Foundations of Data Science[*]

Avrim Blum, John Hopcroft, and Ravindran Kannan

Thursday 4th January, 2018

Contents

1 Introduction

Computer science as an academic discipline began in the 1960's. Emphasis was on programming languages, compilers, operating systems, and the mathematical theory that supported these areas. Courses in theoretical computer science covered finite automata, regular expressions, context-free languages, and computability. In the 1970's, the study of algorithms was added as an important component of theory. The emphasis was on making computers useful. Today, a fundamental change is taking place and the focus is more on a wealth of applications. There are many reasons for this change. The merging of computing and communications has played an important role. The enhanced ability to observe, collect, and store data in the natural sciences, in commerce, and in other fields calls for a change in our understanding of data and how to handle it in the modern setting. The emergence of the web and social networks as central aspects of daily life presents both opportunities and challenges for theory.

While traditional areas of computer science remain highly important, increasingly researchers of the future will be involved with using computers to understand and extract usable information from massive data arising in applications, not just how to make computers useful on specific well-defined problems. With this in mind we have written this book to cover the theory we expect to be useful in the next 40 years, just as an understanding of automata theory, algorithms, and related topics gave students an advantage in the last 40 years. One of the major changes is an increase in emphasis on probability, statistics, and numerical methods.

Early drafts of the book have been used for both undergraduate and graduate courses. Background material needed for an undergraduate course has been put in the appendix. For this reason, the appendix has homework problems.

Modern data in diverse fields such as information processing, search, and machine learning is often advantageously represented as vectors with a large number of components. The vector representation is not just a book-keeping device to store many fields of a record. Indeed, the two salient aspects of vectors: geometric (length, dot products, orthogonality etc.) and linear algebraic (independence, rank, singular values etc.) turn out to be relevant and useful. Chapters 2 and 3 lay the foundations of geometry and linear algebra respectively. More specifically, our intuition from two or three dimensional space can be surprisingly off the mark when it comes to high dimensions. Chapter 2 works out the fundamentals needed to understand the differences. The emphasis of the chapter, as well as the book in general, is to get across the intellectual ideas and the mathematical foundations rather than focus on particular applications, some of which are briefly described. Chapter 3 focuses on singular value decomposition (SVD) a central tool to deal with matrix data. We give a from-first-principles description of the mathematics and algorithms for SVD. Applications of singular value decomposition include principal component analysis, a widely used technique which we touch upon, as well as modern applications to statistical mixtures of probability densities, discrete optimization, etc., which are described in more detail.

Exploring large structures like the web or the space of configurations of a large system with deterministic methods can be prohibitively expensive. Random walks (also called Markov Chains) turn out often to be more efficient as well as illuminative. The stationary distributions of such walks are important for applications ranging from web search to the simulation of physical systems. The underlying mathematical theory of such random walks, as well as connections to electrical networks, forms the core of Chapter 4 on Markov chains.

One of the surprises of computer science over the last two decades is that some domain-independent methods have been immensely successful in tackling problems from diverse areas. Machine learning is a striking example. Chapter 5 describes the foundations of machine learning, both algorithms for optimizing over given training examples, as well as the theory for understanding when such optimization can be expected to lead to good performance on new, unseen data. This includes important measures such as the Vapnik-Chervonenkis dimension, important algorithms such as the Perceptron Algorithm, stochastic gradient descent, boosting, and deep learning, and important notions such as regularization and overfitting.

The field of algorithms has traditionally assumed that the input data to a problem is presented in random access memory, which the algorithm can repeatedly access. This is not feasible for problems involving enormous amounts of data. The streaming model and other models have been formulated to reflect this. In this setting, sampling plays a crucial role and, indeed, we have to sample on the fly. In Chapter 6 we study how to draw good samples efficiently and how to estimate statistical and linear algebra quantities, with such samples.

While Chapter 5 focuses on supervised learning, where one learns from labeled training data, the problem of unsupervised learning, or learning from unlabeled data, is equally important. A central topic in unsupervised learning is clustering, discussed in Chapter 7. Clustering refers to the problem of partitioning data into groups of similar objects. After describing some of the basic methods for clustering, such as the k-means algorithm, Chapter 7 focuses on modern developments in understanding these, as well as newer algorithms and general frameworks for analyzing different kinds of clustering problems.

Central to our understanding of large structures, like the web and social networks, is building models to capture essential properties of these structures. The simplest model is that of a random graph formulated by Erdös and Renyi, which we study in detail in Chapter 8, proving that certain global phenomena, like a giant connected component, arise in such structures with only local choices. We also describe other models of random graphs.

Chapter 9 focuses on linear-algebraic problems of making sense from data, in particular topic modeling and non-negative matrix factorization. In addition to discussing well-known models, we also describe some current research on models and algorithms with provable guarantees on learning error and time. This is followed by graphical models and belief propagation.

Chapter 10 discusses ranking and social choice as well as problems of sparse representations such as compressed sensing. Additionally, Chapter 10 includes a brief discussion of linear programming and semidefinite programming. Wavelets, which are an important method for representing signals across a wide range of applications, are discussed in Chapter 11 along with some of their fundamental mathematical properties. The appendix includes a range of background material.

