MATH 6397 - Mathematics of Data Science
From signal processing to Convolutional Neural Networks

Instructor: Demetrio Labate

March 11, 2021

Course Outline

1. Mathematics of signal processing
   1.1 Fourier series
   1.2 Wavelets
   1.3 Shearlets
   1.4 Wavelet Scattering Transform
2. Mathematics of machine learning
   2.1 Geometry of high dimensional data
   2.2 Statistical learning theory
   2.3 Support Vector Machines
   2.4 Convolutional Neural Networks

References:

- *The Mathematics of Signal Processing*, by Damelin and Miller
- *Foundations of Data Science*, by Blum, Hopcroft and Kannan
- *Foundations of Machine Learning*, by Mohri, Rostamizadeh and Talwalkar
- *Deep Learning with PyTorch*, by Stevens, Antiga and Viehmann

Part II

Mathematics of Data Science

Mathematics of data science

The main motivation for the paradigm shift occurring with the current notion of ‘data science’ is the emphasis on *multidimensional data.*

While classical and modern signal analysis was mostly concerned with 1-D (time-series), 2-D (images) and 3-D (videos) signals, emerging applications from medical imaging, electronic surveillance, social networks, etc, typically involve data which are high-dimensional and non-Euclidean.

The classical formalism of Hilbert spaces and function representations is often impractical or inadequate.

Mathematics of data science

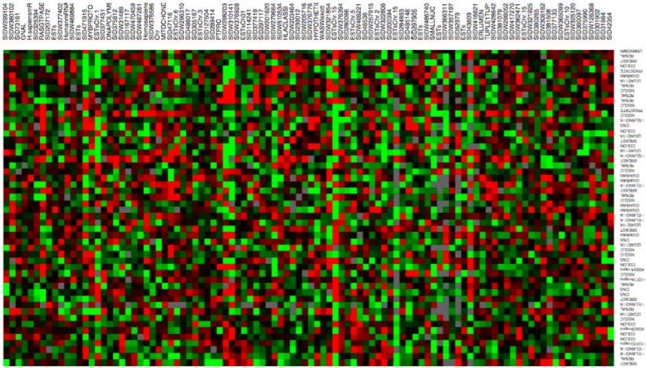


Figure: Computational biology. DNA screening with a few observations and huge number of variables.

Mathematics of data science

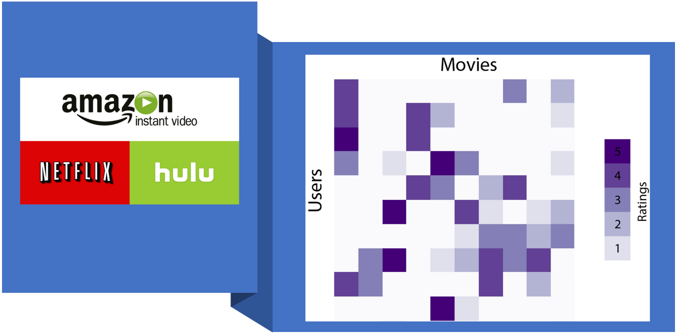


Figure: Netflix challenge (cf. *Netflix Prize*, 2006-2011): to predict users ratings from a sparse incomplete database of ratings given by millions of users on thousands of movies or TV shows.

Geometry of high dimensional data

Geometry of high dimensional data

Two main striking phenomena when one moves from low to high dimensions are:

1. The curse of dimensionality.
2. The concentration of measure.

Both phenomena are manifestations of our difficulty in grasping intuitively the geometry in high dimensions.

Geometry of high dimensional data

Curse of dimensionality [R. Bellman, 1957]: the computational effort associated to many algorithms in R^d become exponentially more onerous as the dimension d grows.

If we want to sample the unit interval such that the distance between adjacent points is at most 0.01, we need 100 evenly-spaced samples.

An equivalent sampling of a 3-dimensional unit hypercube with a grid with a spacing of 0.01 between adjacent points would require 10^6 samples and, similarly, in dimension d, would require 10^{2d} samples.

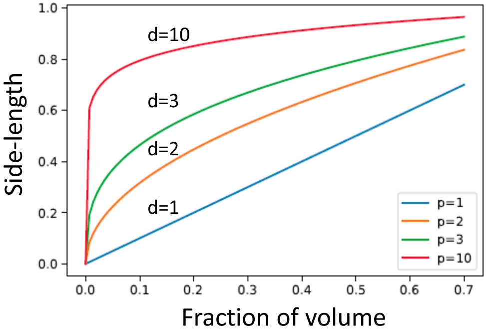
A modest increase in dimensions results in a dramatic increase in required data points to cover the space at the same density.

Geometry of high dimensional data

Notion of **neighborhood**.

To capture a neighborhood that contains a fraction s of the unit hypercube volume, we need the edge length to be $\ell = s^{\frac{1}{d}}$.

- $s = 0.01$, $d = 2$, $\ell = (0.01)^{\frac{1}{2}} = 0.1$
- $s = 0.01$, $d = 3$, $\ell = (0.01)^{\frac{1}{3}} = 0.215...$
- $s = 0.01$, $d = 10$, $\ell = (0.01)^{\frac{1}{10}} = 0.631...$



Geometry of high dimensional data

Notion of **neighborhood**.
Probability is helpful to understand the geometry in high dimensions.

