MATH 6397 - Mathematics of Data Science

Instructor: Demetrio Labate

March 2, 2023

Course Outline

1. Mathematics of machine learning
   1.1 Mathematics of data science
   1.2 Geometry of high dimensional data

Geometry of high dimensional data

Useful references about geometry of high dimensional data:

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
5. Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019

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

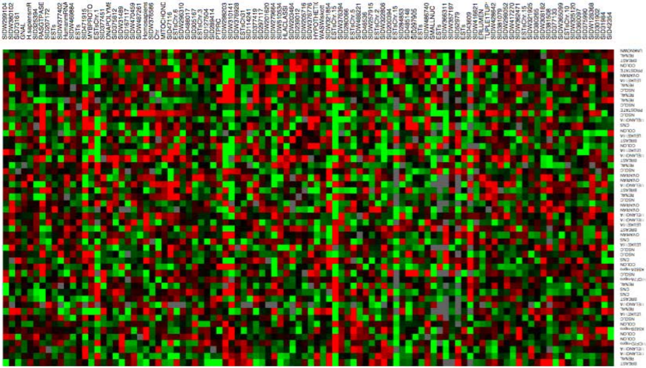


Figure: Computational biology. DNA screening with a few observations and huge number of variables.

Mathematics of data science

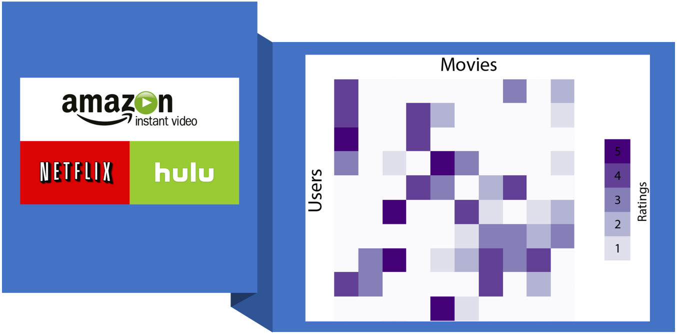


Figure: Netflix challenge (cf. *Netflix Prize*, 2006-2011): to predict users ratings from a sparse incomplete database of ratings given by millions of users on thousands of movies or TV shows.

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.

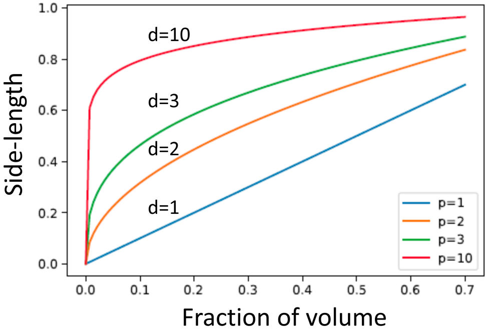
A modest increase in dimensions results in a dramatic increase in required data points to cover the space at the same density.

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

Geometry in high dimensions $\longrightarrow$ **probability**

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.

Pros: Since high-dimensional spaces are sparser, it can 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:

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

- d-dimensional hypersphere of radius R:

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

- 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 $\mathbb{B}^d(R)$ is given by

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

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

Proof. Using polar coordinates,

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

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

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

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

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

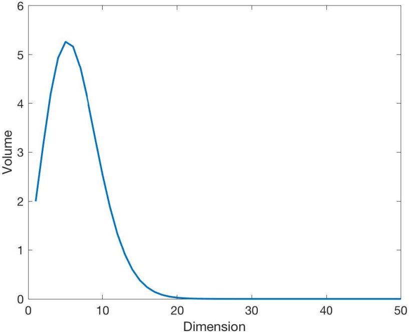
Geometry of high dimensional data

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)

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



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

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

Geometry of high dimensional data

Observation: The volume of a *d*-ball concentrates 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**.

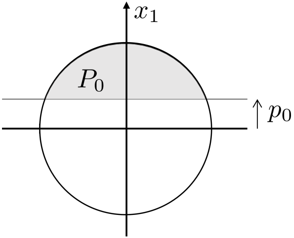
Geometry of high dimensional data

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,

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

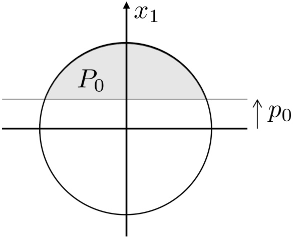


Theorem

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

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

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

Thus

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

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

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

Geometry of high dimensional data

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

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

Thus, from the inequality in page above, we have

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

and, finally,

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

Geometry of high dimensional data

Observation: The volume of a d-ball concentrates on its outer shell.

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

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

Since, for any $\epsilon > 0$, this quantity tends to 0 as $d \to \infty$, it follows that the spherical shell contained between $\mathbb{B}^d(1)$ and $\mathbb{B}^d(1-\epsilon)$ contains most of the volume of $\mathbb{B}^d(1)$, 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.

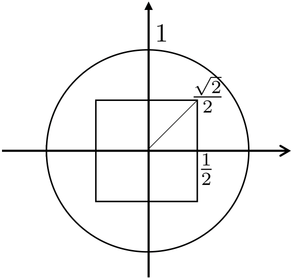
Geometry of high dimensional data

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.)

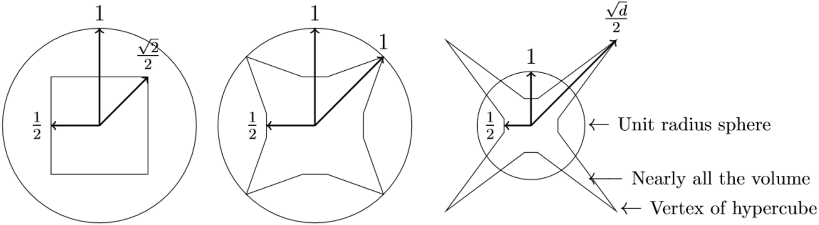


Figure: Relationship between the sphere and the cube in dimensions $d = 2$, $d = 4$ and higher d.

Probability notes

We will discuss below some useful probability estimates.

While results such as the classical Central Limit Theorem provide an **asymptotic** estimate, which is valid when we consider a sum of n random variables with n approaching infinity, concentration inequalities are **non-asymptotic** as they hold for all fixed n.

