Noname manuscript No.
(will be inserted by the editor)

Lower Bounds on Performance of Metric Tree Indexing Schemes for Exact Similarity Search in High Dimensions

Vladimir Pestov

Department of Mathematics and Statistics, University of Ottawa,
585 King Edward Avenue, Ottawa, Ontario K1N 6N5 Canada
e-mail: vpest283@uottawa.ca

The date of receipt and acceptance will be inserted by the editor

Abstract Within a mathematically rigorous model borrowed from statistical learning theory, we analyse the curse of dimensionality for similarity-based information retrieval in the context of a wide class of popular indexing schemes. The datasets X are sampled randomly from a domain Ω, equipped with a distance, ρ, and an underlying probability distribution, μ. The intrinsic dimension of the domain, d, is defined in terms of the concentration of measure phenomenon. For the purposes of asymptotic analysis, we send d to infinity, and assume that the size of a dataset, n, grows faster than any polynomial function in d, yet slower than any exponential function in d. Exact similarity search refers to finding the nearest neighbour in the dataset X to a query point $\omega \in \Omega$, where the query points are subject to the same probability distribution μ as datapoints. Let $\mathscr{F}$ denote a class of all 1-Lipschitz functions on Ω that can be used as decision functions in constructing a hierarchical metric tree indexing scheme. Suppose the VC dimension of the class of subsets defined by inequalities $f \gtreqless a$, $f \in \mathscr{F}$, $a \in \mathbb{R}$ is $d^{O(1)}$. (According to a result of Goldberg and Jerrum, at least for $\Omega = \mathbb{R}^d$ this is a not a serious restriction.) Under those assumptions, we obtain lower bounds on the expected average case performance of hierarchical metric-tree based indexing schemes for exact similarity search in (Ω, X), which bounds are superpolynomial in d.

Introduction

The curse of dimensionality is a well-known phenomenon across the entire computer science, negatively affecting, in particular, the performance of indexing schemes into large datasets for the purpose of similarity-based information retrieval, cf. e.g. Chapter 9 in [21], as well as [3,28].

Paradoxically, there is still no mathematical proof that the above phenomenon is really in the nature of high-dimensional datasets. While the concept of intrinsic dimension of a dataset is open to a discussion (see [18] and references therein), even in cases commonly accepted as "high-dimensional" (e.g. uniformly distributed data in the Hamming cube $\{0,1\}^d$ as $d \to \infty$), the "curse of dimensionality conjecture" for proximity search remains unproven [11]. Diverse results in this direction [4,2,16,5,22], are still preliminary.

Here we will verify the conjecture for a particular class of indexing schemes widely used in similarity search and going back to [24]: metric trees. So are called hierarchical partitioning indexing schemes equipped with 1-Lipschitz (non-expanding) decision functions at every node.

We assume that datapoints are drawn from the domain Ω with regard to an underlying probability measure μ independently of each other. The domain is a metric space, that is, the similarity measure, ρ, satisfies the axioms of a metric. The intrinsic dimension of Ω is defined in terms of concentration of measure as in [18]. This concept agrees with the usual notion of dimension in cases such as the Hamming cube $\{0,1\}^d$ or the Euclidean ball $\mathbb{B}^d$, and is most relevant. A dataset $X \subseteq \Omega$ with n points is modelled by i.i.d. random variables distributed according to μ. We assume, as in [11], that the number of datapoints n grows superpolynomially in dimension d yet subexponentially in d. Using the notation of asymptotic algorithm analysis, this can be written as $n = d^{\omega(1)}$ and $d = \omega(\log n)$.

It is clear that the computational complexity of decision functions used in constructing a metric tree is a major factor in a scheme performance. We take this into account in the form of a combinatorial restriction on the subclass $\mathscr{F}$ of all functions on Ω that are allowed to be used as decision functions, by requiring a well-known parameter of statistical learning theory, the Vapnik-Chervonenkis dimension of $\mathscr{F}$ [25], to be polynomial in d, that is, $\text{VC-dim}\,(\mathscr{F}) = d^{O(1)}$.

A very general class of functions satisfying this VC dimension bound is provided by a theorem of Goldberg and Jerrum [9] about function classes parametrized by elements of $\mathbb{R}^s$ whose computation involves arithmetic operations, conditioning on inequalities, and inputs 0 or 1. Apparently, the decision functions of all indexing schemes used in practice so far in Euclidean (and Hamming cube) domains fall into this class.

Under above assumptions, we prove a superpolynomial in d lower bound on the expected average performance of all possible metric trees. We believe that a lower bound that strong has never been derived before within a mathematically rigorous model and in the present generality.

1 General framework for similarity search

We follow a formalism of [10] as adapted for similarity search [16,19]. A *workload* is a triple $W = (\Omega, X, \mathcal{Q})$, where Ω is the *domain,* whose elements can occur as datapoints and as query points, $X \subseteq \Omega$ is a finite subset (*dataset*, or *instance*), and $\mathcal{Q} \subseteq 2^{\Omega}$ is a family of *queries. Answering a query* $Q \in \mathcal{Q}$ means listing all datapoints $x \in X \cap Q$.

