Indexability, concentration, and VC theory

Vladimir Pestov
Department of Mathematics and Statistics, University of Ottawa
585 King Edward Avenue, Ottawa, Ontario, Canada K1N6N5
vpest283@uottawa.ca

ABSTRACT

Degrading performance of indexing schemes for exact similarity search in high dimensions has long since been linked to histograms of distributions of distances and other 1-Lipschitz functions getting concentrated. We discuss this observation in the framework of the phenomenon of concentration of measure on the structures of high dimension and the Vapnik-Chervonenkis theory of statistical learning.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information search and retrieval—*Search process*; F.2.2 [**Theory of Computation**]: Analysis of algorithms and problem complexity—*Geometrical problems and computations*

1. INTRODUCTION

At an intuitive level, at least for a limited class of indexing schemes the geometric and probabilistic origin of the curse of dimensionality is quite transparent. Let $W = (\Omega, \rho, X)$ denote a similarity workload, where ρ is a metric on a domain Ω and X is a finite subset of Ω (dataset). Let us say we are interested in indexing into W for deterministic, exact range queries. A traditional "distance-based" indexing scheme, stripped down to the bone, consists of a family of real-valued 1-Lipschitz functions f_i, $i \in I$ on Ω, either fully or partially defined:

$$|f_i(x) - f_i(y)| \leq \rho(x, y). \tag{1}$$

(For example, a pivot-based indexing scheme will be using distance functions $\rho(p, -)$ to the pivots p.) Given a query (q, ε), where $q \in \Omega$ and $\varepsilon > 0$, the algorithm chooses recursively a sequence of indices i_n, where i_{n+1} is determined by the previous values $f_{i_k}(q)$, $k \leq n$. The functions f_i serve to discard those datapoints which cannot possibly answer the query. Namely, if $|f_i(q) - f_i(x)| \geq \varepsilon$, then, by the 1-Lipschitz property of f_i, one has $\rho(q, x) \geq \varepsilon$, and so the point x is irrelevant and need not be considered (Figure 1).

SISAP 2010, September 18-19, 2010, Istanbul, Turkey.

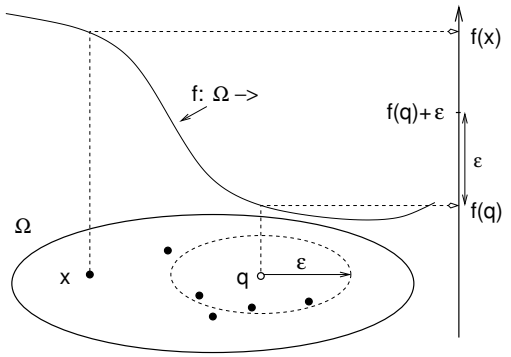

Figure 1: The datapoint x can be discarded.

After the calculation terminates, the algorithm returns all points which cannot be discarded, and checks each one of them against the condition $\rho(x, q) < \varepsilon$.

Next come two standard observations about high-dimensional data. The first one, known as the "empty space paradox," asserts that the average distance $\mathbb{E}(\varepsilon_{NN})$ to the nearest neighbour approaches the average distance $\mathbb{E}(\rho)$ between two datapoints as the dimension d goes to infinity, provided the number of datapoints, n, grows subexponentially in d. Cf. Figure 2, where we illustrate the point with a constant number of points ($n = 10^3$ and $n = 10^5$), and the distances are normalized so that the *characteristic size* of the gaussian space $(\mathbb{R}^n, \gamma^n)$,

$$\text{CharSize}\,(X) = \mathbb{E}_{\mu\otimes\mu}(\rho(x, y)), \tag{2}$$

is one.

The second observation is that the histograms of values of common 1-Lipschitz functions on high-dimensional data are concentrated near their mean (or median) values. This effect is already pronounced in moderate dimensions such as $d = 14$ in Figure 3. Here the function is a distance to a randomly chosen pivot p, and assuming the query point q is at a distance ≈ 1 from p, only the points outside of the region marked by vertical bars can be discarded.

The two properties combined imply that as $d \to \infty$, fewer and fewer datapoints can be discarded for an average range query, and the performance of an indexing scheme degrades rapidly. This mechanism has been discussed repeatedly, e.g. in [9], pp. 35–37, [32], [42], p. 487, to mention just a few.

To make this idea yield rigorous lower performance bounds, one needs to guarantee first that *every* histogram of dis-

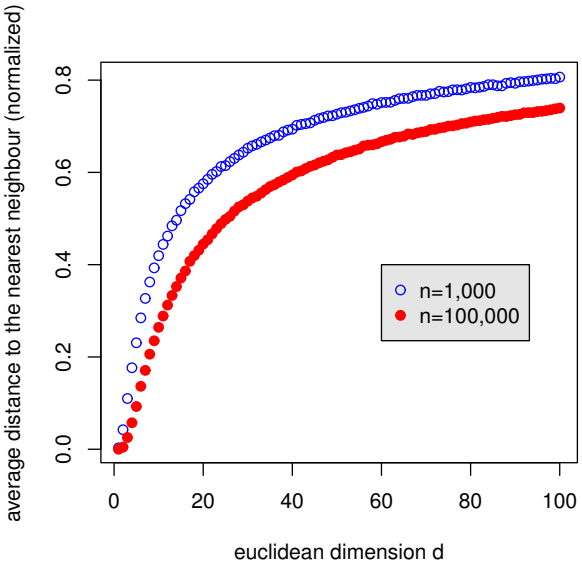

Figure 2: The normalized average distance to the nearest neighbour in a dataset of n points randomly drawn from a gaussian distribution in $\mathbb{R}^d$.

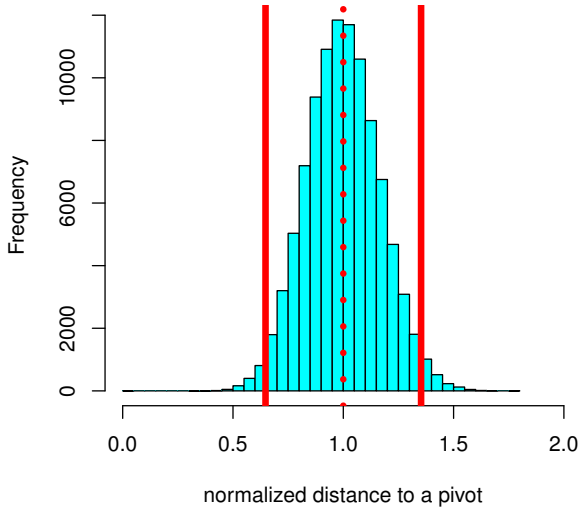

Figure 3: Histogram of distances to a randomly chosen pivot in a dataset X of $n = 10^5$ points drawn from a gaussian distribution in $\mathbb{R}^{14}$. The vertical lines mark the mean normalized distance $1 \pm \varepsilon_{NN}$.

tances of 1-Lipschitz functions used to build an indexing scheme is highly concentrated. In other words, if $\mathscr{F}$ denotes a class of 1-Lipschitz functions from which we choose the f_i, we want a small uniform upper bound on the variances of $f \in \mathscr{F}$. Results of this type are indeed well-known for a variety of geometric objects and are referred jointly as the *phenomenon of concentration of measure*.

