Journal of Discrete Algorithms 13 (2012) 2–18

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Indexability, concentration, and VC theory

Vladimir Pestov [a,b,*,1]

[a] Departamento de Matemática, Universidade Federal de Santa Catarina, Campus Universitário Trindade, CEP 88.040-900 Florianópolis-SC, Brazil
[b] Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Ontario, K1N6N5 Canada

ARTICLE INFO

Article history:
Available online 3 November 2011

Keywords:
Exact similarity search
Indexing schemes
Curse of dimensionality
Lipschitz functions
Concentration of measure
Uniform Glivenko–Cantelli theorem
Pivot tables
Metric trees

ABSTRACT

Degrading performance of indexing schemes for exact similarity search in high dimensions has long since been linked to histograms 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.

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 functions f_i, $i \in I$ on Ω, either fully or partially defined, which satisfy the 1-Lipschitz property:

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

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

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. Fig. 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 $(\mathbb{R}^n, \gamma^n)$,

* Correspondence to: Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Ontario, K1N6N5 Canada.
E-mail address: vpest283@uottawa.ca.

[1] CNPq visiting researcher.

1570-8667/$ – see front matter
doi:10.1016/j.jda.2011.10.002

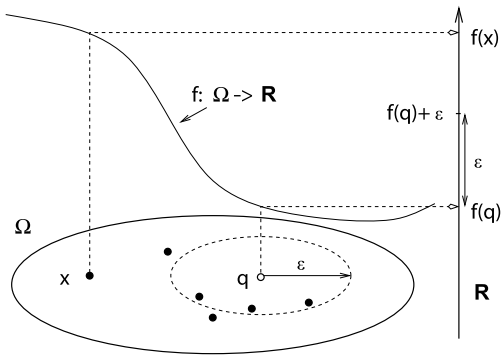

Fig. 1. The datapoint x can be discarded.

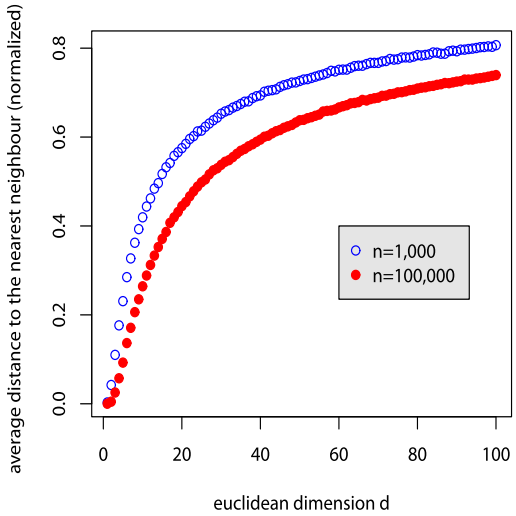

Fig. 2. The normalized average distance to the nearest neighbour in a dataset of n points randomly drawn from a gaussian distribution in $\mathbb{R}^d$.

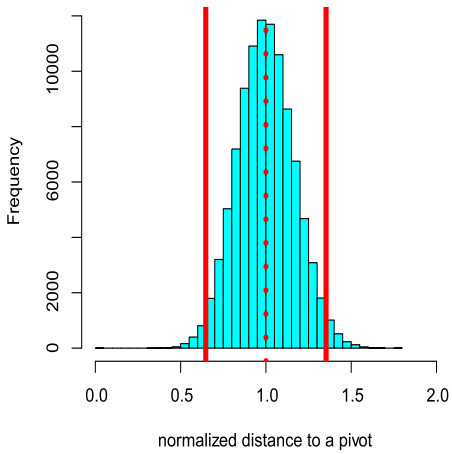

Fig. 3. Histogram of distances to a randomly chosen pivot in a dataset X of $n = 10^5$ points drawn from a gaussian distribution in $\mathbb{R}^{14}$. The vertical lines mark the mean normalized distance $1 \pm \varepsilon_{NN}$.

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

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 Fig. 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. [8], pp. 35–37, [36,46], p. 487, to mention just a few sources.

To make this idea yield rigorous lower performance bounds, one needs to guarantee first that *every* histogram of distances of 1-Lipschitz functions used to build an indexing scheme for a given domain Ω is highly concentrated. In other words, if $\mathcal{F}$ denotes a class of 1-Lipschitz functions from which we can choose the f_i, then we want a low uniform upper bound on the variances of $f \in \mathcal{F}$. 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* [20,34,29].

Next problem is, how to link the concentration of functions f with regard to the presumed *underlying distribution* μ on the domain Ω to concentration with regard to the *empirical measure* μ_n supported on the dataset X (this was essentially a criticism of [41] made in [47])? Here one needs the machinery of *statistical learning theory* of Vapnik and Chervonenkis [52,2,14,56], which can guarantee such results provided the class $\mathcal{F}$ has low combinatorial complexity (e.g., a finite VC

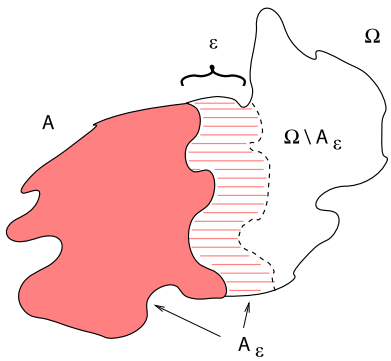

Fig. 4. To the concept of the concentration function $\alpha_\Omega(\varepsilon)$.

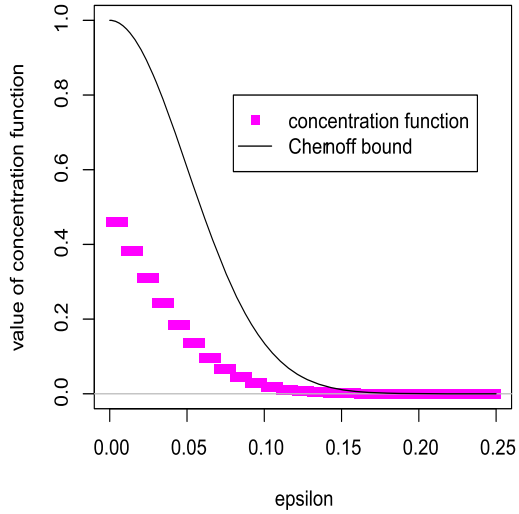

Fig. 5. Concentration function of the Hamming cube of dimension $d = 100$ *vs* Chernoff bound.

dimension). This way, one obtains $\Omega(n/d \lg n)$ lower bounds for the pivot table expected average performance [57], as well as superpolynomial in d lower bounds for metric trees [44].

Approximate NN queries [23,39] 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 in the Hamming cube (the approach of Kushilevitz, Ostrovsky and Rabani [28]), as well as the Euclidean space (the Johnson–Lindenstrauss lemma [24]).

Getting back to exact search, the *Curse of Dimensionality Conjecture* [22] calls for a general statement about lower bounds, 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 a few remarks on the notion of intrinsic dimensionality of data, on a black-box search model of Krauthgamer and Lee [27], as well as on a spatial approximation algorithm based on Delaunay graphs [36].

2. Concentration

2.1. The concentration of measure phenomenon

Informally, the phenomenon can be stated as follows:

On a typical "high-dimensional" structure Ω, every 1-Lipschitz function $f : \Omega \to [0,1]$ has small variation.

Usually, however, concentration is being dealt with using a different dispersion parameter. We proceed to precise definitions.

Let a metric space (Ω, ρ) carry a probability measure μ. Such an object is called a *metric space with measure*. One defines the *concentration function* α_Ω of Ω by setting $\alpha_\Omega(0) = 1/2$ and, for $\varepsilon > 0$,

$$\alpha_\Omega(\varepsilon) = 1 - \inf_{A \subseteq \Omega} \left\{ \mu(A_\varepsilon) \colon \mu(A) \geqslant \frac{1}{2} \right\}.$$

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 $\geqslant 1/2$, 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, one can usually derive gaussian upper bounds of the form

$$\alpha(\varepsilon) \leqslant \exp\bigl(-\Theta\bigl(\varepsilon^2 d\bigr)\bigr),$$

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 $\alpha(\varepsilon) \leqslant \exp(-2\varepsilon^2 d)$ (obtained by combining Harper's isoperimetric inequality, see e.g. [16], with the classical Chernoff bound, cf. [56], 2.2.1). See Fig. 5.