A word about notation in the book. To help the student, we have adopted certain notations, and with a few exceptions, adhered to them. We use lower case letters for scalar variables and functions, bold face lower case for vectors, and upper case letters for matrices. Lower case near the beginning of the alphabet tend to be constants, in the middle of the alphabet, such as i, j, and k, are indices in summations, n and m for integer sizes, and x, y and z for variables. If A is a matrix its elements are a_{ij} and its rows are $\mathbf{a_i}$. If $\mathbf{a_i}$ is a vector its coordinates are a_{ij}. Where the literature traditionally uses a symbol for a quantity, we also used that symbol, even if it meant abandoning our convention. If we have a set of points in some vector space, and work with a subspace, we use n for the number of points, d for the dimension of the space, and k for the dimension of the subspace.

The term "almost surely" means with probability tending to one. We use $\ln n$ for the natural logarithm and $\log n$ for the base two logarithm. If we want base ten, we will use $\log_{10}$. To simplify notation and to make it easier to read we use $E^2(1-x)$ for $\big(E(1-x)\big)^2$ and $E(1-x)^2$ for $E\left((1-x)^2\right)$. When we say "randomly select" some number of points from a given probability distribution, independence is always assumed unless otherwise stated.

2 High-Dimensional Space

2.1 Introduction

High dimensional data has become very important. However, high dimensional space is very different from the two and three dimensional spaces we are familiar with. Generate n points at random in d-dimensions where each coordinate is a zero mean, unit variance Gaussian. For sufficiently large d, with high probability the distances between all pairs of points will be essentially the same. Also the volume of the unit ball in d-dimensions, the set of all points $\mathbf{x}$ such that $|\mathbf{x}| \leq 1$, goes to zero as the dimension goes to infinity. The volume of a high dimensional unit ball is concentrated near its surface and is also concentrated at its equator. These properties have important consequences which we will consider.

2.2 The Law of Large Numbers

If one generates random points in d-dimensional space using a Gaussian to generate coordinates, the distance between all pairs of points will be essentially the same when d is large. The reason is that the square of the distance between two points $\mathbf{y}$ and $\mathbf{z}$,

$$|\mathbf{y}-\mathbf{z}|^2 = \sum_{i=1}^{d} (y_i - z_i)^2,$$

can be viewed as the sum of d independent samples of a random variable x that is distributed as the squared difference of two Gaussians. In particular, we are summing independent samples $x_i = (y_i - z_i)^2$ of a random variable x of bounded variance. In such a case, a general bound known as the Law of Large Numbers states that with high probability, the average of the samples will be close to the expectation of the random variable. This in turn implies that with high probability, the sum is close to the sum's expectation.

Specifically, the Law of Large Numbers states that

$$\text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) \leq \frac{Var(x)}{n\epsilon^2}. \tag{2.1}$$

The larger the variance of the random variable, the greater the probability that the error will exceed ϵ. Thus the variance of x is in the numerator. The number of samples n is in the denominator since the more values that are averaged, the smaller the probability that the difference will exceed ϵ. Similarly the larger ϵ is, the smaller the probability that the difference will exceed ϵ and hence ϵ is in the denominator. Notice that squaring ϵ makes the fraction a dimensionless quantity.

We use two inequalities to prove the Law of Large Numbers. The first is Markov's inequality that states that the probability that a nonnegative random variable exceeds a is bounded by the expected value of the variable divided by a.

Theorem 2.1 (Markov's inequality) *Let x be a nonnegative random variable. Then for $a > 0$,*

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

Proof: For a continuous nonnegative random variable x with probability density p,

$$\begin{aligned} E\left(x\right) &= \int\limits_0^\infty xp(x)dx = \int\limits_0^a xp(x)dx + \int\limits_a^\infty xp(x)dx \\ &\geq \int\limits_a^\infty xp(x)dx \geq a\int\limits_a^\infty p(x)dx = a\text{Prob}(x \geq a). \end{aligned}$$

Thus, $\text{Prob}(x \geq a) \leq \frac{E(x)}{a}$. ∎

The same proof works for discrete random variables with sums instead of integrals.

Corollary 2.2 $Prob\big(x \geq bE(x)\big) \leq \frac{1}{b}$

Markov's inequality bounds the tail of a distribution using only information about the mean. A tighter bound can be obtained by also using the variance of the random variable.

Theorem 2.3 (Chebyshev's inequality) *Let x be a random variable. Then for $c > 0$,*

$$Prob\Big(\,|x - E(x)| \geq c\Big) \leq \frac{Var(x)}{c^2}.$$

Proof: $\text{Prob}\big(|x - E(x)| \geq c\big) = \text{Prob}\big(|x - E(x)|^2 \geq c^2\big)$. Let $y = |x - E(x)|^2$. Note that y is a nonnegative random variable and $E(y) = Var(x)$, so Markov's inequality can be applied giving:

$$\text{Prob}(|x - E(x)| \geq c) = \text{Prob}\left(|x - E(x)|^2 \geq c^2\right) \leq \frac{E(|x - E(x)|^2)}{c^2} = \frac{Var(x)}{c^2}.$$

∎

The Law of Large Numbers follows from Chebyshev's inequality together with facts about independent random variables. Recall that:

$$\begin{aligned} &E(x + y) = E(x) + E(y), \\ &Var(x - c) = Var(x), \\ &Var(cx) = c^2 Var(x). \end{aligned}$$

Also, if x and y are independent, then $E(xy) = E(x)E(y)$. These facts imply that if x and y are independent then $Var(x + y) = Var(x) + Var(y)$, which is seen as follows:

$$\begin{aligned} Var(x + y) &= E(x + y)^2 - E^2(x + y) \\ &= E(x^2 + 2xy + y^2) - \big(E^2(x) + 2E(x)E(y) + E^2(y)\big) \\ &= E(x^2) - E^2(x) + E(y^2) - E^2(y) = Var(x) + Var(y), \end{aligned}$$

where we used independence to replace $E(2xy)$ with $2E(x)E(y)$.

