assume that all the x_i, x_j are positive.

$$\sum_{i,j}^{m} c_i c_j \frac{x_i x_j}{x_i + x_j} = \sum_{i,j}^{m} c_i c_j x_i x_j \int_0^{\infty} e^{-t(x_i + x_j)} dt$$
$$= \int_0^{\infty} \sum_{i,j}^{m} c_i c_j x_i x_j e^{-t(x_i + x_j)} dt$$
$$= \int_0^{\infty} \left(\sum_{i}^{m} c_i x_i e^{-t x_i} \right) \left(\sum_{j}^{m} c_j x_j e^{-t x_j} \right) dt$$
$$= \int_0^{\infty} \left(\sum_{i}^{m} c_i x_i e^{-t x_i} \right)^2 dt \geq 0$$

This therefore gives us that

$$D_{\text{tri}}(v, w) = \sqrt{\sum_i \frac{(v_i - w_i)^2}{v_i + w_i}}$$

which we name as Triangular Distance, is a proper metric such that $(M_+^1(A), D_{\text{tri}})$ is a metric space which is isometrically embeddable in Hilbert space.

5.5. Spaces with Cosine Distance

The term "Cosine" distance does not have a unique meaning in the metric space literature and so requires an explanation.

It has long been known that, for two values v, w in $\mathbb{R}^n$, then the function

$$S_{\text{Cos}}(v, w) = \frac{v \cdot w}{\|v\| \|w\|}$$

gives a convenient estimate of their dimensional correlation. One advantage of this is that it is cheap to calculate, especially when the space is sparse such as applications in information retrieval. This function calculates the cosine of the angle between the vectors, and is best referred to as the Cosine Similarity Coefficient.

As it is bounded in $[0, 1]$, the function $f(v, w) = 1 - S_{\text{Cos}}(v, w)$ gives a bounded divergence coefficient; however this function is not a proper metric, as it lacks triangle inequality. A function which gives the same rank order and is also a proper metric can be simply achieved by converting this value into the angle between two vectors, which can be caused to range within $[0, 1]$ by $d_{\text{Cos}}(v, w) = cos^{-1}(S_{\text{Cos}}(v, w))/\pi$. In the metric

space literature, this function is sometimes referred to as Cosine Distance [Figueroa et al. 2007; Connor and Moss 2012].

This function is a proper metric, but is not isometrically embeddable in Hilbert space. However, there exists another rank-equivalent function based on the Cosine similarity:

$$\widetilde{d}_{\text{Cos}}(v, w) = \sqrt{1 - S_{\text{Cos}}(v, w)}$$

In fact, since $\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2v \cdot w$, the distance $\widetilde{d}_{\text{Cos}}(v, w)$ is equivalent to the Euclidean distance computed on the normalized vectors $v/\|v\|$ and $w/\|w\|$:

$$\widetilde{d}_{\text{Cos}}(v, w) = \widetilde{d}_{\text{Cos}}\left(\frac{v}{\|v\|}, \frac{w}{\|w\|}\right) = \frac{1}{\sqrt{2}}\left\|\frac{v}{\|v\|} - \frac{w}{\|w\|}\right\|$$

and is therefore isometrically 4-embeddable in three dimensional Euclidean space, and hence in a Hilbert space.

5.6. High-Dimensional Euclidean Space

For completeness we reconsider n-dimensional Euclidean space for any n in the context of Hilbert embedding. From Lemmata 5.4 and 5.6 it is sufficient to show that the function $K(v, w) = \sum_i (v_i - w_i)^2$ is a conditionally negative semi-definite semi-metric, which is straightforward to demonstrate using a similar proof to that used in Section 5.4.

5.7. Non-Embeddable Spaces

To complete the picture, it is worth mentioning that not all metric spaces are 4-embeddable in ℓ_2^3; it is therefore necessary to make a proper assessment of the space in question before using Hilbert Exclusion.

Among commonly used proper metrics in the domain of metric search are Manhattan (ℓ_1), Chebyshev (ℓ_∞), Hamming and Levenshtein distances. We briefly show that none of these has the four-point property.

Figure 7 shows, on the left, a square $ABCD$ in the 2D Cartesian plane and, on the right, the ℓ_1 distances between the vertices of the square. Since $\ell_1(A, C) = \ell_1(A, B) + \ell_1(B, C)$, for any isometric embedding of A, B, C in ℓ_2^3 the three points are collinear. However, this is also true for A, D and C, and the distances are the same. Therefore, the four points A, B, C, D cannot be isometrically embedded in ℓ_2^3, as this would require that $\ell_1(B, D) = 0$.

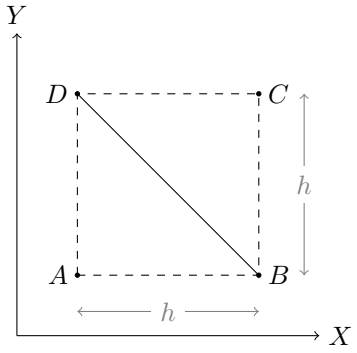
Similarly in Figure 8, $\ell_\infty(A, C) = \ell_\infty(A, B) + \ell_\infty(B, C)$ and so the points are collinear for any isometric embedding in ℓ_2^3; again, D shares the same distances ($\ell_\infty(A, C) = \ell_\infty(A, D) + \ell_\infty(D, C)$), and so a four-point embedding again cannot be achieved as this would require $\ell_\infty(B, D) = 0$.

To show that Hamming distance does not have the four-point property, it can simply be noted that the same distance table as that in Figure 7 is generated by the values $A = 00, B = 10, C = 11, D = 01$. In fact this counterexample gives the same distances also for Levenshtein distance, which therefore also can be seen to not have the four-point property.

5.8. The Four-point Property and Discriminability

In [Blumenthal 1953] it is shown that if (X, d) is a metric space then (X, d^α), with $0 \leq \alpha \leq 1/2$, is isometrically 4-embeddable in ℓ_2^3. It is therefore true that, for any proper metric, a new metric with the four-point property can be formed by taking its square root, and thus used in a metric search structure with Hilbert Exclusion.