Concentration inequalities quantifies how much a random variable X deviates around its mean.

One way in which to control a tail probability $P(X \geq t)$ is by controlling the moments of the random variable X. Gaining control of higher-order moments leads to correspondingly sharper bounds on tail probabilities, ranging from Markov's inequality (which requires only existence of the first moment) to the Chernoff bound (which requires the existence of the moment generating function).

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

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

Then, as $n \to \infty$,

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

Probability notes

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 $Y : S \to \mathbb{R}$ we have

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

Proof. Take any $t > 0$.

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

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

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

Probability notes

Proposition (Markov's inequality)

For any non-negative random variable $Y : S \to \mathbb{R}$ we have

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

Proof. For any $t > 0$, the following holds:

$$\begin{aligned}
E[Y] &= \int_{\mathbb{R}} y\, f(y)\, dy \\
&= \int_0^{\infty} y\, f(y)\, dy \qquad (Y \text{ is positive}) \\
&\geq \int_t^{\infty} y\, f(y)\, dy \\
&\geq \int_t^{\infty} t\, f(y)\, dy \\
&= t\, P(Y \geq t) \qquad \square
\end{aligned}$$

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. By applying Markov's inequality to $Y = (X - \mu)^2$, we have

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

The proof follows by observing that

$$E[Y] = E[(X - \mu)^2] = var(X) = \sigma^2 \qquad \square$$

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

Probability notes

Corollary (Chernoff bound)

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

$$P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t} E[e^{\lambda X}] \quad \text{for } \lambda > 0$$
$$P(X \leq t) = P(e^{\lambda X} \leq e^{\lambda t}) \leq e^{-\lambda t} E[e^{\lambda X}] \quad \text{for } \lambda < 0$$

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

Note: $E[e^{\lambda X}]$ is the moment generating function $M_X(\lambda)$ of X.

Probability notes

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

Theorem (Law of Large Numbers)

Let $X_1, X_2, \dots, X_n$ 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 Chebyshev's inequality, after observing that

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

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.

Probability notes

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

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

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 $E[e^{\lambda(X-\mu)}] = e^{\frac{\sigma^2\lambda^2}{2}}$.

Just as the property of Gaussianity is preserved by linear operations so is the property of sub-Gaussianity.
For instance, if X_1, X_2 are independent sub-Gaussian variables with parameters σ_1 and σ_2, then a simple calculation shows that $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Probability notes

An important example of non-Gaussian but sub-Gaussian random variables are the Rademacher random variables.

Example (Rademacher random variables)

A Rademacher random variable Y takes on the values ± 1 with equal probability and is sub-Gaussian with parameter $\sigma = 1$.

By computing the moment generating function and using the Taylor series expansion for the exponential, we have

$$\begin{aligned} E[e^{\lambda Y}] = \tfrac{1}{2}(e^{\lambda} + e^{-\lambda}) &= \tfrac{1}{2}\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\ &\leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k\, k!} \\ &= e^{\frac{\lambda^2}{2}} \end{aligned}$$

Probability notes

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

Example (Bounded random variables)

Let X be a zero-mean random variable, supported on some interval $[a, b]$. Then X is sub-Gaussian with parameter at most $\sigma = b - a$. In fact the estimate can be sharpened to show that the parameter is at most $\sigma = \frac{b-a}{2}$.

Probability notes

Proposition (Sub-Gaussian tail bounds)

Let X be a sub-Gaussian random variable with parameter σ. For all $t > 0$, we have

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

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

Proof. Using the Chernoff bound and the definition of sub-Gaussian, 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,

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

As the variable $-X$ is also sub-Gaussian we also have

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

which, combined with the other inequality, gives the second statement. □

Probability notes

Using the sub-Gaussian tail bounds and the properties of sub-Gaussianity, we have the following result.

Hoeffding's inequality

Let $X_1, X_2, \ldots, X_n$ be independent sub-Gaussian random variables with mean $E[X_i] = \mu_i$ and sub-Gaussian parameter σ_i, for $i = 1, \ldots, n$. Then

$$P\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right)$$

Probability notes

The Hoeffding's inequality is often stated only for the special case of bounded random variable

Hoeffding's inequality

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

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

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

The notion of sub-Gaussianity is fairly restrictive, so that it is natural to consider various relaxations of it.

Definition

A Random variable X with mean μ is called **sub-exponential** if there exist non-negative quantities 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.

To see that, set $\nu = \sigma$ and $b = 0$ where $\frac{1}{0}$ is interpreted as ∞.

Probability notes

There are sub-exponential random variables that are not sub-Gaussian.

Example

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

For $\lambda < \frac{1}{2}$, noticing that $E[X] = E[Z^2] = 1$, we have

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

For $\lambda > \frac{1}{2}$, the moment generating function is infinite, showing that X is not sub-Gaussian. Also, calculus-type inequalities give

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2} \quad \text{for } |\lambda| < \frac{1}{4}$$

showing that X is sub-exponential with parameters $(\nu, b) = (2, 4)$.

Probability notes

Proposition (Sub-exponential tail bounds)

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

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

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

The proof relies on the Chernoff bound, similar to the proof of the sub-Gaussian tail bounds.

Probability notes

The sub-exponential property can be verified by explicitly computing or bounding the moment-generating function. This direct calculation may be impractical in many settings.
An alternative method is provided by controlling the polynomial moments of the random variable.

Definition

Given a random variable X with mean $\mu = E[X]$ and variance σ^2, we say that **Bernstein's condition with parameter** b holds if

$$|E[(X-\mu)^k] \leq \tfrac{1}{2} k! \sigma^2 b^{k-2} \qquad \text{for } k = 2, 3, \dots$$

One sufficient condition for Bernstein's condition to hold is that X be bounded; in particular, if $|X - \mu| \leq b$, then it is straightforward to verify the condition above.

Probability notes

Proposition

If a random variable X satisfies the Bernstein condition with parameter b, then it is sub-exponential with parameters determined by $\nu = \sigma^2$ and b.