Theorem 2.4 (Law of Large Numbers) *Let $x_1, x_2, \ldots, x_n$ be n independent samples of a random variable x. Then*

$$Prob\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) \leq \frac{Var(x)}{n\epsilon^2}$$

Proof: By Chebychev's inequality

$$\begin{aligned} \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) &\leq \frac{Var\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)}{\epsilon^2} \\ &= \frac{1}{n^2\epsilon^2} Var(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n^2\epsilon^2}\big(Var(x_1) + Var(x_2) + \cdots + Var(x_n)\big) \\ &= \frac{Var(x)}{n\epsilon^2}. \end{aligned}$$

■

The Law of Large Numbers is quite general, applying to any random variable x of finite variance. Later we will look at tighter concentration bounds for spherical Gaussians and sums of 0-1 valued random variables.

One observation worth making about the Law of Large Numbers is that the size of the universe does not enter into the bound. For instance, if you want to know what fraction of the population of a country prefers tea to coffee, then the number n of people you need to sample in order to have at most a δ chance that your estimate is off by more than ϵ depends only on ϵ and δ and not on the population of the country.

As an application of the Law of Large Numbers, let $\mathbf{z}$ be a d-dimensional random point whose coordinates are each selected from a zero mean, $\frac{1}{2\pi}$ variance Gaussian. We set the variance to $\frac{1}{2\pi}$ so the Gaussian probability density equals one at the origin and is bounded below throughout the unit ball by a constant.[1] By the Law of Large Numbers, the square of the distance of $\mathbf{z}$ to the origin will be $\Theta(d)$ with high probability. In particular, there is

[1] If we instead used variance 1, then the density at the origin would be a decreasing function of d, namely $(\frac{1}{2\pi})^{d/2}$, making this argument more complicated.

vanishingly small probability that such a random point $\mathbf{z}$ would lie in the unit ball. This implies that the integral of the probability density over the unit ball must be vanishingly small. On the other hand, the probability density in the unit ball is bounded below by a constant. We thus conclude that the unit ball must have vanishingly small volume.

Similarly if we draw two points $\mathbf{y}$ and $\mathbf{z}$ from a d-dimensional Gaussian with unit variance in each direction, then $|\mathbf{y}|^2 \approx d$ and $|\mathbf{z}|^2 \approx d$. Since for all i,

$$E(y_i - z_i)^2 = E(y_i^2) + E(z_i^2) - 2E(y_i z_i) = Var(y_i) + Var(z_i) + 2E(y_i)E(z_i) = 2,$$

$|\mathbf{y}-\mathbf{z}|^2 = \sum\limits_{i=1}^{d}(y_i - z_i)^2 \approx 2d$. Thus by the Pythagorean theorem, the random d-dimensional $\mathbf{y}$ and $\mathbf{z}$ must be approximately orthogonal. This implies that if we scale these random points to be unit length and call $\mathbf{y}$ the North Pole, much of the surface area of the unit ball must lie near the equator. We will formalize these and related arguments in subsequent sections.

We now state a general theorem on probability tail bounds for a sum of independent random variables. Tail bounds for sums of Bernoulli, squared Gaussian and Power Law distributed random variables can all be derived from this. The table in Figure 2.1 summarizes some of the results.

Theorem 2.5 (Master Tail Bounds Theorem) *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 σ^2. Let $0 \le a \le \sqrt{2}n\sigma^2$. Assume that $|E(x_i^s)| \le \sigma^2 s!$ for $s = 3, 4, \ldots, \lfloor(a^2/4n\sigma^2)\rfloor$. Then,*

$$\textit{Prob}\left(|x| \ge a\right) \le 3e^{-a^2/(12n\sigma^2)}.$$

The proof of Theorem 2.5 is elementary. A slightly more general version, Theorem 12.5, is given in the appendix. For a brief intuition of the proof, consider applying Markov's inequality to the random variable x^r where r is a large even number. Since r is even, x^r is nonnegative, and thus $\text{Prob}(|x| \ge a) = \text{Prob}(x^r \ge a^r) \le E(x^r)/a^r$. If $E(x^r)$ is not too large, we will get a good bound. To compute $E(x^r)$, write $E(x)$ as $E(x_1 + \ldots + x_n)^r$ and expand the polynomial into a sum of terms. Use the fact that by independence $E(x_i^{r_i} x_j^{r_j}) = E(x_i^{r_i})E(x_j^{r_j})$ to get a collection of simpler expectations that can be bounded using our assumption that $|E(x_i^s)| \le \sigma^2 s!$. For the full proof, see the appendix.

