Sampled Angles in High-Dimensional Spaces

Richard Connor✉ and Alan Dearle

rchc@st-andrews.ac.uk al@st-andrews.ac.uk
University of St Andrews, Jack Cole Building, North Haugh, St Andrews,
Fife KY16 9SX, Scotland, UK

Abstract. Similarity search using metric indexing techniques is largely a solved problem in *low-dimensional* spaces. However techniques based only on the triangle inequality property start to fail as dimensionality increases.

Since proper metric spaces allow a finite projection of any three objects into a 2D Euclidean space, the notion of *angle* can be validly applied among any three (but no more) objects. High dimensionality is known to have interesting effects on angles in vector spaces, but to our knowledge this has not been studied in more general metric spaces. Here, we consider the use of angles among objects in combination with distances.

As dimensionality becomes higher, we show that the variance in sampled angles reduces. Furthermore, sampled angles also become correlated with inter-object distances, giving different distributions between query solutions and non-solutions. We show the theoretical underpinnings of this observation in unbounded high-dimensional Euclidean spaces, and then examine how the pure property is reflected in some real-world high dimensional spaces. Our experiments on both generated and real world datasets demonstrate that these observations can have an important impact on the tractability of search as dimensionality increases.

Keywords: metric search · high dimensional space

1 Introduction

The context of interest is searching a (large) finite set of objects S which is a subset of an infinite set U, where (U, d) is a metric space: that is, a pair (U, d), where U is a domain of objects and d is a total distance function $d : U \times U \to \mathbb{R}$, satisfying postulates of non-negativity, identity, symmetry, and triangle inequality. The general requirement is to efficiently find members of S which are similar to an arbitrary member of U given as a query, 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 general mathematical framework, see for example [4, 14]. There are two main types of query: *range* and *nearest-neighbour search.* The *range search* for some query $q \in U$ and threshold $t \in \mathbb{R}$ is defined as having the solution set $R = \{s \in S | \, d(q, s) \leq t\}$. More practical in many contexts is the *nearest-neighbour* (kNN) search where the solution set comprises the k closest objects to a query.

The essence of metric search is to spend time pre-processing the finite set S so that solutions to queries can be efficiently calculated. In all cases distances among members of S and selected *reference* or *pivot* objects are calculated during pre-processing. At query time the relative distances between the query and the same pivot objects can be used to make deductions about which data values may, or may not, be candidate solutions to the query. Such deductions are based upon the *triangle inequality* property of the metric.

1.1 Distances and angles

In this paper we consider not just the measured distances among objects, but also the angles implied by these distances. In any metric space, the triangle inequality property also implies a finite 3-embedding in 2D Euclidean space [5], and so it is valid to discuss the *angles* of a triangle constructed according to the distances among any three objects selected from the space. It is important to stress that, in this paper, this is the only notion of angle that we use; thus our discussion is valid with respect to any proper metric space, not just vector spaces.

In the context of metric search, we are interested in the distribution of angles $\angle pqs_i$ where objects p and q are fixed, and s_i is sampled within a relatively small bounded distance from q. This is typical of a situation where p represents some reference object, q represents a query object, and s_i is sampled from the solution objects of the query. Note that the situation described generalises to both range and nearest-neighbour queries.

In general, we compare this distribution of angles with the alternative distribution of $\angle pqx_i$ where the same p and q are selected, but where x_i is sampled from the entire metric space without constraint. We find that in many cases, especially in high-dimensional spaces, these distributions differ significantly. This information can be used to effect within existing metric access methods, and furthermore gives a geometric explanation to phenomena that have been previously observed, but not previously explained, in approximate indexing techniques.

The main observation of this paper is that the following usually hold in high-dimensional metric spaces:

- if three values p, q, x are randomly sampled, then the mean of the angle $\angle pqx$ is 60 degrees, and the variance in this measurement decreases as dimensionality increases;
- however if values p, q are fixed, and then s_i is sampled from a sufficiently small fixed distance bound of q, then the mean of the angle $\angle pqs_i$ is greater than 60 degrees, and again the variance decreases as dimensionality increases.

We show how these observations can be used to effect in approximate search techniques.

1.2 Contributions

It is generally known that, as the dimension of a vector space increases, the probability of two independently selected vectors being close to orthogonal increases. In Section 3 we show that in an unbounded Euclidean space, for any values of a and b, and c sampled within a distance bound t of b, the mean angle $\angle abc$ is 90°, and the variance decreases according to the dimension of the space. This corollary allows the angles to be calculated using only the distances among three objects, rather than from the values of two vectors, and thus allows the possibility of extension into general metric spaces.