However for practical purposes this is unlikely to be useful. One of the most important efficiency issues with metric search is the discriminability of the space. Whenever

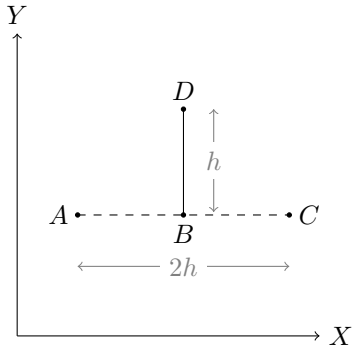


(a)

	A	B	C	D
A		h	$2h$	h
B			h	$2h$
C				h
D				

(b)

Fig. 7: Showing that ℓ_1 does not have the four-point property. Let A, B, C, D be the vertices of a square in ℓ_1^2. On the right we show the ℓ_1 distances between the points. Any isometric embedding in ℓ_2^3 maps A, B, C to collinear points, meaning that the point D cannot be embedded whilst preserving the distances.



(a)

	A	B	C	D
A		h	$2h$	h
B			h	h
C				h
D				

(b)

Fig. 8: The ℓ_∞ distance also does not have the four-point property. On the left, we plot four points in ℓ_∞^2 which are not embeddable ℓ_2^3. On the right we report the ℓ_∞ distances between the points. Once again, any isometric embedding must map A, B, C to collinear points in ℓ_2^3 making D impossible to embed.

any kind of exclusion mechanism is applied during a search, if the search radius is greater than zero, there is always a finite probability of no data being excluded — the lower the discriminability of the space, the higher the probability of no exclusion being made. The quantification of this effect is not yet well understood in detail.

The notion of Intrinsic Dimensionality (IDIM, [Chávez et al. 2001]) is known to give a reasonable estimate of (in)discriminability. IDIM is defined over a sample of distances calculated over randomly selected points from within the space, based on the mean μ and standard deviation σ of these distances, as $\frac{\mu^2}{2\sigma^2}$.

Raising a metric to a small power will cause a significant increase in its IDIM. As any such transform is monotonic increasing and convex, the standard deviation will decrease relative to the median, and threshold distances will increase in relative terms. The advantages of using Hilbert Exclusion are thus likely to be outweighed by these factors.

6. ANALYSIS

As Hilbert Exclusion is strictly weaker than Hyperbolic Exclusion, the performance of any partition-based indexing mechanism would always be expected to be superior. The distance between the pivot points is required as well as the distance between each pivot and the query, however this may always be calculated during the building of any indexing structure and adds nothing to the cost of a query. Query evaluation cost is totally dominated by the number of dynamic distance calculations required and the use of memory where the objects are large; the minor increase in arithmetic cost, and the extra space required to store the distance between pivots, would not normally make a significant difference to the query cost.

The many different index mechanisms reported show that performance is highly dependent on many factors, not least the cost of a distance calculation, the size of the objects, and other factors including the intrinsic dimensionality and the distribution of the data within the space. Furthermore most of the more sophisticated mechanisms use a mixture of hyperplane and cover radius exclusion; it may be that enhanced performance of hyperplane exclusion could make a significant difference to the choice of index. It is not therefore possible to analyse a simple "performance improvement" in general terms.

We therefore give analysis of the improved exclusion condition as follows.

(1) Exclusion power: for a given finite space, we randomly select pairs of pivot points that partition a space into two halves. The exclusion power of each mechanism can then be measured as the probability of a randomly-selected query being able to avoid searching either half of the space based only on its distance from the two points. This is always greater for Hilbert Exclusion than for Hyperbolic Exclusion; in Section 6.2 we give figures for various spaces.
(2) Improvement: for a given metric space, simple data structures relying primarily on hyperplane partitioning are built, namely a generalised hyperplane tree and a monotonic hyperplane tree. The same index structures can be used with either Hilbert or Hyperbolic exclusion; improvement is measured as a simple multiplicative factor between the two. We give results in Section 6.3.
(3) Real-world data: The SISAP forum[5] publishes a number of large data sets drawn from real world contexts which are commonly used as benchmarks for different indexing mechanisms. Results over these have been reported for many different indexing mechanisms. We take the best of these mechanisms, which uses both radius and hyperplane exclusion, and compare it using Hyperbolic and Hilbert exclusion mechanisms. Results for this are given in Section 6.4.

All of our measurements are expressed in terms of the number of distances calculated, and we do not give any measured execution times. While it is often the case that counting only the number of distances is not a good measure of overall efficiency of an indexing mechanism, as we are only measuring a comparison of exclusion mechanisms, the number of distances is the more important outcome. In fact our measured times for the fairly simple experiments we perform are approximately proportional, as the search structures are built over relatively small data sets which fit wholly within main memory and the cost of distance calculations is dominant.

[5] www.sisap.org

6.1. Experimental Method

Any exclusion mechanism will perform better as either the dimensionality of the space, or the threshold of the search, becomes smaller. To give a general overview of the tradeoffs, we perform all tests over a variety of spaces and thresholds.

In all cases, we generated pseudo-random data sets of one million elements within the unit hypercube, uniformly distributed within each dimension, within $\mathbb{R}^d$ for $d \in \{6, 8, 10, 12, 14\}$. In the results presented we name the spaces used based on the metric and the number of Cartesian dimensions, eg `euc_10` for Euclidean distance over $\mathbb{R}^{10}$, `jsd_12` for Jensen-Shannon distance[6] over $\mathbb{R}^{12}$ etc.

Search thresholds were derived by experiment, for each space, as those which would return around n results per million data, for $n \in \{1, 2, 4, 8, 16, 32\}$.

For each space we also calculated the Intrinsic Dimensionality (IDIM, [Chávez et al. 2001]), generally believed to give a reasonably good estimate of the tractability of a space to metric indexing techniques. A common observation is that spaces with an IDIM of greater than around 6 are challenging, and those with an IDIM of greater than about 10 are usually intractable. IDIM is defined over a sample of distances calculated over randomly selected points from within the space, based on the mean μ and standard deviation σ of these distances, as $\frac{\mu^2}{2\sigma^2}$.

