Indexability, concentration, and VC theory

Vladimir Pestov

Department of Mathematics and Statistics
University of Ottawa
Ottawa, Ontario, Canada

SISAP'2010, Istanbul, Sept. 18–19, 2010

Origins of the curse of dimensionality?

Suppose the indexing scheme $=$ family of 1-Lipschitz functions $f_i \colon \Omega \to \mathbb{R}$, $i \in I$ (fully or partially defined):

$$|f_i(x) - f_i(y)| \leq \rho(x, y).$$

Given $q \in \Omega$ and $\varepsilon > 0$, the algorithm chooses recursively $f_{i_1}, f_{i_2}, \ldots, f_{i_m}$, where i_{m+1} is determined by $f_{i_1}(q)$, $f_{i_2}(q)$, $\ldots$, $f_{i_m}(q)$.

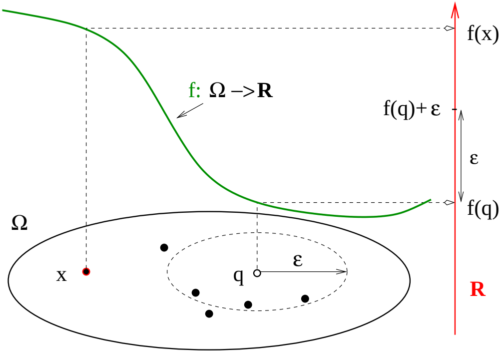
All $x \in X$ with $|f_i(q) - f_i(x)| > \varepsilon$ are discarded.

Points $x \in X$ that are not discarded are returned.

Space complexity: $|I|$

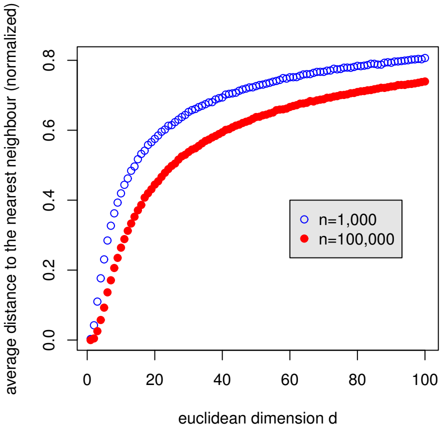
Time complexity: # of function evaluations at q $+$ # of points returned.

Using 1-Lipschitz condition



The datapoint x can be discarded.

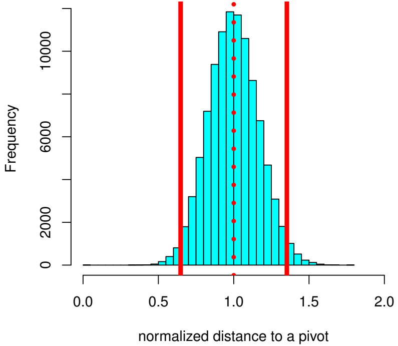
Empty space paradox



Normalized average NN distance in $X \sim$ gaussian distribution in $\mathbb{R}^d$.

1-Lipschitz functions concentrate around their mean

Distances to a random pivot in a dataset X of $n = 10^5$ points $\sim$ gaussian distribution in $\mathbb{R}^{14}$:



$\therefore$ few points are discarded, $\rightsquigarrow$ degrading performance.

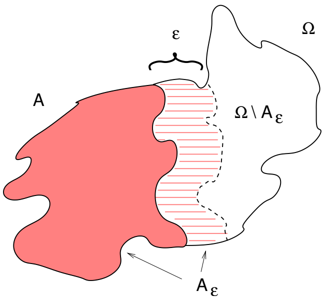
Concentration phenomenon

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

(Ω, ρ) is a metric space, carrying a probability distribution, μ.
(forget datapoints for the moment).