Proof. Using the Bernstein's condition

$$\begin{aligned} E[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{E[(X-\mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2} \\ &= 1 + \frac{\lambda^2\sigma^2}{2} \sum_{k=2}^{\infty} (|\lambda| b)^{k-2} \end{aligned}$$

For $|\lambda| < 1/b$, we can sum the geometric series to obtain

$$E[e^{\lambda(X-\mu)}] \leq 1 + \frac{\lambda^2\sigma^2}{2} \frac{1}{1-b|\lambda|} \leq \exp \frac{\lambda^2\sigma^2/2}{1-b|\lambda|}$$

Probability notes

It follows that

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}} \quad \text{for } |\lambda| < \tfrac{1}{2b},$$

showing that X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$. □

Probability notes

The following result follows from the Proposition above and the sub-exponential tail bounds.

Theorem (Bernstein-type bound)

For any random variable X satisfying the Bernstein condition with parameter b we have

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \qquad \text{for } |\lambda| < \tfrac{1}{b}.$$

Additionally,

$$P(|X-\mu| > t) \leq 2\,e^{-\frac{t^2}{\sigma^2+bt}} \qquad \text{for all } t > 0.$$

Probability notes

Here is another variant of the Bernstein-type bounds (requiring a slightly different argument)

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

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

Note that Bernstein's inequality uses the variance of the summands to improve the tail estimate over Hoeffding's inequality.

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) for all i, $|E[X_i^r]| \leq \sigma^2 r!$ for $r = 3, 4, \ldots, \lfloor \frac{a^2}{4n\sigma^2} \rfloor$.

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

[Sketch of the proof] Apply Markov's inequality to X^r where r is a large even number. Since r is even, x^r is nonnegative, and thus $P(|X| > a) = P(X^r > a^r) \leq E(X^r)/a^r$. If $E(X^r)$ is not too large, we will get a good bound. To compute $E(X^r)$, write $E(X)$ as $E(X_1 + \cdots + X_n)^r$ and expand the polynomial into a sum of terms. Using independence $E(X_i^{r_i} X_j^{r_j}) = E(X_i^{r_i})E(X_j^{r_j})$ so we get a collection of simpler expectations that can be bounded using our assumption that $|E[X_i^r]| \leq \sigma^2 r!$

Geometry of high dimensional data

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

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

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. □

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 $\mathbb{S}^{d-1}(1)$ 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 $\mathbb{S}^{d-1}(1)$.

- 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 $\mathbb{R}^d$ 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 $\mathbb{S}^{d-1}$ at our disposal, we can now give a probabilistic proof that **points on $\mathbb{S}^{d-1}$ 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 $\mathbb{S}^{d-1}$ 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 $X \in \mathbb{S}^{d-1}$, 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 $x = (x_1, \ldots, x_n)$ 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 $(x_1, \ldots, x_n)$ 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

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

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

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

Geometry of high dimensional data

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 $\mathbb{R}^d$ or random vectors chosen from the sphere $\mathbb{S}^{d-1}$.

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.

Geometry of high dimensional data

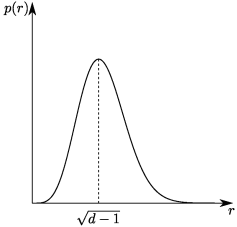
Denoting by $r = \|x\|$ the ℓ^2 distance from the center of the Gaussian, the integral

$$\int_0^1 p(r)\, dr$$

is vanishing small as d increases.

In fact we have that

$$p(r) \approx r^{d-1} e^{r^2/2}$$



We estimate the maximum by setting the derivative to zero

$$\tfrac{d}{dr} p(r) = \tfrac{d}{dr} r^{d-1} e^{r^2/2} = (d-1) r^{d-2} e^{r^2/2} - r^d e^{r^2/2} = 0,$$

showing the maximum occurs at $r = \sqrt{d-1}$

Mass is concentrated about $r \approx \sqrt{d}$.

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 $\sqrt{d}$.

Specifically, 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

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

We have

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

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

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

Using the last inequality, we have

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

With the change of variable $z = \frac{x^2}{2}$, the parenthesis term becomes

$$1 + \tfrac{2^s}{\sqrt{\pi}} \int_0^\infty z^{s-1/2} e^{-z}\, dz = 1 + \tfrac{2^s}{\sqrt{\pi}} \Gamma(s + \tfrac{1}{2}) = 1 + 2^s \prod_{j=0}^{s-1} (j + \frac{1}{2})$$

which can be bound by $2^s s!$

Geometry of high dimensional data

Hence we have

$$|E[w_i^s]| \leq s!$$

which, for the special case $s = 2$, gives that $var(w_i) = E[w_i^2] \leq 2$. 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

Application: Mixture of Gaussians

Problem

Given a mixture of two Gaussian densities

$$p(x) = w_1\, p_1(x) + w_2\, p_2(x), \quad w_1 + w_2 = 1$$

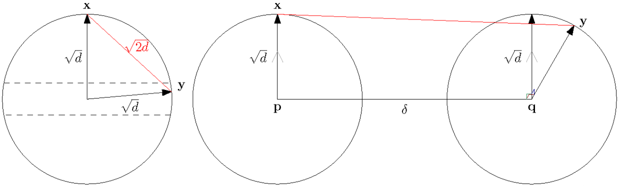
under what conditions the two Gaussians separable?

We claim that the means of the d-dimension spherical unit-variance Gaussians need to be separated by $\Omega(d^{1/4})$.

The idea is that, with high probability, points in the same cluster belong to the same Gaussian because most of the points are concentrated according to the Gaussian Annulus Theorem.

Geometry of high dimensional data

Suppose to randomly select $x, y \sim \mathcal{N}(\mu, I)$ from the same Gaussian.



Observe that most probability mass lies in an annulus of width $O(1)$ and radius $\approx \sqrt{d}$. Rotate the coordinate system so that x is at the North pole. With high probability, y is in the slab $\{(y_1, \dots, y_d) : |y_1| < c\}$ for some $c = O(1)$.
Hence y is nearly orthogonal to x and $\|x - y\| \approx \sqrt{\|x\|^2 + \|y\|^2}$.
Since

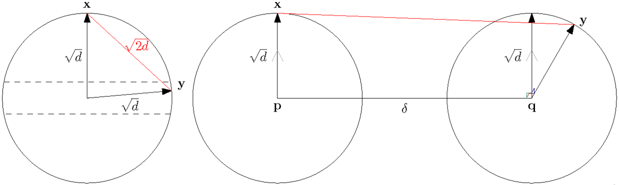
$$x = (\sqrt{d} \pm O(1), 0, \dots, 0), \quad y = (\pm O(1), \sqrt{d} \pm O(1), 0, \dots, 0)$$

then

$$\|x - y\|^2 = (d \pm \sqrt{d}) + (d \pm \sqrt{d}) = 2d \pm \sqrt{d}$$

Geometry of high dimensional data

Suppose now to randomly select $x \sim \mathcal{N}(p, I)$, $y \sim \mathcal{N}(q, I)$ from different Gaussians.