Table I in Appendix B gives values for IDIM and thresholds calculated for each space. Given these values, all experimental results are obtainable through repetition of the experiments described. All results are independent of the computer upon which they are performed, and all figures presented represent mean values where experiments were repeated until the standard error of the mean was less than 1% of the value given.

6.2. Exclusion Power

Figures 9, 10 and 11 illustrate the exclusion power test. Each figure shows the same set of 500 randomly generated points in a 10-dimensional Euclidean space. A futher two points are also generated to act as pivots.

In Figures 9 and 10, the distance between the pivot points is measured as d; an embedded 2D plane is then constructed with these points at $(0, -1/2d)$ and $(0, 1/2d)$ respectively. Each point in the generated set is then measured against these two points, and plotted in the upper half of the plane according to these distances. It can be seen that the same points are plotted in both figures. Note that the relative distances within the plot are of no significance; each point represents a different embedding function. However the position of each point within the space is individually significant with respect to the pivot points.

A query radius is chosen to return around one point per million from a large set. Figure 9 highlights those points which satisfy the Hyperbolic Exclusion condition, and Figure 10 highlights those which satisfy Hilbert Exclusion. As well as noting that the number is substantially greater (201 against 75 in this example), it is instructive to note the shape of the exclusion zones within the two figures; Figure 9 clearly shows the shape of the hyperbola which demarcates the zone, whereas Figure 10 clearly shows parallel lines to either side of the central axis.

To give a reference diagram for single pivot-based exclusion, Figure 11 gives the same plot but highlights those which are more than the same query threshold from the median distance to the left-hand pivot point, which are those that could be excluded according to radius-based exclusion from this point alone; there are 139 of these in this case.

[6]for `euc` and `tri`, each point is normalised so that $\sum_i v_i = 1$

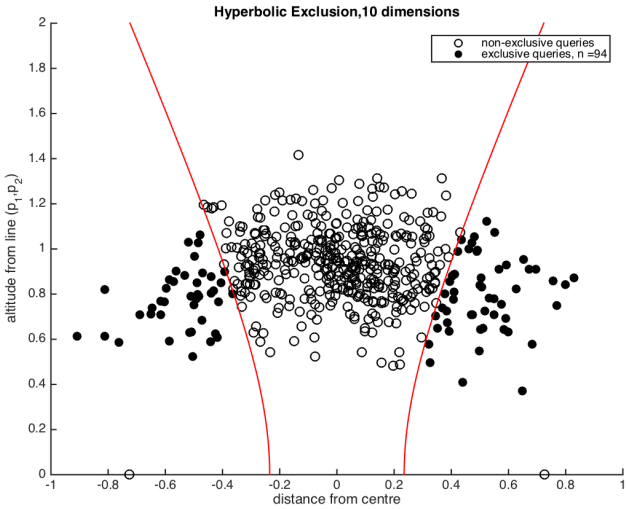


Fig. 9: Hyperbolic Exclusion. Two points are chosen at random from a finite space and placed symmetrically on the X axis, either side of the origin, separated by the distance between them in the original space. The remaining points are plotted in the upper half of the space according to their distance from these two points. Relative distances among these points are not significant as each point represents a different embedding function. Those coloured solidly are those which, were they queries, would allow the semispace on the opposing side to be excluded from a search.

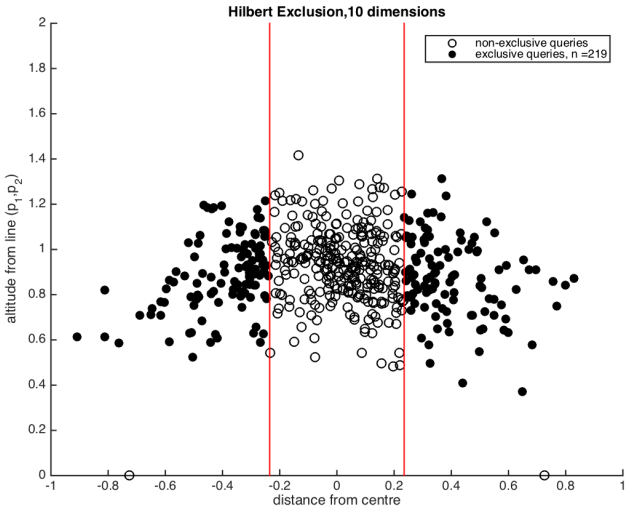


Fig. 10: Hilbert Exclusion. The same plot as in Fig 9; the solidly coloured points represent queries that allow the opposing semispace to be excluded using Hilbert Exclusion. These are now all points at least the threshold distance from the separating hyperplane, which includes many more queries for the same threshold.

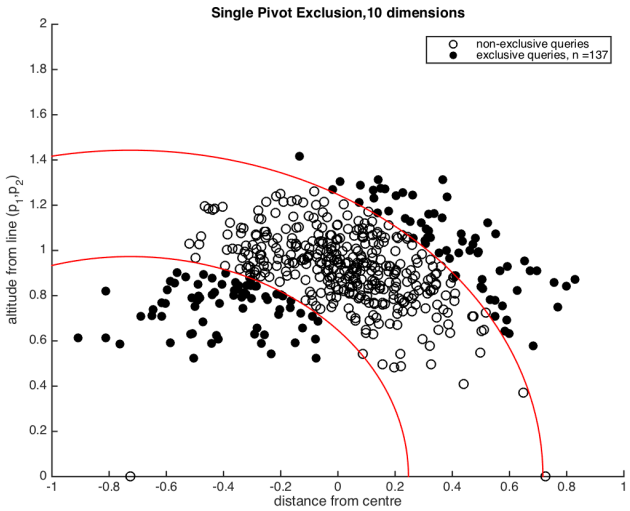


Fig. 11: Pivot Exclusion. The same plot as Figs 9 and 10, but now the left-hand point on the X axis is used to exclude queries based on distance from that alone. Semispaces are defined according to the median distance from this point, and solidly coloured points indicate those whose distance from the pivot point is more than the query threshold away from this median.