Concentration function

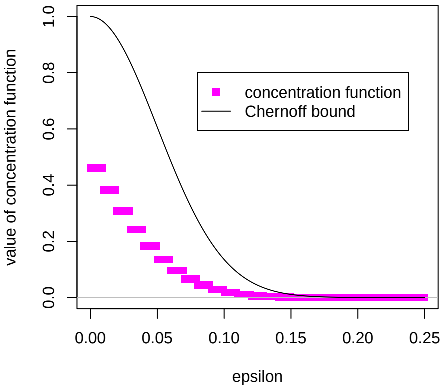
$$\alpha_\Omega(\varepsilon) = \begin{cases} \frac{1}{2}, & \text{if } \varepsilon = 0, \\ 1 - \inf\left\{\mu\left(A_\varepsilon\right) : A \subseteq \Omega,\ \ \mu_\sharp(A) \geq \frac{1}{2}\right\}, & \text{if } \varepsilon > 0. \end{cases}$$



Value $\alpha_\Omega(\varepsilon)$ = upper bound of sizes of $\Omega \setminus A_\varepsilon$.

Typically, $\alpha(\varepsilon) \leq \exp(-O(\varepsilon^2 d))$

Concentration function of the Hamming cube $\{0,1\}^{100}$:



Concentration of 1-Lipschitz functions

Let $f\colon \Omega \to \mathbb{R}$ be 1-Lipschitz, $\varepsilon > 0$. Then:

$$\mu\{x \in \Omega\colon\ |f(x) - M_f| > \varepsilon\} \leq 2\alpha_\Omega(\varepsilon), \tag{1}$$

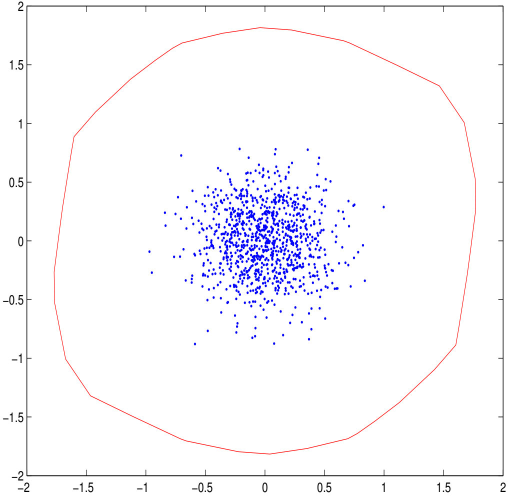
where M_f is the median value of f:

$$\mu[f \geq M_f] \geq 1/2,$$

$$\mu[f \leq M_f] \geq 1/2.$$

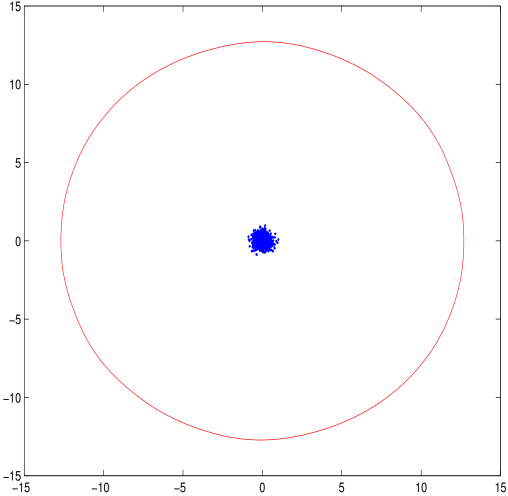
Note: $M_f \approx \mathbb{E}(f)$ as $d \to \infty$.

Random projection of cubes



1000 points in $d = 20$ cube

Random projection of cubes



1000 points in $d = 1000$ cube.

High-dimensional domains

High dimensional objects are all alike to a low-dimensional observer.

- $\mathbb{R}^d$, gaussian measure
- $[0, 1]^d$, uniform measure
- $\mathbb{S}^d$, sphere with the Lebesgue measure
- $\{0, 1\}^d$, Hamming cube.

