Probabilistic Analysis Of Dispersion Function - An Index For Concentration Of Distances In High Dimensional Spaces

AKSHAY GOEL

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Science

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology
Hyderabad

Department of Mathematics

March 2016

Declaration

I, Akshay Goel, do hereby declare that the dissertation entitled 'Probabilistic Analysis Of Dispersion Function' has been undertaken by me for the award of Master of Science in Mathematics & Computing. The study was done under the guidance of Dr. Balasubramaniam Jayaram at IIT Hyderabad. The matter embodied here represents my ideas in my own words, and information, ideas or words derived from the published and unpublished work of others has been acknowledged in the thesis with adequate citations and list of references given in the Biblography. This work has not been submitted in identical or similar manner to any authority and have not been published before.

AKSHAY GOEL
(Student Name)

MA14MSCST11001
(Roll No.)

Acknowledgement

This project would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I am highly indebted to Dr. Balasubramaniam Jayaram for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project. His enthusiasm, encouragement and faith in me throughout have been extremely helpful.

I would like to express my special gratitude towards my beloved parents for their love, encouragement and support financially and mentally that made the completion of this project possible.

My thanks and appreciations also go to my colleague, seniors and juniors who has willingly helped me out with their abilities. I also thank Department of Mathematics, IITH for their support and encouragement.

Lastly, I would like to thank every single person that has been involved directly or indirectly in the completion of this project.

Contents

1. Concentration of Distances (CoD) in High Dimensions

1.1. Introduction

Today we are in such a setting where almost every important data analysis problem is high dimensional. The problem, which we face while working with high dimensional data, is that our intuition about space, that was formed in two and three dimensions, is often misleading in high dimensions. This is sometimes called '*Curse of Dimensionality*'. It is used to refer to the various phenomenon that arises when we work with data in high dimensional space but not so visible in low-dimensional settings such as 2 or 3 dimensions.

There are many aspects of Curse of Dimensionality and their effects are still not well explored and huge amount of research is still going on. Some of the well-known aspects of Curse of Dimensionality are as follows :

(i) *Search Space Complexity:* In some problems, each dimensional variable can take one of several discrete values, or the range of possible values is divided to give a finite number of possibilities. Taking the variables together, a huge number of combinations of values must be considered. This effect is also known as the combinatorial explosion. Even in the simplest case of d binary variables, the number of possible combinations already is $O(2^d)$, exponential in the dimensionality. Naively, each additional dimension doubles the effort needed to try all combinations.

(ii) *Need for greed :* - which refers to the need for atleast a sub-exponential growth in the number of data points as dimension increases for many of the data analysis algorithms to be consistent, see for instance, [11], for more details.

(iii) *Intrinsic vs Embedding*, which refer to the intrinsic and embedding dimensionalities of the data and their influence on the algorithms.

(iv) *Relevance of Dimensions*, which again refers to the presence of irrelevant features that interfere with the performance of similarity queries.

(v) *Hubness Phenomenon* [12], The term was coined after hubs, very frequent neighbor points which dominate among all the occurrences in the k-neighbor sets of inherently high-dimensional data. Most other points either never appear as neighbors or do so very rarely. They are referred to as anti-hubs. This property is usually of a geometric nature and does not reflect the semantics of the data, for instance, in the context of music retrieval. It has been noticed that some songs are very frequently being retrieved, but were unable to attribute these occurrences to any similarity observable by people.

The next section is devoted to another major aspect of Curse of Dimensionality that is *Concentration of Distances* phenomenon, but in detail!

1.2. Survey on CoD in High Dimensions

Concentration of Distances is the phenomenon that, as the data dimensionality increases, all the pairwise distances may converge to the same value. The lack of contrast between the nearest and the farthest points affects each area where high dimensional data processing is required - high dimensional data analysis, database indexing and retrieval, data analysis and statistical machine learning.
To understand this phenomenon, we can do empirically as follows :

- Consider placing 100 points, say $\mathcal{X} = x_1, \ldots, x_{100}$ uniformly at random in a unit interval $[0, 1]$.
- Let us select a point,say x_{i_0} and compute the distance [1] of this point x_{i_0} to all the other points in $\mathcal{X} \setminus \{x_{i_0}\}$. If we plot the histogram of these distances and look at the distribution of these distances, it will be spread throughout the interval $[0, 1]$, see figure 1.2.

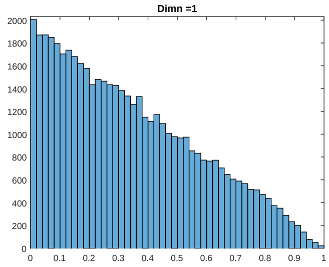
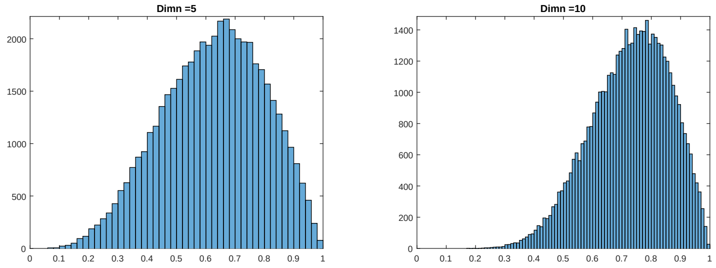


Figure 1: Histogram of Distances in One dimension

Distribution of these distances will be spread throughout $[0, 1]$.

- Now, consider placing 100 points uniformly at random in a unit square $[0, 1] \times [0, 1]$. Each coordinate is generated independently and uniformly at random from the interval $[0, 1]$. Now if we do the same, the spread of the resulting distribution of distances is no more throughout the interval $[0, 1]$.
- If we increase the dimension m and generate the points uniformly at random in a m-dimensional unit hypercube, the distribution of distances becomes concentrated about an average distance. This phenomenon is called the "Concentration of Distances"[8].

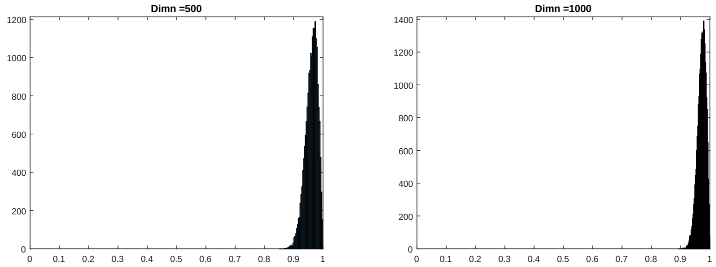


(a) $m = 5$

(b) $m = 10$

As the dimension increases, the spread of the distribution of distances decreases.

[1] We have considered $\mathcal{L}_\infty$ distance function for the above simulation.



(c) $m = 500$

(d) $m = 1000$

As the dimension continues to increase, distribution becomes more and more concentrated about an average value.

1.3. Effects of Concentration of Distances

There are many domains where we have to use different distance functions to measure the proximity and often the data is high dimensional and thus due to CoD, the distance functions which are useful in low dimensions are no longer relevant in higher dimensions. Some of the major areas affected by the CoD phenomenon is Nearest Neighbor Search, Clustering etc.

Nearest Neighbor Search is nothing but to find the object in the database, nearest to the given query point, i.e. the object whose distance to the given query point is minimum. For instance, in face recognition, one needs to search for a picture that is similar to the given query face in a database of images. A picture is made up of thousand of pixels and hence is a high dimensional object. But due to the high dimensionality of the data, all pairwise distances may converge and hence the search might return a lot of candidates similar to our query object, which even does not make any sense. For instance, search for similar pictures of human being might return a picture of table.

This clearly puts a question mark on the usefulness of distance functions in high dimensions as well as meaningfulness of NN query. For more, see[2]

Existing Studies on the CoD Phenomenon

The research studies done on CoD, so far, can be broadly classified into the following three types:

(i) Studies that have theoretically proven the existence of CoD, and also compare different distance functions on the basis of their concentration,

(ii) Studies that have proposed different indices or functions to illustrate or measure the CoD in specific settings,

(iii) Studies that attempt to proposing new distance functions to mitigate the CoD phenomenon.

1.4. Existence of CoD: Theoretical Analysis

Distance functions are known to be sensitive to the dimension of data and hence reduces the efficiency of the search. While searching for the nearest neighbor the obvious approach is to search the database and compute the distance of every data to given query and then to compare the distances. Not only this naive approach is computationally expensive with very large databases, the CoD phenomenon now adds another level of discomfort, since almost all points become equidistant to the query point, i.e., almost all points appear to be the nearest neighbors to the query data, thus questioning the very existence of meaningful nearest neighbor in high dimension.

Beyer *et. al.*[2] were the first to point out that nearest neighbor searching may not always be meaningful when the ratio of the variance of the distance between any two random points, drawn from the data and query distributions, to the expected distance between them converges to zero as dimension goes to infinity by proving the following result. Before stating the result, we have to define some notations, which can be taken as a definition, to state the result.

Definition 1 (**Beyer et. al.,** [2])**.**

- *m is the dimension of the dataset variable which ranges over all positive integers.*
- *$F_{d1}, F_{d2}, \ldots$ is a sequence of data distributions and $F_{q1}, F_{q2}, \ldots$ is a sequence of query distributions.*
- *n is the fixed number of samples (data points) from each distribution.*
- *$\forall m, P_{m1}, P_{m2}, \ldots, P_{mn}$ are n independent data points per m such that $P_{mi} \sim F_{dm}$ and $Q_m \sim F_{qm}$ is a query point chosen independently from all P_{mi}.*
- *$0 < p < \infty$ is a constant.*
- *$\forall m, d_m : F_{dm} \times F_{qm} \to \mathcal{R}^+$ is a (distance) function, and then define :*
  $DMIN_m = min\{d_m(P_{mi}, Q_m) | 1 \leq i \leq n\}$
  $DMAX_m = max\{d_m(P_{mi}, Q_m) | 1 \leq i \leq n\}$

Theorem 1 (**Beyer et. al.,** [2])**.** *Under the conditions in definition 1 if*

$$\lim_{m\to\infty} var\left(\frac{(d_m(P_{m1}, Q_m))^p}{E[(d_m(P_{m1}, Q_m))^p]}\right) = 0 \ , \tag{1}$$

then for all $\varepsilon > 0$,

$$\lim_{m\to\infty} P[DMAX_m \leq (1+\varepsilon)DMIN_m] = 1 \ . \tag{2}$$

Thus, this result shows that under some pre-conditions on the data distribution and distance function the difference between the maximum and minimum distances become very small compared to the minimum distance in high dimension. This means all points are almost equidistant to the query point.