A (*dis*)*similarity measure* on Ω is a function of two arguments $\rho\colon \Omega \times \Omega \to \mathbb{R}$, which we assume to be a metric, as in [30]. A *range similarity query centred at* $\omega \in \Omega$ is a ball of radius ε around the query point:

$$Q = \mathcal{B}_{\varepsilon}(\omega) = \{x \in \Omega\colon \rho(\omega, x) < \varepsilon\}.$$

Equipped with such balls as queries, the triple $W = (\Omega, \rho, X)$ forms a *range similarity workload.*

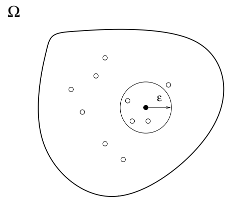

Fig. 1 A range query.

We will assume ρ to be a metric, as in [30], though sometimes one needs to consider more general similarity measures, cf. [8,19].

The *k-nearest neighbours* (k-NN) *query* centred at $\omega \in \Omega$, where $k \in \mathbb{N}$, is normally being reduced to a range query of a suitable search radius.

A workload is *inner* if $X = \Omega$ and *outer* if $|X| \ll |\Omega|$. There is an essential difference between the two types of workloads, and most workloads of practical interest are outer workloads, that is, a typical query point will come from outside the dataset, cf. [19].

2 Hierarchical tree index structures

An *access method* is an algorithm that correctly answers every range query. Principal examples of access methods are *indexing schemes.* A hierarchical tree-based indexing scheme includes a sequence of refining partitions of the domain labelled with a finite rooted tree. For simplicity, we will assume all trees to be binary. This assumption is not really restrictive.

Such a structure will occupy a storage space $O(n)$.

To process a range query $\mathcal{B}_{\varepsilon}(\omega)$, we traverse the tree recursively to the leaf level. Once a leaf B is reached, its contents (i.e., all datapoints $x \in X \cap B$) are accessed, and the condition $x \in \mathcal{B}_{\varepsilon}(\omega)$ verified for each one of them.

Of main interest is what happens at each internal node C. Let us identify C with the corresponding element $C \subseteq \Omega$ of the partition, and suppose that A and B are child nodes of C, so that $C = A \cup B$. A branch descending

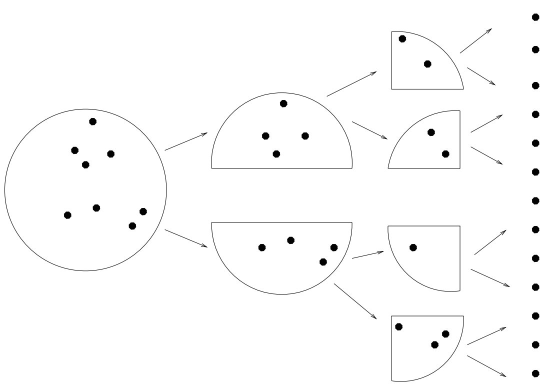

Fig. 2 A refining sequence of partitions of Ω.

from B can be pruned provided $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, because then datapoints contained in B are of no further interest. Equivalently, this is the case where it can be certified that ω is not contained in the ε-neighbourhood of B,

$$\omega \notin B_\varepsilon = \{x \in \Omega \colon d(x,B) < \varepsilon\}.$$

(Cf. Fig. 3, l.h.s.) Similarly, if $\omega \notin A_\varepsilon$, then the sub-tree descending from A can be pruned. However, if the open ball $\mathcal{B}_\varepsilon(\omega)$ meets both A and B or, equivalently, ω belongs to the intersection of ε-neighbourhoods of A and B, pruning is impossible and the search branches out. (Cf. Fig. 3, r.h.s.)

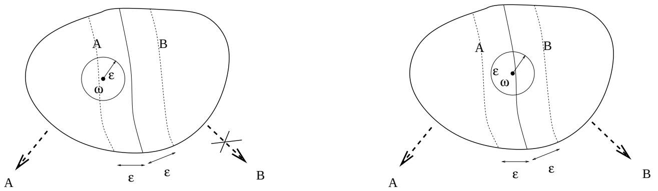

Fig. 3 Pruning is possible (l.h.s.), and impossible (r.h.s.).

In order to "certify" that $\mathcal{B}_\epsilon(\omega) \cap B = \emptyset$, one employs the technique of *decision functions*. Recall that a function $f \colon \Omega \to \mathbb{R}$ is a *1-Lipschitz function* if

$$\forall x,y \in \Omega, \quad |f(x) - f(y)| \leq d(x,y).$$

Assign to every internal mode C a 1-Lipschitz function $f = f_C$ so that $f_C \upharpoonright B \leq 0$ and $f_C \upharpoonright A \geq 0$. It is easily seen that $f_C \upharpoonright B_\varepsilon < \varepsilon$, and so the fact that $\boxed{f_C(\omega) \geq \varepsilon}$ serves as a certificate for $\boxed{\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset}$, assuring that a sub-tree descending from B can be pruned. Similarly, if $f_C(\omega) \leq -\varepsilon$, the sub-tree descending from A can be pruned.

Note that decision functions should have sufficiently low computational complexity in order for the indexing scheme to be efficient.