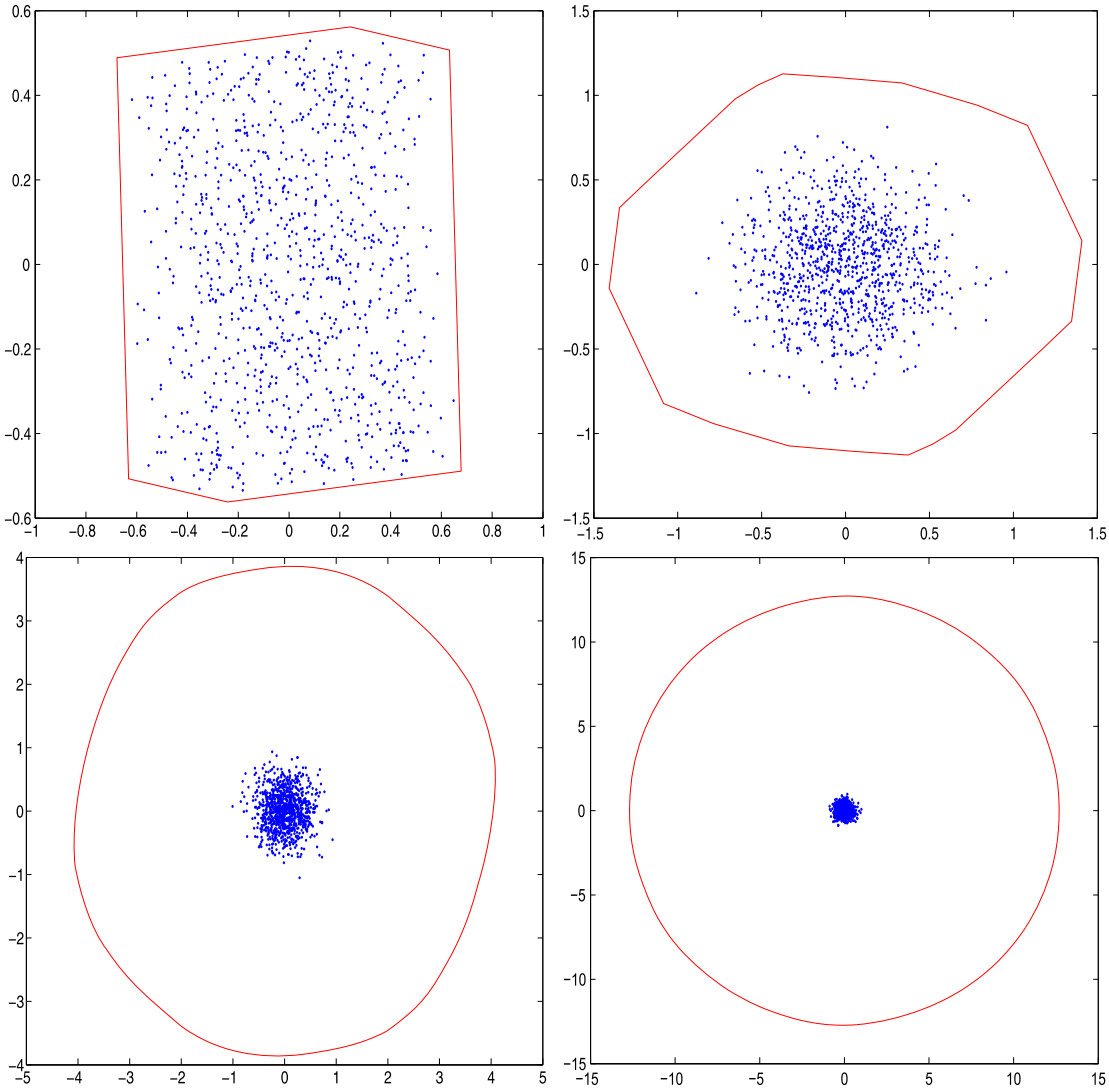

Fig. 6. Orthogonal projection of a unit Euclidean d-cube and of 1000 random points inside the cube on a random 2-subspace, $d = 3$ (top left), $d = 10$ (top right), $d = 100$ (bottom left), $d = 1000$ (bottom right). right). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

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

$$\mu\left\{x \in \Omega\colon \left|f(x) - M_f\right| > \varepsilon\right\} \leqslant 2\alpha_\Omega(\varepsilon), \tag{3}$$

where M_f is the median value of f, that is, a (generally non-unique) real number with the property that for a randomly drawn $x \in \Omega$ the probabilities of the events $[f(x) \geqslant M]$ and $[f(x) \leqslant M]$ are at least 1/2 each. One can further derive uniform upper bounds in terms of α_Ω on the variances of 1-Lipschitz functions on Ω with values in a bounded interval.

The concentration phenomenon admits the following illustration. Draw 1000 points randomly from a high-dimensional geometric object such as the unit cube $\mathbb{I}^d$ centred at the origin, choose a random orthogonal projection onto a two-dimensional subspace, and project both the cube and the chosen points on this subspace. The points will concentrate near the centre, the more so the higher the dimension d is, as seen in Fig. 6 for the values of dimension $d = 3$, 10, 100 and 1000. The red outline is the two-dimensional projection of the cube.

Another noteworthy consequence of concentration is that the shape of the random projection of the cube is getting ever more similar to a disk as $d \to \infty$. In fact, for $d = 1000$ the only visual difference one can spot between a random projection of a unit cube and that of a unit sphere, is the scale of the two projections: 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, a certain precise version of this statement holds true for *all* convex bodies, as recently proved by Klartag [26]. For this reason, for an asymptotic study of performance of 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, [34] is the most reader-friendly, [29] most comprehensive, and [19] contains a wealth of ideas. See also a survey [33].

2.2. Asymptotic assumptions on the similarity workload

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

2.2.1. Domain as a metric space with measure

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 used in [10], which of course agrees with the traditional statistical approach to data modelling.)

2.2.2. Normalization of the distance

The distance ρ on the domain is normalized so that the characteristic size of Ω is constant:

$$\mathrm{CharSize}(\Omega) = \mathbb{E}_{\mu\otimes\mu}(\rho) = O(1).$$

(Every domain can be renormalized in the above fashion unless the expected distance between the two points is infinite, which does not appear to be a realistic assumption anyway.)

2.2.3. Growing intrinsic dimension

Ω has "intrinsic dimension d" in the sense that the concentration function of the metric space with measure (Ω, ρ, μ) admits a gaussian upper bound

$$\alpha_\Omega(\varepsilon) = \exp\bigl(-\Omega\bigl(\varepsilon^2 d\bigr)\bigr).$$

(Such an approach to intrinsic dimensionality is developed in [42,43].)

2.2.4. Size of a dataset

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 indexing schemes for similarity search, cf. [22]. An example of such a rate of growth is $n = 2^{\sqrt{d}}$.)