Theorem 1 clearly discusses only a sufficient condition for concentration, i.e., the distance to the nearest neighbor and the distance to the farthest neighbor tend to converge, in a probabilistic sense, as the dimension m increases. In other words, we get a poor contrast if the spread between the points tends towards 0. However, the question of whether this condition is also necessary was not known. Almost after a decade after the work of Beyer *et al.*, the converse of Theorem 1 was proved by Durrant and Kabán in 2009.

Theorem 2 (**Durrant and Kabán**, [5])**.** *(Converse of theorem 1)*
Assume the sample size n is large enough for $E[(d_m(P_{m1}, Q_m))^p] \in [DMIN_m, DMAX_m]$ to hold. Now, if $\lim_{m\to\infty} P[DMAX_m \leq (1+\varepsilon)DMIN_m] = 1, \forall \varepsilon > 0$,
then

$$\lim_{m\to\infty} var\left(\frac{(d_m(P_{m1}, Q_m))^p}{E[(d_m(P_{m1}, Q_m))^p]}\right) = 0$$

This result, in a sense, tries to answer the question when is nearest neighbour meaningful in high dimensions.

1.5. Study of Concentration of Minkowski-type Norms

Theorem 1 and Theorem 2 provided a necessary and sufficient condition on a general distance function to suffer from concentration in high dimensions. Thus, subsequently, researchers began investigating some indices, which were derived out of these results, for different types of distance functions, specifically for Minkowski Norms ($\mathcal{L}_p$ norms).

Minkowski Norms ($\mathcal{L}_p$ norms) :

Minkowski Norms are the family of p-norms parametrized by exponent $p \in (0, \infty)$ which are defined as For an $\bar{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$:

$$\|\bar{x}\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}}$$

.

- For $p = 1$, it is called the Manhattan norm and is denoted as $\mathcal{L}_1$ norm.
- For $p = 2$, it corresponds to the Euclidean norm and is denoted as $\mathcal{L}_2$ norm.
- If $p = \infty$, it becomes the $\mathcal{L}_\infty$-norm or the sup-norm or the Chebyshev metric.
- For $0 < p < 1$, triangle inequality does not hold for $\mathcal{L}_p$. Hence they are not norms but are called prenorms. An $\mathcal{L}_p$-norm, with $0 < p < 1$, is called a Fractional norm and is denoted by $\mathcal{F}_p$.

1.5.1. Empirical Measure to Illustrate the CoD phenomenon

Motivated by theorem 1, an index, called relative contrast, which is based on the contrast present between minimum and maximum distances, was proposed to illustrate the presence of concentration. The definition is as follows

Definition 2. *Let us consider a similarity workload, (Ω, X, ρ, μ). The Relative Contrast with exponent p is defined as*

$$RC(p, m) = \frac{DMAX_m - DMIN_m}{DMIN_m} \ ,$$

where $DMAX_m$ and $DMIN_m$ are as defined earlier.

While Beyer *et al.* studied the CoD phenomenon for arbitrary norms, the first result for concentration of norms was studied for the Euclidean norms by Demartines in his doctoral thesis, who presented the following imporant theorem.

Theorem 3 (**Demartines, 1994**, [4]). *let $X \subsetneq \mathbb{R}^m$ be an m-dimensional data set, where each dimension is distributed in an* i.i.d. *fashion, i.e., each $X_i \sim \mathcal{F}$ and ρ is the $\mathcal{L}_2$ norm. Then,*

$$E(\rho(\bar{x}, \bar{0})) = E\left(\|\bar{x}\|\right) = \sqrt{am - b} + O\left(\frac{1}{m}\right),$$
$$Var(\rho(\bar{x}, \bar{0})) = Var\left(\|\bar{x}\|\right) = b + O\left(\frac{1}{\sqrt{m}}\right) \ ,$$

where a and b are some constants independent of the dimension m.

This theorem shows that expectation of the distances to the origin increases as dimension increases, but the variance remains a constant. Thus, when the dimension is very large, the variance will still be small as compared to the expected distance, hence the points will be closely packed.

The result of Demartines was generalised to any $\mathcal{L}_p$ norm by Hinneburg *et al.*.

Theorem 4 (**Hinneburg *et. al.***, [1]). *Let $X = \{\bar{x}_1^m, \bar{x}_2^m, ..., \bar{x}_n^m\}$ be n m-dimensional i.i.d. random vectors, ρ be any of the Minkowski norms $\mathcal{L}_p$ with exponent p. Then there exists a constant C_p, independent of the underlying distribution $\mathcal{F}$ of $\bar{x}_i^m$, such that*

$$C_p \leq \lim_{m \to \infty} E\left(\frac{DMAX_m - DMIN_m}{m^{\frac{1}{p} - \frac{1}{2}}}\right) \leq (n-1)C_p \ . \qquad (3)$$

Theorem 4 says that the ratio of contrast to $m^{\frac{1}{p} - \frac{1}{2}}$ is bounded by C_p that depends on the exponent p. Based on (3) Hinneburg *et al.* have made the following observations on the exponent p:

- For $\mathcal{L}_p$ norm ($p \geq 3$), the relative contrast rapidly goes to 0 as m increases. It means that the distance function has lost its discriminative power for $p \geq 3$ in high dimensions.
- For the Euclidean $\mathcal{L}_2$ norm ($p = 2$), contrast remains constant.
- For the Manhattan $\mathcal{L}_1$ norm ($p = 1$), contrast increases as $\sqrt{m}$ increases.
- This tends to imply that the $\mathcal{L}_1$ norm is more preferable than the $\mathcal{L}_2$ norm for high dimensional data as it provides a better contrast than $\mathcal{L}_2$ norm.

This result motivated some researchers to consider the Minkowski norms where the exponent $p \in (0, 1)$, i.e., the Fractional norms $\mathcal{F}_p$. Aggarwal *et al.* further extended Theorem 4 to study the concentration of Fractional norms.

Theorem 5 (**Aggarwal *et al.***, [1]). *$X = \{\bar{x}_1^m, \bar{x}_2^m, ..., \bar{x}_n^m\}$ be n m-dimensional i.i.d. random vectors uniformly distributed over $[0,1]^m$. Then there exists a constant C, independent of p and m, such that*

$$C\sqrt{\frac{1}{2p+1}} \leq \lim_{m\to\infty} E\left(\frac{DMAX_m - DMIN_m}{DMIN_m}\right).\sqrt{m} \leq (n-1).C\sqrt{\frac{1}{2p+1}} \ . \tag{4}$$

From (3), it is clear that the constant C may be independent of p but the bounds for relative contrast depend largely on $\sqrt{\dfrac{1}{2p+1}}$. Hence, they concluded that on an average fractional norms provide better contrast then Minkowski norms.

1.5.2. Theoretical Measure to Illustrate the CoD phenomenon

While RC(p,m) is a good empirical measure to illustrate whether a norm concentrates or not, it is not amenable to theoretical analysis. This motivated François *et al.* [6] to introduce a more theoretical index to measure the concentration in a similarity workload (Ω, X, ρ, μ). Note that this index is also derived from the result of Beyer *et al.*, Theorem 1.

Definition 3 (François *et al.* [6], pg. 877). *Given a similarity workload, (Ω, X, ρ, μ), where Ω is m-dimensional, the relative variance of $\rho(\bar{x}, \bar{0}) = \|\bar{x}\|$ is defined as:*

$$RV(p,m) = \frac{\sqrt{Var\ (\|\bar{x}^m\|^p)}}{E\left(\|\bar{x}^m\|^p\right)} \ .$$

The relative variance $RV(p,m)$ illustrates the concentration of distances by comparing the spread of points with the expectation. If $RV(p,m)$ has small value then it indicates that norms are concentrated and a large value for $RV(p,m)$ denotes a good amount of spread between the points. In some sense it is similar to $RC(p,m)$ as it also compares the measure of spread to measure of location.

In fact, Theorems 1 and 2 can be restated as follows based on the above indices: *If the relative variance is not tending to zero then the relative contrast will also not converge to zero and therefore one does obtain a good separation between points.*

For a fixed but large dimension m, François *et al.* also determined the explicit relation between $RC(p,m)$ and p as follows (see [6], **Theorem 6**):

$$RV(p,m) = \frac{\sqrt{Var\|\bar{x}^m\|^p}}{E\|\bar{x}^m\|^p} \approx \frac{1}{p}\left(\frac{\sigma p}{\nu_p}\right) \ , \tag{5}$$

where $\nu_p = E(|X_i|^p)$ and $\sigma_p = Var(|X_i|^p)$.

The above relation (5) shows that for a fixed large m, as p decreases the relative variance $RV(p,m)$ increases and thus explains why an $\mathcal{F}_p$ norm $(0 < p < 1)$ gives better contrast than other $\mathcal{L}_p$ norms where $p \geq 1$.

While both the indices illustrate the concentration phenomenon well, they do not give any information on the rate at which a norm concentrates. Recently, Pestov [10] introduced a more general mathematical function to measure concentration.

