2. High dimensional data

Consider representing a document by a vector each component of which corresponds to the number of occurrences of a particular word in the document. The English language has on the order of 25,000 words. Thus, a document is represented by a 25,000 dimensional vector. Normalize the vectors so that they are all of unit length. If two documents are similar, the dot product of their corresponding vectors will be close to one. If the documents are not similar, then the dot product will be close to zero. Search engines represent both the content of web pages and also queries by vectors. To respond to a query, the search engine takes the dot product of the query vector with all document vectors to locate the most relevant documents.

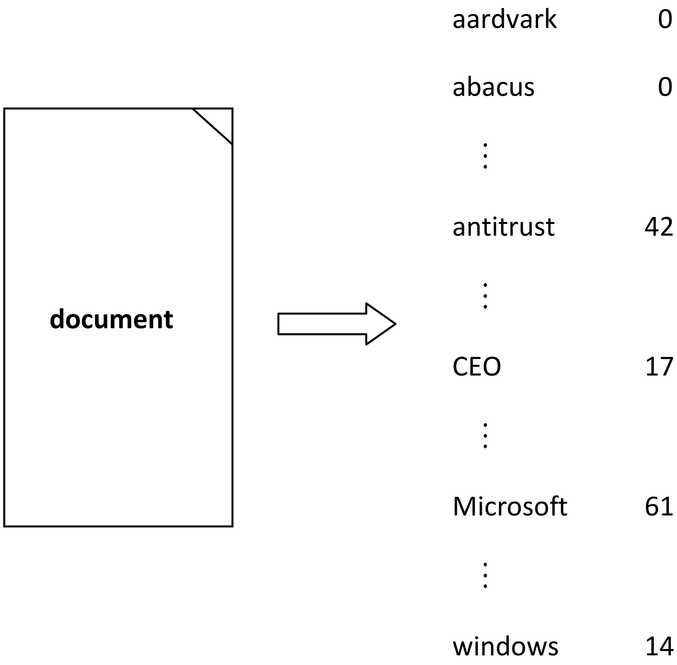


Figure 2.1: A document and its corresponding vector.

The vector space representation of documents gives rise to high dimensional data. Another example arises in determining pairs of products purchased at the same time. If there are 10,000 products for sale in a grocery store, the number of pairs is 10^8. Recording the number of times customers buy a particular pair results in a 10^8 dimensional vector.

Our intuition has been formed in low dimensions and is often misleading when considering high dimensional data. Consider placing 100 points uniformly at random in a unit square. Uniformly at random, means that each coordinate is generated independently and selected uniformly at random from the interval [0, 1]. If we select a point and measure the distance to all other points, we will see a distribution of distances. If we increase the dimension and generate the points uniformly at random in a 100-dimensional unit hypercube, the distribution of distances becomes concentrated about an average distance. The reason for this is the following. Let x and y be points in a d-dimensional space. Then

$$dist(x-y)=\sqrt{\sum_{i=1}^{d}\left(x_i-y_i\right)^2}$$

Since the distribution of each x_i and each y_i is uniform over [0,1], the distribution of each $\left(x_i-y_i\right)^2$ is of bounded variance and by Hoeffding's inequality the distribution of $x-y$ is concentrated about its expected value.

2.1 The high dimensional sphere

One of the interesting facts about a unit radius sphere in high dimensions is that as the dimension increases the volume of the sphere goes to zero. This has important implications. Also the volume of the sphere is essentially all contained in a thin slice at the equator. The volume is also essentially all contained in a narrow annulus at the surface. There is essentially no interior volume. Similarly the surface area is essentially all at the equator.

2.1.1 Volume of the unit hyper sphere and unit hyper cube

Consider the difference between the volume of a unit hypercube and the volume of a unit radius hyper sphere as the dimension, *d*, of the space increases. As the dimension of the hypercube increases, its volume is always one and the maximum possible distance between two points grows as $\sqrt{d}$. In contrast, as the dimension of a hyper sphere increases, its volume goes to zero and the maximum possible distance between two points stays at two.

Note that for *d*=2, the unit square centered at the origin lies completely inside the unit radius circle. The distance from the origin to a vertex of the square is

$$\sqrt{\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2}=\frac{\sqrt{2}}{2}\cong 0.707$$

and thus the square lies inside the circle. At *d*=4 the distance from the origin to a vertex of a unit hypercube centered at the origin is

$$\sqrt{\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2}=1$$

and thus the vertex lies on the surface of the unit 4-sphere centered at the origin. As the dimension *d* increases the distance from the origin to a vertex of the hypercube increases as $\frac{\sqrt{d}}{2}$ and for large *d* the vertices of the hypercube lie far outside the unit sphere. Figure 2.2 illustrates conceptually a hypercube and a hyper sphere. The vertices of the hyper cube are at distance $\frac{\sqrt{d}}{2}$ from the origin and thus lie outside the unit sphere. On the other hand the mid points of each face of the cube are only distance $\frac{1}{2}$ from the origin and thus are inside the sphere. Almost all the volume of the cube is located outside the sphere.

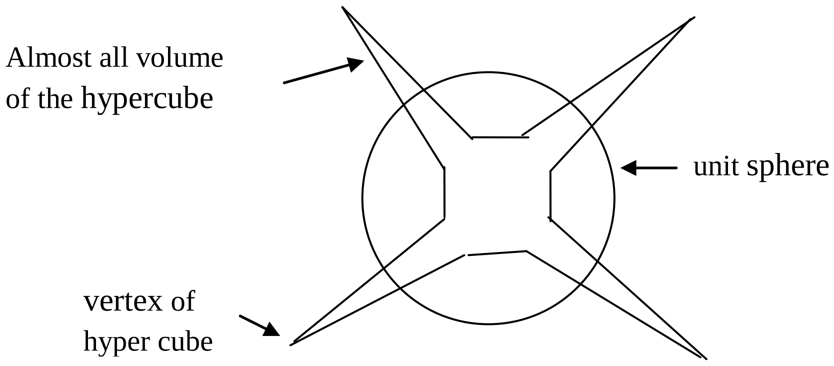


Figure 2.2: Conceptual drawing of hyper sphere and hyper cube

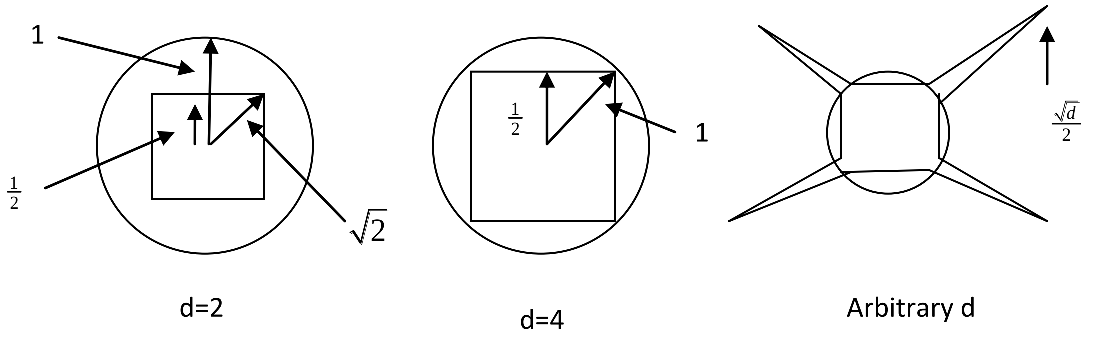


Figure 2.3: Illustration of the relationship between the hyper sphere and hyper cube in 2, 4, and d dimensions

2.1.2 Volume of a hyper sphere

For fixed dimension *d*, the volume of a hyper sphere as a function of its radius grows as r^d . For fixed radius, the volume of a hyper sphere is a function of the dimension of the space. What is interesting is that the volume of a unit hyper sphere goes to zero as the dimension of the sphere increases. To calculate the volume of a hyper sphere, one can integrate in either Cartesian or polar coordinates. In Cartesian coordinates the volume of a unit hyper sphere is given by

$$V(d) = \int_{x_1=-1}^{x_1=1} \int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}} \cdots \int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} dx_d \cdots dx_2 dx_1$$

Since the limits of the integrals are quite complex it is easier to use polar coordinates. Then

$$V(d)=\int_{S^d}\int_{r=0}^{1} r^{d-1}d\Omega dr$$

where S^d is the solid angle extended by the sphere. Since the variables r and Ω do not interact,

$$V(d)=\int_{S^d} d\Omega \int_{r=0}^{1} r^{d-1}dr = \tfrac{1}{d}\int_{S^d} d\Omega$$

The question remains, how do we determine the surface area $A(d)=\int_{S^d} d\Omega$?