Next problem is, how to link the concentration of functions f with regard to the underlying distribution μ on the domain Ω to concentration with regard to the *empirical measure* on the dataset X (this was essentially a criticism of [37] made in [43])? Here one needs the machinery of *statistical learning theory* of Vapnik and Chervonenkis, which can guarantee such results provided the class $\mathscr{F}$ has low capacity (e.g., a finite VC dimension). This way, one obtains $\Omega(n/d \lg n)$ lower bounds for the pivot table expected average performance, as well as superpolynomial in d lower bounds for metric trees.

Approximate NN queries seem to be in some sense free from the curse of dimensionality. In fact, the concentration of measure becomes a positive force here, and we will try to explain why, using the example of random projections method (the Johnson–Lindenstrauss lemma), as well as the approach of Kushilevitz, Ostrovsky and Rabani based on VC theory.

Getting back to exact search, the *Curse of Dimensionality Conjecture* calls for a much more general statement than the lower bounds discussed above, which would apply across the entire range of all possible indexing schemes. The conjecture is still open even for the Hamming cube $\{0,1\}^n$, and we discuss it briefly.

We conclude the article with remarks on the notion of intrinsic dimensionality of data, as well as on a spatial approximation algorithm based on Delaunay graphs.

2. CONCENTRATION

2.1 The concentration of measure phenomenon

Informally, the phenomenon can be stated as follows:

> On a typical "high-dimensional" structure, the variance of every 1-Lipschitz function is small.

This formulation is not quite exact (it is a slightly stronger statement), and in fact a usual way to deal with concentration is through a different dispersion parameter. Let a metric space (Ω, ρ) carry a probability measure μ. One defines the *concentration function* α_Ω of the metric space with measure Ω:

$$\alpha_\Omega(\varepsilon) = \begin{cases} \frac{1}{2}, & \text{if } \varepsilon = 0, \\ 1 - \inf\left\{\mu\left(A_\varepsilon\right) : A \subseteq \Omega, \quad \mu_\sharp(A) \geq \frac{1}{2}\right\}, & \text{if } \varepsilon > 0. \end{cases}$$

The value $\alpha_\Omega(\varepsilon)$ gives a uniform upper bound on the measure of the complement to the ε-neighbourhood A_ε of every subset A of measure $\geq 1/2$, cf. Fig. 4. On a typical high-dimensional geometric object Ω the function $\alpha(\varepsilon)$ drops off steeply near zero. For regular geometric objects such as Hamming cubes, Euclidean unit spheres and so on, there are usually gaussian upper bounds of the form $\alpha(\varepsilon) \leq \exp(-O(\varepsilon^2 d))$, where d is the dimension parameter. For example, the Hamming cube $\{0,1\}^d$ with the normalized Hamming metric and uniform measure satisfies a Chernoff bound $\alpha(\varepsilon) \leq \exp(-2\varepsilon^2 d)$ (obtained by combining Harper's isoperimetric inequality, see e.g. [16], with the classical Chernoff bound, cf. [49], 2.2.1). See Figure 5.

It follows easily that for every real-valued 1-Lipschitz function f on Ω and for each $\varepsilon > 0$ one has

$$\mu\{x \in \Omega\colon\ |f(x) - M_f| > \varepsilon\} \leq 2\alpha_\Omega(\varepsilon), \qquad (3)$$

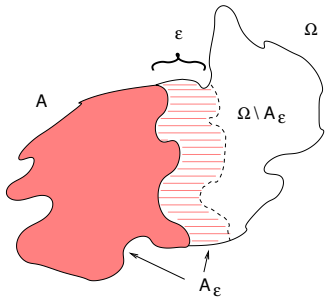

Figure 4: To the concept of the concentration function $\alpha_\Omega(\varepsilon)$.

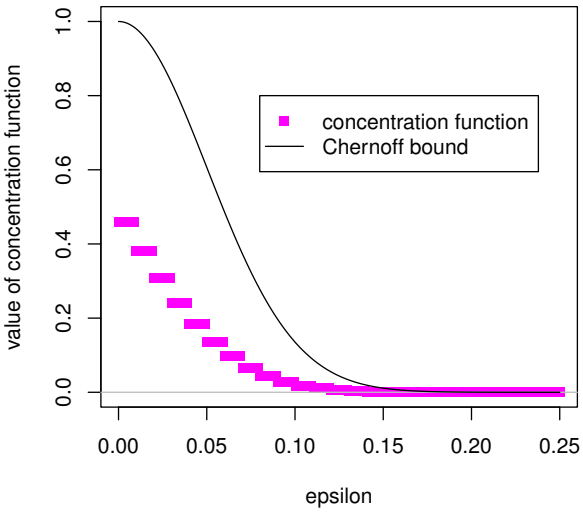

Figure 5: Concentration function of the Hamming cube of dimension $d = 100$ *vs* Chernoff bound.

where M_f is the median value of f. Under mild conditions (e.g. if Ω has finite diameter), this implies a uniform bound on variances. But usually this is Eq. (3) which gives the most useful form of concentration.

The phenomenon admits the following illustration. Draw $1,000$ points randomly from a high-dimensional sphere $\mathbb{S}^d$, choose a random orthogonal projection onto a two-dimensional subspace, and project both the sphere and the chosen points on this subspace. The points will concentrate near the centre, the more so the higher the dimension d, as seen in Figure 6 for $d = 20$ and $d = 1000$... except that the figure shows the results of the same experiment performed on the unit cubes of the same dimension d rather than spheres. However, for $d = 1000$ the only difference one can spot from a unit sphere, is the scale of the projection: the diameter of a unit d-cube is $O(\sqrt{d})$.

This illustrates an interesting feature of geometry of high dimensions: many high-dimensional objects look essentially the same to a low-dimensional observer. For instance, this is true of *all* convex bodies, as recently proved by Klartag [23]. For this reason, for an asymptotic study of lower bounds of

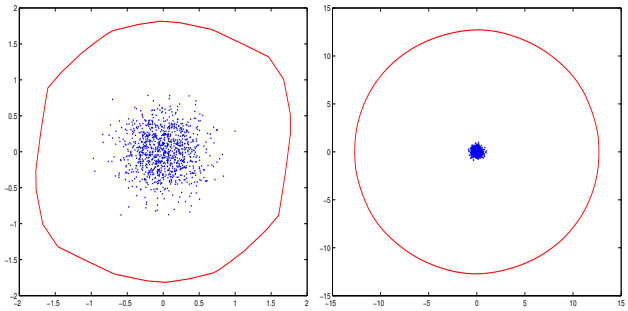

Figure 6: Orthogonal projection of a unit Euclidean d-cube and of $1,000$ random points inside the cube on a random 2-subspace, $d = 20$ (left), $d = 1000$ (right).

indexing schemes when $d \to \infty$ the choice of a particular family of domains (Euclidean spheres, balls, cubes, Hamming cubes) does not matter that much.

Among the books treating the concentration phenomenon, [30] is the most reader-friendly, [26] most comprehensive, and [18] contains a wealth of ideas. See also a survey [29].

2.2 Empty space paradox

Let us agree on the following assumptions on the similarity workload:

- the metric domain (Ω, ρ) is equipped with a probability measure μ, and datapoints are drawn from Ω in an i.i.d. fashion following the distribution μ (this is the model of [11]);
- the distance ρ on the domain is normalized so that the characteristic size of Ω is constant:
$$\mathrm{CharSize}(\Omega) = \mathbb{E}_{\mu\otimes\mu}(\rho) = O(1);$$
- Ω has concentration dimension d in the sense that the concentration function of Ω admits a gaussian upper bound
$$\alpha_\Omega(\varepsilon) = \exp(-\Omega(\varepsilon^2 d));$$
- the number n of datapoints grows faster than any polynomial function in d, but slower than any exponential function in d:
$$n = d^{\omega(1)}, \quad d = \omega(\log n). \tag{4}$$
(This is a standard assumption in the asymptotic analysis of algorithms, cf. [20]. An example of such a rate of growth is $n = 2^{\sqrt{d}}$.)

