Contents

1 High-Dimensional Space

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, such a document is represented by a 25,000-dimensional vector. The representation of a document is called the *word vector model* [**?**]. A collection of n documents may be represented by a collection of 25,000-dimensional vectors, one vector per document. The vectors may be arranged as columns of a $25,000 \times n$ matrix.

Another example of high-dimensional data arises in customer-product data. If there are 1,000 products for sale and a large number of customers, recording the number of times each customer buys each product results in a collection of 1,000-dimensional vectors.

There are many other examples where each record of a data set is represented by a high-dimensional vector. Consider a collection of n web pages that are linked. A link is a pointer from one web page to another. Each web page can be represented by a 0-1 vector with n components where the j^{th} component of the vector representing the i^{th} web page has value 1, if and only if there is a link from the i^{th} web page to the j^{th} web page.

In the vector space representation of data, properties of vectors such as dot products, distance between vectors, and orthogonality often have natural interpretations. For example, the squared distance between two 0-1 vectors representing links on web pages is the number of web pages to which only one of them is linked. In Figure 1.2, pages 4 and 5 both have links to pages 1, 3, and 6 but only page 5 has a link to page 2. Thus, the squared distance between the two vectors is one.

When a new web page is created a natural question is which are the closest pages to it, that is the pages that contain a similar set of links. This question translates to the geometric question of finding nearest neighbors. The nearest neighbor query needs to be answered quickly. Later in this chapter we will see a geometric theorem, called the Random Projection Theorem, that helps with this. If each web page is a d-dimensional vector, then instead of spending time d to read the vector in its entirety, once the random projection to a k-dimensional space is done, one needs only read k entries per vector.

Dot products also play a useful role. In our first example, two documents containing many of the same words are considered similar. One way to measure co-occurrence of words in two documents is to take the dot product of the vectors representing the two documents. If the most frequent words in the two documents co-occur with similar frequencies, the dot product of the vectors will be close to the maximum, namely the product of the lengths of the vectors. If there is no co-occurrence, then the dot product will be close to zero. Here the objective of the vector representation is information retrieval.

Figure 1.1: A document and its term-document vector along with a collection of documents represented by their term-document vectors.

After preprocessing the document vectors, we are presented with queries and we want to find for each query the most relevant documents. A query is also represented by a vector which has one component per word; the component measures how important the word is to the query. As a simple example, to find documents about cars that are not race cars, a query vector will have a large positive component for the word car and also for the words engine and perhaps door and a negative component for the words race, betting, etc. Here dot products represent relevance.

An important task for search algorithms is to rank a collection of web pages in order of relevance to the collection. An intrinsic notion of relevance is that a document in a collection is relevant if it is similar to the other documents in the collection. To formalize this, one can define an ideal direction for a collection of vectors as the line of best-fit, or the line of least-squares fit, i.e., the line for which the sum of squared perpendicular distances of the vectors to it is minimized. Then, one can rank the vectors according to their dot product similarity with this unit vector. We will see in Chapter **??** that this is a well-studied notion in linear algebra and that there are efficient algorithms to find the line of best fit. Thus, this ranking can be efficiently done. While the definition of rank seems ad-hoc, it yields excellent results in practice and has become a workhorse for modern search, information retrieval, and other applications.

Notice that in these examples, there was no intrinsic geometry or vectors, just a collection of documents, web pages or customers. Geometry was added and is extremely useful. Our aim in this book is to present the reader with the mathematical foundations to deal with high-dimensional data. There are two important parts of this foundation. The first is high-dimensional geometry along with vectors, matrices, and linear algebra. The second more modern aspect is the combination with probability. When there is a stochastic model of the high-dimensional data, we turn to the study of random points. Again, there are domain-specific detailed stochastic models, but keeping with our objective of introducing the foundations, the book presents the reader with the mathematical results needed to tackle the simplest stochastic models, often assuming independence and uniform or Gaussian distributions.

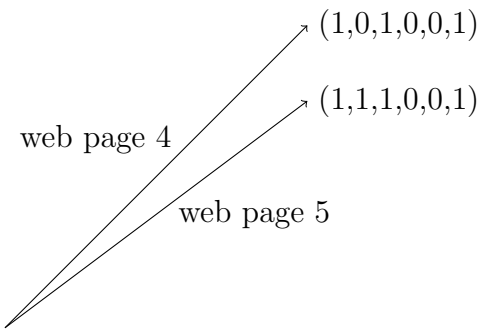

Figure 1.2: Two web pages as vectors. The squared distance between the two vectors is the number of web pages linked to by just one of the two web pages.

1.1 Properties of High-Dimensional Space

Our intuition about space was formed in two and three dimensions and is often misleading in high dimensions. Consider placing 100 points uniformly at random in a unit square. Each coordinate is generated independently and uniformly at random from the interval [0, 1]. Select a point and measure the distance to all other points and observe the distribution of distances. Then increase the dimension and generate the points uniformly at random in a 100-dimensional unit cube. The distribution of distances becomes concentrated about an average distance. The reason is easy to see. Let $\mathbf{x}$ and $\mathbf{y}$ be two such points in d-dimensions. The distance between $\mathbf{x}$ and $\mathbf{y}$ is

$$|\mathbf{x}-\mathbf{y}| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}.$$

Since $\sum_{i=1}^{d} (x_i - y_i)^2$ is the summation of a number of independent random variables of bounded variance, by the law of large numbers the distribution of $|\mathbf{x}-\mathbf{y}|^{\mathbf{2}}$ is concentrated about its expected value. Contrast this with the situation where the dimension is two or three and the distribution of distances is spread out.

For another example, consider the difference between picking a point uniformly at random from the unit-radius circle and from a unit radius sphere in d-dimensions. In d-dimensions the distance from the point to the center of the sphere is very likely to be between $1 - \frac{c}{d}$ and 1, where c is a constant independent of d. Furthermore, the first coordinate, x_1, of such a point is likely to be between $-\frac{c}{\sqrt{d}}$ and $+\frac{c}{\sqrt{d}}$, which we express by saying that most of the mass is near the equator. The equator perpendicular to the x_1 axis is the set $\{\mathbf{x}|x_1 = 0\}$. We will prove these facts in this chapter.

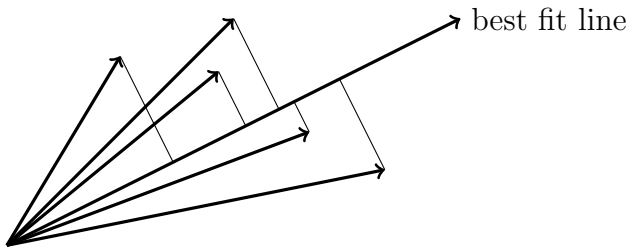

Figure 1.3: The best fit line is that line that minimizes the sum of perpendicular distances squared.

1.2 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 a high-dimensional sphere is essentially all contained in a thin slice at the equator and is simultaneously contained in a narrow annulus at the surface. There is essentially no interior volume. Similarly, the surface area is essentially all at the equator. These facts are contrary to our two or three-dimensional intuition; they will be proved by integration.

1.2.1 The Sphere and the Cube in Higher Dimensions

Consider the difference between the volume of a cube with unit-length sides and the volume of a unit-radius sphere as the dimension d of the space increases. As the dimension of the cube 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 unit-radius 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

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

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

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

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 cube increases as $\frac{\sqrt{d}}{2}$, and for large d, the vertices of the cube lie far outside the unit sphere. Figure 1.5 illustrates conceptually a cube and a sphere. The vertices of the cube are at distance $\frac{\sqrt{d}}{2}$

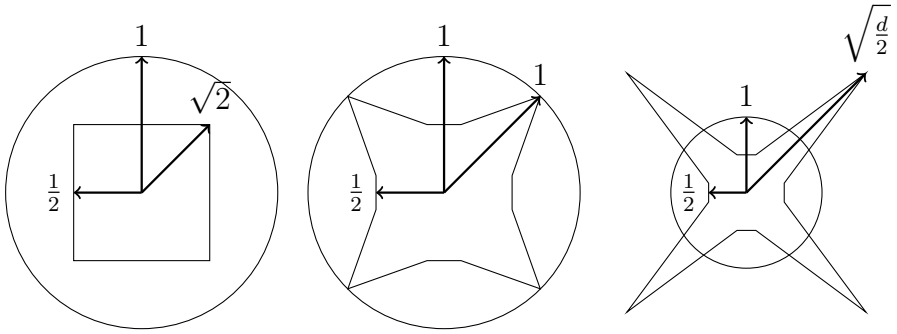

Figure 1.4: Illustration of the relationship between the sphere and the cube in 2, 4, and d-dimensions.

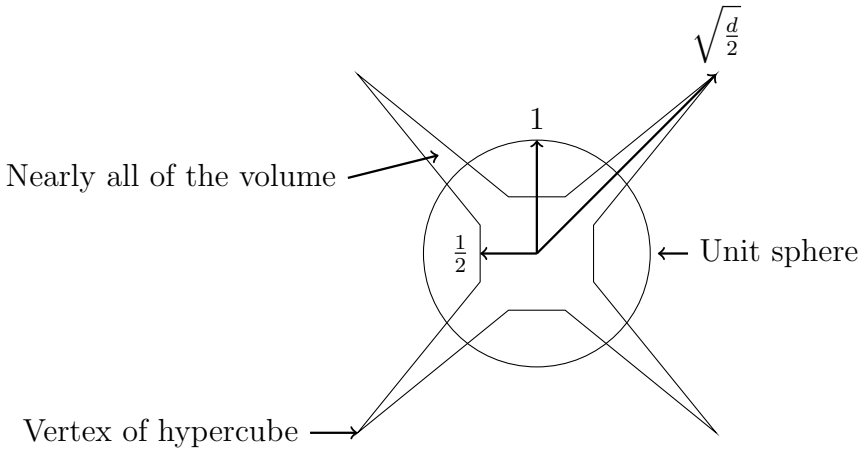

Figure 1.5: Conceptual drawing of a sphere and a cube.

from the origin and for large d lie outside the unit sphere. On the other hand, the mid point of each face of the cube is only distance $1/2$ from the origin and thus is inside the sphere. For large d, almost all the volume of the cube is located outside the sphere.

1.2.2 Volume and Surface Area of the Unit Sphere

For fixed dimension d, the volume of a sphere is a function of its radius and grows as r^d. For fixed radius, the volume of a sphere is a function of the dimension of the space. What is interesting is that the volume of a unit sphere goes to zero as the dimension of the sphere increases.