Consider a different integral

$$I(d)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-\left(x_1^2+x_2^2+\cdots x_d^2\right)}dx_d\cdots dx_2 dx_1$$

Including the exponential allows us to integrate to infinity rather then stopping at the surface of a hyper sphere. This allows us to integrate easily in both Cartesian coordinates and polar coordinates. Integrating in both Cartesian and polar coordinates allows us to solve for the surface area of the unit hyper sphere.

First calculate I(d) by integration in Cartesian coordinates.

$$I(d)=\left[\int_{-\infty}^{\infty} e^{-x^2}dx\right]^d = \left(\sqrt{\pi}\right)^d = \pi^{\frac{d}{2}}$$

Next calculate I(d) by integrating in polar coordinates. Since each side of the differential element is $rd\theta$, the volume of the differential element is $\left(rd\theta\right)^{d-1} dr = r^{d-1}d\Omega dr$. Thus

$$I(d)=\int_{S^d} d\Omega\int_{0}^{\infty} e^{-r^2}r^{d-1}dr$$

The integral $A(d)=\int_{S^n} d\Omega$ is the integral over all solid angles and gives us the surface area, A(d), of a unit hyper sphere. Thus, $I(d)=A(d)\int_0^{\infty} e^{-r^2}r^{d-1}dr$. Evaluating the remaining integral gives

$$\int_0^{\infty} e^{-r^2}r^{d-1}dr = \frac{1}{2}\int_0^{\infty} e^{-t}t^{\frac{d}{2}-1}dt = \frac{1}{2}\Gamma\left(\frac{d}{2}\right)$$

where the gamma function $\Gamma(x)$ is a generalization of the factorial function for non integers values of x. $\Gamma(x)=(x-1)\Gamma(x-1)$, $\Gamma(1)=\Gamma(2)=1$, and $\Gamma\left(\tfrac{1}{2}\right)=\sqrt{\pi}$. For integer x, $\Gamma(x)=(x-1)!$.

Returning to the integral

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

Therefore, the volume of a unit hyper sphere is

$$V(d)=\frac{A(d)}{d}=\frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\Gamma\left(\frac{d}{2}\right)}$$

To check the formula for the volume of a hyper cube note that $V(2)=\pi$ and $V(3)=\frac{2}{3}\frac{\pi^{\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)}=\frac{4}{3}\pi$ which are correct volumes for the unit hyper spheres in two and three dimensions. Note that since $\pi^{\frac{d}{2}}$ is an exponential in $\frac{d}{2}$ and $\Gamma\left(\frac{d}{2}\right)$ grows as the factorial of $\frac{d}{2}$, $\lim_{d\to\infty}V(d)=0$.

2.1.3 Most of the volume is near the equator

Consider a high dimensional unit sphere and fix the North Pole on the x_1 axis at $x_1=1$. Divide the sphere in half by intersecting it with the plane $x_1=0$. The intersection of the plane with the sphere forms a region of one lower dimension, namely $\left\{x\,|\,|x|\le 1, x_1=0\right\}$ called the equator. The intersection is a sphere of dimension *d*-1 and has volume $\mathrm{V}(d-1)$. In three dimensions this region is a circle, in four dimensions the region is a three dimensional sphere, etc. In general, the intersection is a sphere of dimension *d*-1 and has volume $\mathrm{V}(d-1)$.

It turns out that essentially all of the mass of the upper hemisphere sphere lies between the plane $x_1=0$ and a parallel plane, $x_1=\varepsilon$, that is slightly higher. To see this, calculate the volume of the portion of the sphere above the slice lying between $x_1=0$ and $x_1=t_0$. Let $T=\left\{x\,|\,|x|\le 1, x_1\ge t_0\right\}$ be the portion of the sphere above the slice. To calculate the volume of T, integrate over *t* from t_0 to 1. The incremental volume will be a disk of width *dt* whose face is a sphere of dimension *d*-1 of some radius depending on *t*. The radius of the disk is $\sqrt{1-t^2}$ and therefore the surface area of the disk is

$$\left(1-t^2\right)^{\frac{d-1}{2}}\mathrm{V}(d-1).$$

Thus

$$\text{Vol}(T)=\int_{t_0}^{1}\left(1-t^2\right)^{\frac{d-1}{2}}V(d-1)\,dt=V(d-1)\int_{t_0}^{1}\left(1-t^2\right)^{\frac{d-1}{2}}dt$$

Note that $V(d)$ denotes the volume of the d dimensional unit sphere. We use Vol to denote the volume of other sets such as $\text{Vol}(T)$ for the volume of the region T.

The above integral is difficult to integrate so we use some approximations. First, we use the approximation $1+x\le e^x$ for all real x and change the upper bound on the integral to be infinity. This gives

$$\text{Vol}(T)\le V(d-1)\int_{t_0}^{\infty}e^{-\frac{d-1}{2}t^2}dt$$

Since $\frac{t}{t_0}\ge 1$ for $t\ge t_0$, an integral of the form $\int_{t_0}^{\infty}e^{-\lambda t^2}dt$ can be upper bounded by $\int_{t_0}^{\infty}\frac{t}{t_0}e^{-\lambda t^2}dt$ which has value $\frac{1}{2\lambda t_0}e^{-\lambda t_0^2}$. Thus, an upper bound on the volume of T is

$$\text{Vol}(T)\le\frac{1}{d-1}e^{-\frac{d-1}{2}t_0^2}V(d-1) \qquad (2.1)$$

Next we lower bound the volume of the entire upper hemisphere. Taking the ratio of the upper bound on the volume above the slice at t_0 to the lower bound on the volume of the entire hemisphere gives us an upper bound on the fraction of the volume above the slice. Since we believe that most of the volume is in $\left\{x\,|\,|x|\le 1, x_1\le\frac{1}{\sqrt{d-1}}\right\}$, we use the approximation

$$\int_{0}^{1}V(d-1)\left(1-t^2\right)^{\frac{d-1}{2}}dt\ge V(d-1)\int_{0}^{\frac{1}{\sqrt{d-1}}}\left(1-t^2\right)^{\frac{d-1}{2}}dt$$

Using the inequality $(1-\varepsilon)^m\ge 1-m\varepsilon$ for $\varepsilon>0$

$$\int_{0}^{1}V(d-1)\left(1-t^2\right)^{\frac{d-1}{2}}dt\ge V(d-1)\int_{0}^{\frac{1}{\sqrt{d-1}}}\left(1-\frac{d-1}{2}t^2\right)dt$$

Since $t\le\frac{1}{\sqrt{d-1}}$ in the range $\left[0,\frac{1}{\sqrt{d-1}}\right]$ we can replace the t^2 by $\frac{1}{d-1}$ in the integral and thus

$$\begin{aligned}\int_0^1 V(d-1)\left(1-t^2\right)^{\frac{d-1}{2}}dt &\ge V(d-1)\int_0^{\frac{1}{\sqrt{d-1}}}\left(1-\tfrac{d-1}{2}t^2\right)dt\\ &\ge V(d-1)\int_0^{\frac{1}{\sqrt{d-1}}}\left(1-\tfrac{d-1}{2}\tfrac{1}{d-1}\right)dt \qquad (2.2)\\ &\ge \tfrac{1}{2\sqrt{d-1}}V(d-1)\end{aligned}$$

If we compute the ratio of the upper bound on the volume of T, Eq. (2.1), to the lower bound on the volume of the hemisphere, Eq. (2.2), we see that the volume above the disk

$$\left\{x\,|\,|x|\le 1, x_1 \le t_0\right\}$$

is less than $\frac{2}{\sqrt{d-1}}e^{-\frac{d-1}{2}t_0^2}$ of the total volume of the hemisphere.