2.3 The Geometry of High Dimensions

An important property of high-dimensional objects is that most of their volume is near the surface. Consider any object A in R^d. Now shrink A by a small amount ϵ to produce a new object $(1-\epsilon)A = \{(1-\epsilon)x | x \in A\}$. Then the following equality holds:

$$\text{volume}\big((1-\epsilon)A\big) = (1-\epsilon)^d \text{volume}(A).$$

	Condition	Tail bound
Markov	$x \geq 0$	$\text{Prob}(x \geq a) \leq \frac{E(x)}{a}$
Chebychev	Any x	$\text{Prob}\big(\lvert x - E(x)\rvert \geq a\big) \leq \frac{\text{Var}(x)}{a^2}$
Chernoff	$x = x_1 + x_2 + \cdots + x_n$ $x_i \in [0,1]$ i.i.d. Bernoulli;	$\text{Prob}(\lvert x - E(x)\rvert \geq \varepsilon E(x))$ $\leq 3e^{-c\varepsilon^2 E(x)}$
Higher Moments	r positive even integer	$\text{Prob}(\lvert x\rvert \geq a) \leq E(x^r)/a^r$
Gaussian Annulus	$x = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ $x_i \sim N(0,1); \beta \leq \sqrt{n}$ indep.	$\text{Prob}(\lvert x - \sqrt{n}\rvert \geq \beta) \leq 3e^{-c\beta^2}$
Power Law for x_i; order $k \geq 4$	$x = x_1 + x_2 + \ldots + x_n$ x_i i.i.d ; $\varepsilon \leq 1/k^2$	$\text{Prob}\big(\lvert x - E(x)\rvert \geq \varepsilon E(x)\big)$ $\leq (4/\varepsilon^2 kn)^{(k-3)/2}$

Figure 2.1: Table of Tail Bounds. The Higher Moments bound is obtained by applying Markov to x^r. The Chernoff, Gaussian Annulus, and Power Law bounds follow from Theorem 2.5 which is proved in the appendix.

To see that this is true, partition A into infinitesimal cubes. Then, $(1-\varepsilon)A$ is the union of a set of cubes obtained by shrinking the cubes in A by a factor of $1-\varepsilon$. When we shrink each of the $2d$ sides of a d-dimensional cube by a factor f, its volume shrinks by a factor of f^d. Using the fact that $1-x \leq e^{-x}$, for any object A in R^d we have:

$$\frac{\text{volume}\big((1-\epsilon)A\big)}{\text{volume}(A)} \;=\; (1-\epsilon)^d \;\leq\; e^{-\epsilon d}.$$

Fixing ϵ and letting $d \to \infty$, the above quantity rapidly approaches zero. This means that nearly all of the volume of A must be in the portion of A that does not belong to the region $(1-\epsilon)A$.

Let S denote the unit ball in d dimensions, that is, the set of points within distance one of the origin. An immediate implication of the above observation is that at least a $1 - e^{-\epsilon d}$ fraction of the volume of the unit ball is concentrated in $S \setminus (1-\epsilon)S$, namely in a small annulus of width ϵ at the boundary. In particular, most of the volume of the d-dimensional unit ball is contained in an annulus of width $O(1/d)$ near the boundary. If the ball is of radius r, then the annulus width is $O\left(\frac{r}{d}\right)$.

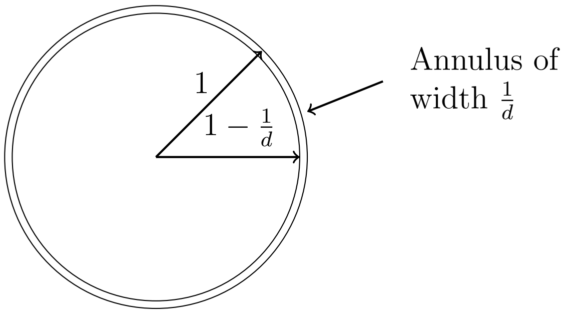


Figure 2.2: Most of the volume of the d-dimensional ball of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

2.4 Properties of the Unit Ball

We now focus more specifically on properties of the unit ball in d-dimensional space. We just saw that most of its volume is concentrated in a small annulus of width $O(1/d)$ near the boundary. Next we will show that in the limit as d goes to infinity, the volume of the ball goes to zero. This result can be proven in several ways. Here we use integration.

2.4.1 Volume of the Unit Ball

To calculate the volume $V(d)$ of the unit ball in R^d, one can integrate in either Cartesian or polar coordinates. In Cartesian coordinates the volume is given by

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

Since the limits of the integrals are complicated, it is easier to integrate using polar coordinates. In polar coordinates, $V(d)$ is given by

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

Since the variables Ω and r do not interact,

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

where $A(d)$ is the surface area of the d-dimensional unit ball. For instance, for $d = 3$ the surface area is 4π and the volume is $\frac{4}{3}\pi$. The question remains, how to determine the surface area $A\left(d\right) = \int\limits_{S^d} d\Omega$ for general d.

Consider a different integral

$$I\left(d\right) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty} e^{-\left(x_1^2+x_2^2+\cdots x_d^2\right)}dx_d\cdots dx_2dx_1.$$

Including the exponential allows integration to infinity rather than stopping at the surface of the sphere. Thus, $I(d)$ can be computed by integrating in both Cartesian and polar coordinates. Integrating in polar coordinates will relate $I(d)$ to the surface area $A(d)$. Equating the two results for $I(d)$ allows one to solve for $A(d)$.