To calculate the volume of a sphere, one can integrate in either Cartesian or polar

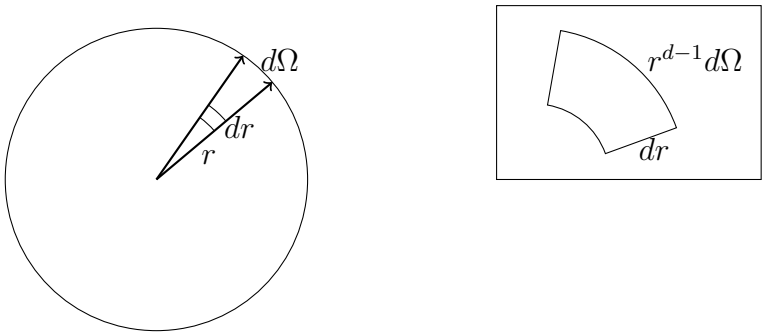

Figure 1.6: Infinitesimal volume in d-dimensional sphere of unit radius.

coordinates. In Cartesian coordinates the volume of a unit 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 complex, it is easier to integrate using polar coordinates. In polar coordinates, $V(d)$ is given by

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

Here, $d\Omega$ is the surface area of the infinitesimal piece of the solid angle S^d of the unit sphere. See Figure 1.6. The convex hull of the $d\Omega$ piece and the origin form a cone. At radius r, the surface area of the top of the cone is $r^{d-1}d\Omega$ since the surface area is $d-1$ dimensional and each dimension scales by r. The volume of the infinitesimal piece is base times height, and since the surface of the sphere is perpendicular to the radial direction at each point, the height is dr giving the above integral.

Since the variables Ω and r do not interact,

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

The question remains, how to 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 one to integrate to infinity rather than stopping at the surface of the sphere. Thus, $I(d)$ can be computed by integrating in Cartesian coordinates. Integrating in polar coordinates relates $I(d)$ to the surface area $A(d)$. Equating the two results for $I(d)$ gives $A(d)$.

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. The volume of the differential element is $r^{d-1} d\Omega dr$. Thus

$$I(d) = \int_{S^d} d\Omega \int_0^{\infty} e^{-r^2} r^{d-1} dr.$$

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

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

and hence, $I(d) = A(d)\frac{1}{2}\Gamma\left(\frac{d}{2}\right)$ where the gamma function $\Gamma(x)$ is a generalization of the factorial function for noninteger values of x. $\Gamma(x) = (x-1)\Gamma(x-1)$, $\Gamma(1) = \Gamma(2) = 1$, and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. For integer x, $\Gamma(x) = (x-1)!$.

Combining $I(d) = \pi^{\frac{d}{2}}$ with $I(d) = A(d)\frac{1}{2}\Gamma\left(\frac{d}{2}\right)$ yields

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

This establishes the following lemma.

Lemma 1.1 *The surface area $A(d)$ and the volume $V(d)$ of a unit-radius sphere in d-dimensions are given by*

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

To check the formula for the volume of a unit sphere, 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 the correct volumes for the unit spheres in two and three dimensions. To check the formula for the surface area of a unit sphere, note that $A(2) = 2\pi$ and $A(3) = \frac{2\pi^{\frac{3}{2}}}{\frac{1}{2}\sqrt{\pi}} = 4\pi$, which are the correct surface areas for the unit sphere in two and three dimensions. Note that $\pi^{\frac{d}{2}}$ is an exponential in $\frac{d}{2}$ and $\Gamma\left(\frac{d}{2}\right)$ grows as the factorial of $\frac{d}{2}$. This implies that $\lim_{d\to\infty} V(d) = 0$, as claimed.

1.2.3 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 $\{\mathbf{x}| \, |\mathbf{x}| \leq 1, x_1 = 0\}$ which we call the equator. The intersection is a sphere of dimension d-1 and has volume $V(d-1)$. 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 $d-1$.

It turns out that essentially all of the mass of the upper hemisphere lies between the plane $x_1 = 0$ and a parallel plane, $x_1 = \varepsilon$, that is slightly higher. For what value of ε does essentially all the mass lie between $x_1 = 0$ and $x_1 = \varepsilon$? The answer depends on the dimension. For dimension d it is $O(\frac{1}{\sqrt{d-1}})$. To see this, calculate the volume of the portion of the sphere above the slice lying between $x_1 = 0$ and $x_1 = \varepsilon$. Let $T = \{\mathbf{x}| \, |\mathbf{x}| \leq 1, x_1 \geq \varepsilon\}$ be the portion of the sphere above the slice. To calculate the volume of T, integrate over x_1 from ε to 1. The incremental volume is a disk of width dx_1 whose face is a sphere of dimension d-1 of radius $\sqrt{1-x_1^2}$ (see Figure 1.7) and, therefore, the surface area of the disk is

$$\left(1-x_1^2\right)^{\frac{d-1}{2}} V(d-1).$$

Thus,

$$\text{Volume}(T) = \int_{\varepsilon}^{1} \left(1-x_1^2\right)^{\frac{d-1}{2}} V(d-1)\, dx_1 = V(d-1) \int_{\varepsilon}^{1} \left(1-x_1^2\right)^{\frac{d-1}{2}} dx_1.$$

Note that $V(d)$ denotes the volume of the d-dimensional unit sphere. For the volume of other sets such as the set T, we use the notation Volume(T) for the volume. The above integral is difficult to evaluate so we use some approximations. First, we use the inequality $1 + x \leq e^x$ for all real x and change the upper bound on the integral to be infinity. Since x_1 is always greater than ε over the region of integration, we can insert x_1/ε in the integral. This gives

$$\text{Volume}(T) \leq V(d-1) \int_{\varepsilon}^{\infty} e^{-\frac{d-1}{2}x_1^2} dx_1 \leq V(d-1) \int_{\varepsilon}^{\infty} \frac{x_1}{\varepsilon} e^{-\frac{d-1}{2}x_1^2} dx_1.$$

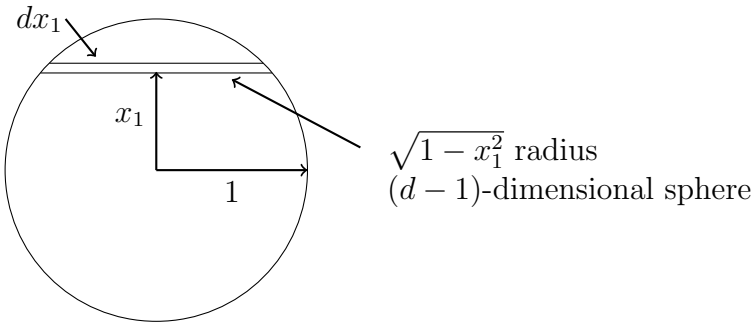

Figure 1.7: The volume of a cross-sectional slab of a d-dimensional sphere.

Now, $\int x_1 e^{-\frac{d-1}{2}x_1^2}\, dx_1 = -\frac{1}{d-1}e^{-\frac{d-1}{2}x_1^2}$ and, hence,

$$\text{Volume}\,(T) \leq \tfrac{1}{\varepsilon(d-1)}e^{-\frac{d-1}{2}\varepsilon^2}V\,(d-1)\,. \tag{1.1}$$

Next, we lower bound the volume of the entire upper hemisphere. Clearly the volume of the upper hemisphere is at least the volume between the slabs $x_1 = 0$ and $x_1 = \frac{1}{\sqrt{d-1}}$, which is at least the volume of the cylinder of radius $\sqrt{1-\frac{1}{d-1}}$ and height $\frac{1}{\sqrt{d-1}}$. The volume of the cylinder is $1/\sqrt{d-1}$ times the $d-1$-dimensional volume of the disk $R = \{\mathbf{x}|\ |\mathbf{x}| \leq 1; x_1 = \frac{1}{\sqrt{d-1}}\}$. Now R is a $d-1$-dimensional sphere of radius $\sqrt{1-\frac{1}{d-1}}$ and so its volume is

$$\text{Volume}(R) = V(d-1)\left(1-\frac{1}{d-1}\right)^{(d-1)/2}.$$

Using $(1-x)^a \geq 1-ax$

$$\text{Volume}(R) \geq V(d-1)\left(1-\frac{1}{d-1}\frac{d-1}{2}\right) = \frac{1}{2}V(d-1).$$

Thus, the volume of the upper hemisphere is at least $\frac{1}{2\sqrt{d-1}}V(d-1)$. The fraction of the volume above the plane $x_1 = \varepsilon$ is upper bounded by the ratio of the upper bound on the volume of the hemisphere above the plane $x_1 = \varepsilon$ to the lower bound on the total volume. This ratio is $\frac{2}{\varepsilon\sqrt{(d-1)}}e^{-\frac{d-1}{2}\varepsilon^2}$ which leads to the following lemma.

Lemma 1.2 *For any $c > 0$, the fraction of the volume of the hemisphere above the plane $x_1 = \frac{c}{\sqrt{d-1}}$ is less than $\frac{2}{c}e^{-c^2/2}$.*

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

For a large constant c, $\frac{2}{c}e^{-c^2/2}$ is small. The important item to remember is that most of the volume of the d-dimensional sphere of radius r lies within distance $O(r/\sqrt{d})$ of the

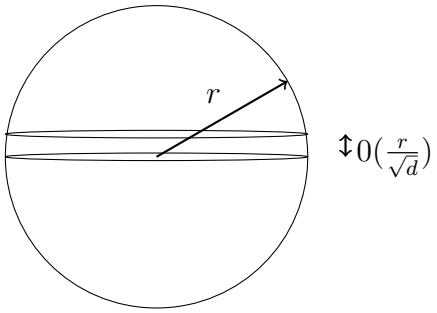

Figure 1.8: Most of the volume of the d-dimensional sphere of radius r is within distance $O(\frac{r}{\sqrt{d}})$ of the equator.

equator as shown in Figure 1.8.

For $c \geq 2$, the fraction of the volume of the hemisphere above $x_1 = \frac{c}{\sqrt{d-1}}$ is less than $e^{-2} \approx 0.14$ and for $c \geq 4$ the fraction is less than $\frac{1}{2}e^{-8} \approx 3 \times 10^{-4}$. 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 the sphere with a $(d-1)$-dimensional plane perpendicular to the unit vector. 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 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.

1.2.4 The Volume is in a Narrow Annulus

The ratio of the volume of a sphere of radius $1-\varepsilon$ to the volume of a unit sphere in d-dimensions is

$$\frac{(1-\varepsilon)^d V(d)}{V(d)} = (1-\varepsilon)^d$$

and thus goes to zero as d goes to infinity, when ε is a fixed constant. In high dimensions, all of the volume of the sphere is concentrated in a narrow annulus at the surface.

Indeed, $(1-\varepsilon)^d \leq e^{-\varepsilon d}$, so if $\varepsilon = \frac{c}{d}$, for a large constant c, all but e^{-c} of the volume of the sphere is contained in a thin annulus of width c/d. The important item to remember is that most of the volume of the d-dimensional sphere of radius $r < 1$ is contained in an annulus of width $O(1 - r/d)$ near the boundary.