Lemma 2.1: For any $c>0$

$$\mathrm{Vol}\left(x\,|\,|x|\le 1, x_1 \ge \tfrac{c}{\sqrt{d-1}}\right)\le \tfrac{2e^{-\frac{c^2}{2}}}{c}\mathrm{V}(d)$$

Proof: Substitute $\frac{c}{\sqrt{d-1}}$ for t_0 in the above.

■

Note that we have shown that essentially all the mass of the sphere lies in a narrow slice at the equator. Note that we selected a unit vector in the x_1 direction and defined the equator to be the intersection of a plane perpendicular to the unit vector and the sphere. However, we could have selected an arbitrary point on the surface of the sphere and considered the vector from the center of the sphere to that point and then defined the equator using the plane through the center perpendicular to this arbitrary vector. Essentially all the mass of the sphere lies in a narrow slice about this equator also.

2.4 Most of the volume of a sphere is in a narrow annulus

The area of a circle is πr^2. Note that one fourth of the area of the circle is within distance one half from the center of the circle. However, in *d* dimensional space, for the sphere $B(0,1-\varepsilon)$ of radius $1-\varepsilon$ centered at the origin

$$\mathrm{Vol}\left(B(0,1-\varepsilon)\right)=(1-\varepsilon)^d\,\mathrm{V}(d)\le e^{-\varepsilon d}\,\mathrm{V}(d)\le \tfrac{1}{4}V(d).$$

provided $\varepsilon \ge 2/d$. Thus, over one fourth of the volume of the *d* dimensional sphere is within distance ε of the surface of the sphere provide $\varepsilon \ge 2/d$.

Lemma 2.2: $\mathrm{Vol}\left(B\left(0,1-\frac{c}{d}\right)\right)\leq e^{-c}\,\mathrm{Vol}\left(d\right)$ for all c.

Proof: Substitute $c=\varepsilon d$ in the above discussion.

■

2.5 Most of the surface area of a sphere is near the equator

Just as a two dimensional circle has an area and a circumference and a three dimensional sphere has a volume and a surface area, a *d* dimensional sphere has a volume and a surface area. The surface area of the hyper sphere is the set $\left\{x\,|\,|x|=1\right\}$. The circumference at the equator is the set $S=\left\{x\,|\,|x|=1,x_1=0\right\}$. The surface area of the sphere is a dimension lower than the volume and the circumference at the equator is two dimensions lower than the volume of the sphere. Just as with volume, essentially all the surface area of the sphere is near the equator. To see this, we calculate the surface area of the slice of the sphere between $x_1=0$ and $x_1=t_0$.

Let $S=\left\{x\,|\,|x|=1,x_1\geq t_0\right\}$. To calculate the surface area of S, integrate over *t* from t_0 to 1. The incremental surface unit will be a band of width *dt* whose edge is a *d*-2 dimensional sphere of some radius depending on *t*. At $x_1=t$ the radius of the edge is $\sqrt{1-t^2}$ and therefore the *d*-2 dimensional circumference of the edge is

$$V\left(d-2\right)\left(1-t^2\right)^{\frac{d-2}{2}}.$$

Thus

$$\mathrm{Vol}\left(S\right)=\mathrm{V}\left(d-2\right)\int_{t_0}^{1}\left(1-t^2\right)^{\frac{d-2}{2}}dt$$

Again the above integral is difficult to integrate and we will use the same approximations as in the earlier section on volume. This leads to the equation

$$\mathrm{Vol}\left(S\right)\leq\frac{1}{d-2}e^{-\frac{d-2}{2}t_0^2}V\left(d-2\right)\qquad(3)$$

Next we lower bound the surface area of the entire upper hemisphere.

$$\int_0^1 V(d-2)\left(1-t^2\right)^{\frac{d-2}{2}} dt \geq V(d-2)\int_0^{\frac{1}{\sqrt{d-2}}}\left(1-t^2\right)^{\frac{d-2}{2}} dt \\ \geq \tfrac{1}{2\sqrt{d-1}}V(d-2) \tag{4}$$

If we compare the upper bound on S, Eq. (3), with the lower bound on the surface area of the hemisphere, Eq. (4), we see that the surface area above the band $\left\{x \mid |x|=1, 0 \leq x_1 \leq t_0\right\}$ is less than $\frac{2}{\sqrt{d-2}}e^{-\frac{d-2}{2}t_0^2}$ of the total surface area.

Lemma 2.3: For any $c > 0$

$$\mathrm{Vol}\left(x \mid |x|=1, x_1 \geq \tfrac{c}{\sqrt{d-2}}\right) \leq \tfrac{2}{c}e^{-\frac{c^2}{2}}\,\mathrm{V}(d-2)$$

■

From the fact that the volume of the sphere is the integral of the surface area of a sphere

$$\mathrm{V}(d) = \int_0^1 \mathrm{V}(d-1)dr$$

we see that surface area is the derivative of the volume with respect to the radius r. In two dimensions the volume of a circle is πr^2 and the circumference is $2\pi r$. In three dimensions the volume is $\frac{4}{3}\pi r^3$ and the surface area is $4\pi r^2$.

2.6 Generating points uniformly at random on a sphere

We now consider how to generate points uniformly at random on the surface of a hyper sphere. First, consider generating random points on a circle of unit radius by the following method. Independently generate each coordinate uniformly at random from the interval $[-1,1]$. This produces points distributed uniformly at random over a square that is large enough to completely contain the unit circle. If we then project each point onto the unit circle, the distribution will not be uniform 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. To solve this problem, discard all points outside the unit circle and project the remaining points onto the circle.

One might generalized this technique in the obvious way to higher dimensions. However, the ratio of the volume of a *d* dimensional unit sphere to the volume of a *d* dimensional unit cube decreases rapidly making the process impractical for high dimensions since almost no points will lie inside the sphere. The solution is to generate *d* Gaussian variables. The probability distribution for a point $x_1, x_2, \cdots, x_d$ is given by

$$P\left(x_1, x_2, \cdots, x_d\right) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{x_1^2+x_2^2+\cdots+x_d^2}{2}}$$

and is spherically symmetric. Thus, normalizing the vector $\left[x_1, x_2, \cdots, x_d\right]$ to a unit vector gives a distribution that is uniform over the sphere.

2.7 Distance between random points on a unit d dimensional sphere

If we pick random points on the surface of a radius one hyper sphere, the distances would again become more concentrated as the dimension increases and would approach a distance of square root two. To see this, randomly generate points on a *d*-dimensional sphere. Rotate the coordinate system so that one of the points is at the North Pole. Since all of the surface area of a high dimensional sphere is in a narrow band about the equator the remaining points are all near the equator and the distance of each of these points to the point at the North Pole is about $\sqrt{2}$.