Asymptotic assumptions

- datapoints are drawn from $\Omega = (\Omega, \rho, \mu)$ in an i.i.d. fashion;
- ρ is normalized:

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

- Ω "has concentration dimension" d:

$$\alpha_\Omega(\varepsilon) = \exp(-\Omega(\varepsilon^2 d));$$

- $n = |X|$ grows faster than any polynomial in d, but slower than any exponential function in d:

$$n = d^{\omega(1)}, \quad d = \omega(\log n).$$

Distance to NN

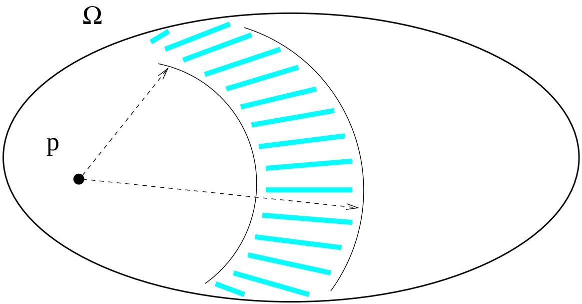
$\varepsilon_{NN}(q) = d(q, X)$, a real function on Ω.

- Highly concentrated around the median, ε_M.
- With high confidence, ε_{NN} converges to "characteristic size" $\mathbb{E}_{\mu\otimes\mu}(d)$ as $d \to \infty$.

Pivot tables and concentration

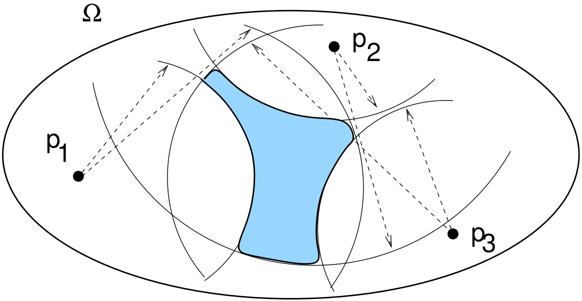
For every $p \in \Omega$, the distance function $\rho(p, -)$ concentrates around its median, M_p. And ε_M is large.

$\therefore$ Spherical shell $\varepsilon_M \pm \varepsilon$ has measure $1 - \exp(-\Omega(d)\varepsilon^2)$.



Intersections of spherical shells

Same goes for intersection, S, of $O(n)$ spherical shells:



If q and x belong to S, *and* $\varepsilon_{NN}(q) \geq \varepsilon_M$ (true for half of query points), *then* x cannot be discarded.

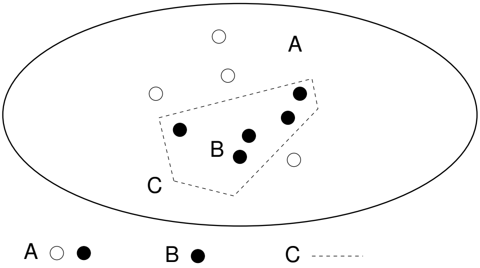
Can we infer that since S is big, it contains *lots of datapoints*?

Shattering

$\mathscr{C}$: a family of subsets of Ω.
Finite $A \subseteq \Omega$ is *shattered* by $\mathscr{C}$ if for every $B \subseteq A$

$$B = A \cap C$$

for a suitable $C \in \mathscr{C}$.



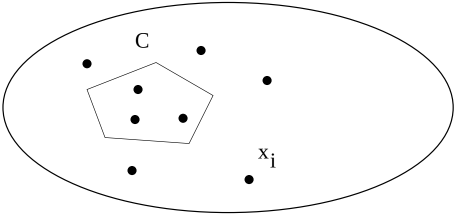
$\mathrm{VC}(\mathscr{C})$: the largest size of $A \subseteq \Omega$ shattered by $\mathscr{C}$.

Vapnik–Chervonenkis dimension: examples

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

Empirical measures

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



Law of Large Numbers $\rightsquigarrow \mu_n(C) \to \mu(C)$ as $n \to \infty$.

If $\mathrm{VC}(\mathscr{C}) < \infty$, the same happens for *every* $C \in \mathscr{C}$, uniformly.