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

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

Here, we have used the fact that $\int_{-\infty}^{\infty} e^{-x^2}\ dx = \sqrt{\pi}$. For a proof of this, see Section 12.3 of the appendix. Next, calculate $I(d)$ by integrating in polar coordinates. The volume of the differential element is $r^{d-1}d\Omega dr$. Thus,

$$I\left(d\right) = \int\limits_{S^d} d\Omega \int\limits_0^{\infty} e^{-r^2}r^{d-1}dr.$$

The integral $\int\limits_{S^d} d\Omega$ is the integral over the entire solid angle and gives the surface area, $A(d)$, of a unit sphere. Thus, $I\left(d\right) = A\left(d\right)\int\limits_0^{\infty} e^{-r^2}r^{d-1}dr$. Evaluating the remaining integral gives

$$\int\limits_0^{\infty} e^{-r^2}r^{d-1}dr = \int\limits_0^{\infty} e^{-t}t^{\frac{d-1}{2}}\left(\tfrac{1}{2}t^{-\frac{1}{2}}dt\right) = \frac{1}{2}\int\limits_0^{\infty} e^{-t}t^{\frac{d}{2}\,-\,1}dt = \frac{1}{2}\Gamma\left(\frac{d}{2}\right)$$

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

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

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

establishing the following lemma.

Lemma 2.6 *The surface area $A(d)$ and the volume $V(d)$ of a unit-radius ball in d dimensions are given by*

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

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

2.4.2 Volume Near the Equator

An interesting fact about the unit ball in high dimensions is that most of its volume is concentrated near its "equator". In particular, for any unit-length vector $\mathbf{v}$ defining "north", most of the volume of the unit ball lies in the thin slab of points whose dot-product with $\mathbf{v}$ has magnitude $O(1/\sqrt{d})$. To show this fact, it suffices by symmetry to fix $\mathbf{v}$ to be the first coordinate vector. That is, we will show that most of the volume of the unit ball has $|x_1|=O(1/\sqrt{d})$. Using this fact, we will show that two random points in the unit ball are with high probability nearly orthogonal, and also give an alternative proof from the one in Section 2.4.1 that the volume of the unit ball goes to zero as $d\to\infty$.

Theorem 2.7 *For $c\geq 1$ and $d\geq 3$, at least a $1-\frac{2}{c}e^{-c^2/2}$ fraction of the volume of the d-dimensional unit ball has $|x_1|\leq\frac{c}{\sqrt{d-1}}$.*

Proof: By symmetry we just need to prove that at most a $\frac{2}{c}e^{-c^2/2}$ fraction of the half of the ball with $x_1\geq 0$ has $x_1\geq\frac{c}{\sqrt{d-1}}$. Let A denote the portion of the ball with $x_1\geq\frac{c}{\sqrt{d-1}}$ and let H denote the upper hemisphere. We will then show that the ratio of the volume of A to the volume of H goes to zero by calculating an upper bound on volume(A) and a lower bound on volume(H) and proving that

$$\frac{\text{volume}(A)}{\text{volume}(H)}\leq\frac{\text{upper bound volume}(A)}{\text{lower bound volume}(H)}=\frac{2}{c}e^{-\frac{c^2}{2}}.$$

To calculate the volume of A, integrate an incremental volume that is a disk of width dx_1 and whose face is a ball of dimension $d-1$ and radius $\sqrt{1-x_1^2}$. The surface area of the disk is $(1-x_1^2)^{\frac{d-1}{2}}V(d-1)$ and the volume above the slice is

$$\text{volume}(A)=\int_{\frac{c}{\sqrt{d-1}}}^{1}(1-x_1^2)^{\frac{d-1}{2}}V(d-1)dx_1$$

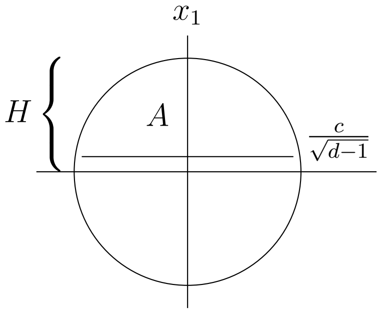


Figure 2.3: Most of the volume of the upper hemisphere of the d-dimensional ball is below the plane $x_1 = \frac{c}{\sqrt{d-1}}$.

To get an upper bound on the above integral, use $1 - x \leq e^{-x}$ and integrate to infinity. To integrate, insert $\frac{x_1\sqrt{d-1}}{c}$, which is greater than one in the range of integration, into the integral. Then

$$\text{volume}(A) \leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} \frac{x_1\sqrt{d-1}}{c} e^{-\frac{d-1}{2}x_1^2} V(d-1)dx_1 = V(d-1)\frac{\sqrt{d-1}}{c}\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} dx_1$$

Now

$$\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} dx_1 = -\frac{1}{d-1} e^{-\frac{d-1}{2}x_1^2} \Big|_{\frac{c}{\sqrt{(d-1)}}}^{\infty} = \frac{1}{d-1} e^{-\frac{c^2}{2}}$$

Thus, an upper bound on volume(A) is $\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}$.