Denote ε_{NN} the nearest neighbour distance function on Ω, given by $\varepsilon_{NN}(\omega) = \rho(\omega, X)$.

THEOREM 2.1. *Under the above assumptions, with high confidence one has for every ε*

$$\mu\{\omega\colon\ |\varepsilon_{NN}(\omega) - \mathrm{CharSize}(\Omega)| > \varepsilon\} < \exp(-O(\varepsilon^2 d)).$$

The result applies to the Hamming cube, the Euclidean cube, the Euclidean space with gaussian measure, the Euclidean ball, etc. For a proof, see [40], Lemma 1.

COROLLARY 2.2. *Under the same assumptions, for every $\varepsilon > 0$ and sufficiently large d with high confidence the pairwise distances between datapoints of X are all in the range* $\mathrm{CharSize}\,(\Omega) \pm \varepsilon$.

For constant $n = |X|$ and a Euclidean domain the result was proved in [19].

3. VC THEORY

3.1 VC dimension

Let $\mathscr{C}$ denote a collection of subsets of the domain Ω. The VC dimension is a measure of combinatorial complexity of $\mathscr{C}$. A finite set $A \subseteq \Omega$ is *shattered* by $\mathscr{C}$ if every subset $B \subseteq A$ can be "carved out" of A with the help of a suitable element C of $\mathscr{C}$:

$$B = A \cap C.$$

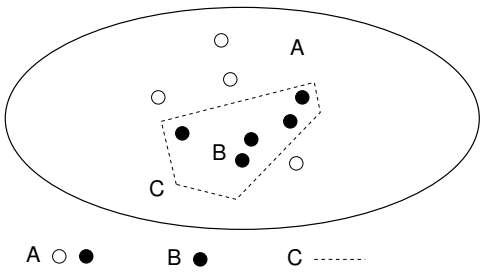

Figure 7: To the concept of a set A shattered by a class $\mathscr{C}$.

The *VC dimension* of $\mathscr{C}$, denoted $\mathrm{VC}(\mathscr{C})$, is the supremum of cardinalities of finite subsets of the domain shattered by $\mathscr{C}$. Here are some classical examples.

Family of sets	VC dimension
Intervals in $\mathbb{R}$	2
Half-spaces in $\mathbb{R}^d$	$d+1$
Euclidean balls of all radii in $\mathbb{R}^d$	$d+1$
Parallelepipeds in $\mathbb{R}^d$	$2d+2$
Family of n sets	$\leq \lg_2 n$
Balls in the Hamming d-cube	$\leq d + \lg_2 d$
Finite unions of intervals in $\mathbb{R}$	∞
Convex polygons in $\mathbb{R}^d$	∞
$\{\Omega \setminus C \colon C \in \mathscr{C}\}$	$\mathrm{VC}(\mathscr{C})$
$\mathscr{C} \cup \mathscr{D}$	$\leq \mathrm{VC}(\mathscr{C}) + \mathrm{VC}(\mathscr{D}) + 1$
k-fold intersections of els of $\mathscr{C}$	$\leq 2k \lg(ek)\mathrm{VC}(\mathscr{C})$

Proofs can be found in [49], Ch. 4. Estimating the VC dimension is often a non-trivial task, and for instance even for the collection of all cubes in $\mathbb{R}^d$ the value of this parameter seems to be unknown.

3.2 Uniform convergence of empirical measures

Now suppose $\mathscr{C}$ consists of *Borel subsets* of Ω, that is, members of the smallest family closed under countable intersections and complements and containing all open balls. (This assumption guarantees that the value $\mu(C)$ is well-defined for every probability measure μ on Ω.) The *empirical measure* of $C \in \mathscr{C}$ with regard to a finite sample $X = \{x_1, \ldots, x_n\}$ is just the normalized counting measure $\mu_n(C) = |\{i \colon x_i \in C\}|/n$. The VC dimension of $\mathscr{C}$ is finite if and only if, with high confidence, the empirical measures of every $C \in \mathscr{C}$ converge uniformly to the true value $\mu(C)$ as the sample size $n \to \infty$, no matter what μ is.

Here is a more exact formulation. The class $\mathscr{C}$ has the property of *uniform convergence of empirical measures*, or is a *uniform Glivenko–Cantelli class,* if there is a function $s(\delta, \varepsilon)$ (*sample complexity* of the class) so that, given a desired precision value $\varepsilon > 0$ and a risk level $\delta > 0$, whenever $n \geq s(\delta, \varepsilon)$, one has

$$\sup_{\mu \in P(\Omega)} P\left\{ \sup_{C \in \mathscr{C}} |\mu(C) - \mu_n(C)| \geq \varepsilon \right\} < \delta.$$

Here $P(\Omega)$ denotes the family of all probability measures on Ω. We quote the following as stated in [49], Theorem 7.8.

THEOREM 3.1. *A concept class $\mathscr{C}$ is uniform Glivenko–Cantelli if and only if $d = \mathrm{VC}(\mathscr{C}) < \infty$, in which case*

$$s(\delta, \varepsilon) \leq \max\left\{ \frac{8d}{\varepsilon} \lg \frac{8e}{\varepsilon}, \frac{4}{\varepsilon} \lg \frac{2}{\delta} \right\}.$$

The proof uses in an essential way the concentration of measure in the Hamming cube $\{0,1\}^n$.

Among a great selection of books treating VC theory, let us mention encyclopaedic sources [49] and [14], a classical monograph [47], and a lighter, but very well-written [2].

4. THE CURSE OF DIMENSIONALITY

4.1 Pivot tables

4.1.1 Reduction and access overhead

Every 1-Lipschitz mapping f from a domain Ω to a domain Υ can be viewed as a *projective reduction* of the exact similarity search problem to the new workload $(\Upsilon, f(X))$. This viewpoint is developed in some detail in [41]. If queries in Υ are easier to process than in Ω, it makes sense to retrieve all datapoints x in the ε-range query of $f(q)$ in Υ and then check them against the condition $\rho_\Omega(q, x) < \varepsilon$. The *access overhead* of the reduction f is defined as

$$\mathrm{acc}_f(q) = |X \cap f^{-1}(B_\varepsilon(f(q)))| - |X \cap B_\varepsilon(q)|.$$

This simple idea on its own can be surprisingly efficient, cf. [45].

4.1.2 Pivot-based reduction to $\ell^\infty(k)$

Every finite collection $f_1, f_2, \ldots, f_k$ of 1-Lipschitz functions on (Ω, ρ) defines a 1-Lipschitz mapping $f = \Delta_{i=1}^k f_i$ from Ω to $\ell^\infty(k)$ via

$$f(x) = (f_1(x), f_2(x), \ldots, f_k(x)).$$

Here $\ell^\infty(k)$ is the vector space $\mathbb{R}^k$ with the norm $\|x\|_\infty = \max_{i=1}^k |x_i|$. If the f_i are distance functions from pivot points $p_i \in \Omega$, the resulting mapping f is of the form

$$f(x) = (d(x, p_1), d(x, p_2), \ldots, d(x, p_k)) \in \ell^\infty(k). \quad (5)$$

In [48], it was proposed to use this reduction in case where the distance computations in Ω are so expensive that even a simple sequential scan of the image $f(X)$ in $\ell^\infty(m)$ is computationally cheaper. This idea was analyzed for more general similarity measures than metrics in [15]. By combining it

with other access methods on the space $\ell^\infty(m)$, further new indexing methods have been developed, see e.g. [8].