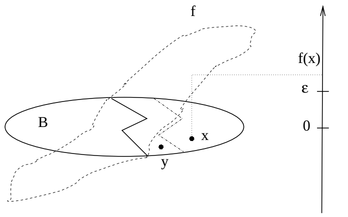

Fig. 4 Graph of a decision function $f = f_C$.

A hierarchical indexing structure employing 1-Lipschitz decision functions at every node is known as a *metric tree.*

3 Metric trees

Here is a formal definition. A metric tree for a metric similarity workload (Ω, ρ, X) consists of

- a finite binary rooted tree $\mathcal{T}$,
- a collection of (possibly partially defined) real-valued 1-Lipschitz functions $f_t \colon B_t \to \mathbb{R}$ for every inner node t (decision functions), where $B_t \subseteq \Omega$,
- a collection of *bins* $B_t \subseteq \Omega$ for every leaf node t, containing pointers to elements $X \cap B_t$,

so that

- $B_{\mathrm{root}(\mathcal{T})} = \Omega$,
- for every internal node t and child nodes t_-, t_+, one has $B_t \subseteq B_{t_-} \cup B_{t_+}$,
- $f_t \restriction B_{t_-} \leq 0$, $f_t \restriction B_{t_+} \geq 0$.

When processing a range query $\mathcal{B}_\epsilon(\omega)$,

- t_- is accessed $\iff$ $f_t(\omega) < \varepsilon$, and
- t_+ is accessed $\iff$ $f_t(\omega) > -\varepsilon$.

Here is the search algorithm in pseudocode.

Algorithm 1

```
on input $(\omega, \varepsilon)$ do
    set $A_0 = \{\mathrm{root}(\mathcal{T})\}$
    for each $i = 0, 1, \ldots, \mathrm{depth}(\mathcal{T}) - 1$ do
        if $A_i \neq \emptyset$
        then for each $t \in A_i$ do
            if $t$ is an internal node
            then do
                if $f_t(\omega) < \varepsilon$
                then $A_{i+1} \leftarrow A_{i+1} \cup \{t_-\}$
                if $f_t(\omega) > -\varepsilon$
            then $A_{i+1} \leftarrow A_{i+1} \cup \{t_+\}$
        else for each $x \in B_t$ do
            if $x \in \mathcal{B}_\varepsilon(\omega)$
            then $A \leftarrow A \cup \{x\}$
return $A$
```

□

Under our assumptions on the metric tree, Algorithm 1 correctly answers every range similarity query for the workload (Ω, ρ, X) and thus is an access method.

For more, see [19], while the survey [5] presents a different perspective. Each of the books [20,21,30] is an excellent reference to indexing structures in metric spaces.

4 Curse of dimensionality

Every similarity query can be answered in time $O(n)$ through a simple linear scan of the dataset X. In practice, a linear scan often outperforms the best known indexing schemes for high-dimensional workloads, though of course there are exceptions, cf. e.g. a relatively efficient scheme developed in [23] for searching large databases of short protein fragments.

As a consequence, the research emphasis in recent years has shifted towards *approximate* similarity search:

- given $\epsilon > 0$ and $\omega \in \Omega$, return a point that is [with high probability] at a distance $< (1+\epsilon)d_{NN}(\omega)$ from ω.

This has led to many spectacular achievements, based on deep results of geometric functional analysis (see e.g. survey [11] and Chapter 7 in [26]). At the same time, research in exact similarity search, especially concerning deterministic algorithms, has slowed down. One of the stumbling blocks is the inability to prove at a mathematically rigorous level that the curse of dimensionality is indeed in the nature of high-dimensional datasets. The following problem remains open.

Conjecture 1 (The curse of dimensionality conjecture, cf. [11]) Let $X \subseteq \{0,1\}^d$ be a dataset with n points, where the Hamming cube $\{0,1\}^d$ is equipped with the Hamming (ℓ^1) distance:

$$d(x,y) = \sharp\{i \colon x_i \neq y_i\}.$$

Suppose $d = n^{o(1)}$, but $d = \omega(\log n)$. (That is, the number of points in X has intermediate growth with regard to the dimension d: it is superpolynomial in d, yet subexponential.) Then any data structure for exact nearest neighbour search in X, with $d^{O(1)}$ query time, must use $n^{\omega(1)}$ space.

Ideally, the conjecture should be proved within the *cell probe model* [15], which is a very general model of computation. The best lower bounds within this model currently known are on the order of $\Omega(d/\log n)$ [2].

5 Concentration of measure

As in [7], we assume the existence of an unknown probability measure μ on Ω, such that both datapoints X and query points ω are being sampled with regard to μ.

On the one hand, this assumption is open to debate: for instance, in a typical university library most books (75 % or more) are never borrowed a single time, so it is reasonable to assume that the distribution of queries in a large dataset will be skewed equally heavily away from data distribution. On the other hand, there is no obvious alternative way of making an apriori assumption about the query distribution, and in some situations the assumption makes sense indeed, e.g. in the context of a large biological database where a newly-discovered protein fragment has to be matched against every previously known sequence.

