1 High-Dimensional Space

In many applications data is in the form of vectors. In other applications, data is not in the form of vectors, but could be usefully represented by vectors. The Vector Space Model [?] is a good example. In the vector space model, a document is represented by a vector, each component of which corresponds to the number of occurrences of a particular term in the document. The English language has on the order of 25,000 words or terms, so each document is represented by a 25,000 dimensional vector. A collection of $n$ documents is represented by a collection of $n$ vectors, one vector per document. The vectors may be arranged as columns of a $25,000 \times n$ matrix. See Figure 1.1. A query is also represented by a vector in the same space. The component of the vector corresponding to a term in the query, specifies the importance of the term to the query. 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.

One needs a measure of relevance or similarity of a query to a document. The dot product or cosine of the angle between the two vectors is an often used measure of similarity. To respond to a query, one computes the dot product or the cosine of the angle between the query vector and each document vector and returns the documents with the highest values of these quantities. While it is by no means clear that this approach will do well for the information retrieval problem, many empirical studies have established the effectiveness of this general approach.

The vector space model is useful in ranking or ordering a large collection of documents in decreasing order of importance. For large collections, an approach based on human understanding of each document is not feasible. Instead, an automated procedure is needed that is able to rank documents with those central to the collection ranked highest. Each document is represented as a vector with the vectors forming the columns of a matrix $A$. The similarity of pairs of documents is defined by the dot product of the vectors. All pairwise similarities are contained in the matrix product $A^T A$. If one assumes that the documents central to the collection are those with high similarity to other documents, then computing $A^T A$ enables one to create a ranking. Define the total similarity of document $i$ to be the sum of the entries in the $i^{th}$ row of $A^T A$ and rank documents by their total similarity. It turns out that with the vector representation on hand, a better way of ranking is to first find the best fit direction. That is, the unit vector $u$, for which the sum of squared perpendicular distances of all the vectors to $u$ is minimized. See Figure 1.2. Then, one ranks the vectors according to their dot product with $u$. The best-fit direction is a well-studied notion in linear algebra. There is elegant theory and efficient algorithms presented in Chapter ?? that facilitate the ranking as well as applications in many other domains.

In the vector space representation of data, properties of vectors such as dot products,
distance between vectors, and orthogonality, often have natural interpretations and this is what makes the vector representation more important than just a book keeping device. For example, the squared distance between two 0-1 vectors representing links on web pages is the number of web pages linked to by only one of the pages. In Figure 1.3, 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. We have seen that dot products measure similarity. Orthogonality of two nonnegative vectors says that they are disjoint. Thus, if a document collection, e.g., all news articles of a particular year, contained documents on two or more disparate topics, vectors corresponding to documents from different topics would be nearly orthogonal.

The dot product, cosine of the angle, distance, etc., are all measures of similarity or dissimilarity, but there are important mathematical and algorithmic differences between them. The random projection theorem presented in this chapter states that a collection of vectors can be projected to a lower-dimensional space approximately preserving all pairwise distances between vectors. Thus, the nearest neighbors of each vector in the collection can be computed in the projected lower-dimensional space. Such a savings in time is not possible for computing pairwise dot products using a simple projection.

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.

High dimensionality is a common characteristic in many models and for this reason much of this chapter is devoted to the geometry of high-dimensional space, which is quite different from our intuitive understanding of two and three dimensions. We focus first on volumes and surface areas of high-dimensional objects like hyperspheres. We will not present details of any one application, but rather present the fundamental theory useful to many applications.

Probability comes in is that many computational problems are hard if our algorithms are required to be efficient on all possible data. In practical situations, domain knowledge often enables the expert to formulate stochastic models of data. In customer-product
Figure 1.2: The best fit line is the line that minimizes the sum of the squared perpendicular distances.

data, a common assumption is that the goods each customer buys are independent of what goods the others buy. One may also assume that the goods a customer buys satisfies a known probability law, like the Gaussian distribution. In keeping with the spirit of the book, we do not discuss specific stochastic models, but present the fundamentals. An important fundamental is the law of large numbers that states that under the assumption of independence of customers, the total consumption of each good is remarkably close to its mean value. The central limit theorem is of a similar flavor. Indeed, it turns out that picking random points from geometric objects like hyperspheres exhibits almost identical properties in high dimensions.

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 $|\mathbf{x} - \mathbf{y}|^2$ 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 a 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 $1 - \frac{c}{\sqrt{d}}$ and 1, where $c$ is a constant independent of $d$. This implies that most of the mass is near the surface of the sphere. 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 results in this chapter, but first a review of some probability.

### 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 simultaneously 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, which are contrary to our two or three-dimensional intuition, will be proved by relating them to probabilities.
1.2.1 The Sphere and the Cube in High 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.

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

$$\sqrt{(\frac{1}{2})^2 + (\frac{1}{2})^2} = \sqrt{\frac{1}{2}} \approx 0.707.$$ 

Here, 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

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

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 $\sqrt{d}$, and for large $d$, the vertices of the cube lie far outside the unit radius sphere. Figure 1.5 illustrates conceptually a cube and a sphere. The vertices of the cube are at distance $\sqrt{\frac{d}{2}}$ 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 The Volume is in a Narrow Annulus

Another big difference between 2 or 3 dimensional spheres and the $d$ dimensional sphere for $d$ large is that most of the volume of the latter lies near the surface. This is the easiest of the properties of the sphere; we will prove it in this section. First note that for $d = 2$, (for $d = 2$, volume is what we generally call “area”) for any $r \in [0, 1]$, the volume of the sphere at distance $r$ from the center is proportional to $r$. To see this, note that in the figure (??), the infinitesimal piece of the cone at distance $r$ has arc length $2\pi r \, d\theta$ and (this is important to note), the radial direction (along which we increment $r$ by $dr$) is orthogonal to the arc. [The orthogonality is necessary because otherwise, the area would also involve the sin of the angle between the arc and the radial direction.] Thus, the expected distance of a random point from the origin in a circle in 2 dimensions is:

$$\int_{r=0}^{1} \frac{2\pi r^2 \, dr}{\pi r^2} = \frac{2}{3}.$$

In 3 dimensions, the volume of the sphere at distance $r$ from the center is proportional to $r^2 \, dr$, since, now instead of an infinitesimal piece of arc, we have an infinitesimal piece of the surface of a 3-dimensional sphere and surface areas are proportional to $r^2$. Again, this infinitesimal surface is orthogonal to the radial direction. So, a similar calculation in 3 dimensions tells us the expected distance of a random point from the origin is

$$\int_{r=0}^{1} \frac{4\pi r^3 \, dr}{(4/3)\pi r^3} = \frac{3}{4}.$$

As the reader can guess, as $d$ goes to infinity, the expected distance goes to 1. So, (at least in expectation), the point is close to the surface. We do not prove this here. Instead, we show that only an exponential fraction of the volume lies in the interior. 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 $\varepsilon$ is a fixed constant. In high dimensions, all of the volume of the sphere is concentrated in a narrow annulus at the surface.
Figure 1.7: 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.

Since, $(1 - \varepsilon)^d \leq e^{-c d}$, 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 unit sphere is contained in an annulus of width $O(1/d)$ near the boundary. If the sphere is of radius $r$, then for sufficiently large $d$, the volume is contained in an annulus of width $O\left(\frac{r}{d}\right)$.

1.3 Volumes and Probabilities

Volumes are related directly to probabilities. For example, to find the volume of a sphere $S$ of radius 1 with the origin as center, enclose the sphere in a cube $C$ of side 2. Pick a random point $\mathbf{x}$ from the cube. [This is easy to do: just pick each coordinate $x_i$ independently, uniformly at random from the interval $[-1, 1]$.] Then,

$$\text{Prob}(\mathbf{x} \in S) = \frac{\text{Vol}(S)}{\text{Vol}(C)} = \frac{\text{Vol}(S)}{2^d}.$$ 

So in principle, by picking many random points from $C$, we can calculate this ratio and hence the volume of $S$. However, this turns out to be very inefficient, since we will see that the ratio goes down exponentially (in $d$) and so we need exponentially many samples to get a good estimate. [Recall from basic Statistics: suppose we are trying to estimate the probability $p$ that a coin comes up heads by tossing it $n$ times and keeping track of the fraction $x$ of heads. The variance of $x$ is $p(1-p)/n$. So we need $n > 1/p$ to get any good estimate. In our situation, $p$ is exponentially small.]

Here we want to estimate the volume of the sphere theoretically. It turns out that a slightly different procedure will be useful for this. We will show that if we pick a point $\mathbf{x}$ uniformly at random from the unit sphere $S$, with high probability, each coordinate $x_i$ satisfies $|x_i| \leq \alpha/\sqrt{d}$ for a small $\alpha$. Intuitively, this is something we expect to be true, since $S = \{\mathbf{x} : x_1^2 + x_2^2 + \cdots + x_d^2 \leq 1\}$ and by symmetry, each of the $x_i^2$ should not exceed $1/d$ (by too much). Once we prove this, it is clear that most of the volume of the sphere is contained in a cube of volume $\alpha^d/d^{d/2}$ which goes to zero since our $\alpha$ will be much smaller than $\sqrt{d}$. But for this, we have to see how to generate a point uniformly at random from
the unit sphere. This task is not as trivial as generating from a cube where we could pick the coordinates independently.

We will relate generating random points from the sphere to generating points from the Gaussian density. We will then introduce some fundamental probability properties which are useful not only for bounding volumes of spheres, but indeed are basic to many topics in the book. We will then return to the sphere in Section (1.6) and prove the assertions we have made here.

1.4 Generating Points Uniformly at Random from a Sphere

First, consider generating points uniformly at random on the surface of a unit-radius sphere. For the 2-dimensional version of generating points on the circumference of a unit-radius circle, first look at 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 2 by 2 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 an independent Gaussian variable—i.e.:

Generate $x_1, x_2, \ldots, x_d$, where the $x_i$ are i.i.d. (independent, identically distributed), each according to the normal (Gaussian) density with mean 0 and variance 1, namely, $\frac{1}{\sigma \sqrt{2\pi}} \exp(-x^2/2)$ on the real line. So the probability density of $x$ is

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

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

Now to generate a point $y$ uniformly over the sphere, we have to scale the point $\frac{x}{|x|}$ generated on the surface by a scaler $\rho \in [0, 1]$. What is the distribution of $\rho$? It is certainly not uniform, even in 2 dimensions. Indeed, we saw that the volume at distance $r$ is proportional to $r$ and so the density of $\rho$ at $r$ is proportional to $r$ for $d = 2$. Similarly, for $d = 3$, it is proportional to $r^2$. You may want to consult the figure (1.8). By similar reasoning, it is easy to see that in $d$ dimensions, the density of $\rho$ at distance $r$ is proportional to $r^{d-1}$. Normalizing, we see that we should pick $\rho$ with probability density (of $\rho = r$) equal to $dr^{d-1}$, over $[0, 1]$. 

9
Figure 1.8: Infinitesimal volume in a $d$-dimensional sphere of unit radius.

$$x \sim c \exp(-|x|^2/2)$$

Figure 1.9: Generating a point from the sphere

Now we have succeeded in generating a point

$$y = \rho \frac{x}{|x|}$$

uniformly at random from the unit sphere $S$. We wish to show that each coordinate of $x/|x|$ is at most $\alpha/\sqrt{d}$ in absolute value for a small $\alpha$. It is easy to bound the numerator - $x_i$ since it has Gaussian density. The difficult part is to estimate $|x|$ or say $|x|^2 = \sum_{i=1}^d x_i^2$. For this, we draw upon an important part of Probability Theory which shows that sums of independent random variables (as $x_1^2 + x_2^2 + \cdots + x_d^2$ is) behave nicely, in the sense that their value is concentrated close to the mean value. A geometric way of saying this is that the Gaussian density concentrates mass in a thin annulus. We defer the proof of this using Probability tail bounds to Sections (1.9) and (1.10); the tail bounds find many applications besides the Gaussian annulus. In the next section, we give intuition for the thin Gaussian annulus and state the precise theorem, which finds immediate use in proving that the mass of the sphere is concentrated near the equator in Section (1.6).

