# Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

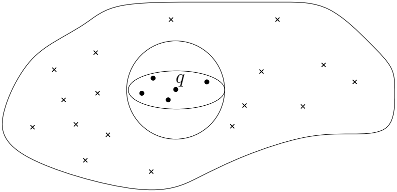
September 21, 2021

Frederic.Cazals@inria.fr

# Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

# Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

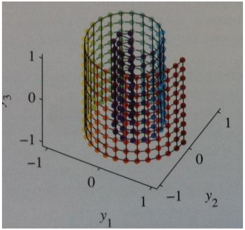
# Nearest neighbors: on the importance of locality




- 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

# 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

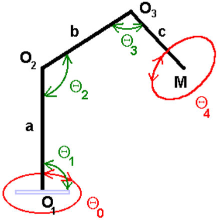
- Consider an $n \times n$ binary image: image $\sim$ point on the hypercube of dimension $n^2$

▷ 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

# Formal notions of intrinsic dimension

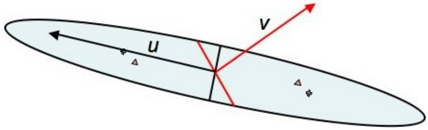
- Natural ones:
  - Affine dimension
  - Manifold dimension
- Requiring (elaborate) calculations:
  - (Local) covariance dimension
  - Assouad - doubling dimension

# 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 $r$.

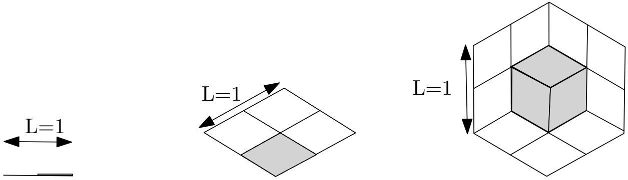



▷ Multi-scale estimation from a point cloud $P$:

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

# Assoud / doubling dimension: intuition

▷ Pick a cube of side length $L$: count how many cubes of side length $L/2$ are needed to cover it



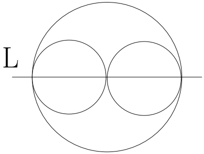

D=1 $N = 2^1$

D=2 $N = 2^2$

D=3 $N = 2^3$

# 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)
- Union of $D$ intervals $[-1, 1]$ in $\mathbb{R}^D$; dim is $\log 2D$
- S = k-dim submanifold of $\mathbb{R}^D$ with finite condition number: Assouad dimension = O(k) in small enough neighborhoods
- S = set of N points: Assouad dimension $\leq logN$

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

# Generalization: doubling dimension and doubling measures

▷ Def.: A metric space $X$ with metric is called *doubling* if there exists $M(X) \in \mathbb{N}$ so that any closed ball $B(x, r)$ can be covered by 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 $\exists C > 0$ such that $\forall x \in X$ and $r > 0$

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

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

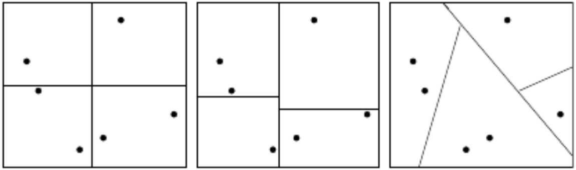
▷ Remarks:

- A metric 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

# Empirical results: contenders

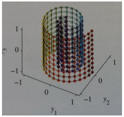
- Contenders / algorithms:
  - dyadic trees aka tries: pick a direction and split at the midpoint; cycle through coordinates.
  - kd-tree: split at median along direction with largest spread.
  - random projection trees: split at the median along a random direction.
  - PD / PCA trees: split at the median 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



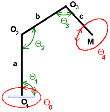
# Real word datasets

▷ 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-seris
- *aw* phoneme from MFCC TIMIT dataset



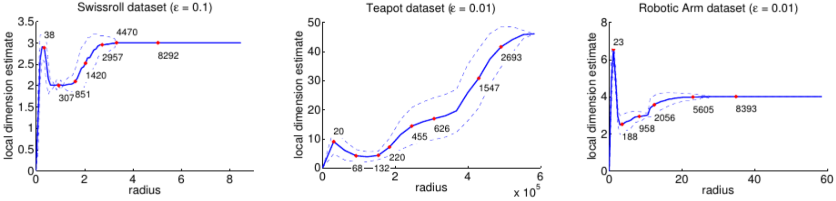





▷Ref: `Verma, Kpotufe, and Dasgupta, UAI 2009.`

# Empirical results: local covariance dimension estimation

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




▷ Observations:

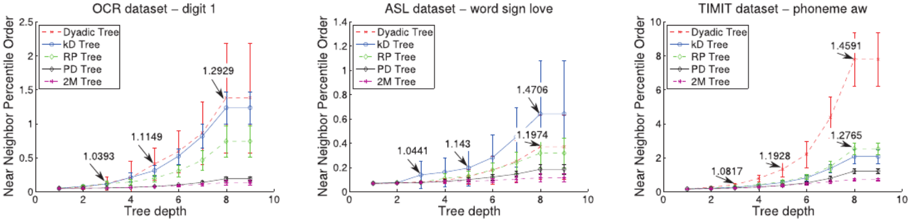
- Swiss roll (ambient space dim is 3): failure at small (noise dominates) and large scales (sheets get blended).
- Teapot: clear small dimensional structure at low scale, but rather 3-4 than 1.
- Robotic arm: tiny spot ($r$ values) to get the correct dimension. . . noise.

▷Ref: Verma, Kpotufe, and Dasgupta, UAI, 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; max is 100%)
- tree depth: NN sought at each level in the tree
- decorating numbers: distance ratio $\|q - nn(q)\| \,/\, \|q - p_{(1)}\|$




▷ Observations:

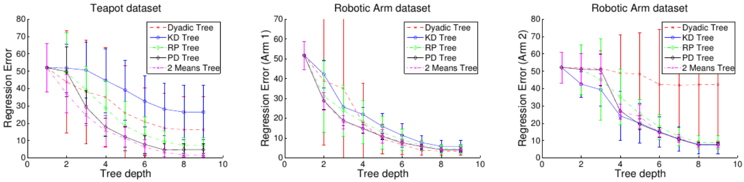
- percentile order deteriorates with depth – separation does occur
- yet, the distance ratio remains *small* even at *high* percentile orders
- 2M and PD (i.e. PCA trees) consistently yield better nearest neighbors: better adaptation to the intrinsic dimension

▷`Ref: Verma, Kpotufe, and Dasgupta, UAI, 2009`

# Empirical results: regression

▷ Regression:

- predicting 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



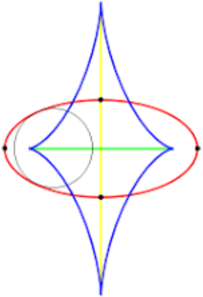

▷ Observations:

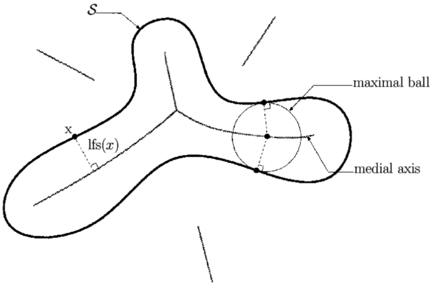
- Small tree depth: averaging over many neighbors is detrimental
- Best results for 2M trees, PD (i.e., PCA) trees, and RP trees.

▷Ref: N. Verma, S. Kpotufe, and S. Dasgupta, UAI, 2009

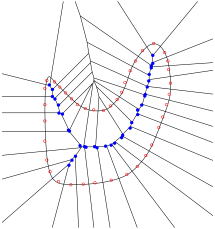
# Intermezzo: medial axis of an open set

▷ Def.:






▷ Construction from Voronoi: idea



# References


- Dasgupta, Sanjoy, and Yoav Freund. Random projection trees and low dimensional manifolds. Proceedings of the fortieth annual ACM symposium on Theory of computing. ACM, 2008.
- Verma, Nakul, Samory Kpotufe, and Sanjoy Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension?. Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 2009.
- J. Heinonen, Lectures on analysis on metric spaces, Springer, 2001.

# Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

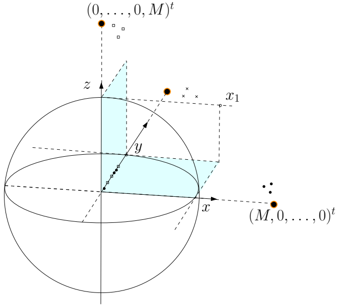
# Random projection trees and nearest neighbors

- 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 $\Phi$

$$\Phi(q, P) = \frac{1}{n} \sum_{i=2}^{n} \frac{\left\| q - x_{(1)} \right\|_2}{\left\| q - x_{(i)} \right\|_2}. \qquad (1)$$

# 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 $(0,1)$

▷ Query point $q$: the origin

▷ kd-trees separate $q$ and $x_1$, 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:

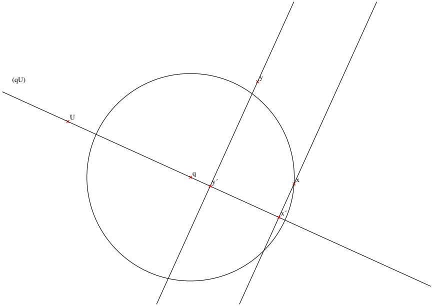
$$\frac{1}{n}\left(n-\frac{n}{d}\right) = 1-\frac{1}{d}$$

▷ Coming next: RPTrees work well in this case; randomness is needed.