We show that this effect can be used to determine an upper bound on the probability of a randomly selected point lying in the intersection of hyperspheres centred on a and b, and how this probability may be used to construct approximate search mechanisms. We show that this probability is related to the function $\sin^n \theta$ where n is the Euclidean dimension. This can lead to very low probabilities in some cases, and thus highly accurate approximations.

In Section 4 we show by experiment that this theory holds perfectly in a bounded uniform Euclidean space, as long as the position of b and the distance d are fixed to ensure that the hypersphere described by $<b, d>$ is fully enclosed in the space. However, in many high-dimensional search spaces, this does not hold. This is because the hypersphere $<q, t>$, where q is a query object and t is a distance bound which includes elements of the finite search space, may include a significant region that lies outside the boundaries of the space. Nonetheless for such spaces there still exists a predictable distribution of sampled angles which is different from randomly sampled angles. We introduce an observed correlation between *outlierness* and the distribution of angles in these spaces.

Finally in Section 5 we study some "real-life" high-dimensional metric search spaces to check if the theoretical observations still hold. We find that, while compromised from the pure model, there is still a useful distinction between the angle distributions of query solutions and non-solutions, and furthermore this is observable in non-Euclidean metric spaces as well as Euclidean spaces. We show some experiments which use a variant of LAESA to demonstrate a practical application of our observations. For all of the datasets used, a relaxation on the exclusion condition based on the angle-enhanced analysis allows substantially more exclusion while still maintaining almost perfect accuracy.

2 Related work

The distribution of vectors within high-dimensional vector spaces is discussed in a book chapter by Hopcroft and Kannan [2]. This introduces the notion of the 90° degree angle norm in discussion of an "annulus" within the hypersphere. However, the subtleties of the hyperspace being only partially embedded within the data space are overlooked.

In [3], the authors state that it is "a matter of folklore" that "all high dimensional random vectors are almost always nearly orthogonal to each other".

They quantify this with a probability density function, directly proportional to $\sin^{(n-2)}\theta$, for the angle between two vectors randomly sampled from a uniform distribution on the surface of an n-dimensional hypersphere. This is consistent with our distribution, which is proportional to $\sin^n\theta$, considering points uniformly distributed within a segment of the hypersphere. They also observe, without explanation, the close relationship between the function $\sin^n\theta$ and a related normal distribution, as we use in Section 3.

Pramanik et al [12] give an expression for the volume of a hypersphere in an n-dimensional Euclidean space, and imply a derived PDF for angles which is proportional to $\sin^n\theta$. They do not give a derivation of their volume formula. Perhaps as a result of this, their implication of the relationship between volume and probability is incorrect, and in fact the density function should be proportional to $\sin^{(n-2)}\theta$ as above.

They go on to use an angle-based relaxation of a ball partitioning mechanism to improve performance of a single-pivot mechanism, the *AB Tree*, which they show to be effective in terms of increased performance versus a small loss in accuracy. There are however a number of issues with their presentation which we clarify in our work. First, they overlook the fact that as dimensions increase the theoretical distribution becomes ever less true due to the inability of a bounded data space to contain the query hypersphere. They present graphs showing a perfect distribution over a search space which we have been unable to reproduce. Most importantly, the conceptual basis of their optimisation depends on the angle $\angle qp_is_j$, where q is the query, p_i is the centre of a ball partition, and s_j is a potential solution to the query. This implies that the radius of the search space around the query object is larger than the radius of the data which is being pivoted by p_i, which could not occur in a high-dimensional space.

We have been able to synthesise large numbers of uniformly distributed values within very low-volume hyperspheres thanks to a technique shown by Voelker et al. [13]. This has been extremely valuable as, in high dimensions, it is effectively impossible to otherwise find uniformly distributed points within a given hypersphere.

Finally, there are many papers that give results based on relaxing the strict condition of triangle inequality in ball partitioning, increasing the efficiency of the mechanisms at cost of giving giving approximate results. We believe that our work here goes a long way to explain the effectiveness of such mechanisms in high-dimensional spaces.

3 Unbounded Euclidean spaces

The orthogonality of vectors in high dimensional spaces, once quantified, can give useful insights with respect to single pivot exclusion, in particular towards assessing the probability of an object lying within the intersection of two hyperspheres. Traditional metric search techniques allow only that this is zero when the sum of the radii is greater than the distance between the centres; we are interested in quantifying the probability of an object from the finite space lying within an intersection of the infinite space. However given a high probability of restricted angles in high dimensional spaces, many overlapping hyperspheres will have a very low probability of the geometric intersection containing any elements of a given finite space.