The triple (Ω, ρ, μ) is known in a mathematical context as a *metric space with measure*. This concept opens the way to systematically using the *phenomenon of concentration of measure on high-dimensional structures*, also known as the *“Geometric Law of Large Numbers.”* This phenomenon arguably plays an important part in explaining away the course of dimensionality and can be informally summarized as follows:

> for a typical “high-dimensional” structure Ω, if A is a subset containing at least half of all points, then the measure of the ε-neighbourhood A_ε of A is overwhelmingly close to 1 already for small $\varepsilon > 0$.

Here is a rigorous way for dealing with the phenomenon. Define the *concentration function* α_Ω of a metric space with measure Ω by

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

The value of $\alpha_\Omega(\varepsilon)$ gives un 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. 5.

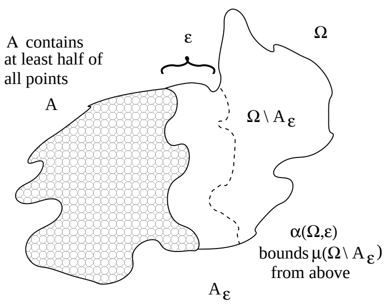

Fig. 5 To the concept of concentration function $\alpha_\Omega(\epsilon)$.

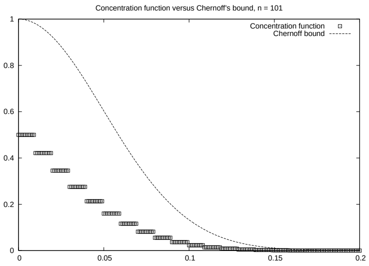

Fig. 6 Concentration function of $\{0,1\}^{101}$ *vs* gaussian bound.

For example, let $\Omega = \{0,1\}^d$ be the Hamming cube equipped with the normalized Hamming distance

$$d(x,y) = \frac{1}{d}\sharp\{i\colon x_i \neq y_i\}$$

and the uniform (normalized counting) measure

$$\mu_\sharp(A) = \frac{\sharp A}{2^d}.$$

Then the concentration function of Ω satisfies a gaussian upper estimate (Chernoff bound):

$$\alpha_{\{0,1\}^n}(\varepsilon) \leq e^{-2\varepsilon^2 n}.$$

For an example in dimension $d = 101$, see Fig. 6.

Similar bounds hold for Euclidean spheres $\mathbb{S}^n$, cubes $\mathbb{I}^n$, and many other structures of both continuous and discrete mathematics, equipped with suitably normalized distances and canonical probability measures. The concentration phenomenon can be expressed by saying that for "typical" high-dimensional metric spaces with measure, Ω, the concentration function $\alpha_\Omega(\varepsilon)$ drops off sharply as $\dim \Omega \to \infty$ [14,12].

6 Workload assumptions

We are ready to make standing assumptions on the workload for the rest of the article.

Let (Ω, ρ, μ) be a domain equipped with a metric ρ and a probability measure μ. We assume that the expected distance between two points of Ω is normalized so as to become asymptotically constant:

$$\mathbb{E}\,\rho(x,y) = \Theta(1). \tag{1}$$

We further assume that Ω has "concentration dimension d" in the sense that the concentration function α_Ω is gaussian with exponent $\Theta(d)$;

$$\alpha_\Omega(\varepsilon) = \exp\left(-\Theta(\varepsilon^2 d)\right). \tag{2}$$

(This approach to intrinsic dimension is developed in [17,18].)

A dataset $X \subseteq \Omega$ contains n points, where the rate of growth of n and d is as follows:

$$n = d^{\omega(1)}, \tag{3}$$

$$d = \omega(\log n). \tag{4}$$

In other words, the rate of growth of n as $d \to \infty$ is faster than any polynomial function Cd^k, $C > 0$, $k \in \mathbb{N}$, but slower than any exponential function e^{cd}, $c > 0$. (An example of this rate of growth is the function $n = 2^{\sqrt{d}}$.) Such assumptions are natural for the purposes of asymptotic analysis of search algorithms, cf. the survey paper [11].

Datapoints are modelled by a sequence of i.i.d. random variables distributed according to the measure μ:

$$X_1, X_2, \ldots, X_n \sim \mu.$$

The instances of datapoints will be denoted with corresponding lower case letters $x_1, x_2, \ldots, x_n$.

Finally, the query centres $\omega \in \Omega$ have the same distribution μ:

$$\omega \sim \mu.$$

7 Query radius and branching

As a well-known concequence of concentration, in high-dimensional domains the distance to the nearest neighbour is close to the average distance between two points (cf. e.g. [3] for a particular case). Denote $\varepsilon_{NN}(\omega)$ the distance from $\omega \in \Omega$ to the nearest point in X. The function ε_{NN} is 1-Lipschitz, and so it concentrates near its median value. From here, one deduces in a standard way:

Lemma 1 *Under our assumptions on the domain Ω and a random sample X, with confidence approaching* 1 *one has for all δ*

$$\mu\left\{\omega\colon |\varepsilon_{NN}(\omega) - \mathbb{E}\,\rho(x,y)| > \delta\right\} < \exp(-O(\delta^2 d)).$$

□

What happens at an internal node C when a metric tree is being traversed? Let α_C denote the concentration function of C equipped with the metric induced from Ω and a probability measure μ_C which is the normalized restriction of the measure μ from Ω:

$$\text{for } A \subseteq C, \quad \mu_C(A) = \frac{\mu(A)}{\mu(C)}.$$

Suppose for the moment that our tree is perfectly balanced, in the sense that $\mu_C(A) = \mu_C(B) = \frac{1}{2}$. Then the size of the ε-neighbourhood of A is at least $1 - \alpha_C(\varepsilon)$, and the same is true of the ε-neighbourhood of B. One concludes: for all query points $\omega \in C$ except a set of measure $\leq 2\alpha_C(\varepsilon)$, the search algorithm 1 branches out at the node C. (Cf. Fig. 7.)