# 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/`




▷ Event E to avoid: $< y, U >$ falls strictly in-between $< q, U >$ and $< x, U >$

- ▷ 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:

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

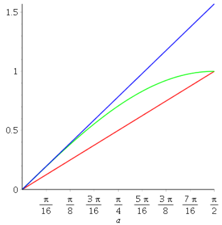
Corollary 2.

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

# 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:

$$\frac{\|q-x\|}{\|q-y\|}\sqrt{1-\mathrm{coll}(q,x,y)^2} \leq \frac{\|q-x\|}{\|q-y\|}$$

Then, apply $(2\phi/\pi) \leq \phi$ to $\phi = \arcsin\|q-x\| / \|q-y\|$.

# Random projections: separation of neighbors

▷ Recall that for $m \geq 1$

$$\Phi_m(q, P) = \frac{1}{m} \sum_{i=2}^{m} \frac{\left\| q - p_{(1)} \right\|_2}{\left\| q - p_{(i)} \right\|_2}. \tag{6}$$

**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 $p_{(1)}$ is at most

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

▷ Proof. Let $Z_i$ be the event : "$p_{(i)}$ *falls between* $q$ *and* $p_{(1)}$ *in the projection*". By the corollary 2, $\mathbb{P}\left[Z_i\right] \leq (1/2) \left\| q - p_{(1)} \right\| / \left\| q - p_{(i)} \right\|$. Then, apply the linearity of expectation to $\sum Z_i / n$ (divide by $n$ to get the fraction).

**Theorem 4.** Let $S \subset P$ with $p_{(1)} \in S$. If $U$ is chosen uniformly at random, then for any $0 < \alpha < 1$, the proba. (over $U$) that a fraction $\alpha$ of the projected points in $S$ fall between $q$ and $p_{(1)}$ is

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

▷ Proof. $\Phi$ is maximized when $S$ consists of the points closest to $q$. Then, previous Thm + Markov's inequality.

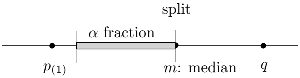
# Regular spill trees–i.e. redundant storage

▷ 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:

$$\mathbb{P}\,[Err] \leq \frac{1}{2\alpha} \sum_{i=0,\ldots,l} \Phi_{\beta^i n}(q, P) \qquad (7)$$




▷ Proof, steps:

- Internal node at depth $i$ contains $\beta^i n$ points
- For such a node: proba to have $q$ separated from $p_{(1)}$

  $p_{(1)}$ 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 upper-bounded by $(1/2\alpha)\Phi_{\beta^i n}(q, P)$
- 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:

$$\mathbb{P}\,[Err] \leq \frac{1}{2\alpha} \sum_{i=0,\ldots,l} \Phi_{\beta^i n}(q, P) \tag{8}$$

▷ 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

# Spill trees: probability of NN search failure

**Theorem 7.** (Spill trees) Consider a spill tree of depth $l = \log 1_{1/\beta}(n/n_0)$, with

- $\beta = 1/2 + \alpha$ for regular spill trees,
- and $\beta = 1/2$ for virtual spill trees.

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

$$\mathbb{P}[Err] \leq \frac{1}{2\alpha} \sum_{i=0,\ldots,l} \Phi_{\beta^i n}(q, P) \tag{9}$$

Nb: $\beta^i n$: number of data points found in an internal node at depth $i$

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

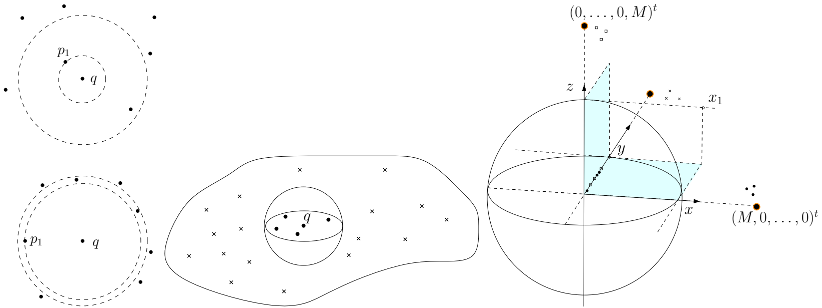
$$\mathbb{P}\left[\text{NN query does not return } p_{(1)}\right] \leq \sum_{i=0,\ldots,l} \Phi_{\beta^i n} \ln \frac{2e}{\Phi_{\beta^i n}} \tag{10}$$

▷ 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.