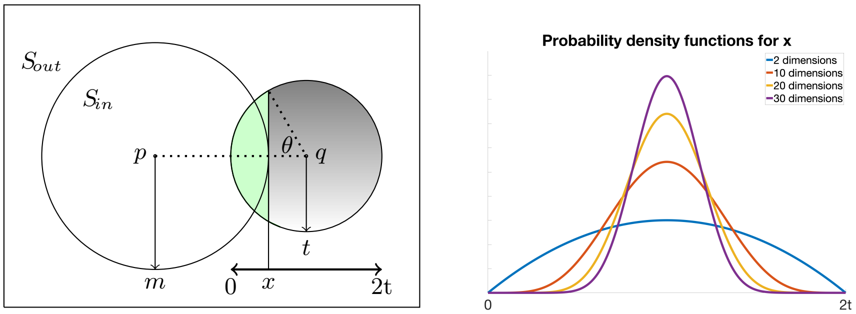

Fig. 1. On the left hand side, the figure shows how the intersection of the hyperspheres is contained in the segment of the hypersphere around q defined by θ. The graph on the right hand side shows the probability of a randomly selected point from within the hypersphere also being in this segment, as x in the left hand figure varies between 0 and $2t$. This therefore gives an upper bound on the probability of such an object in the intersection.

The left hand side of Figure 1 shows how a restriction of angles is useful in metric search. p and q represent hypersphere centres, where a finite metric space S has been divided during pre-processing into S_{out} and S_{in} according to a distance m from a pivot p. q represents a query, to which solutions are being sought within the threshold distance t. Since the hyperspheres intersect, according to the distance $d(p, q)$ calculated at query time, S_{in} cannot be excluded using the metric properties alone.

With respect to the hypersphere centred around q, consider the segment defined by the angle θ. If, for all elements $s_i \in S$ such that $d(q, s_i) \leq t$, the angle $\angle pqs_i$ is greater than θ, then the finite intersection is empty and the set S_{in} can be excluded from the search. If there is a high probability of vectors pq and qs_i being close to orthogonal, there will be a correspondingly high probability of the intersection being empty.

In a general Euclidean space of course this can never be guaranteed; however as we will show, as the dimension of the space increases, the probability of an individual point from a uniform distribution being within the intersection may become very small. The right hand side of Figure 1 gives probability density functions (PDFs), in various dimensions of Euclidean space, for the displacement x for a randomly selected point within the solution space.

It can be seen that, for higher dimensions, the probability of a point lying with the intersection is very low. We will proceed to give a quantification of an upper bound which is easily calculated. In the remainder of this section, we derive a PDF and quantify the examples shown in the figure.

3.1 Volume of a hypersphere

The volume of a hypersphere of radius r can be expressed in terms of the volume of the unit hypersphere (i.e. $r = 1$) as

$$V_n(r) = v_n r^n \tag{1}$$

where v_n is the volume of the unit hypersphere. Equation (1) is well known in a more general context, and straightforward to demonstrate[1].

The intersection of a hyperplane in $\mathbb{R}^n$ with a n-ball is an $(n-1)$-ball. Considering a unit $(n-1)$-ball b_{n-1} centred on the origin, the volume of the unit n-ball can be written as an integral of volumes of $(n-1)$-balls by considering hyperplanes orthogonal to the X_1-axis:

$$v_n = \int_b dx_1 \, \ldots \, dx_n = \int_{-1}^{1} \left(\int_{b_{n-1} \cap \{X_1 = z\}} dx_2 \ldots dx_n \right) dz \tag{2}$$

As depicted in the left-hand side of Figure 2, the intersection $b_n \cap \{X_1 = x\}$ is an (n-1)-ball of radius $r = \sqrt{1-x^2}$, thus its volume is $V_{n-1}(\sqrt{1-x^2})$ and Equation (2) can be rewritten as

$$v_n = \int_{-1}^{1} V_{n-1}(\sqrt{1-x^2}) dx$$

which then, according to Equation 1 gives

$$v_n = v_{n-1} \int_{-1}^{1} \left(\sqrt{1-x^2} \right)^{n-1} \, dx$$

as v_{n-1} may be removed from the integral as it is a constant.

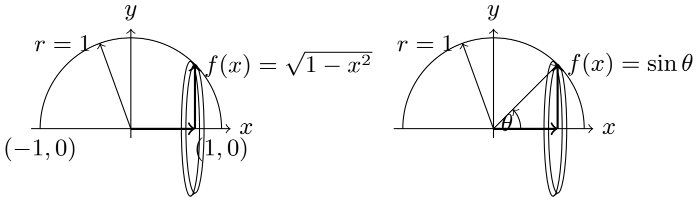

Fig. 2. Volume of a (3D) unit sphere: $\int_{-1}^{1} \pi \left(\sqrt{1-z^2}\right)^2 dz = \int_0^{\pi} \pi \sin^2 \theta \; d\theta$