Lemma 2 *Let C be a subset of a metric space with measure (Ω, ρ, μ). Denote α_C the concentration function of C with regard to the induced metric $\rho \upharpoonright C$ and the induced probability measure $\mu/\mu(C)$. Then for all $\varepsilon > 0$*

$$\alpha_C(\varepsilon) \leq \frac{\alpha_\Omega(\varepsilon/2)}{\mu(C)}.$$

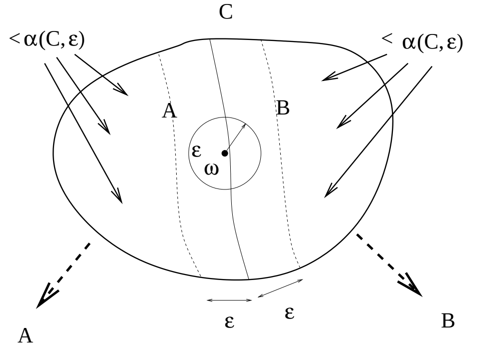

Fig. 7 Search algorithm branches out for most query points ω at a node C if the value $\alpha_C(\varepsilon)$ is small.

Proof Let $\varepsilon > 0$ be any, and let $\delta < \alpha_C(\varepsilon)$. Then there are subsets $D, E \subseteq C$ at a distance $\geq \varepsilon$ from each other, satisfying $\mu(D) \geq \mu(C)/2$ and $\mu(E) \geq \delta\mu(C)$, in particular the measure of either set is at least $\delta\mu(C)$. Since the $\varepsilon/2$-neighbourhoods of D and E in Ω cannot meet by the triangle inequality, the complement, F, to at least one of them, taken in Ω, has the property $\mu(F) \geq 1/2$, while $\mu(F_{\varepsilon/2}) \leq 1 - \delta\mu(C)$, because $F_{\varepsilon/2}$ does not meet one of the two original sets, D or E. We conclude: $\alpha_\Omega(\varepsilon/2) \geq \delta\mu(C)$, and taking suprema over all $\delta < \alpha_C(\varepsilon)$,

$$\alpha_\Omega(\varepsilon/2) \geq \alpha_C(\varepsilon)\mu(C),$$

that is, $\alpha_C(\varepsilon) \leq \alpha_\Omega(\varepsilon/2)/\mu(C)$, as required. □

Since the size of the indexing scheme is $O(n)$, a typical size of a set C will be on the order $\Omega\left(n^{-1}\right)$, while $\alpha_\Omega(\varepsilon)$ will go to zero as $o\left(n^{-1}\right)$.

8 A "naive" average $O(n)$ lower bound

As a first approximation to our analysis, we present a (flawed) heuristic argument, allowing *linear* in n asymptotic lower bounds on the search performance of a metric tree. As we will see, in order to become a rigorous proof, it still lacks an important component.

Let a workload (Ω, ρ, X) be indexed with a balanced metric tree of depth $O(\log n)$, having $O(n)$ bins of roughly equal size in the sense of the probability measure μ underlying the datapoint distribution.

For at least half of all query points, the distance ε_{NN} to the nearest neighbour in X is at least as large as ε_M, the median NN distance. Let ω be such a query centre. For every element C of level t partition of Ω, one has, using Lemmas 2 and 1 and the assumption in Eq. (2),

$$\alpha_C(\varepsilon_M) \leq \frac{\alpha_\Omega(\varepsilon_M/2)}{\mu(C)^{-1}} = \Theta(2^t)e^{-\Theta(1)\varepsilon_M^2 d} = e^{-\Theta(d)},$$

where the constants *do not depend* on a particular internal node C. An argument in Section 7 implies that branching *at every internal node* occurs for all ω except a set of measure

$$\leq \sharp(\text{nodes}) \times 2 \sup_C \alpha_C(\varepsilon) = O(n^2) e^{-\Theta(d)} = o(1),$$

because $d = \omega(\log n)$ and so $e^{\Theta(d)}$ is superpolynomial in n. Thus, the expected average performance of an indexing scheme as above is linear in n.

The problem with arguments of this kind (seen from time to time in data engineering papers) is this. We have replaced the value of the *empirical measure*,

$$\mu_n(C) = \frac{|C|}{n},$$

with $\mu(C)$, implicitely assuming that the two are close to each other:

$$\mu_n(C) \approx \mu(C).$$

But the scheme is being chosen *after* seeing an instance X, and it is reasonable to assume that the choice of indexing partitions will take advantage of large random clusters always present in uniformly distributed data. (Fig. 8 illustrates this point in dimension $d = 2$.) Thus, some elements of indexing partitions, while having large measure μ, may contains few datapoints, and vice versa.

