Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

Frederic.Cazals@inria.fr
Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Nearest neighbors: on the importance of locality

Typical settings:
- Regression – estimating a response variable from neighbors
- Supervised classification using neighbors
- Manifold / shape learning: learning a mathematical model for the data (e.g. simplicial complex)

Samples used at a given location $q$:
- nearest neighbors
- points in a cell of a spatial partition e.g. a RPTree

Key questions:
- robustness of neighbors via random projections in RPTree
- locality assessed by the diameter of cells in RPTree
Intermezzo: data and their intrinsic dimension (I)

- **Intrinsic dimension**: in many real world problems, features may be correlated, redundant, causing data to have low *intrinsic dimension*, i.e., data lies close to a low-dimensional manifold.

- Example: binary ie B&W image
  - Consider an $n \times n$ binary image: image $\sim$ point on the hypercube of dimension $n^2$.

- Example: rotating an image
  - Consider an $n \times n$ pixel image, with each pixel encode in the RGB channels: 1 image $\sim$ on point in dimension $d = 3n^2$.
  - Consider $N$ rotated versions of this image: $N$ point in $\mathbb{R}^{3n^2}$.
  - But these points intrinsically have one degree of freedom (that of the rotation).
Intermezzo: data and their intrinsic dimension (II)

▷ Example: 2D robotic arm with 3 d.o.f.

▷ Example: human body motion capture
  ▷ N markers attached to body (typically N=100).
  ▷ each marker measures position in 3 dimensions, 3N dimensional feature space.
  ▷ But motion is constrained by a dozen-or-so joints and angles in the human body.

▷ Ref: Verma et al. Which spatial partitions are adaptive to intrinsic dimension? UAI 2009
Local covariance dimension and its multi-scale estimation

▷ Def.: a set $T \subset \mathbb{R}^D$ has covariance dimension $(d, \epsilon)$ if the largest $d$ eigenvalues of its covariance matrix satisfy

$$\sigma_1^2 + \cdots + \sigma_d^2 \geq (1 - \epsilon) \cdot (\sigma_1^2 + \cdots + \sigma_D^2).$$

▷ Def.: Local covariance dimension with parameters $(d, \epsilon, r)$: the previous must hold when restricting $T$ to balls of radius $r$.

▷ Multi-scale estimation from a point cloud $P$:

For each datapoint $p$ and each scale $r$

- Collect samples in $B(x, r)$
- Compute covariance matrix
- Check how many eigenvalues are required: yields the dimension
Assouad dimension

▷ Def: Set $S \subset \mathbb{R}^D$ has Assouad dimension $\leq d$: for any ball $B$, subset $S \cap B$ can be covered by $2^d$ balls of half the radius. Also called doubling dimension.

▷ Examples:
  - $S =$ line: Assouad dimension $= 1$
  - $S =$ k-dimensional affine subspace: Assouad dimension $= O(k)$
  - $S =$ set of $N$ points: Assouad dimension $\leq \log N$
  - $S =$ k-dim submanifold of $\mathbb{R}^D$ with finite condition number: Assouad dimension $= O(k)$ in small enough neighborhoods

▷ Rmk: if $S$ has Assouad dim $\leq d$, so do subsets of $S$

▷ Hardness: computing doubling dimensions and constants is generally hard: related to packing problems.
Doubling dimension and doubling measures

- **Def.:** A metric space $X$ with metric is called *doubling* if any ball $B(x, r)$ with $x \in X$ and $r > 0$ is contained in at most $M$ balls of radius $r/2$. The *doubling dimension* is $\log_2 M$.

- **Def.:** A measure $\mu$ on a metric space $X$ is called *doubling* if

  $$\mu(B(x, 2r)) \leq C \mu(B(x, r)).$$

  Equivalently, $\exists d_0$ such that $\forall \alpha > 0$:

  $$\mu(B(x, 2r)) \leq \alpha^{d_0} \mu(B(x, r)).$$

  The *dimension* of the doubling measure satisfies $d_0 = \log_2 C$.

- **Remarks:**
  - A measure space supporting a doubling measure is necessarily a doubling metric space, with dimension depending on $C$.
  - Conversely, any complete doubling metric space supports a doubling measure.
Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Empirical results: contenders

- **Contenders / algorithms:**
  - dyadic trees aka tries: pick a direction and split at the midpoint; cycle through coordinates.
  - kd-tree: split at median along direction with largest spread.
  - random projection trees: split at the median along a random direction.
  - PD / PCA trees: split at the median along the principal eigenvector of the covariance matrix.
  - two means trees: solve the 2-means; pick the direction spanned by the centroids, and split the data as per cluster assignment.

