Analysis of hierarchical metric-tree indexing schemes

for similarity search in high-dimensional datasets

Vladimir Pestov

vpest283@uottawa.ca

http://aix1.uottawa.ca/˜vpest283

Department of Mathematics and Statistics

University of Ottawa

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and
- $\mathcal{Q} \subseteq 2^{\Omega}$ is the set of *queries.*

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and
- $\mathcal{Q} \subseteq 2^{\Omega}$ is the set of *queries.*

Answering a query $Q \in \mathcal{Q}$

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and
- $\mathcal{Q} \subseteq 2^{\Omega}$ is the set of *queries.*

Answering a query $Q \in \mathcal{Q}$ is listing all $x \in X \cap Q$.

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and
- $\mathcal{Q} \subseteq 2^{\Omega}$ is the set of *queries.*

Answering a query $Q \in \mathcal{Q}$ is listing all $x \in X \cap Q$.

A (*dis*)*similarity measure* $s\colon \Omega \times \Omega \to \mathbf{R}$,

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and
- $\mathcal{Q} \subseteq 2^{\Omega}$ is the set of *queries.*

Answering a query $Q \in \mathcal{Q}$ is listing all $x \in X \cap Q$.

A (*dis*)*similarity measure* $s\colon \Omega \times \Omega \to \mathbf{R}$,
e.g. a metric, or a pseudometric.

General setting

Workload: $W = (\Omega, X, \mathcal{Q})$, where:

- Ω is the *domain,*
- $X \subset \Omega$ finite subset (*dataset*, or *instance*), and
- $\mathcal{Q} \subseteq 2^{\Omega}$ is the set of *queries.*

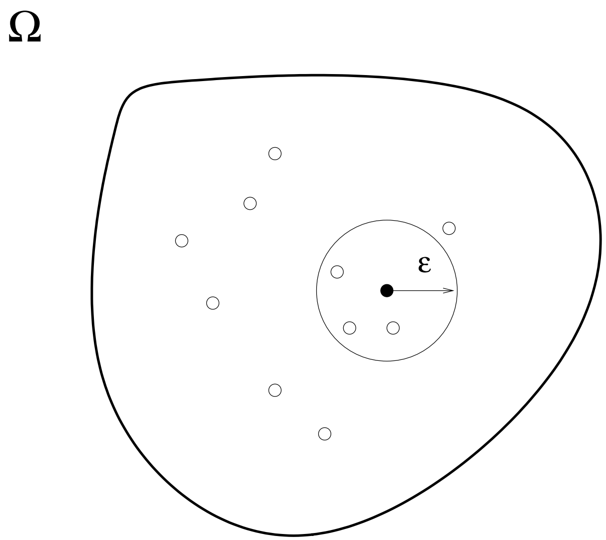
Answering a query $Q \in \mathcal{Q}$ is listing all $x \in X \cap Q$.

A (*dis*)*similarity measure* $s\colon \Omega \times \Omega \to \mathbf{R}$,
e.g. a metric, or a pseudometric.

A *range similarity query centred at* $\omega \in \Omega$:

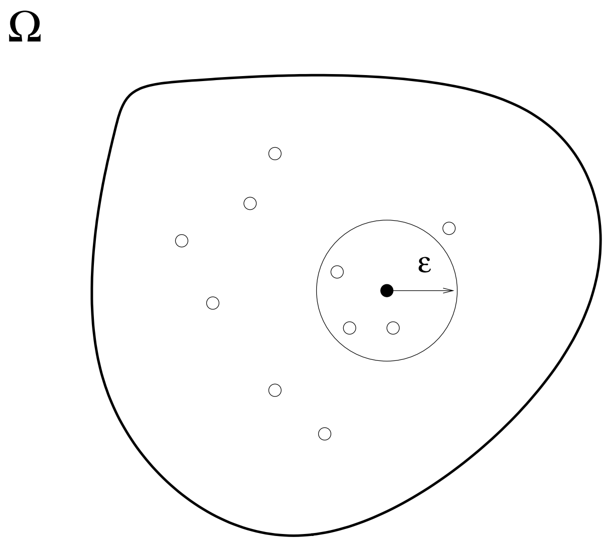
$$Q = \{x \in \Omega\colon s(\omega, x) < \varepsilon\}$$

Similarity workloads



$$W = (\Omega, d, X, \{\mathcal{B}_\varepsilon(x)\})$$

Similarity workloads



$$W = (\Omega, d, X, \{\mathcal{B}_\varepsilon(x)\})$$

- *k-nearest neighbours (k-NN) query* centred at $x^* \in \Omega$, where $k \in \mathbb{N}$.

Example

Example

- $\Omega =$ strings of length $m = 10$ from the alphabet Σ of 20 standard amino acids: $\Omega = \Sigma^{10}$.

Example

- $\Omega =$ strings of length $m = 10$ from the alphabet Σ of 20 standard amino acids: $\Omega = \Sigma^{10}$.
- $X =$ all peptide fragments of length 10 in the SwissProt database (as of 19-Oct-2002). $|X| = 23,817,598$.

Example

- $\Omega =$ strings of length $m = 10$ from the alphabet Σ of 20 standard amino acids: $\Omega = \Sigma^{10}$.
- $X =$ all peptide fragments of length 10 in the SwissProt database (as of 19-Oct-2002). $|X| = 23,817,598$.
- *Similarity measure* given by the most common scoring matrix in sequence comparison, BLOSUM62, by $s(a,b) = \sum_{i=1}^{m} s(a_i, b_i)$ (the *ungapped* score).

Example

- $\Omega =$ strings of length $m = 10$ from the alphabet Σ of 20 standard amino acids: $\Omega = \Sigma^{10}$.
- $X =$ all peptide fragments of length 10 in the SwissProt database (as of 19-Oct-2002). $|X| = 23,817,598$.
- *Similarity measure* given by the most common scoring matrix in sequence comparison, BLOSUM62, by $s(a,b) = \sum_{i=1}^{m} s(a_i, b_i)$ (the *ungapped* score).
- Converted into quasi-metric $d(a,b) = s(a,a) - s(a,b)$, generating the same set of queries (range and k-NN).

(joint with A. Stojmirović)

Inner *vs* outer

Inner *vs* outer

- *Inner workload* if $X = \Omega$,

Inner *vs* outer

- *Inner workload* if $X = \Omega$,
- *Outer workload* if $|X| \ll |\Omega|$.

Inner *vs* outer

- *Inner workload* if $X = \Omega$,
- *Outer workload* if $|X| \ll |\Omega|$.