1.5 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 $\sigma$ has density function

$$p(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} \exp\left(-\frac{|x|^2}{2\sigma^2}\right) .$$
Figure 1.10: Most of the probability mass of $d$ dimensional Gaussian of radius $r$ is contained in an annulus of width $O(r/\sqrt{d})$.

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 following theorem states that the mass is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{d}$. It will be proved in Section (1.10). But we will use it extensively in the next few sections. First, note that

$$E(|x|^2) = \sum_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d.$$  

So the mean squared distance of a point from the center is $d$. We call the square root of the mean squared distance, namely $\sqrt{d}$ here, the radius of the Gaussian.

**Theorem 1.1 Gaussian Annulus Theorem** For a $d$–dimensional unit variance spherical Gaussian, for any positive real number $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the mass lies within the annulus $\sqrt{d} - \beta \leq r \leq \sqrt{d} + \beta$, where, $c$ is a fixed positive constant.

1.6 Volume of a sphere

1.6.1 Volume is near the equator

Consider a high-dimensional unit-radius 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 \( \{x \mid |x| \leq 1, x_1 = 0\} \), called the equator. In three dimensions this region is a circle, in four dimensions the region is a 3-dimensional sphere, etc. In our terminology, a circle is a 2-dimensional sphere and its volume is what one usually refers to as the area of a circle. The surface area of the 2-dimensional sphere is what one usually refers to as the...
Figure 1.11: Most of the volume of the $d$-dimensional sphere of radius $r$ is within distance $O(\frac{r}{\sqrt{d}})$ of the equator.

circumference of a circle.

It turns out that essentially all of the volume 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 $\varepsilon$ does essentially all the volume lie between $x_1 = 0$ and $x_1 = \varepsilon$? The answer depends on the dimension. For dimension $d$, it is $O(\frac{1}{\sqrt{d}})$. Before we prove this, some intuition is in order. Since $|x|^2 = x_1^2 + x_2^2 + \cdots + x_d^2$ and by symmetry, we expect the $x_i$’s to be generally equal (or close to each other), we expect each $x_i^2$ to be at most $O(\frac{1}{d})$. Now we prove this formally.

Lemma 1.2 For $\gamma \geq 1$, the fraction of the volume of the unit hemisphere above the plane $x_1 = \frac{\gamma}{\sqrt{d}}$ is less than $3e^{-c\gamma^2}$, where, $c$ is a fixed constant.

Proof: Recall the way of generating a random point $y$ from the unit sphere described in section (1.4): Pick $x = (x_1, x_2, \ldots, x_d)$, i.i.d., each according to $N(0,1)$ and let $y = \rho \frac{x}{|x|}$, where, $\rho \in [0,1]$ is picked according to the density: $\text{Prob}(\rho = r) = dr^{d-1}$. Applying Theorem (1.1), we get that with probability at least $1 - \exp(-cd)$, we have $|x| \geq \sqrt{d}/2$. Also from the definition of gaussian density, we see that

$$\text{Prob}(|x_1| \geq \gamma/2) = \frac{2}{\sqrt{2\pi}} \int_{z=\gamma/2}^{\infty} e^{-z^2/2} dz \leq \frac{2}{\gamma} e^{-\gamma^2/8},$$

where, we have employed an useful trick of putting in the extra term $z/(\gamma/2)$ which is at least 1 to make it integrable. Using the union bound, we get (provided $\gamma \in [1, \sqrt{d}]$)

$$\text{Prob}(|y_1| \geq \gamma/\sqrt{d}) \leq \text{Prob}(|x| < \sqrt{d}/2) + \text{Prob}(|x_1| \geq \gamma/2) \leq e^{cd} + \frac{2}{\gamma} e^{-\gamma^2/8} \leq 3e^{-c\gamma^2},$$

the last inequality using $\gamma \in [1, \sqrt{d}]$. If $\gamma > \sqrt{d}$, the volume above $\{x_1 = \gamma/\sqrt{d}\}$ is zero and there is nothing to prove.
1.6.2 Surface Area is near the equator

Essentially the same line of proof shows that the surface area of the unit sphere is close to the equator, since we picked a random point from the surface of the unit sphere first (using Gaussian samples). We state the result without proof here.

**Lemma 1.3** For $\gamma > 0$, the fraction of the surface area of the unit hemisphere above the plane $x_1 = \frac{\gamma}{\sqrt{d}}$ is less than $\frac{2}{\gamma} e^{-c\gamma^2}$, where, $c$ is a fixed constant.

1.6.3 Volume of the sphere goes to zero

For a fixed dimension $d$, the volume of the sphere is a function of its radius and grows as $r^d$. For fixed radius, the volume of the sphere is a function of the dimension of the space. What is interesting is that the volume of the unit sphere goes to zero as the dimension of the sphere increases. We will show this now. As in the last section, recall the process of generating a random point from the sphere: generate $x$, with $x_i$ independent from $N(0,1)$, then, set $y = \frac{x}{|x|}$. By Theorem (1.1) again, we see that with probability at least $1 - \exp(-cd)$, we have $|x| \geq \sqrt{d}/2$. Also as in the last section, from the definition of gaussian density, for each fixed $i, i \in \{1, 2, \ldots, d\}$, and a positive real number $a$, the probability that $|x_i| \geq a$ is at most $\frac{1}{a} \exp(-a^2/2)$. Set $a = 2\sqrt{\ln d}$ and use the union bound over all $i$ to get that

$$\text{Prob}(|x_i| \geq 2\sqrt{\ln d} \text{ for some } i) \leq \frac{d}{2\sqrt{\ln d}} \exp(-2\ln d) \leq \frac{1}{d}.$$  

Thus, with probability at least $1 - \frac{1}{d} - \exp(-cd)$, we have that $|y_i| \leq 4\sqrt{\ln d}/\sqrt{d}$ for all $i$. Thus, at least half the volume of the sphere lies in the cube

$$\{ y : |y_i| \leq 4\sqrt{\ln d},$$

which has side length $4\sqrt{\ln d}/\sqrt{d}$ and has volume $4^d(\ln d)^{d/2}/d^{d/2}$. So we get:

**Lemma 1.4** The volume $V(d)$ of a unit-radius sphere in $d$ dimensions is at most

$$4^{d+1}(\ln d)^{d/2}/d^{d/2}$$

and so $V(d) \to 0$ as $d \to \infty$.

We remark that this is not the best upper bound on the volume of the sphere. Indeed, the $\sqrt{\ln d}$ can be done away with. One way to do that is to note that the unit sphere is contained in the “octohedron” $\sum_{i=1}^d |x_i| \leq \sqrt{d}$. It turns out that the volume of the octohedron can be explicitly found and it is $(\sqrt{d})^d/d!$ which is at most $c^d/d^{d/2}$ by Stirling approximation. We do not do this here.


1.7 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

\[ R = \{ x \mid l_1 \leq x_1 \leq u_1, l_2 \leq x_2 \leq u_2, \ldots, l_d \leq x_d \leq u_d \} \]

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

\[ P = \{ x \mid l \leq A x \leq u \} \]

where \( A \) is an invertible \( d \times d \) matrix, and \( l \) and \( u \) are lower and upper bound vectors, respectively. The statements \( l \leq A x \) and \( A x \leq u \) 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. Let

\[ R = \{ y \mid l \leq y \leq u \}. \]

The map \( x = A^{-1} y \) 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 \( \{(x,y) \mid 0 \leq y \leq x \leq 1\} \). 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

\[ S = \{ x \mid 1 \geq x_1 \geq x_2 \geq \cdots \geq x_d \geq 0 \}. \]

How many copies of this simplex exactly fit into the unit square, \( \{x \mid 0 \leq x_i \leq 1\} \)? Every point in the square has some ordering of its coordinates. 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 \( y \) in \( R \) is mapped to an \( x \) 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

14
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 \( v_1, v_2, \ldots, v_{d+1} \), then translating the simplex by \( -v_{d+1} \) results in vertices \( v_1 - v_{d+1}, v_2 - v_{d+1}, \ldots, v_d - v_{d+1}, 0 \). Let \( A \) be the \( d \times d \) matrix with columns \( v_1 - v_{d+1}, v_2 - v_{d+1}, \ldots, v_d - v_{d+1} \). Then, \( A^{-1}T = R \) and \( AR = T \) where \( R \) is the right angle simplex. Thus, the volume of \( T \) is \( \frac{1}{d!} |\text{Det}(A)| \).

### 1.8 Random Projection and Johnson-Lindenstrauss Theorem

One of the most frequently used subroutines for high dimensional data is the Nearest Neighbor Search (NNS) problem. In NNS, we are given a database of \( n \) points in \( \mathbb{R}^d \), where, usually, \( n, d \) are large. The database can be preprocessed and stored in an efficient data structure. Thereafter, we are presented “query” points in \( \mathbb{R}^d \) and are to find the nearest or approximately nearest database point to the query point. Since the number of queries is often large, query time (time to answer a single query) should be very small (ideally a small function of \( \log n, \log d \)), whereas, preprocessing time could be larger (a polynomial function of \( n, d \)). For this and other problems, dimension reduction, where, one projects the database points to a \( k \) dimensional space with \( k < < d \) (usually dependent on \( \log d \)) is useful provided comparing distances in the projected space gives us (approximately) correct answers to distances in \( \mathbb{R}^d \). We will see using the Gaussian Annullus theorem that such a projection indeed exists and is simple.

The projection \( f: \mathbb{R}^d \to \mathbb{R}^k \) with desirable properties is defined as follows. Pick in i.i.d. trials \( k \) vectors \( u_1, u_2, \ldots, u_k \), each with Gaussian distribution \( \frac{1}{(2\pi)^{d/2}} \exp(-|x|^2/2) \). We then define for any vector \( v \), the projection \( f(v) \) by:

\[
f(v) = (u_1 \cdot v, u_2 \cdot v, \ldots, u_k \cdot v).
\]

So, \( f(v) \) is just a vector of dot products of \( v \) with the \( u_i \). It will be easy to show that \( |f(v)| \approx \sqrt{k}|v| \), so if we have to find the distance \( |v_1 - v_2| \) between two vectors \( v_1, v_2 \) in \( \mathbb{R}^d \), it will suffice instead to compute \( |f(v_1) - f(v_2)| \) in \( k \) dimensional space (since the factor of \( \sqrt{k} \) is known and we can just divide by it.)

**Theorem 1.5 (The Random Projection Theorem)** Let \( v \) be a fixed vector in \( \mathbb{R}^d \) and let \( f \) be defined as above. Then, for \( \varepsilon \in (0, 1) \),

\[
\text{Prob} \left( \left| \left| f(v) \right| - \sqrt{k}|v| \right| \geq \varepsilon \sqrt{k}|v| \right) \leq 3e^{-ck\varepsilon^2}.
\]

**Proof:** By scaling both sides by \( |v| \), we may assume that \( |v| = 1 \). The sum of independent normally distributed real variables is also normally distributed; the means and variances just sum up. Since \( u_i \cdot v = \sum_{j=1}^d u_{ij}v_j \), we see that the random variable \( u_i \cdot v \) has Gaussian density with mean 0 and variance equal to \( \sum_{j=1}^d v_j^2 = |v|^2 = 1 \). Further, \( u_1 \cdot v, u_2 \cdot v, \ldots, u_k \cdot v \) are independent. So the current theorem follows from the Gaussian annulus theorem (1.1). \( \square \)
The random projection theorem establishes that the probability of the length of the projection of a single vector differing significantly from its expected value is exponentially small in $k$, the dimension of the target subspace. By a union bound, the probability that any of $O(n^2)$ pairwise differences among $n$ vectors differs significantly from their expected values is small, provided $k \varepsilon^2$ is $\Omega(\ln n)$. Thus, the projection to a random subspace preserves all relative pairwise distances between points in a set of $n$ points. This is the content of the Johnson-Lindenstrauss theorem.