Uniform convergence of empirical measures

A class $\mathscr{C}$ on Ω has *UCEM property* if there is a function $s(\delta, \varepsilon)$so that, given $\varepsilon > 0$ (precision) and $\delta > 0$ (risk), whenever $n \geq s(\delta, \varepsilon)$, one has

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

for *every* μ on Ω. Here $\mu_n(C) = \sharp\{i : x_i \in C\}/n$ is the empirical measure on the sample $\bar{x}$.

thm. (Vapnik–Chervonenkis) $\mathscr{C}$ has UCEM property if and only if $d = \mathrm{VC}(\mathscr{C}) < \infty$. In this case,

$$s(\delta, \varepsilon) = O(d\varepsilon^{-1}(-\log\varepsilon - \log\delta)).$$

Curse of dimensionality for pivot tables

$\therefore$ The class $\mathscr{C}$ of all possible k-fold intersections of spherical shells in Ω has VC dimension $\leq 2k\lg(ek)O(d)$.

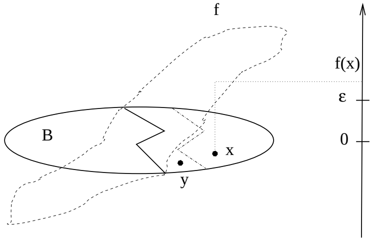
If $k = o(n/d \log n)$, this is $o(n)$. It follows that, with high confidence $1 - \delta$, the proportion of datapoints in *every* k-fold intersection of shells is close to its measure.

Apply to *giant* intersections to get:

thm. (Pestov–Volnyansky) For Ω Hamming cube, Euclidean cube, ... , the average total complexity of the performance of any pivot table is $\Omega\,(n/d \log n)$.

Metric trees

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



Curse of dimensionality for metric trees

thm. Let $\mathscr{F}$ be the class of 1-Lipschitz functions used to construct a particular type of metric tree. If the VC dimension of the class θf, $f \in \mathscr{F}$ is $\mathrm{poly}(d)$, then the expected average performance of the metric tree is superpolynomial in d.

Theorem of Goldberg and Jerrum

How sensible is the assumption $\mathrm{VC}(\mathscr{F}) = \mathrm{poly}(d)$?

thm. 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 for each $x \in \mathbb{R}^n$, there is an algorithm that computes $f(\theta, x)$ in no more than t operations of the following types:

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

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

Curse of dimensionality conjecture

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

Cell probe model

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

f_t is defined on (a subset of) Ω and takes a b-bit string σ as a parameter, except if $t = 0$ is the root.

$f_t(\sigma; q)$ is a pair (τ, s), with τ a b-bit string and s a child of t.

If $i = i(t)$ where t is an inner node, C_i can hold a b-bit string.

If $i = i(t)$ where t is a leaf, then C_i can hold a datapoint $x \in X$.

Initialization: memory image of cells. Reaching leaf, read off x contained in $C_{i(s)}$.

Best results to date

Best lower bound currently known:

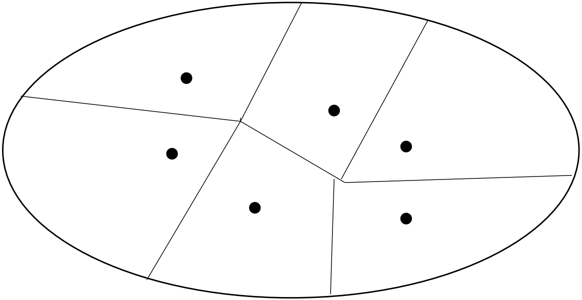
$$O\left(d / \log \frac{sd}{n}\right),$$

where $s =$ number of cells (Pătrascu and M. Thorup 2006).