Let X, Y be independent random variables with uniform distribution in $[0,1]^d$.
The mean square distance $\|X-Y\|^2$ satisfies

$$E[\|X-Y\|^2] = \frac{d}{6} \quad \text{and} \quad \mathrm{var}(\|X-Y\|^2) \approx \frac{d}{25}.$$

The notion of nearest neighborhood - which is used in many numerical algorithms - vanishes in high dimensions.

On the other hand, since high-dimensional spaces are sparser, it should be easier to separate points in high-dimensional space with an adapted classifier.

Geometry of high dimensional data

Our geometric intuition about space is naturally based on $d = 2$ and $d = 3$.

This intuition can often be misleading in high dimensions as properties of even very basic objects become counterintuitive. Understanding these paradoxical properties is essential in data analysis.

We consider:

- d-dimensional hyperball of radius R:

$$B^d(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 \leq R^2\}$$

- d-dimensional hypersphere of radius R:

$$S^{d-1}(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 = R^2\}$$

- d-dimensional hypercube of side $2R$:

$$C^d(R) = [-R, R] \times \cdots \times [-R, R] \quad (d \text{ times product})$$

Geometry of high dimensional data

Theorem. The volume of $B^d(R)$ is given by

$$\mathrm{vol}(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\Gamma(\frac{d}{2})}$$

where $\Gamma(n) = \int_0^\infty r^{n-1} e^{-r} dr$ is the *Gamma function*.

Proof. Using polar coordinates,

$$\mathrm{vol}(B^d(R)) = \int_{S^{d-1}(1)} d\Omega \int_{r=0}^{R} r^{d-1} dr = \frac{A_d R^d}{d}$$

where A_d is the surface area of the unit d-sphere $B^d(1)$.
A direct calculation gives

$$\begin{aligned} I(d) &= \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} e^{-(x_1^2 + x_2^2 \ldots + x_d^2)} dx_1 \ldots dx_d \\ &= (\int_{\mathbb{R}} e^{-u^2} du)^d \\ &= \pi^{\frac{d}{2}} \end{aligned}$$

Geometry of high dimensional data

By computing the same integral using polar coordinates, we have

$$\begin{aligned} I(d) &= \int_{S^{d-1}(1)} d\Omega \int_0^\infty e^{-r^2} r^{d-1} dr \\ &= A_d \int_0^\infty e^{-t} t^{\frac{d-1}{2}} \left(\tfrac{1}{2} t^{-\frac{1}{2}}\right) dt \\ &= A_d \, \tfrac{1}{2} \int_0^\infty t^{\frac{d}{2}-1} e^{-t} dt \\ &= A_d \, \tfrac{1}{2} \, \Gamma\left(\tfrac{d}{2}\right). \end{aligned}$$

By comparing with the above calculation of $I(d)$, we conclude that

$$A_d = \frac{\pi^{\frac{d}{2}}}{\frac{1}{2}\Gamma\left(\frac{d}{2}\right)}.$$

Hence

$$\mathrm{vol}(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\,\Gamma\left(\frac{d}{2}\right)} \qquad \square$$

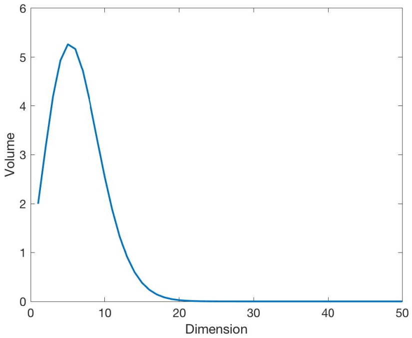
Geometry of high dimensional data

For positive integers n, the have $\Gamma(n) = (n-1)!$ Hence, by Sterling's formula,

$$\Gamma(n) \approx \sqrt{\frac{2\pi}{n}} \left(\frac{n}{e}\right)^n.$$

It follows that, for large d, we have (approximately)

$$\mathrm{vol}(B^d(R)) \approx \frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d}\right)^{\frac{d}{2}}.$$



The volume of the d-sphere reaches its maximum for $d = 5$.

For $d > 5$, the **volume decreases rapidly to zero**.

Geometry of high dimensional data

Observation: Concentration of the volume of a d-ball near its equator

Assume we want to cut off a slab around the equator of the d-unit ball such that 99% of its volume is contained inside the slab.

In two dimensions the width of the slab has to be almost 2, so that 99% of the volume are captured by the slab.
However, as the dimension d increases, the width of the slab gets rapidly smaller.

Indeed, in high dimensions the thickness of the slab shrinks asymptotically to 0, since nearly all the volume of the unit ball lies a very small distance away from the equator.

This phenomenon is a manifestation of the concentration of measure.

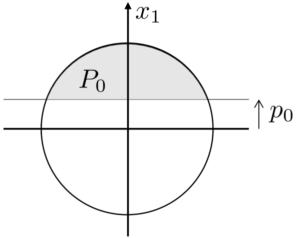
Geometry of high dimensional data

To illustrate more precisely this form of concentration of measure, we examine the unit d-ball.

Without loss of generality, let us first choose a vector x_1 to be the *north pole* so that we can define the *equator* by the intersection with the plane $x_1 = 0 : \{x \in \mathbb{R}^d : \|x\| \leq 1, x_1 = 0\}$.
Hence te equator is a sphere of dimension $d - 1$.