Note that randomly drawing a single dataset $X \subseteq \Omega$ with n points amounts to randomly drawing a single point in the nth power of the domain, Ω^n, equipped with the product probability measure $\mu^{\otimes n}$. In order to perform asymptotic analysis of indexing scheme performance, we will in fact be choosing an *infinite sequence* of datasets $X_d \subseteq \Omega_d$, $d = 1, 2, \ldots$. This is equivalent to drawing a single point $\bar{x}$ (*sample path*) in the infinite product

$$\Omega_1^{n_1} \times \Omega_2^{d_2} \times \cdots \times \Omega_d^{n_d} \times \cdots,$$

with regard to the corresponding infinite product of probability measures:

$$\bar{X} \sim \mu_1^{\otimes n_1} \otimes \mu_2^{\otimes d_2} \otimes \cdots \otimes \mu_d^{\otimes n_d} \otimes \cdots.$$

When talking about *confidence*, we will mean the product probability in the above infinite product space. Specifically, a statement $Q(d, \bar{x})$ parametrized by the dimension d and taking as a variable the sample path $\bar{x}$ *occurs with* (*asymptotically*) *high confidence* if for every $\delta > 0$ there is D so that

$$P\bigl[Q(d, \bar{X}) \text{ is true}\bigr] > 1 - \delta$$

whenever $d \geqslant D$.

At the same time, in order to keep the notation simple, we will suppress the dimension index d and talk just of a single domain Ω and a dataset $X \subseteq \Omega$.

2.3. Empty space paradox

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

Theorem 2.1. *Under our standing assumptions on the workload, for every $\varepsilon > 0$ one has with asymptotically high confidence that for all points $\omega \in \Omega$ except for a set of measure* $\exp(-\Omega(\varepsilon^2 d))$

$$\bigl|\varepsilon_{NN}(\omega) - \mathrm{CharSize}(\Omega)\bigr| < \varepsilon.$$

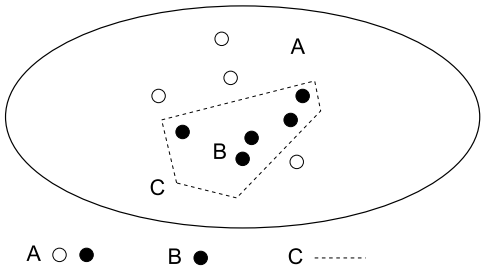

Fig. 7. To the concept of a set A shattered by a class $\mathcal{C}$.

The result applies to the Hamming cube, the Euclidean cube, the Euclidean space with gaussian measure, the Euclidean ball, etc.

As a byproduct of the technique, one obtains:

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

For constant $n = |X|$ and the case of a Euclidean domain the result was established in [21].

Our proofs can be found in Appendix A.

3. VC theory

3.1. VC dimension

Let $\mathcal{C}$ denote a collection of subsets of the domain Ω. The VC dimension is an important measure of combinatorial complexity of $\mathcal{C}$. A finite set $A \subseteq \Omega$ is *shattered* by $\mathcal{C}$ (Fig. 7) if every subset $B \subseteq A$ can be "carved out" of A with the help of a suitable element C of $\mathcal{C}$:

$$B = A \cap C.$$

The *VC dimension* of $\mathcal{C}$, denoted $\mathrm{VC}(\mathcal{C})$, is the supremum of cardinalities of all finite subsets of the domain which are shattered by $\mathcal{C}$. 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 in $\mathbb{R}^d$	$d+1$
Parallelepipeds in $\mathbb{R}^d$	$2d+2$
Convex polygons in $\mathbb{R}^d$	∞
Any family with n sets	$\leqslant \lg_2 n$
Hamming balls in $\{0,1\}^d$	$\leqslant d + \lg_2 d$
If $\mathrm{VC}(\mathcal{C}) = c$, $\mathrm{VC}(\mathcal{D}) = d$:	
$\{\Omega \setminus C\colon\ C \in \mathcal{C}\}$	c
$\mathcal{C} \cup \mathcal{D}$	$\leqslant c + d + 1$
k-fold intersections of members of $\mathcal{C}$	$\leqslant 2k \lg(ek)c$

Proofs can be found e.g. in [56], Ch. 4.

Estimating the VC dimension of a particular family of sets is often a non-trivial task. For example, the value of this parameter does not seem to be known for the collection of all cubes in $\mathbb{R}^d$ with sides parallel to the coordinate hyperplanes. More generally, it is tempting to conjecture that the VC dimension of the family of all balls (either open or closed) in a Banach space of finite dimension d equals $d+1$, but the author is unaware of any results beyond the Euclidean case.

3.2. Uniform convergence of empirical measures

Recall that the *Borel sigma-algebra* of subsets of a separable metric space Ω is the smallest family closed under countable intersections and complements and containing all open balls. Elements of the Borel sigma-algebra are called simply *Borel subsets*. We will restrict our attention to those families $\mathcal{C}$ whose elements are Borel subsets of Ω. This assumption guarantees that the value $\mu(C)$ is well-defined for every probability measure μ on Ω.

The *empirical measure* of $C \in \mathcal{C}$ with regard to a finite sample $X = \{x_1, \ldots, x_n\}$ is just the normalized counting measure

$$\mu_n(C) = \frac{1}{n}\big|\{i\colon x_i \in C\}\big|.$$

The VC dimension of $\mathcal{C}$ is finite if and only if, with high confidence, the empirical measures of every $C \in \mathcal{C}$ converge uniformly to the true value $\mu(C)$ as the sample size n goes to infinity, no matter what the underlying measure μ is.

Here is a more exact formulation. A class $\mathcal{C}$ 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 $n \geqslant s(\delta, \varepsilon)$, one has

$$\sup_{\mu \in P(\Omega)} P\Big\{\sup_{C \in \mathcal{C}}\big|\mu(C) - \mu_n(C)\big| \geqslant \varepsilon\Big\} < \delta.$$

Here $P(\Omega)$ denotes the family of all probability measures on Ω. We quote the following as stated in [56], Theorem 7.8.

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

$$s(\delta, \varepsilon) \leqslant \max\Big\{\frac{8d}{\varepsilon}\lg\frac{8e}{\varepsilon}, \frac{4}{\varepsilon}\lg\frac{2}{\delta}\Big\}.$$

One of the components of the proof is the concentration of measure in the Hamming cube $\{0, 1\}^n$.

Let us remark that similar results can be stated and proved for *function classes*, that is, collections $\mathcal{F}$ of functions from the domain Ω to the interval $[0, 1]$. The role of VC dimension is taken over by other combinatorial parameters, such as the *fat shattering dimension*. We will not enter into details.

Among a great selection of books treating VC theory, let us mention encyclopaedic sources [56] and [14], a classical monograph [52], and a lighter, but very well-written [2].

4. The curse of dimensionality

4.1. Pivot tables

4.1.1. Reduction and access overhead

Let (Ω, ρ, X) be a similarity workload, Υ a metric space, and $f : \Omega \to \Upsilon$ a 1-Lipschitz function. If queries in Υ are easier to process than in Ω, then it makes sense, given a range query (q, ε) in (Ω, X), to run an $(f(q), \varepsilon)$ range query in $(\Upsilon, f(X))$, retrieving all datapoints x with $f(x)$ within the distance of ε of $f(q)$, and then check them against the condition $\rho_\Omega(q, x) < \varepsilon$. The 1-Lipschitz property of f guarantees that no true hits will be missed.

In this way, the function f 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 e.g. in [45]. The *access overhead* of the reduction f is defined as

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

This simple and well-known idea on its own can be surprisingly efficient, cf. [49].

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 the formula

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

Here $\ell^\infty(k)$ is the vector space $\mathbb{R}^k$ equipped 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

$$f(x) = \big(d(x, p_1), \ldots, d(x, p_k)\big) \in \ell^\infty(k). \tag{5}$$

In [55], it was suggested to use a reduction of this form 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. [7].

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

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

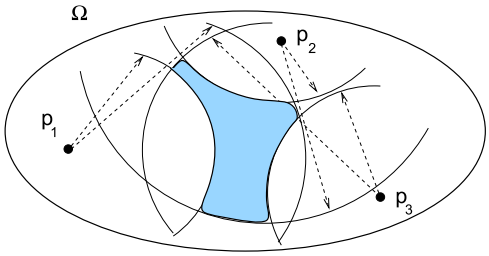

Fig. 8. An intersection of spherical shells.

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. A classical paper on optimizing the pivot selection is [5].