The volume of the hemisphere below the plane $x_1 = \frac{1}{\sqrt{d-1}}$ is a lower bound on the entire volume of the upper hemisphere and this volume is at least that of a cylinder of height $\frac{1}{\sqrt{d-1}}$ and radius $\sqrt{1 - \frac{1}{d-1}}$. The volume of the cylinder is $V(d-1)(1 - \frac{1}{d-1})^{\frac{d-1}{2}} \frac{1}{\sqrt{d-1}}$. Using the fact that $(1-x)^a \geq 1 - ax$ for $a \geq 1$, the volume of the cylinder is at least $\frac{V(d-1)}{2\sqrt{d-1}}$ for $d \geq 3$.

Thus,

$$\text{ratio} \leq \frac{\text{upper bound above plane}}{\text{lower bound total hemisphere}} = \frac{\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}}{\frac{V(d-1)}{2\sqrt{d-1}}} = \frac{2}{c} e^{-\frac{c^2}{2}}$$

■

One might ask why we computed a lower bound on the total hemisphere since it is one half of the volume of the unit ball which we already know. The reason is that the volume of the upper hemisphere is $\frac{1}{2}V(d)$ and we need a formula with $V(d-1)$ in it to cancel the $V(d-1)$ in the numerator.

Near orthogonality. One immediate implication of the above analysis is that if we draw two points at random from the unit ball, with high probability their vectors will be nearly orthogonal to each other. Specifically, from our previous analysis in Section 2.3, with high probability both will be close to the surface and will have length $1 - O(1/d)$. From our analysis above, if we define the vector in the direction of the first point as "north", with high probability the second will have a projection of only $\pm O(1/\sqrt{d})$ in this direction, and thus their dot-product will be $\pm O(1/\sqrt{d})$. This implies that with high probability, the angle between the two vectors will be $\pi/2 \pm O(1/\sqrt{d})$. In particular, we have the following theorem that states that if we draw n points at random in the unit ball, with high probability all points will be close to unit length and each pair of points will be almost orthogonal.

Theorem 2.8 *Consider drawing n points $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ at random from the unit ball. With probability $1 - O(1/n)$*

1. $|\mathbf{x_i}| \geq 1 - \frac{2\ln n}{d}$ *for all i, and*
2. $|\mathbf{x_i} \cdot \mathbf{x_j}| \leq \frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ *for all $i \neq j$.*

Proof: For the first part, for any fixed i by the analysis of Section 2.3, the probability that $|\mathbf{x_i}| < 1 - \epsilon$ is less than $e^{-\epsilon d}$. Thus

$$\text{Prob}\big(|\mathbf{x_i}| < 1 - \frac{2\ln n}{d}\big) \leq e^{-(\frac{2\ln n}{d})d} = 1/n^2.$$

By the union bound, the probability there exists an i such that $|\mathbf{x_i}| < 1 - \frac{2\ln n}{d}$ is at most $1/n$.

For the second part, Theorem 2.7 states that the probability $|\mathbf{x_i}| > \frac{c}{\sqrt{d-1}}$ is at most $\frac{2}{c}e^{-\frac{c^2}{2}}$. There are $\binom{n}{2}$ pairs i and j and for each such pair if we define $\mathbf{x_i}$ as "north", the probability that the projection of $\mathbf{x_j}$ onto the "north" direction is more than $\frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ is at most $O(e^{-\frac{6\ln n}{2}}) = O(n^{-3})$. Thus, the dot-product condition is violated with probability at most $O\left(\binom{n}{2}n^{-3}\right) = O(1/n)$ as well. ∎

Alternative proof that volume goes to zero. Another immediate implication of Theorem 2.7 is that as $d \to \infty$, the volume of the ball approaches zero. Specifically, consider a small box centered at the origin of side length $\frac{2c}{\sqrt{d-1}}$. Using Theorem 2.7, we show that for $c = 2\sqrt{\ln d}$, this box contains over half of the volume of the ball. On the other hand, the volume of this box clearly goes to zero as d goes to infinity, since its volume is $O((\frac{\ln d}{d-1})^{d/2})$. Thus the volume of the ball goes to zero as well.

By Theorem 2.7 with $c = 2\sqrt{\ln d}$, the fraction of the volume of the ball with $|x_1| \geq \frac{c}{\sqrt{d-1}}$ is at most:

$$\frac{2}{c}e^{-\frac{c^2}{2}} = \frac{1}{\sqrt{\ln d}}e^{-2\ln d} = \frac{1}{d^2\sqrt{\ln d}} < \frac{1}{d^2}.$$

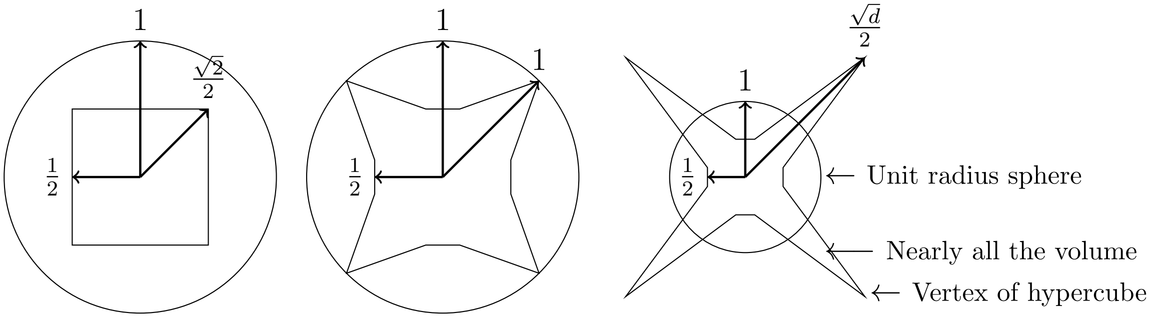


Figure 2.4: Illustration of the relationship between the sphere and the cube in 2, 4, and d-dimensions.