# Error bound depends on $\Phi$?

- $\Phi$ qualifies the hardness of the query situations
- Focus: pathological cases versus settings with some regularity



# 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 $p_1, \ldots, p_n \sim \mu$. Let $0 < \delta < 1/2$. With probability $\geq 1 - 3\delta$:

$$\forall m \in [2, n] : \quad \Phi_m(q, P) \leq 6\big(\frac{2}{m} \ln \frac{1}{\delta}\big)^{1/d_0}$$

**Theorem 10.** Under the same hypothesis, with $k$ the num. of NN sought:
– For both variants of the spill trees:

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

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

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

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

# References

DS13 S. Dasgupta and K. Sinha. Randomized partition trees for exact nearest neighbor search. JMLR: Workshop and Conference Proceedings, 30:1–21, 2013.

V12 S. Vempala. Randomly-oriented kd Trees Adapt to Intrinsic Dimension. FSTTCS, 2012.

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

# 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.

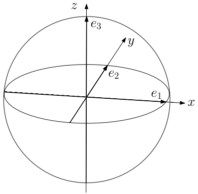
# Doubling dimension – Assouad dimension – locality

- 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 $S = \cup_{i=1,\ldots,D}\{t\ e_i, -1 \leq t \leq 1\}$.
- $S \subset B(0,1)$ and covered by $2D$ balls $B(\cdot, 1/2)$ (this num. is minimal)
- Assouad dimension is $\log 2D$



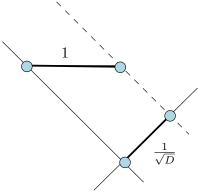

▷ Observation: kd-trees requires

– $d$ 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 $1/\sqrt{D}$




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

$$\mathbb{P}\left[ |< x, U >| \leq \alpha \frac{\|x\|}{\sqrt{D}} \right] \leq \frac{2}{\pi}\alpha \qquad (11)$$

$$\mathbb{P}\left[ |< x, U >| \geq \beta \frac{\|x\|}{\sqrt{D}} \right] \leq \frac{2}{\beta} e^{-\beta^2/2} \qquad (12)$$

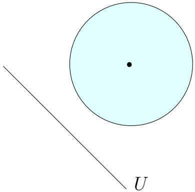
▷ Rmk: these are so-called concentration inequalities, see later.

# 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 $S$?

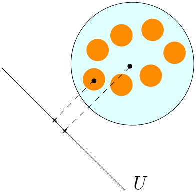
▷ $S$ full dimensional:

$$\text{diam(projection)} \leq \text{diam}(S)$$




▷ $S$ has Assouad dimension $d$:
(then, with high probability...)

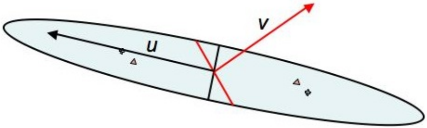
$$\text{diam(projection)} \leq \text{diam}(S)\sqrt{d/D}$$

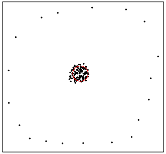



▷ Rmk:

Cover $S$ with $2^d$ balls of radius $1/2$
$4^d$ balls of radius $1/4$
$(1/\varepsilon)^d$ balls of radius $\varepsilon$

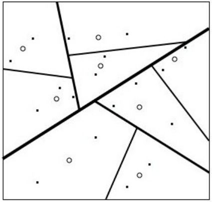
# Random projection trees algorithm: rationale



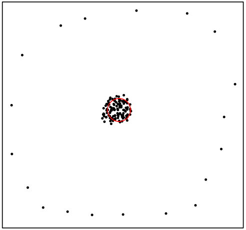
- Keep the good properties of PCA at a much lower cost
  - intuition: splitting along a random direction is not that different since it will have some component in the direction of the principal component
- Generally works, but in some cases fails to reduce diameter
  - Think of a dense spherical cluster 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 $[-1, 1] \|x - y\| / \sqrt{d}$
  $Rule(x) := x \cdot v \leq (\mathbf{median}_{z \in S}(z \cdot v) + \delta)$



- Combined split:

  {Split by projection: no outlier}
  **ChooseRule**$(S)$
  **if** $\Delta^2(S) \leq c \cdot \Delta_A^2(S)$ **then**
  choose a random unit direction $v$
  $Rule(x) := x \cdot v \leq \mathbf{median}_{z \in S}(z \cdot v)$
  **else**
  {Spherical cut: remove *outliers*}
  $Rule(x) := \|x - \mathbf{mean}(S)\| \leq \mathbf{median}_{z \in S}(\|z - \mathbf{mean}(S)\|)$