- **dyadic trees, kd-trees, RP trees**
Real word datasets

- **Datasets:**
  - Swiss roll
  - Teapot dataset: rotated images of a teapot (1 B&W image: 50x30 pixels); thus, 1D dataset in ambient dimension 1500.
  - Robotic arm: dataset in $\mathbb{R}^{12}$; yet, robotic arm has 2 joints: (noisy) 2D dataset in ambient dimension 12.
  - 1 from the MNIST OCR dataset; 20x20 B&W images, i.e. points in ambient dimension 400.
  - *Love* cluster from Australian Sign Language time-series
  - *aw* phoneme from MFCC TIMIT dataset

Empirical results: local covariance dimension estimation

- **Conventions:** bold lines: estimate $d(r)$; dashed lines: std dev; numbers: ave. num of samples in balls of the given radius

- **Observations:**
  - Swiss roll (ambient space dim is 3): failure at small (noise dominates) and large scales (sheets get blended).
  - Teapot: clear small dimensional structure at low scale, but rather 3-4 than 1.

**Ref:** N. Verma, S. Kpotufe, and S. Dasgupta, 2009
Empirical results: performance for NN searches

- Searching \( p(1) \): performance is the order of the NN found / dataset size
  - percentile order: order of NN found / dataset size (the smaller the better)
  - tree depth: NN sought at each level in the tree
  - decorating numbers: distance ratio \( \| q - \text{nn}(q) \| / \| q - p(1) \| \)

Observations:
  - percentile order deteriorates with depth – separation does occur
  - yet, the distance ratio remains small even at high percentile orders
  - 2M and PD (i.e. PCA trees) consistently yield better nearest neighbors: better adaptation to the intrinsic dimension

Empirical results: regression

Regression:
- one predicts the rotation angle (response variable) from the average values found in the cell containing the query point
- performance is $l_2$ error on the response variable
- theory says that best results are expected for data structure adapting to the intrinsic dimension

Observations:
- Small tree depth: averaging over many neighbors is detrimental
- Best results for 2M trees, PD (i.e., PCA) trees, and RP trees.

References


Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Recap:

- Points iteratively projected on random directions
- Risks jeopardizing the search strategy: points far away (from the NN) squeeze in-between \( q \) and \( nn(q) \)
- Hardness of the NN search: function \( \Phi \)

\[
\Phi(q, P) = \frac{1}{n} \sum_{i=2}^{n} \frac{\|q - x^{(1)}\|_2}{\|q - x^{(i)}\|_2}.
\]
Projections on random directions for separation

Separation property fails in using coordinate axis (kd-trees)

▷ Consider the following point set \( \{x_1, \ldots, x_n\} \):

- \( x_1 \): the all-ones vector

- For each \( x_i, i > 1 \): pick a random coord and set it to a large value \( M \); set the remaining coords to uniform random numbers is \((0, 1)\)

kd-trees separate \( q \) and \( p_{(1)} \), even though function \( \Phi \) is arbitrarily small:

- The NN of \( q \) (=origin) is \( x_1 \)

- But by growing \( M \), function \( \Phi \) gets close to 0 \( \Rightarrow \) random projections will work well

- However, any coord. projection separates \( q \) and \( x_1 \): on average, the fraction of points falling in-between \( q \) and \( x_1 \) is arbitrarily large:

\[
\frac{1}{n} \left( n - \frac{n}{d} \right) = 1 - \frac{1}{d}
\]
Demo with DrGeo

Compulsory tools for geometers

- In the sequel: Consider 3 points \( q, x, y \) with \( \| q - x \| \leq \| q - y \| \).
- In projection on a random direction \( U \): probability to have the projection of \( y \) nearest to \( q \) than the projection of \( x \)?

- DrGeo: http://www.drgeo.eu/

NB: also of interest: IPE, http://ipe.otfried.org/
Random projections: relative position of three points

In the sequel: \( q, x, y \): 3 points with \( \|q - x\| \leq \|q - y\| \)

Colinearity index \( q, x, y \):

\[
\text{coll}(q, x, y) = \frac{<q - x, y - x>}{\|q - x\| \|y - x\|}
\]  \( \tag{2} \)

Event E: \( <y, U> \) falls strictly in-between \( <q, U> \) and \( <x, U> \)

Lemma 1. Consider \( q, x, y \in \mathbb{R}^d \) and \( \|q - x\| \leq \|q - y\| \). The proba. over random directions \( U \), of \( E \), satisfies:

\[
\mathbb{P}[E] = \frac{1}{\pi} \arcsin \left( \frac{\|q - x\|}{\|q - y\|} \sqrt{1 - \text{coll}(q, x, y)^2} \right)
\]  \( \tag{3} \)

Corollary 2.