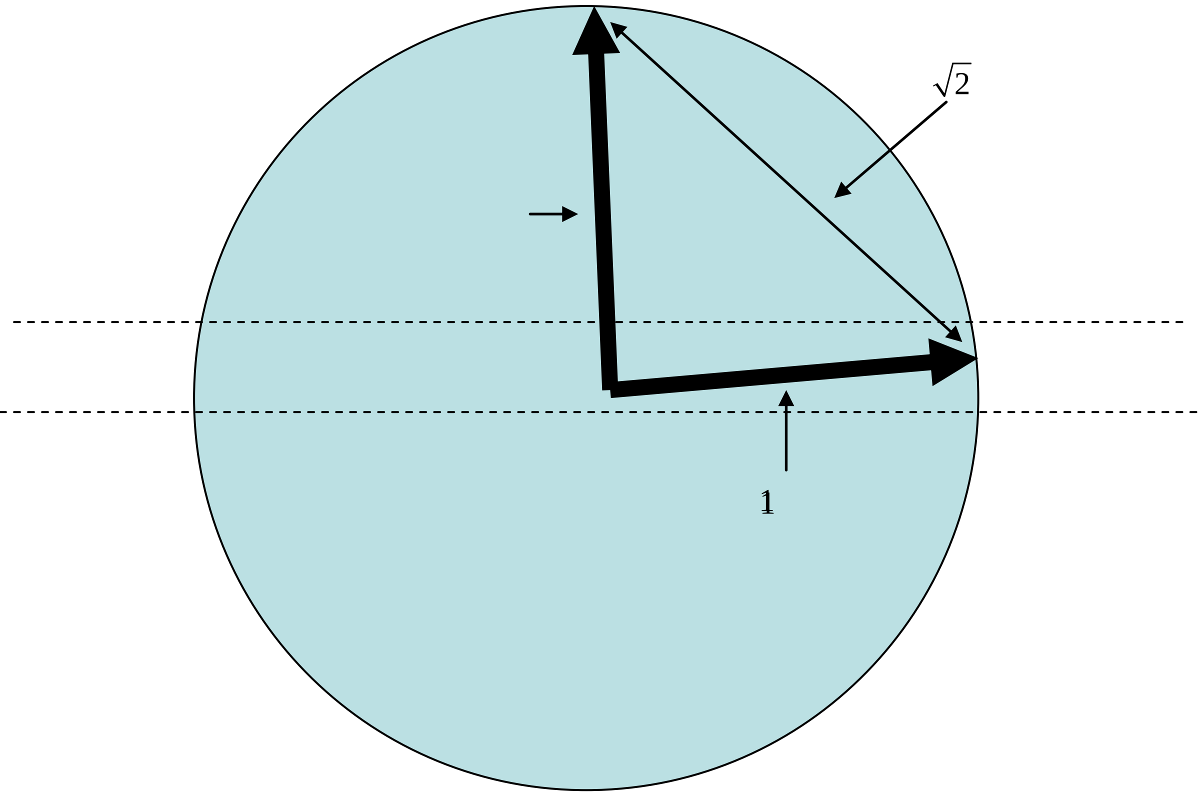


Figure 2.4: Two randomly chosen points in high dimension are almost surely orthogonal.

Distance between two points on two different spheres in high dimension

Given two unit radius spheres in high dimension with centers P and Q separated by a distance δ, what is the distance between a randomly chosen point x on the surface of the first sphere and a randomly chosen point y on the surface of the second sphere? We can write $y-x$ as $(P-x)+(Q-P)+(y-Q)$. We claim that the three segments will be pair-wise nearly orthogonal. To see this, first $Q-P$ is a fixed (not random) vector and by the fact that most of the surface area of a sphere is close to (any) equator, $P-x$ and $y-Q$ are nearly orthogonal to $Q-P$. Further, since x and y are independent, we can pick them in any order; so pick x first. Then, when y is picked, both $P-x$ and $Q-P$ are fixed vectors. Now, there is very little surface area of sphere 2 far away from the equator perpendicular to each of $P-x$ and $Q-P$ separately. But then with a factor of two, we get by the union bound that there is little surface area far from either equator; thus we get mutual orthogonality. Thus, by Pythagoras Theorem we have

$$|x-y| \approx \sqrt{|P-x|^2+|Q-P|^2+|y-Q|^2}=\sqrt{\delta^2+2}.$$

REWRITE ABOVE PARAGRAPH

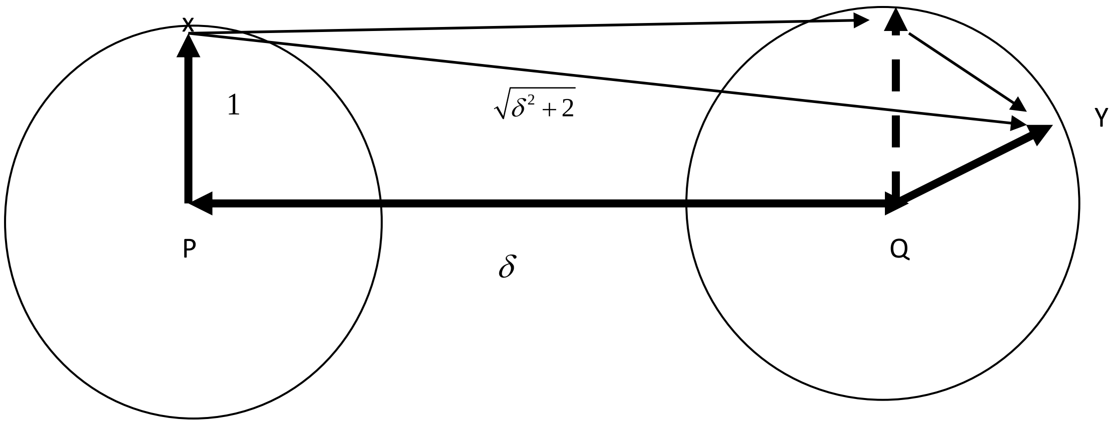


Figure 2.5: Distance between a pair of random points from two different Gaussians

2.8 Gaussians in high dimension

A one dimensional Gaussian has its mass close to the origin. However, as we increase the dimension something different happens. The *d*-dimensional spherical Gaussian with zero mean and variance σ has density function

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

Although the value of the Gaussian is maximum at the origin, there is very little volume there. Integrating the probability density over a unit sphere centered at the origin, yields zero mass since the volume of the sphere is zero. In fact, one would need to increase the radius of the sphere to

$$\Omega(\sqrt{d}\sigma)$$

before one would have a nonzero volume and hence a nonzero probability mass. If one increases the radius beyond $\sqrt{d}$, the integral ceases to increase even though the volume increases since the probability density is dropping off at a much higher rate. Thus, the natural scale for the Gaussian is in units of $\sigma\sqrt{d}$.

Expected squared distance of point from center of a Gaussian

Consider a *d*-dimensional Gaussian centered at the origin with variance σ^2. For a point $x=[x_1,x_2,\cdots,x_d]$ chosen at random from the Gaussian, what is the expected squared magnitude of *x*?

$$E\left(x_1^2+x_2^2+\cdots+x_d^2\right)=d\ E\left(x_1^2\right)=d\int_{-\infty}^{\infty}\tfrac{x_1^2}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{x_1^2}{\sigma^2}}dx_1=d\sigma^2$$

For large *d*, the value of the squared magnitude of *x* is tightly concentrated about its mean. We will call the square root of the expected squared distance (namely $\sqrt{d}\sigma$) the ``radius'' of the Gaussian.