Fragment example: outer,

$$|X|/|\Omega| = 23,817,598/20^{10} \approx 0.0000023$$

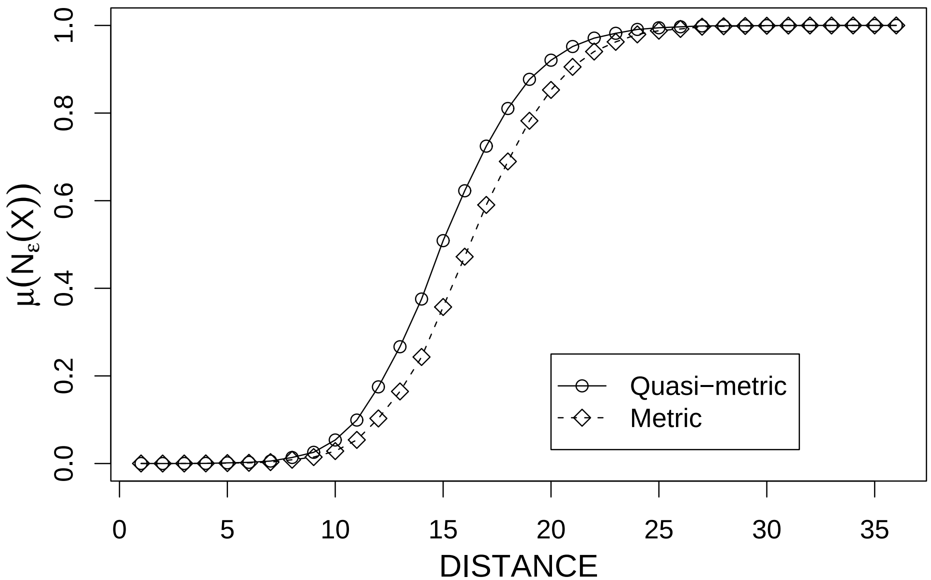
Inner *vs* outer

- *Inner workload* if $X = \Omega$,
- *Outer workload* if $|X| \ll |\Omega|$.

Fragment example: outer,

$$|X|/|\Omega| = 23,817,598/20^{10} \approx 0.0000023$$

Most points $\omega \in \Omega$ have NN $x \in X$ within $\varepsilon = 25$ (high biological relevance).



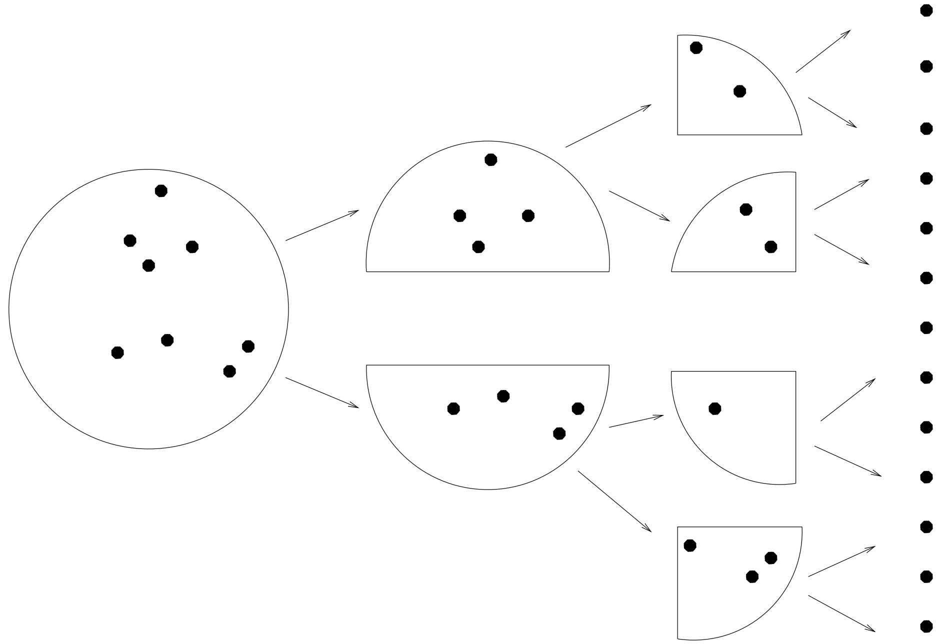
Hierarchical tree index structures

Hierarchical tree index structures

A sequence of refining partitions of the domain:

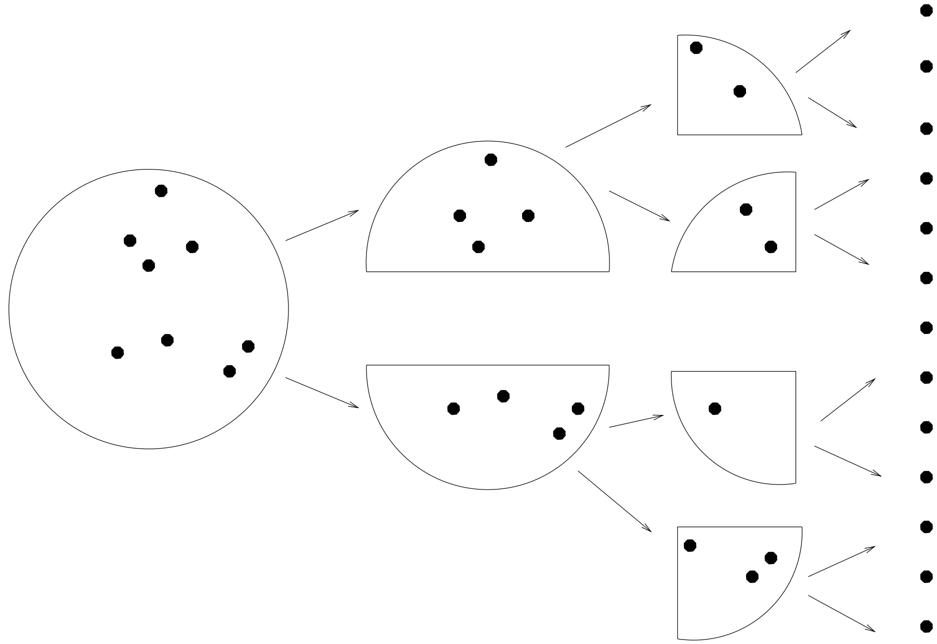
Hierarchical tree index structures

A sequence of refining partitions of the domain:



Hierarchical tree index structures

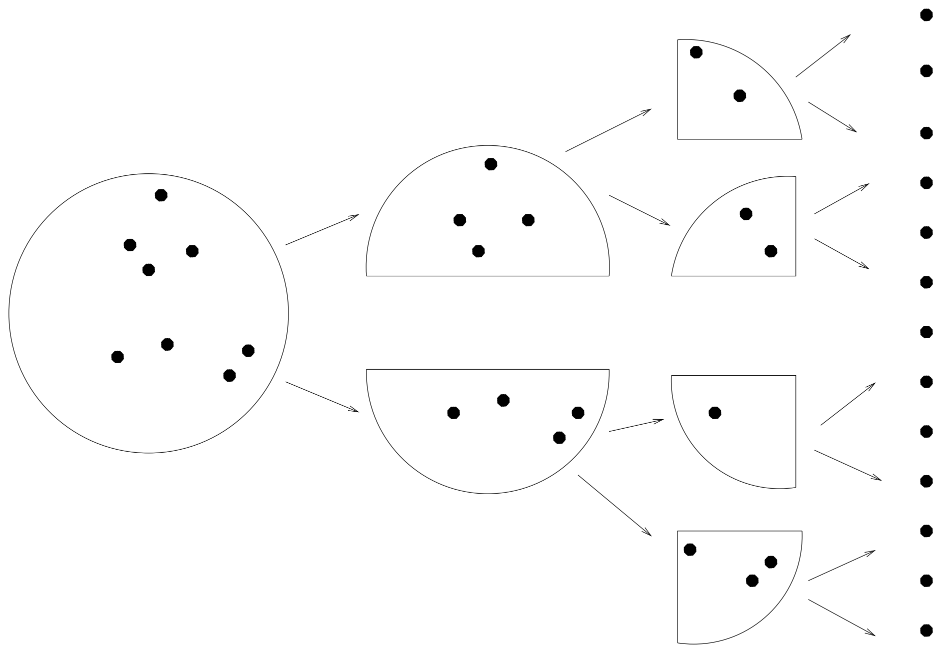
A sequence of refining partitions of the domain:



Space $O(n)$.