With high probability, x and y lie in an annulus of width $O(1)$ and radius $\approx \sqrt{d}$ centered at p and q respectively.
Also, $(x-p), (p-q), (q-y)$ are nearly mutually perpendicular.
Hence,

$$\|x-y\|^2 \approx \|x-p\|^2 + \|p-q\|^2 + \|q-y\|^2 = 2d \pm O(\sqrt{d}) + \|p-q\|^2$$

Thus if $\|p-q\|^2 = \Omega(\sqrt{d})$ we can separate points

Geometry of high dimensional data

Random Projections.
Nearest neighbor search routines are frequently used in applications.

In nearest neighbor search problem, we are given a set 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

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

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 apply the Gaussian Annulus Theorem to show such a projection exists and is simple to compute.

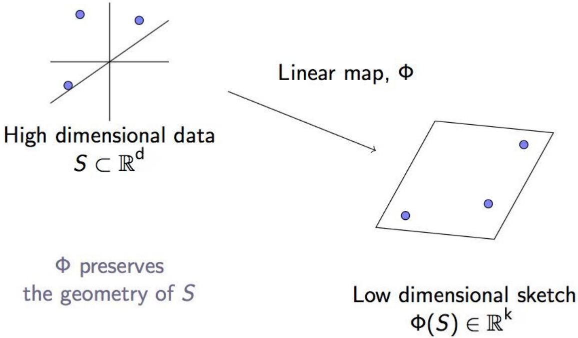
Geometry of high dimensional data

Let $u_1, \ldots, u_k$ be independent random vectors in $\mathbb{R}^d$ drawn from the spherical Gaussian with unit variance $\mathcal{N}(0, I)$.

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

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

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



Linear map, Φ

High dimensional data $S \subset \mathbb{R}^d$

Φ preserves the geometry of S

Low dimensional sketch $\Phi(S) \in \mathbb{R}^k$

To check if v' is close to v in $\mathbb{R}^d$, then it is sufficient to compute

$$|\Phi(v) - \Phi(v')| = |\Phi(v - v')| \approx \sqrt{k}|v - v'|$$

Geometry of high dimensional data

Theorem (Random Projection Theorem)

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

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

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 $|v| = 1$. In particular, for each $i = 1, \ldots, k$,

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

is normally distributed with zero mean and variance 1.

Geometry of high dimensional data

In fact, we have that

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

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

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

$$\int_{(1-\epsilon)\sqrt{k}v<|x=\Phi(v)|<(1+\epsilon)\sqrt{k}v} p(x)\, dx \leq 1 - 3\, e^{-ck\epsilon^2}$$

□

Geometry of high dimensional data

Theorem (Johnson-Lindenstrauss Lemma)

For any $0 < \epsilon < 1$ and any $n > 0$, 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 $f(v_i) - f(v_j) = f(v_i - v_j)$.
The inequality above is 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, we have

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

provided $k \geq \frac{3 \ln n}{c\epsilon^2}$. Hence, there are $\binom{n}{2} < \frac{n^2}{2}$ pairs of points, the probability that the above inequality holds for any pair of points (union bound) is less than $\frac{3}{n^3} \frac{n^2}{2} = \frac{3}{2n}$. □

Geometry of high dimensional data

Despite the dimensionality reduction, the application of the Johnson-Lindenstrauss Lemma is still computationally expensive.

After we draw the random projection matrix, say $M \in \mathbb{R}^{d \times k}$, for each data point $v \in \mathbb{R}^d$, we have to compute Mx which has a computational cost of

$$O(\epsilon^{-2} \log(n) d)$$

since M has kd entries and $k = O(\epsilon^{-2} \log(n))$.

In some applications this might be too expensive, raising the natural question of whether one can do better. Moreover, storing a large-scale dense matrix M is not very desirable either

We might try to replace the dense random matrix M by a sparse matrix M_S.

Geometry of high dimensional data

We consider a sparse $m \times k$ matrix M_S where each row of M_S has just one single non-zero entry of value $\sqrt{k/d}$ at a location drawn uniformly at random.
It follows that for any $x \in \mathbb{R}^d$

$$E_i[(M_S x)_i^2] = \sum_{j=1}^{k} P(i=j)\,\frac{k}{m}\,x_j^2 = \frac{1}{m}\|x\|_2^2$$

Hence

$$E[\|M_S x\|_2^2] = E[\sum_{i=1}^{m}(M_S x_i)^2] = \|x\|_2^2$$

This result show M_S is satisfactory with respect to expectation. However it is not with respect to the variance.

Geometry of high dimensional data

If one coordinate of x is much larger (in absolute value) than all its other coordinates, then we will need a rather large value for k to guarantee that $\|M_S x\|_2 \approx \|x\|_2$.

We can quantify the "peakiness" of a vector via the peak-to-average ratio measured by the quantity $\frac{\|x\|_\infty}{\|x\|_2}$.
It is easy to see that - assuming x is not the zero-vector - we have

$$\frac{1}{\sqrt{d}} \leq \frac{\|x\|_\infty}{\|x\|_2} \leq 1$$

The upper bounds is achieved by vectors with only one non-zero entry, while the lower bound is met by constant-modulus vectors.
Thus, if we have $\frac{\|x\|_\infty}{\|x\|_2} \approx \frac{1}{\sqrt{d}}$ we can hope that sparse subsampling of x will preserve its Euclidean norm.
This suggests to include a preprocessing step by applying a rotation so that sparse vectors become non-sparse in the new basis, thereby reducing their ∞ norm (while their 2-norm remains invariant under rotation)

Geometry of high dimensional data

Definition

The **Fast Johnson-Lindenstrauss Transform** is the map $\Psi : \mathbb{R}^d \to \mathbb{R}^k$, defined by $\Psi = M_S F D$ where M_S and D are random matrices and F is a deterministic matrix. In particular:

- M_S is a $k \times d$ matrix, where each row of M_S has just one single non-zero entry of value $\sqrt{k/d}$ at a location drawn uniformly at random.
- F is either the $d \times d$ DFT matrix or the $d \times d$ Hadamard matrix (if it exists), in each case normalized by $1/\sqrt{d}$ to obtain a unitary matrix.
- D is a $d \times d$ diagonal matrix whose entries are drawn independently from $\{-1, +1\}$ with probability $1/2$.