Definition 4 (Pestov, [10])**.** *Let us be given a measurable metric space (Ω, ρ, μ). The concentration function $\alpha_\Omega : \mathbb{R}^{\geq 0} \to [\frac{1}{2}, 1]$ is defined as follows:*

$$\alpha_\Omega(\varepsilon) = \begin{cases} 1 - \inf\{\mu(O_\varepsilon(A)) : A \subseteq \Omega \ \textit{is} \ \ \textit{Borel} \ \& \ \mu(A) \geq 1/2\} \ , & \textit{if } \varepsilon > 0 \ , \\ \frac{1}{2}, & \textit{if } \varepsilon = 0 \ , \end{cases}$$

where

$$O_\varepsilon(A) = \{x \in \Omega : \textit{for} \ \ \textit{some } a \in A, \rho(x, a) < \varepsilon\} \ .$$

The value $\alpha_\Omega(\varepsilon)$ gives an upper bound on the measure of the complement to the ε-neighborhood A_ε of every subset A of measure greater than or equal to $\frac{1}{2}$.

2. Motivation and Intent of the work

2.1. Motivation

Recently, [9] a new index to measure the concentration of different distance functions called the *dispersion index*, denoted by τ_ρ where ρ is the distance function, has been proposed which is explained in detail below.

Notations :

Let us fix some notations before going further :

Symbol	Explanation
$\mathcal{X}$	Given Data set,
m	Dimension of Data,
N	Total number of data points,
δ_i	Nearest Neighbor Distance of i^{th} point,
Δ_p^i	Random variable of δ_i with respect to $\mathcal{L}_p$ distance,
$F_{\Delta_p^i}$	Cumulative Distribution Function of Δ_p^i,
$f_{\Delta_p^i}$	Probability Density function of Δ_p^i,
Δ_p^0	Maximum of $\Delta_p^1, \Delta_p^2, \ldots, \Delta_p^n$,
$f_{\Delta_p^0}$	Probability Density function of Δ_p^0,
δ_0	Expectation of Δ_p^0,
$Pr(A)$	Probability of event A,
$Pr(A, B)$	Probability of $A \cap B$,
$Pr(A\|B)$	Probability of A given B.

Dispersion Index and Related Concepts

All the definitions and theorems in this section has been taken from [9].

Definition 5. *Let $(\Omega, \mathcal{X}, \rho, \mu)$ be the similarity workload. Given a query point $q \in \mathcal{X}$ and $\epsilon \in \mathcal{R}^+$, a range query is said to be ϵ unstable if*

$$\#\big\{x \in \mathcal{X} : \rho(q, x) \geq (1+\epsilon)\delta\big\} \geq \frac{\#(\mathcal{X})}{2}$$

where $\delta = \min\{\rho(q, x) : x \in \mathcal{X}\}$

Definition 6. *Let $\mathcal{X}$ be given data set whose cardinality is N i.e., $\mu_c(\mathcal{X}) = N$. Let $x \in \mathcal{X}$ and let δ denote the nearest neighbor distance of x. For any $g \in \mathcal{R}^+$, the $g\delta$ nbd of x is defined as:*

$$N_{g,\delta} = N_g(x, \delta) = \{x' \in \mathcal{X} : \rho(x', x) \leq g\delta\}$$

Definition 7. *Define a function $C : \mathcal{X} \times \mathcal{R}^+ \to N_n$ as*

$$C(x, g) = \mu_c(N_{g\delta}(x))$$

$C(x, g)$ is called the $g\delta$ count of the point x.

If $C(x, g)$ values of most of the $x \in \mathcal{X}$ for small values of g are high, the more points are lying in the dilated $g\delta$ nbd of each $x \in \mathcal{X}$ and hence we can say the data is distributed very close to each other and relative distance between the data points will be small.

Definition 8. *Define $C^* : \mathcal{X} \times \mathcal{R}^+ \to [0, 1]$, called the normalized complement of C, as*

$$C^*(x, g) = \frac{(\#\mathcal{X} - C(x, g))}{\#\mathcal{X}} = 1 - \frac{C(x, g)}{\#\mathcal{X}}$$

Definition 9 ([9], Definition 4.2)**.** *Dispersion Function (λ_ρ): Let $(\Omega, \mathcal{X}, \rho, \mu)$ be the similarity workload and $\mu_c(\mathcal{X}) = n$. Define $\lambda_\rho : (-1, \infty] \to [0, 1]$ as*

$$\lambda_\rho(\epsilon) = avg_{x_i \in \mathcal{X}}(C^*(x_i, (1+\epsilon)\delta_0))$$

For a given $\epsilon > 0$, λ_ρ returns the average of the fraction of the data set which is not captured by a data point in its dilated $(1+\epsilon)\delta_0$ neighborhood. Thus, when n is large, high values of λ_ρ indicate that a large part of the data set are such that most of the data are lying at a distance greater than $(1+\epsilon)\delta_0$ to each of them.

If we take ϵ to be small, then $(1+\epsilon)\delta_0 \approx \delta_0$ and therefore remaining points are at least δ_0 distance away from each point and so data will still be well separated. Thus λ_ρ can be considered as a statistical measure of the dispersion as measured by the distance function ρ.

Theorem 6 ([9], Theorem 5.1)**.** *For any given similarity workload $(\Omega, \mathcal{X}, \rho, \mu)$, λ_ρ is a decreasing function i.e., if $\epsilon_1 \leq \epsilon_2$ then $\lambda_\rho(\epsilon_1) \geq \lambda_\rho(\epsilon_2)$.*

Proof. Let $\epsilon_1 \leq \epsilon_2$ for $\epsilon_1, \epsilon_2 \in (-1, \infty]$.

$$\begin{aligned}
&\epsilon_1 \leq \epsilon_2 \\
&\implies (1+\epsilon_1)\delta_0 \leq (1+\epsilon_2)\delta_0 \\
&\implies N(x_i, (1+\epsilon_1)\delta_0) \subset N(x_i(1+\epsilon_2)\delta_0 \qquad\qquad \forall i \\
&\implies \mu_c(N(x_i, (1+\epsilon_1)\delta_0)) \leq \mu_c(N(x_i(1+\epsilon_2)\delta_0) \quad \forall i \\
&\implies C(x_i, (1+\epsilon_1)\delta_0) \leq C(x_i, (1+\epsilon_2)\delta_0) \qquad\qquad \forall i \\
&\implies C^*(x_i, (1+\epsilon_1)\delta_0) \geq C^*(x_i, (1+\epsilon_2)\delta_0) \qquad\qquad \forall i \\
&\implies avg_{x_i \in \mathcal{X}}(C^*(x_i, (1+\epsilon_1)\delta_0)) \geq avg_{x_i \in \mathcal{X}}(C^*(x_i, (1+\epsilon_2)\delta_0)) \\
&\implies \lambda_\rho(\epsilon_1) \geq \lambda_\rho(\epsilon_2).
\end{aligned}$$

□

Definition 10. *Let $(\Omega, \mathcal{X}, \rho, \mu)$ be the similarity workload and λ_ρ be the corresponding* dispersion function*. Define $\epsilon_\rho^+, \epsilon_\rho^-$ as*

$$\begin{aligned}
\epsilon_\rho^+ &= \sup\{\epsilon \in [-1, \infty) : \lambda_\rho(\epsilon) = 1\} \\
\epsilon_\rho^- &= \inf\{\epsilon \in [-1, \infty) : \lambda_\rho(\epsilon) = 0\}
\end{aligned}$$

Definition 11 ([9], Definition 6.1)**.** *Given a similarity workload $(\Omega, \mathcal{X}, \rho, \mu)$ and the corresponding dispersion function λ_ρ, define the index τ_ρ as*

$$\tau_\rho = \int_{\epsilon_\rho^+}^{\epsilon_\rho^-} \lambda_\rho(\epsilon) d\epsilon$$

τ_ρ is called the dispersion index*.*

Clearly, τ_ρ calculates the area under λ_ρ over the interval $[\epsilon_\rho^+, \epsilon_\rho^-]$.

Empirically the author has also shown in the paper [9] that the above dispersion index can be used to compare different distance functions on the basis of their concentration in high dimensions. Specifically if $\tau_{\rho_1} > \tau_{\rho_2}$ then ρ_1 is less concentrated than ρ_2. Using this index the above claim about the family of $\mathcal{L}_p$ or Minkowski norms was validated empirically.

2.2. Intent of this work

In this work, we are interested in doing a theoretical, largely probabilistic, analysis of the dispersion index, specifically, for the $\mathcal{L}_\infty$ distance function. This is specific case of the $\mathcal{L}_p$ distance function.

Consider the following setting:

- Let $\mathcal{X}$ be the given data set in m dimension and with cardinality N, i.e., let the total number of points in $\mathcal{X}$ be N.
- $\mathcal{X} \sim \mathcal{U}[0,1]^m$. Each dimension $\mathcal{X}_i$ is independent of the other dimensions and identically distributed as $\mathcal{U}[0,1]$.
- Let $\mathcal{L}_p$ denote the p^{th}-Minkowski norm where $p \in (0, \infty]$.

Given the above, we intend to analyze probabilistically the dispersion function for the $\mathcal{L}_\infty$ distance function. This will provide the mathematical backup to the above index and also helps to verify the consistency of the dispersion function mathematically.

2.3. Method

Study above concepts gives idea to divide our work into the following steps:

- **Step 1:** Calculation of the probability density function of nearest neighbor distance (δ_i) for each i^{th} point of the given dataset $\mathcal{X}$ i.e., calculation of $f_{\Delta_p^i}$ for each i^{th} point.
- **Step 2:** Calculation of the probability density function of the maximum of $\Delta_p^1, \Delta_p^1 \ldots, \Delta_p^N$ i.e., calculation of $f_{\Delta_p^0}$.
- **Step 3:** Calculation of Expectation of Δ_p^0 i.e., calculation of δ_0.
- **Step 4:** Let $\epsilon \in [-1, \infty)$ be fixed. Let μ_{x_i} denote the number of points in $((1+\epsilon)\delta_0)(= r)$. Clearly μ_{x_i} is discrete random variable. Thus this step is about the calculation of probability mass function of μ_{x_i}. Hence calculation of probability mass function of $\mu_{x_i}^c$ which is the complement of the μ_{x_i}.
- **Step 5:** Calculation of Expectation of $\mu_{x_i}^c$. Hence calculation of $avg_{x_i \in \mathcal{X}} C^*(x_i, (1+\epsilon)\delta_0)$ which is nothing but $\lambda_\rho(\epsilon)$.