We define the *polar cap* P_0 as the region of the sphere above the slab of width $2p_0$ around the equator,

$$P_0 = \{x \in \mathbb{R}^d : \|x\| \leq 1, x_1 \geq p_0\}$$



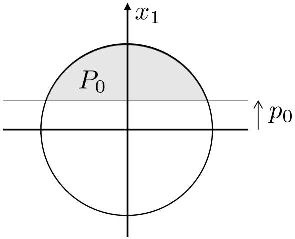
Theorem.

$$\frac{2\,\mathrm{vol}(P_0)}{\mathrm{vol}(B^d(1))} \leq e^{-\frac{d-1}{2}p_0^2}$$

Geometry of high dimensional data

Proof. To compute the volume of the cap P_0 we integrate over all slices of the cap from p_0 to 1.
Each slice is a $(d-1)$-ball of radius $r(x_1) = \sqrt{1 - x_1^2}$.
Hence, the volume of such a slice is

$$(1 - x_1^2)^{\frac{d-1}{2}} \mathrm{vol}(B^{d-1}(1))$$



Thus

$$\mathrm{vol}(P_0) = \mathrm{vol}(B^{d-1}(1)) \int_{p_0}^{1} (1 - x_1^2)^{\frac{d-1}{2}} \, dx_1$$

Using inequalities $1 + x \leq e^x$ and $\mathrm{erfc}(x) \leq e^{-x^2}$, we have

$$\mathrm{vol}(P_0) \leq \mathrm{vol}(B^{d-1}(1)) \int_{p_0}^{\infty} e^{-\frac{(d-1)x_1^2}{2}} \, dx_1 \leq \frac{\mathrm{vol}(B^{d-1}(1))}{d-1} e^{-\frac{(d-1)p_0^2}{2}}$$

Geometry of high dimensional data

From the theorem above, we have that $\mathrm{vol}(B^d(1)) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\Gamma(\frac{d}{2})}$.
It follows that

$$\mathrm{vol}(B^{d-1}(1)) = \frac{\pi^{-\frac{1}{2}}d}{d-1}\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})}\mathrm{vol}(B^d(1)) \leq \frac{d-1}{2}\mathrm{vol}(B^d(1))$$

Thus, from the inequality in page above, we have

$$\mathrm{vol}(P_0) \leq \frac{\mathrm{vol}(B^d(1))}{2}\, e^{-\frac{(d-1)p_0^2}{2}}$$

and, finally,

$$\frac{2\,\mathrm{vol}(P_0)}{\mathrm{vol}(B^d(1))} \leq e^{-\frac{d-1}{2}p_0^2} \qquad \square$$

Geometry of high dimensional data

Observation: Concentration of the volume of a d-ball on shells

Using the formula of the volume of a ball, we obtain

$$\frac{\mathrm{vol}(B^d(1-\epsilon)}{\mathrm{vol}(B^d(1))} = (1-\epsilon)^d \leq e^{-\epsilon d}$$

Since, for any $\epsilon > 0$, this quantity tends to 0 as $d \to \infty$, it follows that the spherical shell contained between $B^d(1)$ and $B^d(1-\epsilon)$ contains most of the volume of $B^d(1)$, for large enough d, even if ϵ is very small.

Setting $\epsilon = \frac{1}{d}$, the estimate shows that at least $(1 - e^{-1})$ of the volume is concentrated in a shell of width $\frac{1}{d}$.

Remark. A similar property holds for d-hypercube. As d increases, most of the volume is concentrated near the surface.

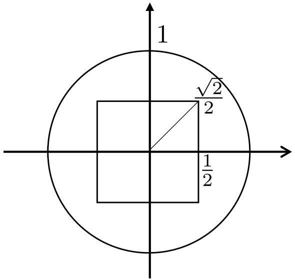
Geometry of high dimensional data

Also the hypercube exhibits an interesting volume concentration behavior.

Proposition. The unit hypercube $C^d(\frac{1}{2})$ has volume 1 and diameter $\sqrt{d}$.

It follows that corners will "stretch out" more and more as the dimension d increases, while the rest of the cube must "shrink" to keep the volume constant.

For $d = 2$, the unit square is completely contained in the unit sphere. The distance from the center to a vertex (radius of the circumscribed sphere) is $\frac{\sqrt{2}}{2}$ and the apothem (the radius of the inscribed sphere) is $\frac{1}{2}$.



Geometry of high dimensional data

For $d = 4$, the distance from the center to a vertex is 1, so the vertices of the cube touch the surface of the sphere. However, the apothem is still $\frac{1}{2}$. The result, when projected in two dimensions no longer appears convex even though all hypercubes are convex.

For $d > 4$, the distance from the center to a vertex is $\frac{\sqrt{2}}{2} > 1$ and thus the vertices of the hypercube extend outside the sphere. (For large d, most of the volume is located in the corners.)

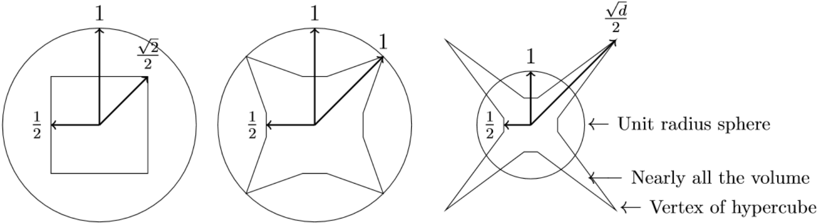


Figure: Relationship between the sphere and the cube in dimensions $d = 2$, $d = 4$ and higher d.