A m-NN similarity query is processed in (Ω, d, X) in time

$$k + \ell + (\mathrm{acc}_f(q) + m).$$

Here the first term stands for the calculation of k distances from a query point q to the pivots and ℓ is the processing time of a rectangular query in $\ell^\infty(k)$, while the latter expression lists the number of distance computations in Ω needed to separate false hits from k true positives. There is only one term above that can be minimized through the choice of the pivots. A classical paper on pivot selection is [6].

4.1.3 Lower query time bounds for pivot tables

Our next result (a slightly corrected version of the main theorem in [50]) is valid for the Hamming cube which is a testbed for asymptotic analysis of performance of indexing schemes, but also for the Euclidean space $\mathbb{R}^d$ with the gaussian measure, the cube $[0,1]^d$, and so forth.

THEOREM 4.1. *In addition to the assumptions of Subs. 2.2, suppose also that the VC dimension of the family of all balls in Ω is $O(d)$. Any pivot table with $k = o(n/d\log n)$ pivots will return an expected average number of $\Omega(n)$ datapoints. Consequently, the average total complexity of the performance of any pivot table for the resulting workload is $\Omega\left(n/d\log n\right)$.*

PROOF. Let ε_M denote the median value of the function ε_{NN}, so that for at least half query points q the distance to the NN in X is $\geq \varepsilon_M$. For each pivot p_i, $i = 1, 2, \ldots, k$, denote ρ_i^M the median value of the distance function $\rho(p_i, -)$. Because of concentration, the mesure of the spherical shell

$$S_i = \{q \in \Omega \colon \rho_i^M - \varepsilon_M/2 < \rho(p_i, q) < \rho_i^M + \varepsilon_M/2\},$$

is $1 - \exp(-\Omega(\varepsilon_M^2 d))$, and the measure of the intersection, $S = \cap_i S_i$, of all k shells is

$$1 - o(n/d)\exp(-\Omega(\varepsilon_M^2 d)) = 1 - \exp(-\Omega(\varepsilon_M^2 d)),$$

since n is subexponential in d. Thus, among all k-fold intersections of spherical shells (Figure 8), we have found a giant one, whose μ-measure is nearly one. To assure that it contains an accordingly high proportion of datapoints, consult the table in Subs. 3.1 to notice the family of all k-fold intersections of spherical shells in Ω has VC dimension not exceeding $2k\lg(ek)O(d) = o(n)$. By Theorem 3.1, the empirical measure $\mu_n(S)$ approaches $\mu(S)$ and therefore 1 with high confidence as $d \to \infty$. One concludes: about half query points will belong to S, and for them all datapoints belonging to S, that is, most datapoints, have to be returned. This gives expected average total complexity $\Omega(n)$. □

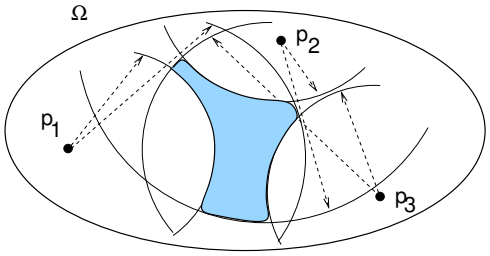

Figure 8: An intersection of spherical shells.

Notice that we allow the pivots p_i to be any points of the domain Ω. If we require that pivots be chosen from the dataset X, then the set S in the above proof is guaranteed to contain $n - k$ datapoints by Theorem 2.1 and Corollary 2.2, and we obtain (without using VC theory):

COROLLARY 4.2. *Under the assumptions of Subs. 2.2, if all pivots p_i belong to the dataset X, then the expected total complexity of the performance of the resulting pivot table is $\Omega(n)$.*

4.1.4 A remark on results of [15]

The above lower bounds agree with an exponential in d upper bound of $k + c^d$ derived in the influential paper [15], Theorem 3 within a similar model, with no restriction on a number n of datapoints, and with d a dimension parameter defined by a certain measure distribution density condition verified, e.g. by the Hamming cube $\{0,1\}^d$ or the Euclidean sphere $\mathbb{S}^d$. Here c is a constant depending on Ω, the smallest distortion parameter of a 1-Lipschitz embedding $f\colon \Omega \to \ell^\infty(k)$:

$$\forall x, y \in \Omega, \quad \|f(x) - f(y)\|_\infty \leq \rho(x,y) \leq c\,\|f(x) - f(y)\|_\infty .$$

However, the usefulness of the result is limited because of an imprecise claim (*loc. cit.*, Example 1) that for a bounded subset X of $\ell^2(d)$ there always exists a 1-Lipschitz function $f\colon X \to \ell^\infty(d+2)$ having distortion $c \leq 2$. In fact, an optimal constant here is on the order $O(\sqrt{d})$, and so the query performance estimate for the Euclidean domains made in Remark after the main Theorem 3, *loc.cit.*, becomes superexponential in d and thus meaningless.

This has created a misconception which led the authors of [6] to believe that using the prior knowledge of Euclideanness of the domain would make the construction of an optimal pivot-based scheme easy (cf. their remarks on p. 2358, end of first paragrph on the r.h.s., and at the beginning of Section 5).

4.2 Hierarchical metric tree schemes

4.2.1 Metric trees

For a finite rooted tree T we denote $L(T)$ the set of leaves of T and $I(T)$ the set of inner nodes. The symbol $*$ will denote the root node of T.

Let $\mathscr{F}$ be a class of 1-Lipschitz functions on Ω (possibly partially defined). A *metric tree* (*of type* $\mathscr{F}$) for a workload (Ω, ρ, X) is a hierarchical indexing structure consisting of

- a finite binary rooted tree $\mathcal{T}$,
- an assignment of a function $f_t \in \mathscr{F}$ (a *pruning,* or *decision function*) to every inner node $t \in I(T)$, and
- a collection of subsets $B_t \subseteq \Omega$, $t \in L(T)$ (*bins*), covering the dataset: $X \subseteq \cup_{t \in L(T)} B_t$.

For simplicity, we will assume that the tree T is binary and thus can be identified with a sub-tree of the prefix tree, that is, a subset of binary strings $\varepsilon_1\varepsilon_2\ldots\varepsilon_k$, $0 \leq k \leq n$, where $\varepsilon_i = \pm 1$ for all i. The indexing scheme is constructed so as to assure the following condition. Let $s = \varepsilon_1\varepsilon_2\ldots\varepsilon_m$ be a leaf node, and let x belong to the bin B_s labelled with s. Then for each $0 \leq k \leq m$, the value of $f_{\varepsilon_1\varepsilon_2\ldots\varepsilon_i}$ at x equals ε_{i+1}.