\[
\frac{1}{\pi} \frac{\|q - x\|}{\|q - y\|} \sqrt{1 - \text{coll}(q, x, y)^2} \leq \mathbb{P}[E] \leq \frac{1}{2} \frac{\|q - x\|}{\|y - x\|}
\]  \( \tag{4} \)
Proof of the corollary

Proof of the corollary

Using the Inequality:

\[ \theta \in [0, \pi/2] : \frac{2\theta}{\pi} \leq \sin \theta \leq \theta \quad (5) \]

Lower bound of the corr.: from the upper bound of Eq. (5): \( \theta \leq \arcsin \theta \) applied to \( \mathbb{P}[E] \)

Upper bound of the corr.:
First note that:

\[
\frac{\|q - x\|}{\|q - y\|} \sqrt{1 - \text{coll}(q, x, y)^2} \leq \frac{\|q - x\|}{\|q - y\|}
\]

Then, apply \((2\phi/\pi) \leq \phi\) to \( \phi = \arcsin \|q - x\| / \|q - y\| \).
Random projections: separation of neighbors

Theorem 3. Consider \( q, p_1, \ldots, p_n \in \mathbb{R}^d \), and a random direction \( U \).
The expected fraction of the projected \( p_i \) that fall between \( q \) and \( p^\top) \) is at most

\[
\frac{1}{2} \Phi(q, P).
\]

\( \triangleright \) **Proof.** Let \( Z_i \) be the event: “\( p^\top(i) \) falls between \( q \) and \( p^\top(1) \) in the projection”.
By the corollary 2, \( \mathbb{P}[Z_i] \leq (1/2) \frac{\|q - p^\top(1)\|}{\|q - p^\top(i)\|} \). Then, apply the linearity of expectation to \( \sum Z_i/n \) (divide by \( n \) to get the fraction).

Theorem 4. Let \( S \subseteq P \) with \( p^\top(1) \in S \). If \( U \) is chosen uniformly at random, then \( 0 < \alpha < 1 \), the probability that an \( \alpha \) fraction of the projected points in \( S \) fall between \( q \) and \( p^\top(1) \) is bounded by

\[
\frac{1}{2\alpha} \Phi_{|S|}(q, P).
\]

\( \triangleright \) **Proof.** Previous Thm + Markov’s inequality.
Regular spill trees—i.e. redundant storage

Recap:

- Storage: point possibly stored twice using overlapping split with parameter $\alpha$; depth is $O(\log n/n_0)$
- Query routing: routing to a single leaf

Theorem 5. Let $\beta = 1/2 + \alpha$. The error probability is:

$$\Pr[\text{Err}] \leq \frac{1}{2\alpha} \sum_{i=0,\ldots,l} \Phi_{\beta_i n}(q, P)$$

Proof, steps:

- Internal node at depth $i$ contains $\beta^i n$ points
- For such a node: proba to have $q$ separated from $p(1)$ transmitted to one side of the split $\Rightarrow$ a fraction $\alpha$ of the points of the cell fall between $q$ and the median $m \Rightarrow$ a fraction $\alpha$ of the points of the cell fall between $q$ and $p(1)$: this occurs with proba $(1/2\alpha)\Phi_{\beta_i n}(q, P)$
- To conclude: union-bound over all levels $i$
Virtual spill trees

- Recap:
  - Storage: each point stored in a single leaf with median splits; depth is $O(\log n/n_0)$
  - Query routing: with overlapping splits of parameter $\alpha$

Theorem 6. Let $\beta = 1/2$. The error probability is:

$$P[Err] \leq \frac{1}{2\alpha} \sum_{i=0,...,l} \Phi_{\beta^i n}(q, P)$$

- Proof, mutatis mutandis:
  - Consider the path root - leaf of $p_{(1)}$
  - For a level, bound the proba. to have $q$ routed to one side only
  - Add up for all levels
Random projection trees

Recap:

- Pick a random direction and project points onto it
- Split at the $\beta$ fractile for $\beta \in (1/4, 3/4)$
- Storage: each point mapped to a single leaf
- Query routing: query point mapped to a single leaf too

Theorem 7. Consider an RP tree for $P$. Define $\beta = 3/4$, and $l = \log_{1/\beta}(n/n_0)$. One has:

$$\mathbb{P}\left[\text{NN query does not return } p_{(1)}\right] \leq \sum_{i=0,\ldots,l} \Phi_{\beta^i n} \ln \frac{2e}{\Phi_{\beta^i n}}$$

Proof, key steps:

- $F$: fraction of points separating $q$ and $p_{(1)}$ in projection
- Since split chosen at random in interval of mass $1/2$: it separates $q$ and $p_{(1)}$ with proba. $F/(1/2)$
- Integrating yields the result for one level; then, union bound.
Theorem 8. (Spill trees) Consider a spill tree of depth \( l = \log \frac{1}{1/\beta} \left( \frac{n}{n_0} \right) \), with

- \( \beta = \frac{1}{2} + \alpha \) for regular spill trees,
- and \( \beta = \frac{1}{2} \) for virtual spill trees.