Probability notes

Theorem (Integrated tail probability expectation formula) For any integrable (i.e., finite-mean) random variable X

$$E[X] = \int_0^\infty P(X > x)\,dx - \int_{-\infty}^0 P(X < x)\,dx$$

Proof. We first assume that X is a non-negative random variable. We use the 'layer cake representation' of a non-negative measurable function

$$X = \int_0^X dx = \int_0^\infty \chi_{\{x<X\}}\,dx$$

By interchanging the order of expectation and integration

$$E[X] = \int_0^\infty E[\chi_{\{X>x\}}]\,dx = \int_0^\infty P(X > x)\,dx$$

Probability notes

If X is a general random variable, then we consider its positive and negative parts separately by writing $X = X_+ - X_-$, where $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$.

Using the calculation above,

$$E[X_+] = \int_0^\infty P(X > x)dx; \quad E[X_-] = \int_0^\infty P(X < -x)dx = \int_{-\infty}^0 P(X < x)dx$$

Hence, by the integrability of X,

$$E[X] = E[X_+] - E[X_-] = \int_0^\infty P(X > x)\,dx - \int_{-\infty}^0 P(X < x)\,dx \qquad \square$$

Probability notes

Proposition (Markov's inequality). For any non-negative random variable $X : S \to \mathbb{R}$ we have

$$P(X \geq t) \leq \frac{E[X]}{t}, \quad \text{for all } t > 0.$$

Proof.

$$E[X] = E[X|X < t]\,P(X < t) + E[X|X \geq t]\,P(X \geq t)$$

Since X is non-negative, $E[X|X < t]\,P(X < t) \geq 0$.
Also, $E[X|X \geq t] \geq t$.
Thus

$$E[X] \geq E[X|X \geq t]\,P(X \geq t) \geq t\,P(X \geq t). \qquad \square$$

Probability notes

Corollary: Chebyshev's inequality). Let X be a random variable with mean μ and variance σ^2. For any $t > 0$,

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Proof. Apply Markov's inequality to $Y = (X - \mu)^2$.

Chebyshev's inequality is a form of concentration inequality: X *must be close to its mean whenever the variance is small.*

Corollary - Chernoff bound. Let X be a random variable with a moment generating function in a n-hood of zero. For any $t > 0$,

$$P(|X - \mu| \geq t) = P(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \frac{E[e^{\lambda(X-\mu)}]}{e^{\lambda t}}.$$

Proof. Apply Markov's inequality to $Y = e^{\lambda(X-\mu)}$.

Probability notes

The Law of Large Numbers is a consequence of Chebychev's inequality.

Theorem (Law of Large Numbers). Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Proof. Proof follows directly from Chebychev's inequality, after observing that

$$\text{var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) = \frac{\sigma^2}{n}$$

Probability notes

As an application of the Law of Large Numbers, let Z be a d-dimensional random point whose coordinates are each selected from a zero mean, $\frac{1}{2\pi}$ variance Gaussian.
We set such value of the so the Gaussian probability density equals one at the origin and is bounded below throughout the unit ball by a constant.
By the Law of Large Numbers, the square of the distance of Z to the origin will be of the order of d with high probability. In particular, there is vanishingly small probability that such a random point z would lie in the unit ball. This implies that the integral of the probability density over the unit ball must be vanishingly small. On the other hand, the probability density in the unit ball is bounded below by a constant. We thus conclude that the unit ball must have vanishingly small volume.

Proposition (Gaussian tail bounds). Let $X \sim \mathcal{N}(\mu, \sigma^2)$. For all $t > 0$, we have

$$P(|X - \mu| \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

Proof. The moment-generating function is $E[e^{\lambda X}] = e^{\lambda\mu} e^{\lambda^2 \frac{\sigma^2}{2}}$. In fact, for $Y = X - \mu$, a direct calculation shows

$$\begin{aligned} E[e^{\lambda Y}] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} e^{\lambda y - \frac{y^2}{2\sigma^2}}\, dy = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda\sigma z - \frac{z^2}{2}}\, dz \\ &= \frac{e^{\frac{\lambda^2\sigma^2}{2}}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(z-\lambda\sigma)^2}{2}}\, dz = e^{\frac{\lambda^2\sigma^2}{2}} \end{aligned}$$

Using the Chernoff bound, we obtain

$$P(|X - \mu| > t) \leq E[e^{\lambda(X-\mu)}]\, e^{-\lambda t} = e^{-\lambda t} e^{\lambda^2 \frac{\sigma^2}{2}}.$$

Minimizing this expression over λ gives $\lambda = \frac{t}{\sigma^2}$ and thus

$$P(|X - \mu| > t) \leq e^{-\frac{t^2}{2\sigma^2}} \qquad \square$$

Probability notes

Definition. A Random variable X with mean μ is called **sub-Gaussian** if there exists a positive number σ such that

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \quad \text{for all } \lambda \in \mathbb{R}.$$

Any Gaussian random variable with variance σ^2 is also a sub-Gaussian random variable with parameter σ.
In fact, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $E[e^{\lambda(X-\mu)}] = e^{\frac{\sigma^2\lambda^2}{2}}$.