1.2.5 The Surface Area is Near the Equator

Just as a 2-dimensional circle has an area and a circumference and a 3-dimensional sphere has a volume and a surface area, a d-dimensional sphere has a volume and a surface area. The surface of the sphere is the set $\{\mathbf{x}| \, |\mathbf{x}| = 1\}$. The surface of the equator is the

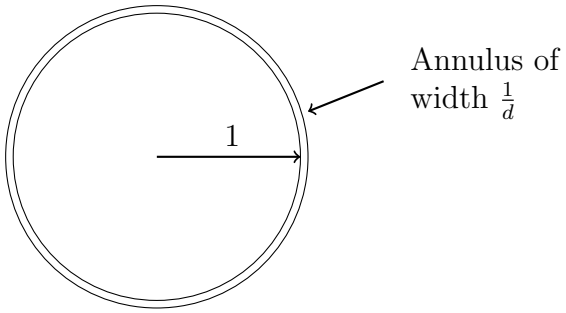

Figure 1.9: Most of the volume of the d-dimensional sphere of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

set $S = \{\mathbf{x}| \, |\mathbf{x}| = 1, x_1 = 0\}$ and it is the surface of a sphere of one lower dimension, i.e., for a 3-dimensional sphere, the circumference of a circle. Just as with volume, essentially all the surface area of a high-dimensional sphere is near the equator. To see this, calculate the surface area of the slice of the sphere between $x_1 = 0$ and $x_1 = \varepsilon$.

Let $S = \{\mathbf{x}| \, |\mathbf{x}| = 1, x_1 \geq \varepsilon\}$. To calculate the surface area of S, integrate over x_1 from ε to 1. The incremental surface unit will be a band of width dx_1 whose edge is the surface area of a $d-1$-dimensional sphere of radius depending on x_1. The radius of the band is $\sqrt{1-x_1^2}$ and therefore the surface area of the $(d-1)$-dimensional sphere is

$$A\left(d-1\right)\left(1-x_1^2\right)^{\frac{d-2}{2}}$$

where $A(d-1)$ is the surface area of a unit sphere of dimension $d-1$. Thus,

$$\text{Area}\left(S\right) = A\left(d-1\right)\int_{\varepsilon}^{1}\left(1-x_1^2\right)^{\frac{d-2}{2}}\,dx_1.$$

Again the above integral is difficult to integrate and the same approximations as in the earlier section on volume leads to the bound

$$\text{Area}\left(S\right) \leq \tfrac{1}{\varepsilon(d-2)}e^{-\frac{d-2}{2}\varepsilon^2}A\left(d-1\right). \tag{1.2}$$

Next we lower bound the surface area of the entire upper hemisphere. Clearly the surface area of the upper hemisphere is greater than the surface area of the side of a d-dimensional cylinder of height $\frac{1}{\sqrt{d-2}}$ and radius $\sqrt{1-\frac{1}{d-2}}$. The surface area of the cylinder is $\frac{1}{\sqrt{d-2}}$ times the circumference area of the d-dimensional cylinder of radius $\sqrt{1-\frac{1}{d-2}}$ which is $A(d-1)(1-\frac{1}{d-2})^{\frac{d-2}{2}}$. Using $(1-x)^a \geq 1-ax$, the surface area of the hemisphere is at

most

$$\frac{1}{\sqrt{d-2}}(1-\frac{1}{d-2})^{\frac{d-2}{2}}A(d-1) \geq \frac{1}{\sqrt{d-2}}(1-\frac{d-2}{2}\frac{1}{d-2})A(d-1)$$
$$\geq \frac{1}{2\sqrt{d-2}}A(d-1) \tag{1.3}$$

Comparing the upper bound on the surface area of S, in (1.2), with the lower bound on the surface area of the hemisphere in (1.3), we see that the surface area above the band $\{\mathbf{x}|\,|\mathbf{x}|=1, 0\leq x_1\leq\varepsilon\}$ is less than $\frac{2}{\varepsilon\sqrt{d-2}}e^{-\frac{d-2}{2}\varepsilon^2}$ of the total surface area.

Lemma 1.3 *For any $c>0$, the fraction of the surface area above the plane $x_1=\frac{c}{\sqrt{d-2}}$ is less than or equal to $\frac{2}{c}e^{-\frac{c^2}{2}}$.*

Proof: Substitute $\frac{c}{\sqrt{d-2}}$ for ε in the above. ∎

So far we have considered unit-radius spheres of dimension d. Now fix the dimension d and vary the radius r. Let $V(d,r)$ denote the volume and let $A(d,r)$ denote the surface area of a d-dimensional sphere. Then,

$$V(d,r)=\int_{x=0}^{r}A(d,x)dx.$$

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

1.3 The High-Dimensional Cube and Chernoff Bounds

We can ask the same questions about the d-dimensional unit cube $C=\{\mathbf{x}|0\leq x_i\leq 1,\ i=1,2,\ldots,d\}$ as we did for spheres. First, is the volume concentrated in an annulus? The answer to this question is simple. If we shrink the cube from its center $(\frac{1}{2},\frac{1}{2},\ldots,\frac{1}{2})$ by a factor of $1-(c/d)$ for some constant c, the volume clearly shrinks by $(1-(c/d))^d\leq e^{-c}$, so much of the volume of the cube is contained in an annulus of width $O(1/d)$. See Figure 1.10. We can also ask if the volume is concentrated about the equator as in the sphere. A natural definition of the equator is the set

$$H=\left\{\mathbf{x}\;\middle|\;\sum_{i=1}^{d}x_i=\frac{d}{2}\right\}.$$

We will show that most of the volume of C is within distance $O(1)$ of H. See Figure 1.11. The cube does not have the symmetries of the sphere, so the proof is different. The starting point is the observation that picking a point uniformly at random from C is

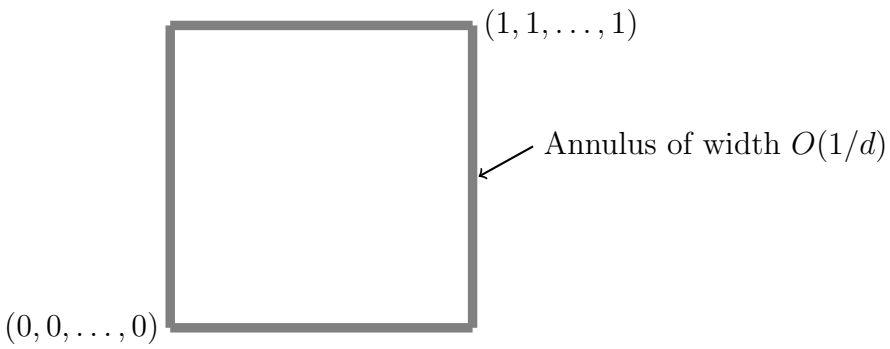

Figure 1.10: Most of the volume of the cube is in an $O(1/d)$ annulus.

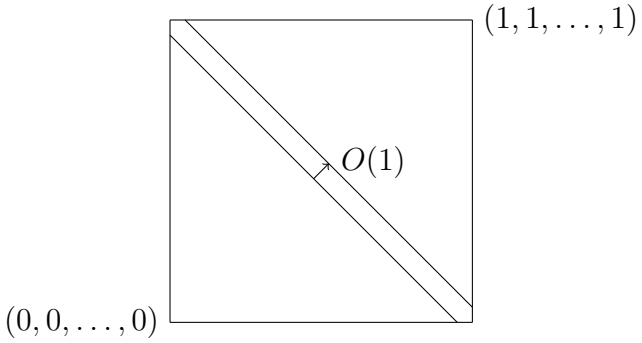

Figure 1.11: Most of the volume of the cube is within $O(1)$ of equator.

equivalent to independently picking $x_1, x_2, \ldots, x_d$, each uniformly at random from $[0, 1]$. The perpendicular distance of a point $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ to H is

$$\frac{1}{\sqrt{d}} \left| \left(\sum_{i=1}^{d} x_i \right) - \frac{d}{2} \right|.$$

Note that $\sum_{i=1}^{d} x_i = c$ defines the set of points on a hyperplane parallel to H. The perpendicular distance of a point $\mathbf{x}$ on the hyperplane $\sum_{i=1}^{d} x_i = c$ to H is $\frac{1}{\sqrt{d}} \left(c - \frac{d}{2} \right)$ or $\frac{1}{\sqrt{d}} \left| \left(\sum_{i=1}^{d} x_i \right) - \frac{d}{2} \right|$. The expected squared distance of a point from H is

$$\frac{1}{d} E \left[\left(\left(\sum_{i=1}^{d} x_i \right) - \frac{d}{2} \right)^2 \right] = \frac{1}{d} \text{Var} \left(\sum_{i=1}^{d} x_i \right).$$

By independence, the variance of $\sum_{i=1}^{d} x_i$ is the sum of the variances of the x_i, so $\text{Var}(\sum_{i=1}^{d} x_i) = \sum_{i=1}^{d} \text{Var}(x_i) = d/4$. Thus, the expected squared distance of a point

from H is 1/4. By Markov's inequality

$$\text{Prob}\left(\begin{array}{l}\text{distance from } H \text{ is greater}\\ \text{than or equal to t}\end{array}\right) = \text{Prob}\left(\begin{array}{l}\text{distance squared from } H \text{ is}\\ \text{greater than or equal to } t^2\end{array}\right) \leq \frac{1}{4t^2}.$$

Lemma 1.4 *A point picked at random in a unit cube will be within distance t of the equator defined by $H = \left\{ \mathbf{x} \left| \sum_{i=1}^{d} x_i = \frac{d}{2} \right.\right\}$ with probability at least $1 - \frac{1}{4t^2}$.*

The proof of Lemma 1.4 basically relied on the fact that the sum of the coordinates of a random point in the unit cube will, with high probability, be close to its expected value. We will frequently see such phenomena and the fact that the sum of a large number of independent random variables will be close to its expected value is called the *law of large numbers* and depends only on the fact that the random numbers have a finite variance. How close is given by a Chernoff bound and depends on the actual probability distributions involved.

The proof of Lemma 1.4 also covers the case when the x_i are Bernoulli random variables with probability 1/2 of being 0 or 1 since in this case $\text{Var}(x_i)$ also equals 1/4. In this case, the argument claims that at most a $1/(4t^2)$ fraction of the corners of the cube are at distance more than t away from H. Thus, the probability that a randomly chosen corner is at a distance t from H goes to zero as t increases, but not nearly as fast as the exponential drop for the sphere. We will prove that the expectation of the r^{th} power of the distance to H is at most some value a. This implies that the probability that the distance is greater than t is at most a/t^r, ensuring a faster drop than $1/t^2$. We will prove this for a more general case than that of uniform density for each x_i. The more general case includes independent identically distributed Bernoulli random variables.

We begin by considering the sum of d random variables, $x_1, x_2, \ldots, x_d$, and bounding the expected value of the r^{th} power of the sum of the x_i. Each variable x_i is bounded by $0 \leq x_i \leq 1$ with an expected value p_i. To simplify the argument, we create a new set of variables, $y_i = x_i - p_i$, that have zero mean. Bounding the r^{th} power of the sum of the y_i is equivalent to bounding the r^{th} power of the sum of the $x_i - p_i$. The reader my wonder why the $\mu = \sum_{i=1}^{d} p_i$ appears in the statement of Lemma 1.5 since the y_i have zero mean. The answer is because the y_i are not bounded by the range $[0, 1]$, but rather each y_i is bounded by the range $[-p_i, 1 - p_i]$.