If this tree is used to answer a query \( q \), then:

\[
P[\text{Err}] \leq \frac{1}{2\alpha} \sum_{i=0,...,l} \Phi_{\beta_i n}(q, P) \tag{9}
\]

Nb: \( \beta_i n \): number of data points found in an internal node at depth \( i \)
Bounding function $\Phi$ in specific settings

Improving the bound $\Phi \leq 1$

▷ Perspective: assume that $x_1, \ldots, x_n$ are drawn i.i.d. from a doubling measure. Can this regularity be used?

**Theorem 9.** Let $\mu$ be a continuous measure on $\mathbb{R}^d$, a doubling measure of dimension $d_0 \geq 2$. Assume $p_1, \ldots, p_n \sim \mu$. With probability $\geq 1 - 3\delta$, for all $2 \leq m \leq n$:

$$\Phi_m(q, P) \leq 6 \left( \frac{2}{m} \ln \frac{1}{\delta} \right)^{1/d_0}.$$

**Theorem 10.** Under the same hypothesis:

– For both variants of the spill trees:

$$\mathbb{P}[Err] \leq \frac{c_0 k d_0}{\alpha} \left( \frac{8 \max(k, \ln 1/\delta)}{n_0} \right)^{1/d_0}$$

– For random projection trees with $n_0 \geq c_0 (3k)^{d_0} \max(k, \ln 1/\delta)$:

$$\mathbb{P}[Err] \leq c_0 k (d_0 + \ln n_0) \left( \frac{8 \max(k, \ln 1/\delta)}{n_0} \right)^{1/d_0}$$

▷ Rmk: failure proba. can be made arbitrarily small by taking $n_0$ large enough.
References


**VKD09** N. Verma, S. Kpotufe, S. Dasgupta, Which spatial partitions are adaptive to intrinsic dimension? UAI 2009.
Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Partitioning rules that adapt to intrinsic dimension

- **Principal component analysis**: split the data at the median along the principal direction of covariance.
  - Drawback 1: estimation of principal component requires a significant amount of data and only about $\frac{1}{2^l}$ fraction of data remains at a cell at level $l$
  - Drawback 2: computationally too expensive for some applications

- **2-means i.e. solution of $k$-means with $k = 1$**: compute the 2-means solution, and split the data as per the cluster assignment
  - Drawback 1: 2-means is an NP-hard optimization problem
  - Drawback 2: the best known $(1 + \epsilon)$-approximation algorithm for 2-means (A. Kumar, Y. Sabharwal, and S. Sen, 2004) would require a prohibitive running time of $O(2^{d^{O(1)}} Dn)$, since we need $\epsilon \approx 1/d$.
    - Approximate solution can be obtained using Loyd iterations.
Doubling dimension – Assouad dimension – locality

- Assouad and doubling dimensions (seen earlier)
- On the importance of locality: see examples of the accuracy of regressors based on nearest neighbors (seen earlier)
Recursive splits: how many splits are required to halve the diameter of a point set?

- A set defined along coordinate axis in $\mathbb{R}^D$:
  - Consider $S = \bigcup_{i=1,...,D} \{ t e_i, -1 \leq t \leq 1 \}$.
  - $S \subseteq B(0,1)$ and covered by $2D$ balls $B(\cdot, 1/2)$ (this num. is minimal)
  - Assouad dimension is $\log 2D$

- Observation: kd-trees requires
  - $d$ splits / levels to halve the diameter of $S$
  - this requires in turn $\geq 2^d$ points

- Fact: RPTree will halve the diameter faster ($d \log d$ levels with $d$ the intrinsic dim.)
Random projections and distances

▶ In $\mathbb{R}^D$: distance roughly get shrunk by a factor $1/\sqrt{D}$

\[
\begin{aligned}
&P\left[ |\langle x, U \rangle| \leq \alpha \frac{\|x\|}{\sqrt{D}} \right] \leq \frac{2}{\pi} \alpha \\
&P\left[ |\langle x, U \rangle| \geq \beta \frac{\|x\|}{\sqrt{D}} \right] \leq \frac{2}{\beta} e^{-\beta^2/2}
\end{aligned}
\]

Lemma 11. Fix any vector $x \in \mathbb{R}^d$. Pick any random unit vector $U$ on $S^{d-1}$. One has:

▶ Rmk: these are so-called concentration inequalities, see later.
Random projections and diameter

▷ Projecting a subset \( S \subset \mathbb{R}^d \) along a random direction: how does the diameter of the projection compares to that of \( S \)?