Since this is true for each of the d dimensions, by a union bound at most a $O(\frac{1}{d}) \leq \frac{1}{2}$ fraction of the volume of the ball lies outside the cube, completing the proof.

Discussion. One might wonder how it can be that nearly all the points in the unit ball are very close to the surface and yet at the same time nearly all points are in a box of side-length $O\left(\frac{\ln d}{d-1}\right)$. The answer is to remember that points on the surface of the ball satisfy $x_1^2 + x_2^2 + \ldots + x_d^2 = 1$, so for each coordinate i, a typical value will be $\pm O\left(\frac{1}{\sqrt{d}}\right)$. In fact, it is often helpful to think of picking a random point on the sphere as very similar to picking a random point of the form $\left(\pm\frac{1}{\sqrt{d}}, \pm\frac{1}{\sqrt{d}}, \pm\frac{1}{\sqrt{d}}, \ldots \pm \frac{1}{\sqrt{d}}\right)$.

2.5 Generating Points Uniformly at Random from a Ball

Consider generating points uniformly at random on the surface of the unit ball. For the 2-dimensional version of generating points on the circumference of a unit-radius circle, 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.

In higher dimensions, this method does not work since the fraction of points that fall inside the ball drops to zero and all of the points would be thrown away. The solution is to generate a point each of whose coordinates is an independent Gaussian variable. Generate $x_1, x_2, \ldots, x_d$, using a zero mean, unit variance Gaussian, namely, $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ on the real line.[2] Thus, the probability density of $\mathbf{x}$ is

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

and is spherically symmetric. Normalizing the vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ to a unit vector, namely $\frac{\mathbf{x}}{|\mathbf{x}|}$, 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.

To generate a point $\mathbf{y}$ uniformly over the ball (surface and interior), scale the point $\frac{\mathbf{x}}{|\mathbf{x}|}$ generated on the surface by a scalar $\rho \in [0, 1]$. What should the distribution of ρ be as a function of r? It is certainly not uniform, even in 2 dimensions. Indeed, the density of ρ at r is proportional to r for $d = 2$. For $d = 3$, it is proportional to r^2. By similar reasoning, the density of ρ at distance r is proportional to r^{d-1} in d dimensions. Solving $\int_{r=0}^{r=1} cr^{d-1}dr = 1$ (the integral of density must equal 1) one should set $c = d$. Another way to see this formally is that the volume of the radius r ball in d dimensions is $r^dV(d)$. The density at radius r is exactly $\frac{d}{dr}(r^dV_d) = dr^{d-1}V_d$. So, pick $\rho(r)$ with density equal to dr^{d-1} for r over $[0, 1]$.

We have succeeded in generating a point

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

uniformly at random from the unit ball by using the convenient spherical Gaussian distribution. In the next sections, we will analyze the spherical Gaussian in more detail.

2.6 Gaussians in High Dimension

A 1-dimensional Gaussian has its mass close to the origin. However, as the dimension is increased something different happens. The d-dimensional spherical Gaussian with zero mean and variance σ^2 in each coordinate has density function

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

The value of the density is maximum at the origin, but there is very little volume there. When $\sigma^2 = 1$, integrating the probability density over a unit ball centered at the origin yields almost zero mass since the volume of such a ball is negligible. In fact, one needs

[2] One might naturally ask: "how do you generate a random number from a 1-dimensional Gaussian?" To generate a number from any distribution given its cumulative distribution function P, first select a uniform random number $u \in [0, 1]$ and then choose $x = P^{-1}(u)$. For any $a < b$, the probability that x is between a and b is equal to the probability that u is between $P(a)$ and $P(b)$ which equals $P(b) - P(a)$ as desired. For the 2-dimensional Gaussian, one can generate a point in polar coordinates by choosing angle θ uniform in $[0, 2\pi]$ and radius $r = \sqrt{-2\ln(u)}$ where u is uniform random in $[0, 1]$. This is called the Box-Muller transform.

to increase the radius of the ball to nearly $\sqrt{d}$ before there is a significant volume and hence significant probability mass. If one increases the radius much beyond $\sqrt{d}$, the integral barely increases even though the volume increases since the probability density is dropping off at a much higher rate. The following theorem formally states that nearly all the probability is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{d}$.

Theorem 2.9 (Gaussian Annulus Theorem) *For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.*

For a high-level intuition, note that $E(|\mathbf{x}|^2) = \sum\limits_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d$, so the mean squared distance of a point from the center is d. The Gaussian Annulus Theorem says that the points are tightly concentrated. We call the square root of the mean squared distance, namely $\sqrt{d}$, the radius of the Gaussian.