NB: $\Delta$: diameter; $\Delta_A$: average interpoint distance

# Random projection trees algorithm: RPTree-max and RPTree-mean

- Algorithm:

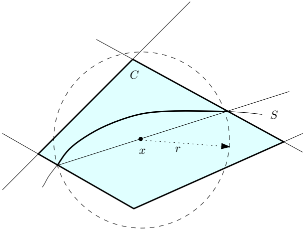
  **MakeTree**(S)
  **if** $|S| <$ MinSize **then**
      **return** $(Leaf)$
  **else**
      $Rule \leftarrow ChooseRule(S)$
      $LeftTree \leftarrow Maketree(\{x \in S : Rule(x) = true\})$
      $RightTree \leftarrow Maketree(\{x \in S : Rule(x) = false\})$
      **return** $[Rule, LeftTree, RightTree]$

- 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): ( $\Delta^2(S) \leq c \cdot \Delta_A^2(S)$ )

$$\mathrm{E}[\Delta_A^2(S \cap C')] \leq (1 - (c_3/d))\Delta_A^2(S \cap C)$$

• by distance i.e. spherical cut (focus on diameter):

$$\mathrm{E}[\Delta^2(S \cap C')] \leq c_2 \Delta^2(S \cap C)$$

▷ NB: the expectation is over the randomization in splitting $C$ and the choice of $x \in S \cap C$.

# Bibliography


- Results presented:
  - Dasgupta S, Freund Y. Random projection trees and low dimensional manifolds. ACM STOC 2008.
- Related:
  - Kpotufe S. k-NN regression adapts to local intrinsic dimension. NIPS 2011.
  - Chaudhuri K, Dasgupta S. Rates of convergence for nearest neighbor classification. NIPS 2014. (NB: k-NN based classification.)

# Diameter reduction again: the revenge of kd-trees

- ▷ Diameter reduction property: holds for kd-trees on randomly rotated data
- ▷ Rmk: one random ration suffices

▷Ref: Vempala. Randomly-oriented kd Trees Adapt to Intrinsic Dimension. FSTTCS. Vol. 18. 2012.

# Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

# 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, $p > 0$:

$$\|X\|_p = (\sum_i^i \mid X_i \mid^p)^{1/p} \qquad (13)$$

Unit balls: see plots

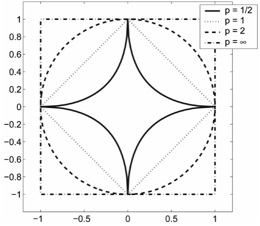



Fig. 2. Two-dimensional unit balls for several values of the parameter of the $p$-norm.

▷ 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.

▷ Study the variation of $\|\|_p$ as a function of $d$

# Concentration of the Euclidean norm: Observations

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

$$\min \|\cdot\|_2, \quad \mathbb{E}\left[\|\cdot\|_2\right] - \sigma\left[\|\cdot\|_2\right], \quad \mathbb{E}\left[\|\cdot\|_2\right], \mathbb{E}\left[\|\cdot\|_2\right] + \sigma\left[\|\cdot\|_2\right], \quad \max \|\cdot\|_2, M = \sqrt{d} \tag{14}$$

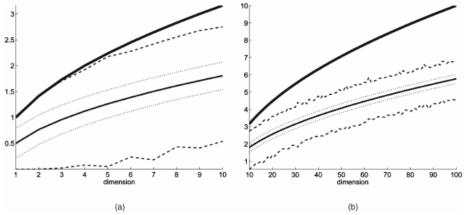



Fig. 1. From bottom to top: minimum observed value, average minus standard deviation, average value, average plus standard deviation, maximum observed value, and maximum possible value of the Euclidean norm of a random vector. The expectation grows, but the variance remains constant. A small subinterval of the domain of the norm is reached in practice.