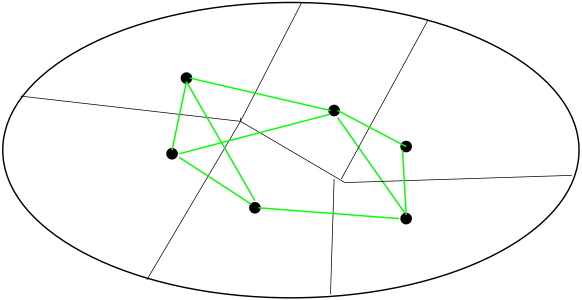
$\rightsquigarrow$ earlier bound $\Omega(d / \log n)$ for polynomial space data structures (Barkol and Rabani 2000), and

$\rightsquigarrow$ lower bound $\Omega(d / \log d)$ for "near linear space" $n \log^{O(1)} n$.

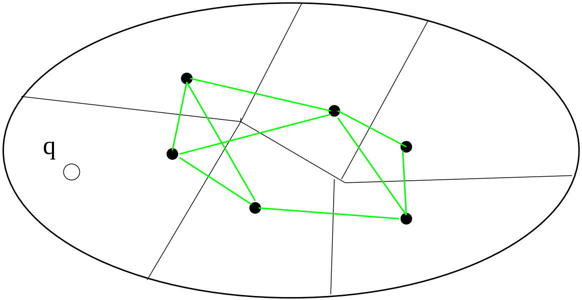
Delaunay graph-based indexing



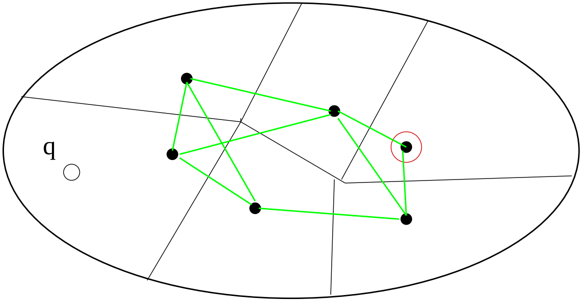
Delaunay graph-based indexing



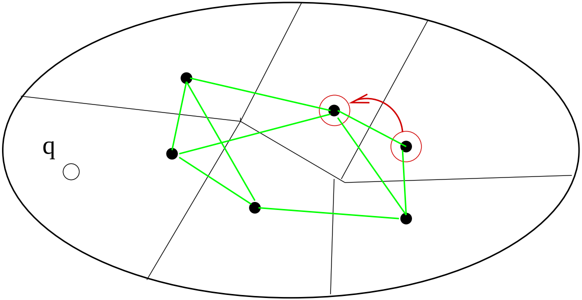
Delaunay graph-based indexing



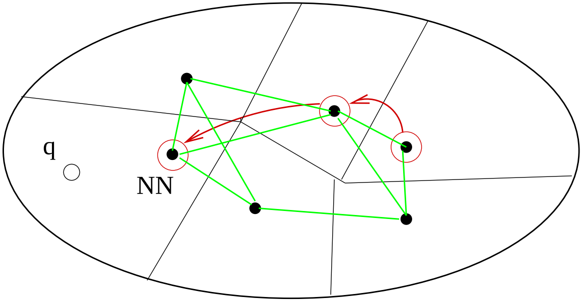
Delaunay graph-based indexing



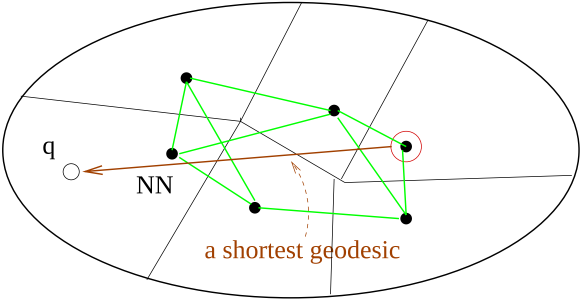
Delaunay graph-based indexing



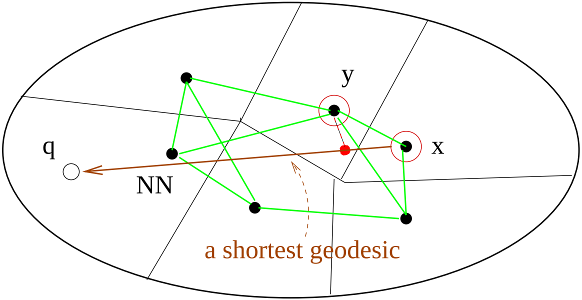
Delaunay graph-based indexing



Delaunay graph-based indexing



Delaunay graph-based indexing



Delaunay graph-based indexing

thm. (Gonzalo Navarro) If X is a finite metric space, $x, y \in X$, then one can embed X into a metric space Ω where x, y are Delaunay-adjacent.