In the rest of this section, we consider spherical Gaussians with $\sigma=1$; all results can be scaled up by σ. The probability mass of a Gaussian as a function of the distance from its center is given by $e^{-\frac{r^2}{2}}r^{d-1}$ **IS THIS CORRECT OR IS THERE A CONSTANT C?** where *r* is the distance from the center and *d* is the dimension of the space. The probability mass function has its maximum at

$$r=\sqrt{d-1}$$

which can be seen as follows

$$\tfrac{\partial}{\partial r}e^{-\frac{r^2}{2}}r^{d-1}=(d-1)e^{-\frac{r^2}{2}}r^{d-2}-r^de^{-\frac{r^2}{2}}=0\quad\Rightarrow\quad d-1=r^2.$$

Calculation of width of annulus

The function $e^{-r^2/2}r^{d-1}$ drops off fast away from its maximum. In fact, most of the mass of the Gaussian will be contained in a narrow annulus of width $O(\sigma)$. Consider the ratio of the probability mass as a function of *r* for $r=\sqrt{d-1}$. where the probability mass is maximized, and $r=\sqrt{d-1}+k$.

$$\frac{e^{-\frac{d-1+2k\sqrt{d-1}+k^2}{2}}\left(\sqrt{d-1}+k\right)^{d-1}}{e^{-\frac{d-1}{2}}\left(\sqrt{d-1}\right)^{d-1}}=e^{-k\sqrt{d-1}-\frac{k^2}{2}}\left(1+\frac{k}{\sqrt{d-1}}\right)^{d-1}$$

For large *d*, $\left(1+\frac{k}{\sqrt{d-1}}\right)^{d-1}=e^{k\sqrt{d-1}}$. Thus, the ratio of probability mass drops off as $e^{-\frac{k^2}{2}}$. So for k a large constant independent of d, the annulus between radii $\sqrt{d-1}$ and $\sqrt{d-1}+k$ contains most of the mass. So, as d gets large, we have

$$\frac{\text{width of the annulus}}{\text{radius of the spherical Gaussian}}\rightarrow\frac{1}{\sqrt{d}}.$$ **USE O NOTATION WE DO NOT KNOW CONSTANT**

Thus, similar to the situation for the hyper-sphere, most of the mass is concentrated in a thin annulus (for the sphere, the ratio was 1/d, rather than $1/\sqrt{d}$.)

Separating Gaussians

Consider two spherical unit variance Gaussians. The distance between two points generated by the same Gaussian is $\sqrt{2d}$. If two points come from different Gaussians separated by δ , then the distance between them is $\sqrt{\delta^2+2d}$. Here we have made an approximation that the points lie on a sphere of radius $\sqrt{d}$ and thus there is some approximation error in the distances. Let c bound the approximation error. Then δ needs to be large enough so that

$$\sqrt{\delta^2+2d}\geq\sqrt{2d}+c$$

Since

$$\sqrt{\delta^2+2d}=\sqrt{2d}\sqrt{1+\frac{\delta^2}{2d}}=\sqrt{2d}\left(1+\frac{1}{2}\frac{\delta^2}{2d}-\cdots\right)$$

in order to determine whether two points are from the same or different Gaussians. This requires that $\frac{1}{2}\frac{\delta^2}{\sqrt{2d}}$ to be of order 1 or $\delta>d^{\frac{1}{4}}$ in order to determine if two points were generated by the same or different Gaussians. Thus, mixtures of spherical Gaussians can be separated provided their centers are separated by more than $d^{\frac{1}{4}}$.

Algorithm: Calculate all pair wise distances between points. The cluster of smallest pair wise distances must come from a single Gaussian. Remove these points and repeat the process. In Chapter 4, we will see an algorithm to separate a mixture of k spherical Gaussians.

Fitting a single spherical Gaussian to Data

Given a set of sample points, $x_1, x_2, \cdots, x_n$, in a *d* dimensional space, we wish to find the spherical Gaussian that best fits the points. Let *F* be the unknown Gaussian. The probability of picking these very points when we sample according to *F* is given by

$$ce^{-\frac{(x_1-u)^2+(x_2-u)^2+\cdots+(x_n-u)^2}{2\sigma^2}}$$

where the normalizing constant c is $\left[\int e^{-\frac{|x-\mu|^2}{2\sigma^2}}dx\right]^n$. Note that c is really independent of μ and is equal

$$\text{to}\left[\int e^{-\frac{|x|^2}{2\sigma^2}}dx\right]^n.$$

The Maximum Likelihood Estimator (MLE) of F given the samples $x_1, x_2, \cdots, x_n$ is the F that maximizes the above probability.

Lemma 2.4 : Let $\{x_1, x_2, \cdots, x_n\}$ be a set of points in d-space. Then $(x_1-u)^2+(x_2-u)^2+\cdots+(x_n-u)^2$ is minimized when μ is the centroid of the points $x_1, x_2, \cdots, x_n$, namely $\mu = \frac{x_1+x_2+\cdots+x_n}{n}$.

Proof: Setting the derivative of $(x_1-u)^2+(x_2-u)^2+\cdots+(x_n-u)^2$ to zero yields

$$-2(x_1-u)-2(x_2-u)-\cdots-2(x_d-u)=0.$$

Solving for u gives $\mu = \frac{x_1+x_2+\cdots+x_n}{n}$.

■

Thus, in the maximum likelihood estimate for F, μ is set to the centroid. Next we will show that σ is set to the standard deviation of the sample. Substitute $v=\frac{1}{2\sigma^2}$ and $a=(x_1-u)^2+(x_2-u)^2+\cdots+(x_n-u)^2$ in the sample probability formula. This gives

$$\frac{e^{-av}}{\left[\int\limits_x e^{-\frac{x}{v}}dx\right]^m}$$

Now, a is fixed and v is to be determined. Taking logs, the expression to maximize is

$$-av-m\ln\left[\int\limits_x e^{-vx^2}dx\right]$$

To find the maximum, differentiate with respect to v, set the derivative to zero, and solve for σ. The derivative is

$$-a+m\frac{\int\limits_x x^2e^{-vx^2}dx}{\int\limits_x e^{-vx^2}dx}.$$

Setting $y=\sqrt{v}x$ in the derivative, yields

$$-a+\frac{m}{v}\frac{\int\limits_{y} y^2e^{-y^2}dy}{\int\limits_{y} e^{-y^2}dy}.$$

Since the ratio of the two integrals is the expected distance squared of a *d* dimensional spherical Gaussian of standard deviation $\frac{1}{\sqrt{2}}$ to its center and this is known to be $\frac{d}{2}$ we get

$$-a+\tfrac{md}{2v}=-a+md\sigma^2$$

It is easy to see that the maximum occurs when $\sigma=\frac{\sqrt{a}}{\sqrt{md}}$. Note that this quantity is the square root average distance squared of the samples to their mean in a coordinate direction, which is the sample standard deviation. Thus we get the following Lemma.

Lemma 2.5: The maximum likelihood spherical Gaussian for a set of samples is the one with center equal to the sample mean and standard deviation equal to the standard deviation of the sample.

ADD NOTE THAT USING AVERAGE VALUE FOR MEAN BRINGS IN DEPENDENCY

2.9 The Random Projection Theorem and the Nearest Neighbor problem

Many problems often involve high dimensional data. One such problem is the Nearest Neighbor problem in which we are given a set of *n* points in *d* dimensions. The points are processed and stored in a database. Presented with a set of query points, for each query, report the nearest point from the database. Variations of the problem where we have to report all nearby points are also of interest. The aim is often to minimize the query response time, perhaps at the cost of some extra pre-processing time.

One place the problem arises is in web search. Web pages are represented in the vector space model as points in high dimensional space. As a web-crawler discovers web pages, it processes them into some data structure. A query *q* is also a point in the high dimensional space. We wish to find the points closest to the point *q* quickly.