▷ Observation:

- The average value increases with the dimension $d$
- The standard deviation seems to be constant; likewise for the min-max values
- 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

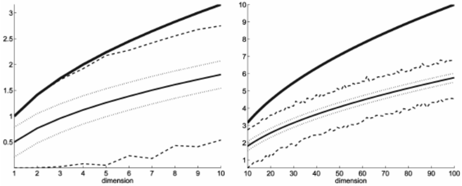
**Theorem 14.** Let $X \in \mathbb{R}^d$ be a random vector with iid components $X_i \sim \mathcal{F}$. There exist constants $a$ and $b$ that do not depend on the dimension (they depend on $\mathcal{F}$), such that:

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

$$\mathrm{Var}\left[\|X\|_2\right] = b + O(1/\sqrt{d}). \quad (16)$$

▷ 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:

$$\text{Contrast}_a := D_{max} - D_{min} = \max_j \left\| p^{(j)} \right\|_p - \min_j \left\| p^{(j)} \right\|_p . \tag{17}$$

and the *relative contrast* is defined by:

$$\text{Contrast}_r = \frac{D_{max} - D_{min}}{D_{min}} . \tag{18}$$

▷ Variation of the contrast $\mid D_{max} - D_{min} \mid$ for various $p$ and **increasing** $d$:

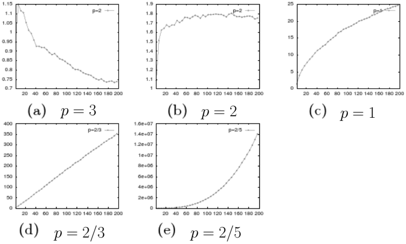



**Fig. 1.** $|Dmax - Dmin|$ depending on $d$ for different metrics (uniform data)

# Contrast and Relative Contrast: the case of Minkowski norms

**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:

$$C_p \leq \lim_{d \to \infty} \mathbb{E}\left[\frac{D_{max} - D_{min}}{d^{1/p-1/2}}\right] \leq (n-1)C_p. \tag{19}$$

▷ Observations:

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

| Metric | Contrast $D_{max} - D_{min}$ |
|---|---|
| $L_1$ | $C_1\sqrt{d}$ |
| $L_2$ | $C_2$ |
| $L_3$ | 0 |

- 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

- The concentration of distances:
  - The first NN (of the origin) is well defined – cf the min curve
  - But in seeking $k$-NN: the concentration is likely to yield a large number of points at the *same* distance – these points are equivalent distance-wise.
- Complexity-wise: the curse of dimensionality:
  - Exact strategies (cf kd-trees, metric trees): likely to trigger a visit of almost all nodes in the tree: the concentration of distance can be such that a method does no better than the linear scan.
  - In contrast: defeatist search strategies suffice.

- Sanity check: in running a NN query, make sure that distances are meaningful: multi-modality (at least bi-modality) of the distribution of distance is a good sanity check to ensure some samples are really closer.

- If possible: use less concentrated metrics, with more discriminative power – see also feature selection.

# A wise use of distances

- Distance filtering:
  - *What is the nearest neighbor in high dimensional spaces?*, Hinneburg et al, VLDB 2000.
  - *Using sketch-map coordinates to analyze and bias molecular dynamics simulations*, Parrinello et al, PNAS 109, 2012.
- Feature selection:
  - *Random Forests*, Breiman, Machine learning 2001
  - *Principal Differences Analysis: Interpretable Characterization of Differences between Distributions*, Mueller et al, NIPS 2015

# References


AHK01 Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. Springer, 2001.

CTP11 M. Ceriotti, G. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. PNAS, 108(32):13023-13028, 2011.

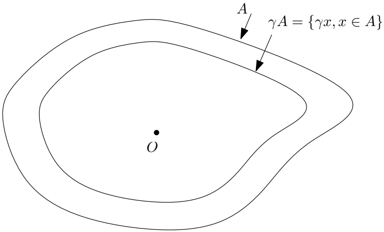
FWV07 D. Francois, V. Wertz, and M. Verleysen. The concentration of fractional distances. IEEE Trans. on Knowledge and Data Engineering, 19(7):873-886, 2007.

HAK00 Alexander Hinneburg, Charu C Aggarwal, and Daniel A Keim. What is the nearest neighbor in high dimensional spaces? VLDB 2000.

# Nearest Neighbors Algorithms in Euclidean and Metric Spaces: Analysis

# Geometry in high dimension: scaled bodies and their volume

▷ Scaling a body from $\mathbb{R}^d$:




▷ For $\gamma = 1 - \varepsilon$ [1]:

$$\frac{\text{Vol}((1-\varepsilon)A)}{\text{Vol}(A)} = (1-\varepsilon)^d \leq e^{-\varepsilon d}. \tag{20}$$

▷ 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

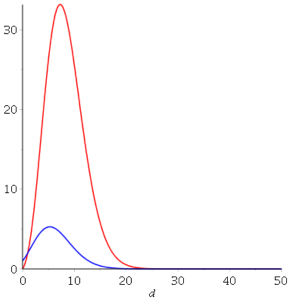
▷ The Gamma function Γ:

$$\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds. \tag{21}$$

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

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

$$A(d) = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \quad V(d) = \frac{A(d)}{d}. \tag{22}$$