**Theorem 1.6 (Johnson-Lindenstrauss Theorem)** For any $0 < \varepsilon < 1$ and any integer $n$, let $k$ satisfy $k \varepsilon^2 \in \Omega(\ln n)$. For any set $P$ of $n$ points in $\mathbb{R}^d$, a random projection $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ defined above has the property that for all $u$ and $v$ in $P$ with probability at least $1 - (1.5/n)$,

$$(1 - \varepsilon) \sqrt{\frac{k}{d}} |u - v| \leq |f(u) - f(v)| \leq (1 + \varepsilon) \sqrt{\frac{k}{d}} |u - v|.$$  

**Proof:** Applying the random projection theorem (Theorem 1.5), for any fixed $u$ and $v$, the probability that $|f(u) - f(v)|$ is outside the range

$$\left[ (1 - \varepsilon) \sqrt{\frac{k}{d}} |u - v|, (1 + \varepsilon) \sqrt{\frac{k}{d}} |u - v| \right]$$

is at most

$$3e^{-ck\varepsilon^2} \leq \frac{3}{n^3}$$

for $k \geq \frac{\varepsilon \ln n}{\varepsilon^2}$. By the union bound, the probability that some pair has a large distortion is less than $\binom{n}{2} \times \frac{3}{n^3} \leq \frac{1.5}{n}$.

**Remark:** It is important to note that the conclusion of Theorem 1.6 asserts 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 $v$ might 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, take $n = n_1 + n_2$ and project the database to a random $k$-dimensional space, where $k \geq \frac{102\ln n}{\varepsilon^2}$. On receiving a query, project the query to the same subspace and compute nearby database points. The Johnson Lindenstrauss theorem says that with high probability this will yield the right answer whatever the query. Note that the exponentially small in $k$ probability in Theorem 1.5 was useful here in making $k$ only dependent on $\ln n$, rather than $n$. 

16
1.9 Bounds on Tail Probability

Markov’s inequality bounds the tail probability of a nonnegative random variable $x$ based only on its expectation. For $a > 0$,

$$\text{Prob}(x > a) \leq \frac{E(x)}{a}.$$  

As $a$ grows, the bound drops off as $1/a$. Given the second moment of $x$, Chebyshev’s inequality, which does not assume $x$ is a nonnegative random variable, gives a tail bound falling off as $1/a^2$

$$\text{Prob}(|x - E(x)| \geq a) \leq \frac{E((x - E(x))^2)}{a^2}.$$

Higher moments yield bounds by applying either of these two theorems. For example, if $r$ is a nonnegative even integer, then $x^r$ is a nonnegative random variable even if $x$ takes on negative values. Applying Markov’s inequality to $x^r$,

$$\text{Prob}(|x| \geq a) = \text{Prob}(x^r \geq a^r) \leq \frac{E(x^r)}{a^r},$$

a bound that falls off as $1/a^r$. The larger the $r$, the greater the rate of fall, but a bound on $E(x^r)$ is needed to apply this technique.

For a random variable $x$ that is the sum of a large number of independent random variables, $x_1, x_2, \ldots, x_n$, one can derive bounds on $E(x^r)$ for high even $r$. There are many situations where the sum of a large number of independent random variables arises. For example, $x_i$ may be the amount of a good that the $i^{th}$ consumer buys, the length of the $i^{th}$ message sent over a network, or the indicator random variable of whether the $i^{th}$ record in a large database has a certain property. Each $x_i$ is modeled by a simple probability distribution. Gaussian, exponential (probability density at any $t > 0$ is $e^{-t}$), or binomial distributions are typically used, in fact, respectively in the three examples here. If the $x_i$ have 0-1 distributions, there are a number of theorems called Chernoff bounds, bounding the tails of $x = x_1 + x_2 + \cdots + x_n$, typically proved by the so-called moment-generating function method (see Section ?? of the appendix). But exponential and Gaussian random variables are not bounded and these methods do not apply. However, good bounds on the moments of these two distributions are known. Indeed, for any integer $s > 0$, the $s^{th}$ moment for the unit variance Gaussian and the exponential are both at most $s!$.

Given bounds on the moments of individual $x_i$ the following theorem proves moment bounds on their sum. We use this theorem to derive tail bounds not only for sums of 0-1 random variables, but also Gaussians, exponentials, Poisson, etc.

The gold standard for tail bounds is the Central Limit Theorem for independent, identically distributed random variables $x_1, x_2, \cdots, x_n$ with zero mean and $\text{Var}(x_i) = \sigma^2$.
that states as $n \to \infty$ the distribution of $x = (x_1 + x_2 + \cdots + x_n) / \sqrt{n}$ tends to the Gaussian density with zero mean and variance $\sigma^2$. Loosely, this says that in the limit, the tails of $x = (x_1 + x_2 + \cdots + x_n) / \sqrt{n}$ are bounded by that of a Gaussian with variance $\sigma^2$. But this theorem is only in the limit, whereas, we prove a bound that applies for all $n$.

In the following theorem, $x$ is the sum of $n$ independent, not necessarily identically distributed, random variables $x_1, x_2, \ldots, x_n$, each of zero mean and variance at most $\sigma^2$. By the central limit theorem, in the limit the probability density of $x$ goes to that of the Gaussian with variance at most $n\sigma^2$. In a limit sense, this implies an upper bound of $ce^{-a^2/(2n\sigma^2)}$ for the tail probability $\text{Prob}(|x| > a)$ for some constant $c$. The following theorem assumes bounds on higher moments, but asserts a quantitative upper bound of $3e^{-a^2/(12n\sigma^2)}$ on the tail probability, not just in the limit, but for every $n$. We will apply this theorem to get tail bounds on sums of Gaussian, binomial, and power law distributed random variables.

**Theorem 1.7** Let $x = x_1 + x_2 + \cdots + x_n$, where $x_1, x_2, \ldots, x_n$ are mutually independent random variables with zero mean and variance at most $\sigma^2$. Assume for $s = 3, 4, \ldots, \lfloor (a^2/4n\sigma^2) \rfloor$, $|E(x_i^s)| \leq \sigma^2 s!$, then for $0 \leq a \leq \sqrt{2}n\sigma^2$,

$$\text{Prob}(|x| \geq a) \leq 3e^{-a^2/(12n\sigma^2)}.$$

**Proof:** We first prove an upper bound on $E(x^r)$ for any even positive integer $r \leq s$ and then use Markov’s inequality as discussed earlier. Expand $(x_1 + x_2 + \cdots + x_n)^r$.

$$(x_1 + x_2 + \cdots + x_n)^r = \sum \binom{r}{r_1, r_2, \ldots, r_n} x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n} = \sum r! x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}$$

where the $r_i$ range over all nonnegative integers summing to $r$. By independence

$$E(x^r) = \sum \frac{r!}{r_1!r_2!\cdots r_n!} E(x_1^{r_1}) E(x_2^{r_2}) \cdots E(x_n^{r_n}).$$

If in a term, any $r_i = 1$, the term is zero since $E(x_i) = 0$. Assume henceforth that $(r_1, r_2, \ldots, r_n)$ runs over sets of nonzero $r_i$ summing to $r$ where each nonzero $r_i$ is at least two. Let

$$J = \{(r_1, r_2, \ldots, r_n) : r_i \in \{0, 2, 3, \ldots \} ; \sum_{i=1}^n r_i = r\}.$$

Since $|E(x_i^{r_i})| \leq \sigma^2 r_i!$,

$$E(x^r) \leq r! \sum_{(r_1, r_2, \ldots, r_n) \in J} \sigma^2 \text{number of nonzero } r_i \text{ in set}. $$

Collect terms of the summation with $t$ nonzero $r_i$ for $t = 1, 2, \ldots, r/2$. Let

$$J_t = \{(r_1, r_2, \ldots, r_n) \in J : \text{number of non-zero } r_i = t\}.$$
So,

\[ E(x^r) = r! \sum_{t=1}^{r/2} |J_t| \sigma^{2t}. \]

We now bound \(|J_t|\). There are \(\binom{n}{t}\) subsets of \(\{1, 2, \ldots, n\}\) of cardinality \(t\). Once a subset is fixed as the set of \(t\) values of \(i\) with nonzero \(r_i\), set each of the \(r_i \geq 2\). That is, allocate two to each of the \(r_i\) and then allocate the remaining \(r - 2t\) to the \(t\) \(r_i\) arbitrarily. The number of such allocations is just \(\binom{r - 2t - 2 + t - 1}{t - 1} = \binom{r - t - 1}{t - 1}\). So,

\[ |J_t| \leq \binom{n}{t} \binom{r - t - 1}{t - 1}, \]

\[ E(x^r) \leq r! \sum_{t=1}^{r/2} \binom{n}{t} \binom{r - t - 1}{t - 1} \sigma^{2t} \leq \sum_{t} \frac{(n \sigma^2)^t}{t!} 2^{r-t-1}. \]

Let \(h(t) = \frac{(n \sigma^2)^t}{t!} 2^{r-t-1}\). In the hypotheses of the theorem \(a \leq \sqrt{2} n \sigma^2\) and \(s \leq \frac{a^2}{2 n \sigma^2}\). Thus, \(r\) is at most \(n \sigma^2/2\). For \(t \leq r/2\), increasing \(t\) by one, increases \(h(t)\) by at least \(n \sigma^2/2\), which is at least two. This gives

\[ E(x^r) = r! \sum_{t=1}^{r/2} f(t) \leq r! h(r/2) (1 + \frac{1}{2} + \frac{1}{4} + \cdots) \leq \frac{r!}{(r/2)!} 2^{r/2} (n \sigma^2)^{r/2}. \]

Applying Markov inequality,

\[ \text{Prob}(|x| > a) = \text{Prob}(|x|^r > a^r) \leq \frac{r! (n \sigma^2)^{r/2} 2^{r/2}}{(r/2)! a^r} \leq \left( \frac{2n \sigma^2}{a^2} \right)^{r/2}. \]

The bound applies for any \(r \leq s\). Take \(r\) to be the largest even integer less than or equal to \(a^2/(6n \sigma^2)\). [By Calculus, we see that the function \(f(x) = (cx)^{x/2}\) is minimized at \(x = 1/ec\) (just differentiate \(\ln(f(x))\)). So, \(r = a^2/(2en \sigma^2)\) minimizes the upper bound. Our choice here replaces \(2e\) by \(6\).] The tail probability is at most \(e^{-r/2}\), which is at most \(e^{-a^2/(12n \sigma^2)} \leq 3 \cdot e^{-a^2/(8n \sigma^2)}\), proving the theorem.

\[ \square \]

1.10 Applications of the tail bound

Calculation of width of the Gaussian annulus

Let \((y_1, y_2, \ldots, y_d)\) be a unit variance Gaussian centered at the origin. We argue that the mass of the Gaussian is in a narrow annulus of width \(O(1)\) of a sphere of radius approximately \(\sqrt{d}\). It is easier to deal with squared distance to the origin rather than distance. Thus, we ask what is the probability that \(|y_1^2 + y_2^2 + \cdots + y_d^2 - d| \geq \beta\)? Let \(x_i = y_i^2 - 1\) and change the question to what is the probability that \(|x_1 + x_2 + \cdots + x_d| \geq \beta\) to which we can apply Theorem 1.7.
Theorem 1.7 requires bounds on the moments of the $x_i$. For $|y_i| \leq 1$, $|x_i|^s \leq 1$ and for $|y_i| \geq 1$, $|x_i|^s \leq |y_i|^2s$. Thus

$$|E(x_i^s)| = E(|x_i|^s) \leq E(1 + y_i^{2s}) = 1 + E(y_i^{2s})$$

$$= 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty y^{2s} e^{-y^2/2} dy$$

Using the substitution $y^2 = 2z$,

$$|E(x_i^s)| = 1 + \frac{1}{\sqrt{\pi}} \int_0^\infty 2^s z^{s-(1/2)} e^{-z} dz$$

$$\leq 2^s s!.$$  

The last inequality is from the Gamma integral.

$E x_i = 0$ and so $\text{Var}(x_i) = E(x_i^2) \leq 2^2 2 = 8$. But to make $|E(x_i^s)| \leq 8s!$ as required in theorem (1.7), we use $w_i = x_i/2$. Then, $\text{Var}(w_i) = 2$ and $|E(w_i^s)| \leq 2s!$.