Hierarchical tree index structures

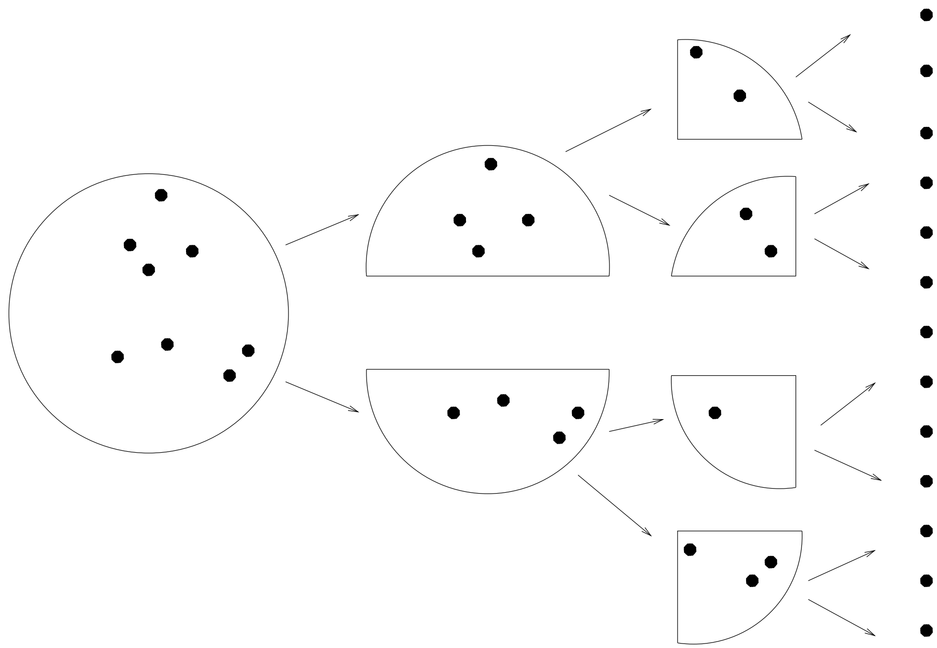
A sequence of refining partitions of the domain:



Space $O(n)$.
To process a range query $\mathcal{B}_\varepsilon(\omega)$, we traverse the tree all the way down to the leaf level.

Hierarchical tree index structures

A sequence of refining partitions of the domain:



Space $O(n)$.
To process a range query $\mathcal{B}_\varepsilon(\omega)$, we traverse the tree all the way down to the leaf level.

What happens in each node?

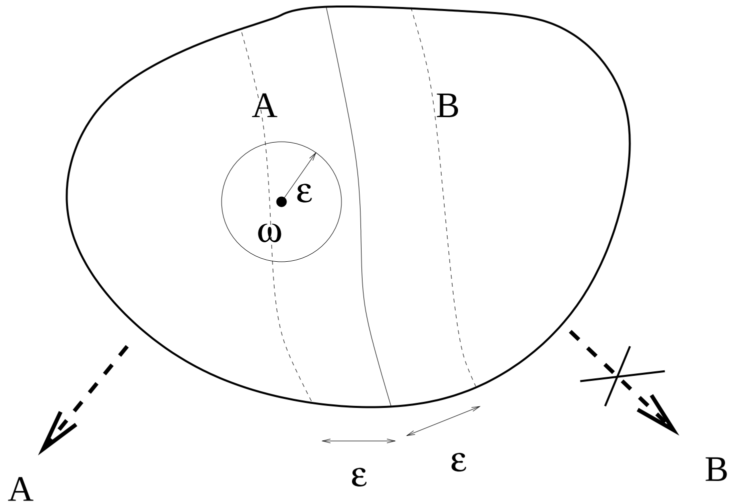
Pruning

Pruning

- If $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, the sub-tree descending from the node B can be pruned:

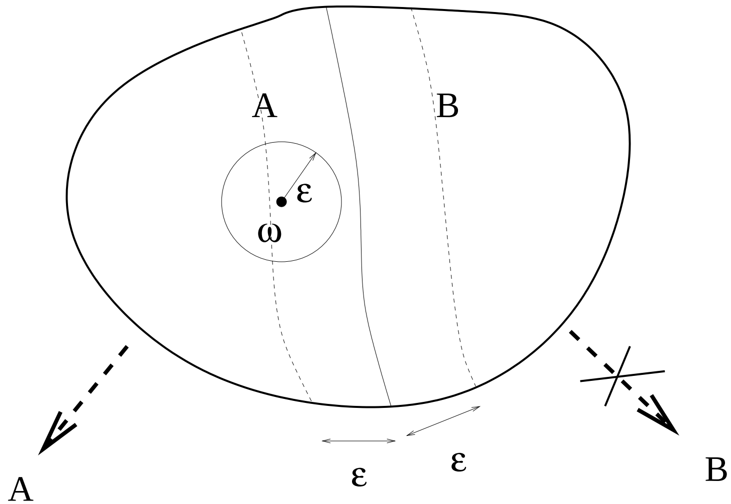
Pruning

- If $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, the sub-tree descending from the node B can be pruned:



Pruning

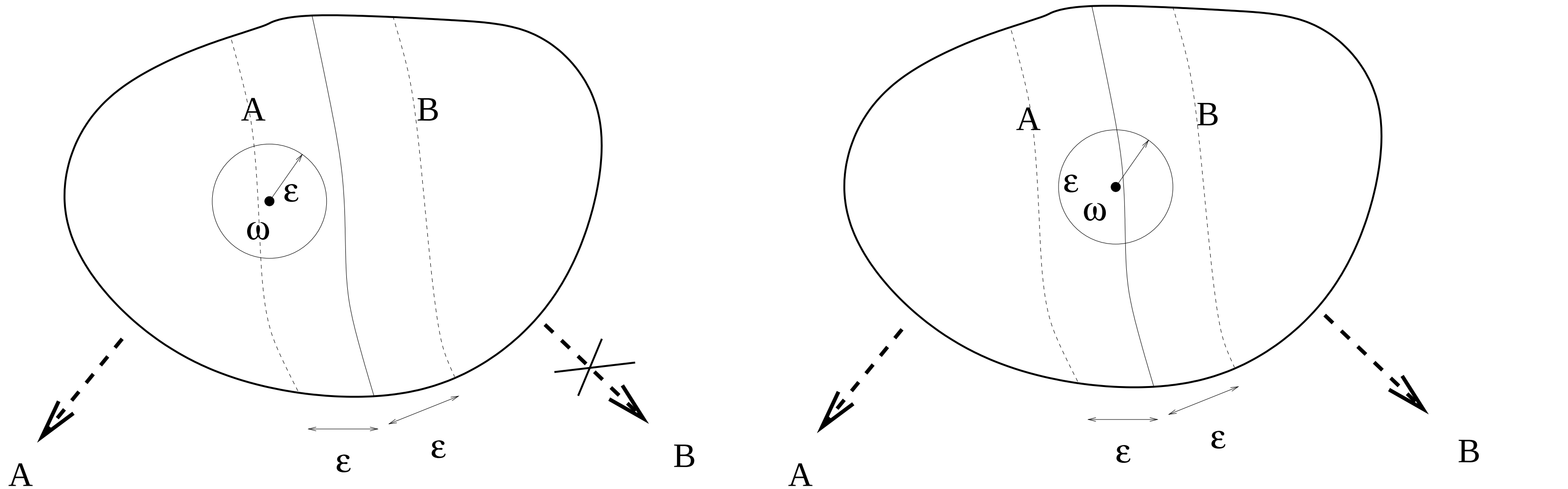
- If $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, the sub-tree descending from the node B can be pruned:



that is, if it can be certified that $\omega \notin B_\varepsilon = \{x \in \Omega \colon d(x, B) < \varepsilon\}$.