▷ \( S \) full dimensional:

\[
diam(\text{projection}) \leq diam(S)
\]

▷ \( S \) has Assouad dimension \( d \):

(then, with high probability . . . )

\[
diam(\text{projection}) \leq diam(S)\sqrt{d/D}
\]

▷ Rmk:

Cover \( S \) with:

- \( 2^d \) balls of radius 1/2
- \( 4^d \) balls of radius 1/4
- \( (1/\varepsilon)^d \) balls of radius \( \varepsilon \)
Random projection trees algorithm: rationale

- Keep the good properties of PCA at a much lower cost
  - Intuition: splitting along a random direction is not that different since it will have some component in the direction of the principal component
- Generally works, but in some cases fails to reduce diameter
  - Think of a dense spherical cluster around the mean containing most of the data and a concentric shell of points much farther away (think: outliers)
  - Characterized by the average interpoint distance $\Delta_A$ within cell being much smaller than its diameter $\Delta$
  - $\Rightarrow$ another split is used, based on distance from the mean
Linear versus spherical cuts

▷ **Linear split with jitter:**

{Split by projection: no outlier}

**ChooseRule**($S$)

- choose a random unit direction $v$
- pick any $x \in S$ at random
- let $y \in S$ its furthest neighbor
- choose $\delta$ at random in $[-1, 1] ||x - y|| / \sqrt{d}$

Rule($x$) := $x \cdot v \leq (\text{median}_{z \in S}(z \cdot v) + \delta)$

▷ **Combined split:**

{Split by projection: no outlier}

**ChooseRule**($S$)

\[
\text{if } \Delta^2(S) \leq c \cdot \Delta_A^2(S) \text{ then} \\
\text{choose a random unit direction } v \\
\text{Rule}(x) := x \cdot v \leq \text{median}_{z \in S}(z \cdot v)
\]

\[
\text{else} \\
\{\text{Spherical cut: remove outliers}\} \\
\text{Rule}(x) := ||x - \text{mean}(S)|| \leq \text{median}_{z \in S}(||z - \text{mean}(S)||)
\]

NB: $\Delta$: diameter; $\Delta_A$: average interpoint distance
Random projection trees algorithm: RPTree-max and RPTree-mean

Algorithm:
\[
\text{MakeTree}(S) \\
\text{if } |S| < \text{MinSize} \text{ then} \\
\quad \text{return } (\text{Leaf}) \\
\text{else} \\
\quad \text{Rule } \leftarrow \text{ChooseRule}(S) \\
\quad \text{LeftTree } \leftarrow \text{Maketree} \{x \in S : \text{Rule}(x) = \text{true}\} \\
\quad \text{RightTree } \leftarrow \text{Maketree} \{x \in S : \text{Rule}(x) = \text{false}\} \\
\text{return } [\text{Rule}, \text{LeftTree}, \text{RightTree}] \\
\]

Two options

- RPTree-max: linear split with jitter
- RPTree-mean: combined split
Performance guarantee:

amortized (i.e., global) result for RPTree-max

Def.: radius of a cell $C$ of a RPTree: smallest $r > 0$ such that $S \cap C \subset B(x, r)$ for some $x \in C$.

Theorem 12. (RPTree-max) Consider a RPTree-max built for a dataset $S \subset \mathbb{R}^d$. Pick any cell $C$ of the tree; assume that $S \cap C$ has Assouad dimension $\leq d$. There exists a constant $c_1$ such that with proba. $\geq 1/2$, for every descendant $C'$ more than $c_1 d \log d$ levels below $C$, one has $\text{radius}(C') \leq \text{radius}(C)/2$.

Summary: $d \log d$ levels suffice to halve the diameter (with high probability)
Intermezzo: complexity analysis in computer science

- Various complexities used to analyse the performances of an algorithm:
  - Worst-case - best-case.
    Example: quicksort.
  - Average case: averaged over some randomness hypothesis.
    Example: quicksort.
  - Amortized: averaged over a sequence of operations. A costly operation can help *reorganize* / *optimize* the data structure - construction, which helps future operations.
    Example: insertion into a red-black tree.