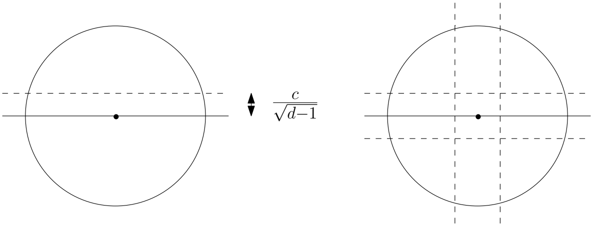



Variation of the surface area (red) and volume (blue) of the unit sphere, as a function of the dimension $d$

# Unit ball: volume concentration near the equator

▷ Thm: (Slab 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: With $c = 2\sqrt{\ln d}$, a fraction at least $1 - O(\frac{1}{d}) \geq 1/2$ of the volume of the unit ball lies in a cube of half side length $c/\sqrt{d-1} = 2\sqrt{\ln d}/\sqrt{d-1}$.
Since the vol. of this cube $\to 0$, the volume of the unit ball goes to 0 when $d \to \infty$.




Proof: apply the Them with $c = 2\sqrt{\ln d}$. Details on the blackboard.

Nb: Vertices of the cube are outside the ball. This does not matter since the Thm integrates slices *up to* $c/\sqrt{d-1}$.

# Unit ball:

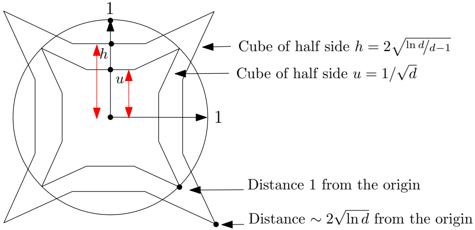
## are points near the surface of within a small cubic core?

▷ Apparent contradiction:

- Argument from body scaling: mass located near the surface of the unit sphere
- Previous argument: $\geq 1/2$ of the mass located *near* the equator, within a cube of side length $4\sqrt{\ln d / d-1}$

▷ Explanation:

- cube whose vertices are on the unit sphere: half side $1/\sqrt{d}$
- corners of the cube of half side length $h = 2\sqrt{\ln d / d-1}$ are at distance $\sim 2\sqrt{\ln d}$ from the origin. this cube covers a significant portion of the unit ball.




The cube of *small* side length $h$ *projects* vertices far away from the unit sphere.

# Random points are almost orthogonal with high probability

▷ Thm. Consider $n$ points $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ drawn uniformly at random from the unit ball. The following holds with probability $1 - O(1/n)$:

1. $\mathbb{P}\left[\|\boldsymbol{x}_i\| \geq 1 - \frac{2\ln n}{d}\right] \geq 1 - O(1/n), \forall i$
2. $\mathbb{P}\left[|< \boldsymbol{x}_i, \boldsymbol{x}_j >| \leq \sqrt{\frac{6\ln n}{d-1}}\right] \geq 1 - O(1/n), \forall i \neq j.$

# Generating random points on/inside $S^{d-1}$

- Generate a point $\boldsymbol{x} = (x_1, \ldots, x_d)^t$ whose coordinates are iid Gaussians:
  - Generate $x_1, \ldots, x_d$ iid Gaussian with $\mu = 0$ and $\sigma = 1$
    - distribution is spherically symmetric (on a sphere of given radius).
    - random vector has arbitrary norm
  - The density of $X$ is

$$f_G(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{x_1^2 + x_2^2 + \cdots + x_d^2}{2}} = \frac{1}{(2\pi)^{d/2}} e^{-\|\boldsymbol{x}\|^2/2}. \tag{23}$$

  - To obtain a unit vector: $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$. NB: its coordinates are not independent.
- Inside the unit ball: the point $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ needs to be scaled by a density $\rho(r) = dr^{d-1}$.

# The Gaussian annulus theorem

## for an isotropic $d$ dimensional Gaussian

▷ Density of the isotropic Gaussian: Gaussian of zero mean and $\sigma^2$ along each dir.:

$$f_G(X) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{x_1^2+x_2^2+\cdots+x_d^2}{2}}. \quad (24)$$

▷ Expectation of $\|X\|^2$:

$$\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[\sum_{i=1,\ldots,d} x_i^2\right] = \sum_{i=1,\ldots,d} \mathbb{E}\left[x_i^2\right] = d\mathbb{E}\left[x_1^2\right] = d. \quad (25)$$

▷ Thm. Consider an isotropic $d$ dimensional Gaussian with $\sigma = 1$ in each direction. For any $\beta \leq \sqrt{d}$, consider the annulus defined by

$$\mathcal{A} = \{X \text{ such that } \sqrt{d} - \beta \leq \|X\| \leq \sqrt{d} + \beta\}. \quad (26)$$

There exists a fixed positive constant $c$ such that

$$\mathbb{P}(\mathcal{A}^c) \leq 3e^{-c\beta^2}. \quad (27)$$

▷ Rmk: how come the mass concentrates around $\sqrt{d}$?