At each inner node $t = \varepsilon_1\varepsilon_2\ldots\varepsilon_l$ the value of the pruning function f_t at the query center q is evaluated. The condition $f_t(q) > \varepsilon$ gurantees that the child node $t(-1) =$

$\varepsilon_1\varepsilon_2\ldots\varepsilon_l(-1)$ need not be visited, because all elements x of the bins indexed with the descendants of $t(-1)$ are at a distance $> \varepsilon$ from q. Indeed, assuming $x \in B_\varepsilon(q)$, one has

$$|f_t(x) - f_1(q)| \leq d(x,q) \leq \varepsilon.$$

Similarly, if $f_t(q) < -\varepsilon$, then the node $t1 = \varepsilon_1\varepsilon_2\ldots\varepsilon_k 1$ can be pruned, because no bin labelled with descendants of $t1$ can possibly contain a point within the range ε from q.

However, if $f_t(q) \in [-\varepsilon, \varepsilon]$, then no pruning is possible and both children nodes of t have to be visited. The search branches out. In the presence of concentration, the amount of branching is considerable, and results in dimensionality curse.

The M-tree [10] is by now a classical example of a metric tree. However, metric tree-type indexing schemes are very numerous, cf. Sections 2.1-2.4 in [51] or Section 4.5 in [42].

4.2.2 Lower bounds for metric trees

THEOREM 4.3. *In addition to the assumptions of Subs. 2.2, suppose that the VC dimension of the class $\mathscr{F}$ of* 1*-Lipschitz functions used to construct a particular type of metric tree is* $\mathrm{poly}(d)$*. Then the expected average performance of a metric tree indexing structure is superpolynomial in d.*

That the assumption $\mathrm{VC}(\mathscr{F}) = \mathrm{poly}(d)$ is sensible, follows from a theorem of Goldberg and Jerrum [17]. Consider the parametrized class

$$\mathscr{F} = \{x \mapsto f(\theta, x) \colon \theta \in \mathbb{R}^s\}$$

for some $\{0,1\}$-valued function f. Suppose that, for each input $x \in \mathbb{R}^n$, there is an algorithm that computes $f(\theta, x)$, and this computation takes no more than t operations of the following types:

- the arithmetic operations $+, -, \times$ and / on real numbers,
- jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisons of real numbers, and
- output 0 or 1.

Then $\mathrm{VC}(\mathscr{F}) \leq 4s(t+2)$.

ON THE PROOF OF THEOREM 4.3. (For details, see [40].) As the total content of bins B_t indexed with strings t of length superpolynomial in d has to be asymptotically negligible, we can assume the total number of bins to be $2^{\mathrm{poly}(d)}$. Statistical learning estimates imply that measures of bins cannot be too skewed, for a bin of large measure has to contain many points, leading to the desired estimate. Now concentration is used to prove that at least $\mathrm{poly}(d)$ bins B_t have size so large that the ε_M-neighbourhood of B_t has almost full measure. One deduces further that query centres q whose ε_{NN}-neighbourhood meets at least $d^{\omega(1)}$ bins have measure $1/2 - o(1)$. Processing the nearest neighbour query with such a centre q requires accessing all of these bins, let even to verify that it is empty. □

4.3 The curse of dimensionality conjecture

4.3.1 The problem

Of course the above are just particular results only applicable to specific indexing schemes. If one wants to validate the curse of dimensionality once and for all, here is a fascinating open problem.

CONJECTURE 4.4 (CF. [20]). *Let X be a dataset with n points in the Hamming cube $\{0,1\}^n$. Suppose $d = n^{o(1)}$ and $d = \omega(\log n)$. Then any data structure for exact nearest neighbour search in X, with $d^{O(1)}$ query time, must use $n^{\omega(1)}$ space.*

The data structure and algorithm are understood in the sense of the *cell probe model* of computation (cf. [31, 5]). It is a natural choice when one is interested in lower bounds on the performance of general indexing schemes.

4.3.2 Cell probe model

In the context of similarity search, the model can be described as follows. An abstract indexing structure for a domain Ω consists of

- a collection of functions f_t indexed with inner nodes of a rooted tree T,
- a collection of cells C_i, indexed with a set I, and
- a mapping $t \mapsto i(t)$ from T to I (not necessarily one-to-one).

Every function f_t is defined on a subset of Ω and besides takes a b-bit string σ as a parameter, except if $t = 0$ is the root. A value $f_t(\sigma; q)$ is a pair (τ, s), consisting of a b-bit string τ and a child s of the node t. If $i = i(t)$ corresponds to an inner node, the cell C_i can hold a b-bit string. If $i = i(t)$ where t is a leaf, then C_i can hold a datapoint $x \in X$. Often the problem is replaced with a weaker *decision version,* whereby a range parameter $\varepsilon_0 > 0$ is fixed and the algorithm is expected to tell whether there is an $x \in X$ at a distance $< \varepsilon_0$ from the query point. In such a case, C_i will hold a single bit (a "yes" or "no" answer).

Building the data structure at the preprocessing stage, given a dataset X, consists in storing in every inner node cell a b-bit string, and in every leaf node cell (a bin) a pointer to a datapoint $x \in X$ (or, in the decision version, a bit).

A memory image of the indexing structure $C_i, i \in I$ is created when the algorithm is initialized. Given a query point $q \in \Omega$, the tree is traversed down to the leaf level beginning with the root. At the inner node t, the content σ of the cell $C_{i(t)}$ is read and passed on to the function f_t as a parameter. The computed value $f_t(\sigma; q)$ consists of a string, to be written in the cell $C_i(t)$ instead of the current value, and of a child s of t to follow at the next step. When a leaf s is reached, the algorithm halts. The query time is the length of the branch traversed, or equivalently the number of cells probed during the execution of the algorithm.

The cell probe model is very liberal. Indeed, the cost of storing functions f_t is not even taken into account, the space is just the number of cells C_i. Similarly, the cost of computing the values of f_t is disregarded. For this reason, any lower bound obtained under the cell probe model will likely hold under any other model of computation. Since the nearest neighbour problem is stronger than the decision version ("near neighbour problem"), any lower bound for the decision version will serve as a lower bound for the conjecture.

As an example, the pivot table will be encoded within the cell probe model in a rather cumbersome way, because coordinates of each pivot p in the Hamming cube will occupy $\lceil d/b \rceil$ cells, and the distance function $\rho(p, -)$ will need to be computed recursively in $O(d/b)$ steps, with values of partial distances on substrings stored and then passed on further as parameters. For the same reason, our lower bound on the

query time of the pivot algorithm becomes $\Omega(n/\lg n)$.

4.3.3 *Current state of the problem*

The best lower bound currently known is $O(d/\log \frac{sd}{n})$, where s is the number of cells used by the data structure [36]. In particular, this implies the earlier bound $\Omega(d/\log n)$ for polynomial space data structures [3], as well as the bound $\Omega(d/\log d)$ for near linear space (namely $n \log^{O(1)} n$).

5. APPROXIMATE NN SEARCH AND DIMENSIONALITY REDUCTION

Approximate nearest neighbour search [35] is often said to be free from the curse of dimensionality, and the reason is that the (dimensionality) reduction maps f used in indexing are no longer 1-Lipschitz, but what may be called "probably approximately 1-Lipschitz". They no longer exhibit a strong concentration around their means. The price to pay is that we may lose some relevant datapoints, as some distances may get distorted. Curiously, the construction of reduction maps is often based on the concentration phenomemon and/or the VC theory.

5.1 Random projections

Let $\mathbb{S}^{d-1}$ denote the Euclidean sphere of unit radius sitting in the space $\mathbb{R}^d$. The projection π_1 on the first coordinate is a 1-Lipschitz function. For all pairs of points $x, y \in \mathbb{S}^{d-1}$, one has $|\pi_1(x) - \pi_1(y)| \le \|x - y\|$, and for antipodal pairs of points the equality is achieved. Now let $x, y \in \mathbb{S}^{d-1}$ be drawn at random. What is the expected value of the distortion $|\pi_1(x) - \pi_1(y)| / \|x - y\|$?