**Proof:** (of Theorem (1.1)) Let $r$ be the distance to a point generated by the Gaussian. If $|r - \sqrt{d}| \geq \beta$, then since $|r + \sqrt{d}| \geq \sqrt{d}$, $|r^2 - d| = |r - \sqrt{d}||r + \sqrt{d}| \geq \beta \sqrt{d}$. Thus $|y_1^2 + y_2^2 + \ldots + y_n^2 - d| \geq \beta \sqrt{d}$ and hence $|x_1 + x_2 + \ldots + x_d| \geq \beta \sqrt{d}$ or $|w_1 + w_2 + \ldots + w_d| \geq \frac{3\sqrt{d}}{2}$. Applying Theorem 1.7 where $\sigma^2 = 2$ and $n = d$, this occurs with probability less than or equal to $3e^{-\beta^2 / 8}$.  

**Chernoff Bounds**

Chernoff bounds deal with sums of Bernoulli random variables. Here we apply Theorem 1.7 to derive similar bounds.

**Theorem 1.8** Suppose $y_1, y_2, \ldots, y_n$ are independent 0-1 random variables with $E(y_i) = p$ for all $i$. Let $y = y_1 + y_2 + \ldots + y_n$. Then for any $c \in [0, 1]$,

$$\text{Prob}(|y - E(y)| \geq cnp) \leq 3e^{-np^2 / 8}.$$  

**Proof:** Let $x_i = y_i - p$. Then, $E(x_i) = 0$ and $E(x_i^2) = E(y_i - p)^2 = p(1 - p)$. For $s \geq 3$,

$$|E(x_i^s)| = |E((y_i - p)^s)|$$

$$= |p(1-p)^s + (1-p)(0-p)^s|$$

$$= |p(1-p)((1-p)^{s-1} + p^{s-1})|$$

$$\leq p(1-p).$$

Apply Theorem 1.7 with $a = np$. Noting that $a < \sqrt{2} np$, completes the proof.  

The appendix contains a different proof that uses a standard method based on moment-generating functions, which gives a better constant in the exponent.
Power Law Distributions

The power law distribution of order $k$ where $k$ is a positive integer is

$$f(x) = \frac{k-1}{x^k} \quad \text{for} \quad x \geq 1.$$  

The power law is the hypothesized distribution in many practical settings. For example, if we plot the how many words occur with a certain frequency in a document (against frequency), the so-called Zipf’s law postulates that the plot obeys a power law.

If a random variable $x$ has this distribution for $k \geq 4$, then

$$\mu = E(x) = \frac{k-1}{k-2} \quad \text{and} \quad \text{Var}(x) = \frac{k-1}{(k-2)^2(k-3)}.$$  

**Theorem 1.9** Suppose $y$ obeys a power law of order $k \geq 4$ and $x_1, x_2, \ldots, x_n$ are independent random variables, each with the same distribution as $y - E(y)$. Let $x = x_1 + x_2 + \cdots + x_n$. For any nonnegative $a \leq \frac{1}{\sqrt{n}} \sqrt{\frac{k}{k-2}}$,  

$$\text{Prob}(|x| \geq a) \leq e^{-\frac{a^2}{8\text{Var}(x)}}.$$  

**Proof:** For integer $s$, the $s^{th}$ moment of $x_i$, namely, $E(x_i^s)$, exists if and only if $s \leq k-2$. For $s \leq k-2$,  

$$E(x_i^s) = (k-1) \int_{1}^{\infty} \frac{(y - \mu)^s}{y^k} dy$$  

Figure 1.12: Zipf’s Law: Number of words versus frequency.
Using the substitution of variable $z = \mu/y$

$$\frac{(y - \mu)^s}{y^k} = y^{s-k}(1 - z)^s = \frac{z^{k-s}}{\mu^{k-s}}(1 - z)^s$$

As $y$ goes from 1 to $\infty$, $z$ goes from $\mu$ to 0, and $dz = -\frac{\mu}{y^2}dy$. Thus

$$E(x_i^n) = (k-1) \int_1^\infty \frac{(y - \mu)^s}{y^k} dy$$

$$= \frac{k-1}{\mu^{k-s-1}} \int_0^1 (1-z)^s z^{k-s-2} dz + \frac{k-1}{\mu^{k-s-1}} \int_1^\mu (1-z)^s z^{k-s-2} dz.$$ 

The first integral is just the standard integral of the beta function and its value is $\frac{s!(k-2-s)!}{(k-1)!}$. To bound the second integral, note that for $z \in [1, \mu]$, $|z-1| \leq \frac{1}{\mu^{k-2}}$ and

$$z^{k-s-2} \leq (1 + (1/(k-2)))^{k-s-2} \leq e^{(k-s-2)/(k-2)} \leq e.$$ 

Apply Theorem 1.7 requires bounding $|E(x_i^n)|$ for $3 \leq s \leq \left\lfloor \frac{a^2}{4n \text{Var}(x_i)} \right\rfloor$. Since $a \leq \frac{1}{10} \sqrt{\frac{\pi}{k}}$, it follows that

$$\left\lfloor \frac{a^2}{4n \text{Var}(x_i)} \right\rfloor \leq \frac{n}{100k} \frac{(k-2)^2(k-3)}{k-1} \leq \frac{(k-2)^2(k-3)}{k(k-1)} \leq k-2.$$ 

So it suffices to prove that $|E(x_i^n)| \leq s! \text{Var}(x)$ for $3 \leq s \leq \ldots, k-2$. If $k = 4$, $s$ can go only up to 2 and there is nothing to prove. So assume $k \geq 5$. Since $\mu > 1$,

$$|E(x_i^n)| \leq \frac{(k-1)!}{(k-1)!} \frac{(k-2-s)!}{(k-2)!} + \frac{e(k-1)}{(k-2)^{s+1}} \leq s! \text{Var}(y) \left( \frac{1}{k-4} + \frac{e}{3!} \right) \leq s! \text{Var}(x).$$ 

Now, the theorem follows from Theorem 1.7.

\[\square\]

### 1.11 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 $p_1(x)$ and $p_2(x)$ where data is drawn from the mixture $p(x) = w_1p_1(x) + w_2p_2(x)$ with positive weights $w_1$ and $w_2$ summing to one. Assume that $p_1$ and $p_2$ are spherical with unit variance. If their means are very close, then given data from the mixture, one cannot tell for each data point whether it came from $p_1$ or $p_2$. The question arises as to how much separation is needed between the means to determine which Gaussian generated which data point. We will see that a separation of $\Omega(q^{1/4})$ suffices. The algorithm to separate two Gaussians is simple. Calculate the distance between all pairs of points. Points whose distance apart is smaller are from the same Gaussian, points whose distance is larger are from different Gaussians. Later, we will see that with more sophisticated algorithms, even a separation
of $\Omega(1)$ suffices.

Consider two spherical unit-variance Gaussians. From Theorem 1.7, most of the probability mass of each Gaussian lies on an annulus of width $O(1)$ at radius $\sqrt{d-1}$. Also $e^{-|x|^2/2} = \prod_i e^{-x_i^2/2}$ and almost all of the mass is within the slab $\{ x \mid -c \leq x_1 \leq c \},$ for $c \in O(1)$. Pick a point $x$ from the first Gaussian. After picking $x$, rotate the coordinate system to make the first axis point towards $x$. Independently pick a second point $y$ also from the first Gaussian. The fact that almost all of the mass of the Gaussian is within the slab $\{ x \mid -c \leq x_1 \leq c, c \in O(1) \}$ at the equator implies that $y$’s component along $x$’s direction is $O(1)$ with high probability. Thus, $y$ is nearly perpendicular to $x$. So, $|x - y| \approx \sqrt{|x|^2 + |y|^2}$. See Figure 1.13. More precisely, since the coordinate system has been rotated so that $x$ is at the North Pole, $x = (\sqrt{d} \pm O(1), 0, \ldots, 0)$. Since $y$ is almost on the equator, further rotate the coordinate system so that the component of $y$ that is perpendicular to the axis of the North Pole is in the second coordinate. Then $y = (O(1), \sqrt{d} \pm O(1), 0, \ldots, 0)$. Thus,

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

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

Given two spherical unit variance Gaussians with centers $p$ and $q$ separated by a distance $\delta$, the distance between a randomly chosen point $x$ from the first Gaussian and a randomly chosen point $y$ from the second is close to $\sqrt{\delta^2 + 2d}$, since $x - p, p - q$, and $q - y$ are nearly mutually perpendicular. Pick $x$ and rotate the coordinate system so that $x$ is at the North Pole. Let $z$ be the North Pole of the sphere approximating the second Gaussian. Now pick $y$. Most of the mass of the second Gaussian is within $O(1)$ of the equator perpendicular to $q - z$. Also, most of the mass of each Gaussian is within
distance $O(1)$ of the respective equators perpendicular to the line $q - p$. See Figure 1.14. Thus,

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

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 $d^{1/4}$. One can actually separate Gaussians where the centers are much closer. Chapter 4 contains 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. The remaining points come from the second Gaussian.

**Fitting a single spherical Gaussian to data**

Given a set of sample points, $x_1, x_2, \ldots, x_n$, 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 $\mu$ and variance $\sigma^2$ in each direction. The probability of picking these points when sampling according to $F$ is given by

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

where the normalizing constant $c$ is the reciprocal of $\int e^{-\frac{|x-\mu|^2}{2\sigma^2}} \, dx$. In integrating from $-\infty$ to $\infty$, one could shift the origin to $\mu$ and thus $c$ is $\int e^{-\frac{|x|^2}{2\sigma^2}} \, dx = \frac{1}{(2\pi)^{\frac{n}{2}}}$ and is independent of $\mu$.

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

**Lemma 1.10** Let $\{x_1, x_2, \ldots, x_n\}$ be a set of $n$ points in $d$-space. Then $(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2$ is minimized when $\mu$ is the centroid of the points $x_1, x_2, \ldots, x_n$, namely $\mu = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.

**Proof:** Setting the gradient of $(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2$ with respect $\mu$ to zero yields

$$-2(x_1 - \mu) - 2(x_2 - \mu) - \cdots - 2(x_n - \mu) = 0.$$

Solving for $\mu$ gives $\mu = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.

To determine the maximum likelihood estimate of $\sigma^2$ for $F$, set $\mu$ to the true centroid. Next, we show that $\sigma$ is set to the standard deviation of the sample. Substitute $\nu = \frac{1}{2\sigma^2}$ and $a = (x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2$ into the formula for the probability of picking the points $x_1, x_2, \ldots, x_n$. This gives

$$e^{-a\nu} \left( \int e^{-\nu x^2} \, dx \right)^n.$$

Now, $a$ is fixed and $\nu$ is to be determined. Taking logs, the expression to maximize is

$$-a\nu - n \ln \left( \int e^{-\nu x^2} \, dx \right).$$

To find the maximum, differentiate with respect to $\nu$, set the derivative to zero, and solve for $\sigma$. The derivative is

$$\frac{\int |x|^2 e^{-\nu x^2} \, dx}{\int e^{-\nu x^2} \, dx} - a + n \frac{\int x^2 e^{-\nu x^2} \, dx}{\int e^{-\nu x^2} \, dx}.$$
Setting \( y = |\sqrt{\nu}x| \) in the derivative, yields

\[-a + \frac{n}{\nu} \int \frac{y^2 e^{-y^2}}{y} dy.
\]

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 \(-a + \frac{nd}{\nu} \). Substituting \( \sigma^2 \) 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 1.11** 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 from the true mean.

Let \( x_1, x_2, \ldots, x_n \) be a sample of points generated by a Gaussian probability distribution. \( \mu = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \) 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 \( \mu = \frac{1}{n-1}(x_1 + x_2 + \cdots + x_n) \) when estimating the variance. See Section ?? of the appendix.

### 1.12 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 and Section ?? of the appendix. There is vast literature on the 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 [?].
1.13 Experiments

Exercise 1.1

1. 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 $E((x - y)^2)$?

2. Let $x$ and $y$ be independent random variables with uniform distribution in $[-\frac{1}{2}, \frac{1}{2}]$. What is the expected value $E(x)$, $E(x^2)$, $E(x - y)$, $E(xy)$, and $E((x - y)^2)$?