Pruning

- If $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, the sub-tree descending from the node B can be pruned:

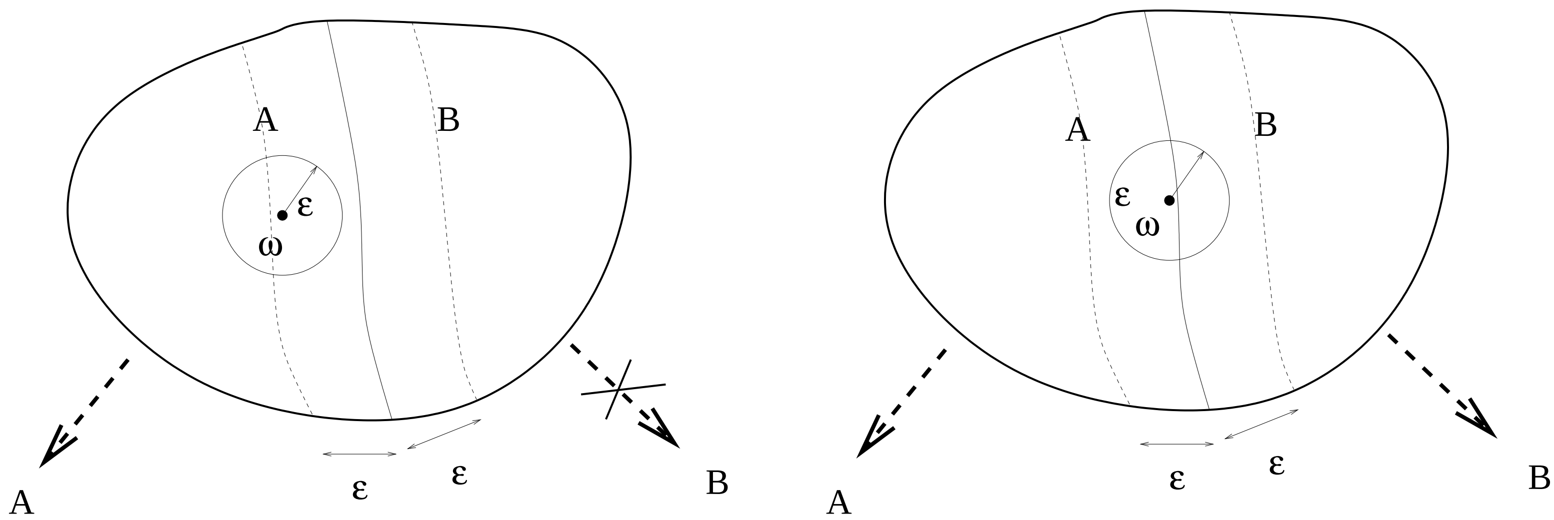


that is, if it can be certified that
$\omega \notin B_\varepsilon = \{x \in \Omega \colon d(x, B) < \varepsilon\}$.

- Otherwise the search branches out.

Pruning

- If $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, the sub-tree descending from the node B can be pruned:



that is, if it can be certified that
$\omega \notin B_\varepsilon = \{x \in \Omega \colon d(x, B) < \varepsilon\}$.

- Otherwise the search branches out.

How to “certify” that $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$?

Decision functions

Decision functions

Let $f: \Omega \to \mathbf{R}$ be a 1-Lipschitz function,

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

Decision functions

Let $f : \Omega \to \mathbf{R}$ be a 1-Lipschitz function,

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

such that $f \upharpoonright B \leq 0$.

Decision functions

Let $f\colon \Omega \to \mathbf{R}$ be a 1-Lipschitz function,

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

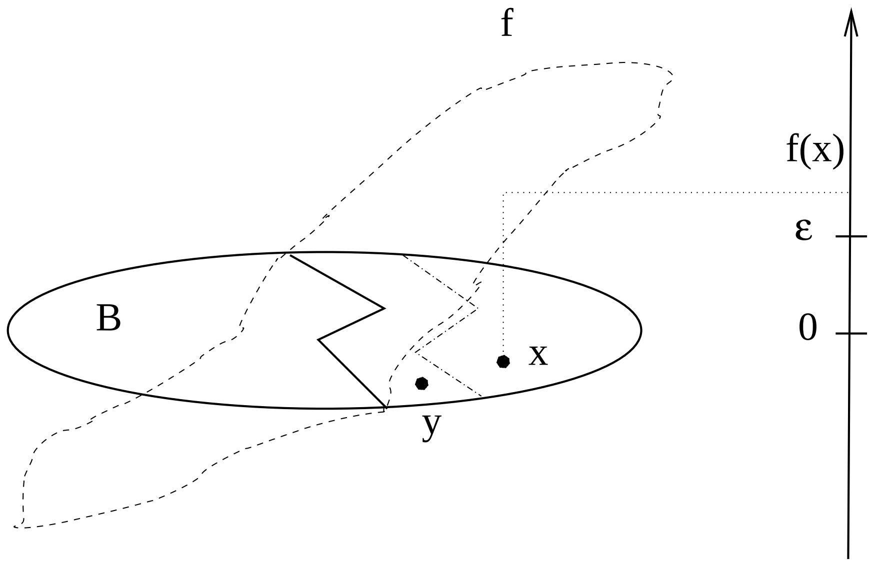
such that $f \upharpoonright B \leq 0$. Then $f \upharpoonright B_\varepsilon < \varepsilon$,

Decision functions

Let $f\colon \Omega \to \mathbf{R}$ be a 1-Lipschitz function,

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

such that $f \restriction B \leq 0$. Then $f \restriction B_\varepsilon < \varepsilon$,

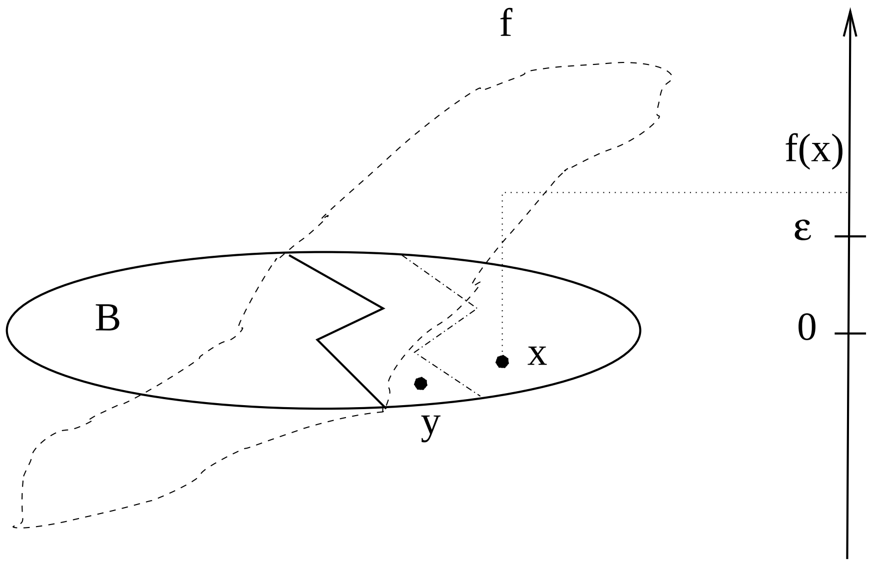


Decision functions

Let $f : \Omega \to \mathbf{R}$ be a 1-Lipschitz function,

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

such that $f \restriction B \leq 0$. Then $f \restriction B_\varepsilon < \varepsilon$,