Lemma 1.5 *Let $x_1, x_2, \ldots, x_d$ be independent random variables with $0 \leq x_i \leq 1$ and $E(x_i) = p_i$. Let $y_i = x_i - p_i$ and $\mu = \sum_{i=1}^{d} p_i$. For any positive integer r,*

$$E\left[\left(\sum_{i=1}^{d} y_i\right)^r\right] \leq \text{Max}\left[(2r\mu)^{r/2}\ ,\ r^r\right]. \tag{1.4}$$

Proof: There are d^r terms in the multinomial expansion of $E\left((y_1 + y_2 + \cdots y_d)^r\right)$; each term a product of r not necessarily distinct y_i. Focus on one term. Let r_i be the number

of times y_i occurs in that term and let I be the set of i for which r_i is nonzero. The r_i associated with the term sum to r. By independence

$$E\left(\prod_{i\in I} y_i^{r_i}\right) = \prod_{i\in I} E(y_i^{r_i}).$$

If any of the r_i equals one, then the term is zero since $E(y_i) = 0$. Thus, assume each r_i is at least two implying that $|I|$ is at most $r/2$. Now,

$$E(|y_i^{r_i}|) \leq E(y_i^2) = E(x_i^2) - p_i^2 \leq E(x_i^2) \leq E(x_i) = p_i.$$

Thus, $\prod_{i\in I} E(y_i^{r_i}) \leq \prod_{i\in I} E(|y_i^{r_i}|) \leq \prod_{i\in I} p_i$. Let $p(I)$ denote $\prod_{i\in I} p_i$. So,

$$E\left[\left(\sum_{i=1}^{d} y_i\right)^r\right] \leq \sum_{\substack{I\\ |I|\leq r/2}} p(I)n(I),$$

where $n(I)$ is number of terms in the expansion of $\left(\sum_{i=1}^{d} y_i\right)^r$ with I as the set of i with nonzero r_i. Each term corresponds to selecting one of the variables among y_i, $i \in I$ from each of the r brackets in the expansion of $(y_1 + y_2 + \cdots + y_d)^r$. Thus $n(I) \leq |I|^r$. Also,

$$\sum_{\substack{I\\ |I|=t}} p(I) \leq \left(\sum_{i=1}^{d} p_i\right)^t \frac{1}{t!} = \frac{\mu^t}{t!}.$$

To see this, do the multinomial expansion of $\left(\sum_{i=1}^{d} p_i\right)^t$. For each $I with |I| = t$, we get $\prod_{i\in I} p_i$ exactly $t!$ times. We also get other terms with repeated p_i, hence the inequality. Thus, using the Stirling approximation $t! \cong \sqrt{2\pi t}\left(\frac{t}{e}\right)^t$,

$$E\left[\left(\sum_{i=1}^{d} y_i\right)^r\right] \leq \sum_{t=1}^{r/2} \frac{\mu^t t^r}{t!} \leq \sum_{t=1}^{r/2} \frac{\mu^t}{\sqrt{2\pi t}t^t e^{-t}} t^r \leq \frac{1}{\sqrt{2\pi}} \left(\mathrm{Max}_{t=1}^{r/2} f(t)\right) \sum_{t=1}^{r/2} t^r,$$

where $f(t) = \frac{(e\mu)^t}{t^t}$. Taking logarithms and differentiating, we get

$$\begin{aligned} \ln f(t) &= t\ln(e\mu) - t\ln t \\ \frac{d}{dt}\ln f(t) &= \ln(e\mu) - 1 + \ln t \\ &= \ln(\mu) - \ln(t) \end{aligned}$$

Setting $\ln(\mu) - \ln(t)$ to zero, we see that the maximum of $f(t)$ is attained at $t = \mu$. If $\mu < r/2$, then the maximum of $f(t)$ occurs for $t = u$ and $\mathrm{Max}_{t=1}^{r/2} f(t) \leq e^{\mu} \leq e^{r/2}$. If

$\mu \geq r/2$, then $\text{Max}_{t=1}^{r/2} f(t) \leq \frac{(2e\mu)^{r/2}}{r^{r/2}}$. The geometric sum $\sum_{t=1}^{r/2} t^r$ is bounded by twice its last term or $2(r/2)^r$. Thus,

$$\begin{aligned} E\left[\left(\sum_{i=1}^{d} y_i\right)^r\right] &\leq \frac{2}{\sqrt{2\pi}} \text{Max}\left[\left(\frac{2e\mu}{r}\right)^{\frac{r}{2}}, e^{\frac{r}{2}}\right]\left(\frac{r}{2}\right)^r \\ &\leq \text{Max}\left[\left(\frac{er\mu}{2}\right)^{\frac{r}{2}}, \left(\frac{e}{4}r^2\right)^{\frac{r}{2}}\right] \\ &\leq \text{Max}\left[(2r\mu)^{r/2}\ ,\ r^r\right] \end{aligned}$$

proving the lemma. ∎

Theorem 1.6 ***(Chernoff Bounds):*** *Suppose x_i, y_i, and μ are as in the Lemma 1.5. Then*

$$\text{Prob}\left(\left|\sum_{i=1}^{d} y_i\right| \geq t\right) \leq 3e^{-t^2/12\mu}, \qquad \text{for} \quad 0 < t \leq 3\mu$$

$$\text{Prob}\left(\left|\sum_{i=1}^{d} y_i\right| \geq t\right) \leq 4 \times 2^{-t/3}, \qquad \text{for} \quad t > 3\mu.$$

Proof: Let r be a positive even integer. Let $y = \sum_{i=1}^{d} y_i$. Since r is even, y^r is nonnegative. By Markov inequality

$$\text{Prob}\left(|y| \geq t\right) = \text{Prob}\left(y^r \geq t^r\right) \leq \frac{E(y^r)}{t^r}.$$

Applying Lemma 1.5,

$$\text{Prob}\left(|y| \geq t\right) \leq \text{Max}\left[\frac{(2r\mu)^{r/2}}{t^r}, \frac{r^r}{t^r}\right]. \tag{1.5}$$

Since this holds for every even positive integer r, choose r to minimize the right hand side. By calculus, the r that minimizes $\frac{(2r\mu)^{r/2}}{t^r}$ is $r = t^2/(2e\mu)$. This is seen by taking logarithms and differentiating with respect to r. Since the r that minimizes the quantity may not be an even integer, choose r to be the largest even integer that is at most $t^2/(2e\mu)$. Then,

$$\left(\frac{2r\mu}{t^2}\right)^{r/2} \leq e^{-r/2} \leq e^{1-(t^2/4e\mu)} \leq 3e^{-t^2/12\mu}$$

for all t. When $t \leq 3\mu$, since r was choosen such that $r \leq \frac{t^2}{2eu}$,

$$\frac{r^r}{t^r} \leq \left(\frac{t}{2e\mu}\right)^r \leq \left(\frac{3\mu}{2e\mu}\right)^r \leq \left(\frac{2e}{3}\right)^{-r} \leq \left(\sqrt{e}\right)^{-r} \leq e^{-r/2},$$

which completes the proof of the first inequality.

For the second inequality, choose r to be the largest even integer less than or equal to $2t/3$. Then, Max $\left[\frac{(2r\mu)^{r/2}}{t^r}, \frac{r^r}{t^r}\right] \leq 2^{-r/2}$ and the proof is completed similar to the first case. ∎

Concentration for heavier-tailed distributions

The only place $0 \leq x_i \leq 1$ is used in the proof of (1.4) is in asserting that $E|y_i^k| \leq p_i$ for all $k = 2, 3, \ldots, r$. Imitating the proof above, one can prove stronger theorems that only assume bounds on moments up to the r^{th} moment and so include cases when x_i may be unbounded as in the Poisson or exponential density on the real line as well as power law distributions for which only moments up to some r^{th} moment are bounded. We state one such theorem. The proof is left to the reader.

Theorem 1.7 *Suppose $x_1, x_2, \ldots, x_d$ are independent random variables with $E(x_i) = p_i$, $\sum_{i=1}^{d} p_i = \mu$ and $E|(x_i - p_i)^k| \leq p_i$ for $k = 2, 3, \ldots, \lfloor t^2/6\mu \rfloor$. Then,*

$$Prob\left(\left|\sum_{i=1}^{d} x_i - \mu\right| \geq t\right) \leq \text{Max}\left(3e^{-t^2/12\mu}, 4 \times 2^{-t/e}\right).$$

1.4 Volumes of Other Solids

There are very few high-dimensional solids for which there are closed-form formulae for the volume. The volume of the rectangular solid

$$R = \{\mathbf{x} | l_1 \leq x_1 \leq u_1,\ l_2 \leq x_2 \leq u_2,\ \ldots,\ l_d \leq x_d \leq u_d\},$$

is the product of the lengths of its sides. Namely, it is $\prod_{i=1}^{d}(u_i - l_i)$.

A parallelepiped is a solid described by

$$P = \{\mathbf{x} \,|\, \mathbf{l} \leq A\mathbf{x} \leq \mathbf{u}\},$$

where A is an invertible $d \times d$ matrix, and $\mathbf{l}$ and $\mathbf{u}$ are lower and upper bound vectors, respectively. The statements $\mathbf{l} \leq A\mathbf{x}$ and $A\mathbf{x} \leq \mathbf{u}$ are to be interpreted row by row asserting $2d$ inequalities. A parallelepiped is a generalization of a parallelogram. It is easy to see that P is the image under an invertible linear transformation of a rectangular solid. Indeed, let

$$R = \{\mathbf{y} \,|\, \mathbf{l} \leq \mathbf{y} \leq \mathbf{u}\}.$$

Then the map $\mathbf{x} = A^{-1}\mathbf{y}$ maps R to P. This implies that

$$\text{Volume}(P) = |\text{Det}(A^{-1})|\, \text{Volume}(R).$$

Simplices, which are generalizations of triangles, are another class of solids for which volumes can be easily calculated. Consider the triangle in the plane with vertices $\{(0,0),(1,0),(1,1)\}$ which can be described as $\{(x,y)\,|\,0 \leq y \leq x \leq 1\}$. Its area is $1/2$ because two such right triangles can be combined to form the unit square. The generalization is the simplex in d-space with $d+1$ vertices,

$$\{(0,0,\ldots,0),(1,0,0,\ldots,0),(1,1,0,0,\ldots 0),\ldots,(1,1,\ldots,1)\},$$

which is the set

$$S = \{\mathbf{x}\,|\,1 \geq x_1 \geq x_2 \geq \cdots \geq x_d \geq 0\}.$$

How many copies of this simplex exactly fit into the unit square, $\{\mathbf{x}\,|\,0 \leq x_i \leq 1\}$? Every point in the square has some ordering of its coordinates and since there are $d!$ orderings, exactly $d!$ simplices fit into the unit square. Thus, the volume of each simplex is $1/d!$. Now consider the right angle simplex R whose vertices are the d unit vectors $(1,0,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,0,0,\ldots,0,1)$ and the origin. A vector $\mathbf{y}$ in R is mapped to an $\mathbf{x}$ in S by the mapping: $x_d = y_d$; $x_{d-1} = y_d + y_{d-1}$; $\ldots$; $x_1 = y_1 + y_2 + \cdots + y_d$. This is an invertible transformation with determinant one, so the volume of R is also $1/d!$.