Note 1. *We have described the above method for the general $\mathcal{L}_p$ distance function, though we will do only for the $\mathcal{L}_\infty$ distance function.*

3. Distribution of Nearest Neighbor Distances

We have started by focusing on 1^{st} step of our method described above with respect to the $\mathcal{L}_\infty$ norm i.e., calculation of $f_{\Delta^i_\infty}$ for each i^{th} point.

3.1. Layman's View

Let x_i be a point at the center of the hypercube $[0,1]^m$. Let us consider an $\mathcal{L}_\infty$ hypercube of radius r centered at x_i. Clearly r varies from 0 to $\frac{1}{2}$

Case 1: when $r \leq 0$, $Pr(\delta_i \leq r) = 0$.

Case 2: when $0 \leq r \leq \frac{1}{2}$,$Pr(\delta_i \leq r) = 1 - Pr(\delta_i > r)$

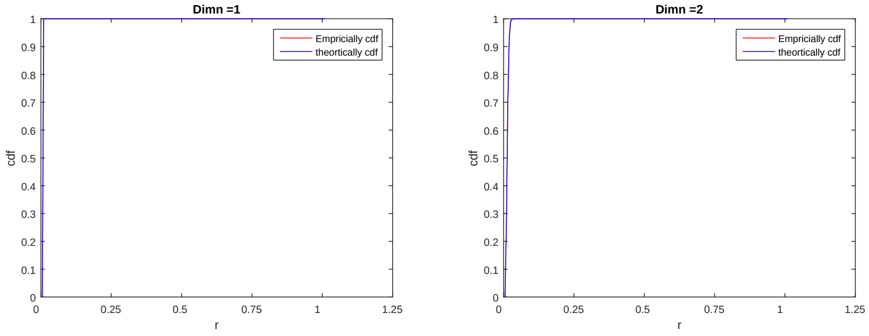
Now, the volume contained in the above hypercube in m-dimensions is $(2r)^m$ and since $\mathcal{X}$ is uniformly distributed the probability of finding a single point outside of this hypercube is $1-(2r)^m$. Hence, the probability of finding all the remaining $N-1$ points outside of this hypercube is $Pr(\delta_i > r) = (1-(2r)^m)^{(N-1)}$ and hence the probability that there exists at least one point at a distance of r from the point x_i is given by

$$Pr(\delta_i \leq r) = 1 - (1-(2r)^m)^{(N-1)} \ .$$

Case 3: when $r \geq \frac{1}{2}$, $Pr(\delta_i \leq r) = 1$.

$$\text{Thus, } F_{\Delta^i_\infty} = \begin{cases} 0, & r \leq 0 \ , \\ 1-(1-(2r)^m)^{(N-1)}, & 0 \leq r \leq \frac{1}{2} \ , \\ 1, & \frac{1}{2} \leq r \ . \end{cases} \tag{6}$$

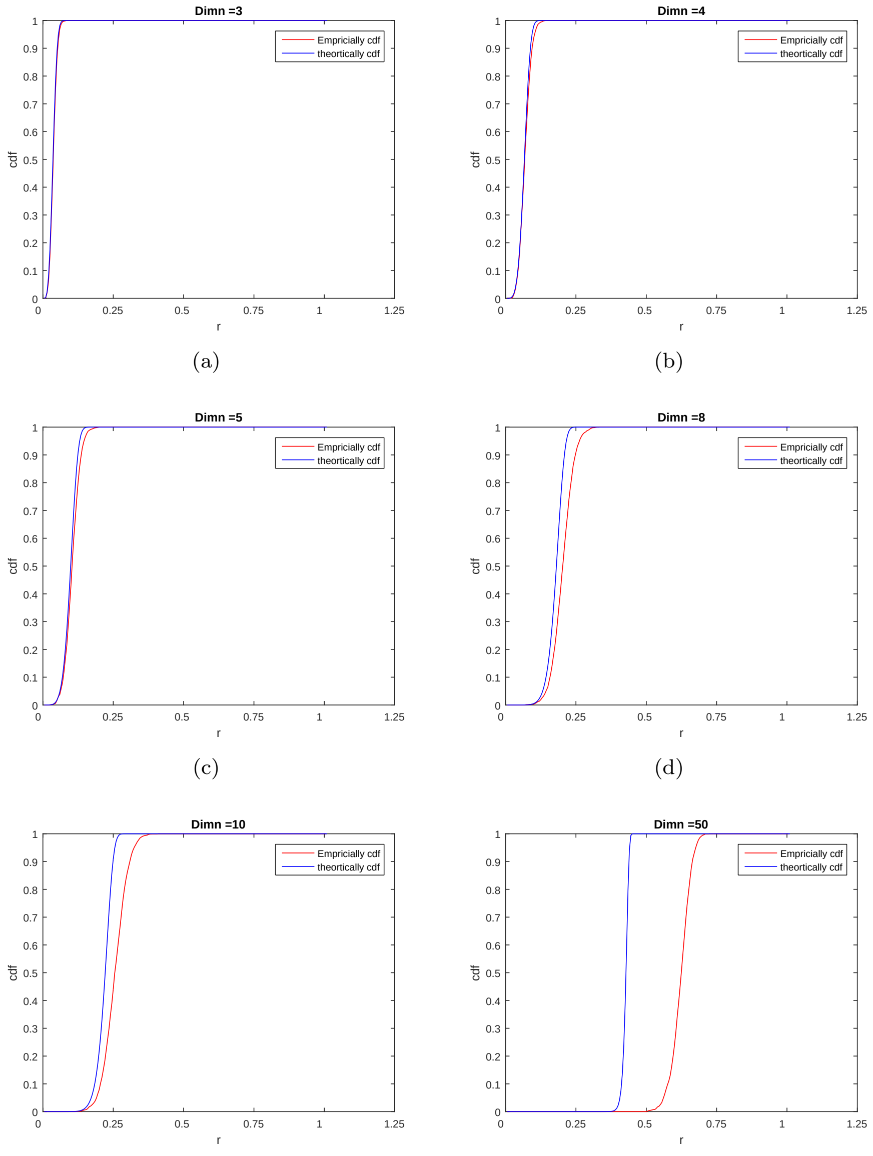
Now, to check the correctness of the equation (6), we will compare the (6) with the simulations for the dimensions $m = 1,2,3,4,5,8,10,50$.



(e)

(f)

Figure 2: Comparison of the CDF as calculated with (6) and the empirical CDF for $m = 1,2,$.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: Comparison of the CDF as calculated with (6) and the empirical CDF for $m = 3, 4, 5, 8, 10, 50$.

From the distribution function $F_{\Delta^i_\infty}$ a probability density function $f_{\Delta^i_\infty}$ can be derived by differentiation:

$$f_{\Delta^i_\infty} = \frac{\partial F_{\Delta^i_\infty}}{\partial r} = 2m(n-1)(2r)^{(m-1)}(1-(2r)^m)^{(n-2)} \ , \qquad 0 \leq r \leq \frac{1}{2} \ . \tag{7}$$

3.1.1. Observations From Simulations

- From the above figures, it is very clear that up to 5 dimensional datasets, the above calculated cumulative density function (6) seems to match almost exactly with the empirically calculated cumulative density function of δ_i.

- But as we increase the dimension of the data set, the empirically calculated CDF moves away from our theoretical CDF (see Figure 3).

- Thus we can say our calculated CDF does not correctly capture the phenomenon when we move to high dimensions.

Therefore, in the next section we will try to find out some likely causes that lead to this situation in higher dimensions.

3.2. Effects in High-Dimensional Data Spaces

In this section we describe one of the many effects occurring in high-dimensional data spaces which are not accordingly captured by (6) and also try to modify (6) accordingly.

Problems specific to higher Dimensional Data spaces

The effect occurring especially in high-dimensional data space is that all data points are likely to be near by the boundary of the data space. This effect is known as Boundary Effect[7].

Intuitively this can be reasoned as follows :
Let us have 100 dimensional data set where points are generated uniformly in $[0,1]^{100}$ hypercube and each dimension is independent of the others. Let $x = (x_1, x_2, ..., x_{100})$ be a point generated where $0 \leq x_i \leq 1, i \in \{1,2,3....100\}$. Since the dimensions are independent of each other, we can think coordinates of x as generating 100 points uniformly in $[0,1]$ interval. Since distribution of points is uniform, it is very likely to have at least one point near 0 or near1, say y and this y is one of the ordinates of x say x_j. Therefore the distance between x_j and j^{th} axis is very likely either to be close to zero or to be close to 1. Thus in the both cases, x will be pulled or pushed by j^{th} to the boundary.

It is very clear from the above reasoning that" Boundary effect is due to the assumption of the uniform distribution of the data points and independence between the dimensions."

The same can also be answered probabilistically as follows :
Let $P_s(r)$ be the probability that a point randomly taken from a uniform and independent distribution in a m-dimensional data space has a distance of r or below to the space boundary.

Case 1: when $r \leq 0$, $P_s(r) = 0$.

Case 2: when $0 \leq r \leq \frac{1}{2}$, $P_s(r) = 1 - (1-2r)^m$.

Case 3: when $r \geq \frac{1}{2}$, $P_s(r) = 1$.

Therefore when m is large even for smaller values of r, $P_s(r)$ is close to 1 which implies the same effect.
The same phenomenon can also be said in terms of the volume of the hypercube as most of the volume of the hypercube is contained in annulus of width ϵ near the boundary where ϵ is inversely related to the dimension of the hypercube and in the uniform distribution case, volume of a region in nothing but the number of points contained in the region divided by total number of points.