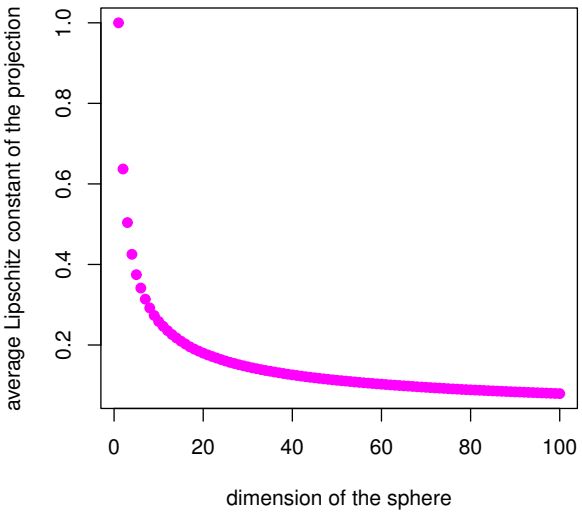

Figure 9: The expected distortion of one-dimensional projection of the d-dimensional sphere $\mathbb{S}^{d-1}$ over all pairs of points.

Figure 9 shows that for a vast majority of pairs of points, the projection distorts distances by the factor $O(1/\sqrt{d})$. A geometric explanation is very simple. A randomly chosen pair of points on the high-dimensional sphere, because to concentration, are at a distance $\approx \sqrt{2}$. At the same time, half of the points of the sphere project on the interval of length $O(1/\sqrt{d})$, and so are contained in the equatorial region (Figure 10).

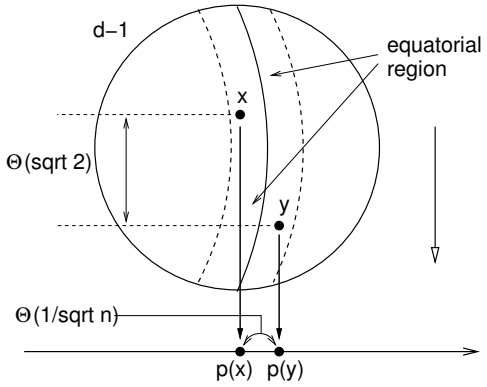

Figure 10: To the geometry of random projections.

If we define a new mapping

$$f(x) = C\sqrt{n}\pi(x),$$

for a suitably chosen constant C, it will be approximately 1-Lipschitz for a finite fraction of pairs of points. Now, in order to achieve a distortion in the range $1 \pm \varepsilon$ with high confidence, it is enough to combine $k = O(\log n/\varepsilon^2)$ mutually orthogonal projections as above, that is, to project on a randomly chosen k-dimensional subspace. This is the famous *Johnson–Lindenstrauss lemma* [22]. (See [27] for an up-to-date state of the lemma.) The projection is not quite as good as a genuine 1-Lipschitz map, because the distortion of a distance can exceed one, and occasionally very considerably. Yet, as a reduction mapping for *approximate NN search,* the projection map is quite OK. And its histogram is concentrated *no more.* This explains the efficiency of the random projection method for *approximate* NN search. Combined with an indexing scheme in a low-dimensional space $\mathbb{R}^k$, this dimensionality reduction leads to an indexing scheme for an $(1+\varepsilon)$-approximate NN search (Indyk and Motwani [21]).

5.2 Algorithm of Kushilevitz, Ostrovsky and Rabani

The indexing scheme constructed in [25] at the same time and independently from [21] is an application of VC theory.

Think of the Hamming cube $\{0,1\}^d$ as the set of all binary functions in the space $\ell^1(d) = L^1([d])$, where $[d] = \{1, 2, \ldots, d\}$ supports a uniform measure. If the dataset $X \subseteq \{0,1\}^d$ contains n points, then the VC dimension of X, viewed as a concept class on $\{1, 2, \ldots, d\}$, does not exceed $\lg_2 n$. According to the Glivenko–Cantelli theorem, if $O(\varepsilon^{-2} \lg_2 n)$ coordinates of the Hamming cube are chosen at random, then with high confidence the restriction mapping from X to the Hamming cube $\{0,1\}^{O(\varepsilon^{-2} \lg_2 n)}$ preserves the pairwise distances to within a factor of $1 \pm \varepsilon$. Cf. Figure 11.

A further analysis allows to conclude that the same will hold for the distances within $X \cup \{q\}$, for a *vast majority* of query points q in $\{0,1\}^d$. Since the new cube only contains $2^{O(\varepsilon^{-2} \lg_2 n)}$ points, a hash table storing nearest neighbourhs,

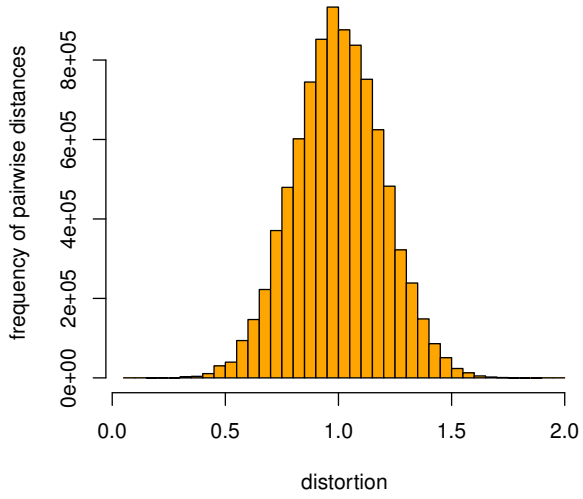

Figure 11: Histogram of distortions of all pairwise distances in a random dataset of $n = 3,000$ points in the $d = 500$ Hamming cube under a projection to a Hamming cube on randomly chosen $k = 25$ bits.

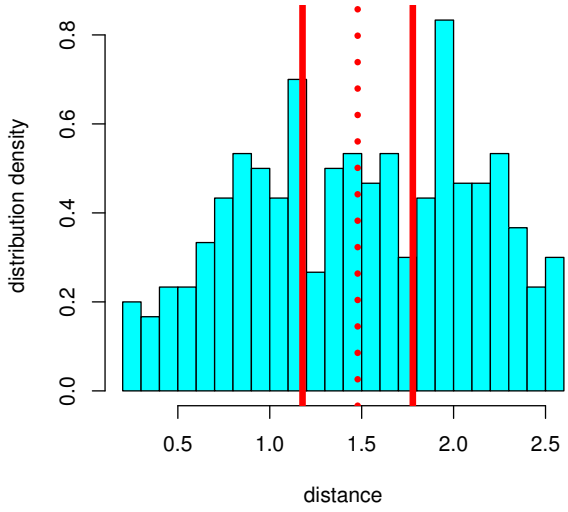

Figure 12: Empirical density histogram of distances from a pivot having the highest found value of dissipation for the NASA dataset. Vertical lines mark the mean $\pm$ tolerance range $\varepsilon = 0.275$. The ϵ-dissipation (0.747) is the area outside of extreme lines.

together with the reduction map f, produces an indexing scheme taking space polynomial in n and answering $(1+\varepsilon)$-approximate NN queries in time $O(\varepsilon^{-2} \lg_2 n)$.