[1] $V_n(R) = \int_{B_n(R)} 1 \, dx_1 \, \ldots \, dx_n = \int_{B_n(1)} R^n \, dy_1 \, \ldots \, dy_n = R^n V_n(1)$, where we integrate by substitution with $x_i = Ry_i$ for all i

Finally, by considering, as shown on the right hand side of Figure 2, that x can be written as $\cos\theta$ and then $f(x) = \sin\theta$, integrating by substitution we have

$$\int_{-1}^{1} \left(\sqrt{1-x^2}\right)^n dx = \int_{\pi}^{0} \left(\sqrt{(1-\cos^2\theta)}\right)^{n-1} (-\sin\theta)d\theta = \int_{0}^{\pi} (sin\theta)^n d\theta$$

So finally putting all the pieces together we have an expression for the volume of a hypersphere of radius r in n dimensions:

$$V_n(r) = r^n k \int_0^{\pi} sin^n x dx \tag{3}$$

for a constant k.

3.2 Derivation of the PDF

To construct a PDF, we note that Equation 3 derives from a Riemann integral of infinitesimal hyperspheres, in $n-1$ dimensions, each orthogonal to a diameter through the centre of the n-dimensional hypersphere. Considering the left-hand side of Figure 1, the integration may notionally be performed along the axis pq within any $(n-1)$-dimensional hyperplane containing p and q. Then the angle θ in the figure corresponds to the integral variable x in Equation 3. Thus, the volume in the green-shaded area of the figure is given by the definite integral $t^n k \int_0^{\theta} sin^n x dx$. Within a uniformly populated space, the PDF of a point being within the defined segment, with respect to the angle θ, is therefore directly proportional to $h(x) = sin^n x, x \in [0, \pi]$.

The PDFs shown in the right hand side of Figure 1 are produced by applying this function to $\cos\frac{x}{t}$, where x is the distance from q along the line pq, in order to convert the angular dependence to a distance along the pq axis. The outcome is then divided by the volume of the hypersphere around q to normalise the area under the curves.

Quantifying this PDF is non-trivial. However, for high values of n, the function $\sin^n x$ becomes almost indistinguishable from a related Gaussian, and in turn the related PDF becomes almost indistinguishable from that of a normal distribution, and thus readily available. For large n, for example $n > 15$, the PDF function is almost indistinguishable from that of a normal distribution with $\mu = \frac{\pi}{2}$ and $\sigma = \frac{1}{\sqrt{n}}$. This observation has also been made by [3], and is discussed in [1].

3.3 Examples of overlap in unbounded Euclidean spaces

Table 1 gives probability calculations, for Euclidean spaces of various dimensions, for the situation shown in Figure 1, where $d(p,q) = m + \frac{t}{2}$. These figures correspond with the probability density functions shown in the Figure. Two points are notable: first, how small the probabilities become as dimensions increase, even with this significant amount of overlap; secondly, how the normal distribution estimate gives an increasingly small error as the dimension increases.

Table 1. Probability of Inclusion

Dimension	Probability	Normal estimate
2	0.195	0.228
10	0.0405	0.0479
20	0.0074	0.0093
30	0.0015	0.0020

Table 2. Proportion of queries within unit cube

Metric	*Inside cube*	*Outside Cube*
Euc10	87.57%	12.43%
Euc20	51.15%	48.85%
Euc30	35.88%	64.12%

4 Experiments in generated Euclidean spaces

In the following experiments we use a number of different generated Euclidian spaces with individual coordinates drawn from a Gaussian distribution. Data points in these spaces have 10, 20 and 30 coordinates and are referred to as EUC10, EUC20 and EUC30.

In the following experiments we examine the mean and variance of angles *abc* within various spaces. In the first experiment *a*, *b* and *c* are sampled uniformly from within the space. The results of this experiment are shown in the brown (left hand) distributions in Figure 3. In all cases the average angle is close to 60° with standard deviations of 16.5°, 11.25°, 9.11° for Euc10,20 and 30 respectively. As can be observed from the figure the standard deviation drops as the dimensionality of the data-set increases.

In the second experiment *a* and *b* are sampled uniformly from the space but the third point *c* is constrained to be both within a threshold of the query point *b* and within the unit cube. For each experiment the radius of the hypersphere is calibrated to return one-millionth of the data-set. For EUC10, EUC20 and EUC30 these are: 0.229, 0.602 and, 0.727 respectively.[2] The results of this experiment are shown in the blue (right hand) distributions in Figure 3. As can be seen the angle is close to 90° and like the earlier experiment the standard distribution of angles reduces with increasing dimensionality of the data-set.