In all spaces that we have measured, the single-pivot method has more exclusion power than Hyperbolic exclusion, but less power than Hilbert Exclusion. In metric indexing things are not this simple, as in particular hyperplane separation is normally used to effect in conjunction with cover radius exclusion. The greater exclusion potential of Hilbert Exclusion requires two distance calculations, against a single calculation for pivot-based exclusion; however many indexes have ways of amortising this extra cost. Finally, plane partitioning is very effective when the space is amenable to geometric separation, as it tends to cluster subsets which are relatively closer to each other, whereas ball partitioning tends to be less effective in this respect.

In all there is a hint that, when applicable, the new condition appears to enjoy the best of all worlds in this respect; at least it may make a significant difference to the choice of mechanism for a given data set, and may possibly inspire new mechanisms to be developed.

6.2.1. Results. Table II in Appendix C gives outcomes of the exclusion power test for the three given Hilbert-embeddable metrics over spaces of various dimensions, using various query thresholds. These results are graphically summarised in Figure 12 for Euclidean spaces; the other two metrics give very similar patterns. The left-hand figure shows the exclusion percentage obtained at various dimensions and thresholds; it can be seen that Hilbert Exclusion performs much better than Hyperbolic Exclusion, and is much more tolerant to increases in both dimensionality and query threshold; that is, it performs relatively better as the space becomes less tractable.

The right hand graphs illustrates this in terms of improvement of Hilbert over Hyperbolic exclusion, which again can be seen to increase sharply as the space becomes less tractable.

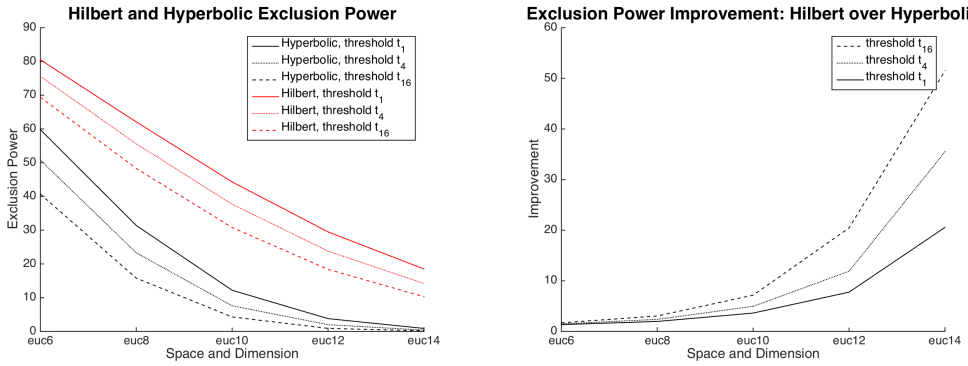


Fig. 12: Exclusion Power tests: Each figure shows five different dimensionalities, and three different search thresholds, for Euclidean spaces. Jensen-Shannon and Triangular spaces gives similar results. Left figure is percentage exclusion, right is relative improvement of Hilbert over Hyperbolic.

6.3. Improvement

To give a more practical measurement of performance improvement, the two exclusion mechanisms have also been tested over metric indexes built over actual data sets. The indexes used are the general hyperplane tree (GHT, [Uhlmann 1991]) and the monotonic hyperplane tree (MHT, [Noltemeier et al. 1992])[7], which are in a sense the most "pure" (and certainly the simplest) hyperplane indexing structures. In these experiments, for each data set used the same data structure is created, the only difference is in the exclusion mechanism used.

It should be noted here that the notions of "bisector" and "hyperplane" tree are conceptually different; although they share the same construction algorithm, bisector trees use a cover radius for pivot-based exclusion, and hyperplane trees use, normally, hyperbolic exclusion. In our experiments we use both cover radius and hyperplane exclusion mechanisms, as would be normal in practice, and compare the use of hyperbolic exclusion with Hilbert exclusion.

6.3.1. Results. Table III in Appendix C shows, for various metrics and dimensionalities, the cost of indexing two hyperplane-based metric index structures with the different exclusion strategies. Figure 13 shows some of the results in graphical form.

It can be seen that, for all spaces, Hilbert Exclusion always gives better performance than Hyperbolic Exclusion; this is expected, as the exclusion condition is strictly weaker. Table III shows that, under Hyperbolic Exclusion, the MHT always gives marginally improved performance over the GHT; again, this is already known and understood. It can also be seen that the GHT under Hilbert Exclusion gives equal or better performance than the MHT under Hyperbolic Exclusion. Interestingly however, the improvement given by using Hilbert Exclusion with the MHT is dramatically better than the improvement given over the GHT, for which we do not currently have a reason.

Another interesting observation is shown on the right of Figure 13, which gives the ratio of the number of distances calculated by the MHT for the two exclusion mechanisms; it can be seen that, for all search thresholds, this reaches a maximum at around 10 dimensions and then decreases again. This can be explained by the fact

[7]Originally named the "Monotonous Bisector* Tree"; the term "monotonic" is generally agreed to be a better description of the concept

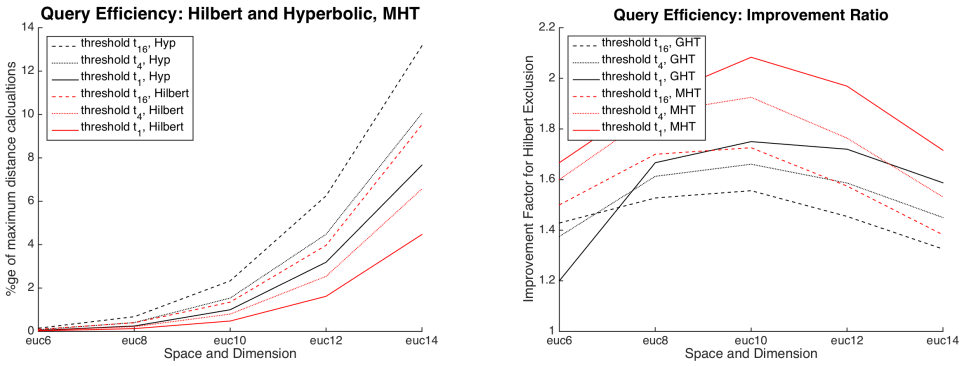


Fig. 13: The left hand graph shows absolute performance data for the MHT, at various dimensionalities and thresholds, for the two exclusion mechanisms. The right hand graph shows the same results interpreted as an improvement ratio, and also includes data from the GHT. In the left hand graph, lines of the same pattern represent the same data, and the same index structures, only the query exclusion mechanism is different.

that, for very tractable spaces, both mechanisms function very well; there is not therefore a great improvement. For intractable spaces, neither mechanism can do well and so again the relative improvement becomes less. The observation is in keeping with the left hand diagram shown in Figure 12, where it be seen that the gap in exclusion power of the two mechanisms is greatest at around the same range of dimensions.