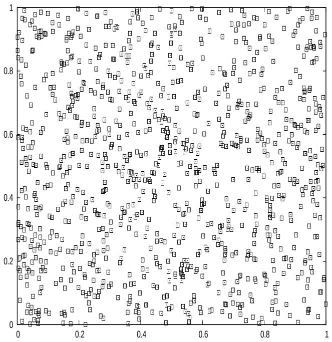

Fig. 8 1000 points randomly and uniformly distributed in the square $[0,1]^2$.

An equivalent consideration is that we only know the concentration function of the domain Ω, but not of a randomly chosen dataset X. It seems the research problem of estimating the concentration function of a random sample has not been systematically treated.

In order to be able to estimate the empirical measure in terms of the underlying distribution, one needs to invoke an approach of statistical learning.

9 Vapnik–Chervonenkis theory

Let $\mathscr{A}$ be a family of subsets of a set Ω (a *concept class*). One says that a subset $B \subseteq \Omega$ is *shattered* by $\mathscr{A}$ (cf. Fig. 9) if for each $C \subseteq B$ there is $A \in \mathscr{A}$ such that

$$A \cap B = C.$$

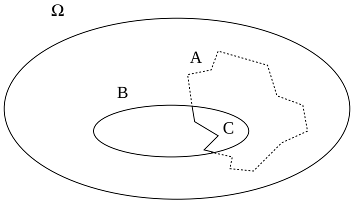

Fig. 9 A set B is shattered by the class $\mathscr{A}$.

The *Vapnik–Chervonenkis dimension* VC-dim$(\mathscr{A})$ of a class $\mathscr{A}$ is the largest cardinality of a set $B \subseteq \Omega$ shattered by $\mathscr{A}$.

Estimating the VC dimension is often non-trivial, and here are some examples.

1. The VC dimension of the class of all Euclidean balls in $\mathbb{R}^d$ is $d+1$.
2. The class of all parallelepipeds in $\mathbb{R}^d$ has VC dimension $2d+2$.
3. The VC dimension of the class of all ℓ^1-balls in the Hamming cube $\{0,1\}^d$ is bounded from above by $d + \lfloor \log_2 d \rfloor$.
   (As every ball is determined by its centre and radius, the total number of pairwise different balls in $\{0,1\}^d$ is $d2^d$. Now one uses an obvious observation: the VC dimension of a finite concept class $\mathscr{A}$ is bounded above by $\log_2 |\mathscr{A}|$.)

Here is a deeper and very general observation.

Theorem 2 (Goldberg and Jerrum [9]) *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 VC-dim$(\mathscr{F}) \leq 4s(t+2)$. □

Now, a typical result of statistical learning theory (see [1,25,27] for more).

Theorem 3 *Let $\mathscr{A} \subseteq 2^{\Omega}$ be a concept class of finite VC dimension, d. Then for all $\epsilon, \delta > 0$ and every probability measure μ on Ω, if n datapoints in X are drawn randomly and independently acoording to μ, then with confidence $1-\delta$*

$$\forall A \in \mathscr{A}, \quad \left| \mu(A) - \frac{|X \cap A|}{n} \right| < \epsilon,$$

provided n is large enough:

$$n \geq \frac{128}{\epsilon^2} \left(d \log \left(\frac{2e^2}{\epsilon} \log \frac{2e}{\epsilon} \right) + \log \frac{8}{\delta} \right).$$

For statistical learning theory, we refer to [1,25,27], or a set of lecture notes [13].

Here is one of many existing analogues of the concept of VC dimension for classes of functions. Let $\mathscr{F}$ be a class of (possibly partially defined) real-valued functions on Ω. Denote by $\mathscr{F}_{\substack{\geq \\ <}}$ the class of all subsets of Ω of the form

$$\{\omega \in \operatorname{dom} f \colon f(\omega) \geq a\} \text{ or } \{\omega \in \operatorname{dom} f \colon f(\omega) \leq a\}, \quad f \in \mathscr{F}, \quad a \in \mathbb{R}. \tag{5}$$

The *Vapnik pseudodimension of $\mathscr{F}$* is the VC dimension of the concept class $\mathscr{F}_{\substack{\geq \\ <}}$.

For example, if $\mathscr{F}$ is the class of all distance functions to points of $\mathbb{R}^d$, the Vapnik pseudodimension of $\mathscr{F}$ is $2(d+1)$. It is usually easy to estimate pseudodimention of function classes where decision functions of metric trees of various types come from.

10 Examples of indexing schemes

10.1 vp-tree

The *vp-tree* [29] uses decision functions of the form

$$f_t(\omega) = (1/2)(\rho(x_{t_+}, \omega) - \rho(x_{t_-}, \omega)),$$

where $t_{\pm}$ are two children of t and $x_{t_{\pm}}$ are the *vantage points* for the node t.