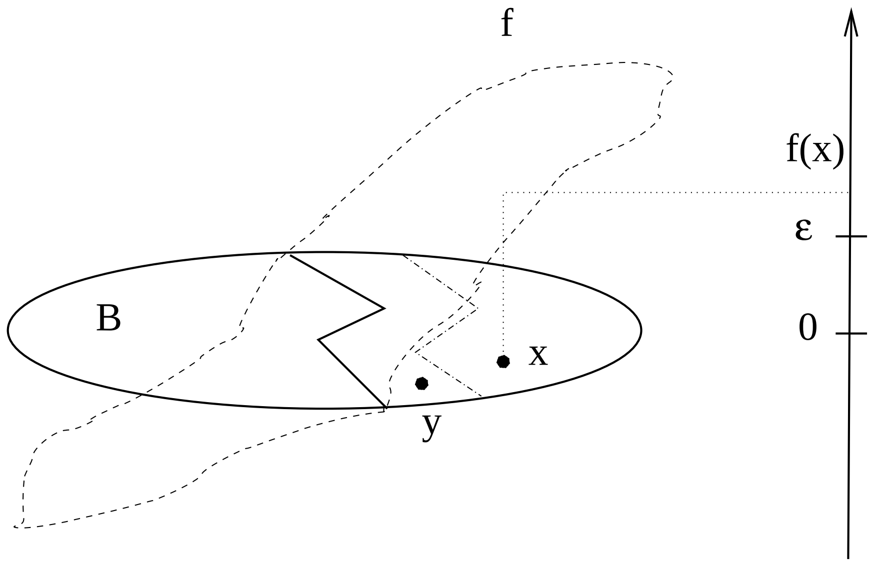
that is, $\boxed{f(\omega) \geq \varepsilon}$

Decision functions

Let $f\colon \Omega \to \mathbf{R}$ be a 1-Lipschitz function,

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

such that $f \restriction B \leq 0$. Then $f \restriction B_\varepsilon < \varepsilon$,



that is, $\boxed{f(\omega) \geq \varepsilon}$ is a certificate that $\boxed{\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset}$

Metric trees

A *metric tree* for a metric similarity workload (Ω, ρ, X):

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

such that

- $B_{root(T)} = \Omega$,
- $\forall$ inner node t and child nodes t_-, t_+, $B_t \subseteq B_{t_-} \cup B_{t_+}$.

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

- t_- [t_+] is accessed $\iff f_t(\omega) < \varepsilon$ [resp. $f_t(\omega) > -\varepsilon$].

What happens in practice?

What happens in practice?

The best indexing schemes for exact similarity search in high-dimensional *outer datasets* are often (not always!) outperformed by linear scan.

$$* * *$$

What happens in practice?

The best indexing schemes for exact similarity search in high-dimensional *outer datasets* are often (not always!) outperformed by linear scan.

$* * *$

The emphasis has shifted towards *approximate* similarity search:

What happens in practice?

The best indexing schemes for exact similarity search in high-dimensional *outer datasets* are often (not always!) outperformed by linear scan.

$$* * *$$

The emphasis has shifted towards *approximate* similarity search:

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

The curse of dimensionality conjecture

The curse of dimensionality conjecture

Conjecture.

The curse of dimensionality conjecture

Conjecture. Let $X \subseteq \{0,1\}^d$ be a dataset with n points, where the Hamming cube is equipped with the Hamming (ℓ^1) distance:

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

The curse of dimensionality conjecture

Conjecture. Let $X \subseteq \{0,1\}^d$ be a dataset with n points, where the Hamming cube 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 $d = \omega(\log n)$.

The curse of dimensionality conjecture

Conjecture. Let $X \subseteq \{0,1\}^d$ be a dataset with n points, where the Hamming cube 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 $d = \omega(\log n)$. Any data structure for exact nearest neighbour search in X,

The curse of dimensionality conjecture

Conjecture. Let $X \subseteq \{0,1\}^d$ be a dataset with n points, where the Hamming cube 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 $d = \omega(\log n)$. Any data structure for exact nearest neighbour search in X, with $d^{O(1)}$ query time,

The curse of dimensionality conjecture

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

$$* * *$$

The curse of dimensionality conjecture

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

$* * *$

The *cell probe model*: $\Omega(d/\log n)$ lower bound (Barkol–Rabani, 2000).

Concentration of measure

Concentration of measure

The *phenomenon of concentration of measure on high-dimensional structures* (*“Geometric LLN”*):

Concentration of measure

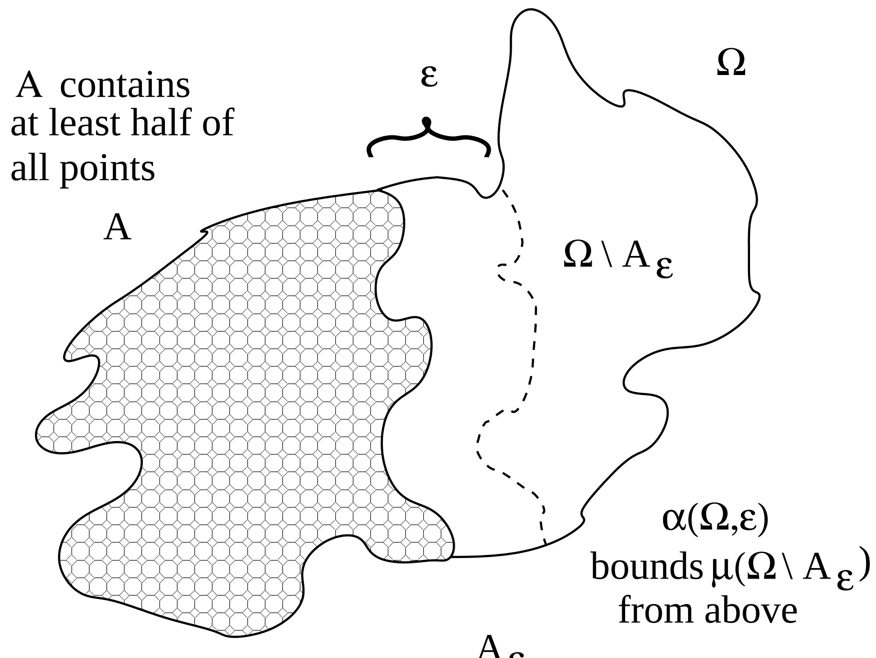
The *phenomenon of concentration of measure on high-dimensional structures* (*"Geometric LLN"*):

> 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 $\varepsilon > 0$.

Concentration of measure

The *phenomenon of concentration of measure on high-dimensional structures* (*“Geometric LLN”*):

> 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 $\varepsilon > 0$.



Concentration function

Concentration function

Let $\Omega = (\Omega, d, \mu)$ be a metric space with measure.