6.4. "Real-world" data

There are many different contexts for metric search, and no mechanism is generally believed to be best for all purposes. The most competitive comparator at the time of writing is the Distal Spatial Approximation Tree (DiSAT) [Chávez et al. 2016] which has been shown to perform better than a large range of other mechanisms. The authors write:

> "Our data structure has no parameters to tune-up and a small memory footprint. In addition it can be constructed quickly. Our approach is among the most competitive, those outperforming DiSAT achieve this at the expense of larger memory usage or an impractical construction time."

We can therefore take this mechanism as the state of the art in metric indexing, and as it uses hyperplane partitioning we can test the effect of applying Hilbert Exclusion against the Hyperbolic Exclusion with which it has been defined. In their publication, the authors test the DiSAT very extensively and it is in almost all cases the best performing index.

The SISAP forum[8] publishes a number of large data sets drawn from real world contexts which are commonly used as benchmarks for different indexing mechanisms, and results for the DiSAT were given with respect to these. We have implemented the DiSAT as described in [Chávez et al. 2016] and measured the same results over Euclidean spaces; therefore we need only compare this structure with the two different exclusion criteria.

[8] www.sisap.org

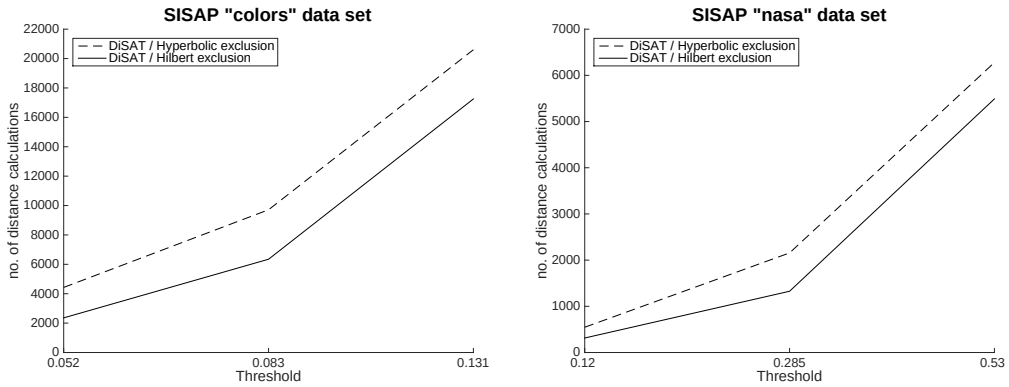


Fig. 14: Comparing Hyperbolic and Hilbert Exclusion Conditions for the DiSAT. The two graphs represent benchmark applications of the "state of the art" DiSAT index over SISAP benchmark data sets, with significant improvements achieved through changing only the exclusion condition.

The same experimental context was used: the SISAP "colors" and "nasa" data sets are used to build instances of DiSATs. The "nasa" data set contains a set of 40,700 20-dimensional feature vectors, generated from images downloaded from NASA. The "colors" set contains 112,682 feature vectors of dimension 112, representing color histograms from an image database.

In each case ten percent of the data is used as queries over remaining 90 percent of the set, at threshold values which return 0.01%, 0.1% and 1% of the data sets respectively. Figure 14 shows the outcome of these experiments. It is clear that using Hilbert exclusion greatly improves the performance.

6.5. Correctness

It is finally worth mentioning that during the course of the experiments described in this paper, over one million queries have been executed over sets of at least one million data using a number of different indexes, including those using both Hyperbolic and Hilbert exclusion; all queries over the same sets, using different mechanisms, have been checked against each other and in all cases the results were identical. While we are confident about the correctness of the mathematical derivations given, it is nonetheless comforting to have such experimental validation.

7. THE EFFECTS OF INCREASING DIMENSIONALITY

The results given have shown how the relative advantage of Hilbert Exclusion over Hyperbolic Exclusion increases as the spaces become less tractable, that is as the intrinsic dimensionality increases.

A reason for this can be seen from studying the geometry of the two mechanisms in the three dimensional embeddings. As the dimensionality increases, there are three well-known effects: the mean distances between randomly sampled points increases; the standard deviation of these distances decreases,and query thresholds greatly increase. This last gives the greatest effect in terms of the tractability of indexing mechanisms, and is an effect of the relative ratio of the volume of the unit hypercube and the unit hypersphere as dimensions increase. The volume of the unit hypersphere in $2k$ dimensions is $\frac{\pi^k}{k!}$, which decreases very rapidly after three dimensions, whereas the volume of the unit hypercube remains as 1, independent of the dimension.

As can be seen from Table I, in 6-dimensional Euclidean space the radius of a hypersphere with a volume of 10^{-6} is 0.076; in 14-dimensional Euclidean space it is 0.386. This has the effect of not only making the hyperbola wide, but also causing it to veer sharply away from the central hyperplane.

Figure 15 illustrates this effect by illustrating the siutation in both 6 and 14 dimensions for a small set of 500 randomly generated points in the unit hypercube.

8. CONCLUSIONS AND FURTHER WORK

We have shown that many common metric spaces have a further, stronger, property: namely, as well as the ability to isometrically embed any three points in two-dimensional Euclidean space, they also have the ability to isometrically embed any four points in three-dimensional Euclidean space. We have shown how the stronger geometric guarantee allows more effective metric indexing, and also that any metric space which is isometrically embeddable in Hilbert space has the stronger property. Such spaces include those most commonly used, including spaces of any dimension governed by Euclidean, Jensen-Shannon, Triangular or Cosine distance.

We have shown that, for such spaces, the most popular, state-of-the-art indexing mechanisms have significantly better performance, and that the improvement increases as the dimensionality of the space increases, which is an important result in this field.