10.2 M-tree

The *M-tree* [6] employs decision functions

$$f_t(\omega) = \rho(x_t, \omega) - \sup_{\tau \in B_t} \rho(x_t, \tau),$$

where B_t is a block corresponding to the node t, x_t is a datapoint chosen for each node t, and suprema on the r.h.s. are precomputed and stored.

For both schemes, if the domain $\Omega = \mathbb{R}^d$, then the Vapnik pseudodimension of the class of all possible decision functions is $d+1$. A similar conclusion holds for the Hamming cube.

11 Rigorous lower bounds

In this Section we prove the following theorem under general assumptions of Section 6.

Theorem 4 *Let the domain Ω equipped with a metric ρ and probability measure μ have concentration dimension $\Theta(d)$ (cf. Eq. (2)) and expected distance between two points $\mathbb{E}d(x,y)=1$. Let $\mathscr{F}$ be a class of all 1-Lipschitz functions on the domain Ω that can be used as decision functions for metric tree indexing schemes of a given type. Suppose the Vapnik pseudodimention p of $\mathscr{F}$ is polynomial in d:*

$$p = d^{O(1)}.$$

Let X be an i.i.d. random sample of Ω according to μ, having n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$. Then, with confidence asymptotically approaching 1, an optimal metric tree indexing scheme for the similarity workload (Ω, ρ, X) has expected average performance $d^{\omega(1)}$. In other words, the average search time for a nearest neighbour is superpolynomial in dimension d.

The following is an immediate consequence of Lemma 4.2 in [16].

Lemma 3 ("Bin Access Lemma") *Let $\varepsilon > 0$ and $m \geq 4$ be such that $\alpha_{\Omega}(\varepsilon) \leq m^{-1}$, and let γ be a collection of subsets $A \subseteq \Omega$ of measure $\mu(A) \leq m^{-1}$ each, satisfying $\mu(\cup\gamma) \geq 1/2$. Then the 2ε-neighbourhood of every point $\omega \in \Omega$, apart from a set of measure at most $\frac{1}{2}m^{-\frac{1}{2}}$, meets at least $\frac{1}{2}m^{\frac{1}{2}}$ elements of γ.*

Here is the next step in the proof.

Lemma 4 *Denote $\mathscr{B}$ the class of all subsets $B \subseteq \Omega$ appearing as bins of metric trees of depth $\leq h$ built using certification functions from a class $\mathscr{F}$ of Vapnik pseudodimension $\leq p$. Then*

$$\textit{VC-dim}\,(\mathscr{B}) \leq 2hp\log(hp) = O(hp).$$

Proof Every such set B is an intersection of a family of $\leq h$ sets of the form (5). Now one uses Th. 4.5 in [27]: if $\mathscr{A}$ is a concept class of VC dimension $\leq p$, then the VC dimension of the class of all sets obtained as intersections of $\leq h$ sets from $\mathscr{F}$ is bounded by $2hp\log(hp)$. □

Let us prove Theorem 4. Without loss in generality, suppose that for any value $0 < c < 1$ such as e.g. $c = 1/4$, for all points ω except in a set of measure $\leq c$ the depth of the search tree is polynomial in d, uniformly in ω, for otherwise there is nothing to prove.

Using Eq. (1) and Lemma 1, pick any $\varepsilon' > 0$ such that, for sufficiently high values of d, for most points ω the value of $\varepsilon_{NN}(\omega)$ exceeds ε'. Let $0 < \beta < 1/2$. Again without losing generality, we can assume that the measure of the set of query centres ω whose ε'-neighbourhood meets at least one bin with $\geq n^{1/2-\beta}$ points is $\leq 1/4$.

Combining the two assumptions together, we deduce that for at least half of all query centres ω the ε'-ball around ω only meets bins with fewer than $n^{1/2-\beta}$ points. By Theorem 3 and Lemma 4, the value of measure μ for each of these bins is $\leq 2n^{1/2-\beta}$ if n is sufficiently large. Lemma 3, applied with $m = 2n^{1/2-\beta}$ and $\varepsilon = \varepsilon'/2$, implies that for all ω from a set of measure $1 - o(1)$ the ε'-neighbourhood of ω meets at least $O(n^{1/4-\beta/2}) = d^{\omega(1)}$ bins. Since accessing each bin requires at least one operation (let even to check that a bin is empty), the theorem is proved. □

Combining our Theorem 4 with Theorem 2 of Goldberg and Jerrum shows that for all practical purposes the worst-case average performance of metric trees is superpolynomial in dimension of the domain.

Theorem 5 *Let the domain $\Omega = \mathbb{R}^d$ be equipped with a probability measure μ_d in such a way that $(\mathbb{R}^d, \mu_d)$ form a normal Lévy family and the μ_d-expected value of the Euclidean distance is $\Theta(1)$. Let $\mathscr{F}_d$ denote a class of functions $f(\theta, x)$ on $\mathbb{R}^d$ parametrized with θ taking values in a space $\mathbb{R}^{\mathrm{poly}\,(d)}$ and such that computing each value $f(\theta, x)$ takes $d^{O(1)}$ operations of the type described in Thm. 2. Let X be an i.i.d. random sample of $\mathbb{R}^d$ according to μ_d, having n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$. Then, with confidence asymptotically approaching 1, an optimal metric tree indexing scheme for the similarity workload (Ω, ρ, X) whose decision functions belong to the parametrized class $\mathscr{F}$ has expected average performance $d^{\omega(1)}$.* □

Two remarks are in order to explain the strength of the above result.