4.1.3. Lower query time bounds for pivot tables

Our next result (a slightly corrected version of the main theorem in [57]) is valid not only 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 Section 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 $\Omega(n/d\log n)$.*

Proof. Assume the number of pivots k is $o(n/d\log n)$. 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 $\geqslant \varepsilon_M$. For each pivot p_i, $i = 1, 2, \ldots, k$, denote by ρ_i^M the median value of the distance function $\rho(p_i, -)$. Because of concentration, the measure of the spherical shell

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

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

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

since n is subexponential in d. Thus, among all k-fold intersections of spherical shells (Fig. 8), we have found a giant one, whose μ-measure is nearly one.

To assure that this intersection contains an accordingly high proportion of datapoints, consult the table in Section 3.1 to deduce that 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 $d \to \infty$.

The measure of the set Q of query points $q \in S$ whose distance to the nearest neighbour in X is greater than or equal to ε_M is at least $1/2 - \exp(-O(d)\varepsilon^2)$. For every non-empty range query (q, ε) where $q \in Q$, all datapoints belonging to S, that is, most datapoints of X, have to be returned. This gives an expected average total complexity $\Omega(n)$ under our assumption on the number of pivots. □

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

Corollary 4.2. *Under the assumptions of Section 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 $n(1 - o(n))$.*

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 $f : \Omega \to \ell^\infty(k)$:

$$\forall x, y \in X, \quad \left\|f(x) - f(y)\right\|_\infty \leqslant \rho(x, y) \leqslant c\left\|f(x) - f(y)\right\|_\infty.$$

However, the usefulness of the result is limited because of an imprecise claim ([15], Example 1) that for a bounded subset X of $\ell^2(d)$ there always exists a 1-Lipschitz function $f : X \to \ell^\infty(d+2)$ having distortion $c \leqslant 2$. In fact, an optimal constant here is on the order $\Omega(\sqrt{d})$ (see Appendix B). As a result, the query performance estimate for the Euclidean domains made in Remark after the main Theorem 3, [15], becomes superexponential in d and thus meaningless.

This misconception has led to some further confusion, cf. remarks made in [5] (p. 2358, end of first paragraph 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 by $L(T)$ the set of leaves of T and $I(T)$ the set of inner nodes. The symbol $*$ will denote the root node of T.

Let $\mathcal{F}$ be a class of 1-Lipschitz functions on Ω (possibly partially defined). A *metric tree* (*of type* $\mathcal{F}$) 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 $f_t \in \mathcal{F}$ (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: $X \subseteq \bigcup_{t \in L(T)} B_t$.

Since we assume that the tree T is binary, it can be identified with a sub-tree of the prefix tree, that is, a subset of binary strings $\varepsilon_1\varepsilon_2\ldots\varepsilon_k$, $0 \leqslant k \leqslant n$, where $\varepsilon_i = \pm 1$ for all i.

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 $f_t(q) > \varepsilon$ guarantees 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

$$\left|f_t(x) - f_1(q)\right| \leqslant d(x,q) \leqslant \varepsilon.$$

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 [9] 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 [58] or Section 4.5 in [46].

4.2.2. Lower bounds for metric trees

For a function f and a real number t, denote $\mathbf{1}_{f \leqslant 1} = \{x \in \mathrm{dom}(f)\colon\ f(x) \leqslant t\}$.

Theorem 4.3. *In addition to the assumptions of Section 2.2, let $\mathcal{F}$ be a class of 1-Lipschitz functions on the domain Ω such that the VC dimension of the family of sets $\mathbf{1}_{f \leqslant t}$, $f \in \mathcal{F}$, $t \in \mathbb{R}$, is* poly(d). *Then the expected average performance of every metric tree indexing structure of type $\mathcal{F}$ is superpolynomial in d.*

That the above combinatorial assumption on the class $\mathcal{F}$ is sensible, follows from a theorem of Goldberg and Jerrum [18]. Consider a parametrized class

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

for some $\{0,1\}$-valued function f. Suppose that, for each input $x \in \mathbb{R}^s$, 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 $>$, $\geqslant$, $<$, $\leqslant$, $=$, and $\neq$ comparisons of real numbers, and
- output 0 or 1.

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

Essentially, the above result states that a class of binary functions that can be computed in polynomial time taking a parameter value of polynomial length will have a polynomial VC dimension.

On the proof of Theorem 4.3. (For details, see [44].) Suppose the conclusion is false, and fix a particular poly(d) rate, $f(d)$, bounding from above the performance of a metric tree on any sample path. As the total content of bins B_t indexed with strings t of length exceeding the rate $f(d)$ has to be asymptotically negligible, we can assume without loss in generality that the indexing tree has depth $f(d)$.

Without loss in generality every bin can be replaced with an intersection of a family of sets of the form $\mathbf{1}_{f \leqslant t}$, $f \in \mathcal{F}$, and their complements. This provides a poly(d) upper bound on the VC dimension on the family of all possible bins.

With high confidence, a bin of a large measure will contain many data points, contradicting the poly(d) performance bound. This leads to conclude that measures of bins cannot be too skewed. Now concentration of measure is used to prove that at least 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 $\geqslant 1/2 - o(1)$. Processing a nearest neighbour query with such a centre q requires accessing all of these bins, let even to verify that some of them are empty. This leads to a contradiction with the assumed uniform performance bound on the algorithm. □

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 an interesting open problem.

Conjecture 4.4. *(Cf. [22].) Let X be a dataset with n points in the Hamming cube $\{0,1\}^d$. 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. [35,4]).

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 cells C_i, indexed with a set I,
- a dictionary $T = W^*$ over an alphabet $W = \{0,1\}^b$, viewed as a rooted prefix tree,
- a computable mapping $t \mapsto i(t)$ from T to I (*cell selector*), and
- a computable function $f = f_t(\sigma; q)$ (either partially or fully) defined on $T \times \{0,1\}^b \times \Omega$ and taking values in W.

For a $t \in T$, one can think of each f_t as a function defined on a subset of Ω and taking a b-bit string σ as a parameter, except if $t = *$ is the root. A value $f_t(\sigma; q)$ is a child s of the node t.

For every i, the cell C_i can hold a b-bit string. Sometimes b is regarded as constant, but often it is assumed that $b = \Theta(\lg n)$, so that a cell corresponding to a leaf node can store a pointer to a datapoint $x \in X$. Occasionally the nearest neighbour problem is replaced with a weaker *decision version* (known as *near neighbour problem*), 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, a leaf node cell 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 node cell a b-bit string.

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 prefix tree $T = W^*$ 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 $f_t(\sigma; q) = s \in W = \{0,1\}^b$ indicates a child of t to follow at the next step. When a leaf l is reached, the algorithm halts and returns the contents of $C_{i(l)}$. The query time is the length of the branch traversed, or equivalently the number of cell probes during the execution of the algorithm. The space requirement of the model is the total number of cells, $|I|$.

The cell probe model is very liberal, as the cost of computing the values of f is disregarded. For this reason, any lower bound obtained under the cell probe will likely hold under any other model of computation.

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 [40]. 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 [39] 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. Rather, they are what may be called "probably approximately 1-Lipschitz", and sometimes only on a certain distance scale. Such maps no longer exhibit a strong concentration around their means. The price to pay is that we may lose some relevant datapoints, as some distances are typically getting distorted, and so such maps cannot be used for exact NN search.

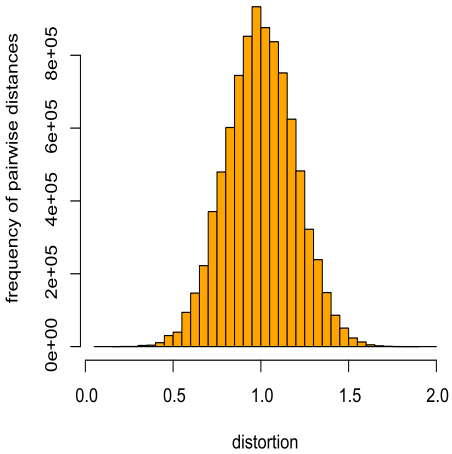

Fig. 9. Histogram of distortions of all pairwise distances in a random dataset of $n = 3000$ points in the $d = 500$ Hamming cube under a projection to a Hamming cube on randomly chosen $k = 25$ bits.