Justification

In this section we will try to justify that (6) will not work correctly in high dimension space mainly because of the above discussed Boundary Effect phenomenon.

Let us try to explain this with fixing a point (say x). Now in high Dimension Space this point is more likely to near the boundary due to Boundary effect. So let us fix x near the boundary in two dimensional space (Figure 3.2). Now at a radius (r) greater than the minimum of the distance of the point x from the boundaries of the square, $B(x,r)$ (ball centered at x, of radius r) will not completely lie inside the square. In other words, $B(x,r)$ is no more uniform. Let V_a be the volume of the $B(x,r)$ which completely lie inside the square.

$$\begin{aligned}
&\text{Clearly, } V(B(x,r)) \geq V_a \\
&\qquad \implies (1 - V(B(x,r))) \leq (1 - V_a) \\
&\qquad \implies 1 - (1 - V(B(x,r))) \geq 1 - (1 - V_a)
\end{aligned}$$

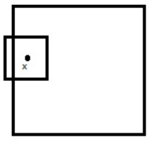


Figure 4: Boundary Effect

Thus, according to (6) Probability of nearest neighbor distance of x less than or equal to r is greater than the actual probability which implies in the Figure 3 the blue curve raises above the x axis before the red curve in high dimension.

The variation or deviation of the blue curve from the red curve is more in high dimensions (see Figure 3) because with the increase in dimension the difference $V(B(x,r)) - V_a$ increases when $V_a > 0$.

3.3. Modification

It is clear from above discussion that due to boundary effect in higher dimensions we have to consider also the position of the point while calculating the probability density function of nearest neighbor distance for that point.

Let us begin by considering one dimensional data set spread over the unit interval $[0,1]$. Consider the intersection of the interval $(x-r, x+r)$ with the unit interval $[0,1]$, where $r \in [0, \frac{1}{2}]$ is fixed and x is the position of the point in the interval $[0,1]$. The function $g : [0,1] \to [0,1]$ given below plots the amount of intersection:[3]

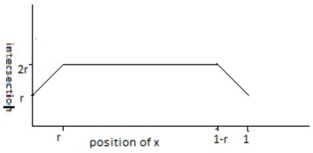


Figure 5: Amount of intersection of the interval $(x-r, x+r)$ with the unit interval $[0,1]$ for fixed $r \in \left[0, \frac{1}{2}\right]$

$$g(x) = \begin{cases} x+r, & 0 \le x \le r \ , \\ 2r, & r \le x \le (1-r) \ , \\ r-x+1, & (1-r) \le x \le 1 \ . \end{cases} \tag{8}$$

Thus when $r \in \left[0, \frac{1}{2}\right]$, for a fixed position x of the point, the intersection of the interval $(x-r, x+r)$ with the interval $[0,1]$ is given by $g(x)$.
Observe that the above graph or definition of function $g(x)$ is valid only when fixed $r \in \left[0, \frac{1}{2}\right]$. Therefore Lets fix $r \in \left[\frac{1}{2}, 1\right]$. Now see below the intersection of the interval $(x-r, x+r)$ with the unit interval $[0,1]$, where $r \in [\frac{1}{2}, 1]$ is fixed and x is the position of the point in the interval $[0,1]$. The function $h : [0,1] \to [0,1]$ given below plots the amount of intersection:

$$h(x) = \begin{cases} x+r, & 0 \le x \le (1-r) \ , \\ 1, & (1-r) \le x \le r \ , \\ r-x+1, & r \le x \le 1 \ . \end{cases} \tag{9}$$

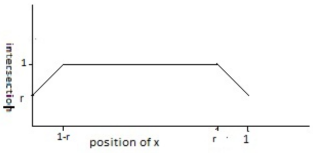


Figure 6: Amount of intersection of the interval $(x - r, x + r)$ with the unit interval $[0, 1]$ for fixed $r \in \left[\frac{1}{2}, 1\right]$

Thus when $r \in \left[\frac{1}{2}, 1\right]$, for a fixed position x of the point, the intersection of the interval $(x - r, x + r)$ with the interval $[0, 1]$ is given by $h(x)$.
We can also write the two functions $g(x)$ and $h(x)$ together and introduce a new function $f : [0.1] \to [0, 1]$ as

$$f(x) = \begin{cases} x + r, & 0 \leq x \leq \min(r, (1 - r)) \ , \\ \min(2r, 1), & \min(r, (1 - r)) \leq x \leq \max(r, (1 - r)) \ , \\ r - x + 1, & \max(r, (1 - r)) \leq x \leq 1 \ . \end{cases} \tag{10}$$

Let $\mathcal{V}(m, q, r)$ denote the volume of the m dimensional hypercube centered at q of radius r, completely contained in $[0, 1]^m$ hypercube .

Therefore, From equation (10), $\mathcal{V}(1, q, r)$ is nothing but equal to the $f(q)$.

Since the dimensions are independent of each other, we can write for a $q = (q^1, q^2 \ldots, q^m)$,

$$\mathcal{V}(m, q, r) = \prod_{j=1}^{m} f(q^j)$$

Thus for a particular point, say $x_i = (x_i^1, x_i^2, \ldots, x_i^m)$ in the m dimensional data set we can calculate the distribution of the nearest neighbor distance δ_i as follows :

Case 1: when $r \leq 0$, $Pr_{x_i}(\delta_i \leq r) = 0$.

Case 2: when $0 \leq r \leq 1$, $Pr_{x_i}(\delta_i \leq r) = 1 - [1 - \mathcal{V}(m, x_i, r)]^{(n-1)}$

$$= 1 - [1 - \prod_{j=1}^{m} f(x_i^j)]^{(n-1)} \ .$$

Case 3: when $r \geq 1$, $Pr_{x_i}(\delta_i \leq r) = 1$.

Now since the position of a point matters, we will compare the above calculated CDF with the simulations by taking three different positions of a point in dimensions $5, 10, 50, 100$.

- When the point is at the origin. In the case

$$\mathcal{V}(m, q, r) = \begin{cases} r^m, & 0 \leq r \leq 1 \ , \\ 1, & r \geq 1 \ . \end{cases}$$

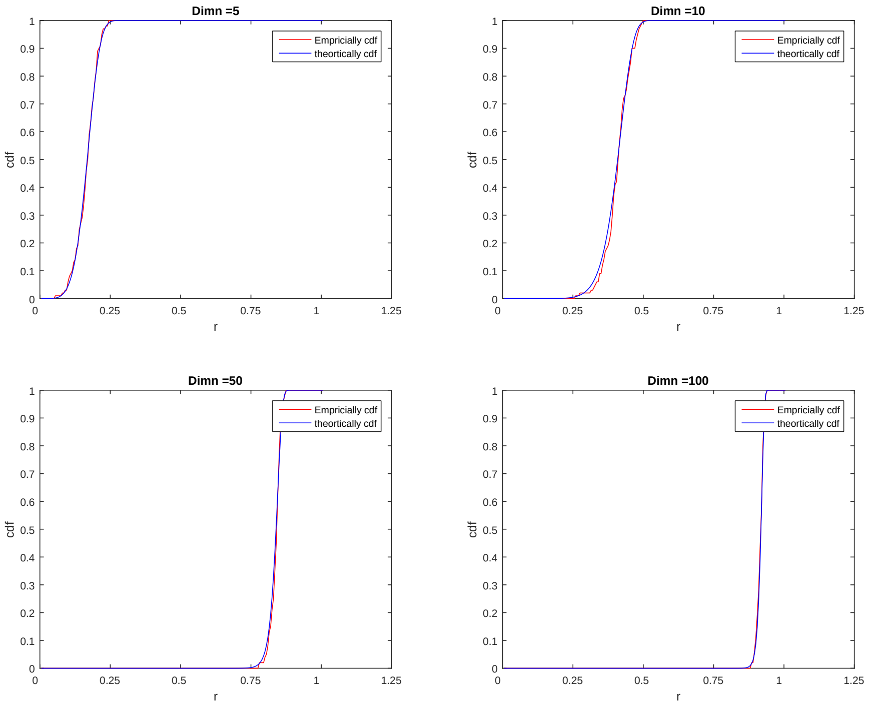
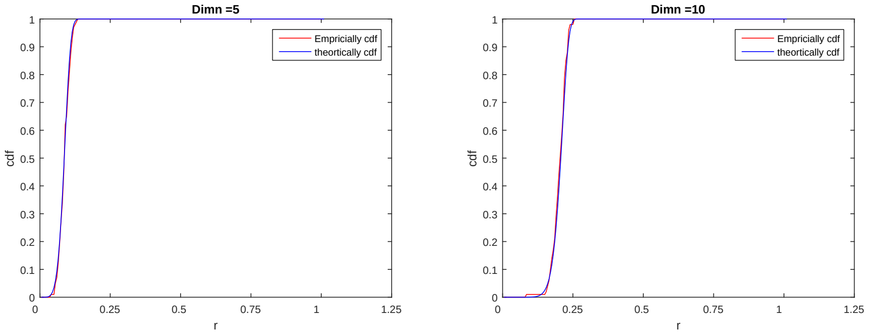


Figure 7: Point is at the origin

- When the point is at the center of the hypercube. In the case

$$\mathcal{V}(m,q,r)=\begin{cases}(2r)^m, & 0\le r\le \frac{1}{2}\ ,\\ 1, & \frac{1}{2}\le r\ .\end{cases}$$