- Concentration thm: the mass concentrates near $\sqrt{\mathbb{E}\left[\|X\|^2\right]} = \sqrt{d}$
- 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. (prop. seen earlier.)
- In going well beyond $\sqrt{d}$: the density $f_G$ gets too small.

# Projecting onto a (random) affine subspace

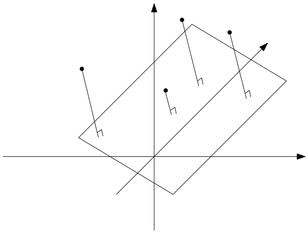
▷ $k$-dimensional affine subspace: matrix $R : d \times k$ whose vectors define an (orthonormal) basis

▷ 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 $R$ – dim is $d \times k$):

- Draw the $d \times k$ entries at random, using a the normal distribution (Gaussian with 0 mean and unit variance)
- Then $f(\boldsymbol{v}) = (\boldsymbol{u}_1 \cdot \ \boldsymbol{v}, \boldsymbol{u}_2 \cdot \ \boldsymbol{v}, \ldots, \boldsymbol{u}_k \cdot \ \boldsymbol{v})^{\mathsf{T}}$



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

$$f(\boldsymbol{v}) = R^t \cdot \boldsymbol{v}. \qquad (28)$$

NB: $f(\boldsymbol{v})$ has dimensions $(k \times d)(d \times 1) = k \times 1$

# Projection theorem
# onto a random dimension $k$ affine subspace

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

▷ Rmks:

- The distance/norm $\|f\|(\cdot)$ 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 $\varepsilon > 0$:

$$\mathbb{P}\left[\left|\,\|f(\boldsymbol{v})\| - \sqrt{k}\,\|\boldsymbol{v}\|\,\right| \geq \varepsilon\sqrt{k}\,\|\boldsymbol{v}\|\right] \leq 3e^{-ck\varepsilon^2}. \tag{29}$$

NB: the constant $c$ comes from the Gaussian annulus them.

▷ Proof: blackboard.

▷ NB: versions where matrix $R$ is orthonormal also exist. See the bibliography.

# Application: the Johnson-Lindenstrauss lemma

▷ Rationale: project a point set $P = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from $\mathbb{R}^d$ to $\mathbb{R}^k$ while preserving distances / with low distortion.

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

$$k \geq \frac{3}{c\varepsilon^2} \ln n. \tag{30}$$

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

$$\mathcal{E} : (1 - \varepsilon)\sqrt{k} \leq \frac{\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|}{\|\mathbf{x}_i - \mathbf{x}_j\|} \leq (1 + \varepsilon)\sqrt{k}, \forall(\mathbf{x}_i, \mathbf{x}_j). \tag{31}$$

One has:

$$\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \frac{3}{2n}. \tag{32}$$

▷ Proof: blackboard.

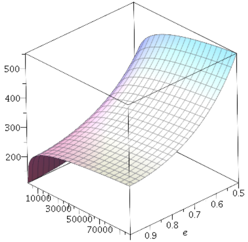
▷ NB: the only *property* of data used while defining the projection is the number of samples.

# Johnson-Lindenstrauss: lower bound

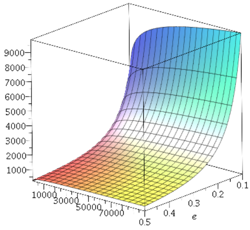
▷ Embedding dimension $k$:

$$k = \frac{3}{c\varepsilon^2} \ln n. \qquad (33)$$

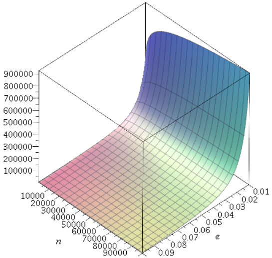
▷ Large: $\varepsilon \in [0.5 - 0.99]$




▷ Medium: $\varepsilon \in [0.1 - 5]$




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



# Bibliography


- S. Dasgupta and A. Gupta, an elementary proof of a theorem of Johnson and Lindenstrauss, Random structures and algorithms, 2003.
- S. Vempala, The random projection method, AMS, 2005.
- S. Levy, Flavors of geometry, Cambridge, 1997
- A. Blum, J. Hopcroft, R. Kannan, Foundations of Data Science, Cambridge, 2020.