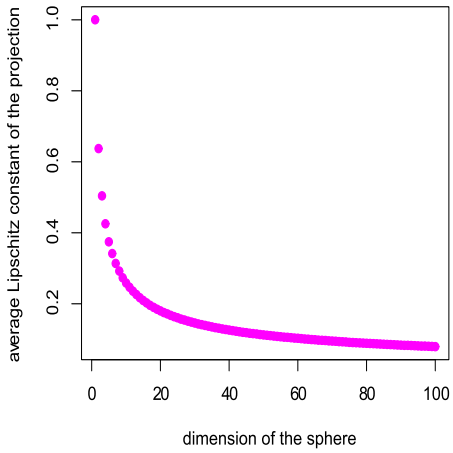

Fig. 10. The expected distortion of one-dimensional projection of the d-dimensional sphere $\mathbb{S}^{d-1}$ over all pairs of points.

5.1. Random projections in the Hamming cube

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. In other words, we normalize the Hamming distance $d(x, y) = \sharp\{i\colon\ x_i \neq y_i\}$ by multiplying it by $1/d$. Of course such a normalization has no effect on similarity search. 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 uniform Glivenko–Cantelli Theorem 3.1, 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 $\{0,1\}^{O(\varepsilon^{-2} \lg_2 n)}$ (under its own normalized Hamming distance) preserves the pairwise distances to within $\pm\varepsilon$. Cf. Fig. 9.

The error of $\pm\varepsilon$ is additive rather than multiplicative, so the random sampling of the coordinates is only appropriate for ANN search in the range on the order of $d/2$. The construction has to be generalized for all possible ranges $\ell = 1, 2, \ldots, d$. Such a generalization was developed in [28].

Projecting on a randomly sampled subset of k coordinates of the Hamming cube essentially amounts to a linear transformation $x \mapsto xA$, where A is a $d \times k$ matrix with i.i.d. Bernoulli entries assuming values 1 and 0 with probabilities $1/d$ and $1 - 1/d$, respectively. (The operations are carried mod 2.) One of the key observations of [28] — in the form given to it in [53], 7.2 — is that if the probability $1/d$ is replaced with $1/\ell$, then a random linear transformation $x \mapsto xA \bmod 2$, under a suitable normalization, preserves distances on the scale $\ell/2$, $\ell = 1, 2, \ldots, d$, to within an additive error $\ell\varepsilon$, and on a larger scale — away from it. Since the new cube only contains $2^{O(\varepsilon^{-2} \lg_2 n)}$ points, a hash table storing nearest neighbours, together with the reduction map f, produces an indexing scheme for ℓ-range search taking space polynomial in n and answering $(1+\varepsilon)$-approximate queries in time $O(\varepsilon^{-2} \lg_2 n)$.

Another discovery of [28] is that if on every scale ℓ one employs a sufficiently large series of independent projections onto k-cubes, then with high confidence one can assure that *every* ANN query — as opposed to *most* ANN queries — will be answered correctly. Finally, a separate indexing scheme is constructed for every range ℓ. The overall space requirement is still polynomial in n, and the running time of the algorithm is $O(d \operatorname{polylog}(dn))$.

5.2. Random projections in the Euclidean space

Let $\mathbb{S}^{d-1}$ denote the Euclidean sphere of unit radius 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 $|\pi_1(x) - \pi_1(y)| \leqslant \|x - y\|$, and for exactly one pair of antipodal 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 of distances $|\pi_1(x) - \pi_1(y)|/\|x - y\|$?

Fig. 10 shows that for a vast majority of pairs of points, the projection distorts distances by the factor $\Theta(1/\sqrt{d})$. A geometric explanation, at least at an intuitive level, is simple. Two randomly chosen points on the high-dimensional sphere, because of concentration of measure, are at a distance $\approx \sqrt{2}$ from each other. 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 (Fig. 11).

It follows that the expected absolute value of the norm of a projection of a given point x in a random direction is of the order $\Theta(1/\sqrt{d})$. Now let $X = \{x_1, \ldots, x_n\}$ be a finite subset of points of the sphere. Denote by Y the set of all vectors of the form $x_i - x_j$ whose length is normalized to one. Each $y \in Y$ can be identified with the function $\tilde{y}\colon z \mapsto |\langle y, z\rangle|$ on the unit sphere. If we now think of $\mathbb{S}^d$ as the domain (consisting of one-dimensional projections), then Y plays the role of a finite *function class*. Just like for finite concept classes, the combinatorial dimension of Y is of the order $O(\log n)$, and so, by VC

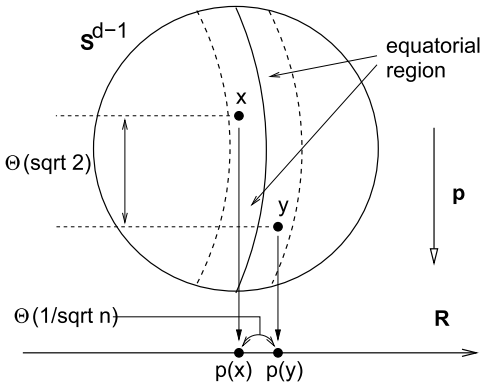

Fig. 11. To the geometry of random projections.

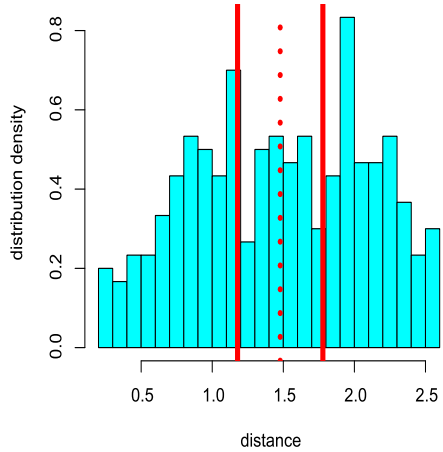

Fig. 12. Empirical density histogram of distances from a pivot having the highest found value of dissipation for the NASA dataset. Vertical lines mark the mean $\pm$ tolerance range $\varepsilon = 0.275$. The ϵ-dissipation (0.747) is the area outside of extreme lines.

theory, the empirical mean on a random sample of $\Theta(\varepsilon^{-1}\log n)$ directions will estimate the expectations of all $\tilde{y}$, $y \in Y$ to within a factor of ε with high confidence. A small number of randomly chosen directions are likely to be nearly pairwise orthogonal because of concentration, so we can instead choose an orthogonal projection to a randomly chosen space of dimension $\Theta(\varepsilon^{-1}\lg n)$. Since the projection is a linear map, we get the same estimate, but with a *multiplicative error* ε, for all pairwise distances between the points of X. It remains to work out the meaning of the empirical mean in the above setting in order to obtain the following famous result.

Theorem 5.1 *(Johnson–Lindenstrauss lemma). (See [24].) Let $\varepsilon \in (0, 1/2)$ be a real number, and $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n points in $\mathbb{R}^n$. Let k be an integer with $k \geqslant C\varepsilon^{-2}\log n$, where C is a sufficiently large absolute constant. Then there is a mapping $f : \mathbb{R}^n \to \mathbb{R}^k$ such that*

$$(1-\varepsilon)\left\|f(x_i) - f(x_j)\right\| \leqslant \|x_i - x_j\| \leqslant (1+\varepsilon)\left\|f(x_i) - f(x_j)\right\|$$

for all $i, j = 1, 2, \ldots, n$. Moreover, as f, one can with high confidence choose a suitably renormalized random projection from $\mathbb{R}^n$ to a k-dimensional Euclidean subspace.