An important example of non-Gaussian but sub-Gaussian random variables are the **Rademacher random variables**.
A Rademacher random variable Y takes on the values ± 1 with equal probability and is sub-Gaussian with parameter $\sigma = 1$.

One can show that any bounded random variable is sub-Gaussian.

Probability notes

Proposition (Sub-Gaussian tail bounds). Let X be a sub-Gaussian random variable with parameter σ. For all $t > 0$, we have

$$P(|X - \mu| \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

Proof. Using the Chernoff bound and the definition, we obtain

$$P(|X - \mu| \geq t) \leq e^{-\lambda t} E[e^{\lambda(X-\mu)}] \leq e^{-\lambda t} e^{\frac{\sigma^2\lambda^2}{2}}$$

Minimizing this expression over λ gives $\lambda = \frac{t}{\sigma^2}$ and thus

$$P(|X - \mu| \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}. \qquad \square$$

Probability notes

Definition. A Random variable X with mean μ is called **sub-exponential** if there exist numbers ν, b such that

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}}, \quad \text{for all } \lambda \leq \frac{1}{b}.$$

A sub-Gaussian random variable is also sub-exponential (set $\nu = \sigma$ and $b = 0$ where $\frac{1}{b}$ is interpreted as ∞).
However, the converse is not true in general.

Let $Z = X^2$, where $X \sim \mathcal{N}(0,1)$. One can show that Z is sub-exponential but is not sub-Gaussian.

Proposition (Sub-exponential tail bounds). Let X be a sub-exponential random variable with parameters ν, b. Then

$$P(|X-\mu| \geq t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{b} \\ e^{-\frac{t}{2b}} & \text{if } t > \frac{\nu^2}{b} \end{cases}$$

Probability notes

Theorem (Master Tail bound). Let $X_1, \ldots, X_n$ are independent random variables with zero mean and variance at most σ^2.
Suppose

(i) $a \in [0, \sqrt{2}n\sigma^2]$;

(ii) s is a positive integers satisfying $s \in [\frac{a^2}{4n\sigma^2}, \frac{n\sigma^2}{2}]$;

(iii) for all i, $|E[X_i^r]| \leq \sigma^2 r!$ for $r = 3, 4, \ldots, s$.

Then

$$P(|\sum_{i=1}^{n} X_i| \geq a) \leq 3\, e^{-\frac{a^2}{12n\sigma^2}}$$

Probability notes

The celebrated **central limit theorem** shows that the limiting distribution of a sum of i.i.d. random variables is always Gaussian.

Lindeberg-Levy Central Limit Theorem. Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Denote

$$S_n = X_1 + X_2 + \cdots + X_n$$

and consider the normalized random variable

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{\text{var}(S_n)}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Then, as $n \to \infty$,

$$Z_n \to \mathcal{N}(0, 1) \quad \text{in distribution.}$$

Probability notes

Concentration inequalities quantifies how much a sum of independent random variables deviates around its mean. Unlike the classical central limit theorem, the concentration inequalities below are non-asymptotic in the sense that they hold for all fixed n and not just as $n \to \infty$.

Hoeffding's inequality. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with mean $E[X_i] = 0$ and satisfying $|X_i| \leq a_i$, for $i = 1, \ldots, n$. Then

$$P\left(|\sum_{i=1}^{n} X_i| > t \right) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2} \right)$$

Remark. The inequality implies that fluctuations larger than $O(\sqrt{n})$ have small probability. For example, if $a_i = a$ for all i, then setting $t = a\sqrt{2n \ln n}$ yields

$$P\left(|\sum_{i=1}^{n} X_i| > a\sqrt{2n \ln n} \right) \leq \frac{2}{n}$$

Probability notes

Bernstein's inequality, uses the variance of the summands to improve over Hoeffding's inequality.

Bernstein's inequality. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables satisfying $|X_i| \leq a$ and $E[X_i^2] = \sigma^2$, for $i = 1, \ldots, n$. Then

$$P\left(|\sum_{i=1}^{n} X_i| > t\right) \leq 2 \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right)$$

Probability notes

Theorem. Almost all the volume of the high-dimensional cube is located in its corners.

Proof. Let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ where each $x_i \in [-\frac{1}{2}, \frac{1}{2}]$ is chosen uniformly at random. The event that x also lies in the sphere means

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \leq 1.$$

Let $z_i = x_i^2$ and observe that

$$E[z_i] = \int_{-\frac{1}{2}}^{\frac{1}{2}} t^2 dt = \frac{t^3}{3}\bigg|_{-\frac{1}{2}}^{\frac{1}{2}} = \frac{1}{12} \quad \Rightarrow \quad E[\|x\|_2^2] = \sum_{i=1}^{d} E[z_i] = \frac{d}{12}.$$

Geometry of high dimensional data

Using Hoeffding's inequality, for sufficiently large d, we have that

$$
\begin{aligned}
P(\|x\|_2 \leq 1) &= P\left(\sum_{i=1}^{d} x_i^2 \leq 1\right) \\
&= P\left(\sum_{i=1}^{d} (z_i - E[z_i]) \leq 1 - \frac{d}{12}\right) \\
&= P\left(\sum_{i=1}^{d} (E[z_i] - z_i) \geq \frac{d}{12} - 1\right) \\
&\leq 2\exp\left(-\frac{(\frac{d}{12} - 1)^2}{2d\,(\frac{1}{6})^2}\right) \\
&\leq 2\,e^{-\frac{d}{8}}
\end{aligned}
$$

As this values goes to 0 when $d \to \infty$, this shows random points in d-cubes are most likely outside the sphere. That is, almost all the volume of a d-cube concentrates in its corners. $\square$

Geometry of high dimensional data

Problem: How to generate random points on a sphere?