3. What is the expected squared distance between two points generated at random inside a unit d-dimensional cube centered at the origin?

4. Randomly generate a number of points inside a d-dimensional unit cube centered at the origin and plot distance between and the angle between the vectors from the origin to the points for all pairs of points.

Solution:

1. $E(x) = \int_0^1 x \, dx = \frac{1}{2} x^2 \big|_0^1 = \frac{1}{2}$,
   $E(x^2) = \int_0^1 x^2 \, dx = \frac{1}{3} x^3 \big|_0^1 = \frac{1}{3}$.
   Note $E(x^2) \neq E^2(x)$,
   $E(x - y) = E(x) - E(y) = 0$.
   Since $x$ and $y$ are independent $E(xy) = E(x)E(y) = \frac{1}{4}$,
   $E[(x - y)^2] = E(x^2) - 2E(xy) + E(y^2) = \frac{1}{3} - 2 \times \frac{1}{4} + \frac{1}{3} = \frac{1}{6}$.

2. $E(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x \, dx = \frac{1}{2} x^2 \big|_{-\frac{1}{2}}^{\frac{1}{2}} = 0$,
   $E(x^2) = \int_0^1 x^2 \, dx = \frac{1}{3} x^3 \big|_{-\frac{1}{2}}^{\frac{1}{2}} = \frac{1}{12}$.
   Note $E(x^2) \neq E^2(x)$,
   $E(x - y) = E(x) - E(y) = 0$.
   Since $x$ and $y$ are independent $E(xy) = E(x)E(y) = 0$,
   $E[(x - y)^2] = E(x^2) - 2E(xy) + E(y^2) = \frac{1}{12} - 2 \times 0 + \frac{1}{12} = \frac{1}{6}$.

3. $E[(x - y)^2] = E[\sum_{i=1}^{d} (x_i - y_i)^2] = dE[(x_1 - y_1)^2] = \frac{d}{6}$

4. The distance between two points should be close to the expected value of $\frac{d}{2}$. The vectors should be close to perpendicular since $E(xy) = 0$ and thus if $\cos(\theta) = 0$ $\theta = \frac{\pi}{2}$.

Exercise 1.2 Consider two random 0-1 vectors in high dimension. What is the angle between them?
Solution: With high probability each vector has approximately half 1’s and thus the length squared is approximately $\frac{n}{2}$ and the dot product is approximately $\frac{n}{4}$. The dot product of the two normalized vectors is approximately $\frac{1}{2}$ and corresponds to an angle of 30 deg.

Exercise 1.3 In Section 1.1 on properties of high-dimensional space, we state that the distance of a point to the center of a sphere in $d$-dimensions is likely to be between $1 - \frac{c}{\sqrt{d}}$ and 1. We also claim that the first coordinate of such a point is likely to be between $-\frac{c}{\sqrt{d}}$ and $\frac{c}{\sqrt{d}}$. Justify the role of $d$ in these statements. Why is the $d$ in the denominator linear in one case and in the other appears as a square root.

Solution:

1. Set $c = \frac{1}{2}$. Then half of the volume is in the sphere of radius $\frac{c}{\sqrt{d}}$ independent of $d$ and the point will with probability $\frac{1}{2}$ be within distance $1 - \frac{c}{\sqrt{d}}$. If we had not placed the $d$ in the denominator of the fraction $\frac{c}{\sqrt{d}}$, then the volume would have gone to zero.

2. As the dimension increases the expected magnitude for a coordinate so that the vector is of unit length is $\frac{1}{\sqrt{d}}$. Then $\sum_{i=1}^{d} \left( \frac{1}{\sqrt{d}} \right)^2 = 1$. Thus, the $\sqrt{d}$ in the denominator of $\frac{c}{\sqrt{d}}$. I would like to say something about the variance but once the vector is normalized the coordinates are no longer independent.

Exercise 1.4 Show that Markov’s inequality is tight by showing the following:

1. For each of $a = 2, 3,$ and 4 give a probability distribution for a nonnegative random variable $x$ where $\text{Prob} \left( x \geq aE(x) \right) = \frac{1}{a}$.

2. For arbitrary $a \geq 1$ give a probability distribution for a nonnegative random variable $x$ where $\text{Prob} \left( x \geq aE(x) \right) = \frac{1}{a}$.

Solution: For parts one and two, let $\text{Prob}(0) = 1 - \frac{1}{a}$ and $\text{Prob}(a) = \frac{1}{a}$.

Exercise 1.5 In what sense is Chebyshev’s inequality tight?

Solution: Let $\text{Prob}(0) = \frac{1}{2}$ and $\text{Prob}(2) = \frac{1}{2}$. Then Chebyshev is tight as shown in the figure below. Extend to more general probability as in previous exercise.
Exercise 1.6 Consider the probability function \( p(x) = c \frac{1}{x^4} \), \( x \geq 1 \), and generate 100 random samples. How close is the average of the samples to the expected value of \( x \)?

Solution:
\[
c \int_{x=1}^{\infty} \frac{1}{x^4} dx = - \frac{c}{3} x^{-3} \bigg|_{1}^{\infty} = \frac{c}{3}
\]
Thus \( c = 3 \).

\[
E(x) = 3 \int_{x=1}^{\infty} \frac{1}{x^3} dx = - \frac{3}{2} x^{-2} \bigg|_{1}^{\infty} = \frac{3}{2}
\]
The cumulative distribution is \( 3 \int_{1}^{x} x^{-4} dx = -x^{-3} \bigg|_{1}^{x} = 1 - x^{-3} \). Generate a random number \( z \) uniformly in \([0, 1]\) and solve \( z = 1 - x^{-3} \). \( x = \left( \frac{1}{1-z} \right)^{\frac{1}{3}} \)

```matlab
function m = homework(n);
%compute data for exercise 2.6
z = rand(1, n);
a = ((1./(1-z))). & (1/3));
m = sum(a)/n
mx = max(a)
end
```

The result for five runs averaging 100 values per run. 1.4514, 1.3874, 1.4742, 1.5311, 1.4110.

Exercise 1.7 Consider the portion of the surface area of a unit radius, 3-dimensional sphere with center at the origin that lies within a circular cone whose vertex is at the origin. What is the formula for the incremental unit of area when using polar coordinates to integrate the portion of the surface area of the sphere that is lying inside the circular cone? What is the formula for the integral? What is the value of the integral if the angle of the cone is \( 36^\circ \)? The angle of the cone is measured from the axis of the cone to a ray on the surface of the cone.

Solution: The incremental area is a circular anulus of width \( d\theta \) and radius \( \sin \theta \) centered on the axis of the cone. The incremental area is \( 2\pi \sin \theta d\theta \). The integral
\[
2\pi \int_{\theta=0}^{\frac{\pi}{10}} \sin \theta d\theta = -2\pi \cos \theta \bigg|_{0}^{\frac{\pi}{10}} = -2\pi (\cos \frac{\pi}{10} - 1) = 0.3075
\]

Exercise 1.8 For what value of \( d \) does the volume, \( V(d) \), of a \( d \)-dimensional unit sphere take on its maximum?

Hint: Consider the ratio \( \frac{V(d)}{V(d-1)} \).
**Solution:** The maximum is at $d=5$ where the volume is 5.26. Calculate the ratio $V(d)/V(d+1)$. The ratio can be expressed as $\frac{d+1}{\sqrt{\pi d}} \Gamma\left(\frac{d+1}{2}\right)$ and is monotonically increasing. At $d=5$ the ratio becomes greater than one. One may need to break the problem into two for even $d$ and odd $d$.

**Alternative solution**

The volume of a $d$-dimensional sphere is

$$V(d) = \frac{2\pi^{d/2}}{d\Gamma(d/2)}$$

For $d=4, \quad V(4) = \frac{2\pi^2}{6\Gamma(2)} = \frac{1}{2} \pi^2 = 4.93$

For $d=5, \quad V(4) = \frac{2\pi^2\sqrt{\pi}}{5\Gamma(2)} = \frac{8\pi^2}{15} = 5.26$ Note $\Gamma\left(\frac{5}{2}\right) = \frac{3}{2} \sqrt{\pi}$

For $d=6, \quad V(4) = \frac{2\pi^3}{6\Gamma(3)} = \frac{\pi^3}{6} = 5.16$

Thus, the maximum volume occurs for $d = 5$.

**Exercise 1.9** Write a recurrence relation for $V(d)$ in terms of $V(d-1)$ by integrating using an incremental unit that is a disk of thickness $dr$.

**Solution:**

$$V(d) = 2 \int_{r=0}^{1} (1 - r^2)^{d-1} V(d-1) dr = 2V(d-1) \int_{r=0}^{1} (1 - r^2)^{d-1} dr$$

Start with the standard integral

$$\int_{x=0}^{1} x^{a-1}(1-x)^{\beta-1} dx = \frac{\Gamma(a)\Gamma(\beta)}{\Gamma(a+\beta)}$$

Set $x = y^2$ and $dx = 2ydy$.

$$\int_{x=0}^{1} x^{a-1}(1-x)^{\beta-1} dx = 2 \int_{y=0}^{1} y^{2a-1}(1 - y^2)^{\beta-1} dy = \frac{\Gamma(a)\Gamma(\beta)}{\Gamma(a+\beta)}$$

Then setting $a = \frac{1}{2}$

$$2 \int_{y=0}^{1} (1 - y^2)^{\beta-1} dy = \frac{\Gamma(\frac{1}{2})\Gamma(\beta)}{\Gamma(\beta + \frac{1}{2})}$$

Setting $\beta = \frac{d+1}{2}$

$$2 \int_{y=0}^{1} (1 - y^2)^{\frac{d-1}{2}} dy = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2} + 1)}$$
Thus

\[ V(d) = 2 \frac{1}{2} \sqrt{\pi} \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2} + 1\right)} V(d - 1) \]

Check of recurrence

\[ V(d - 1) = \frac{2}{(d - 1) \Gamma\left(\frac{d-1}{2}\right)} \pi^{\frac{d-1}{2}} \]

Thus

\[ V(d) = \frac{\sqrt{\pi} \Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2} + 1\right)} \frac{2}{(d - 1) \Gamma\left(\frac{d-1}{2}\right)} \pi^{\frac{d}{2}} \]

\[ = \frac{2\pi^{\frac{d}{2}} \Gamma\left(\frac{d+1}{2}\right)}{(d - 1) \Gamma\left(\frac{d}{2} + 1\right) \Gamma\left(\frac{d-1}{2}\right)} \]

\[ = \frac{2\pi^{\frac{d}{2}} \Gamma\left(\frac{d}{2}\right)}{(d - 1) \Gamma\left(\frac{d}{2}\right) \Gamma\left(\frac{d-1}{2}\right)} \]

\[ = \frac{2\pi^{\frac{d}{2}}}{d \Gamma\left(\frac{d}{2}\right)} \]

**Exercise 1.10** 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?

**Solution:** The volume of a sphere of radius 2 is \(2^d V(d)\). For radius \(r\), the volume is \(r^d V(d)\). To determine the function of \(d\), assume \(d\) is even and use Sterling approximation for \(\Gamma\left(\frac{d}{2}\right)\). For even \(d\), \(\Gamma\left(\frac{d}{2}\right) = \left(\frac{d}{2} - 1\right)\). Thus, \(\frac{d}{2} \Gamma\left(\frac{d}{2}\right) = \frac{d!}{2}\). Sterling’s approximation for \(\frac{d!}{2}\) is \(\sqrt{2\pi d} \left(\frac{n}{e}\right)^n\). Thus, the volume is given by \(\frac{\pi^{\frac{d}{2}}}{\left(\frac{d}{2e}\right)^{\frac{d}{2}}} r^d\). Thus, \(r\) must equal some constant times \(\sqrt{d}\).

**Exercise 1.11** A 3-dimensional cube has vertices, edges, and faces. In a \(d\)-dimensional cube, these components are called faces. A vertex is a 0-dimensional face, an edge a 1-dimensional face, etc. For \(0 \leq i \leq d\), how many \(i\)-dimensional faces does a \(d\)-dimensional hyper cube have? What is the total number of faces of all dimensions? The \(d\)-dimensional face is the cube itself which you can include in your count.