A general simplex is obtained by a translation (adding the same vector to every point) followed by an invertible linear transformation on the right simplex. Convince yourself that in the plane every triangle is the image under a translation plus an invertible linear transformation of the right triangle. As in the case of parallelepipeds, applying a linear transformation A multiplies the volume by the determinant of A. Translation does not change the volume. Thus, if the vertices of a simplex T are $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_{d+1}$, then translating the simplex by $-\mathbf{v}_{d+1}$ results in vertices $\mathbf{v}_1-\mathbf{v}_{d+1},\mathbf{v}_2-\mathbf{v}_{d+1},\ldots,\mathbf{v}_d-\mathbf{v}_{d+1},\mathbf{0}$. Let A be the $d\times d$ matrix with columns $\mathbf{v}_1-\mathbf{v}_{d+1},\mathbf{v}_2-\mathbf{v}_{d+1},\ldots,\mathbf{v}_d-\mathbf{v}_{d+1}$. Then, $A^{-1}T = R$ and $AR = T$. Thus, the volume of T is $\frac{1}{d!}|\mathrm{Det}(A)|$.

1.5 Generating Points Uniformly at Random on the surface of a Sphere

Consider generating points uniformly at random on the surface of a unit-radius sphere. First, consider the 2-dimensional version of generating points on the circumference of a unit-radius circle by the following method. Independently generate each coordinate uniformly at random from the interval $[-1,1]$. This produces points distributed over a square that is large enough to completely contain the unit circle. Project each point onto the unit circle. The distribution is not 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 of the square 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 generalize 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 a point each of whose coordinates is a Gaussian variable. The probability distribution for a point $(x_1, x_2, \ldots, x_d)$ is given by

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

and is spherically symmetric. Normalizing the vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ to a unit vector gives a distribution that is uniform over the sphere. Note that once the vector is normalized, its coordinates are no longer statistically independent.

1.6 Gaussians in High Dimension

A 1-dimensional Gaussian has its mass close to the origin. However, as the dimension is increased something different happens**.** The d-dimensional spherical Gaussian with zero mean and variance σ has density function

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

The value of the Gaussian is maximum at the origin, but there is very little volume there. When $\sigma = 1$, integrating the probability density over a unit sphere centered at the origin yields nearly zero mass since the volume of such a sphere is negligible. In fact, one needs to increase the radius of the sphere to $\sqrt{d}$ before there is a significant 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. The natural scale for the Gaussian is in units of $\sigma\sqrt{d}$.

Expected squared distance of a point from the center of a Gaussian

Consider a d-dimensional Gaussian centered at the origin with variance σ^2. For a point $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ chosen at random from the Gaussian, the expected squared length of $\mathbf{x}$ is

$$E\left(x_1^2 + x_2^2 + \cdots + x_d^2\right) = d\,E\left(x_1^2\right) = d\sigma^2.$$

For large d, the value of the squared length of $\mathbf{x}$ is tightly concentrated about its mean. We call the square root of the expected squared distance (namely $\sigma\sqrt{d}$) 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 unit variance Gaussian as a function of the distance from its center is given by $r^{d-1}e^{-r^2/2}$ times some constant normalization factor 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 from setting the derivative equal to zero

$$\frac{\partial}{\partial r} e^{-\frac{r^2}{2}} r^{d-1} = (d-1)e^{-\frac{r^2}{2}} r^{d-2} - r^d e^{-\frac{r^2}{2}} = 0$$

which implies $r^2 = d-1$.

Calculation of width of the annulus

We now show that most of the mass of the Gaussian is within an annulus of constant width and radius $\sqrt{d-1}$. The probability mass of the Gaussian as a function of r is $g(r) = r^{d-1}e^{-r^2/2}$. To determine the width of the annulus in which $g(r)$ is nonnegligible, consider the logarithm of $g(r)$

$$f(r) = \ln g(r) = (d-1)\ln r - \frac{r^2}{2}.$$

Differentiating $f(r)$,

$$f'(r) = \frac{d-1}{r} - r \quad \text{and} \quad f''(r) = -\frac{d-1}{r^2} - 1 \leq -1.$$

Note that $f'(r) = 0$ at $r = \sqrt{d-1}$ and $f''(r) < 0$ for all r. The Taylor series expansion for $f(r)$ about $\sqrt{d-1}$, is

$$f(r) = f(\sqrt{d-1}) + f'(\sqrt{d-1})(r - \sqrt{d-1}) + \frac{1}{2}f''(\sqrt{d-1})(r - \sqrt{d-1})^2 + \cdots .$$

Thus,

$$f(r) = f(\sqrt{d-1}) + f'(\sqrt{d-1})(r - \sqrt{d-1}) + \frac{1}{2}f''(\zeta)(r - \sqrt{d-1})^2$$

for some point ζ between $\sqrt{d-1}$ and r.[1] Since $f'(\sqrt{d-1}) = 0$, the second term vanishes and

$$f(r) = f(\sqrt{d-1}) + \frac{1}{2}f''(\zeta)(r - \sqrt{d-1})^2.$$

Since the second derivative is always less than -1,

$$f(r) \leq f(\sqrt{d-1}) - \frac{1}{2}(r - \sqrt{d-1})^2.$$

Recall that $g(r) = e^{f(r)}$. Thus

$$g(r) \leq e^{f(\sqrt{d-1}) - \frac{1}{2}(r-\sqrt{d-1})^2} = g(\sqrt{d-1})e^{-\frac{1}{2}(r-\sqrt{d-1})^2}.$$

[1] see Whittaker and Watson 1990, pp. 95-96

Let c be a positive real and let I be the interval $[\sqrt{d-1}-c, \sqrt{d-1}+c]$. We calculate the ratio of an upper bound on the probability mass outside the interval to a lower bound on the total probability mass. The probability mass outside the interval I is upper bounded by

$$\begin{aligned}
\int_{r\notin I} g(r)\,dr &\le \int_{r=0}^{\sqrt{d-1}-c} g(\sqrt{d-1})e^{-(r-\sqrt{d-1})^2/2}\,dr + \int_{r=\sqrt{d-1}+c}^{\infty} g(\sqrt{d-1})e^{-(r-\sqrt{d-1})^2/2}\,dr \\
&\le 2g(\sqrt{d-1})\int_{r=\sqrt{d-1}+c}^{\infty} e^{-(r-\sqrt{d-1})^2/2}\,dr \\
&= 2g(\sqrt{d-1})\int_{y=c}^{\infty} e^{-y^2/2}\,dy \\
&\le 2g(\sqrt{d-1})\int_{y=c}^{\infty} \frac{y}{c}e^{-y^2/2}\,dy \\
&= \frac{2}{c}g(\sqrt{d-1})e^{-c^2/2}.
\end{aligned}$$

To get a lower bound on the probability mass in the interval $[\sqrt{d-1}-c, \sqrt{d-1}+c]$, consider the subinterval $[\sqrt{d-1}, \sqrt{d-1}+\frac{c}{2}]$. For r in the subinterval $[\sqrt{d-1}, \sqrt{d-1}+\frac{c}{2}]$, $f''(r) \ge -2$ and

$$f(r) \ge f(\sqrt{d-1}) - (r-\sqrt{d-1})^2 \ge f(\sqrt{d-1}) - \frac{c^2}{4}.$$

Thus

$$g(r) = e^{f(r)} \ge e^{f\sqrt{d-1}}e^{-\frac{c^2}{4}} = g(\sqrt{d-1})e^{\frac{c^2}{4}}.$$

Hence, $\int_{\sqrt{d-1}}^{\sqrt{d-1}+\frac{c}{2}} g(r)\ dr \ge \frac{c}{2}g(\sqrt{d-1})e^{-c^2/4}$ and the fraction of the mass outside the interval is

$$\frac{\frac{2}{c}g(\sqrt{d-1})e^{-c^2/2}}{\frac{c}{2}g(\sqrt{d-1})e^{-c^2/4} + \frac{2}{c}g(\sqrt{d-1})e^{-c^2/2}} = \frac{\frac{2}{c^2}e^{-\frac{c^2}{4}}}{\frac{c}{2}+\frac{2}{c^2}e^{-\frac{c^2}{4}}} = \frac{e^{-\frac{c^2}{4}}}{\frac{c^2}{4}+e^{-\frac{c^2}{4}}} \le \frac{4}{c^2}e^{-\frac{c^2}{4}}.$$

This establishes the following lemma.

Lemma 1.8 *For a d-dimensional spherical Gaussian of variance 1, all but $\frac{4}{c^2}e^{-c^2/4}$ fraction of its mass is within the annulus $\sqrt{d-1}-c \le r \le \sqrt{d-1}+c$ for any $c>0$.*

Separating Gaussians

Gaussians are often used to model data. A common stochastic model is the mixture model where one hypothesizes that the data is generated from a convex combination of simple probability densities. An example is two Gaussian densities $F_1(\mathbf{x})$ and $F_2(\mathbf{x})$ where data is drawn from the mixture $F(\mathbf{x}) = w_1F_1(\mathbf{x}) + w_2F_2(\mathbf{x})$ with positive weights w_1 and w_2 summing to one. Assume that F_1 and F_2 are spherical with unit variance. If their

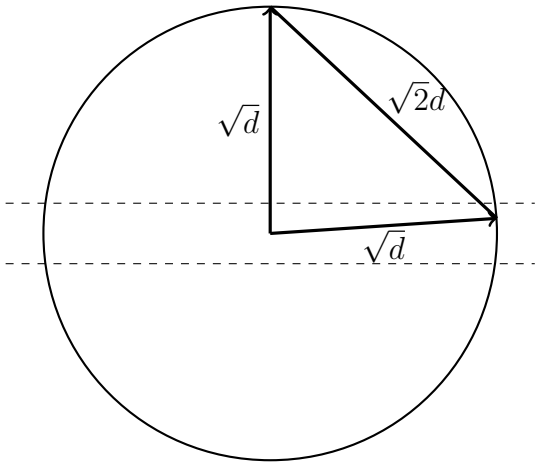

Figure 1.12: Two randomly chosen points in high dimension are almost surely nearly orthogonal.

means are very close, then given data from the mixture, one cannot tell for each data point whether it came from F_1 or F_2. The question arises as to how much separation is needed between the means to tell which Gaussian generated which data point. We will see that a separation of $\Omega(d^{1/4})$ suffices. Later, we will see that with more sophisticated algorithms, even a separation of $\Omega(1)$ suffices.