4.1 Query regions lying outwith the unit cube

In the above experiment we constrained the third points *c* to be within the sampled space. To determine the proportion of the query ball that lies outwith the unit cube we performed the following experiment on each of the Euclidean spaces. We randomly sampled one thousand points from within the space. For each point we uniformly sampled a further thousand points from within the

[2] If the radius of the hypersphere is constrained to be within the unit cube rather than at the defined radii the results are identical.

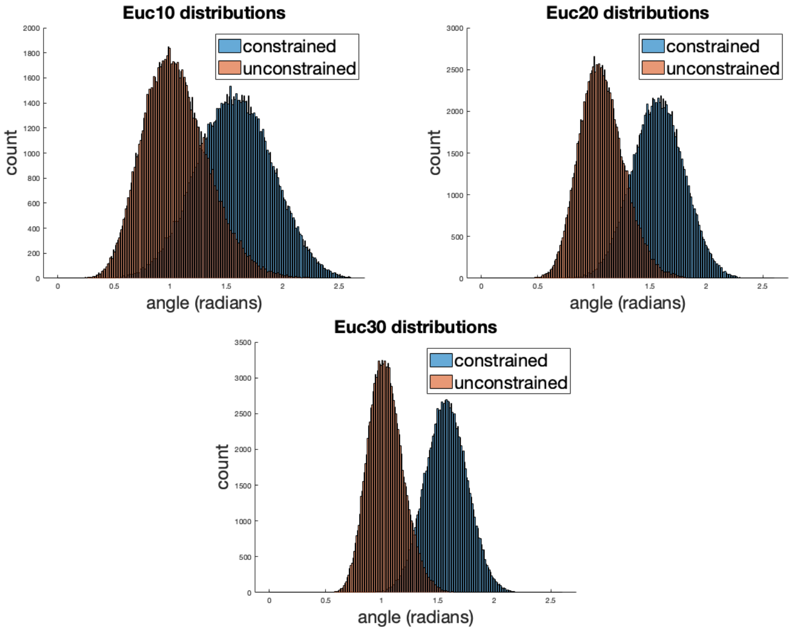

Fig. 3. Distance distributions for constrained and unconstrained triples

hypercube of radius set to be the standard thresholds (described above) and measured if the point is within the unit cube or not. The results of running the experiment is shown in Table 2. As can be seen, the proportion of the hypersphere outwith the unit cube increases dramatically as the dimensionality of the dataset increases.

4.2 Prediction of the angle distribution

To understand the effect of the relationship between where queries are in the space and the resultant angles we conducted the following experiment. We sweep a hyper sphere up the diagonal of the unit hyper cube (in some dimension) from the origin to the opposite corner (1,..,1) in intervals of 0.01. The radius of the hypersphere is set using the standard thresholds used above. We examine the mean and variance of angles $\angle abc$ as follows. a is a fixed viewpoint which is always the centre of the unit cube[3], b is set to be a point along the diagonal of the cube $(0, 0, ..., 0)$ to $(1, 1, ..., 1)$, and for each instance of b, c is sampled from within a fixed hypersphere centred around b as before. As before, we discard any points that are not within the (0,..,0)-(1,..,1) hypercube – i.e. those points that cannot be legal solutions to the query. In each case 1 million points from within the hyper sphere are chosen randomly and those lying outwith the unit cube are discarded. During this process we also measure *outlierness* using a Local

[3] We separately established that the viewpoint does not affect the measured angles.

Intrinsic Dimensionality (LIDIM) maximum likelihood estimator due to Levina & Bickel [9]. To determine *outlierness*, we apply this formula using distances to a set of reference points rather than calculating LIDIM using a set of neighbouring points.