This argument naturally extends to a framework where the metric ρ on the domain is an L^p-distance for some p in the range $1 \leq p < \infty$. In other words, there is a measure space (S, ν) with $\Omega \subseteq L^p(S, \nu)$.

The dimensionality reduction methods have been shown to be near optimal in the cell probe model [1].

6. CONCLUDING REMARKS

6.1 Intrinsic dimensionality

Merits of asymptotic analysis of indexing algorithms using artificial datasets sampled from theoretical high-dimensional distributions should be clear from [33]. At the same time, it is an often held belief that the real data does not have very high intrinsic dimension. This corresponds to the existence of 1-Lipschitz functions that are highly dissipating. Figure 12 shows the distance distribution to the points of the SISAP benchmark dataset of NASA images $X \subseteq \ell^2(20)$ of $40,149$ vectors in a 20-dimensional Euclidean space [6, 44] from a highly dissipating pivot, selected from a gaussian cloud around X with standard deviation on the order of the tolerance range $\varepsilon = 0.275$ retrieving on average 0.1% of data. This has to be compared to Figure 3.

Of a great variety of approaches to intrinsic dimension [13], at least two specifically measure the amount of concentration in data. The first one is the intrinsic dimension by Chávez *et al.* [9]

$$\dim_{dist}(X) = \frac{1}{2\mathrm{var}\,(d)}. \tag{6}$$

The second is the concentration dimension, studied within an axiomatic approach of [38, 39]:

$$\dim_{\alpha}(X) = \frac{1}{\left[2\int_0^1 \alpha_X(\varepsilon)\; d\varepsilon\right]^2}. \tag{7}$$

(In both cases we assume that $\mathrm{CharSize}(X) = 1$.) The value (7) is convenient for asymptotic analysis in the spirit of this paper, but is nearly impossible to estimate for a given dataset. On the other hand, (6) is readily calculated by sampling (e.g. $\dim_{dist}(X) = 5.18$ for NASA images) and forms a good statistical estimator for the dimension of the hypothetical underlying measure μ in the most (only?) interesting case where metric balls have low VC dimension. The shortcoming of (6) is that the parameter estimates the concentration/dissipation behaviour of a *typical* pivot distance function, while it is a few most dissipating pivots that really matter for indexing. One may envisage the emergence of further concepts of intrinsic dimension in the same spirit. For instance, the *local dimension* of Ollivier [34], Definition 3, deserves a close look.

6.2 Black box search model and Urysohn space

The *black box model* of similarity search was studied by Krauthgamer and Lee [24]. Given a finite metric space (X, d), a query is a one-point metric space extension $X \cup \{q\}$, where every distance $d(q, x)$, $x \in X$ is accessible via the distance oracle and can be evaluated in constant (unit) time. A preprocessing phase is allowed, and the indexing scheme occupies poly (n) space. The query time is a number of calls to the distance oracle.

Recall that the *Assouad* (or *doubling*) *dimension* of a met-

ric space (X, d) is the minimum value $\rho \geq 0$ such that every set A in X can be covered by 2^ρ balls of half the diameter of A. Denote this parameter by $\dim_{dbl}(X)$. The above authors have shown that a metric space (X, d) admits an algorithm requirying poly (n) space and taking polylog (n) time to answer a $(1 + \varepsilon)$-approximate nearest neighbour query, where $\varepsilon < 2/5$, if and only if

$$\dim_{dbl}(X, d) = O(\log \log n).$$

The *Urysohn metric space,* $\mathbb{U}$ [28], is a complete separable metric space uniquely defined by the following *one-point extension property:* suppose X is a finite subset of $\mathbb{U}$ and q a one-point metric space extension of X. Then $\mathbb{U}$ contains a point q' so that the distances from q and from q' to any point $x \in X$ are the same. The black-box model can be restated as a classical NN search problem with $\Omega = \mathbb{U}$ as the domain. A simple information-theoretic argument shows that any deterministic algorithm for exact similarity search in a finite metric space X within this model will take the worst case time $n = |X|$. However, if one requires the queries to follow the same underlying distribution as datapoints, the problem becomes more subtle, and we do not know the answer.

6.3 Indexing via Delaunay graph

By far not every indexing scheme for exact similarity search is "distance-based." Here is an example. The *Voronoi cell* $V(x)$ of a datapoint $x \in X$ in a domain Ω consists of all points $q \in \Omega$ having x as the nearest neighbour. The *Delaunay graph* has X as the set of vertices, with x, y being adjacent if their Voronoi cells intersect. Suppose the domain has the property that every two points x, y can be joined by a shortest geodesic path, not necessarily unique. (All the domains previously considered are such.) Then for any $q \in \Omega$ and $x \in X$, either x is the nearest neighbour to q, or else one of the datapoints y adjacent to x is closer to q than x is. (Proof: start moving along a shortest geodesic from x towards q.) This makes the Delaunay graph of X into an indexing scheme for exact nearest neighbour search. Denote S_x the list of points adjacent to each $x \in X$. Given a query q, start with an arbitrary $x_0 \in X$, and find

$$x_1 = \arg\min_{y \in S_x} d(q, y).$$

If $x_1 \neq x_0$, move to x_1 and repeat the procedure. Once $x_{i+1} = x_i$, the algorithm halts and returns x_i. This algorithm, already mentioned in [12], was studied for general metric spaces by Navarro [32]. See also [42], 4.1.6.

In order to be efficient, the average degree of the Delaunay graph has to be small. Navarro had observed (*loc. cit.,* Theorem 1) that this is not the case in general metric spaces. In fact, one can deduce from Corollary 2.2 that if Ω be either $\mathbb{R}^n$, or the sphere $\mathbb{S}^n$, or the Hamming cube, then under the assumptions of Subs. 2.2 the Delaunay graph of X is, with high probability, the complete graph on n vertices.

Thus, the indexing scheme in question suffers from the curse of dimensionality because of concentration of measure considerations, but the argument seems to be of a different nature from that either for pivots or for trees. What would a common proof for all three types of schemes look like? This highlights the difficulty of proving in a uniform way lower bounds in a general setting of the cell probe model for all possible indexing schemes at once.

To finish on an optimistic note, for real data the complexity of the Delaunay graph is lower than in an artificial asymptotic setting, and Voronoi diagrams are being successfully used for data mining algorithms in high dimensions, cf. [46]. It further seems the above algorithm could be especially efficient in hyperbolic metric spaces, and Alain Connes had suggested [7], pp. 138–141, that a long-term human memory is organized as a hyperbolic simplicial complex, where a search is performed in a manner similar to the above.

7. REFERENCES

[1] A. Andoni, P. Indyk, M. Pătrascu, On the optimality of the dimensionality reduction method, in: Proc. 47th IEEE Symp. on Foundations of Computer Science, pp. 449–458, 2006.

[2] M. Anthony and P. Bartlett, *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, Cambridge, 1999.

[3] O. Barkol and Y. Rabani. Tighter lower bounds for nearest neighbor search and related problems in the cell probe model. In: *Proc. 32nd ACM Symp. on the Theory of Computing,* 2000, pp. 388–396.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, When is "nearest neighbor" meaningful?, in: *Proc. 7-th Intern. Conf. on Database Theory (ICDT-99),* Jerusalem, pp. 217–235, 1999.

[5] A. Borodin, R. Ostrovsky, and Y. Rabani, Lower bounds for high-dimensional nearest neighbor search and related problems, in: *Proc. 31st Annual ACS Sympos. Theory Comput.*, pp. 312–321, 1999.