Consider two spherical unit variance Gaussians. From Lemma 1.8, most of the probability mass of each Gaussian lies on an annulus of width $O(1)$ at radius $\sqrt{d-1}$. Also $e^{-|\mathbf{x}|^2/2}$ factors into $\prod_i e^{-x_i^2/2}$ and almost all of the mass is within the slab $\{\mathbf{x}|-c \leq x_1 \leq c\}$, for $c \in O(1)$. Pick a point $\mathbf{x}$ from the first Gaussian. After picking $\mathbf{x}$, rotate the coordinate system to make the first axis point towards $\mathbf{x}$. Then, independently pick a second point $\mathbf{y}$ also from the first Gaussian. The fact that almost all of the mass of the Gaussian is within the slab $\{\mathbf{x}| - c \leq x_1 \leq c, c \in O(1)\}$ at the equator says that $\mathbf{y}$'s component along $\mathbf{x}$'s direction is $O(1)$ with high probability. Thus, $\mathbf{y}$ is nearly perpendicular to $\mathbf{x}$. So, $|\mathbf{x}-\mathbf{y}| \approx \sqrt{|\mathbf{x}|^2 + |\mathbf{y}|^2}$. See Figure 1.12. More precisely, since the coordinate system has been rotated so that $\mathbf{x}$ is at the North Pole, $\mathbf{x} = (\sqrt{(d)} \pm O(1), 0, \ldots)$. Since $\mathbf{y}$ is almost on the equator, further rotate the coordinate system so that the component of $\mathbf{y}$ that is perpendicular to the axis of the North Pole is in the second coordinate. Then $\mathbf{y} = (O(1), \sqrt{(d)} \pm O(1), \ldots)$. Thus,

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

and $|\mathbf{x}-\mathbf{y}| = \sqrt{(2d)} \pm O(1)$.

Given two spherical unit variance Gaussians with centers $\mathbf{p}$ and $\mathbf{q}$ separated by a distance δ, the distance between a randomly chosen point $\mathbf{x}$ from the first Gaussian and a randomly chosen point $\mathbf{y}$ from the second is close to $\sqrt{\delta^2 + 2d}$, since $\mathbf{x}-\mathbf{p}$, $\mathbf{p}-\mathbf{q}$, and $\mathbf{q}-\mathbf{y}$

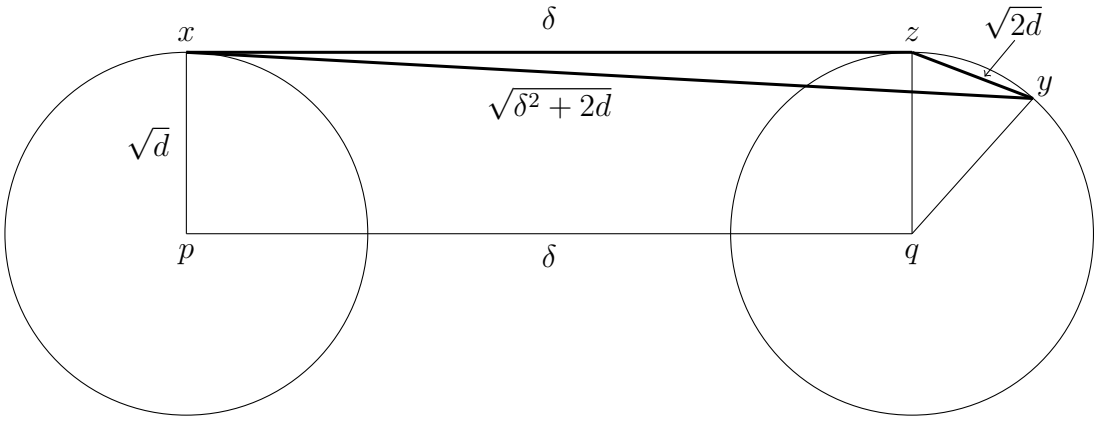

Figure 1.13: Distance between a pair of random points from two different unit spheres approximating the annuli of two Gaussians.

are nearly mutually perpendicular. To see this, pick $\mathbf{x}$ and rotate the coordinate system so that $\mathbf{x}$ is at the North Pole. Let $\mathbf{z}$ be the North Pole of the sphere approximating the second Gaussian. Now pick $\mathbf{y}$. Most of the mass of the second Gaussian is within $O(1)$ of the equator perpendicular to $\mathbf{q} - \mathbf{z}$. Also, most of the mass of each Gaussian is within distance $O(1)$ of the respective equators perpendicular to the line $\mathbf{q} - \mathbf{p}$. Thus,

$$\begin{aligned} |\mathbf{x} - \mathbf{y}|^2 &\approx \delta^2 + |\mathbf{z} - \mathbf{q}|^2 + |\mathbf{q} - \mathbf{y}|^2 \\ &= \delta^2 + 2d \pm O(\sqrt{d})). \end{aligned}$$

To ensure that the distance between two points picked from the same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of distance between points from different Gaussians. This requires that $\sqrt{2d} + O(1) \leq \sqrt{2d + \delta^2} - O(1)$ or $2d + O(\sqrt{d}) \leq 2d + \delta^2$, which holds when $\delta \in \Omega(d^{1/4})$. Thus, mixtures of spherical Gaussians can be separated provided their centers are separated by more than $d^{\frac{1}{4}}$. One can actually separate Gaussians where the centers are much closer. In Chapter 4, we will see an algorithm that separates a mixture of k spherical Gaussians whose centers are much closer.

Algorithm for separating points from two Gaussians

> Calculate all pairwise distances between points. The cluster of smallest pairwise distances must come from a single Gaussian. Remove these points and repeat the process.

Fitting a single spherical Gaussian to data

Given a set of sample points, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, in a d-dimensional space, we wish to find the spherical Gaussian that best fits the points. Let F be the unknown Gaussian with mean $\boldsymbol{\mu}$ and variance σ^2 in every direction. The probability of picking these very points when sampling according to F is given by

$$c \exp\left(-\frac{(\mathbf{x_1}-\boldsymbol{\mu})^2+(\mathbf{x_2}-\boldsymbol{\mu})^2+\cdots+(\mathbf{x_n}-\boldsymbol{\mu})^2}{2\sigma^2}\right)$$

where the normalizing constant c is the reciprocal of $\left[\int e^{-\frac{|\mathbf{x}-\boldsymbol{\mu}|^2}{2\sigma^2}}dx\right]^n$. In integrating from $-\infty$ to ∞, one could shift the origin to $\boldsymbol{\mu}$ and thus c is $\left[\int e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}}dx\right]^{-n}$ and is independent of $\boldsymbol{\mu}$.

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

Lemma 1.9 *Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n points in d-space. Then $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ is minimized when $\boldsymbol{\mu}$ is the centroid of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, namely $\boldsymbol{\mu}=\frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$.*

Proof: Setting the derivative of $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ to zero yields

$$-2(\mathbf{x}_1-\boldsymbol{\mu})-2(\mathbf{x}_2-\boldsymbol{\mu})-\cdots-2(\mathbf{x}_d-\boldsymbol{\mu})=0.$$

Solving for $\boldsymbol{\mu}$ gives $\boldsymbol{\mu}=\frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$. ∎

In the maximum likelihood estimate for F, $\boldsymbol{\mu}$ is set to the centroid. Next we show that σ is set to the standard deviation of the sample. Substitute $\nu=\frac{1}{2\sigma^2}$ and $a=(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ into the formula for the probability of picking the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. This gives

$$\frac{e^{-a\nu}}{\left[\int\limits_x e^{-|\mathbf{x}|^2\nu}dx\right]^n}.$$

Now, $\mathbf{a}$ is fixed and ν is to be determined. Taking logs, the expression to maximize is

$$-a\nu-n\ln\left[\int\limits_{\mathbf{x}} e^{-\nu|\mathbf{x}|^2}dx\right].$$

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

$$-a+n\frac{\int\limits_{\mathbf{x}}|\mathbf{x}|^2e^{-\nu|\mathbf{x}|^2}dx}{\int\limits_{\mathbf{x}}e^{-\nu|\mathbf{x}|^2}dx}.$$

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

$$-a + \frac{n}{\nu}\frac{\int_{\mathbf{y}} \mathbf{y}^2 e^{-\mathbf{y}^2} dy}{\int_{\mathbf{y}} e^{-\mathbf{y}^2} dy}.$$

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 $-a + \frac{nd}{2\nu}$. Substituting σ^2 for $\frac{1}{2\nu}$ gives $-a + nd\sigma^2$. Setting $-a + nd\sigma^2 = 0$ shows that the maximum occurs when $\sigma = \frac{\sqrt{a}}{\sqrt{nd}}$. Note that this quantity is the square root of the average coordinate distance squared of the samples to their mean, which is the standard deviation of the sample. Thus, we get the following lemma.

Lemma 1.10 *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.*

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample of points generated by a Gaussian probability distribution. $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$ is an unbiased estimator of the expected value of the distribution. However, if in estimating the variance from the sample set, we use the estimate of the expected value rather than the true expected value, we will not get an unbiased estimate of the variance since the sample mean is not independent of the sample set. One should use $\boldsymbol{\mu} = \frac{1}{n-1}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$ when estimating the variance. See appendix.

1.7 Random Projection and the Johnson-Lindenstrauss Theorem

Many high-dimensional problems such as the nearest neighbor problem can be sped up by projecting the data to a random lower-dimensional subspace and solving the problem there. This technique requires that the projected distances have the same ordering as the original distances. If one chooses a random k-dimensional subspace, then indeed all the projected distances in the k-dimensional space are approximately within a known scale factor of the distances in the d-dimensional space. We first show that for one distance pair, the probability of its projection being badly represented is exponentially small in k. Then we use the union bound to argue that failure does not happen for any pair.

Project a fixed (not random) unit length vector $\mathbf{v}$ in d-dimensional space onto a random k-dimensional space. By the Pythagoras theorem, the length squared of a vector is the sum of the squares of its components. Thus, we would expect the squared length of the projection to be $\frac{k}{d}$. The following theorem asserts that the squared length of the projection is very close to this quantity.

Theorem 1.11 (The Random Projection Theorem): *Let* $\mathbf{v}$ *be a fixed unit length vector in a* d*-dimensional space and let* W *be a random* k*-dimensional subspace. Let* $\mathbf{w}$ *be the projection of* $\mathbf{v}$ *onto* W*. For any* $0 \leq \varepsilon \leq 1$*, Prob* $\left(\left||\mathbf{w}|^2 - \frac{k}{d}\right| \geq \varepsilon\frac{k}{d}\right) \leq 4e^{-\frac{k\varepsilon^2}{64}}$.

Proof: A random subspace is generated by selecting a set of basis vectors. Working with such a set of dependent vectors is difficult. However, projecting a fixed vector onto a random subspace is the same as projecting a random vector onto a fixed subspace. That is, the probability distribution of $\mathbf{w}$ in the theorem is the same as the probability distribution of the vector obtained by taking a random unit length vector $\mathbf{z}$ and projecting it onto the fixed subspace U spanned by its first k coordinate vectors. This is because one can rotate the coordinate system so that a set of basis vectors for W are the first k coordinate axes.