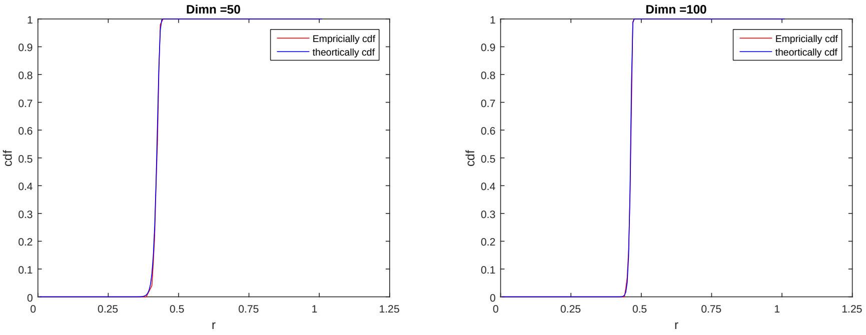


Figure 8: Point is at the center of the hypercube

- When the point is at the center of the one face of hypercube.In this case

$$\mathcal{V}(m,q,r)=\begin{cases} r(2r)^{(m-1)}, & 0\le r\le \frac{1}{2}\ , \\ r, & \frac{1}{2}\le r\le 1\ , \\ 1, & r\ge 1\ . \end{cases}$$

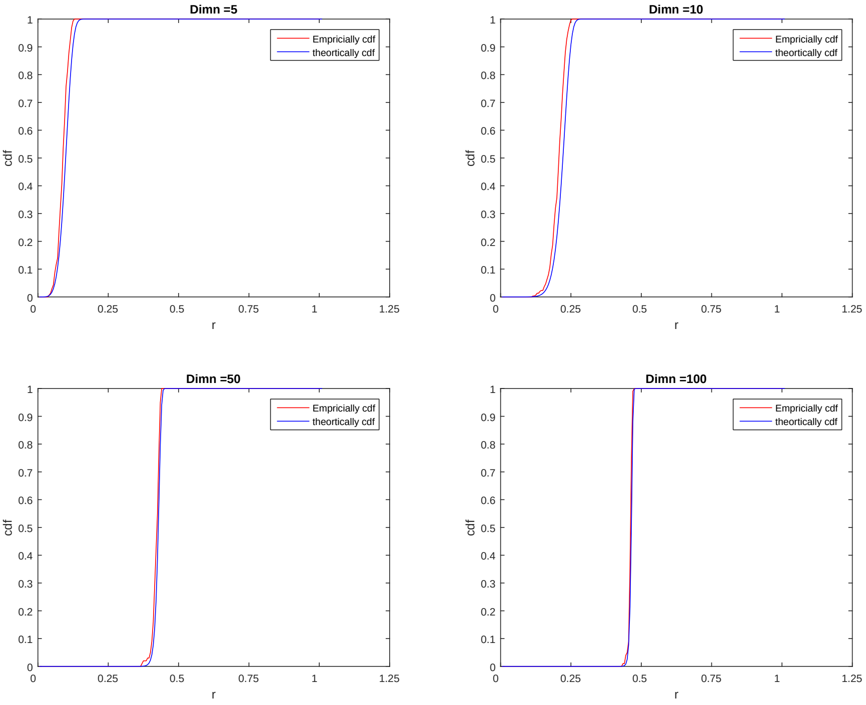


Figure 9: Point is at the center of the face of the hypercube

3.4. Position of the Point: Unknown!

From the above figures we have seen that boundary effect phenomenon can be resolved by considering the position of the point. But for the second step to be solved, above formulas does not help much. The reason simply that it depends on the position of the point which in actual we do not know. Therefore we need some formula like we have above in Layman's view which is independent of the position of the point and also works well in high dimensions.

To address this issue, we consider on an average what is the volume of intersection of interval $(x-r, x+r)$ with the interval $[0,1]$ where $x \in [0,1]$ is the position of the point and r is fixed i.e., nothing but the average of functions $g(x)$ and $h(x)$. Let V_g and V_h denote the average of $g(x)$ and $h(x)$ respectively.

$$\begin{aligned} V_g &= \int_0^1 g(x)\ dx \\ &= \int_0^r (x+r)\ dx + \int_r^{1-r} 2r\ dx + \int_{1-r}^1 (r-x+1)\ dx \\ &= 2r - r^2 \qquad where \ \ r \in \left[0, \frac{1}{2}\right]\ . \end{aligned}$$

$\therefore V_g = 2r - r^2 \qquad where \ \ r \in \left[0, \frac{1}{2}\right]$. Similarly,

$$\begin{aligned} V_h &= \int_0^1 h(x)\ dx \\ &= \int_0^{1-r} (x+r)\ dx + \int_{1-r}^{r} 1\ dx + \int_{r}^1 (r-x+1)\ dx \\ &= 2r - r^2 \qquad where \ \ r \in \left[\frac{1}{2}, 1\right]\ . \end{aligned}$$

$\therefore V_h = 2r - r^2 \qquad where \ \ r \in [\frac{1}{2}, 1]$

Thus On an average the volume of intersection of the interval $(x-r, x+r)$ with the interval $[0,1]$ where $x \in [0,1]$ and fixed $r \in [0,1]$ is determined by the formula :

$$V_{avg} = 2r - r^2 \qquad where \ \ r \in [0,1]$$

Since the dimensions are independent, In the case of m dimensional dataset average volume of intersection is given by :

$$V_{avg}^m = (2r - r^2)^m \qquad where \ \ r \in [0,1]$$

Thus we can replace (6) and can write :

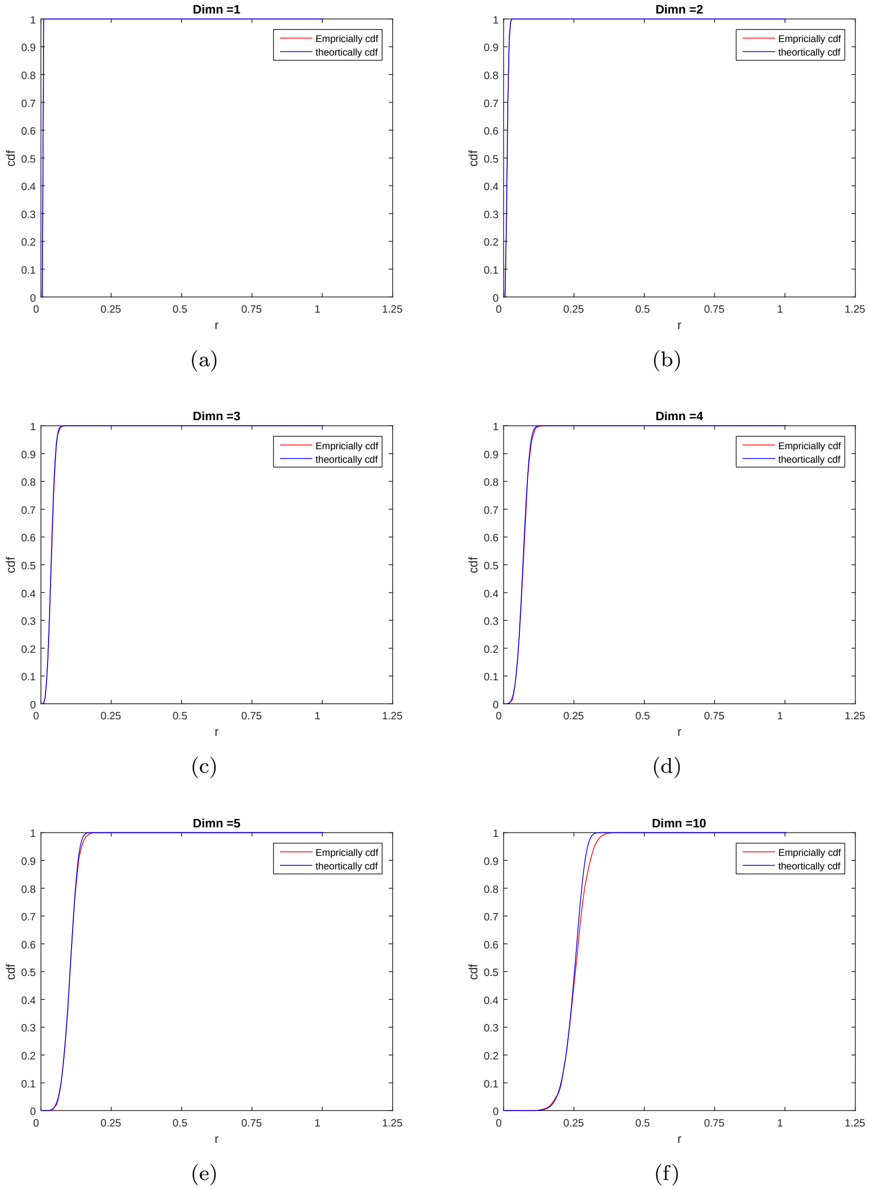
Case 1: when $r \leq 0$, $Pr(\delta_i \leq r) = 0$.

Case 2: when $0 \leq r \leq 1$, $Pr(\delta_i \leq r) = 1 - (1 - (2r - r^2)^m)^{(N-1)}$.

Case 3: when $r \geq 1$, $Pr(\delta_i \leq r) = 1$.

$$\text{Thus, } F_{\Delta_\infty^i} = \begin{cases} 0, & r \leq 0\ , \\ 1 - (1 - (2r - r^2)^m)^{(N-1)}, & 0 \leq r \leq 1\ , \\ 1, & 1 \leq r\ . \end{cases} \tag{11}$$

Now again to check the correctness, we will compare the equation (11) by the simulations for the dimensions $m = 1, 2, 3, 4, 5, 10, 50, 100, 500, 1000, 1200, 1500$.