Here is an approach when $d = 2$.

To generate a point (x, y), we select x and y coordinates uniformly at random from $[-1, 1]$. This yields points that are distributed uniformly at random in a square that contains the unit circle. We next project these points onto the circle.

The resulting distribution will not be uniform on the circle since more points fall on a line from the origin to a vertex of the square, than fall on a line from the origin to the midpoint of an edge due to the difference in length of the diagonal of the square to its side length.
To remedy this problem, we discard all points outside the unit circle and only project the remaining points onto the circle.

Geometry of high dimensional data

- The above construction fails in higher dimensions.

As we have shown above, the ratio of the volume of $S^{d-1}(1)$ to the volume of $C^d(1)$ decreases rapidly as the dimension d increases.

As a result, for large d, almost all the generated points will be discarded in this process as they lay outside the unit d-ball and we end up with essentially no points inside the d-ball and thus, after projection, with essentially no points on $S^{d-1}(1)$.

- Instead we can proceed as follows.

Recall that the multivariate Gaussian distribution is symmetric about the origin - which is exactly what we need.

Hence, we construct a vector in R^d whose entries are independently drawn from a univariate Gaussian distribution. We then normalize the resulting vector to lie on the sphere. This gives a distribution of points that is uniform over the sphere.

Geometry of high dimensional data

Having a method of generating points uniformly at random on S^{d-1} at our disposal, we can now give a probabilistic proof that **points on S^{d-1} concentrate near its equator.**

Without loss of generality we pick an arbitrary unit vector x_1 which represents the north pole and the intersection of the sphere with the plane $x_1 = 0$ forms our equator.

We extend x_1 to an orthonormal basis $x_1, \ldots, x_d$.

Using the method presented above, we generate random points X on S^{d-1} by fist sampling $(Z_1, \ldots, Z_n) \in \mathcal{N}(0,1)$, and then normalizing $X = (X_1, \ldots, X_d)$ where $X_i = \frac{1}{\sum_{k=1}^{d} Z_k^2} Z_i$.

Geometry of high dimensional data

Since $X \in S^{d-1}$, then $\sum_{k=1}^{d} \langle X, x_k \rangle^2 = 1$
We also have that

$$E[\sum_{k=1}^{d} \langle X, x_k \rangle^2] = E[1] = 1$$

hence, by symmetry, $E[\langle X, x_1 \rangle^2] = \frac{1}{d}$.

By Markov's inequality,

$$P(|\langle X, x_1 \rangle| > \epsilon) = P(|\langle X, x_1 \rangle|^2 > \epsilon^2) \leq \frac{E[\langle X, x_1 \rangle^2]}{\epsilon^2} = \frac{1}{d\epsilon^2}.$$

For fixed ϵ we can make this probability arbitrarily small by increasing the dimension d.

This proves our claim that points on the high-dimensional sphere concentrate near its equator.

Geometry of high dimensional data

Properties of random vectors in high dimensions.

Suppose we generate a vector $(x_1, \ldots, x_n)$ where each coordinate is an independent random variable with zero mean and unit variance. Then

$$E[\|x\|^2] = E\left[\sum_{i=1}^{n} x_i^2\right] = \sum_{i=1}^{n} E[x_i^2] = n.$$

Hence we expect the length $\|x\|$ of x is $\sqrt{n}$.

This does not imply that the typical length is about $\sqrt{n}$. For that, we need to derive a concentration inequality.

Geometry of high dimensional data

We assume that the coordinates x_i of the vector $(x_1, \dots, x_n)$ are $x_i \sim \mathcal{N}(0,1)$.

It follows that $Z = \sum_{i=1}^n x_i^2$ has a χ^2 distribution with n degrees of freedom.

It turns out that Z is sub-exponential with parameters $(2\sqrt{n}, 4)$. Hence, using the sub-exponential tail bounds formula, we have

$$P(|\frac{1}{n}\sum_{i=1}^n x_i^2 - 1| \geq t) \leq \begin{cases} 2e^{-\frac{nt^2}{8}} & \text{if } 0 < t \leq 1 \\ 2e^{-\frac{nt}{8}} & \text{if } t > 1 \end{cases} \quad \leq 2e^{-\frac{n}{8}\min(t,t^2)}$$

Geometry of high dimensional data

Observation: Two randomly drawn vectors in high dimensions are almost perpendicular.