Let $\tilde{\mathbf{z}}$ be the projection of $\mathbf{z}$ onto U. We will prove that $|\tilde{\mathbf{z}}|^2 \approx \frac{k}{d}$. Now $\left||\tilde{\mathbf{z}}|^2 - \frac{k}{d}\right|$ is greater than or equal to $\varepsilon\frac{k}{d}$ if either $|\tilde{\mathbf{z}}|^2 \leq (1-\varepsilon)\frac{k}{d}$ or $|\tilde{\mathbf{z}}|^2 \geq (1-\varepsilon)\frac{k}{d}$. Let $\beta = 1-\varepsilon$. We will prove that Prob $\left(|\tilde{\mathbf{z}}|^2 < \beta\frac{k}{d}\right) \leq 2e^{-\frac{k\varepsilon^2}{4}}$. The other case is similar and is omitted. Together they imply Prob $\left(\left||\mathbf{w}|^2 - \frac{k}{d}\right| \geq \varepsilon\frac{k}{d}\right) \leq 4e^{-\frac{k\varepsilon^2}{64}}$.

Pick a random vector $\mathbf{z}$ of length one by picking independent Gaussian random variables $x_1, x_2, \ldots, x_d$, each with mean zero and variance one. Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and take $\mathbf{z} = \mathbf{x}/|\mathbf{x}|$. This yields a random vector $\mathbf{z}$ of length one.

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

If $k\varepsilon^2 < 64$, then $4e^{-\frac{k\varepsilon^2}{64}} > 1$, so the probability upper bound asserted in the theorem is greater than one and there is nothing to prove. So assume that $k\varepsilon^2 \geq 64$ which implies $\varepsilon \geq \frac{8}{\sqrt{k}}$. Define $c = \varepsilon\sqrt{k}/4$ and since $\varepsilon \geq \frac{8}{\sqrt{k}}$, it follows that $c \geq 2$.

Now if $\beta k\left(x_1^2 + x_2^2 + \cdots + x_d^2\right) - d\left(x_1^2 + x_2^2 + \cdots + x_k^2\right) > 0$, then one of the following inequalities must hold with $c = \varepsilon\sqrt{k}/4$:

$$\beta k(x_1^2 + x_2^2 + \cdots + x_d^2) > \beta k(\sqrt{d-1} + c)^2 \tag{1.6}$$

$$d(x_1^2 + x_2^2 + \cdots + x_k^2) < \beta k(\sqrt{d-1} + c)^2. \tag{1.7}$$

Using Lemma 1.8 we will prove that the probability of each of (1.6) and (1.7) is at most $\frac{4}{c^2}e^{-\frac{c^2}{4}} = \frac{64}{k\varepsilon^2}e^{-\frac{k\varepsilon^2}{64}} \leq e^{-k\varepsilon^2/64}$ so that the probability of at least one of (1.6) or (1.7) is less than or equal to $2e^{-\frac{k\varepsilon^2}{64}}$ which proves the theorem. Lemma 1.8 implies

$$(x_1^2 + x_2^2 + \cdots x_d^2)^2 \geq (\sqrt{d-1} + c)^2$$

with probability less than or equal to $\frac{4}{c^2}e^{-c^2/4} \leq e^{-k\varepsilon^2/64}$ from which (1.6) follows. For (1.7), from Lemma 1.8,

$$\text{Prob}(d(x_1^2 + x_2^2 + \cdots + x_k^2) < d(\sqrt{k-1} - c)^2)$$

with probability less than or equal to $\frac{4}{c^2}e^{-c^2/4} \leq e^{-k\varepsilon^2/64}$. Since $\beta k < d$, $\beta k(\sqrt{d-1} - c)^2 \leq d(\sqrt{k-1} - c)^2$ and thus

$$\begin{aligned}\text{Prob}\left(d(x_1^2 + x_2^2 + \cdots + x_k^2) < \beta k(\sqrt{d-1} - c)^2\right) \\ \leq \text{Prob}\left(d(x_1^2 + x_2^2 + \cdots + x_k^2) < d(\sqrt{k-1} - c)^2\right) \\ \leq e^{-k\varepsilon^2/64}\end{aligned}$$

completing the proof of Theorem 1.11. ∎

The Random Projection Theorem enables us to argue, using the union bound, that the projection to order $\log n$ dimensions preserves all relative pairwise distances between a set of n points. This is the content of the Johnson-Lindenstrauss Lemma.

Theorem 1.12 (Johnson-Lindenstrauss Lemma): *For any $0 < \varepsilon < 1$ and any integer n, let k be an integer such that*

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

For any set P of n points in R^d, a random projection f mapping $f : R^d \to R^k$ has the property that for all $\mathbf{u}$ and $\mathbf{v}$ in P,

$$(1-\varepsilon)\frac{k}{d}\left|\mathbf{u}-\mathbf{v}\right|^2 \leq \left|f(\mathbf{u}) - f(\mathbf{v})\right|^2 \leq (1+\varepsilon)\frac{k}{d}\left|\mathbf{u}-\mathbf{v}\right|^2$$

with probability at least 9/10.

Proof: Let S be a random k-dimensional subspace and let $f(\mathbf{u})$ be the projection of $\mathbf{u}$ onto S multiplied by the scalar $\sqrt{\frac{d}{k}}$. Applying the Random Projection Theorem 1.11, for any fixed $\mathbf{u}$ and $\mathbf{v}$, the probability that $\|f(\mathbf{u}) - f(\mathbf{v})\|^2$ is outside the range

$$\left[(1-\varepsilon)\tfrac{k}{\beta}|\mathbf{u}-\mathbf{v}|^2, (1+\varepsilon)\tfrac{k}{\beta}|\mathbf{u}-\mathbf{v}|^2\right]$$

is at most

$$e^{-\frac{k\varepsilon^2}{16}} = e^{-4\ln n} = \frac{1}{n^4}.$$

By the union bound the probability that any pair has a large distortion is less than $\binom{n}{2} \times \frac{1}{n^4} \leq \frac{1}{n}$. ∎

Remark: It is important to note that the conclusion of Theorem 1.12 is asserted for all **u** and **v** in P, not just for most **u** and **v**. The weaker assertion for most **u** and **v** is not that useful, since we do not know which **v** would end up being the closest point to **u** and an assertion for most may not cover the particular **v**. A remarkable aspect of the theorem is that the number of dimensions in the projection is only dependent logarithmically on n. Since k is often much less than d, this is called a dimension reduction technique.

For the nearest neighbor problem, if the database has n_1 points and n_2 queries are expected during the lifetime, then take $n = n_1 + n_2$ and project the database to a random k-dimensional space, where $k \geq \frac{64 \ln n}{\varepsilon^2}$. On receiving a query, project the query to the same subspace and compute nearby database points. The theorem says that this will yield the right answer whatever the query with high probability. Note that the exponentially small in k probability in Theorem 1.11 was useful here in making k only dependent on $\log n$, rather than n**.**

1.8 Bibliographic Notes

The word vector model was introduced by Salton [**?**]. Taylor series remainder material can be found in Whittaker and Watson 1990, pp. 95-96. There is vast literature on Gaussian distribution, its properties, drawing samples according to it, etc. The reader can choose the level and depth according to his/her background. For Chernoff bounds and their applications, see [**?**] or [**?**]. The proof here and the application to heavy-tailed distributions is simplified from [**?**]. The original proof of the Random Projection Theorem by Johnson and Lindenstrauss was complicated. Several authors used Gaussians to simplify the proof, see [**?**] for details and applications of the theorem. The proof here is due to Das Gupta and Gupta [**?**].

The SVD based algorithm for identifying the space spanned by centers of spherical Gaussians is from Vempala and Wang [**?**]. The paper handles more complex densities besides spherical Gaussians, but the restricted case of spherical Gaussians contains the key ideas.

1.9 Exercises

Exercise 1.1 *Assume you have 100 million documents each represented by a vector in the word vector model. How would you represent the documents so that given a query with a small number of words you could efficiently find the documents with the highest dot product with the query vector? Given the number of documents you do not have time to look at every document.*

Exercise 1.2 *Let x and y be independent random variables with uniform distribution in $[0, 1]$. What is the expected value $E(x)$? $E(x^2)$? $E(x - y)$? $E(xy)$? and $E((x - y)^2)$?*

Exercise 1.3 *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 cube in 100 dimensions? Give a qualitative answer.*

Exercise 1.4 *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 sphere? What is the cosine of the angle between them?*

Exercise 1.5 *Consider two random 0-1 vectors in high dimension. What is the angle between them? What is the probability that the angle is less than 45?*

Exercise 1.6 *Place two unit-radius 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 1.7 *Generate a 1,000 points at vertices of a 1,000-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 a point k differing from i in only one coordinate so that dist (i, k) and dist (j, k) are both less than dist (i, j). Then continue the path by the same algorithm from k to j. What is the expected length of the path?*

Exercise 1.8 (Overlap of spheres) *Let $\mathbf{x}$ be a random sample from the unit sphere in d-dimensions with the origin as center.*

1. *What is the mean of the random variable x? The mean, denoted $E(\mathbf{x})$, is the vector, whose i^{th} component is $E(x_i)$.*
2. *What is the component-wise variance of $\mathbf{x}$?*
3. *Show that for any unit length vector $\mathbf{u}$, the variance of the real-valued random variable $\mathbf{u}^T \cdot \mathbf{x}$ is $\sum_{i=1}^{d} u_i^2 E(x_i^2)$. Using this, compute the variance and standard deviation of $\mathbf{u}^T\mathbf{x}$.*

4. *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*

$$\frac{4e^{-\frac{a^2(d-1)}{8}}}{a\sqrt{d-1}}$$

times the volume of each sphere.
Hint: Relate the volume of the intersection to the volume of a cap, then, use Lemma 1.2.

5. *From (4), conclude that if the inter-center separation of the two spheres of radius r is $\Omega(r/\sqrt{d})$, 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 cluster is exactly the set of points generated from one sphere. The actual classification requires an efficient algorithm to achive this. Note that the inter-center separation required here goes to zero as d gets larger, provided the radius of the spheres remains the same. So, it is easier to tell apart spheres (of the same radii) in higher dimensions.*

Exercise 1.9 *Derive an upper bound on $\int_{x=\alpha}^{\infty} e^{-x^2}\,dx$ where α is a positive real. Discuss for what values of α this is a good bound.*
Hint: Use $e^{-x^2} \leq \frac{x}{\alpha}e^{-x^2}$ for $x \geq \alpha$.

Exercise 1.10 *What is the formula for the incremental unit of area in using polar coordinates to integrate the area of a circle that lies in a 2-dimensional cone whose vertex is at the center of the circle? What is the formula for the integral? What is the value of the integral if the cone is 36°?*

Exercise 1.11 *For what value of d is the volume, $V(d)$, of a d-dimensional unit sphere maximum?*
Hint: Consider the ratio $\frac{V(d)}{V(d-1)}$ for odd and even values of d.