**Solution:** A \(d\)-dimensional hyper cube has \(2^d\) vertices with \(d\) edges incident to each vertex. The number of \(i\)-dimensional faces can be computed as follows. Select a vertex
and $i$ edges incident to that vertex. This defines an $i$-dimensional face. There are $2^d \binom{d}{i}$ ways to do this. However, since an $i$-dimensional face has $2^i$ vertices, we will count each $i$-dimensional face $2^i$ times. Thus, there are $2^{d-i} \binom{d}{i}$ $i$-dimensional faces in a $d$-dimensional hyper cube. The sum of all the faces is

$$
\sum_{k=0}^{d} 2^{d-k} \binom{d}{k} = 2^d \sum_{k=0}^{d} \frac{1}{2^k} \binom{d}{k}
$$

Thus, the sum is $2^d$ times $(1 + x)^d$ with $\frac{1}{2}$ substituted for $x$. This is $2^d \left(\frac{3}{2}\right)^d = 3^d$. The $d$-dimensional face is the cube itself so we should subtract one if we wanted to exclude it.

**Exercise 1.12** For $i \leq i \leq d$, how many $i$-dimensional faces does a $d$-dimensional tetrahedron have?

**Solution:** A $d$-dimensional tetrahedron has $d+1$ vertices or 0-dimensional faces. Select a vertex and $i$ edges incident to the vertex. This defines an $i$ dimensional face. There are $(d+1)\binom{d}{i}$ ways to do this? However, since an $i$-dimensional face has $i+1$ vertices, we will count each face $i+1$ times. Thus, there are $\frac{d+1}{i+1} \binom{d}{i}$ $i$-dimensional faces in a $d$-dimensional tetrahedron.

**Exercise 1.13** Consider a unit radius, circular cylinder in 3-dimensions of height one. The top of the cylinder could be an horizontal plane or half of a circular sphere. Consider these two possibilities for a unit radius, circular cylinder in 4-dimensions. In 4-dimensions the horizontal plane is 3-dimensional and the half circular sphere is 4-dimensional. In each of the two cases, what is the surface area of the top face of the cylinder? You can use $V(d)$ for the volume of a unit radius, $d$-dimension sphere and $A(d)$ for the surface area of a unit radius, $d$-dimensional sphere. An infinite length, unit radius, circular cylinder in 4-dimensions would be the set \( \{(x_1, x_2, x_3, x_4)|x_2^2 + x_3^2 + x_4^2 \leq 1\} \) where the coordinate $x_1$ is the axis.

**Solution:** For the horizontal plane case the surface area is the volume of a 3-dimensional sphere or $V(3)$. For the half sphere case the surface area is the surface area of a 4-dimensional sphere or $\frac{1}{2}A(4)$.

**Exercise 1.14** What is the surface area of a $d$-dimensional cylinder of radius two and height one in terms of $V(d)$ and $A(d)$?

**Solution:** The top and bottom of the cylinder are $(d-1)$-dimensional spheres of radius two. Thus their surface area is $2V(d-1)2^{d-1}$. The remaining surface area is $A(d-1)2^{d-2}$.

**Exercise 1.15** 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.
Solution: Instead of a sphere in $d$-dimensions consider generating points at random where each coordinate of a point is independent and distributed according to a unit variance Gaussian. The set of points generated will lie in a thin annulus of width $O(1)$ at radius $\sqrt{d}$. We will use this annulus as a good approximation to a sphere of radius $\sqrt{d}$. If we now project this annulus onto the first coordinate axis, we get a unit variance Gaussian. Thus, the surface area of a $d$-dimensional sphere of radius $\sqrt{d}$ projected onto a line through the origin gives rise to a unit variance Gaussian. Thus, almost all the surface area of a radius $\sqrt{d}$ sphere in $d$-dimensions lies in a band of width three about the equator.

Exercise 1.16 Place two unit-radius spheres in $d$-dimensions, one at $(-2,0,0,\ldots,0)$ and the other at $(2,0,0,\ldots,0)$. Give an upper bound on the probability that a random line through the origin will intersect the spheres.

Solution: One could estimate by a computer program that randomly generated line and tested if the line intersected the spheres. To generate a random line, generate a random vector $(a_1, a_2, \ldots, a_d)$ and let the line be the set of points $(x_1, \frac{a_2}{a_1}x_1, \frac{a_3}{a_1}x_1, \ldots, \frac{a_d}{a_1}x_1)$. Ask if this set of points intersects the sphere $(x_1 - 2)^2 + x_2^2 + \cdots + x_d^2 = 1$. If the components of the vector are selected uniformly at random over the range $[0,1]$, the direction of the line will not be uniformly distributed. One needs to generate the components as Gaussians to get a spherical distribution. For many applications, selecting the coordinates of the vector uniformly from $[0,1]$ may not matter. In fact we might be able to use

$$a_i = \begin{cases} 1 & \frac{1}{3} \\ 0 & \frac{1}{2} \\ 1 & \frac{1}{4} \end{cases}$$

Asking if the line intersects the sphere is equivalent to asking if there exists an $x_1$ such that

$$(x_1 - 2)^2 + \left(\frac{a_2}{a_1}x_1\right)^2 + \left(\frac{a_3}{a_1}x_1\right)^2 + \cdots + \left(\frac{a_d}{a_1}x_1\right)^2 = 1$$

This simplifies to

$$(a_1^2 + a_2^2 + \cdots + a_d^2) x_1^2 - 4a_1^2 x_1 + 3a_1^2 = 0$$

Applying the $b^2 - 4ac$ test for the existence of a root of the above quadratic equation, there is a solution provided $16a_1^2 \geq 12$ or $a_1^2 \geq \frac{3}{4}$. Use the fact that most of the mass is concentrated about the equator to derive a bound on the probability of this. HOW?

As a check in two dimensions this requires that $a_1 \geq \sqrt{3}a_2$. We see from the picture below that this is correct.

At tangency $a_1^2 + a_2^2 = 3$ and $(2 - a_1)^2 + a_2^2 = 1$ or $a_1 = \frac{3}{2}$ and $a_2 = \frac{\sqrt{5}}{2}$. That is $a_1 = \sqrt{3}a_2$.

Exercise 1.17 Let $x$ be a random sample from the unit sphere $\{x \mid |x| \leq 1\}$ in $d$-dimensions with the origin as center.
1. What is the mean of the random variable $x$? The mean, denoted $E(x)$, is the vector, whose $i^{th}$ component is $E(x_i)$.

2. What is the component-wise variance of $x$?

3. For any unit length vector $u$, the variance of the real-valued random variable $u^T x$ is $\sum_{i=1}^{d} u_i^2 E(x_i^2)$. Note that the $x_i$ are not independent. Using (2), simplify this expression for the variance of $x$.

4. * Given two spheres in $d$-space, both of radius one whose centers are distance $s$ apart, show that the volume of their intersection is at most

   \[
   4e^{-\frac{s^2(d-1)}{8}} \frac{s}{s\sqrt{d-1}}
   \]

   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 achieve 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.

6. * In this part, you will carry out the same exercise for Gaussians. First, restate the shared mass of two spheres as $\int_{x \in \text{space}} \min(f(x), g(x)) dx$, where $f$ and $g$ are just the uniform densities in the two spheres respectively. Make a similar definition for the shared mass of two spherical Gaussians. Using this, show that for two spherical Gaussians, each with standard deviation $\sigma$ in every direction and with centers at distance $s$ apart, the shared mass is at most $(c_1/s) \exp(-c_2 s^2/\sigma^2)$, where $c_1$ and $c_2$ are constants. This translates to “if two spherical Gaussians have centers which are $\Omega(\sigma)$ apart, then they share very little mass”. Explain.

Solution:

1. $E(x_i) = 0$ for all $i$, so $E(x) = 0$.

2. $\text{Var}(x_i) = E(x_i^2) = \frac{1}{d} E(|x|^2)$ 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, $\text{Var}(x_i) = \frac{1}{d+2}$.

3. The proof is by induction on $d$. It is clear for $d = 1$.

$$\text{Var}(\sum_i u_i x_i) = E((\sum_i u_i x_i)^2),$$  

since the mean is 0. Now,

$$E((\sum_i u_i x_i)^2) = E(\sum_i u_i^2 x_i^2) + 2E(\sum_{i \neq j} u_i u_j x_i x_j)$$  

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

$$E(\sum_{i \neq j} u_i u_j x_i x_j) = E \sum_{i \geq 2} u_1 x_1 u_i x_i + E \sum_{i \neq j; i,j \geq 2} u_i u_j x_i x_j$$

$$= E_{x_1} \left( u_1 x_1 E_{x_2,x_3,...,x_d} \left( \sum_{i \geq 2} u_i x_i | x_1 \right) \right) + E_{x_1} \left( E_{x_2,x_3,...,x_d} \left( \sum_{i \neq j; i,j \geq 2} u_i x_i u_j x_j | x_1 \right) \right)$$

[Notation: $E(Y|x_1)$is some function $f$ of $x_1$; it is really short-hand for writing $f(a) = E(Y|x_1 = a)$.]  

Now, for every fixed value of $x_1$, $E(x_i|x_1) = 0$ for $i \geq 2$, so the first term is zero. Since a section of the sphere is just a $d$-1sphere, the second term is zero by induction on $d$.

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

5. Simple.

6. If a spherical Gaussian has standard deviation $\sigma$ 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 $\sigma$ (as we show in Chapter 4 (or 5) ??). Let $a>0$ and let the centers be $0$ and $(a,0,0,...,0)$ 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

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

We bound this by using

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

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.

---

**Exercise 1.18** Prove that $1 + x \leq e^x$ for all real $x$. For what values of $x$ is the approximation $1 + x \approx e^x$ good?

**Solution:** Let $f(x) = e^x - x - 1$. Then $f'(x) = e^x - 1$ and $f''(x) = e^x \geq 0$. Thus, $f(x)$ is convex and has a unique minimum at $x = 0$ where $f'(x) = 0$. Thus, $f(x)$ is always nonnegative and $1 + x \leq e^x$. The approximation is good close to $x = 0$ and until $\frac{1}{2}x^2$ becomes significant.

---

**Exercise 1.19** Derive an upper bound on $\int_{x=a}^{\infty} e^{-x^2/2\sigma} dx$ where $a$ is a positive real. Discuss for what values of $a$ this is a good bound.

**Hint:** Use $e^{-x^2/2} \leq \frac{2}{a} e^{-\frac{a^2}{2}}$ for $x \geq a$.

**Solution:**

$$
\int_{x=a}^{\infty} e^{-x^2/2\sigma} dx \leq \frac{1}{a} \int_{x=a}^{\infty} xe^{-x^2/2\sigma} dx = \left. -\frac{1}{a} e^{-x^2/2} \right|_{a}^{\infty} = \frac{1}{a} e^{-\frac{a^2}{2\sigma}}
$$

<table>
<thead>
<tr>
<th>$a$</th>
<th>$1 - \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2\sigma}}$</th>
<th>$\frac{1}{\sqrt{2\pi}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7588</td>
<td>0.8413</td>
</tr>
<tr>
<td>2</td>
<td>0.9730</td>
<td>0.9772</td>
</tr>
<tr>
<td>3</td>
<td>0.9985</td>
<td>0.9987</td>
</tr>
</tbody>
</table>
Exercise 1.20 Verify the formula \( V(d) = 2 \int_0^1 V(d-1)(1-x_1^2)^{d-1} dx_1 \) for \( d = 1 \) and \( d = 2 \) by integrating and comparing with \( V(2) = \pi \) and \( V(3) = \frac{4}{3} \pi \).

Solution:

\[
V(3) = 2V(2) \int_0^1 \sqrt{1-x_1^2} dx_1 = 2\pi \left(x - \frac{x^3}{3}\right) = \frac{4}{3} \pi \\
V(2) = 2V(1) \int_0^1 \sqrt{1-x_1^2} dx_1 = 4 \int_0^\pi \cos^2 \theta d\theta = 4 \int_0^\pi \frac{1}{2}(1 + \cos 2\theta) d\theta \\
= 2 \int_0^\pi (1 + \cos 2\theta) d\theta = 2 \left(\theta + \frac{\sin 2\theta}{2}\right)_0^\pi = \frac{\pi}{2} = \pi
\]

Exercise 1.21 What is the volume of a radius \( r \) cylinder of height \( h \) in \( d \)-dimensions?