(1) Measures μ_d satisfying the above assumption include, for instance, the normal gaussian distribution $\mathcal{N}(0, 1)$, the uniform measures on the unit ball, on the unit sphere, etc.

(2) A polynomial upper bound on the size of the parameter θ for $\mathscr{F}$ is dictated by the obvious restriction that reading off a parameter of superpolynomial length leads to a superpolynomial lower bound on the length of computation.

Conclusion

In this paper, we have obtained superpolynomial lower bounds on the performance of a wide class of indexing schemes for similarity-based information retrieval in datasets of high intrinsic dimension. The results were obtained both in great generality and within mathematically exacting standards of statistical learning. In particular, we have stressed the importance of using statistical learning methods (Vapnik-Chernonenkis theory) in order to justify heuristic arguments often used in data engineering for the purpose of algorithm analysis.

The significance of superpolynomial lower bounds on the performance of various indexing schemes is not that they rule out using the schemes in quesion, but rather provide a better insight on how they function. Indeed, most data practitioners seem to believe that the intrinsic dimension of real-life datasets does not exceed as few as perhaps seven or ten dimensions. A deeper understanding of underlying geometry of workloads and its interplay with compleixty is called for in order to learn to detect and use this low dimensionality efficiently, and asymptotic analysis of algorithm performance in an artificial setting of very high dimensions is contributing towards this goal.

We believe that a glimpse into the underlying geometric and probabilistic nature of the curse of dimensionality offered by this article can be useful for the challenges faced by data engineering.

References

1. M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, Cambridge, 1999.
2. 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.
3. 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.
4. 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.*, 312–321, 1999.
5. E. Chávez, G. Navarro, R. Baeza-Yates and J. L. Marroquín. Searching in metric spaces. *ACM Computing Surveys* 33:273–321, 2001.
6. 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.
7. 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.
8. A. Faragó, T. Linder, and G. Lugosi, Fast nearest neighbor search in dissimilarity spaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 957–962, 1993.
9. 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.
10. J. M. Hellerstein, E. Koutsoupias, D. P. Miranker, C. Papadimitriou, and V. Samoladas. On a model of indexability and its bounds for range queries. *Journal of the ACM (JACM)*, 49(1):35–55, 2002.
11. 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.

12. M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2001.
13. S. Mendelson, A few notes on statistical learning theory. In: S. Mendelson, A.J. Smola, Eds., *Advanced Lectures in Machine Learning*, LNCS 2600, pp. 1–40, Springer, 2003.
14. V.D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces*, volume 1200 of *Lecture Notes in Mathematics.* Springer, 1986.
15. P.B. Miltersen, *Cell probe complexity - a survey.* In: 19th Conference on the Foundations of Software Technology and Theoretical Computer Science (FSTTCS), 1999. Advances in Data Structures Workshop.
16. V. Pestov. On the geometry of similarity search: dimensionality curse and concentration of measure. *Inform. Process. Lett.*, 73:47–51, 2000.
17. 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. pp. 1775–1780, 2007.
18. V. Pestov. An axiomatic approach to intrinsic dimension of a dataset. *Neural Networks,* 21:204–213, 2008.
19. V. Pestov and A. Stojmirović. Indexing schemes for similarity search: an illustrated paradigm. *Fund. Inform.*, 70:367–385, 2006.
20. H. Samet. *Foundations of Multidimensional and Metric Data Structures.* Morgan Kaufmann Publishers Inc., San Francisco, CA, 2005.
21. S. Santini, *Exploratory Image Databases: Content-Based Retrieval,* Academic Press, Inc. Duluth, MN, USA, 2001.
22. U. Shaft and R. Ramakrishnan. Theory of nearest neighbors indexability. *ACM Transactions on Database Systems (TODS),* 31:814–838, 2006.
23. A. Stojmirović and V. Pestov. Indexing schemes for similarity search in datasets of short protein fragments. *Information Systems,* 32:1145–1165, 2007.
24. J.K. Uhlmann. Satisfying general proximity/similarity queries with metric trees, *Information Processing Letters* 40:175–179, 1991.
25. V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.
26. S.S. Vempala. *The Random Projection Method.* DIMACS Series in Discrete Mathematics and Theoretical Computer Science, **65**, Amer. Math. Soc., Providence, R.I., 2004.
27. M. Vidyasagar. *Learning and Generalization, With Applications to Neural Networks.* Second Ed. Springer-Verlag, London, 2003.
28. R. Weber, H.-J. Schek, and S. Blott, A quantatitive analysis and performance study for similarity-search methods in high-dimensional spaces. in: *Proceedings of the 24-th VLDB Conference,* New York, pp. 194–205, 1998.
29. P. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces, in: *Proc. 3rd Annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 311–321, 1993.
30. P. Zezula, G. Amato, Y. Dohnal, and M. Batko. *Similarity Search. The Metric Space Approach.* Springer Science + Business Media, New York, 2006.