Ref: Cormen, Leiserson, Rivest; Introduction to algorithms; MIT press
Performance guarantee:
per-level result for RPTree-mean, with adaptation to covariance dimension

**Theorem 13.** (RPTree-mean) There exists constants $0 < c_1, c_2, c_3 < 1$ for which the following holds.

- Consider any cell $C$ such that $S \cap C$ has **covariance dimension** $(d, \epsilon)$, $\epsilon < c_1$.
- Pick $x \in S \cap C$ at random, and let $C'$ be the cell containing it at the next level down.
- Then, if $C$ is split:
  - by projection (focus on interpoint distance): \( \Delta^2(S) \leq c \cdot \Delta^2_A(S) \)
    \[ E[\Delta^2_A(S \cap C')] \leq (1 - (c_3/d))\Delta^2_A(S \cap C) \]
  - by distance i.e. spherical cut (focus on diameter):
    \[ E[\Delta^2(S \cap C')] \leq c_2\Delta^2(S \cap C) \]

**NB:** the expectation is over the randomization in splitting $C$ and the choice of $x \in S \cap C$. 
Diameter reduction again: the revenge of kd-trees

- Diameter reduction property: holds for kd-trees on randomly rotated data
- Rmk: one random rotation suffices
Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPRTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
p-norms and Unit Balls

- **Notations:**
  - \(d\): the dimension of the space
  - \(\mathcal{F}\): a 1d distribution
  - \(X = (X_1, \ldots, X_d)\) a random vector such that \(X_i \sim \mathcal{F}\)
  - \(P = \{p^{(j)}\}\): a collection on \(n\) iid realizations of \(X\)

- **Generalizations of \(L_p\) norms, \(p > 0\):**

  \[
  \|X\|_p = \left( \sum_i |X_i|^p \right)^{1/p}
  \]  
  \[\text{(12)}\]

  Unit balls: see plots

- **Cases of interest in the sequel:**
  - Minkowski norms: \(p\), an integer \(p \geq 1\):
  - fractional p-norms: \(0 < p < 1\). NB: triangle inequality not respected; NB: balls not convex for \(p < 1\). sometimes called pre-norms.
Concentration of the Euclidean Norm: Observations

Plotting the variation of the following for random points in $[0, 1]^d$:

$$\min \| \cdot \|_2, \; \mathbb{E} [\| \cdot \|_2] - \sigma [\| \cdot \|_2], \; \mathbb{E} [\| \cdot \|_2], \; \mathbb{E} [\| \cdot \|_2] + \sigma [\| \cdot \|_2], \; \max \| \cdot \|_2, \; M = \sqrt{d}$$

(13)

Observation:

- The average value increases with the dimension $d$
- The standard deviation seems to be constant
- For $d \leq 10$ i.e. $d$ small: the min and max values are close to the bounds: lower bound is 0, upper bound is $M = \sqrt{d}$
- For $d$ large say $d \geq 10$, the norm concentrates within a small portion of the domain; the gap wrt the bounds widens when $d$ increases.
Concentration of the Euclidean Norm: Theorem

**Theorem 14.** Let $X \in \mathbb{R}^d$ be a random vector with iid components $X_i \sim F$. There exist constants $a$ and $b$ that do not depend on the dimension (they depend on $F$), such that:

$$\mathbb{E} \left[ \|X\|_2 \right] = \sqrt{ad} - b + O(1/d) \quad (14)$$
$$\text{Var} \left[ \|X\|_2 \right] = b + O(1/\sqrt{d}). \quad (15)$$

**Remarks:**

- The variance is small wrt the expectation, see plot
- The error made in using $\mathbb{E} \left[ \|X\|_2 \right]$ instead of $\|X\|_2$ becomes negligible: it looks like points are on a sphere of radius $\mathbb{E} \left[ \|X\|_2 \right]$.
- The results generalize even if the $X_i$ are not independent; then, $d$ gets replaced by the number of degrees of freedom.
Contrast and Relative Contrast: Definition

Contrast and relative contrast of \( n \) iid random draws from \( X \). The annulus centered at the origin and containing the points is characterized by:

\[
\text{Contrast}_a := D_{\text{max}} - D_{\text{min}} = \max_j \left\| p^{(j)} \right\|_p - \min_j \left\| p^{(j)} \right\|_p.
\]  

(16)

and the relative contrast is defined by:

\[
\text{Contrast}_r = \frac{D_{\text{max}} - D_{\text{min}}}{D_{\text{min}}}.
\]  

(17)

Variation of the contrast \( |D_{\text{max}} - D_{\text{min}}| \) for various \( p \) and increasing \( d \):

Fig. 1. \(|D_{\text{max}} - D_{\text{min}}|\) depending on \( d \) for different metrics (uniform data)
Theorem 15. Consider $n$ points which are iid realization of $X$. There exists a constant $C_p$ such that the absolute contrast of a Minkowski norm satisfies:

$$C_p \leq \lim_{d \to \infty} \mathbb{E} \left[ \frac{D_{\text{max}} - D_{\text{min}}}{d^{1/p - 1/2}} \right] \leq (n - 1)C_p.$$  (18)

Observations:

- The contrast grows as $d^{1/p - 1/2}$

<table>
<thead>
<tr>
<th>Metric</th>
<th>Contrast $D_{\text{max}} - D_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>$C_1 \sqrt{d}$</td>
</tr>
<tr>
<td>$L_2$</td>
<td>$C_2$</td>
</tr>
<tr>
<td>$L_3$</td>
<td>0</td>
</tr>
</tbody>
</table>

- The Manhattan metric: only one for which the contrast grows with $d$.
- For the Euclidean metric, the contrast converge to a constant.
- For $p \geq 3$, the contrast converges to zero: the distance does not discriminate between the notions of close and far.
- NB: the bounds depend on $n$; it makes sense to try to exploit the particular coordinates at hand (cf later).
- NB: Thm also exist for the relative contract and other $p$-norms.
Practical Implications for Exact NN Queries

- **Curse of dimensionality**: the concentration of distance can be such that a method with expected logarithmic cost performs no better than a linear scan. Recall the non defeatist search strategies in kd-tree and metric trees: if all distances are comparable, a linear number of nodes gets visited.

- **Use less concentrated metrics, with more discriminative power** – see also feature selection.

- **Sanity check**: in running a NN query, make sure that distances are meaningful: bi-modality of the distribution of distance is a good sanity check.

- **Example**: selected features yield two modes rather than a continuum of distances:

  ![Graphs showing bi-modal distribution](https://via.placeholder.com/150)

  (a) all rel. Dim.  
  (b) one non-rel. Dim.  
  (c) two non-rel. Dim.
A wise use of distances

Distance filtering:

Feature selection:
References


Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

Intrinsic dimension?

Selected experiments on NN, regression, dimension estimation

RPTrees: search performance analysis

Random projections and intrinsic dimension

Concentration phenomena: application to nearest neighbor searches

Concentration phenomena: key properties
Geometry in high dimension:
scaled bodies and their volume

Scaling a body from \( \mathbb{R}^d \):

\[ \gamma A = \{ \gamma x, x \in A \} \]

For \( \gamma = 1 - \varepsilon \):

\[ \frac{\text{Vol}((1 - \varepsilon)A)}{\text{Vol}(A)} = (1 - \varepsilon)^d \leq e^{-\varepsilon d}. \] (19)

Fix \( \varepsilon \) and let \( d \to \infty \): the ratio tends to zero. That is: nearly all the volume of \( A \) belongs to the annulus of width \( \varepsilon \).

---

1 Use \( e^{-x} \geq 1 - x \)
Unit sphere: surface area and volume

▷ The Gamma function $\Gamma$:  

$$
\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds.
$$

(20)

NB: for integers $\Gamma(n) = (n-1)!$

▷ The surface area and volume of the unit sphere $S^d$ are given by:

$$
A(d) = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \quad V(d) = \frac{V(d)}{d}.
$$

(21)

Variation of the surface area and volume of the unit sphere, as a function of the dimension $d$
Unit sphere: volume concentration near the equator

▷ **Thm:** For $c \geq 1$ and $d \geq 3$, at least a fraction $1 - \frac{2}{c} e^{-c^2/2}$ of the volume of the unit ball satisfies $|x_1| \leq \frac{c}{\sqrt{d-1}}$.

▷ **Corr:** a fraction at least $1 - O\left(\frac{1}{d}\right)$ of the volume of the ball lies in a cube of side length $2^{\frac{2\sqrt{\ln d}}{\sqrt{d-1}}}$.

Proof: apply the Them with $c = 2\sqrt{\ln d}$. Details on the blackboard.
Unit ball: are points near the surface of within a small cubic core?

▷ Apparent contradiction:
  ▷ Argument from body scaling: mass located near the surface of the unit sphere
  ▷ Previous argument: mass located within a box of length $O\left(\sqrt{\frac{\ln d}{d-1}}\right)$

▷ Explanation: typical coordinates of points on the surface are $\pm O\left(\frac{1}{\sqrt{d}}\right)$

Indeed:

$$x_1^2 + x_2^2 + \cdots + x_d^2 = 1.$$  \hfill (22)

Unit sphere versus cube of side 1: $d = 2$, $d = 4$, arbitrary $d
Generating random points on/inside $S^{d-1}$

▷ Generate a point $X = (x_1, \ldots, x_d)^t$ whose coordinates are iid Gaussians:
  - Generate $x_1, \ldots, x_d$ iid Gaussian with $\mu = 0$ and $\sigma = 1$
  - The density of $X$ is
    \[
    f_G(X) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{x_1^2 + x_2^2 + \cdots + x_d^2}{2}}. \tag{23}
    \]
    NB: this density is spherically symmetric (on a sphere of given radius).
  - To obtain a unit vector: $\frac{X}{\|X\|}$. NB: its coordinates are not independent.
  - Inside the unit ball: the point $\frac{X}{\|X\|}$ needs to be scaled by a density $\rho(r) = dr^{d-1}$. 