(a)

(b)

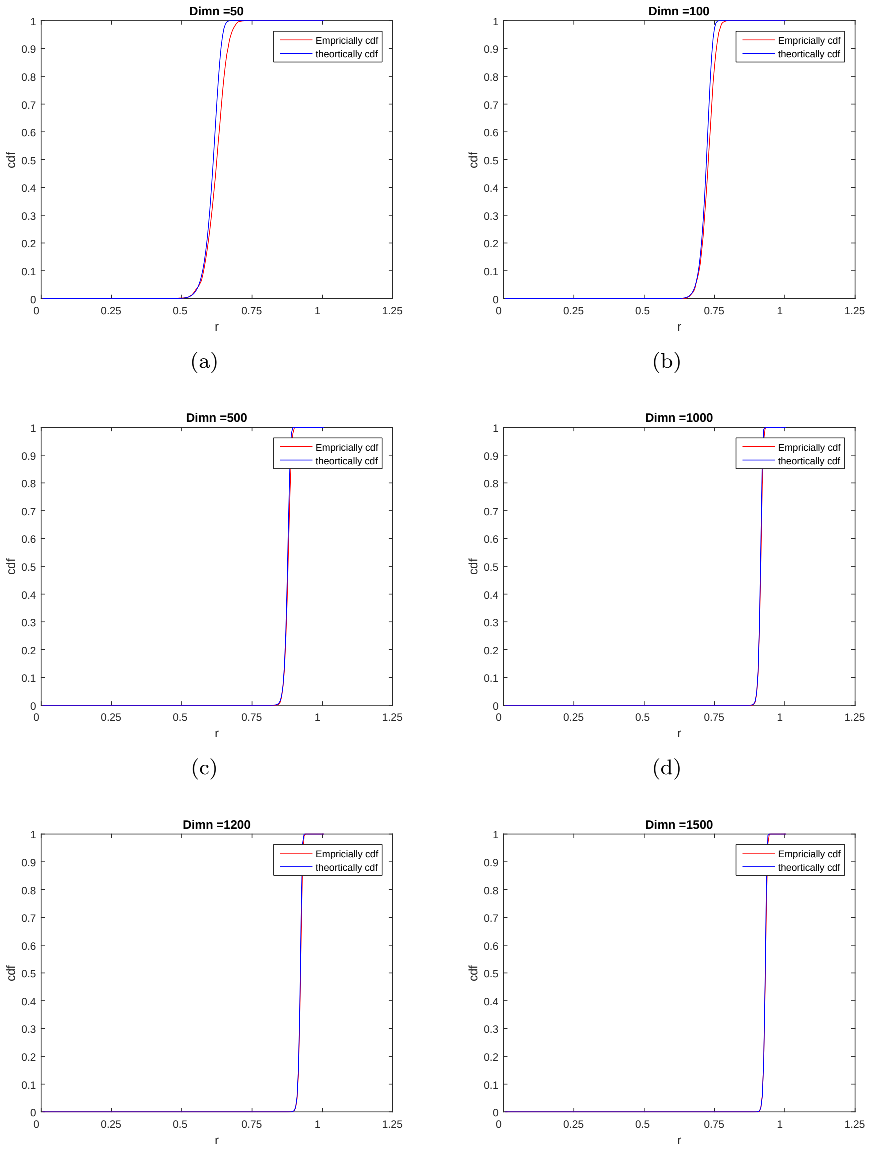
(c)

(d)

(e)

(f)

Figure 10: Comparison of the CDF as calculated with (11) and the empirical CDF for $m = 1, 2, 3, 4, 5, 10$.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 11: Comparison of the CDF as calculated with (11) and the empirical CDF for $m = 50, 100, 500, 1000, 1200, 1500$.

Remarks

So far more or less we have completed the first step of the problem and also address some non intuitive phenomenon (*Boundary Effect*) occurring in High Dimension Space. We modify our formulas to account for this phenomenon.

Also we have noticed though not emphasized much that this Boundary Effect phenomenon is due to the assumptions of the uniformity of the dataset and independency of the dimensions. So it is also interesting to see what happens when we remove either or both the assumptions .

4. Dispersion Function: A Theoretical Approximation

In this section, we carry out the remaining steps, i.e., Steps (2)–(5).

4.1. Distribution of the Maximum NN distance (Step 2)

To calculate the step(2), see 2.3, we will use the concept of order statistics which is as follows :

Definition 12. *For $X_1, X_2, \ldots, X_n$ iid random variables, $X_{(k)}$ is the k^{th} smallest X, usually called the k^{th} order statistic.*

In our case we need only maximum order statistic i.e. $X_{(n)} = max(X_1, X_2, \ldots, X_n)$, thus need to have only the density of the maximum order statistic.

Density of the maximum order statistic ($X_{(n)}$)

For $X_1, X_2, \ldots, X_n$ iid continuous[2] random variables with PDF f and CDF F, the density of the maximum order statistic is :

$$\begin{aligned}
Pr(X_{(n)} \in [x, x+\epsilon]) &= Pr(\text{one of the X's} \in [x, x+\epsilon] \text{ and all others } < x) \\
&= \sum_{i=1}^{n} Pr(X_i \in [x, x+\epsilon] \text{ and all others} < x) \\
&= n\ Pr(X_1 \in [x, x+\epsilon])\ Pr(\text{and all others} < x) \\
&= n\ Pr(X_1 \in [x, x+\epsilon])\ Pr(X_2 < x)\ Pr(X_3 < x) \ldots Pr(X_n < x) \\
&= n\ f(x)\ \epsilon\ (F(x))^{(n-1)}
\end{aligned}$$

$$\implies f_{(n)}(x) = n\ f(x)\ (F(x))^{(n-1)}$$

where $f_{(n)}$ is the PDF of $X_{(n)}$. Thus the CDF, i.e $F_{(n)}$ of $X_{(n)}$ is as follows

$$F_{(n)} = (F(x))^n \tag{12}$$

Using equation (12), CDF of maximum nearest neighbor distance is as follows :

$$F_{\Delta^0_\infty} = \begin{cases} 0, & r \leq 0\ , \\ (1-(1-(2r-r^2)^m)^{(N-1)})^N, & 0 \leq r \leq 1\ , \\ 1, & 1 \leq r\ . \end{cases} \tag{13}$$

4.2. Maximum Nearest Distance on an Average (Step 3)

Now the next step is to calculate the expectation of the Δ^0_∞, i.e. δ_0, which will be calculated by integrating equation (13) from $r = 0$ to 1, i.e.

$$\delta_0 = 1 - \int_0^1 (1-(1-(2r-r^2)^m)^{(N-1)})^N\ dr$$

But it is not possible to find out the closed form solution in terms of m and N. So we move to the next steps, i.e. step (4) and (5) with assuming the maximum nearest neighbor distance on an average is δ_0. Finally, we shall find the values of δ_0, for different values of m and N numerically, to verify the results, which we will obtain in the next section.

[2] In the case of continuous random variable X, $Pr(X_{(i)} = X_{(j)}) = 0\ \ \forall\ \ i \neq j$

4.3. Distribution of $C(x_i, (1+\epsilon)\delta_0)$ (Step 4)

Let us look at the distribution of $C(x_i, (1+\epsilon)\delta_0)$, which is nothing but the number of points lying in the neighborhood of point x_i of radius $(1+\epsilon)\delta_0$.

Let C_i be the random variable corresponding to $C(x_i, (1+\epsilon)\delta_0)$. Clearly, it is a binomial random variable whose probability mass function is as follows :

$$Pr(C_i = K) = \binom{N-2}{K-2} \left(\mathcal{V}(m, x_i, (1+\epsilon)\delta_0)\right)^{(K-2)} \left(1 - \mathcal{V}(m, x_i, (1+\epsilon)\delta_0)\right)^{(N-K)}$$

The reason to choose $(K-2)$ out of $(N-2)$ is that the points x_i and its nearest neighbor always present in the neighborhood of point x_i of radius $(1+\epsilon)\delta_0$ $\forall \epsilon \in (-1, \infty]$.

But to make analysis easier, we shall use $\mathcal{V}^m_{avg}((1+\epsilon)\delta_0)$ instead of $\mathcal{V}(m, x_i, (1+\epsilon)\delta_0)$, so that $C_1, C_2, \ldots, C_N$ will become iid binomial random variables. Also when N is large, $(N-2) \sim N$.

Thus, Probability mass function of C_i is as follows :

$$Pr(C_i = K) = \binom{N}{K} \left(\mathcal{V}^m_{avg}((1+\epsilon)\delta_0)\right)^{K} \left(1 - \mathcal{V}^m_{avg}((1+\epsilon)\delta_0)\right)^{(N-K)}$$

where $\mathcal{V}^m_{avg}(r) = (2r - r^2), \quad r \in [0,1]$

4.4. Theoretical Form of Dispersion function (Step 5)

Now, dispersion function (see Definition 9) can also be written as :

$$\lambda_\infty(\epsilon) = avg_{x_i \in \mathcal{X}} \left(1 - \frac{C(x_i, (1+\epsilon)\delta_0)}{N}\right)$$

which, in terms of probability, is nothing as :