Here, we illustrate how finding the nearest neighbour can be made efficient by first projecting the points in the database to a randomly chosen lower dimensional space. The central result is a theorem that asserts that distances between pairs of points are preserved (up to a known scale factor) by a random projection onto a subspace provided the dimension of the subspace is not too low. Clearly one could not project a 3-dimensional object onto a 1-dimensional subspace and preserve distances between most pairs of points.

We begin by proving that projecting any fixed d-dimensional vector into a random k-dimensional subspace of $\mathbf{R}^d$, results in a vector of length very close to $\sqrt{\frac{k}{d}}$ times the length of the original vector. The projection is done by randomly choosing a basis whose first k axes span the subspace we are projecting onto. Since the subspace is random as is the basis for the subspace the squared length of each component of the vector in the new coordinate system should be equal and be 1/d times the whole. Since the projection keeps only the first k coordinates, the sum of the squared value of the projection's coordinates would be $\frac{k}{d}$ times the whole. The theorem states that random subspaces behave nicely. In fact, it asserts that the probability that the length squared of the projection deviates from $\frac{k}{d}$ falls off exponentially with the deviation.

To show that the probability falls off exponentially fast it would be convenient if the subspace was fixed and the vector was random. Thus we observe that projecting a fixed vector onto a random subspace is the same as projecting a random vector onto a fixed sub space. Let $\mathbf{v}$ be a fixed (not random) vector in $\mathbf{R}^d$ and V be a random k-dimensional subspace of $\mathbf{R}^d$. The length of the projection of $\mathbf{v}$ onto V is the same random variable as the length of a random vector $\mathbf{z}$ of the same length as v projected onto the subspace U spanned by the first k unit vectors of the coordinate system. Let $\tilde{z} = (z_1, z_2, \ldots, z_k)$. The expected value of $|\tilde{z}|^2$ is clearly $\frac{k}{d}$. We will show that the value of $|\tilde{z}|^2$ is tightly concentrated around $\frac{k}{d}$.

Theorem (The Random Projection Theorem): Let **z** be a random unit length vector in *d*-dimensions and $\tilde{z}$ be the vector of its first k components. For $0 < \varepsilon < 1$

$$\Pr\left(\left| \, |\tilde{z}|^2 - \frac{k}{d} \right| \geq \text{ò} \frac{k}{d} \right) \leq e^{-\frac{k\text{ò}^2}{16}}.$$

Proof: We need the following fact. If x is a normally distributed real random variable with zero mean and variance one, that is $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, then

$$E[e^{tx^2}] = \int_{-\infty}^{\infty} e^{tx^2} p(x) dx = 2\int_0^{\infty} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-(\frac{1}{2}-t)x^2} dx$$

Now for $t < \frac{1}{2}$,

$$E[e^{tx^2}] = \frac{2}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2\sqrt{\frac{1}{2}-t}} = \frac{1}{\sqrt{1-2t}}.$$

One way of picking a random vector $\mathbf{z}$ of length 1 is to pick independent Gaussian random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$, each with mean 0 and variance 1 and take $\mathbf{z} = \mathbf{x} / |\mathbf{x}|$. This yields the random vector **z** of length one. Thus, when $|\tilde{x}|^2 < k/d$, with $\beta = 1 - \text{ò}$ we have

$$\text{Prob}\left[\left|\tilde{z}\right|^2 < \beta\frac{k}{d}\right] = \text{Prob}\left[x_1^2 + x_2^2 + \cdots x_k^2 < \beta\frac{k}{d}\left(x_1^2 + x_2^2 + \cdots x_d^2\right)\right] .$$
$$= \text{Prob}\left[\beta k\left(x_1^2 + x_2^2 + \cdots x_d^2\right) - d\left(x_1^2 + x_2^2 + \cdots x_k^2\right) > 0\right] .$$

Thus for any $t > 0$

$$\text{Prob}\left[\left|\tilde{z}\right|^2 < \beta\frac{k}{d}\right] = \text{Prob}\left[t\left(\beta k\left(x_1^2 + x_2^2 + \cdots x_d^2\right) - d\left(x_1^2 + x_2^2 + \cdots x_k^2\right) > 0\right)\right]$$
$$= \text{Prob}\left[e^{t\left(\beta k\left(x_1^2 + x_2^2 + \cdots x_d^2\right) - d\left(x_1^2 + x_2^2 + \cdots x_k^2\right)\right)} > 1\right]$$

Applying Markov's inequality which states that $\text{Prob}\left[y > 1\right] \le E\left(y\right)$

$$\text{Prob}\left[\left|\tilde{z}\right|^2 < \beta\frac{k}{d}\right] \le E\left[e^{t\left(\beta k\left(x_1^2 + x_2^2 + \cdots x_d^2\right) - d\left(x_1^2 + x_2^2 + \cdots x_k^2\right)\right)}\right]$$
$$= E\left[e^{t(\beta k - d)\left(x_1^2 + x_2^2 + \cdots x_k^2\right)} e^{t\beta k\left(x_{k+1}^2 + x_{k+2}^2 + \cdots x_d^2\right)}\right]$$
$$= E\left[e^{t(\beta k - d)x_1^2}\right]^k E\left[e^{t\beta k x_1^2}\right]^{d-k}$$
$$= \left(\frac{1}{\sqrt{1 - 2t\beta k}}\right)^{(d-k)} \left(\frac{1}{\sqrt{1 - 2t\left(\beta k - d\right)}}\right)^k .$$

where *t* is restricted so that $2t\beta k < 1$. Now select *t* to minimize the probability. Let

$$g(t) == \left(\frac{1}{\sqrt{1 - 2t\beta k}}\right)^{(d-k)} \left(\frac{1}{\sqrt{1 - 2t\left(\beta k - d\right)}}\right)^k$$

Minimizing g is the same as maximizing

$$f(t) == \left(\sqrt{1 - 2t\beta k}\right)^{(d-k)} \left(\sqrt{1 - 2t\left(\beta k - d\right)}\right)^k$$

The maximum of *f*(*t*) occurs for $t_0 = \dfrac{1 - \beta}{2\beta(d - k\beta)}$. It is easy to check that $2t_0\beta k < 1$**.** Set $t = t_0 = \dfrac{1 - \beta}{2\beta(d - k\beta)}$**,** then

$$\text{Prob}\left[\left|\tilde{z}\right|^2 < \beta\tfrac{k}{d}\right] \le \beta^{\frac{k}{2}}\left(\frac{d - k\beta}{d - k}\right)^{\frac{d-k}{2}} \le \beta^{\frac{k}{2}}\left(\frac{d - k}{d - k} + \frac{k - k\beta}{d - k}\right)^{\frac{d-k}{2}} \le \beta^{\frac{k}{2}}\left(1 + \frac{(1 - \beta)k}{(d - k)}\right)^{\frac{d-k}{2}} \le e^{\frac{k}{2}(\ln\beta + 1 - \beta)}$$

using $1 + x \le e^x$ for all real x. Now by power series expansion, we have $\ln\beta = \ln(1 - ò) \le -ò - (1/2)ò^2$ from which the lemma follows for the case $|\mathbf{z}|^2 < k/d$.

The proof for the case when $|\mathbf{z}|^2 > k/d$ is similar and is omitted.

■

The Random Projection Theorem enables us to argue (using the union bound) that the projection to order $\log(n)$ dimensions preserves all pairwise distances between a set of *n* points consisting of the database and the query points, so that we get the answers right for all the queries. This is the content of the Johnson-Lindenstrauss lemma.