The angle $\theta_{x,y}$ between two vectors x and y in $\mathbb{R}^d$ satisfies

$$\cos\theta_{x,y} = \frac{\langle x, y\rangle}{\|x\|\|y\|}$$

Theorem. Let $x, y \in \mathbb{R}^d$ be two random vectors with i.i.d. Rademacher variables (that is, the entries x_i, y_i take values ± 1 with equal probability).
Then

$$P\left(|\cos\theta_{x,y}| \geq \sqrt{\frac{2\ln d}{d}}\right) \leq \frac{2}{d}$$

Proof. Observe that $\langle x, y\rangle = \sum_i x_i y_i$ is a sum of i.i.d. Rademacher variables, hence $E[\langle x, y\rangle] = \sum_i E[x_i y_i] = 0$. By the Hoeffding's inequality

$$(\text{Recall: } P(|\sum_{i=1}^{d} X_i| > a\sqrt{2d \ln d}) \leq \tfrac{2}{d})$$

observing that $a = |x_i y_i| \leq 1$ we have

$$P(|\frac{\langle x, y\rangle}{\|x\|\|y\|}| > \sqrt{\frac{2 \ln d}{d}}) = P(|\langle x, y\rangle| > \sqrt{2d \ln d}) \leq \tfrac{2}{d} \qquad \square$$

Remark. A similar result holds for Gaussian random vectors in R^d or random vectors chosen from the sphere S^{d-1}.

Geometry of high dimensional data

Remark. Let $x_1, x_2, \ldots, x_m$ be random vectors whose entries are i.i.d. Rademacher variables. By refining the argument in the proof above, we obtain that for any pair of vector x_i, x_j,

$$P\left(|\cos\theta_{x_i,x_j}| \geq \sqrt{\frac{2\ln c}{d}}\right) \leq \frac{2}{c},$$

where $c > 0$ is a constant.
By choosing $m = \sqrt{c}/4$ (using the union bound) we have that with high probability

$$\max_{i,j,i\neq j} |\cos\theta_{x_i,x_j}| \leq \sqrt{\frac{2\ln c}{d}}$$

If we choose $c = e^{d/200}$, then any two vectors are almost orthogonal in the sense that $|\cos\theta_{x_i,x_j}| \leq \frac{1}{10}$.

Geometry of high dimensional data

Gaussians in High Dimension
A one-dimensional Gaussian has its mass close to the origin. However, the behavior is different when the dimension d increases.

The d-dimensional spherical Gaussian with zero mean and variance σ^2 in each coordinate has density function

$$p(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \, e^{-\frac{|x|^2}{2\sigma^2}}$$

The value of the density is maximum at the origin, but there is very little volume there.

When $\sigma = 1$, integrating the probability density over a unit ball centered at the origin yields almost zero mass, since the volume of such a ball is negligible.
One needs to increase the radius of the ball to about $\sqrt{d}$ before there is a significant volume.

Geometry of high dimensional data

Theorem (Gaussian Annulus Theorem) Let $p(x)$ be a d-dimensional spherical Gaussian with unit variance in each direction. For any $\beta \leq \sqrt{d}$

$$\int_{\sqrt{d}-\beta \leq |x| \leq \sqrt{d}+\beta} p(x)\, dx \geq 1 - 3e^{-c\beta^2},$$

where c is a fixed positive constant.

The Gaussian Annulus Theorem states that volume concentrates about a thin annulus of radius $\sqrt{d}$.

More precisely, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta$.

Note that $E(|x|^2) = \sum_{i=1}^{d} |x_i|^2 = d$, hence the mean squared distance of a point from the center is d.

Geometry of high dimensional data

Proof. Let $x = (x_1, \ldots, x_d)$ be a point selected from a unit variance Gaussian centered at the origin and let $r = |x|$.

The domain of integration can be expressed as $|r - \sqrt{d}| \leq \beta$
We examine the complementary region $|r - \sqrt{d}| > \beta$
If $|r - \sqrt{d}| > \beta$ then

$$|r^2 - d| = |r + \sqrt{d}||r - \sqrt{d}| \geq (r + \sqrt{d})\beta \geq \beta\sqrt{d} \tag{1}$$

We have

$$\begin{aligned} |r^2 - d| &\geq \beta\sqrt{d} \\ |x_1^2 + \ldots + x_d^2 - d| &\geq \beta\sqrt{d} \\ |(x_1^2 - 1) + \ldots + (x_d^2 - 1)| &\geq \beta\sqrt{d} \\ |w_1 + \ldots + w_d| &\geq \frac{\beta\sqrt{d}}{2} \end{aligned}$$

where, in the last step, we used the change of variable $w_i = \frac{x_i^2 - 1}{2}$

Note that $E[w_i] = \frac{1}{2}E[x_i^2 - 1] = \frac{1}{2}(E[x_i^2] - 1) = 0$

Geometry of high dimensional data

In order to apply the Master Tail Bound theorem, we verify the bound on high order moments.

Let s be a positive integer. If $|x_i| \leq 1$, then $|x_i^2 - 1|^s \leq 1$ and, if $|x_i| > 1$, then $|x_i^2 - 1|^s \leq |x_i|^{2s}$.

It follows that

$$|w_i|^s = \left(\frac{|x_i^2 - 1|}{2}\right)^s \leq \frac{1 + x_i^{2s}}{2^s}.$$

Using the last inequality, we have

$$\begin{aligned}
|E[w_i^s]| \leq 2^{-s} E(1 + x_i^{2s}) &= 2^{-s}\left(1 + E(x_i^{2s})\right) \\
&= 2^{-s} + 2^{-s}\sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-\frac{x^2}{2}}\, dx \\
&\leq s! \qquad \text{[using the Gamma integral]}
\end{aligned}$$

Geometry of high dimensional data

From the calculation above, we have $var(w_i) = E[w_i^2] \leq 2$.
This implies:

$$|E[w_i^s]| \leq 2s! := \sigma^2 s!$$

where $\sigma^2 = 2$ is the bound on the variance of the variables w_i.