Geometry of high dimensional data

Theorem (Fast Johnson-Lindenstrauss Transform)

For any $\epsilon > 0$, there is a random matrix Ψ of size $k \times d$ with $k = O(\frac{1}{\epsilon^2} \log \frac{d}{\delta} \log \frac{1}{\delta})$ such that, for each $x \in \mathbb{R}^d$

$$(1-\epsilon)\|x\|_2 \leq \|\Psi x\|_2 \leq (1+\epsilon)\|x\|_2$$

holds with probability at least $1-\delta$.
Matrix-vector multiplication with Ψ takes $O(d \log d + k)$ operations.

Geometry of high dimensional data

The proof of the Fast Johnson-Lindenstrauss Transform Theorem follows from the two lemmas below.

We first show that with high probability the random rotation FD produces vectors with a sufficiently low peak-to-average ratio.

Lemma

Let $y = FDx$, where F and D are as in the definition above. Then

$$P\left(\frac{\|y\|_\infty}{\|y\|_2} \geq \frac{2\log(4d/\delta)}{d}\right) \leq \frac{\delta}{2}$$

Next we apply the following result.

Lemma

Conditioned on the event that $\|y\|_\infty \gtrsim \frac{2\log(4d/\delta)}{d}$, it holds that

$$P\left(\|M_S y\|_2^2 - 1 \leq \epsilon\right) \leq 1 - \frac{\delta}{2}$$

Geometry of high dimensional data

The above results show that, to preserve the distances between n points up to ϵ accuracy, it suffices to randomly project them to $k = O(\epsilon^{-2}\log(n))$ dimensions.
This follows from the observation that a random projection approximately preserves the norm of every point in a set S if it projects into $k = O(\epsilon^{-2}\log|S|)$ dimensions.

Questions:

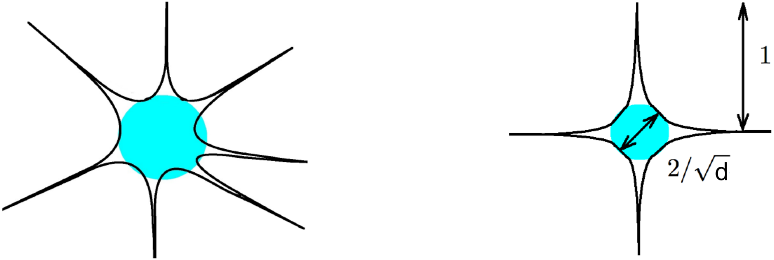
- Can we improve this estimate if S has a special structure?
- How can we measure the complexity of S in a way that explains how many dimensions one needs to project on and still approximately preserve the norms of points in S?

As we have seen above, the geometry of sets in high dimensions is often counter-intuitive.

Geometry of high dimensional data

How can we measure the **complexity of a set** S in $\mathbb{R}^d$?

Convex bodies consist of two parts: the “bulk” and the “outliers”, where the bulk makes up most of the volume, but has small diameter (usually looks like a ball); the outliers contribute little to volume but are large in diameter.



(a) A general convex set

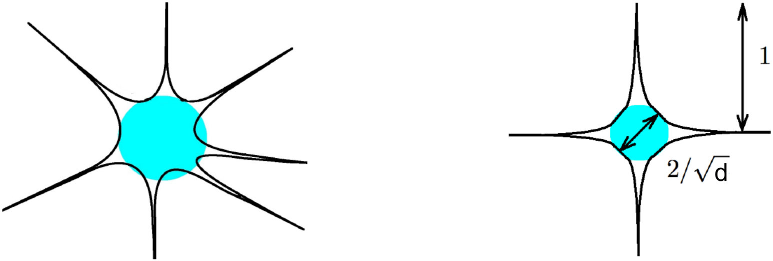
(b) The ℓ_1 ball

Geometry of high dimensional data

For instance, the Euclidean ball $\mathbb{B}_2 \in \mathbb{R}^d$ inscribed within the ℓ^1 ball $\mathbb{B}_1^d = \{x \in \mathbb{R}^d : ||x||_1 \leq 1\}$ and has radius $1/\sqrt{d}$ but

$$vol(\mathbb{B}_2)^{1/d} \asymp vol(\mathbb{B}_1)^{1/d} \asymp \frac{1}{d}$$

indicating that the ball $\mathbb{B}_2$, perhaps inflated by a constant factor, forms the bulk of $\mathbb{B}_1$. The outliers of $\mathbb{B}_1$ are the spikes shown in the figure, which extend far beyond $\mathbb{B}_2$ in the coordinate directions.



(a) A general convex set

(b) The ℓ_1 ball

Geometry of high dimensional data

Let us compare the unit ℓ_1- and ℓ_∞-balls

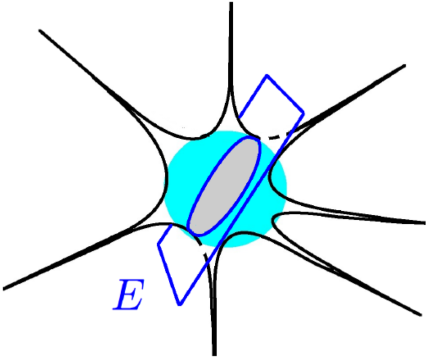
$$
\begin{aligned}
\mathbb{B}_1^d &= \{x \in \mathbb{R}^d : ||x||_1 \leq 1\} \\
\mathbb{B}_\infty^d &= \{x \in \mathbb{R}^d : ||x||_\infty \leq 1\},
\end{aligned}
$$

Though these balls have the same unit radius, the ℓ_1-ball $\mathbb{B}_1^d$ has $2d$ vertices whereas the ℓ_∞-ball B_∞^d has 2^d vertices.
The polytope $\mathbb{B}_\infty^d$ is significantly more complex than $\mathbb{B}_1^d$.