In fact: X can be embedded into a metric space (Urysohn metric space $\mathbb{U}$) in which *every two distinct* x, y *are Dalaunay–adjacent*.

Even better: under our assumptions, for d large enough every two points in X are Delaunay-adjacent for all common domains.

The curse of dimensionality is present, but apparently for different reasons. How to give a common proof?

Approximate NN search

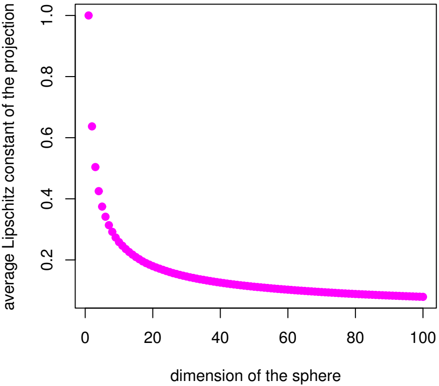
(r, ε)-ANN search: given r and $\varepsilon > 0$, for a $q \in \Omega$, if $\varepsilon_{NN}(q) \leq r$, then return a datapoint $x \in X$ at a distance $d(q, x) < (1 + \varepsilon)r$.

Often said: “free from the curse of dimensionality”.

Not quite true... yet, much more efficient algorithms are known.

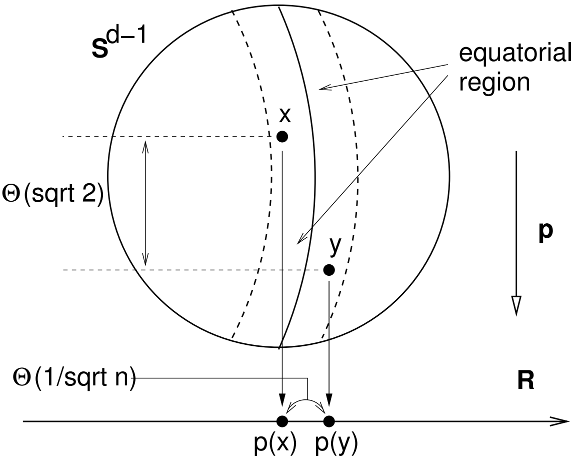
Often randomized.

Expected distortion of a random projection



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

Geometry of random projections



Johnson–Lindenstrauss lemma

Renormalized projection:

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

is *approximately* 1-*Lipschitz* for a *finite fraction of pairs*.

To achieve distortion in the range $1 \pm \varepsilon$ with high confidence, combine $k = O(\log n/\varepsilon^2)$ mutually orthogonal projections as above, that is, project on a randomly chosen k-dimensional subspace.

(*Johnson–Lindenstrauss lemma*.)

Indyk–Motwani: combined such a projection on $\mathbb{R}^{O(\log n/\varepsilon^2)}$ with an indexing scheme in a low-dimensional space.

$\rightsquigarrow$ LSH for ANN search.

Scheme of Kushilevitz, Ostrovsky and Rabani

Let $\Omega = \{0,1\}^d$, $X \subseteq \{0,1\}^d$ a dataset.