However we believe that the so-called four point property will turn out to also be of value in other areas of similarity search. Although not yet fully investigated, we have included here the observation that our Hilbert Exclusion has better properties than normal pivot-based exclusion over a single object, and while Hilbert exclusion has the disadvantage of requiring two reference points, it has been seen (for example in monotonic bisector trees) how this extra cost can be amortised by reusing the pivot points. We have also made some early but promising observations that the four-point property can be used to effect beyond indexing structures, for example in the use of locality-sensitive hashing and permutation ordering, which we are currently investigating further.

In essence, almost the entire literature of metric search is based upon the property of 3-embeddability in two dimensional space; almost every derived result in the whole domain can be usefully re-examined in terms of the stronger property of 4-embeddability in three dimensional space.

Finally, it is also the case that any Hilbert space with the four-point property in fact has the ability to embed any n points with $(n-1)$-dimensional Euclidean space; we are currently trying to understand if this property gives rise to further uses within metric indexes.

REFERENCES

Aleksandr Danilovich Aleksandrov, Andre Nikolaevich Kolmogorov, and Mikhail Alekseevich Lavrent'ev. 1999. *Mathematics: Its Content, Methods and Meaning (Dover Books on Mathematics)*. Dover Publications.

Leonard M. Blumenthal. 1933. A note on the four-point property. *Bull. Amer. Math. Soc.* 39, 6 (1933), 423–426.

Leonard M. Blumenthal. 1953. *Theory and applications of distance geometry*. Clarendon Press. 347 pages.

Edgar Chávez, Verónica Ludueña, Nora Reyes, and Patricia Roggero. 2014. Faster proximity searching with the distal SAT. In *Similarity Search and Applications - 7th International Conference, SISAP 2014, Los Cabos, Mexico, October 29-31, 2014. Proceedings (Lecture Notes in Computer Science)*, Agma Juci Machado Traina, Caetano Traina, and Robson Leonardo Ferreira Cordeiro (Eds.). Springer International Publishing, 58–69. DOI:http://dx.doi.org/10.1007/978-3-319-11988-5_6

Edgar Chávez, Verónica Ludueña, Nora Reyes, and Patricia Roggero. 2016. Faster proximity searching with the distal SAT. *Information Systems* 59 (2016), 15 – 47. DOI:http://dx.doi.org/10.1016/j.is.2015.10.014

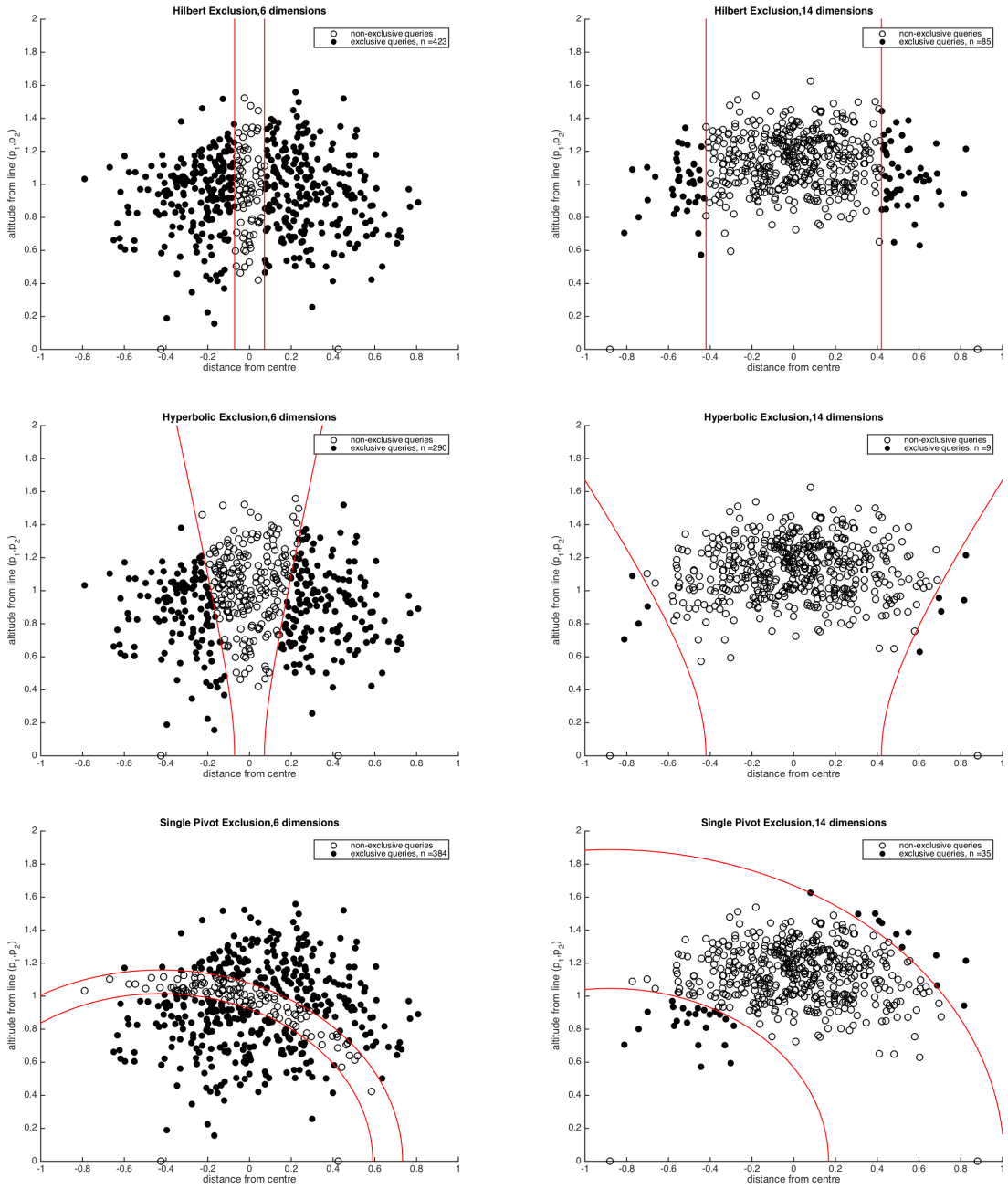


Fig. 15: The effect of dimensionality increase on the three "power" plots. At 6 dimensions (left hand column) the hyperbola can be clearly seen to disadvantage Hyperbolic Exclusion (middle row) against the parallel lines given by Hilbert Exclusion (top row.) At 14 dimensions however, Hyperbolic Exclusion excludes only a handful of points, and Hilbert Exclusion achieves significantly more exclusion than single-point pivoting (bottom row.) Query thresholds are chosen to return one per million objects.