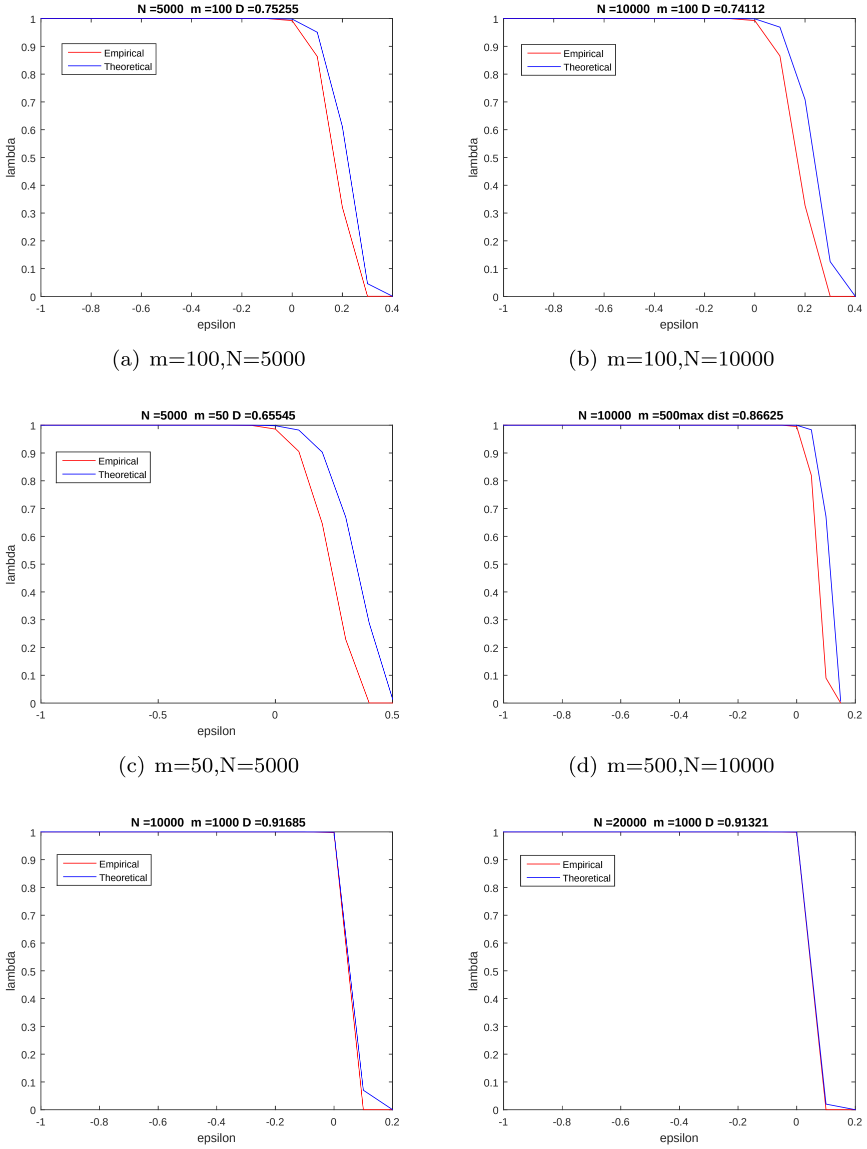
$$\lambda_\infty(\epsilon) = 1 - \frac{E[C_i]}{N}$$

$$\lambda_\infty(\epsilon) = 1 - \mathcal{V}^m_{avg}((1+\epsilon)\delta_0)$$

. Thus,

$$\lambda_\infty(\epsilon) = \begin{cases} 1 - (2(1+\epsilon)\delta_0 - ((1+\epsilon)\delta_0)^2)^m, & -1 \le \epsilon \le \frac{1}{\delta_0} - 1 \ , \\ 0, & otherwise \ . \end{cases} \tag{14}$$

Simulations to verify the above :



(a) m=100,N=5000

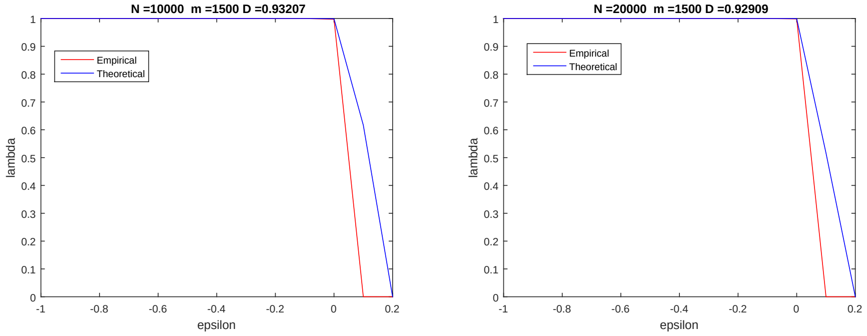
(b) m=100,N=10000

(c) m=50,N=5000

(d) m=500,N=10000

(e) m=1000,N=10000

(f) m=1000,N=20000



(g) m=1500,N=10000

(h) m=1500,N=20000

Figure 12: Comparison of the (14) with Definition (9)

4.5. Dispersion Function for General Minkowski norms

Let $\mathcal{V}^m_{(avg,p)}(r)$, denote the average volume of intersection with $[0,1]^m$, when $\mathcal{L}_p$ distance function is considered. Thus $\mathcal{V}^m_{(avg,\infty)}(r)$ is nothing but $\mathcal{V}^m_{avg}(r)$ as stated above.
Observe that for each $p \in (0,\infty]$ and for each $r \in [0,\infty)$, $\mathcal{V}^m_{(avg,p)}(r) \le 1$.

Theorem 7. *Let δ^p_0 be the maximum nearest neighbor distance on an average when $\mathcal{L}_p$ is considered. Then the dispersion function $\lambda_p : (-1,\infty] \to [0,1]$ can also be defined as :*

$$\lambda_p(\epsilon) = \begin{cases} 1 - \mathcal{V}^m_{(avg,p)}((1+\epsilon)\delta^p_0), & -1 \le \epsilon \le \frac{m^{\frac{1}{p}}}{\delta^p_0} - 1 \ , \\ 0, & otherwise \ . \end{cases} \tag{15}$$

With the understanding that $\frac{1}{\infty} = 0$.

Proof. Proof is similar as that of equation 14. □

Observe that if we want to have closed form formula for the $\mathcal{V}^m_{(avg,p)}(r)$, for distance functions other than $\mathcal{L}_\infty$, it will become very difficult because we cannot simply generalize the volume of intersection from one dimension to any dimension as in the case of $\mathcal{L}_\infty$ distance function (see **Section** 3.3). This suggests some qualitative analysis has to be done so that one can at least compare the values of the dispersion function, for a given ϵ, for two different distance functions.

By "Qualitative Analysis" we mean , not to calculate the CDF of nearest neighbor for each different distance functions or any formula analytically, but to give general results, which points one towards some possible answers about the ordering between values of the dispersion function, for a given ϵ, for two different distance functions.

For instance, still confining ourselves to the data setting as given in Section 2.2, we have the following result:

Proposition 1. *Let us consider $\mathcal{L}_p$ and $\mathcal{L}_q$ distance functions such that $p < q$. Let $B(x,r)$ be the ball centered at x and of radius r completely contained in $[0,1]^m$. Then $Pr_p(\delta \le r) < Pr_q(\delta \le r)$*

Proof.

$$\begin{aligned} Clearly, \quad & V_p(B(x,r)) \le V_q(B(x,r)) \\ & \implies 1 - V_p(B(x,r)) \ge 1 - V_q(B(x,r)) \\ & \implies (1 - V_p(B(x,r)))^{N-1} \ge (1 - V_q(B(x,r)))^{N-1} \\ & \implies 1 - (1 - V_p(B(x,r)))^{N-1} \le 1 - (1 - V_q(B(x,r)))^{N-1} \\ & \implies Pr_p(\delta \le r) \le Pr_q(\delta \le r). \end{aligned}$$

□

The above proposition tells us that if we plot the CDF of nearest neighbor distances for two different distance functions, say $\mathcal{L}_p$ and $\mathcal{L}_q$ such that $p < q$ in the same figure then the graph of CDF corresponding to $\mathcal{L}_q$ is far behind the graph of the CDF of nearest neighbor distances corresponding to the $\mathcal{L}_p$ distance function. We will verify the same by simulation.

Simulation to verify the above proposition

Here we consider three different distance functions, which are $\mathcal{L}_\infty, \mathcal{L}_2\ and\ \mathcal{L}_1$. So, by above proposition, graph of CDF for $\mathcal{L}_\infty$ should come first, then graph for $\mathcal{L}_2$ and then for $\mathcal{L}_1$ distance function. (see Figure 4.5)

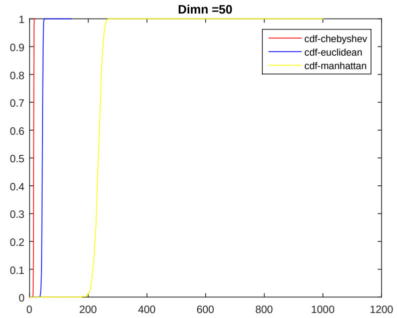


Figure 13: Comparison of the CDF of NN distance wrt $\mathcal{L}_\infty, \mathcal{L}_2\ and\ \mathcal{L}_1$ distance functions respectively.

References

[1] Aggarwal, C. C., Hinneburg, A., Keim, D. A., 2001. On the surprising behavior of distance metrics in high dimensional spaces. In: Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings. pp. 420–434.

[2] Beyer, K. S., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful? In: Database Theory - ICDT '99, 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings. pp. 217–235.

[3] Böhm, C., 1998. Efficiently indexing high-dimensional data spaces. Ph.D. thesis.

[4] Demartines, P., 1994. Analyse de donné,es par ré,seaux de neurones auto-organisé,s. PhD dissertation, Institut Nat',l Polytechnique de Grenoble, Grenoble, France.

[5] Durrant, R. J., Kabán, A., 2009. When is 'nearest neighbour' meaningful: A converse theorem and implications. J. Complexity 25 (4), 385–397.

[6] François, D., Wertz, V., Verleysen, M., 2007. The concentration of fractional distances. IEEE Trans. Knowl. Data Eng. 19 (7), 873–886.

[7] Jiřina, M., Marcel Jiřina, j., July 2004. Features of nearest neighbors distances in high-dimensional space. Tech. rep., Institute of Computer Science , Academy of Sciences of the Czech Republic.

[8] John Hopcroft, R. K., 2014. Foundations of Data Science.

[9] Kumari, S., Jayaram, B., 2015. An efficient and empirical index for measuring the con. IEEE Trans. Knowl. Data Eng., Under Revision.

[10] Pestov, V., 2000. On the geometry of the similar search: dimensionality curse and concentration of measure. Information Processing Letters 73 (10), 47–51.

[11] Pestov, V., 2013. Is the kk-nn classifier in high dimensions affected by the curse of dimensionality? Computers & Mathematics with Applications 65 (10), 1427–1437.

[12] Radovanovic, M., Nanopoulos, A., Ivanovic, M., 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11, 2487–2531.

Approval Sheet

This Thesis entitled 'Probabilistic Analysis Of Dispersion Function - An Index For Concentration Of Distances In High Dimensional Spaces' by Akshay Goel is approved for the degree of Master of Science from IIT Hyderabad.

06/05/2016.

(Dr. Balasubramaniam Jayaram)
Supervisor
Deptarment of Mathematics
IITH