Turn things on their head: take as the domain

$\Omega = [d] = \{1,2,3,\ldots,d\}$,
and view datapoints x as *concepts* (subsets of $[d]$):

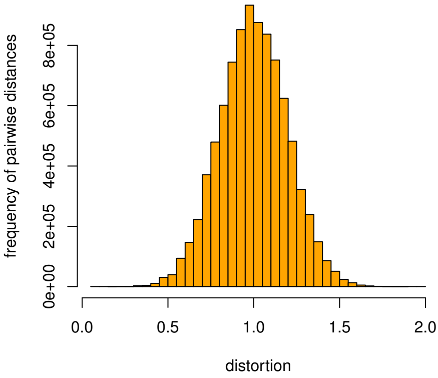
$$x \mapsto \{i : x_i = 1\} \subseteq [d].$$

$\mathrm{VC}(X) \leq \lg n$.

If $\varepsilon > 0$, by VC theory, pairwise distances in X are within $\pm\varepsilon$ of distances in the Hamming cube on $O(\varepsilon^{-2} \lg n)$ randomly sampled coordinates (with high confidence!)
Same holds for $X \cup \{q\}$ for most q.

Hash table for NN in $\{0,1\}^{O(\varepsilon^{-2} \lg n)}$ (size: $n^{O(\varepsilon^{-2})}$) $\rightsquigarrow$ efficient (r,ε)-ANN search for r is a "reasonable" range.

Illustration to Kushilevitz–Ostrovsky–Rabani



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

Discussion: query stability

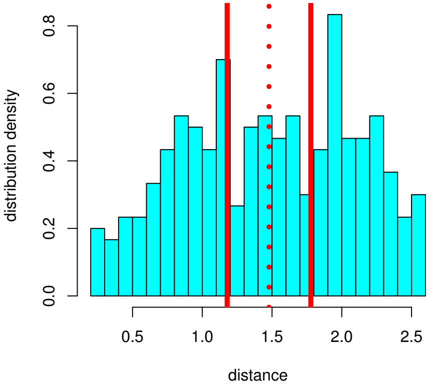
An NN query with centre q is ε-*unstable* (Beyer–Golsdtein–Ramakrishnan–Shaft) if the $(1+\varepsilon)\varepsilon_{NN}(q)$-ball around q contains $\geq n/2$ datapoints.

Asymptotically, under our assumptions, most queries are ε-unstable.

In particular, (r,ε)-ANN search becomes meaningless for ε-unstable queries: pick x at random, and with high confidence, the result is correct.

The curse of dimensionality conjecture $\approx$ "unstable queries are impossible to answer." (believable, yes; relevant, ?; unproven.)

Discussion: intrinsic dimension



Empirical density histogram of distances from a pivot having the highest found value of dissipation for the NASA dataset. Lines: the mean $\pm$ tolerance range $\varepsilon = 0.275$.

Discussion: derandomization

Wigderson and others: if $P \neq NP$, then one can use any hard function as a source of random bits, and this turns randomized algorithms (correct with high confidence) into deterministic (provably correct).

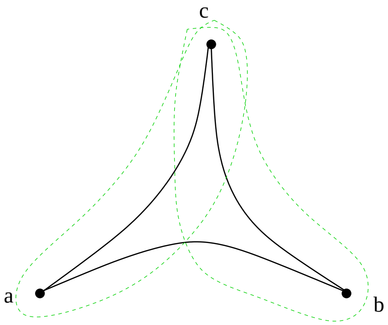
Say NN query is *stable* if the $(1+\varepsilon)\varepsilon_{NN}(q)$-ball around q contains a manageable number of data (say, $\mathrm{poly}\,(d)$).

Derandomizing Kushilevitz—Ostrovsky—Rabani, one would be able to answer *stable* queries (the ones worth answering...) in $\mathrm{poly}\,(d)$ time.

(Won't contradict the curse of dimensionality conjecture which is all about impossibility to answer *unstable* queries).

Discussion: hyperbolicity

A metric space is *hyperbolic* if there is $\delta > 0$ so that for each geodesic triangle a, b, c the side $[a, b]$ is contained in the δ-neighbourhood of $[b, c] \cup [a, c]$.



Alain Connes suggested that long-term memory uses Delaunay graph of a hyperbolic simplicial complex.