We can now apply the Master Tail Bound theorem with $\sigma^2 = 2$ (according to the notation of the Theorem where σ^2 denotes the bound on the variance of the random variables w_i) to obtain

$$P(|w_1 + \ldots + w_d| \geq \frac{\beta\sqrt{d}}{2}) \leq 3\, e^{-\frac{\beta^2}{96}} \qquad \square$$

Geometry of high dimensional data

Random Projections.

Nearest neighbor search routines are frequently used in applications.

In nearest neighbor search, we are given a database of n points in $\mathbb{R}^d$ where n and d are usually large. The task is to find the nearest or approximately nearest database point to a query point.

To speed up the search, it is convenient to reduce the dimensionality of the problem by projecting

$$f : \mathbb{R}^d \to \mathbb{R}^k, \qquad k \ll d$$

This should be carried out while maintaining the geometry of the problem. That is, if points were close in $\mathbb{R}^d$ then they should remain close in $\mathbb{R}^k$.

We will see, using the Gaussian Annulus Theorem, that such a projection exists and is simple to compute.

Geometry of high dimensional data

Let $u_1, \ldots, u_k$ be independent random vectors in $\mathbb{R}^d$ drawn from the spherical Gaussian with unit variance.

For any $v \in \mathbb{R}^d$, we define the projection $f : \mathbb{R}^d \to \mathbb{R}^k$ by

$$f(v) = (u_1 \cdot v, \ldots, u_k \cdot v).$$

We will show that, with high probability, $|f(v)| \approx \sqrt{k}|v|$.

If this is the case, it follows that if we want to measure $|v_1 - v_2|$, we can compute

$$|f(v_1) - f(v_2)| = |f(v_1 - v_2)| \approx \sqrt{k}|v_1 - v_2|$$

Geometry of high dimensional data

Theorem (Random Projection Theorem) Let $v \in \mathbb{R}^d$ and the projection f be defined as above. There exists $c > 0$ s.t., for any $\epsilon \in (0,1)$,

$$P\left(\left||f(v)| - \sqrt{k}|v|\right| \geq \epsilon\sqrt{k}|v|\right) \leq 3e^{-ck\epsilon^2}$$

where P is taken over the random draws of the vectors u_i.

Proof. By rescaling both sides of the inequality by $|v|$, we can assume $|v| = 1$. We observe that, for each $i = 1, \ldots, k$,

$$u_i \cdot v = \sum_{j=1}^{d} u_{ij} v_j$$

has Gaussian density zero mean and variance 1; in particular, follows that

$$var(u_i \cdot v) = var(\sum_{j=1}^{d} u_{ij} v_j) = \sum_{j=1}^{d} var(u_{ij}) v_j^2 = \sum_{j=1}^{d} v_j^2 = |v^2| = 1$$

Geometry of high dimensional data

Since $u_1 \cdot v, \ldots, u_k \cdot v$ are independent Gaussian random variables, $f(v)$ is a random vector from a k-dimensional spherical Gaussian with unit variance in each coordinate.

The proof is completed by applying the Gaussian Annulus Theorem with $d = k$ and $\beta = \epsilon\sqrt{k}$. $\square$

Geometry of high dimensional data

Theorem (Johnson-Lindenstrauss Lemma) For any $0 < \epsilon < 1$ and any integer n, let $k \geq \frac{3}{c\epsilon^2} \log n$, where c is as in the Random Projection Theorem. For any set of n points in $\mathbb{R}^d$, the random projection $f : \mathbb{R}^d \to \mathbb{R}^k$ defined above has the property that, for any pair $v_i, v_j \in \mathbb{R}^d$, with probability at least $1 - \frac{3}{2n}$,

$$(1-\epsilon)\sqrt{k}|v_i - v_j| \leq |f(v_i) - f(v_j)| \leq (1+\epsilon)\sqrt{k}|v_i - v_j|.$$

Proof. Observe that $f(v_i) - f(v_j) = f(v_i - v_j)$.
Inequalities above are equivalent to

$$|f(v_i)-f(v_j)|-\sqrt{k}|v_i-v_j| = |f(v_i-v_j)|-\sqrt{k}|v_i-v_j| \geq \epsilon\sqrt{k}|v_i-v_j|.$$

By applying the Random Projection Theorem

$$P(|f(v_i - v_j)| - \sqrt{k}|v_i - v_j| \geq \epsilon\sqrt{k}|v_i - v_j|) \leq 3\, e^{-cl\epsilon^2} \leq \frac{3}{n^3}$$

Hence, for $\binom{n}{2} < \frac{n^2}{2}$ pairs of points, the probability that the above inequality holds for any pair of points is less than $\frac{3}{n^3}\,\frac{n^2}{2} = \frac{3}{2n}$. $\square$

Geometry of high dimensional data

High-dimensional data analysis: bibliography

1. Avrim Blum, John Hopcroft, Ravindran Kannan. *Foundations Of Data Science.* Cambridge University Press, 2020.
2. David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality, AMS Conference on Math challenges of the 21st century, 2000.
3. Michael Mitzenmacher and Eli Upfal. *Probability and Computing - Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.
4. Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, 2018