Theorem (Johnson-Lindenstrauss lemma): For any $0 < \varepsilon < 1$ and any integer *n*, let *k* be a positive integer such that

$$k \geq \frac{64 \ln n}{\text{ò}^2}.$$

Then for any set P of *n* points in R^d, there is a map $f : R^d \to R^k$ such that for all *u* and *v* in P,

$$(1-\varepsilon)|u-v|^2 \leq |f(u) - f(v)|^2 \leq (1+\varepsilon)|u-v|^2$$

Further this map can be found in randomized polynomial time.

Proof: If $d \leq k$ the theorem is trivial. Let S be a random *k*-dimensional subspace and let f(u) be the projection of *u* onto S. Let $r = \|f(u) - f(v)\|^2$. Applying the above Random Projection theorem, for any fixed *u* and *v*, the probability that *r* is outside the range $\left[(1-\varepsilon)|u-v|^2, (1+\varepsilon)|u-v|^2\right]$ is at most $\frac{2}{n^3}$. By the union bound the probability that any pair has a large distortion is less than $\binom{n}{2} \times \frac{2}{n^3} \leq \frac{1}{n}$.

■

For the nearest neighbor problem, if the database has n_1 points in it and we expect n_2 queries during the lifetime, then take $n = n_1 + n_2$ and project the database to a random k dimensional space, where, $k \geq \frac{64 \ln n}{\text{ò}^2}$. On receiving a query, project the query to the same subspace and compute nearby database points. The theorem says that with high probability, this will yield the right answer, whatever the query.

In general, nearest neighbor algorithms first find a set of candidate nearby points and then choose the nearest point from the set of candidates. Suppose the number of candidates is *m*. Without the projection, working in the whole d dimensional space would have taken time *md* to compare the query point to each candidate. But with the projection, we take only d time to project the query to the subspace and then *mk* time to compare it against the candidates. Since k<<d, this saves time. We do not go into the details of how to ensure that *m* is not too large here.

Exercise: (Overlap of spheres) Let X a be a random sample from the unit sphere in *d*-dimensions with the origin as center.

(a) What is the mean of this random variable? The mean, denoted $E(X)$, is vector, whose i^{th} component is the mean of the i^{th} component of the random sample

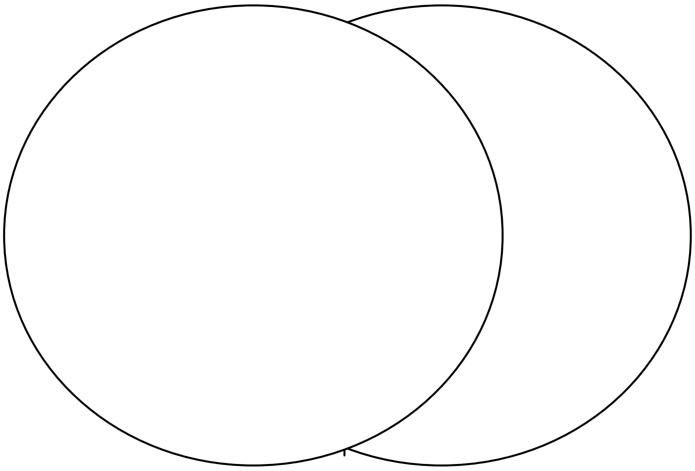
(b) What is the variance of X (again component-wise)?

(c) Show that for any unit length vector u, the variance of the real-valued random variable $u^T X$ is $\sum_{i=1}^{d} u_i^2 E(X_i^2)$. Using this, compute the variance and standard deviation of $u^T X$.

(d) Given two spheres in d space, both of radius one whose centers are distance a apart. Show that the volume of their intersection is at most

$$\frac{4e^{-\frac{a^2(d-1)}{2}}}{a\sqrt{d-1}}$$

times the volume of each one. [Hint: See picture and also use Lemma 2.1]



(e) From (d), conclude that if the inter-center separation of the two spheres is $\Omega(\text{radius}/\sqrt{d})$, then they share very small mass. Theoretically, at this separation, given randomly generated points from the two distributions, one inside each sphere, it is possible to tell which sphere contains which point, i.e., classify them into two clusters so that each is exactly the set of points generated from one sphere. The actual separation requires an efficient algorithm to achieve this. Note that the inter-center separation required in (e) goes to zero as d gets larger provided the radius of the spheres remains the same. So it is easier tell apart spheres (of the same radii) in higher dimensions.

(f) Derive the required separation for a pair of d dimensional spherical Gaussians, both with the same standard deviation.

Solution: (a) $E(X_i) = 0$ for all i, so $E(X) = \mathbf{0}$.

(b) $\text{Var}(X_i) = E(X_i^2) = \frac{1}{d}E(|X|^2)$ by symmetry. Let $V(d)$ denote the volume of the unit sphere and $A(d)$ denoting the surface area of the sphere of radius one. The infinitesimal volume of an annulus of width dr at radius r has volume $A(d)r^{d-1}dr$. So we have

$$E(|X|^2)=\frac{1}{V(d)}\int_{r=0}^{1}A(d)r^{d-1}r^2dr=\frac{A(d)}{V(d)(d+2)}=\frac{d}{d+2}.$$

Thus, $\text{Var}(X_i)=\frac{1}{d+2}$.

(c) The proof is by induction on d. It is clear for $d=1$.

$\text{Var}(\sum_i u_iX_i)=E((\sum_i u_iX_i)^2)$, since the mean is 0. Now,

$$E((\sum_i u_iX_i)^2)=E(\sum_i u_i^2X_i^2)+2E(\sum_{i\neq j}u_iu_jX_iX_j)$$

If the X_i had been independent, then the second term would be zero. But they are obviously not. So we take each section of the sphere cut by a hyperplane of the form $X_1=$constant, first integrate over this section, then integrate over all sections. In probability notation, this is taking the ``conditional expectation'' conditioned on (each value of) X_1 and then taking the expectation over all values of X_1. Doing this, we get

$$E(\sum_{i\neq j}u_iu_jX_iX_j)=E\sum_{i\geq 2}u_1X_1u_iX_i+E\sum_{i\neq j;i,j\geq 2}u_iu_jX_iX_j$$

$$=E_{X_1}\left(u_1X_1E_{X_2,X_3,\ldots X_d}\left(\sum_{i\geq 2}u_iX_i\mid X_1\right)\right)+E_{X_1}\left(E_{X_2,X_3,\ldots X_d}\left(\sum_{i\neq j;i,j\geq 2}u_iX_iu_jX_j\mid X_1\right)\right)$$

[Notation: $E(Y\mid X_1)$ is some function f of X_1; it is really short-hand for writing $f(a)=E(Y\mid X_1=a)$.]

Now, for every fixed value of X_1, $E(X_i\mid X_1)=0$ for $i\geq 2$, so the first term is zero. Since a section of the sphere is just a $d-1$sphere, the second term is zero by induction on d.

(d) Looking at the picture, by symmetry, we see that the volume of the intersection of the two spheres is just twice the volume of the section of the first sphere given by: $\{x:|x|\leq 1;x_1\geq a/2\}$ if we assume without loss of generality that the center of the second sphere is at $(a,0,0,\ldots 0)$.

(e) Simple.