The Gaussian annulus theorem

\( f_G(X) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{x_1^2 + x_2^2 + \cdots + x_d^2}{2}} \).

(24)

Expectation of \( \|X\|^2 \): 

\[ E[\|X\|^2] = E\left[ \sum_{i=1}^{d} x_i^2 \right] = \sum_{i=1}^{d} E[x_i^2] = dE[x_1^2] = d. \] 

(25)

Thm. Consider an isotropic \( d \) dimensional Gaussian with \( \sigma = 1 \). For any \( \beta \leq \sqrt{d} \), consider the annulus defined by 

\[ \mathcal{A} = \{ X \text{ such that } \sqrt{d} - \beta \leq \|X\| \leq \sqrt{d} + \beta \}. \]

(26)

There exists a fixed positive constant \( c \) such that 

\[ \mathbb{P}(\mathcal{A}^c) \leq 3e^{-c\beta^2}. \]

(27)

Rmks:

- Concentration thm: the mass concentrates near \( \sqrt{E[\|X\|^2]} = \sqrt{d} \).
- The density \( f_G \) is max. at the origin; but integrating over the unit ball ... no mass since the volume of the unit ball tends to 0.
- In going well beyond \( \sqrt{d} \): the density \( f_G \) gets too small.
Projecting onto a (random) affine subspace

- $k$-dimensional affine subspace: matrix $R : n \times k$ whose vectors define the basis
- To obtain such an orthonormal matrix $R$:
  - draw $k$ (unit) random vectors (see above)
  - perform a Gram–Schmidt orthonormalization
    NB: the orthonormalization process complicates things, since entries of the matrix are no longer independent
- To get a randomized dimension-$k$ matrix $R$:
  - draw the $n \times k$ entries at random, using a the normal distribution (Gaussian with 0 mean and unit variance)

Projection $f(v)$ of a vector $v$ onto a (random) affine space of dimension $k$, in matrix form:

$$f(v) = R^t \cdot v.$$ 

(28)

NB: $f(v)$ has dimensions $(k \times n)(n \times) = k \times 1$
Random projection theorem
onto a dimension $k$ affine subspace

- **Goal**: we shall prove that in projection $\|f(v)\| \sim \sqrt{k}\|v\|$

- **Rmks**:
  - The distance/norm $\|f\|$ increases since the vectors defining the affine space are not unit length.
  - The basis defined by $R$ is not orthonormal.
  - BUT: the analysis are much simpler!

- **Thm.** Let $v$ be a vector from $\mathbb{R}^d$. Consider a random affine subspace as defined on the previous slide. Then, for any $\varepsilon > 0$:

$$
\mathbb{P} \left[ \left| \|f(v)\| - \sqrt{k}\|v\| \right| \geq \varepsilon \sqrt{k}\|v\| \right] \leq 3e^{-ck\varepsilon^2}.
$$

(29)

**NB**: the constant $c$ comes from the Gaussian annulus then.

- **Proof**: blackboard.

- **NB**: versions where matrix $R$ is orthonormal also exist. See the bibliography.
Application: the Johnson-Lindenstrauss lemma

▶ Rationale: project a point set $P$ from $\mathbb{R}^d$ to $\mathbb{R}^k$ while preserving distances with low distorsion.

▶ Thm / lemma: Johnson-Lindenstrauss For any $\varepsilon \in (0, 1)$, consider

$$k \geq \frac{3}{c\varepsilon^2} \ln n.$$  \hspace{1cm} (30)

(NB: $c$ from the Gaussian annulus Thm.) For a random projection onto an affin space of dim. $k$, define the event:

$$\mathcal{E} : 1 - \varepsilon \leq \frac{\|f(p_i) - f(p_j)\|}{\|p_i - p_j\|} \leq 1 + \varepsilon, \forall (p_i, p_j).$$  \hspace{1cm} (31)

One has:

$$\mathbb{P} [\mathcal{E}] \geq 1 - \frac{3}{2n}.$$  \hspace{1cm} (32)

▶ Proof: blackboard.

▶ NB: the only property of data used while defining the projection is the number of samples.
Johnson-Lindenstrauss: lower Bound

- Embedding dimension $k$:
  \[ k = \frac{3}{c\varepsilon^2} \ln n. \]  \hspace{1cm} (33)

- Large: $\varepsilon \in [0.5 - 0.99]

- Medium: $\varepsilon \in [0.1 - 5]

- Small: $\varepsilon \in [0.01 - 0.1]


S. Levy, Flavors of geometry, Cambridge, 1997