Edgar Chávez and Gonzalo Navarro. 2005. Metric Databases. In *Encyclopedia of Database Technologies and Applications*, Laura C. Rivero, Jorge Horacio Doorn, and Viviana E. Ferraggine (Eds.). Idea Group, 366–371.

Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. 2001. Searching in Metric Spaces. *ACM Comput. Surv.* 33, 3 (Sept. 2001), 273–321. DOI:http://dx.doi.org/10.1145/502807.502808

Richard Connor, Franco Alberto Cardillo, Robert Moss, and Fausto Rabitti. 2013. *Similarity Search and Applications: 6th International Conference, SISAP 2013, A Coruña, Spain, October 2-4, 2013, Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Evaluation of Jensen-Shannon Distance over Sparse Data, 163–168. DOI:http://dx.doi.org/10.1007/978-3-642-41062-8_16

Richard Connor and Robert Moss. 2012. A Multivariate Correlation Distance for Vector Spaces. In *Similarity Search and Applications*, Gonzalo Navarro and Vladimir Pestov (Eds.). Lecture Notes in Computer Science, Vol. 7404. Springer Berlin Heidelberg, 209–225. DOI:http://dx.doi.org/10.1007/978-3-642-32153-5_15

Dominik Maria Endres and Johannes E. Schindelin. 2003. A new metric for probability distributions. *Information Theory, IEEE Transactions on* 49, 7 (2003), 1858–1860. DOI:http://dx.doi.org/10.1109/TIT.2003.813506

Karina Figueroa, Gonzalo Navarro, and Edgar Chávez. 2007. Metric Spaces Library. www.sisap.org/library/manual.pdf. (2007).

Bent Fuglede and Flemming Topsøe. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *IEEE International Symposium on Information Theory*. 31–31.

Karl Menger. 1931. New Foundation of Euclidean Geometry. *American Journal of Mathematics* 53, 4 (1931), 721–745. http://www.jstor.org/stable/2371222

Gonzalo Navarro. 2002. Searching in metric spaces by spatial approximation. *The VLDB Journal* 11, 1 (2002), 28–46. DOI:http://dx.doi.org/10.1007/s007780200060

Gonzalo Navarro and Nora Reyes. 2002. *String Processing and Information Retrieval: 9th International Symposium, SPIRE 2002 Lisbon, Portugal, September 11–13, 2002 Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Fully Dynamic Spatial Approximation Trees, 254–270. DOI:http://dx.doi.org/10.1007/3-540-45735-6_23

Hartmut Noltemeier, Knut Verbarg, and Christian Zirkelbach. 1992. *Data structures and efficient algorithms: Final Report on the DFG Special Joint Initiative*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Monotonous Bisector* Trees — a tool for efficient partitioning of complex scenes of geometric objects, 186–203. DOI:http://dx.doi.org/10.1007/3-540-55488-2_27

David Novak, Michal Batko, and Pavel Zezula. 2011. Metric Index: An efficient and scalable solution for precise and approximate similarity search. *Information Systems* 36, 4 (2011), 721 – 733. DOI:http://dx.doi.org/10.1016/j.is.2010.10.002 Selected Papers from the 2nd International Workshop on Similarity Search and Applications {SISAP} 2009.

Ferdinand Österreicher and Igor Vajda. 2003. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.* 55 (2003), 639–653. DOI:http://dx.doi.org/10.1007/BF02517812

Isaac J. Schoenberg. 1938. Metric Spaces and Completely Monotone Functions. *Annals of Mathematics* 39, 4 (1938), 811–841. http://www.jstor.org/stable/1968466

Sebastian Scholtes. 2013. A characterisation of inner product spaces by the maximal circumradius of spheres. *Archiv der Mathematik* 101, 3 (2013), 235–241. DOI:http://dx.doi.org/10.1007/s00013-013-0556-6

Flemming Topsøe. 2000. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory* 46, 4 (Jul 2000), 1602–1609. DOI:http://dx.doi.org/10.1109/18.850703

Flemming Topsøe. 2003. Jenson-shannon divergence and norm-based measures of discrimination and variation. *preprint* (2003).

Jeffrey K. Uhlmann. 1991. Satisfying general proximity / similarity queries with metric trees. *Inform. Process. Lett.* 40, 4 (1991), 175 – 179. DOI:http://dx.doi.org/10.1016/0020-0190(91)90074-R

Wallace A Wilson. 1932. A relation between metric and euclidean spaces. *American Journal of Mathematics* 54, 3 (1932), 505–517. http://www.jstor.org/stable/2370894

Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. 2006. *Similarity search: the metric space approach*. Advances in Database Systems, Vol. 32. Springer.

9. APPENDICES

A. ALGEBRAIC PROOF OF WEAKNESS

Here we prove that the Hilbert Exclusion Condition is weaker than the Hyperbolic Exclusion Condition. The intuition behind this is clear from the geometric derivation but the algebraic proof is straightforward.

We require to prove that

$$\frac{d(q,p_1)^2 - d(q,p_2)^2}{2\,d(p_1,p_2)} > t$$

is a weaker condition than

$$\frac{d(q,p_1) - d(q,p_2)}{2} > t$$

for which it is sufficient to show that

$$\frac{d(q,p_1)^2 - d(q,p_2)^2}{2\,d(p_1,p_2)} \geq \frac{d(q,p_1) - d(q,p_2)}{2}$$

Using the triangle inequality property on q, p_1 and p_2, this requirement can be stated as

$$\frac{a^2 - b^2}{2\,c} \geq \frac{a-b}{2} \quad , \quad c \leq a + b$$

and so

$$\frac{(a+b)(a-b)}{2\,c} \geq \frac{a-b}{2}$$

which is clear when $c \leq a + b$.