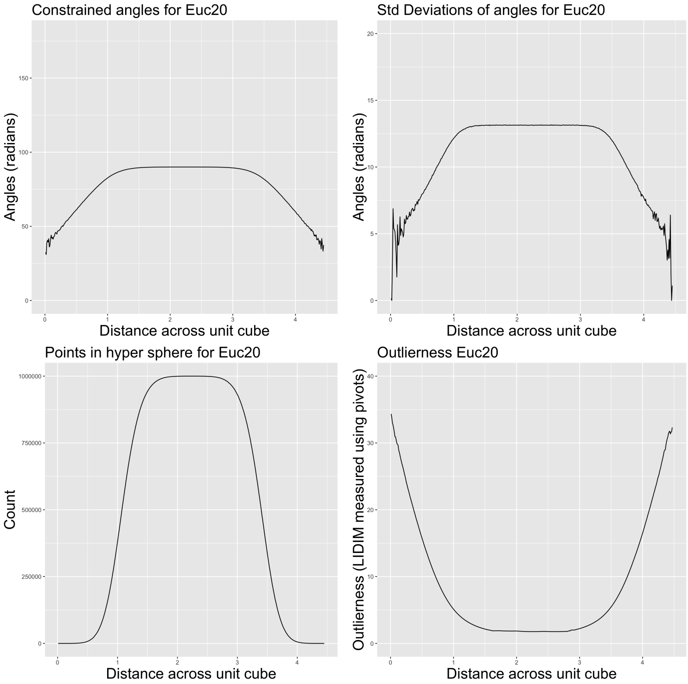

Fig. 4. Constrained Angles in the EUC20 Dataset

The results of these experiments are shown in Figure 4. These four plots demonstrate a number of different interesting facets of the query solutions. Firstly, it can can be seen that angles $\angle abc$ are far from uniform. At the edge of the unit cube they rise and fall rapidly as the hyper sphere approaches the vertices of the unit cube. When the hypersphere approaches the centre of the cube the angles tend towards 90°. Secondly the distribution of angles are not constant. As the sphere approaches the centre of the cube they rise to a relatively constant variation of approximately 13.1°. Thirdly, the number of points inside the cube vary greatly. Close to the vertices of the cube the number of legal points tend towards zero whereas in the centre all of the 1 million sampled points are within the cube. Lastly, a good approximation to how much of an outlier a point is can be made by using the LIDIM formula with a fixed set of reference points as described above.

5 Experiments in other spaces

In this section we test the value of the angle analysis by examining pivot exclusion, enhanced by angle analysis, in a LAESA-like framework. We should stress that these experiments are to test the proof of the concept only; we believe that more sophisticated mechanisms can take advantage of the angle information to much better effect.

Table 3. Data Sets

Name	Dimensions	Derivation	Preparation	Metric
MfAlex	4096	MirFlickr	fc6 layer AlexNet, no RELU	Euclidean
DeCaf	4096	Profiset	fc7 layer AlexNet, post-RELU	Euclidean
AnnSift	128	MirFlickr	ℓ_2 normalised	Euclidean
MfGist	480	MirFlickr	ℓ_1 normalised	Jensen-Shannon

We used four different high-dimensional data sets with different properties, as summarised in Table 3. *MfAlex* is derived from the application of the AlexNet [8] convolutional neural network on the MirFlickr image collection[4]. The data used is extracted from the first fully connected layer (fc6). *DeCaf* descriptors [7] are extracted from the *Profiset* image collection[5] using AlexNet, from which the fc7 *post-Relu* layer is extracted. *AnnSift* descriptors [10] are taken from the *ANN_SIFT1M* dataset[6]. Although queried with the ℓ_2 distance, these vectors are ℓ_2 normalised and thus this metric acts as a proxy for Cosine distance. *MfGist* is derived using GIST [11] image descriptors over the MirFlickr 1M image collection. These descriptors are queried using the Jensen-Shannon distance, which has been shown to be the best metric for near-duplicate detection [6].

Of the four data sets, only the first therefore represents a true Euclidean space where each dimension contains a range of positive and negative values. We have deliberately chosen this range of data sets to examine whether the angular properties which are clear in unbounded Euclidean spaces follow in more general metric spaces. The four spaces all contain one million objects, and in each case a ground truth is known for one thousand queries, each of which has 100 known nearest neighbours.

The queries are divided into two equal sets, the first of which is used to perform analysis over the space, and the second of which is used to test a search mechanism using that analysis.

5.1 Correlation of outlierness and angle

Our hypothesis is that for any high-dimensional space, the distribution of angles $\angle p_i q s_j$, where p_i is selected from a set of reference points, q is a query and

[4] https://press.liacs.nl/mirflickr

[5] http://disa.fi.muni.cz/profiset/

[6] http://corpus-texmex.irisa.fr/

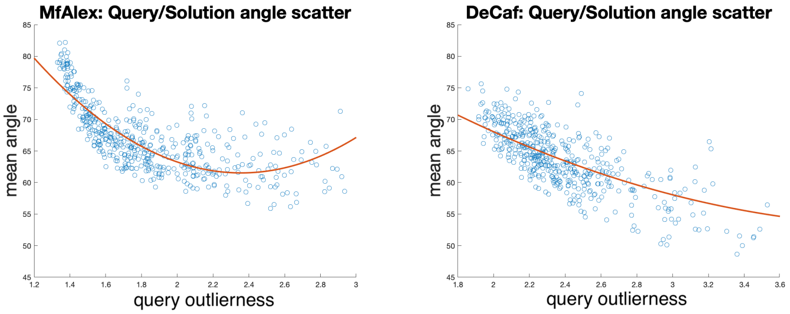

Fig. 5. Correlation between query outlierness and mean angle $\angle p_i q s_j$, where s_j is a solution to query q. The lines show the best-fit quadratics, used in the experiments in Section 5.