Concentration function

Let $\Omega = (\Omega, d, \mu)$ be a metric space with measure.
The concentration function of Ω:

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

Concentration function

Let $\Omega = (\Omega, d, \mu)$ be a metric space with measure.
The concentration function of Ω:

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

For $\Omega = \Sigma^n$, the Hamming cube (normalized distance + unif. measure):

$$\alpha_{\Sigma^n}(\varepsilon) \leq e^{-2\varepsilon^2 n}.$$

Concentration function

Let $\Omega = (\Omega, d, \mu)$ be a metric space with measure.
The concentration function of Ω:

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

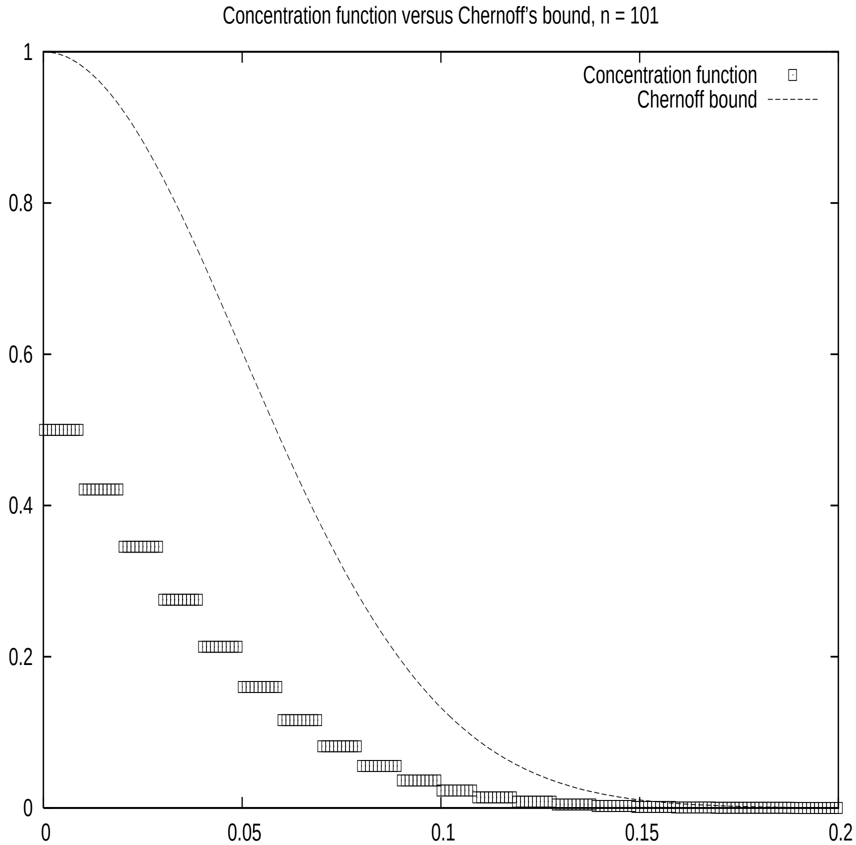
For $\Omega = \Sigma^n$, the Hamming cube (normalized distance + unif. measure):

$$\alpha_{\Sigma^n}(\varepsilon) \leq e^{-2\varepsilon^2 n}.$$

Gaussian estimates are typical

(Euclidean spheres $\mathbb{S}^n$, cubes $\mathbb{I}^n$, ...)

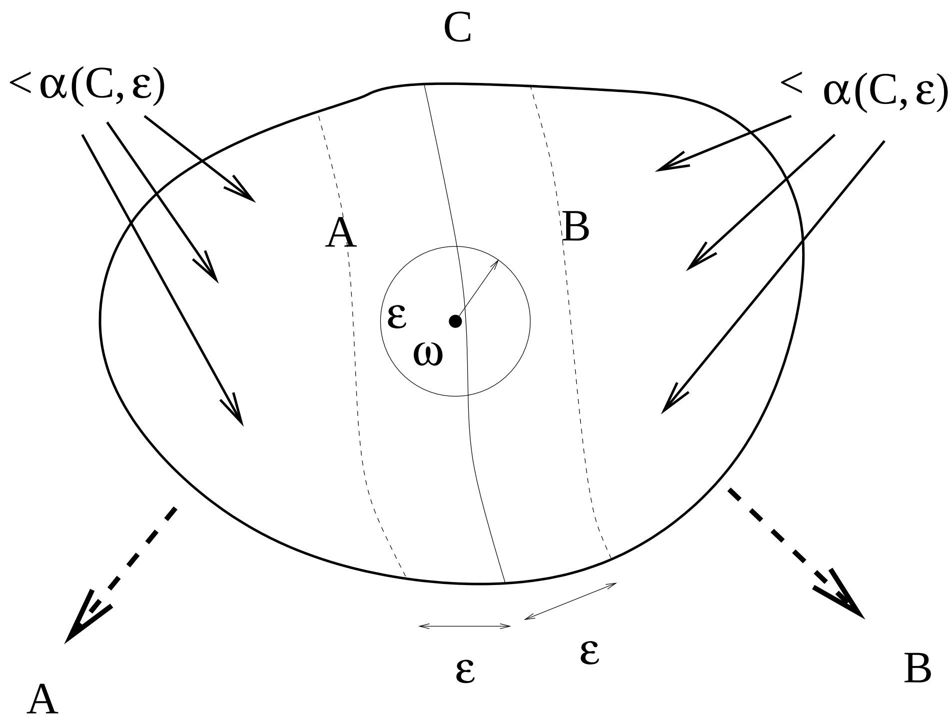
Example: the Hamming cube



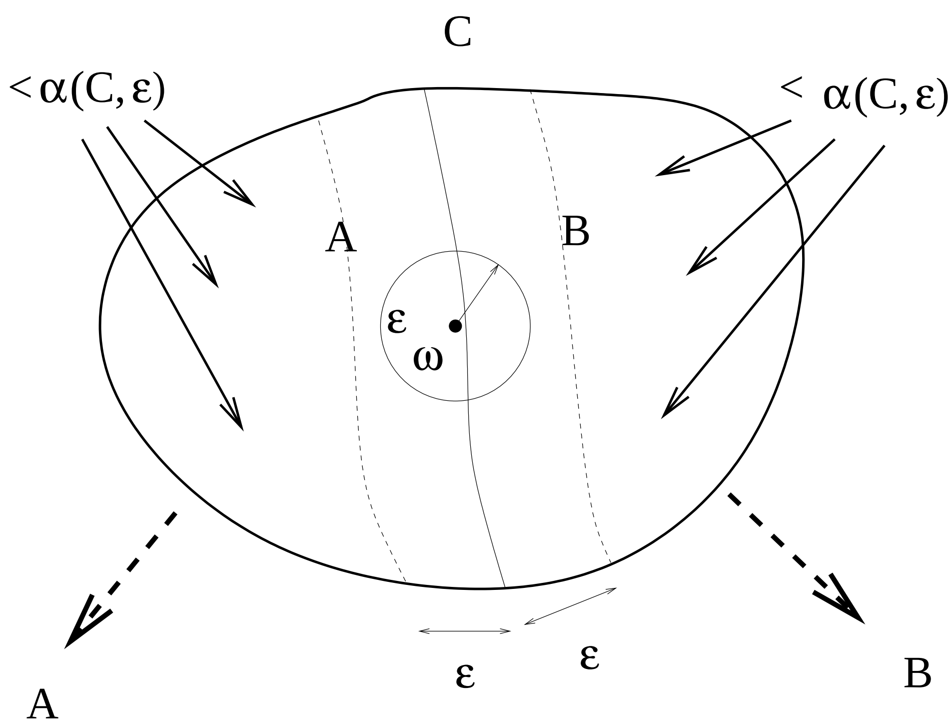
Concentration function $\alpha(\Sigma^{101}, \varepsilon)$ versus Chernoff bound