An even simpler proof using concentration can be found in [30], Section 15.2, and an up-to-date survey of the lemma, in [31].

The normalized projection is not quite as good as a genuine 1-Lipschitz map, because the distortion of a distance can exceed one, and on rare occasions 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 a suitable indexing scheme in a lower-dimensional space $\mathbb{R}^k$, or rather a collection of such schemes, the random projection method leads to an efficient indexing scheme for a $(1+\varepsilon)$-approximate NN search (Indyk and Motwani [23]).

The articles [28] and [23] have appeared independently and at about the same time, and afterwards the dimensionality reduction methods have been shown [1] to be near optimal in the cell probe model.

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 [37]. 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. Fig. 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 [5,48] 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 Fig. 3.

Of a great variety of approaches to intrinsic dimension [12], at least two specifically measure the amount of concentration in data. The first one is the intrinsic dimension by Chávez et al. [8]

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

The second is the concentration dimension, studied within an axiomatic approach of [42,43]:

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

(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, such as the *local dimension* of Ollivier [38], Definition 3.

6.2. Black box search model and Urysohn space

The *black box model* of similarity search was studied by Krauthgamer and Lee [27]. Given a metric space (instance) (X, d), a query is a one-point metric space extension $X \cup \{q\}$, where the distances $d(q, x)$, $x \in X$, are accessible via the distance oracle. Each $d(q, x)$ can be evaluated in constant (unit) time. A preprocessing phase is allowed, under the condition that an indexing scheme occupies $\mathrm{poly}(n)$ space. The efficiency of an algorithm for (exact or approximate) similarity search is estimated as a number of calls to the distance oracle necessary to answer a query.

This is a "black-box model" in the sense that, formally speaking, there is no obvious domain (though we will see shortly that the domain is a well-defined separable metric case, and the setting is, in fact, classical). A remarkable feature of the model is that the problem of characterizing workloads admitting approximate NN queries in terms of an intrinsic dimension parameter receives a complete answer.

Recall that the *Assouad* (or *doubling*) *dimension* of a metric space (X, d) is the minimum value $\rho \geqslant 0$ such that every set A in X can be covered by 2^{ρ} balls of half the diameter of A. (The *diameter* of a set A is the supremum of $d(x, y)$, $x, y \in A$.) Denote this parameter by $\dim_{dbl}(X)$.

Theorem 6.1. *(See Krauthgamer and Lee [27].) A metric space (X, d) admits an algorithm requiring* $\mathrm{poly}(n)$ *space and taking* $\mathrm{polylog}(n)$ *time to answer a $(1+\varepsilon)$-approximate nearest neighbour query, where $\varepsilon < 2/5$, if and only if*

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

Here we will show that, on the contrary, an *exact* NN search in this context exhibits the curse of dimensionality even if the metric space (X, d) is contained in the unit interval $[0, 1]$ with the usual distance. With this purpose, we first convert the black box model into a conventional setting of searching in a metric domain.

The *universal Urysohn metric space*, $\mathbb{U}$, [19,32] is a complete separable metric space uniquely defined by the *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. See Fig. 13.

An equivalent definition is that if $X \subseteq \mathbb{U}$ is finite and $f : X \to \mathbb{R}$ satisfies

$$\left|f(x) - f(y)\right| \leqslant d_X(x, y) \leqslant f(x) + f(y) \tag{8}$$

for all $x, y \in X$, then there is $q \in \mathbb{U}$ with $f(x) = d(q, x)$ for all $x \in X$. (The functions satisfying (8) are called *Katětov functions*.)

This remarkable object has recently received plenty of attention in metric geometry. It is a *random*, or *generic*, metric space, in a sense that by equipping the integers with a randomly chosen metric ρ and taking a completion, one obtains $\mathbb{U}$ almost surely [54]. The space $\mathbb{U}$ contains an isometric copy of every separable metric space Ω. For this reason, one can use $\mathbb{U}$ as a "universal domain," and the black-box model can be restated as a classical similarity search problem in the domain $\Omega = \mathbb{U}$.

Theorem 6.2. *Let X be a finite metric space. Denote $n = |X|$. Then any deterministic algorithm for exact similarity search in X within the black box model will take the worst case time n.*

The result is true for simple information-theoretic reasons. We will produce for every $k < n$ a query q'' with a uniquely defined nearest neighbour in X which cannot be answered in time k.

Without loss in generality, we can assume that $\mathrm{diam}(X) = 1$. Let initially q be a query having the property that $d(q, x) = 1$ for all $x \in X$. Suppose that the algorithm has made $k < n$ calls to the distance oracle. Denote $x_1, x_2, \ldots, x_k \in X$ the points whose distance to q has been accessed. Since $d(q, x_i) = 1$ for all $i \leqslant k$, the algorithm clearly cannot halt at this stage. Let Q

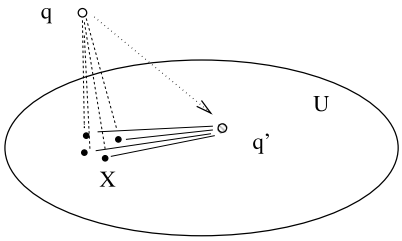

Fig. 13. One-point extension property.

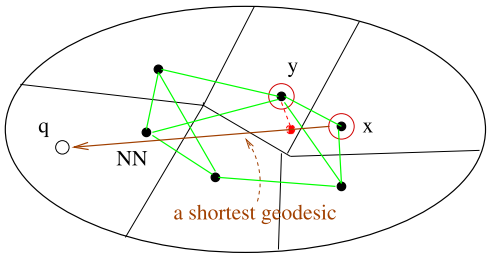

Fig. 14. NN search using Delaunay graph.

be the set of all $q' \in \mathbb{U}$ with $1 = d(q', x_i)$ for all $i = 1, 2, \ldots, k$. Since the algorithm is deterministic, we can replace q with any $q' \in Q$, and the sequence of executed calls to the oracle up until the step k will be the same.

Now denote $Y = \{x_1, x_2, \ldots, x_k\}$ and fix an $x_0 \in X \setminus Y$. The function

$$f(x) = \max\{1 - d(x, Y), d(x_0, Y) - d(x, x_0)\}$$

is Katětov, and thus it is the distance function from some q''. Clearly, $q'' \in Q$, and q'' admits a unique nearest neighbour in X, namely x_0. Thus, the search cannot be concluded in k steps even if it started with the well-defined query q''.

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

Here is an example of an indexing scheme for exact similarity search which is still "distance-based" but of a rather different type from either pivots or metric trees.

The *Voronoi cell* $V(x)$ of a datapoint $x \in X$ in a metric 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 $x, y \in \Omega$ can be joined by a shortest geodesic path, not necessarily unique. (All the domains previously considered in this article are such, including even the Urysohn space.) 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 Delaunay-adjacent to x is strictly closer to q than x is. (Proof: start moving along a shortest geodesic from x towards q, cf. Fig. 14, and use the triangle inequality.)

This observation turns the Delaunay graph of X in Ω 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 [11], was studied for general metric spaces by Navarro [36]. See also [46], 4.1.6.

In order for the algorithm to be efficient, the average vertex degree of the Delaunay graph has to be small. Navarro had observed ([36], Theorem 1) that this is not the case in general metric spaces. Specifically, he proved that for every two elements $a, b \in X$ there exists a finite metric space $Y = Y_{a,b}$ containing X as a subspace in which $a, b \in X$ are connected in the Delaunay graph of X. The result by Navarro translates immediately into:

Theorem 6.3. *Let X be a finite metric subspace of the universal Urysohn space $\mathbb{U}$. Then every two elements $a, b \in X$ are adjacent in the Delaunay graph of X in $\mathbb{U}$.*

In fact, the same remains true in less exotic situations, as one can deduce from Proposition 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 Section 2.2 the Delaunay graph of X is, with high confidence, a complete graph on n vertices.