Geometry of high dimensional data

To capture the complexity of a set $S \subset \mathbb{R}^d$, we could examine the intersections of S with randomly oriented low-dimensional subspaces.

According to the above observation, if E is a random low-dimensional subspace, we should expect that E misses the spikes of a convex set S and the intersection $E \cap S$ looks like a ball.



Geometry of high dimensional data

The above observation is the content of Dvoretsky‘s theorem.

Theorem (Dvoretsky‘s Theorem)

Let $S \subset \mathbb{R}^d$ be an origin symmetric convex body such that the maximal volume ellipsoid is the Euclidean ball. Let $\epsilon \in (0, 1)$ and E be a uniform random subspace (with respect to the Haar measure) of dimension $k = c\,\epsilon^{-2} \log d$. Then there exists an $R > 0$ such that, with high probability, we have

$$(1 - \epsilon)\mathbb{B}_2(R) \subset S \cap E \subset (1 + \epsilon)\mathbb{B}_2(R)$$

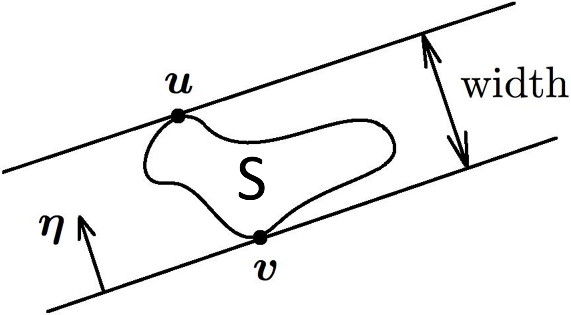
where $\mathbb{B}_2(R) \subset E$ is the Euclidean ball of radius R in E.

Note: John’s theorem guarantees that every convex body contains an ellipsoid of maximal volume. Also, any ellipsoid may be mapped to a Euclidean ball through an affine transformation. Thus, up to affine transformation, the assumptions of Dvoretsky‘s theorem are pretty mild.

Geometry of high dimensional data

To capture the complexity of a set $S \subset \mathbb{R}^d$ we will look at intersections with higher-dimensional subspaces that are more likely to intersect the spikes of S.
Note: Below, we no longer assume that S is a convex body, but any bounded set.

We defined the **width** of S in the direction of a unit vector $\eta \in \mathbb{S}^{d-1}$ as the smallest slab between two parallel hyperplanes with normals η that contains S



Geometry of high dimensional data

Analytically, we can express the width of $S \subset \mathbb{R}^d$ in the direction $\eta \in \mathbb{S}^{d-1}$ as

$$\sup_{u,\, v \in S} \langle \eta, u - v \rangle = \sup_{z \in S - S} \langle \eta, z \rangle \,.$$

where the $S - S = \{u - v : u, v \in S\}$ is the Minkowski sum of the sets S and $-S$.

This shows that width may be expressed through the **support function** of S - a fundamental object in convex analysis:

$$\sup_{z \in S - S} \langle \eta, z \rangle = \sigma_S(\eta) + \sigma_S(-\eta),$$

where $\sigma_S(\eta) = \sup_{z \in S} \langle \eta, z \rangle$.

Geometry of high dimensional data

By averaging over all directions we obtain the following notion.

Definition

The **spherical mean width** of $S \subset \mathbb{R}^d$ is obtained by averaging the width uniformly over all directions $\eta \in \mathbb{S}^{d-1}$, that is,

$$\overline{\omega}(S) := \mathbb{E}\left[\sup_{z \in S-S} \langle \eta, z \rangle\right].$$

In many applications, e.g., statistical learning theory, it is convenient to replace the spherical random vector $\eta \sim \text{Unif}(\mathbb{S}^{d-1})$ by the spherical Gaussian random vector $g \sim \mathcal{N}(0, I_d)$.

Geometry of high dimensional data

Definition (Gaussian mean width)

Given a bounded set $S \subset \mathbb{R}^d$, its **Gaussian mean width** $\omega(S)$ is defined as

$$\omega(S) = E\left[\sup_{x \in S-S} \langle g_d, x\rangle\right] = E\left[\sup_{x \in S-S} [g_d^t x]\right], \text{ where } g_d \sim \mathcal{N}(0, I_d)$$

One advantage of using $g_d \sim \mathcal{N}(0, I_d)$ rather than $\eta \sim \text{Unif}(\mathbb{S}^{d-1})$ is that g_d has independent coordinates while η does not.

The mean Gaussian width is invariant under translations, orthogonal linear transformations, and taking convex hulls.

By the last property, the Gaussian width does not distinguish between convex and nonconvex sets: $\omega(S) = \omega(conv(S))$

Geometry of high dimensional data

Rotation invariance of the Gaussian distribution shows that the random variable $\|g_d\|$ is independent from the random vector $\eta = \frac{g_d}{\|g_d\|}$, which happens to be uniformly distributed on the sphere. Thus, for $S \in \mathbb{R}^d$,

$$\omega(S) = E\left[\sup_{z \in S-S} \|g_d\| \langle \eta, x \rangle\right] = E[\|g_d\|]\overline{\omega}(S).$$

Since $E[\|g_d\|] \asymp \sqrt{d}$, then $\omega(S) \asymp \sqrt{d}\,\overline{\omega(S)}$.
Hence, in high dimensions, the standard norm distribution is close to the uniform distribution on the sphere of radius $\sqrt{d}$, that is,

$$\mathcal{N}(0, 1_d) \approx \mathrm{Unif}(\sqrt{d}\,\mathbb{S}^{d-1}).$$

For g_d fixed, we have

$$\sup_{x \in S-S} \langle g_d, x \rangle = \langle g_d, x_1 - x_2 \rangle = ||g_d|| \cdot ||x_1 - x_2||_2 \leq \sqrt{d} \cdot \mathrm{diam}(S).$$

Geometry of high dimensional data

As remarked above, variance of the Gaussian width are commonly used. We called the following one Gaussian width to distinguish it from the mean Gaussian width. Its properties are very close to the mean Gaussian width.