Effects of concentration on branching

Effects of concentration on branching



Effects of concentration on branching



For all query points $\omega \in C$ except a set of measure

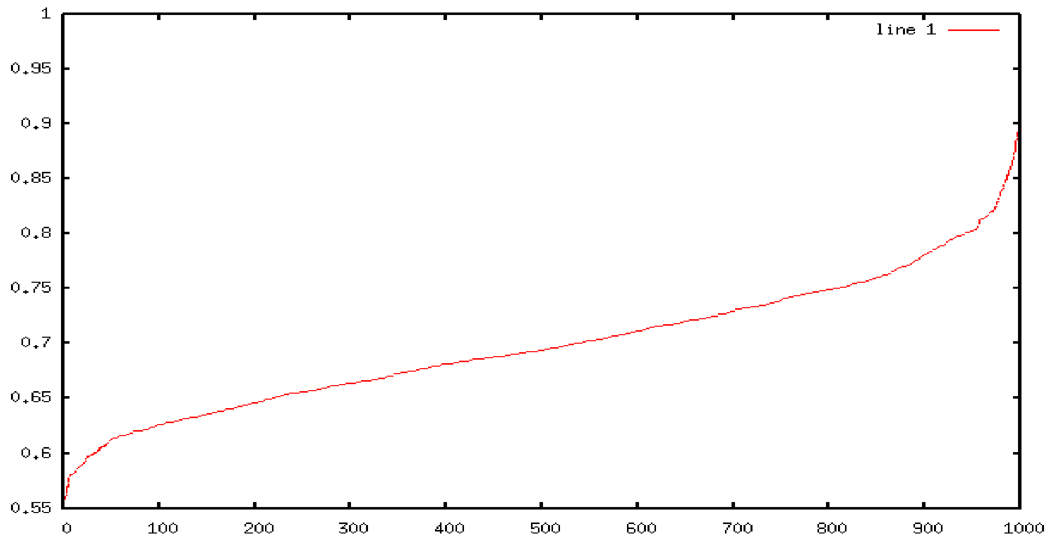
$$\leq 2\alpha(C, \varepsilon),$$

the search algorithm branches out at the node C.

Search radius

- $\varepsilon_{NN}(\omega)$ is a 1-Lipschitz function, so concentrates near the median value, ε_M;
- $\varepsilon_M \to \mathbb{E}_{\mu\otimes\mu} d(x,y) = O(1)$.

Example: 1000 pts $\sim [0,1]^{10}$, the ℓ^2-ε_{NN}:



$$\varepsilon_M = 0.69419 \qquad \mathbb{E}d(x,y) = 1.2765.$$

A naive average $O(n)$ lower bound

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...
...as well as query points.

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...
...as well as query points.
A balanced metric tree of depth $O(\log n)$, with $O(n)$ bins of roughly equal size (μ-measure).

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...
...as well as query points.
A balanced metric tree of depth $O(\log n)$, with $O(n)$ bins of roughly equal size (μ-measure).
in $1/2$ the cases, $\varepsilon_{NN} \geq \varepsilon_M = O(1)$, the median NN dist.

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...
...as well as query points.
A balanced metric tree of depth $O(\log n)$, with $O(n)$ bins of roughly equal size (μ-measure).
in $1/2$ the cases, $\varepsilon_{NN} \geq \varepsilon_M = O(1)$, the median NN dist.
For every element A of level t partition,

$$\alpha(A, \varepsilon_M) \leq 2\mu(A)^{-1}\alpha(\Omega, \varepsilon_M/2) = O(2^t)e^{-O(1)\varepsilon_M^2 d}.$$

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...
...as well as query points.
A balanced metric tree of depth $O(\log n)$, with $O(n)$ bins of roughly equal size (μ-measure).
in $1/2$ the cases, $\varepsilon_{NN} \geq \varepsilon_M = O(1)$, the median NN dist.
For every element A of level t partition,

$$\alpha(A, \varepsilon_M) \leq 2\mu(A)^{-1}\alpha(\Omega, \varepsilon_M/2) = O(2^t)e^{-O(1)\varepsilon_M^2 d}.$$

$\rightsquigarrow$ branching at every node occurs for all ω except

A naive average $O(n)$ lower bound

Suppose datapoints are distributed according to $\mu \in P(\Omega)$...
...as well as query points.
A balanced metric tree of depth $O(\log n)$, with $O(n)$ bins of roughly equal size (μ-measure).
in $1/2$ the cases, $\varepsilon_{NN} \geq \varepsilon_M = O(1)$, the median NN dist.
For every element A of level t partition,

$$\alpha(A, \varepsilon_M) \leq 2\mu(A)^{-1}\alpha(\Omega, \varepsilon_M/2) = O(2^t)e^{-O(1)\varepsilon_M^2 d}.$$

$\rightsquigarrow$ branching at every node occurs for all ω except

$$\sharp(\text{nodes}) \times 2\sup_A \alpha(A, \varepsilon) = O(n^2)e^{-O(1)d} = o(1),$$

because $d = \omega(\log n)$, $\rightsquigarrow e^{-O(1)d}$ is superpoly(n).

What's wrong?

What's wrong?

A dataset X is modeled by a sequence of i.i.d. r.v. $X_i \sim \mu$.

What's wrong?

A dataset X is modeled by a sequence of i.i.d. r.v. $X_i \sim \mu$.

Implicit assumption: empirical measure $\mu_n(A) = \frac{|A|}{n} \approx \mu(A)$.

What's wrong?

A dataset X is modeled by a sequence of i.i.d. r.v. $X_i \sim \mu$.

Implicit assumption: empirical measure $\mu_n(A) = \frac{|A|}{n} \approx \mu(A)$.

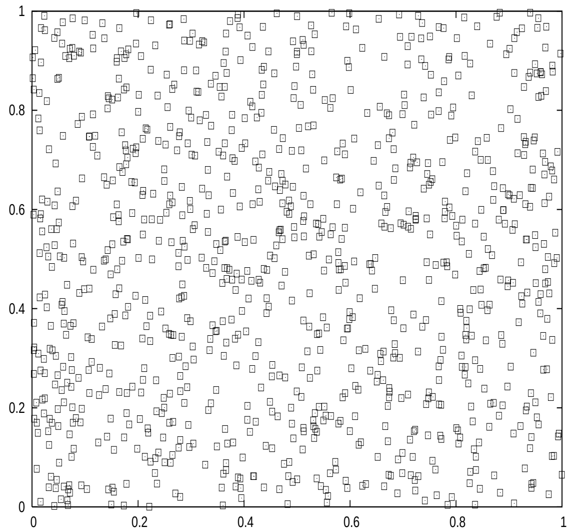
But the scheme is chosen *after* seeing an instance X!

What's wrong?

A dataset X is modeled by a sequence of i.i.d. r.v. $X_i \sim \mu$.

Implicit assumption: empirical measure $\mu_n(A) = \frac{|A|}{n} \approx \mu(A)$.

But the scheme is chosen *after* seeing an instance X!



How much can be said of concentration in (Ω, μ_n)?