Exercise 1.12 *How does the volume of a sphere of radius two behave as the dimension of the space increases? What if the radius was larger than two but a constant independent of d? What function of d would the radius need to be for a sphere of radius r to have approximately constant volume as the dimension increases?*

Exercise 1.13

1. *What is the volume of a sphere of radius r in d-dimensions?*
2. *What is the surface area of a sphere of radius r in d-dimensions?*
3. *What is the relationship between the volume and the surface area of a sphere of radius r in d-dimensions?*

4. *Why does the relationship determined in (3) hold?*

Exercise 1.14 *Consider vertices of a d-dimensional cube of width two centered at the origin. Vertices are the points* $(\pm 1, \pm 1, \ldots, \pm 1)$. *Place a unit-radius sphere at each vertex. Each sphere fits in a cube of width two and, thus, no two spheres intersect. Show that the probability that a point of the cube picked at random will fall into one of the* 2^d *unit-radius spheres, centered at the vertices of the cube, goes to 0 as d tends to infinity.*

Exercise 1.15 *Consider the power law probability density*

$$p(x) = \frac{c}{Max(1, x^2)}$$

over the nonnegative real line.

1. *Determine the constant c.*

2. *For a nonnegative random variable x with this density, does* $E(x)$ *exist? How about* $E(x^2)$*?*

Exercise 1.16 *Consider d-space and the following density over the positive orthant:*

$$p(\mathbf{x}) = \frac{c}{Max(1, |\mathbf{x}|^\alpha)}.$$

Show that $\alpha > d$ *is necessary for this to be a proper density function. Show that* $\alpha > d+1$ *is a necessary condition for a (vector-valued) random variable* $\mathbf{x}$ *with this density to have an expected value* $E(|\mathbf{x}|)$. *What condition do you need if we want* $E(|\mathbf{x}|^2)$ *to exist?*

Exercise 1.17 *Consider the upper hemisphere of a unit-radius sphere in d-dimensions. What is the height of the maximum volume cylinder that can be placed entirely inside the hemisphere? As you increase the height of the cylinder, you need to reduce the cylinder's radius so that it will lie entirely within the hemisphere.*

Exercise 1.18 *What is the volume of a radius r cylinder of height h in d-dimensions?*

Exercise 1.19 *For a 1,000-dimensional unit-radius sphere centered at the origin, what fraction of the volume of the upper hemisphere is above the plane* $x_1 = 0.1$*? Above the plane* $x_1 = 0.01$*?*

Exercise 1.20 *Almost all of the volume of a sphere in high dimensions lies in a narrow slice of the sphere at the equator. However, the narrow slice is determined by the point on the surface of the sphere that is designated the North Pole. Explain how this can be true if several different locations are selected for the North Pole.*

Exercise 1.21 *Explain how the volume of a sphere in high dimensions can simultaneously be in a narrow slice at the equator and also be concentrated in a narrow annulus at the surface of the sphere.*

Exercise 1.22 *How large must ε be for 99% of the volume of a d-dimensional unit-radius sphere to lie in the shell of ε-thickness at the surface of the sphere?*

Exercise 1.23

1. *Write a computer program that generates n points uniformly distributed over the surface of a unit-radius d-dimensional sphere.*
2. *Generate 200 points on the surface of a sphere in 50 dimensions.*
3. *Create several random lines through the origin and project the points onto each line. Plot the distribution of points on each line.*
4. *What does your result from (3) say about the surface area of the sphere in relation to the lines, i.e., where is the surface area concentrated relative to each line?*

Exercise 1.24 *If one generates points in d-dimensions 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 1.25 *Project the surface area of a sphere of radius $\sqrt{d}$ in d-dimensions onto a line through the center. For d equal 2 and 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 1.26 *Generate 500 points uniformly at random on the surface of a unit-radius sphere in 50 dimensions. Then randomly generate five additional points. For each of the five new points calculate a narrow band at the equator assuming the point was the North Pole. How many of the 500 points are in each band corresponding to one of the five equators? How many of the points are in all five bands?*

Exercise 1.27 *We have claimed that a randomly generated point on a sphere lies near the equator of the sphere independent of the point selected for the North Pole. Is the same claim true for a randomly generated point on a cube? To test this claim, randomly generate ten ± 1 valued vectors in 128 dimensions. Think of these ten vectors as ten choices for the North Pole. Then generate some additional ± 1 valued vectors. To how many of the original vectors is each of the new vectors close to being perpendicular, that is how many of the equators is each new vectors close to.*

Exercise 1.28 *Consider a slice of a 100-dimensional sphere that lies between two parallel planes each equidistant from the equator and perpendicular to the line from the North to South Pole. What percentage of the distance from the center of the sphere to the poles must the planes be to contain 95% of the surface area?*

Exercise 1.29 *Place n points at random on a d-dimensional unit radius sphere. Assume d is large. Pick a random vector and let it define two parallel hyperplanes on opposite sides of the origin that are equal distance from the origin. How far apart can the hyperplanes be moved and still have no points between them?*

Exercise 1.30 *Consider two random vectors in a high-dimensional space. Assume the vectors have been normalized so that their lengths are one and thus the points lie on a unit sphere. Assume one of the vectors is the North pole. Prove 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. Thus, the probability that the angle between two random vectors is at most 45° goes to zero. How does this relate to the result that most of the volume is near the equator?*

Exercise 1.31 *Project the vertices of a high-dimensional cube onto a line from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$. 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 of the line as center.*

Exercise 1.32 *Draw a 3-dimensional cube and illustrate the equator as defined in Section 1.3.*

Exercise 1.33

1. *What is the surface area of a unit cube in d-dimensions?*
2. *Is the surface area of a unit cube concentrated close to the equator as is the case with the sphere?*

Exercise 1.34 *Consider the simplex*

$$S = \{\mathbf{x} \,|\, x_i \geq 0, 1 \leq i \leq d; \ \sum_{i=1}^{d} x_i \leq 1\}.$$

For a random point $\mathbf{x}$ picked with uniform density from S, find $E(x_1 + x_2 + \cdots + x_d)$. Find the centroid of S.

Exercise 1.35 *How would you sample uniformly at random from the parallelepiped*

$$P = \{\mathbf{x} \,|\, \mathbf{0} \leq A\mathbf{x} \leq \mathbf{1}\},$$

where A is a given nonsingular matrix? How about from the simplex

$$\{\mathbf{x} \,|\, 0 \leq (A\mathbf{x})_1 \leq (A\mathbf{x})_2 \cdots \leq (A\mathbf{x})_d \leq 1\}?$$

Your algorithms must run in polynomial time.

Exercise 1.36 *Randomly generate a 100 points on the surface of a sphere in 3-dimensions and in 100-dimensions. Create a histogram of all distances between the pairs of points in both cases.*

Exercise 1.37 *We have claimed that in high dimensions a unit variance Gaussian centered at the origin has essentially zero probability mass in a unit-radius sphere centered at the origin since the unit-radius sphere has no volume. Explain what happens if the Gaussian has an extremely small variance and has no probability mass outside the unit-radius sphere? How small must the variance be in terms of d for this to happen?*

Exercise 1.38 *Consider two unit-radius spheres in d-dimensions whose centers are distance δ apart where δ is a constant independent of d. Let $\mathbf{x}$ be a random point on the surface of the first sphere and $\mathbf{y}$ a random point on the surface of the second sphere. Prove that the probability that $|\mathbf{x}-\mathbf{y}|^2$ is more than $2+\delta^2+a$ falls off exponentially with a.*

Exercise 1.39 *The Cauchy distribution in one dimension is $Prob(x) = \frac{1}{c+x^2}$. What would happen if one tried to extend the distribution to higher dimensions by the formula $Prob(r) = \frac{1}{1+r^2}$ where r is the distance from the origin? What happens when you try to determine a normalization constant c?*

Exercise 1.40 *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 1.41 *Pick a point $\mathbf{x}$ uniformly at random from the following set in d-space:*

$$K = \{\mathbf{x} | x_1^4 + x_2^4 + \cdots + x_d^4 \leq 1\}.$$

1. *Show that the probability that $x_1^4 + x_2^4 + \cdots + x_d^4 \leq \frac{1}{2}$ is $\frac{1}{2^{d/4}}$.*
2. *Show that with high probability, $x_1^4 + x_2^4 + \cdots + x_d^4 \geq 1 - O(1/d)$.*
3. *Show that with high probability, $|x_1| \leq O(1/d^{1/4})$.*

Exercise 1.42 *Suppose there is an object moving at constant velocity along a straight line. You receive the gps coordinates corrupted by Gaussian noise every minute. How do you estimate the current position?*

Exercise 1.43 *Generate ten values by a Gaussian probability distribution with zero mean and variance one. What is the center determined by averaging the points? What is the variance? In estimating the variance use both the real center and the estimated center. When using the estimated center to estimate the variance, use both $n = 10$ and $n = 9$. How do the three estimates compare?*

Exercise 1.44 *Suppose you want to estimate the (unknown) center of a Gaussian in d-space which has variance one in each direction. Show that $O(\log d/\varepsilon^2)$ random samples from the Gaussian are sufficient to get an estimate $\tilde{\mu}$ of the true center μ so that with probability at least 99/100, we have*

$$|\mu - \tilde{\mu}|_\infty \leq \varepsilon.$$

How many samples are sufficient to ensure that

$$|\mu - \tilde{\mu}| \leq \varepsilon?$$

Exercise 1.45 *Let G be a d-dimensional spherical Gaussian with variance $\frac{1}{2}$ centered at the origin. Derive the expected squared distance to the origin.*

Exercise 1.46 *Show that the maximum of $f(t) = \left(\sqrt{1-2t\beta k}\right)^{(d-k)} \left(\sqrt{1-2t\left(\beta k - d\right)}\right)^k$ is attained at $t = \frac{1-\beta}{2\beta(d-k\beta)}$.*
Hint: Maximize the logarithm of $f(t)$ by differentiating.

Exercise 1.47 *Generate 20 points uniformly at random on a 1000-dimensional sphere of radius 100. Calculate the distance between each pair of points. Then project the data onto subspaces of dimension k=100, 50, 10, 5, 4, 3, 2, 1 and calculate the sum of squared error between $\frac{k}{d}$ times the original distances and the new pair wise distances for each of the above values of k.*

Exercise 1.48 *You are given two sets, P and Q, of n points each in n-dimensional space. Your task is to find the closest pair of points, one each from P and Q, i.e., find $\mathbf{x}$ in P and $\mathbf{y}$ in Q such that $|\mathbf{x} - \mathbf{y}|$ is minimum.*

1. *Show that this can be done in time $O(n^3)$.*
2. *Show how to do this with relative error 0.1% in time $O(n^2 \ln n)$, i.e., you must find a pair $\mathbf{x} \in P, \mathbf{y} \in Q$ so that the distance between them is, at most, 1.001 times the minimum possible distance. If the minimum distance is 0, you must find $\mathbf{x} = \mathbf{y}$.*

Exercise 1.49 *Given n data points in d-space, find a subset of k data points whose vector sum has the smallest length. You can try all $\binom{n}{k}$ subsets, compute each vector sum in time $O(kd)$ for a total time of $O\left(\binom{n}{k}kd\right)$. Show that we can replace d in the expression above by $O(k \ln n)$, if we settle for an answer with relative error .02%.*

Exercise 1.50 *Create a list of the five most important things that you learned about high dimensions.*

Exercise 1.51 *Write a short essay whose purpose is to excite a college freshman to learn about high dimensions.*