Definition (Gaussian width)

Given a compact set $S \subset \mathbb{R}^d$, its **Gaussian width** $w(S)$ is defined as

$$w(S) = E \max_{x \in S} \langle g_d, x \rangle = E \max_{x \in S} [g_d^t x], \quad \text{where } g_d \sim \mathcal{N}(0, I_d)$$

One can show that

$$\frac{1}{\sqrt{2\pi}} \operatorname{diam}(S) \leq w(S) \leq \frac{\sqrt{d}}{2} \operatorname{diam}(S)$$

Geometry of high dimensional data

Examples:

- ℓ^2 ball $\quad w(\mathbb{S}^{d-1}) = w(\mathbb{B}_2^d) = E[\|g_d\|_2] \asymp \sqrt{d}$
- ℓ^1 ball $\quad w(\mathbb{B}_1^d) \asymp \sqrt{\log d}$
- ℓ^∞ ball $\quad w(\mathbb{B}_\infty^d) = E\|g_d\|_1 \, d = \sqrt{2/\pi}\, d$
- finite set $\quad w(S) \leq C\sqrt{\log|S|}\,\mathrm{diam}(S)$
- Hypercube $Q = [-1,1]^d \quad w(Q) = \sqrt{\frac{2}{\pi}}\, d$
- Sparse set $K = \{x \in \mathbb{R}^d : \|x\| = 1, \|x\|_0 \leq s\}$ $w(K) \asymp \sqrt{s \log 2d/s}$

Note: since $\overline{w(S)} \asymp \frac{w(S)}{\sqrt{d}}$, then $\overline{w(\mathbb{B}_1^d)} \asymp \sqrt{\frac{\log d}{d}}$ showing that the spherical width of $\mathbb{B}_1^d$ is much smaller than its diameter.

Geometry of high dimensional data

Theorem (Gordon's Theorem, 1988)

Let $G \in \mathbb{R}^{k \times d}$ be a random matrix with independent entries in $\mathcal{N}(0,1)$ and $S \in \mathbb{S}^{d-1}$ be a closed subset. Then

$$E \max_{x \in S} \| \tfrac{1}{a_k} Gx \| \leq 1 + \frac{w(S)}{a_k}$$

$$E \min_{x \in S} \| \tfrac{1}{a_k} Gx \| \geq 1 - \frac{w(S)}{a_k}$$

where $a_k = E\|g_k\|$, with $g_k \sim \mathcal{N}(0, I_{k \times k})$ and $w(S)$ is the Gaussian width of S.
Note that we have $\sqrt{\frac{k}{k+1}}\sqrt{k} \leq a_k \leq \sqrt{k}$.

The theorem shows that the linear map $\frac{1}{a_k} G$ preserves the norm of the points in the set S up to $1 \pm \frac{w(S)}{a_k}$.

Geometry of high dimensional data

We remark that the function $f(G) = \max_{x \in S} \|Gx\|$ is 1-Lipschitz:

$$\begin{aligned} |\max_{x \in S} \|G_1 x\| - \max_{x \in S} \|G_2 x\|| &\leq \max_{x \in S} |\|G_1 x\| - \|G_2 x\|| \\ &\leq \max_{x \in S} |\|(G_1 - G_2)x\| \\ &\leq \|G_1 - G_2\| \\ &\leq \|G_1 - G_2\|_F \end{aligned}$$

Similarly, the function $\tilde{f}(G) = \min_{x \in S} \|Gx\|$ is 1-Lipschitz.
Hence, using Gaussian concentration with Gordon's theorem we get

$$P\left(\max_{x \in S} \|Gx\| \geq a_k + w(S) + t\right) \leq \exp(-\frac{t^2}{2})$$

$$P\left(\min_{x \in S} \|Gx\| \geq a_k - w(S) + t\right) \leq \exp(-\frac{t^2}{2})$$

for any $t < 0$.

Geometry of high dimensional data

Using the last observation, with $\epsilon = \frac{w(S)+t}{a_k}$ we obtain the following

Theorem

Let $G \in \mathbb{R}^{k\times d}$ be a random matrix with independent entries in $\mathcal{N}(0,1)$ and $S \in \mathbb{S}^{d-1}$ be a closed subset. Then, for $\epsilon > \sqrt{\frac{w(S)^2}{a_k^2}}$, with probability larger than $1 - 2\exp\left(-\frac{a_k^2}{2}(\epsilon - \frac{w(S)}{a_k})^2\right)$, we have

$$(1-\epsilon)\|x\| \leq \|\frac{1}{a_k} Gx\| \leq (1+\epsilon)\|x\|$$

where $a_k = E\|g_k\|$, with $g_k \sim \mathcal{N}(0, I_{k\times k})$ and $w(S)$ is the Gaussian width of S.
Recall that $\frac{k}{k+1}k \leq a_k^2 \leq k$.

Geometry of high dimensional data

Remarks:
Since $w(S) \leq C\sqrt{\log|S|}$, this theorem essentially implies the Johnson Lindenstrauss theorem; not exactly though, since $\frac{1}{a_k}Gx$ is not a projection.

In fact, under the assumptions $\epsilon > \sqrt{\frac{w(S)^2}{a_k^2}}$ and $k \geq a_k^2$, we have that $k \geq \frac{w(S)^2}{\epsilon^2}$.
For a finite set S, the Johnson Lindenstrauss theorem claims the existence of of an almost isometric map from $\mathbb{R}^d$ into $\mathbb{R}^k$ provided $k = O(\epsilon^{-2}\log|S|)$.
This is consistent with the last theorem requiring $k = O(\epsilon^{-2}w(S)^2$.
Recall in fact that $w(S)^2 = O(\log|S|))$ for a finite set S.

Geometry of high dimensional data

The last theorem suggests that if $w(S) \leq a_k$, a uniformly chosen random subspace of $\mathbb{R}^n$ of dimension $n - k$ (which can be seen as the nullspace of G) avoids a set S with high probability.