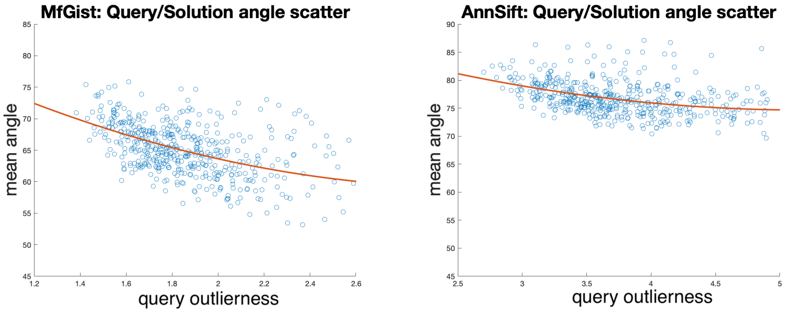

Fig. 6. Correlation between query outlierness and mean angle $\angle p_i q s_j$ for the non-Euclidean spaces.

s_j is selected from solutions to that query, will be constrained in comparison with randomly sampled angles from the space, and will be correlated with the *outlierness* of the query.

A randomly selected set of 256 objects was selected from the dataset to act as reference objects. For each query q, for each p_i in the reference set, and for each s_j in the known solution set, the angle $\angle p_i q s_j$ was measured and the mean and variance recorded.

For each q, an approximation to the outlierness was calculated based on the distances $d(p_i, q)$ for each reference object using the maximum likelihood estimator as described above [9]. The scatter plots in Figure 6 show a clear relationship between query outlierness and the mean angle. It is clear in all cases that the majority of angles are greater 60° and are thus distinct from the angles within a randomly selected triplet. The angles depicted on the scatters are averaged over the query's 100 nearest neighbours, and in all cases the standard deviation is quite low - almost always less than 10°. The implication, in terms of the analysis shown in Section 3, is that exclusion may safely occur in many situation where its safety cannot be guaranteed by triangle inequality alone. We quantify this in the next section.

5.2 Use in querying

In this section, a simple search mechanism is applied in order to give experimental validation of the principles outlined[7].

The search mechanism used is a variant of LAESA. From the data set S, a randomly selected subset P of 256 reference objects[8] is removed. The preprocessing phase comprises the calculation of a distance table between each reference object $p_i \in P$ and each remaining member s_j of S. At query time, the distance between the query and p_i is calculated. The possibility of exclusion of each object s_j is determined by scanning the appropriate row of the pre-calculated distance table. In the normal LAESA algorithm, exclusion may occur if and only if $|d(p_i, s_j) - d(p_i, q)| > t$, where t is the threshold distance for that query. In that case, it is impossible for the hypersphere of radius t centred on q to contain s_j, and the distance $d(q, s_j)$ does not require to be calculated.

The pure LAESA mechanism is adapted to perform exclusion even in some cases where $|d(p_i, s_j) - d(p_i, q)| \leq t$, as depicted in Figure 1. For each query q, the set of distances $d(q, p_i)$ is calculated as usual. These distances are first used to measure an estimate of outlierness, as described in Section 5.1, and thus to determine an estimate γ of the mean angle $\angle pqs_i$ in cases where $d(q, s_i)$ is small. In our experiments, a fixed amount of variance τ is allowed, with the intent that, for all solutions, the angle $\angle pqs_i$ is highly likely to lie within the bounds $\gamma \pm \tau$. Now, for all values s_j from the finite set, and for each p_i, the angle θ is calculated[9] from the values $d(p_i, s_j), d(p_i, q)$ and t. If the angle θ lies outside the range $\gamma \pm \tau$, then s_j is excluded without performing the calculation $d(q, s_j)$.

In the experiments over all the data sets we report outcomes using a range of fixed tolerances between 0.3 and 0.65 radians. Finally, each experiment is repeated with a tolerance of $\frac{\pi}{2}$ radians, which effectively makes exclusion impossible other than when allowed by the pure LAESA mechanism.

Results for the four data sets are shown in Figure 7. The left hand graph shows the cost per query for the different tolerances; as expected, a smaller tolerance, resulting in a larger cutoff angle θ, gives a lower cost. The right hand graph shows recall for the same tolerances. As noted, this mechanism always gives perfect precision.

The results fully justify our derived model. For all data sets, a tolerance value of around 0.6 radians gives almost perfect recall. This value implies that all query solutions, in these spaces, do indeed lie within an arc of 1.2 radians in

[7] Source code is available from https://github.com/aldearle/SISAP2020_angles or from the authors.

[8] the same number is used across all sets to allow fair comparison, even although for example the cost of performing an exclusion for *AnnSift* may be greater than the cost of measuring the distance directly.