Solution: \( V(d-1)hr^{d-1} \). ■

Exercise 1.22 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.

Solution: The volume of the circular cylinder is \( hr^{d-1}V(d-1) \) where \( h = \sqrt{1-r^2} \). Since \( V(d-1) \) is a constant we can set the derivative of \( \sqrt{1-r^2}r^{d-1} \) to 0 and solve for \( r \).

\[
\frac{\partial}{\partial r} \sqrt{1-r^2}r^{d-1} = -(1-r^2)^{-\frac{1}{2}}r^d + (1-r^2)^{\frac{1}{2}}(d-1)r^{d-2} = 0
\]

Multiplying by \( \sqrt{1-r^2} \) and dividing by \( r^{d-2} \) gives

\[-r^2 + (1-r^2)(d-1) = 0\]

or

\[r^2 = \frac{d-1}{d}\]

Thus, \( h = \sqrt{1 - \frac{d-1}{d^2}} = \frac{1}{\sqrt{d}} \). ■

Exercise 1.23 What is the volume of the maximum size \( d \)-dimensional hypercube that can be placed entirely inside a unit radius \( d \)-dimensional sphere?

Solution: Let \( 2a \) be the length of a side of the cube. Then \( a^2d = 1 \), \( a = \frac{1}{\sqrt{d}} \), and the volume is \( \left(\frac{2}{\sqrt{d}}\right)^d \). ■
Exercise 1.24 In showing that the volume of a unit sphere was near the equator we obtained an upper bound on the volume of the upper hemisphere above the slice of
\[
\frac{1}{\epsilon(d-1)} e^{\frac{d-1}{2} \epsilon^2} V(d-1)
\]
and a lower bound on the volume of the upper hemisphere of \(\frac{1}{2\sqrt{d-1}} V(d-1)\). Show that for a radius \(r\) sphere these bounds become
\[
\frac{r^{d+1}}{\epsilon(d-1)} e^{\frac{d-1}{2} \epsilon^2} V(d-1) \quad \text{and} \quad \frac{r^d}{2\sqrt{d-1}} V(d-1)
\]
and that the ratio is \(\frac{2r}{\epsilon \sqrt{d-1}} e^{-\frac{d-1}{2} \epsilon^2}\).

Solution: The upper bound on the volume above the slice is
\[
\int_{\epsilon}^{r} (r^2 - x_1^2)^{\frac{d-1}{2}} V(d-1)dx_1 = V(d-1)r^{d-1} \int_{\epsilon}^{r} \left( 1 - \left( \frac{x_1}{r} \right)^2 \right)^{\frac{d-1}{2}}
\]
\[
\leq V(d-1)r^{d-1} \int_{\epsilon}^{\sqrt{r^2 - \epsilon^2}} \frac{x_1}{\epsilon} e^{-\frac{d-1}{2} \frac{r^2}{\epsilon^2}} dx_1
\]
\[
= V(d-1)d^{d-1} \left( -\frac{r^2}{\epsilon} (d-1) \right) e^{-\frac{d-1}{2} \frac{r^2}{\epsilon^2}} \bigg|_{\epsilon}^{\sqrt{r^2 - \epsilon^2}}
\]
\[
= V(d-1) \frac{r^{d-1}}{\epsilon(d-1)} e^{-\frac{d-1}{2} \epsilon^2}.
\]
The lower bound on the volume of the upper hemisphere is the volume of a disk of radius \(\sqrt{r^2 - \frac{\epsilon^2}{d-1}}\) and height \(\frac{r}{\sqrt{d-1}}\), which is
\[
V(d-1)r^{d-1} \left( 1 - \frac{1}{d-1} \right)^{\frac{d-1}{2}} \frac{r}{\sqrt{d-1}} \geq V(d-1)r^d \left( 1 - \frac{1}{d-1} \frac{d-1}{2} \right) \frac{1}{\sqrt{d-1}}
\]
\[
= V(d-1) \frac{r^d}{2\sqrt{d-1}}
\]
Thus, the ratio is \(\frac{2r}{\epsilon \sqrt{d-1}} e^{-\frac{d-1}{2} \epsilon^2}\).

Exercise 1.25 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 \(x_1 = 0.01\)?

Solution:

Exercise 1.26 Let \(\{ x \mid |x| \leq 1 \}\) be a \(d\)-dimensional, unit radius sphere centered at the origin. What fraction of the volume is the set \(\{(x_1, x_2, \ldots, x_d) \mid \forall i \mid x_i \leq \frac{1}{\sqrt{d}}\}\)?
Solution: The volume of the cube is \( \left( \frac{2}{\sqrt{d}} \right)^d \) and \( V(d) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \). Thus, the ratio is
\[
\frac{2^d}{d^\frac{d}{2}} \frac{d\Gamma(\frac{d}{2})}{2\pi^\frac{d}{2}}
\]

Exercise 1.27 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 location of the North Pole giving rise to different equators.

Solution: Select \( x_1 = 1, x_2 = 1, \) and \( x_3 = 1 \) as positions for the North Pole. Then the equators are \( x_1 = 0, x_2 = 0, \) and \( x_3 = 0. \) If \( |x| \leq 1, \) then with high probability \( x_1, x_2, \) and \( x_3 \) are all close to \( \frac{1}{\sqrt{d}}. \)

Also notice that no matter what point one picks for the North Pole, the mathematics will give the same answer since the math is the same.

Exercise 1.28 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.

Solution: If \( |x| \leq 1, \) then with high probability \( |x| \geq 1 - \epsilon \) and \( x_1 \leq \frac{1}{\sqrt{d}}. \)

Exercise 1.29 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.

Solution: The projection can be thought of as taking the dot product with the unit vector \( (1/\sqrt{d}, 1/\sqrt{d}, \ldots, 1/\sqrt{d}) \) along the line. By the Central Limit Theorem, this dot product minus its mean should behave as the Gaussian. The variance of \( \sum_{i=1}^{d} (x_i - \frac{1}{2}) \frac{1}{\sqrt{d}} \) (for independent 0-1 variables \( x_i \)) is the sum of the variances which is clearly \( O(1) \).

Exercise 1.30

1. A unit cube has vertices, edges, faces, etc. How many k-dimensional objects are in a d-dimensional cube?

2. What is the surface area of a unit cube in d-dimensions?

3. What is the surface area of the cube if the length of each side was 2?
4. Prove that the volume of a unit cube is close to its surface.

Solution:

1. At each vertex select \( k \) edges to specify the \( k \)-dimensional object. Each \( k \)-dimensional object will be specified at \( 2^k \) vertices so divide by \( 2^k \). Thus, the answer is \( 2^{d-k} \binom{d}{k} \).

2. There are \( 2d, d - 1 \)-dimensional faces each of area one. Thus, the surface area is \( 2d \).

3. The surface area of each face increases by \( 2^{d-1} \) so multiply \( 2d \) by \( 2^{d-1} \) to get \( 2^d \).

4. The volume of a unit cube is close to the surface since a cube of dimension \( 1 - \epsilon \) has volume \((1 - \epsilon)^d\), which goes to zero as \( d \to \infty \).

Exercise 1.31 Define the equator of a \( d \)-dimensional unit cube to be the hyperplane \( \left\{ x \mid \sum_{i=1}^{d} x_i = \frac{d}{2} \right\} \).

1. Are the vertices of a unit cube concentrated close to the equator?

2. Is the volume of a unit cube concentrated close to the equator?

3. Is the surface area of a unit cube concentrated close to the equator?

Solution:

1. A vertex is a point \( x = (x_1, x_2, \ldots, x_d) \) where \( x_i \in \{0, 1\} \).

\[
\begin{align*}
E(x_i) &= \frac{1}{2} \\
\text{var}(x_i) &= E\left(x_i - \frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{4}
\end{align*}
\]

By Chebyshev’s inequality

\[
\text{Prob}\left( |x_1 + x_2 + \cdots + x_d - \frac{d}{2}| > \epsilon \right) \leq \frac{\text{var}(x_i)}{\epsilon^2} = \frac{d}{4\epsilon^2}
\]

For \( \epsilon > \sqrt{d} \) the probability that the vertex is more than \( \epsilon \) from the equator goes to zero. Hence for large \( d \) a vertex is almost surely close to the equator.

2. Let \( (x_1, x_2, \ldots, x_d) \). The \( E(x_i) = \frac{1}{2} \). Thus \( E(x_1, x_2, \ldots, x_d) = \frac{d}{2} \). The equator is \( \{x| x_1 + x_2 + \cdots + x_d = \frac{d}{2}\} \). Given a random point \( x \) in a cube,

\[
\text{Prob}\left( |x_1 + x_2 + \cdots + x_d - \frac{d}{2}| \geq \epsilon \right) \leq \frac{\text{var}(x_i)}{d\epsilon^2}
\]

\[
\text{var}(x_i) = E\left(x_i - \frac{1}{2}\right)^2 = E\left(x_i^2 - x_i + \frac{1}{4}\right) = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{4 - 6 + 3}{12} = \frac{1}{12}
\]
\[
\text{Prob}\left( |x_1 + x_2 + \cdots + x_d - \frac{d}{2}| \geq \epsilon \right) \leq \frac{1}{12d\epsilon^2}
\]

If \( \epsilon > \sqrt{d} \) the probability goes to zero as \( d \) goes to infinity. Thus for large \( d \) a point is within \( \sqrt{d} \) of equator with probability one and thus the volume is concentrated within \( \sqrt{d} \) of the equator.

3. The \( d \)-dimensional cube has \( 2d \) \((d-1)\)-dimensional faces. One such face is the set \( \{ x | x_1 = 0, x_i \in [0,1], 2 \leq i \leq d \} \). Select a point \( x \) uniformly at random on this face. By the above argument
\[
\text{Prob}\left( |0 + x_2 + x_3 + \cdots + x_d - \frac{d-1}{2}| \geq \epsilon \right) \geq \epsilon
\]
foes to zero as \( d \) goes to infinite provided \( \epsilon > \sqrt{d} \). This ays a point on the face is close to a plane parallel to the equator but distance \( \frac{1}{2} \) away. A similar argument applies to points on the face \( \{ x | x = (1, x_2, \ldots, x_d) \} \) and all \( 2d \) faces.

**Exercise 1.32** How large must \( \epsilon \) be for 99% of the volume of a \( d \)-dimensional unit-radius sphere to lie in the shell of \( \epsilon \)-thickness at the surface of the sphere?

**Solution:**

**Exercise 1.33** Calculate the ratio of area above the plane \( x_1 = \epsilon \) of a unit radius sphere in \( d \)-dimensions for \( \epsilon = 0.01, 0.02, 0.03, 0.04, 0.05 \) and for \( d = 100 \) and \( d = 1,000 \). Also calculate the ratio for \( \epsilon = 0.001 \) and \( d = 1,000 \).

**Solution:**

<table>
<thead>
<tr>
<th>( \epsilon )</th>
<th>( d )</th>
<th>100</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.366</td>
<td>4.3 \times 10^{-5}</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>0.133</td>
<td>1.7 \times 10^{-9}</td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>0.048</td>
<td>5.9 \times 10^{-14}</td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td>0.017</td>
<td>1.8 \times 10^{-18}</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.006</td>
<td>5.2 \times 10^{-24}</td>
<td></td>
</tr>
</tbody>
</table>

For \( \epsilon = 0.001 \) and \( d = 1,000 \) the ratio is \( 0.367 \)

**Exercise 1.34**

1. What is the maximum size rectangle that can be fitted in a unit variance Gaussian?

2. What rectangle best approximates a unit variance Gaussian if one measure goodness of fit by how small the symmetric difference of the Gaussian and rectangle is.

**Solution:**
Exercise 1.35  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? How wide do the bands need to be for all points to be in all five bands?