VC dimension

VC dimension

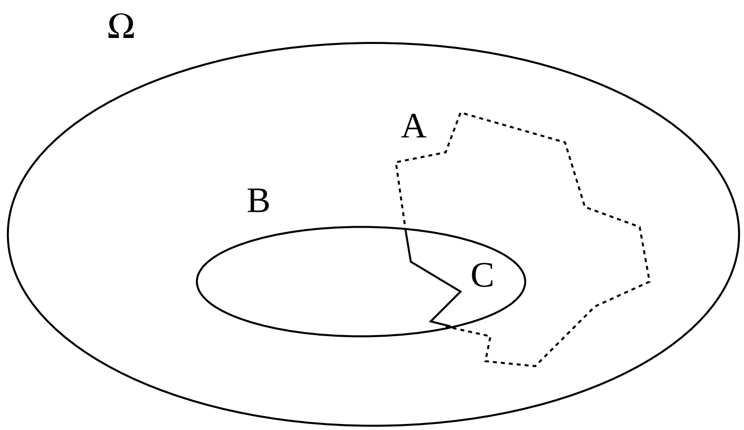
Let $\mathscr{A}$ be a family of subsets of Ω (a *concept class*).
$B \subseteq \Omega$ is *shattered* by $\mathscr{A}$ if for each $C \subseteq B$ there is $A \in \mathscr{A}$ such that

$$A \cap B = C.$$

VC dimension

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

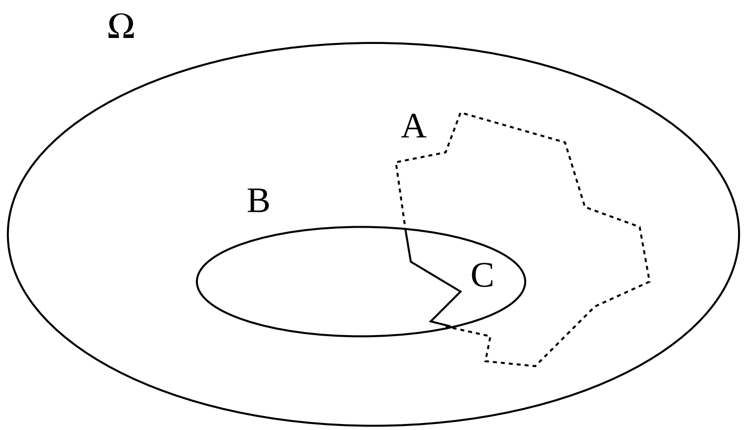
$$A \cap B = C.$$



VC dimension

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

$$A \cap B = C.$$



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

Statistical learning bounds

Statistical learning bounds

Let $\mathscr{A} \subseteq 2^{\Omega}$ be a concept class of finite VC dimension, d.

Statistical learning bounds

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,

Statistical learning bounds

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 $1 - \delta$

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

Statistical learning bounds

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:

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

Bin access lemma

Bin access lemma

Let $\delta > 0$, and let γ be a collection of subsets $A \subseteq \Omega$ of measure $\mu(A) \leq \alpha(\delta) \leq \frac{1}{4}$ each, satisfying $\mu(\cup\gamma) \geq 1/2$.

Bin access lemma

Let $\delta > 0$, and let γ be a collection of subsets $A \subseteq \Omega$ of measure $\mu(A) \leq \alpha(\delta) \leq \frac{1}{4}$ each, satisfying $\mu(\cup\gamma) \geq 1/2$. Then the 2δ-neighbourhood of every point $\omega \in \Omega$, apart from a set of measure at most $\frac{1}{2}\alpha(\delta)^{\frac{1}{2}}$, meets at least $\lceil \frac{1}{2}\alpha(\delta)^{-\frac{1}{2}} \rceil$ elements of γ.

$* * *$

Bin access lemma

Let $\delta > 0$, and let γ be a collection of subsets $A \subseteq \Omega$ of measure $\mu(A) \leq \alpha(\delta) \leq \frac{1}{4}$ each, satisfying $\mu(\cup\gamma) \geq 1/2$. Then the 2δ-neighbourhood of every point $\omega \in \Omega$, apart from a set of measure at most $\frac{1}{2}\alpha(\delta)^{\frac{1}{2}}$, meets at least $\lceil \frac{1}{2}\alpha(\delta)^{-\frac{1}{2}} \rceil$ elements of γ.

$* * *$

If we can now guarantee that the bins are not too large, we get a lower bound on the number of bin accesses.

Bin complexity estimates

Bin complexity estimates

Let $\mathscr{F}$ be a class of 1-Lipschitz functions used for constructing a metric tree of a particular type.

Bin complexity estimates

Let $\mathscr{F}$ be a class of 1-Lipschitz functions used for constructing a metric tree of a particular type.

Let $\mathscr{A}$ be the concept class of all solution sets to inequalities

$$f \mathop{\gtreqless} a, \quad f \in \mathscr{F}, \quad a \in \mathbf{R}.$$

Bin complexity estimates

Let $\mathscr{F}$ be a class of 1-Lipschitz functions used for constructing a metric tree of a particular type.

Let $\mathscr{A}$ be the concept class of all solution sets to inequalities

$$f \gtreqless a, \quad f \in \mathscr{F}, \quad a \in \mathbf{R}.$$

Suppose

$$p = \text{VC-dim}\,(\mathscr{A}) < \infty$$

(*pseudodimension of $\mathscr{F}$ in the sense of Vapnik*).

Bin complexity estimates

Let $\mathscr{F}$ be a class of 1-Lipschitz functions used for constructing a metric tree of a particular type.

Let $\mathscr{A}$ be the concept class of all solution sets to inequalities

$$f \underset{<}{\geq} a, \quad f \in \mathscr{F}, \quad a \in \mathbf{R}.$$

Suppose

$$p = \text{VC-dim}\,(\mathscr{A}) < \infty$$

(*pseudodimension of $\mathscr{F}$ in the sense of Vapnik*).

Denote $\mathscr{B}$ the class of all bins of all possible metric trees of depth $\leq h$ built using $\mathscr{F}$. Then

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

Rigorous lower bounds

Rigorous lower bounds

thm. Let $\mathscr{F}$ be a class of 1-Lipschitz functions on $\{0,1\}^d$ with VC dimension of the class of sets given by inequalities $f \gtreqless a$ being poly(d).

Rigorous lower bounds

thm. Let $\mathscr{F}$ be a class of 1-Lipschitz functions on $\{0,1\}^d$ with VC dimension of the class of sets given by inequalities $f \substack{\geq \\ <} a$ being poly(d).
With probability approaching 1, every metric tree indexing scheme for a random sample X of $\{0,1\}^d$ containing n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$,

Rigorous lower bounds

thm. Let $\mathscr{F}$ be a class of 1-Lipschitz functions on $\{0,1\}^d$ with VC dimension of the class of sets given by inequalities $f \gtreqless a$ being poly(d).
With probability approaching 1, every metric tree indexing scheme for a random sample X of $\{0,1\}^d$ containing n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$, will have the worst-case performance $d^{\omega(1)}$.

Rigorous lower bounds

thm. Let $\mathscr{F}$ be a class of 1-Lipschitz functions on $\{0,1\}^d$ with VC dimension of the class of sets given by inequalities $f \gtreqless a$ being poly(d).
With probability approaching 1, every metric tree indexing scheme for a random sample X of $\{0,1\}^d$ containing n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$, will have the worst-case performance $d^{\omega(1)}$.

$\triangleleft$ Can suppose every bin contains poly(d) datapoints, and the tree depth is poly(d). The VC-dim of all possible bins is poly$(d) = o(n)$. If $\epsilon = n^{1/2-\gamma}$, by learning estimates the measure of each bin of the scheme is $O(n^{-1/2+\gamma})$, so there will be $\Omega(n^{1/4-\gamma}) = d^{\omega(1)}$ bin accesses. $\triangleright$

Example: vp-tree

The *vp-tree* (Yianilos) 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.

If $\Omega = \mathbf{R}^d$, VC dimension is $d + 1$.

Example: M-tree

The *M-tree* (Ciaccia, Patella, Zezula) 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.

If $\Omega = \mathbf{R}^d$, VC-dim is $d+1$; for $\Omega = \{0,1\}^d$, it is $O(d)$.