[9] unless $|d(p_i, s_j) - d(p_i, q)| > t$, when exclusion can occur in any case

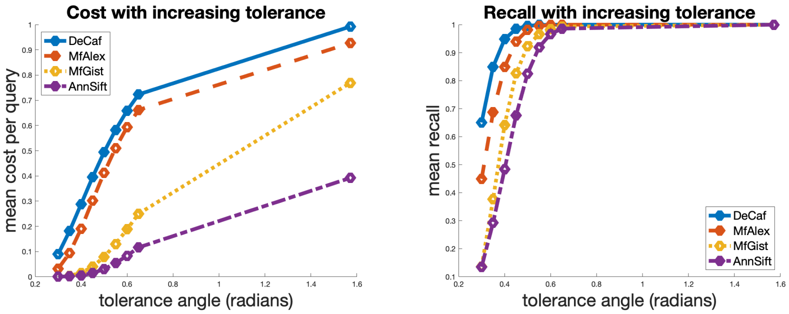

Fig. 7. Test results with LAESA with increasing tolerance. The left figure shows cost, as the number of distance calculations performed divided by the size of the data. The right hand figure shows recall. The values at the right hand side of each graph equate to those for an unmodified LAESA mechanism.

the 2D projection, implying that over half of the angular space is empty. It is also noteworthy that the data sets with higher costs for the unmodified LAESA give perfect recall with lower tolerance levels. It is reasonable to assume that this is a consequence of a higher inherent dimensionality leading to a tighter clustering of the angles within the 2D projection.

6 Conclusions and future work

We have taken an observation from high-dimensional vector spaces and applied it to general metric spaces by way of a derived approximate search paradigm. We have shown an underlying mathematical model which explains a related effect in unbounded, uniform Euclidean spaces, and demonstrated it experimentally. We have shown that, unfortunately, the effect does not hold perfectly in bounded high dimensional search spaces. This is because the radius required to capture query solutions in a finite space far exceeds the boundaries of the space.

We have demonstrated nonetheless an interesting restriction in the distribution of angles in metric spaces, and in particular that the angles from a reference point, via a query, to a query solution are significantly different from angles randomly sampled from the space.

We have outlined how this may be used to conduct a probabilistic search, and a trade-off is shown between query efficiency and accuracy in spaces which are otherwise intractable for exact search. We believe that this topic has much further promise; our present analysis of the spaces is based on a relatively crude measure of query outlierness, and we believe a more sophisticated analysis of the space may result in a finer-grained understanding of the angle distribution, as well as further query mechanisms based on it. In particular, we have not yet examined the effect of the restricted angles on hyperplane exclusion mechanisms, nor in conjunction with the four-point property of supermetric spaces.

References

1. Large powers of sine appear gaussian — why? (Accessed: 2020-06-22), https://math.stackexchange.com/questions/2293330
2. Blum, A., Hopcroft, J., Kannan, R.: High-Dimensional Space, p. 4–28. Cambridge University Press (2020). https://doi.org/10.1017/9781108755528.002
3. Cai, T., Fan, J., Jiang, T.: Distributions of angles in random packing on spheres. Journal of Machine Learning Research **14**(21), 1837–1864 (2013)
4. Chávez, E., Navarro, G.: Metric databases. In: Encyclopedia of Database Technologies and Applications. Idea Group (2005)
5. Connor, R., Cardillo, F.A., Vadicamo, L., Rabitti, F.: Hilbert Exclusion: Improved metric search through finite isometric embeddings. ACM Transactions on Information Systems **35**(3), 17:1–17:27 (Dec 2016). https://doi.org/10.1145/3001583
6. Connor, R.C.H., Cardillo, F.A.: Quantifying the specificity of near-duplicate image classification functions. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP, Rome, Italy, February 27-29, 2016. pp. 647–654 (2016)
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML 2014, Beijing, China. pp. 647–655 (2014)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
9. Levina, E., Bickel, P.J.: Maximum likelihood estimation of intrinsic dimension. In: Advances in Neural Inf. Processing Systems. pp. 777–784. MIT Press (2005)
10. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, Kerkyra, Corfu, Greece, September 20-25, 1999. pp. 1150–1157 (1999)
11. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Comp. Vision **42**, 145–175 (2001)
12. Pramanik, S.K., Li, J., Ruan, J., Bhattacharjee, S.K.: Efficient search scheme for very large image databases. In: Internet Imaging. vol. 3964
13. Voelker, A.R., Gosmann, J., Stewart, T.C.: Efficiently sampling vectors and coordinates from the n-sphere and n-ball. Tech. rep., Centre for Theoretical Neuroscience, Waterloo, ON (01 2017)
14. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity search: the metric space approach, vol. 32. Springer Science & Business Media (2006)