Solution:

Exercise 1.36  We have claimed that a randomly generated point on a sphere lies near the equator of the sphere, wherever we place the North Pole. Is the same claim true for a randomly generated point on a cube? To test this claim, randomly generate ten $\pm 1$ valued vectors in 128 dimensions. Think of these ten vectors as ten choices for the North Pole. Then generate some additional $\pm 1$ 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 vector close to?

Solution:

Exercise 1.37  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?

Solution: For a vector to be within a 45° degree cone whose axis is at the North pole, it must be within 45° degrees of the North pole in each dimension. That is it must lie within a “square” cone containing the circular cone. The probability that in a given dimension a randomly choose vector is within 45° degrees of North is 1/2. Thus, the probability goes to zero as the dimension increases. Thus, with probability one, two randomly chosen vectors are at least 45° degree apart in direction and their dot product, assuming unit length vectors, is at most $\frac{\sqrt{2}}{2}$. This says that two vectors selected at random are close to orthogonal and are distance $\sqrt{2}$ apart.

Exercise 1.38  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 Pole to the 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?
Solution:

Exercise 1.39 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 the probability that none of the $n$ points lands between them be at least .99?

Exercise 1.40 Project the surface area of a $d$-dimensions sphere of radius $\sqrt{d}$ onto a line through the center. For large $d$, give an intuitive argument that the projected surface area should behave like a Gaussian.

Exercise 1.41 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 $S$.

Solution:

Exercise 1.42 How would you sample uniformly at random from the parallelepiped

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

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.

Solution:

Exercise 1.43 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.

Solution:

Exercise 1.44

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?

Solution:

**Exercise 1.45** 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 \(\sqrt{d}\).

1. What is the distribution when the points are projected onto a random line through the origin?

2. If one uses a Gaussian with variance four, where in \(d\)-space will the points lie?

Solution:

1. Since the distribution is spherically symmetrical, rotate the coordinate system so that the line is aligned with the first coordinate axis. The distribution of the projection is the distribution of the first coordinate and hence is a unit variance Gaussian.

2. \(E(x_1^2 + x_2^2 \cdots x_d^2) = dE(x_1^2) = 4d\). Thus, the points will approximately lie on the surface of a sphere of radius \(2\sqrt{d}\).

**Exercise 1.46** 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.

Solution:

**Exercise 1.47** 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. Show that as the variance of the Gaussian goes down, more and more of its mass is contained in the unit-radius sphere. How small must the variance be for 0.99 of the mass of the Gaussian to be contained in the unit-radius sphere?

Solution:

**Exercise 1.48** Consider two unit-radius spheres in \(d\)-dimensions whose centers are distance \(\delta\) apart where \(\delta < 1\) is a constant independent of \(d\). Let \(x\) be a random point on the surface of the first sphere and \(y\) a random point on the surface of the second sphere. Prove that as \(d\) goes to infinity, the probability that \(|x - y|^2\) is more than \(2 + \delta^2 + s\), falls off exponentially with \(s\).
Solution:

Exercise 1.49 Pick a point $x$ uniformly at random from the following set in $d$-space:

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

1. Show that the probability that $x_1^4 + x_2^4 + \cdots + x_d^4 \leq \frac{1}{2}$ is $\frac{1}{2^{d/2}}$.

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, $|x_1| \leq O(1/d^{1/4})$.

Solution: Let $U(d) = \text{Volume}K$ which is the solid $K$ defined earlier. We have

$$\text{Volume}(K \cap \{x_1 \geq a\}) = \int_a^1 \text{Volume}\left((x_2, x_3, \ldots, x_d) | x_2^4 + x_3^4 + \cdots + x_d^4 \leq 1 - t^4\right) \, dt$$

$$= \int_a^1 U(d-1)(1-t^4)^{(d-1)/4} \, dt$$

$$\leq U(d-1) \int_a^1 \frac{t^3}{d^3}(1-t^4)^{(d-1)/4} \ldots \text{integrable}$$

Also

$$\int_{t=0}^1 U(d-1)(1-t^4)^{(d-1)/4} \, dt \geq U(d-1) \int_{t=0}^{1/d^{1/4}} (1-t^4)^{(d-1)/2} \geq U(d-1)e^{-1/4}.$$  

Compare to get ratio. So, conclude that

$$\text{Prob}( |x_1| \geq \frac{c}{d^{1/4}} ) \leq \frac{1}{e^3}e^{-c^4/4}.$$  

Exercise 1.50 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?

Solution:

Exercise 1.51 Let $x_1, x_2, \ldots, x_n$ be independent samples of a random variable $x$ with mean $m$ and variance $\sigma^2$. Let $m_s = \frac{1}{n} \sum_{i=1}^n x_i$ be the sample mean. Suppose one estimates the variance using the sample mean rather than the true mean, that is,

$$\sigma_s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m_s)^2$$

Prove that $E(\sigma_s^2) = \frac{n-1}{n} \sigma^2$ and thus one should have divided by $n - 1$ rather than $n$.

Hint: First calculate the variance of the sample mean and show that $\text{var}(m_s) = \frac{1}{n} \text{var}(x)$. Then calculate $E(\sigma_s^2) = E[\frac{1}{n} \sum_{i=1}^n (x_i - m_s)^2]$ by replacing $x_i - m_s$ with $(x_i - m) - (m_s - m)$.  

45
Solution:

1. The variance of the sample mean is

\[
\text{var}(m_s) = \text{var}\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) = \frac{1}{n^2} \text{var}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n} \text{var}(x).
\]

2. 

\[
E(\sigma_s^2) = E\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - m_s)^2\right]
\]
\[
= E\left[\frac{1}{n} \sum_{i=1}^{n} \left((x_i - m) - (m_s - m)\right)^2\right]
\]
\[
= E\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2 - 2(m_s - m) \sum_{i=1}^{n} (x_i - m) + (m_s - m)^2\right]
\]
\[
= E\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2 - (m_s - m)^2\right]
\]
\[
= \sigma^2 - E(m_s - m)^2
\]
\[
= \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \frac{n-1}{n}
\]

Exercise 1.52 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?

Solution:

Exercise 1.53 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 \(\mu\), so that with probability at least 99/100,

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

How many samples are sufficient to ensure that

\[
|\mu - \tilde{\mu}| \leq \varepsilon?
\]

Solution:
Exercise 1.54 Use the probability distribution \( \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2} (x-5)^2} \) to generate ten points.

(a) From the ten points estimate \( \mu \). How close is the estimate of \( \mu \) to the true mean of 5?

(b) Using the true mean of 5, estimate \( \sigma^2 \) by the formula \( \sigma^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - 5)^2 \). How close is the estimate of \( \sigma^2 \) to the true variance of 9?

(c) Using your estimate of the mean, estimate \( \sigma^2 \) by the formula \( \sigma^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - 5)^2 \). How close is the estimate of \( \sigma^2 \) to the true variance of 9?

(d) Using your estimate of the mean, estimate \( \sigma^2 \) by the formula \( \sigma^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - 5)^2 \). How close is the estimate of \( \sigma^2 \) to the true variance of 9?

Exercise 1.55 The Cauchy distribution in one dimension is \( \text{Prob}(x) = \frac{1}{\pi} \frac{1}{1+x^2} \). What would happen if one tried to extend the distribution to higher dimensions by the formula \( \text{Prob}(r) = c \frac{1}{1+r^2} \), where \( r \) is the distance from the origin? What happens when you try to determine a normalization constant \( c \)?

Solution:

Exercise 1.56 Consider the power law probability density

\[
p(x) = \frac{c}{\max(1, x^2)} = \begin{cases} 
    c & 0 \leq x \leq 1 \\
    \frac{c}{x^2} & x > 1
\end{cases}
\]

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 \( E(x^2) \)?

Solution:

1. \( \int_0^\infty \frac{1}{\max(1,x^2)} = [1 + \int_1^\infty \frac{1}{x^2} \, dx] = [1 - \frac{1}{x}]_1^\infty = 2 \) implying \( c = \frac{1}{2} \).

2.

Exercise 1.57 Consider \( d \)-space and the following density over the positive orthant:

\[
p(x) = \frac{c}{\max(1, |x|^a)}.
\]

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

Exercise 1.58 Assume you can generate a value uniformly at random in the interval [0, 1]. How would you generate a value according to a probability distribution p(x)?

Solution:

Exercise 1.59 Let x be a random variable with probability density \( \frac{1}{4} \) for \( 0 \leq x \leq 4 \) and zero elsewhere.

1. Use Markov’s inequality to bound the probability that \( x > 3 \).
2. Make use of \( \text{Prob}(|x| > a) = \text{Prob}(x^2 > a^2) \) to get a tighter bound.
3. What is the bound using \( \text{Prob}(|x| > a) = \text{Prob}(x^r > a^r) \)?

Solution:

1. \( E(x) = 2 \). Thus, by Markov’s inequality \( \text{Prob}(|x| > 3) \leq \frac{E(x)}{3} < \frac{2}{3} \).
2. \( E(x^2) = \frac{1}{4} \int_0^4 x^2 \, dx = \frac{1}{4} \frac{3}{5} x^3 |_0^4 = \frac{16}{5} \). Thus, \( \text{Prob}(|x| > 3) = \text{Prob}(x^2 > 9) \leq \frac{E(x^2)}{9} = \frac{16}{27} \approx 0.59 \).
3. \( E(x^r) = \frac{4^r}{r+1} \). Thus, \( \text{Prob}(|x| > 3) \leq \frac{1}{r+1} \left( \frac{4}{3} \right)^r \).

Exercise 1.60 Consider the probability distribution \( p(x=0) = 1 - \frac{1}{a} \) and \( p(x=a) = \frac{1}{a} \). Plot the probability that \( x \) is greater than or equal to \( b \) as a function of \( b \) for the bound given by Markov’s inequality and by Markov’s inequality applied to \( x^2 \) and \( x^4 \).

Solution:

Exercise 1.61 Suppose \( x \) and \( y \) are two random 0-1 d-vectors. Show that with high probability the cosine of the angle between them is close to \( \frac{1}{2} \). Hint: Model your proof after that of the random projection theorem.

Solution:

Exercise 1.62 Generate 20 points uniformly at random on a 1,000-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 difference between \( \sqrt{\frac{k}{d}} \) times the original distances and the new pair-wise distances. For each value of \( k \) what is the maximum difference as a percent of \( \sqrt{\frac{k}{d}} \).
Solution:

Exercise 1.63  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 $x$ in $P$ and $y$ in $Q$ such that $|x - y|$ 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 $x \in P, y \in Q$ 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 $x = y$.

Solution: Use the random projection theorem. Project $P, Q$ to a random $O(\ln n)$ dimensional space (which preserves all pairwise distances) and compute all pairwise distances in that space.

Exercise 1.64  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%.

Solution: We want a random projection that will preserve the lengths of all $\binom{n}{k}$ sums of $k$-subsets. For this, it suffices to project to $O(\ln(n^k)) = O(k \ln n)$ dimensions.

Exercise 1.65  In $d$-dimensions there are exactly $d$-unit vectors that are pairwise orthogonal. However, if you wanted a set of vectors that were almost orthogonal you might squeeze in a few more. For example, in 2-dimensions if almost orthogonal meant at least 45 degrees apart you could fit in three almost orthogonal vectors. Suppose you wanted to find 900 almost orthogonal vectors in 100 dimensions where almost orthogonal meant an angle of between 85 and 95 degrees. How would you generate such a set?

Hint: Consider projecting a 1,000 orthonormal vectors to a random 100-dimensional space.

Exercise 1.66  To preserve pairwise distances between $n$ data points in $d$ space, we projected to a random $O(\ln n/\varepsilon^2)$ dimensional space. To save time in carrying out the projection, we may try to project to a space spanned by sparse vectors, vectors with only a few nonzero entries. that is, choose $O(\ln n/\varepsilon^2)$ vectors at random, each with 100 nonzero components and project to the space spanned by them. Will this work (to preserve approximately all pairwise distances) ? Why?

Solution:

Exercise 1.67  Create a list of the five most important things that you learned about high dimensions.

Exercise 1.68  Write a short essay whose purpose is to excite a college freshman to learn about high dimensions.