Thus, the indexing scheme in question still 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 obtaining in a uniform way lower bounds for all possible "distance-based" indexing schemes (after they are formalized in a suitable way), not to mention an even more general setting of the cell probe model for all possible indexing schemes.

This having said, 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. [50].

In fact, it would be interesting to investigate the performance of the spatial approximation algorithm in hyperbolic metric spaces. Recall that a metric space X in which every two points x, y can be joined by a geodesic segment $[x, y]$ is *hyperbolic*

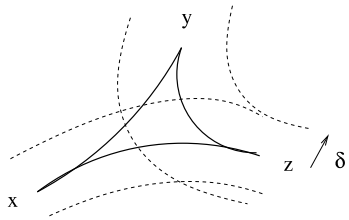

Fig. 15. A δ-thin geodesic triangle.

(in the sense of Rips) [51] if there exists a $\delta > 0$ so that every geodesic triangle is δ-*thin*: each side $[x, y]$ is contained in the δ-neighbourhood of the two other sides, $[x, z]$ and $[y, z]$ (Fig. 15).

Alain Connes has conjectured in [6], 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.

Appendix A. Proof of the empty space paradox (Theorem 2.1 and Proposition 2.2)

Without loss in generality, normalize the observable diameter of Ω to one. Let $\omega \in \Omega$. The distance function $\rho(\omega, -)$ is 1-Lipschitz and so concentrates around its median value, $R(\omega)$. The resulting function $R : \Omega \to \mathbb{R}$, $\omega \mapsto R(\omega)$ is also 1-Lipschitz, and concentrates around its median, R_M. It is easy to check that under our assumptions, the difference between the mean and the median of every 1-Lipschitz function f on Ω converges to zero as $O(\sqrt{d})$ (uniformly in f). Thus, without loss in generality, we can assume that, with high confidence, $R_M \to 1$ as $d \to \infty$. Notice that the above argument concerns the *domain* and not a particular *dataset*.

To prove Proposition 2.2, fix $\varepsilon > 0$ and sample an instance of data, X. With confidence $1 - n\exp(-O(d)\varepsilon^2)$, one has $|R(x) - 1| < \varepsilon/2$ for all $x \in X$. Moreover, since the datapoints are sampled in an i.i.d. fashion, by the union bound one has with confidence $1 - n^2\exp(-O(d)\varepsilon^2)$ that $|\rho(x, y) - R(x)| < \varepsilon/2$ for every pair $x, y \in X$. Since $n = |X|$ is subexponential in d, the statement follows.

To prove Theorem 2.1, again fix $\varepsilon > 0$. Denote ε_M the median value of the function ε_{NN}. Suppose $\liminf_{d\to\infty} \varepsilon_M < 1$. Proceed to a subsequence of domains and find $\gamma > 0$ with $\varepsilon_M \leqslant R_M - \gamma$ for all d. The probability that $R(\omega)$ deviates from R_M by more than $\gamma/2$ is exponentially small in d. Since $n = |X|$ only grows subexponentially in d, with confidence $1 - \exp(-O(\varepsilon^2 d))$ one has for every $x \in X$:

$$R(x) - \varepsilon_M \geqslant \frac{\gamma}{2}.$$

Now we use a technical observation from [20]: if $A \subseteq \Omega$ is such that $\mu(A) > \alpha_\Omega(\gamma)$ for some $\gamma > 0$, then $\mu(A_\gamma) > 1/2$. It follows that

$$\mu\big(B_{\varepsilon_M}(x)\big) \leqslant \alpha(\gamma/2) = \exp\big(-O\big(\varepsilon^2 d\big)\big),$$

and therefore

$$\mu(X_{\varepsilon_M}) \leqslant n\exp\big(-O\big(\varepsilon^2 d\big)\big) = \exp\big(-O\big(\varepsilon^2 d\big)\big),$$

which contradicts the definition of ε_M. This implies: $\liminf_{d\to\infty} \varepsilon_M \geqslant 1$.

To establish the converse inequality $\limsup_{d\to\infty} \varepsilon_M \leqslant 1$, recall that a ball of radius $R(\omega)$ centred at ω has measure $\geqslant 1/2$, and so we have an obvious estimate $\varepsilon_M \leqslant \min_{x\in X} R(x)$. The rest follows from concentration of the function R around one.

Appendix B. Distortion of Lipschitz embeddings $\ell^2(d) \hookrightarrow \ell^\infty(d+k)$

Lemma B.1. *Fix k. Let $c > 0$ be a constant having the property that for every d and each bounded subset X of $\ell^2(d)$ there exists a 1-Lipschitz function $f : X \to \ell^\infty(d+k)$ having distortion c: for all $x, y \in X$,*

$$\big\|f(x) - f(y)\big\|_\infty \leqslant \|x - y\|_2 \leqslant c\big\|f(x) - f(y)\big\|_\infty. \tag{B.1}$$

Then $c = \Omega(\sqrt{d+k}/\sqrt{k})$, that is, $\Omega(\sqrt{d})$ with a constant depending on k.

The proof consists of a series of statements.

1. *There exists a 1-Lipschitz function $f : \ell^2(d) \to \ell^\infty(d+k)$ with the property (B.1).*

For every $n \in \mathbb{N}$, choose a function f_n from the closed n-ball $B_n(0)$ in $\ell^2(d)$ to the n-ball in $\ell^\infty(d+k)$ with distortion c. The Banach space ultrapowers of both participating spaces formed with regard to a non-principal ultrafilter on the

integers (see e.g. p. 55 in [25]) are isometric, respectively, to $\ell^2(d)$ and $\ell^\infty(d+k)$, because the spaces in question are finite-dimensional. The family of 1-Lipschitz functions (f_n) determines in a standard way a 1-Lipschitz function, f, from $\ell^2(d)$ to $\ell^\infty(d+2)$, with the property (B.1) being preserved.

2. *There exists a linear function f with the property* (B.1).

Choose f as in 1. According to the Rademacher theorem (cf. a discussion and references on p. 42 in [25]), f is differentiable almost everywhere with regard to the Lebesgue measure. The differential of f at any point, which we denote T, is a linear operator of norm one having property (B.1). In particular it is injective (though of course not onto), and the inverse has norm $\leqslant c$.

Recall that the *multiplicative Banach–Mazur distance* between two normed spaces E and F of the same dimension is the infimum of all numbers $\|T\| \cdot \|T\|^{-1}$, where T ranges over all isomorphisms between E and F. (See [25], p. 3, and [17], 7.2.) From the previous observation, we conclude:

3. *The Banach–Mazur distance between $\ell^2(d)$ and some d-dimensional subspace of $\ell^\infty(d+k)$ is $\leqslant c$.*

4. *The Banach–Mazur distance between $\ell^2(d+k)$ and $\ell^\infty(d+k)$ is $O(\sqrt{k}c)$.*

There is a projection p from $\ell^\infty(d+k)$ having $T(\ell^2(d))$ as its kernel and such that $\|p\| \leqslant \sqrt{k}$ and $\|1-p\| \leqslant \sqrt{k}$ (combine [13], Corollary on p. 209, with a classical result of Kadec and Snobar on projection constants, cf. [25], p. 71). The Banach–Mazur distance between $\ell^2(k)$ and any other k-dimensional normed space, including the kernel of p, is $O(\sqrt{k})$. Choose an isomorphism S realizing this distance, then it is easy to verify that $T \oplus S$ realizes the distance $O(\sqrt{k}c)$ between $\ell^2(d+k)$ and $\ell^\infty(d+k)$.

Finally, the Banach–Mazur distance between $\ell^2(d+k)$ and $\ell^\infty(d+k)$ is $\sqrt{d+k}$ (cf. [17], p. 766).