[6] B. Bustos, G. Navarro, and E. Chávez, Pivot selection techniques for proximity searching in metric spaces. *Pattern Recogn. Lett.*, 24:2357–2366, 2003.

[7] J.-P. Changeux and A. Connes, *Conversations on Mind, Matter, and Mathematics,* Princeton University Press, Dec. 1998.

[8] E. Chávez, J.L. Marroquín, and R.A. Baeza-Yates, Spaghettis: An Array Based Algorithm for Similarity Queries in Metric Spaces, in: *Proceedings SPIRE/CRIWG,* pp. 38–46, 1999.

[9] E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquín, Searching in metric spaces, *ACM Computing Surveys,* 33:273–321, 2001.

[10] P. Ciaccia, M. Patella and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proceedings of 23rd International Conference on Very Large Data Bases (VLDB'97), (Athens, Greece)*, 426–435, 1997.

[11] P. Ciaccia, M. Patella and P. Zezula. A cost model for similarity queries in metric spaces, in: *Proc. 17-th ACM Symposium on Principles of Database Systems* (PODS'98), Seattle, WA, 59–68, 1998.

[12] K.L. Clarkson. An algorithm for approximate closest-point queries. In: *Proc. 10th symp. Comp. Geom.* Stony Brook, NY, 160–164, 1994.

[13] K.L. Clarkson. Nearest-neighbor searching and metric space dimensions. In: Nearest-Neighbor Methods for Learning and Vision: Theory and Practice, MIT Press, 2006, pp. 15–59.

[14] L. Devroye, L. Györfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York, 1996.

[15] A. Faragó, T. Linder, and G. Lugosi, Fast nearest neighbor search in dissimilarity spaces, *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 18:957–962, 1993.

[16] P. Frankls and Z. Füredi. A short proof for a theorem of Harper about Hamming-spheres, *Discrete Math.* 34:311–313, 1981.

[17] P.W. Goldberg and M.R. Jerrum. Bounding the Vapnik–Chervonenkis dimension of concept classes parametrized by real numbers. *Machine Learning* 18:131-148, 1995.

[18] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces.* Progress in Mathematics **152**. Birkhauser Verlag, 1999.

[19] P. Hall, J.S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67:427–444, 2005.

[20] P. Indyk. Nearest neighbours in high-dimensional spaces, In: J.E. Goodman, J. O'Rourke, Eds., *Handbook of Discrete and Computational Geometry*, Chapman and Hall/CRC, Boca Raton–London–New York–Washington, D.C. 877–892, 2004.

[21] P. Indyk and R. Motwani. Approximate nearest neighbours: towards removing the curse of dimensionality. Proc. 30th ACM Symp. Theory of Computing, 604–613, 1998.

[22] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* 26:189–206, 1984.

[23] B. Klartag. A central limit theorem for convex sets. *Invent. Math.* 168:91–131, 2007.

[24] R. Krauthgamer and J.R. Lee, *The black-box complexity of nearest-neighbor search,* Theoretical Computer Science 348(2):262 - 276, December 2005.

[25] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces. *SIAM Journal on Computing* 30:457–474, 2000.

[26] M. Ledoux. *The concentration of measure phenomenon*. Math. Surveys and Monographs **89**, Amer. Math. Soc., 2001.

[27] J. Matoušek. On variants of the Johnson–Lindensrauss lemma. *Random Structures and Algorithms* 33:142–156, 2008.

[28] J. Melleray. On the geometry of Urysohn's universal metric space. *Topology and its Applications* 154:384-403, 2007.

[29] V. Milman. Topics in asymptotic geometric analysis. In: *Geometric and Functional Analysis,* special volume GAFA2000, pp. 792–815, 2000.

[30] V.D. Milman and G. Schechtman. *Asymptotic theory of finite-dimensional normed spaces (with an Appendix by M. Gromov)*. Lecture Notes in Math., **1200**, Springer, 1986.

[31] P.B. Miltersen. Cell probe complexity - a survey. In: *19th Conf. on the Foundations of Software Technology and Theoretical Computer Science (FSTTCS),* 1999. Advances in Data Structures Workshop.

[32] G. Navarro. Searching in metric spaces by spatial approximation. *The VLDB Journal* 11:28–46, 2002.

[33] G. Navarro. Analysing metric space indexes: what for? Invited paper, in: *Proc. 2nd Int. Workshop on Similarity Search and Applications (SISAP 2009)*, Prague, Czech Republic, 2009, 3–10.

[34] Yann Ollivier. Ricci curvature of metric spaces. *C.R. Math. Acad. Sci. Paris* 345:643–646, 2007.

[35] M. Patella and P. Ciaccia. The many facets of approximate similarity search. Invited paper, in: *Proc. First Int. Workshop on Similarity Search and Applications (SISAP 2008),* Cancun, México, pp. 10–21, 2008.

[36] M. Pătrascu and M. Thorup. Higher lower bounds for near-neighbor and further rich problems. in: *Proc. 47th IEEE Symp. on Foundations of Computer Science*, pp. 646–654, 2006.

[37] V. Pestov, On the geometry of similarity search: dimensionality curse and concentration of measure. *Inform. Process. Lett.*, 73:47–51, 2000.

[38] V. Pestov. Intrinsic dimension of a dataset: what properties does one expect? in: *Proc. of the 22-nd Int. Joint Conf. on Neural Networks (IJCNN'07), Orlando, FL.*, pp. 1775–1780, 2007.

[39] V. Pestov. An axiomatic approach to intrinsic dimension of a dataset. *Neural Networks* 21:204-213, 2008.

[40] V. Pestov. *Lower bounds on performance of metric tree indexing schemes for exact similarity search in high dimensions.* preprint, arXiv:0812.0146 [cs.DS].

[41] V. Pestov and A. Stojmirović. Indexing schemes for similarity search: an illustrated paradigm, *Fund. Inform.*, 70:367–385, 2006.

[42] H. Samet. *Foundations of Multidimensional and Metric Data Structures.* Morgan Kaufmann Publishers Inc., San Francisco, CA, 2005.

[43] U. Shaft and R. Ramakrishnan, Theory of nearest neighbors indexability. *ACM Trans. Database Syst. (TODS),* vol. 31, pp. 814–838, 2006.

[44] SISAP metric space library, `http://sisap.org/Metric_Space_Library.html`

[45] A. Stojmirović and V. Pestov, Indexing schemes for similarity search in datasets of short protein fragments, *Information Systems* 32:1145-1165, 2007.

[46] K. Taşdemir and E. Merényi. Exploiting the data topology in visualizing and clustering of self-organizing maps. *IEEE Trans. Neural Networks* 20(4):549–562, 2009.

[47] V.N. Vapnik, *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.

[48] E. Vidal, An algorithm for finding nearest neighbors in (approximately) constant average time, *Pattern Recognition Letters* 4:145–157, 1986.

[49] M. Vidyasagar, *Learning and Generalization, With Applications to Neural Networks.* Second Ed., Springer-Verlag, London, 2003.

[50] I. Volnyansky and V. Pestov. Curse of dimensionality in pivot-based indexes. In: *Proc. 2nd Int. Workshop on Similarity Search and Applications (SISAP 2009),* Prague, Czech Republic, 2009, pp. 39-46.

[51] P. Zezula, G. Amato, V. Dohnal and M. Batko, *Similarity Search. The Metric Space Approach.* Springer Science + Business Media, New York, 2006.