This proof also neatly demonstrates the fact that the conditions are equivalent only if the query point is colinear with the two pivots p_1 and p_2; in all other cases, the Hilbert Exclusion Condition is strictly weaker.

B. IDIMS AND QUERY THRESHOLDS

Table I: Intrinsic Dimensionality and Thresholds for Experimental Spaces

Space	IDIM	t_1	t_2	t_4	t_8	t_{16}	t_{32}
euc_6	7.698	0.076	0.085	0.095	0.107	0.120	0.135
euc_8	10.40	0.149	0.162	0.177	0.193	0.211	0.230
euc_10	13.36	0.228	0.245	0.262	0.281	0.301	0.323
euc_12	16.23	0.308	0.327	0.346	0.367	0.388	0.412
euc_14	19.13	0.386	0.406	0.426	0.448	0.471	0.495
jsd_6	5.162	0.022	0.026	0.030	0.035	0.040	0.046
jsd_8	7.273	0.045	0.051	0.057	0.064	0.071	0.078
jsd_10	9.486	0.067	0.073	0.079	0.086	0.094	0.102
jsd_12	11.51	0.084	0.091	0.099	0.107	0.114	0.122
jsd_14	13.69	0.103	0.111	0.118	0.126	0.133	0.141
tri_6	5.754	0.025	0.030	0.035	0.041	0.047	0.055
tri_8	8.181	0.053	0.060	0.068	0.075	0.083	0.091
tri_10	10.46	0.078	0.086	0.093	0.101	0.110	0.119
tri_12	13.02	0.098	0.106	0.116	0.125	0.133	0.142
tri_14	15.60	0.120	0.129	0.137	0.146	0.155	0.164

C. EXCLUSION POWER RESULTS

Table II: Exclusion Power results for various metrics, spaces and thresholds.

		Hyperbolic			Hilbert			Pivot		
Data Set	IDIM	t_1	t_4	t_{16}	t_1	t_4	t_{16}	t_1	t_4	t_{16}
euc_6	7.64	59.8	50.8	40.7	80.5	75.6	69.4	74.4	68.1	60.4
euc_8	10.5	31.4	23.3	15.8	62.1	55.6	48.3	51.8	44.2	36.0
euc_10	13.3	12.2	7.6	4.3	44.3	37.7	30.8	31.9	25.1	18.7
euc_12	16.1	3.8	2.0	0.9	29.5	23.8	18.4	17.4	12.7	8.6
euc_14	19.0	0.9	0.4	0.2	18.5	14.2	10.3	8.8	6.0	3.8
jsd_6	5.15	66.1	54.9	42.9	83.8	77.8	70.7	82.4	75.8	68.0
jsd_8	7.26	32.4	21.7	13.5	62.8	53.9	45.2	58.5	48.8	39.3
jsd_10	9.39	11.4	6.3	3.0	42.6	34.4	26.4	36.2	27.7	19.8
jsd_12	11.4	3.5	1.4	0.5	27.4	19.6	13.5	20.8	13.6	8.5
jsd_14	13.7	0.6	0.2	0.1	14.4	9.5	6.0	9.3	5.4	3.0
tri_6	5.76	63.7	51.9	39.7	82.3	75.8	68.2	80.4	73.1	64.6
tri_8	8.25	27.9	17.6	10.3	59.5	50.1	41.0	54.2	43.9	34.2
tri_10	10.6	8.1	4.1	1.8	38.0	29.7	21.8	31.0	22.8	15.4
tri_12	13.0	1.9	0.6	0.2	22.7	15.3	9.9	16.2	9.8	5.7
tri_14	15.5	0.3	0.1	0.0	10.8	6.6	3.8	6.2	3.3	1.6

Table III: Indexing Costs for General Hyperplane and Monotonic Hyperplane Tree: mean number of distance calculations per query as percentage of data size ($n = 10^6$).

	Hyperbolic						Hilbert					
	GHT			MHT			GHT			MHT		
Data Set	t_1	t_4	t_{16}	t_1	t_4	t_{16}	t_1	t_4	t_{16}	t_1	t_4	t_{16}
euc_6	0.06	0.11	0.20	0.05	0.08	0.15	0.05	0.08	0.14	0.03	0.05	0.10
euc_8	0.30	0.50	0.84	0.25	0.41	0.68	0.18	0.31	0.55	0.13	0.22	0.40
euc_10	1.19	1.86	2.91	1.00	1.54	2.33	0.68	1.12	1.87	0.48	0.80	1.35
euc_12	3.87	5.60	7.97	3.19	4.48	6.25	2.25	3.53	5.48	1.62	2.54	3.97
euc_14	9.92	13.18	17.26	7.67	10.06	13.17	6.25	9.09	13.02	4.47	6.57	9.53
tri_6	0.05	0.11	0.21	0.04	0.08	0.16	0.04	0.07	0.15	0.02	0.05	0.11
tri_8	0.40	0.78	1.41	0.32	0.62	1.10	0.23	0.48	0.92	0.17	0.35	0.69
tri_10	1.95	3.29	5.37	1.66	2.73	4.36	1.11	2.05	3.71	0.84	1.57	2.87
tri_12	6.10	9.84	14.49	5.25	8.24	12.04	3.74	6.86	11.27	2.92	5.43	9.04
tri_14	16.63	23.11	30.57	13.95	19.45	26.06	12.02	18.57	26.52	9.68	15.24	22.25
jsd_6	0.05	0.10	0.20	0.04	0.08	0.15	0.04	0.07	0.15	0.02	0.05	0.11
jsd_8	0.32	0.63	1.15	0.26	0.51	0.92	0.20	0.40	0.78	0.14	0.29	0.58
jsd_10	1.50	2.58	4.29	1.35	2.22	3.61	0.90	1.64	2.99	0.68	1.25	2.31
jsd_12	4.67	7.68	11.47	4.17	6.62	9.76	2.84	5.27	8.71	2.22	4.15	6.97
jsd_14	12.4	17.67	23.9	10.77	15.17	20.57	8.62	13.62	19.97	6.94	11.13	16.69