(f) If a spherical Gaussian has standard deviation σ in each direction, then its radius (really the square root of the average squared distance from the mean) is $\sqrt{d}\sigma$. Its projection on any line is again a Gaussian with standard deviation σ (as we show in Chapter 4 (or 5) ??). Let a>0 and let the centers be $\mathbf{0}$ and $(a,0,0,\ldots 0)$ without loss of generality. To find the shared mass, we can use the projection onto the x_1 axis and integrate to get that the shared mass is

$$\int_{x=-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma}}\text{Min}\left(e^{-x^2/2\sigma},e^{-(x-a)^2/2\sigma}\right).$$

We bound this by using

$$\begin{aligned}
&\int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \mathrm{Min}\left(e^{-x^2/2\sigma}, e^{-(x-a)^2/2\sigma}\right) \\
&\le \int_{x=-\infty}^{a/2} \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-a)^2/2\sigma} dx + \int_{x=a/2}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma} dx \\
&= 2\int_{x=a/2}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma} dx \le 2\int_{x=a/2}^{\infty} \frac{2x}{a\sqrt{2\pi\sigma}} e^{-x^2/2\sigma} dx \\
&\le \frac{2}{a} e^{\frac{-a^2}{8\sigma}}
\end{aligned}$$

where in the last step, we are able to integrate xe^{-cx^2} in closed form.
So again, as soon as the inter-center separation goes beyond a few standard deviations, the shared mass goes down exponentially

Exercises

Exercise 2.1: Let *x* and *y* be random variables with uniform distribution in [0,1]. What is the expected value E(*x*)? E(x^2)? E(*x*-*y*)? and E((*x*-*y*)^2)?

Exercise 2.2: What is the distribution of the distance between two points chosen uniformly at random in the interval [0,1]? In the unit square? In the unit hypercube in 100 dimensions?

Exercise 2.3: Integrate using polar coordinates the area of the portion of a circle in a cone of 45°.

Exercise 2.4: For what value of *d* is the volume, *V*(*d*), of a *d*-dimensional hyper sphere maximum?

Exercise 2.5: How does the volume of a hyper sphere of radius two behave as the dimension of the space increases? What if the radius was larger than two but constant independent of *d*? What function of *d* would the radius need to be for a hyper sphere of radius *r* to have approximately constant volume as the dimension increases?

Exercise: (a) What is the volume of a hyper sphere of radius *r* in *d*-dimensions?

(b) What is the surface area of a hyper sphere of radius *r* in *d* dimensions?

(c) What is the relationship between the volume and the surface area of a hyper sphere of radius *r* in *d* dimensions?

(d) Why does the relationship determined in (c) hold?

(e) Geometrically what is the second derivative with respect to the radius of the volume of a hypersphere.

Exercise 2.6: Consider vertices of a hyper cube centered at the origin of width two. Vertices are the points $(\pm 1, \pm 1, \cdots, \pm 1)$. Place a unit radius hyper sphere at each vertex. Each sphere fits in a hyper cube of width two and thus no two spheres intersect. Prove that the volume of all of the spheres is a vanishing fraction of the hyper cube as the dimension goes to zero. That is, a point of the hyper cube picked at random will not fall into any sphere.

Exercise 2.7: How large must ε be for the annulus to contain 99% of the volume of the d dimensional sphere.

Exercise 2.8: Create a histogram of all distances between pairs of 100 points on a sphere in 3-dimensions and 100-dimensions.

Exercise 2.9:

(a) Write a computer program that generates *n* points uniformly distributed over the surface of a *d*-dimensional sphere.

(b) Create a random line through the origin and project the points onto the line. Plot the distribution of points on the line.

(c) What does your result from part b say about the surface area of the sphere in relation to the line, i.e., where is the surface area concentrated relative to the line?

Exercise 2.10: If one generates points with each coordinate a unit variance Gaussian, the points will approximately lie on the surface of a sphere of radius $\sqrt{d}$. What is the distribution when the points are projected onto a random line through the origin?

Exercise 2.11: Quantify the distance between two random points on the surfaces of two unit radius hyperspheres whose centers are separated by δ . I.e., prove that the probability that the distance is more than a away is at most some (exponentially falling) function of a .

Exercise 2.12: Project the surface area of a sphere of radius $\sqrt{d}$ in d dimensions on to a line through the center. For $d = 2,3$, derive an explicit formula for how the projected surface area changes as we move along the line. For large d , argue (intuitively) that the projected surface area should behave like a Gaussian.

Exercise 2.13: In dimension 100 what percentage of the surface area of a sphere is within distance 1/10 of the equatorial zone. Here fix the North and South Poles and ask for two planes perpendicular to the axis from the North to South Pole, what percentage of the distance to the pole must the planes be to contain 95% of the surface area?

Exercise 2.14: Project the vertices of a unit hypercube with a vertex at the origin onto a line from $(0,0,\cdots,0)$ to $(1,1,\cdots,1)$. Argue that the ``density'' of the number of projected points (per unit distance) varies roughly as a Gaussian with variance $O(1)$ with the mid-point as center.

Exercise 2.15: Place two unit spheres in *d*-dimensions, one at (-2,0,0,…,0) and the other at (2,0,0,…,0). Give an upper bound on the probability that a random line through the origin will intersect the spheres?

Exercise 2.16: Given two unit variance Gaussians in high dimensional space whose centers are one unit apart, by how much do their annuli at radius $\sqrt{d}$ of width ò> 0, small, overlap?

Exercise 2.17: How many points do you need in high dimensional space to easily detect clusters? How do you formulate this problem and develop an answer?

Exercise 2.18: Place *n* points at random on a *d*-dimensional unit sphere. Assume *d* is large. Pick a random vector and let it define two parallel hyper planes. How far apart can the hyper planes be moved and still have no points between them?

Exercise 2.19: Generate a 1000 points at vertices of a 1000 dimensional cube. Select two points *i* and *j* at random and find a path from *i* to *j* by the following algorithm. Start at *i* and go to the closest point *k* having the property that $\text{dist}(i, j)$ and $\text{dist}(j,k)$ are both less than $\text{dist}(i,k)$. Then continue the path by the same algorithm from *j* to *k*. What is the expected length of the path?

Exercise 2.20: If one has 1000 points in two dimensions that are within a unit box, one might view them as stepping stones in a pond. Select two points i and j at random and find a path from i to j by the following algorithm. Start at i and go to closest point k having the property that dist(i,j) and dist(k,j) are both less than dist(i,j). Then continue the path by the same algorithm from k to j. A computer simulation suggests that on average the path will be of length 34. If one repeats the experiment for 1000 points in 1000 dimensions on average the path will consist of only 5 hops.

Exercise 2.21: Consider a set of vectors in a high dimensional space. Assume the vectors have been normalized so that their lengths are one. Thus, the points lie on a unit sphere. Select two points at random. Assume one is at the North pole. As the dimension of the space increases the probability that the other point is close to the equator goes to one. To see this note that the ratio of the area of a cone with axis at the North pole of fixed angle say 45° to the area of a hemisphere goes to zero as the dimension increases.

Exercise 2.22: What is the expected distance between two points selected at random inside a *d*-dimensional unit cube? For two points selected at random inside a *d*-dimensional unit hyper sphere? What is cosine of the angle between them?

■

Exercise 2.23: Consider two random 0-1 vectors in high dimension. What is the angle between them? What is probability that angle is less than 45?

Exercise 2.24: Project the surface area of a *d*-dimensional unit hyper sphere onto one of its axes. What is the distribution of projected area on the axis?

■

Exercise 2.25: Where do points generated by a heavy tailed, high dimensional distribution lie? For the Gaussian distribution points lie in an annulus because the probability distribution falls off quickly as the volume increases.

■

Exercise 2.26: Given a cluster of points in *d*-dimensions how many points do we need to average to accurately determine a center?

■

Exercise 2.27: Show that the maximum of $f(t)$ is attained at $t_0 = \frac{1-\beta}{2\beta(d-k\beta)}$.

Hint: Maximize the logarithm of $f(t)$ by differentiating**.**

■

Exercise 2.28: Given the probability distribution $\frac{1}{\sqrt{2\pi}3}e^{-3\frac{(x-5)^2}{2}}$ generate ten points. From the ten points estimate u and σ.

Exercise 2.29: Calculate V(d) by a recursive procedure V(d)=cV(d-1). Develop exercise.

References