Theorem (Gordon's Escape Through a Mesh Theorem)

Let $S \in \mathbb{S}^{d-1}$ be a closed subset. If $w(S) < a_k$, then for a $(d-k)$ dimensional subspace drawn uniformly from the Grassmanian manifold we have

$$P(\Lambda \cap S \neq \varnothing) \leq \frac{7}{2} \exp\left(-\frac{1}{18}(a_k - w(S))^2\right)$$

where $a_k = E\|g_k\|$, with $g_k \sim \mathcal{N}(0, I_{k\times k})$ and $w(S)$ is the Gaussian width of S.

Geometry of high dimensional data

A remarkable application of Gordon's Theorem is that one can use it for sets such as the set of all natural images or the set of all plausible user-product ranking matrices.

In these cases Gordon's Theorem suggests that a measurements corresponding just to a random projection may be enough to keep geometric properties of the data set in question, that is, it may allow for reconstruction of the data point from just the projection.

These phenomenon and the sensing savings that arises from it is at the heart of Compressed Sensing and several recommendation system algorithms.

Geometry of high dimensional data

Let $x \in \mathbb{R}^d$ represent a signal (or image) that we wish to acquire via linear measurements

$$y_i = a_i^t x_i \quad \text{for } a_i \in \mathbb{R}^d$$

In general, one would need d linear one-dimensional measurements to find x, one for each coordinate.

The idea behind Compressed Sensing is that one may be able to significantly decrease the number of measurements needed if we know more about the structure of x, a prime example being when x is known to be sparse, i.e., to have few non-zero entries

Geometry of high dimensional data

We consider the reconstruction problem consisting of recovering $x \in \mathbb{R}^d$ from m linear measurements

$$y = Ax, \quad \text{where } A = \begin{pmatrix} a_1^t \\ a_2^t \\ \dots \\ a_m^t \end{pmatrix} \in \mathbb{R}^{m \times d}$$

where typically $d \gg m$.
We assume that $x \in \mathbb{R}^d$ is s-**sparse**, meaning that x has at most s non-zero entries.

In order for reconstruction to be stable, we will require that A is almost an isometry, meaning that the ℓ^2 distance between Ax_1 and Ax_2 should be comparable to the distances between x_1 and x_2. Since the difference between two s-sparse vectors is a $2s$-sparse vector, we can alternatively ask for A to approximately preserve the norm of $2s$ sparse vectors.

Geometry of high dimensional data

By Gordon's Theorem, we can satisfy the condition above by taking $A \in \mathbb{R}^{m\times d}$ to have i.i.d. Gaussian entries with m chosen to satisfy $m \approx w(S_{2s})^2$ where $S_{2s} = \{x \in \mathbb{S}^{d-1} : \|x\|_0 \leq 2s\}$ is the set of $2s$ sparse vectors, and $w(S_{2s})$ the Gaussian width of S_{2s}.

We have the following result

Proposition

If $s \leq d$, the Gaussian width $w(S_s)$ of $S_s = \{x \in \mathbb{S}^{d-1} : \|x\|_0 \leq s\}$ satisfies

$$w(S_s)^2 \lesssim s \log(\tfrac{d}{s})$$

This results indicates that $m \approx 2s\log(\frac{d}{2s})$ measurements suffice to stably recover a $2s$-sparse vector.
The theory of Compressed Sensing shows this number of measurement is also sufficient to guarantee that the signal in question can be recover with efficient algorithms.

Compressed sensing

Definition (Restricted Isometry Property)

A matrix $A \in \mathbb{R}^{m \times d}$ satisfies the **Restricted Isometry Property** if for any s-sparse vector $x \in \mathbb{R}^d$, there exists a δ_s, such that

$$(1 - \delta_s)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta_s)\|x\|^2$$

We recall that if A is an isometry, then it is a linear transformation that exactly preserves distance or length.
Additionally, all eigenvalues of A are ± 1.
Since an isometry also preserves orthogonality, for any two orthogonal vectors x, y

$$x^t y = 0 \Rightarrow x^t A^t A y = 0$$

Compressed sensing

Proposition

Suppose $A \in \mathbb{R}^{m \times d}$ satisfies the Restricted Isometry Property (RIP). Then

1. for any subset $S \subset [d]$ of columns of A (denote as A_S) with $|S| = s$, the singular values of A_S are all between $(1 - \delta_s)$ and $(1 + \delta_s)$;
2. for any two orthogonal vectors $x, y \in \mathbb{R}^d$, we have that

$$|x^t A^t A y| \leq 2\delta_s \|x\| \|y\|.$$

Proof.

1. Follows directly from the definition of RIP.

Compressed sensing

2. Without loss of generality, assume that $\|x\| = \|y\| = 1$.
Since x and y are orthogonal, then $\|x + y\|^2 = 2$. Hence, by RIP:

$$2(1 - \delta_s) \leq \|A(x + y)\|^2 \leq 2(1 + \delta_s)$$

and

$$(1 - \delta_s) \leq \|Ax\|^2 \leq (1 + \delta_s),\ (1 - \delta_s) \leq \|Ay\|^2 \leq (1 + \delta_s)$$

Hence

$$\begin{aligned} 2x^t A^t A y &= (x + y)^t A^t A(x + y) - x^t A^t A x - y^t A^t A y \\ &= \|A(x + y)\|^2 - \|Ax\|^2 - \|Ay\|^2 \\ &\leq 2(1 + \delta_s) - (1 - \delta_s) - (1 - \delta_s) \\ &= 4\delta_s \end{aligned}$$

so that

$$|x^t A^t A y| \leq 2\delta_s \|x\| \|y\|. \qquad \square$$

Compressed sensing

Theorem [Candès 2005]

Let $y = Ax$ where x is an s-sparse vector. Assume that A satisfies the RIP with $\delta_s < \frac{1}{3}$. Then there is a unique solution $x^* = x$ to the ℓ^1 minimization problem

$$\min_x \|x\|_1 \quad \text{subject to } y = Ax.$$

Theorem

Let $A \in \mathbb{R}^{m \times d}$ be a matrix with i.i.d. standard Gaussian entries and assume there exists a constant C such that $m \geq C\, s \ln \frac{d}{s}$. Then the matrix $\frac{1}{\sqrt{m}} A$ satisfies the RIP with high probability.

Hence, an s-sparse vector can be efficiently recovered with high probability from $O(s \ln \frac{d}{s})$ linear measurements.
Note that, in general, $O(s \ln \frac{d}{s}) \ll d$.