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, 2006, pp. 449–458.
[2] M. Anthony, P. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, Cambridge, 1999.
[3] O. Barkol, 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] A. Borodin, R. Ostrovsky, Y. Rabani, Lower bounds for high-dimensional nearest neighbor search and related problems, in: Proc. 31st Annual ACS Sympos. Theory Comput., 1999, pp. 312–321.
[5] B. Bustos, G. Navarro, E. Chávez, Pivot selection techniques for proximity searching in metric spaces, Pattern Recogn. Lett. 24 (2003) 2357–2366.
[6] J.-P. Changeux, A. Connes, Conversations on Mind, Matter, and Mathematics, Princeton University Press, Dec. 1998.
[7] E. Chávez, J.L. Marroquín, R.A. Baeza-Yates, Spaghettis: An array based algorithm for similarity queries in metric spaces, in: Proceedings SPIRE/CRIWG, 1999, pp. 38–46.
[8] E. Chávez, G. Navarro, R. Baeza-Yates, J.L. Marroquín, Searching in metric spaces, ACM Comput. Surv. 33 (2001) 273–321.
[9] P. Ciaccia, M. Patella, 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, 1997, pp. 426–435.
[10] P. Ciaccia, M. Patella, P. Zezula, A cost model for similarity queries in metric spaces, in: Proc. 17th ACM Symposium on Principles of Database Systems (PODS'98), Seattle, WA, 1998, pp. 59–68.
[11] K.L. Clarkson, An algorithm for approximate closest-point queries, in: Proc. 10th Symp.. Comp. Geom. Stony Brook, NY, 1994, pp. 160–164.
[12] 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.
[13] W.J. Davis, Remarks on finite rank projections, J. Approx. Theory 9 (1973) 205–211.
[14] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996.
[15] A. Faragó, T. Linder, G. Lugosi, Fast nearest neighbor search in dissimilarity spaces, IEEE Trans. Pattern Anal. Machine Intell. 18 (1993) 957–962.
[16] P. Frankls, Z. Füredi, A short proof for a theorem of Harper about Hamming-spheres, Discrete Math. 34 (1981) 311–313.
[17] A.A. Giannopoulos, V.D. Milman, Euclidean structure in finite dimensional normed spaces, in: Handbook of the Geometry of Banach Spaces, vol. 1, North-Holland, Amsterdam, 2001, pp. 707–779.
[18] P.W. Goldberg, M.R. Jerrum, Bounding the Vapnik–Chervonenkis dimension of concept classes parametrized by real numbers, Machine Learning 18 (1995) 131–148.
[19] M. Gromov, Metric Structures for Riemannian and Non-Riemannian Spaces, Progress in Mathematics, vol. 152, Birkhauser Verlag, 1999.
[20] M. Gromov, V.D. Milman, A topological application of the isoperimetric inequality, Amer. J. Math. 105 (1983) 843–854.
[21] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 427–444.
[22] 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., 2004, pp. 877–892.
[23] P. Indyk, R. Motwani, Approximate nearest neighbours: towards removing the curse of dimensionality, in: Proc. 30th ACM Symp. Theory of Computing, 1998, pp. 604–613.
[24] W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, Contemp. Math. 26 (1984) 189–206.
[25] W.B. Johnson, J. Lindenstrauss, Basic concepts in the geometry of Banach spaces, in: Handbook of the Geometry of Banach Spaces, vol. 1, North-Holland, Amsterdam, 2001, pp. 1–84.
[26] B. Klartag, A central limit theorem for convex sets, Invent. Math. 168 (2007) 91–131.
[27] R. Krauthgamer, J.R. Lee, The black-box complexity of nearest-neighbor search, Theoret. Comput. Sci. 348 (2) (December 2005) 262–276.
[28] E. Kushilevitz, R. Ostrovsky, Y. Rabani, Efficient search for approximate nearest neighbor in high-dimensional spaces, SIAM J. Comput. 30 (2000) 457–474.
[29] M. Ledoux, The Concentration of Measure Phenomenon, Mathematical Surveys and Monographs, vol. 89, Amer. Math. Soc., 2001.
[30] J. Matoušek, Lectures on Discrete Geometry, Springer, New York, 2002.
[31] J. Matoušek, On variants of the Johnson–Lindenstrauss lemma, Random Structures Algorithms 33 (2008) 142–156.
[32] J. Melleray, On the geometry of Urysohn's universal metric space, Topology Appl. 154 (2007) 384–403.
[33] V. Milman, Topics in asymptotic geometric analysis, Geometric and Functional Analysis (special volume GAFA2000) (2000) 792–815.

[34] V.D. Milman, G. Schechtman, Asymptotic Theory of Finite-Dimensional Normed Spaces (with an Appendix by M. Gromov), Lecture Notes in Math., vol. 1200, Springer, 1986.
[35] P.B. Miltersen, Cell probe complexity — a survey, in: Advances in Data Structures Workshop, 19th Conf. on the Foundations of Software Technology and Theoretical Computer Science (FSTTCS), 1999.
[36] G. Navarro, Searching in metric spaces by spatial approximation, VLDB J. 11 (2002) 28–46.
[37] 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, pp. 3–10.
[38] Yann Ollivier, Ricci curvature of metric spaces, C. R. Math. Acad. Sci. Paris 345 (2007) 643–646.
[39] M. Patella, 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, 2008, pp. 10–21.
[40] M. Pătrascu, M. Thorup, Higher lower bounds for near-neighbor and further rich problems, in: Proc. 47th IEEE Symp. on Foundations of Computer Science, 2006, pp. 646–654.
[41] V. Pestov, On the geometry of similarity search: dimensionality curse and concentration of measure, Inform. Process. Lett. 73 (2000) 47–51.
[42] V. Pestov, Intrinsic dimension of a dataset: what properties does one expect? in: Proc. of the 22nd Int. Joint Conf. on Neural Networks (IJCNN'07), Orlando, FL., 2007, pp. 1775–1780.
[43] V. Pestov, An axiomatic approach to intrinsic dimension of a dataset, Neural Networks 21 (2008) 204–213.
[44] V. Pestov, Lower bounds on performance of metric tree indexing schemes for exact similarity search in high dimensions, in: Alfredo Ferro (Ed.), Proceedings of the 4th International Conference on Similarity Search and Applications (SISAP 2011), 30 June–1 July 2011, Lipari, Sicily, Italy, ACM, New York, NY, 2011, pp. 25–32.
[45] V. Pestov, A. Stojmirović, Indexing schemes for similarity search: an illustrated paradigm, Fund. Inform. 70 (2006) 367–385.
[46] H. Samet, Foundations of Multidimensional and Metric Data Structures, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2005.
[47] U. Shaft, R. Ramakrishnan, Theory of nearest neighbors indexability, ACM Trans. Database Systems (TODS) 31 (2006) 814–838.
[48] SISAP metric space library, http://sisap.org/Metric_Space_Library.html.
[49] A. Stojmirović, V. Pestov, Indexing schemes for similarity search in datasets of short protein fragments, Inform. Systems 32 (2007) 1145–1165.
[50] K. Taşdemir, E. Merényi, Exploiting the data topology in visualizing and clustering of self-organizing maps, IEEE Trans. Neural Networks 20 (4) (2009) 549–562.
[51] J. Väisälä, Gromov hyperbolic spaces, Expo. Math. 23 (2005) 187–231.
[52] V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, Inc., New York, 1998.
[53] S.S. Vempala, The Random Projection Method, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 65, Amer. Math. Soc., Providence, RI, 2004.
[54] A.M. Vershik, Universality and randomness for the graphs and metric spaces, in: Frontiers in Number Theory, Physics, and Geometry. I, Springer, Berlin, 2006, pp. 245–266.
[55] E. Vidal, An algorithm for finding nearest neighbors in (approximately) constant average time, Pattern Recogn. Lett. 4 (1986) 145–157.
[56] M. Vidyasagar, Learning and Generalization, with Applications to Neural Networks, second ed., Springer-Verlag, London, 2003.
[57] I. Volnyansky, 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.
[58] P. Zezula, G. Amato, V. Dohnal, M. Batko, Similarity Search. The Metric Space Approach, Springer Science + Business Media, New York, 2006.