To prove the Gaussian Annulus Theorem we make use of a tail inequality for sums of independent random variables of bounded moments (Theorem 12.5).

Proof (Gaussian Annulus Theorem): Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ be a point selected from a unit variance Gaussian centered at the origin, and let $r = |\mathbf{x}|$. $\sqrt{d} - \beta \leq |\mathbf{y}| \leq \sqrt{d} + \beta$ is equivalent to $|r - \sqrt{d}| \geq \beta$. If $|r - \sqrt{d}| \geq \beta$, then multiplying both sides by $r + \sqrt{d}$ gives $|r^2 - d| \geq \beta(r + \sqrt{d}) \geq \beta\sqrt{d}$. So, it suffices to bound the probability that $|r^2 - d| \geq \beta\sqrt{d}$.

Rewrite $r^2 - d = (x_1^2 + \ldots + x_d^2) - d = (x_1^2 - 1) + \ldots + (x_d^2 - 1)$ and perform a change of variables: $y_i = x_i^2 - 1$. We want to bound the probability that $|y_1 + \ldots + y_d| \geq \beta\sqrt{d}$. Notice that $E(y_i) = E(x_i^2) - 1 = 0$. To apply Theorem 12.5, we need to bound the s^{th} moments of y_i.

For $|x_i| \leq 1$, $|y_i|^s \leq 1$ and for $|x_i| \geq 1$, $|y_i|^s \leq |x_i|^{2s}$. Thus

$$\begin{aligned} |E(y_i^s)| = E(|y_i|^s) &\leq E(1 + x_i^{2s}) = 1 + E(x_i^{2s}) \\ &= 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-x^2/2} dx \end{aligned}$$

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

$$\begin{aligned} |E(y_i^s)| &= 1 + \frac{1}{\sqrt{\pi}} \int_0^\infty 2^s z^{s-(1/2)} e^{-z} dz \\ &\leq 2^s s!. \end{aligned}$$

The last inequality is from the Gamma integral.

Since $E(y_i) = 0$, $Var(y_i) = E(y_i^2) \leq 2^2 2 = 8$. Unfortunately, we do not have $|E(y_i^s)| \leq 8s!$ as required in Theorem 12.5. To fix this problem, perform one more change of variables, using $w_i = y_i/2$. Then, $Var(w_i) \leq 2$ and $|E(w_i^s)| \leq 2s!$, and our goal is now to bound the probability that $|w_1 + \ldots + w_d| \geq \frac{\beta\sqrt{d}}{2}$. Applying Theorem 12.5 where $\sigma^2 = 2$ and $n = d$, this occurs with probability less than or equal to $3e^{-\frac{\beta^2}{96}}$. ∎

In the next sections we will see several uses of the Gaussian Annulus Theorem.

2.7 Random Projection and Johnson-Lindenstrauss Lemma

One of the most frequently used subroutines in tasks involving high dimensional data is nearest neighbor search. In nearest neighbor search we are given a database of n points in $\mathbf{R}^d$ where n and d are usually large. The database can be preprocessed and stored in an efficient data structure. Thereafter, we are presented "query" points in $\mathbf{R}^d$ and are asked to find the nearest or approximately nearest database point to the query point. Since the number of queries is often large, the time to answer each query should be very small, ideally a small function of $\log n$ and $\log d$, whereas preprocessing time could be larger, namely a polynomial function of n and d. For this and other problems, dimension reduction, where one projects the database points to a k-dimensional space with $k \ll d$ (usually dependent on $\log d$) can be very useful so long as the relative distances between points are approximately preserved. We will see using the Gaussian Annulus Theorem that such a projection indeed exists and is simple.

The projection $f : \mathbf{R}^d \to \mathbf{R}^k$ that we will examine (many related projections are known to work as well) is the following. Pick k Gaussian vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$ in $\mathbf{R}^d$ with unit-variance coordinates. For any vector $\mathbf{v}$, define the projection $f(\mathbf{v})$ by:

$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}).$$

The projection $f(\mathbf{v})$ is the vector of dot products of $\mathbf{v}$ with the $\mathbf{u_i}$. We will show that with high probability, $|f(\mathbf{v})| \approx \sqrt{k}|\mathbf{v}|$. For any two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$, $f(\mathbf{v_1} - \mathbf{v_2}) = f(\mathbf{v_1}) - f(\mathbf{v_2})$. Thus, to estimate the distance $|\mathbf{v_1} - \mathbf{v_2}|$ between two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ in $\mathbf{R}^d$, it suffices to compute $|f(\mathbf{v_1}) - f(\mathbf{v_2})| = |f(\mathbf{v_1} - \mathbf{v_2})|$ in the k-dimensional space since the factor of $\sqrt{k}$ is known and one can divide by it. The reason distances increase when we project to a lower dimensional space is that the vectors $\mathbf{u_i}$ are not unit length. Also notice that the vectors $\mathbf{u_i}$ are not orthogonal. If we had required them to be orthogonal, we would have lost statistical independence.

Theorem 2.10 (The Random Projection Theorem) *Let $\mathbf{v}$ be a fixed vector in $\mathbf{R}^d$ and let f be defined as above. There exists constant $c > 0$ such that for $\varepsilon \in (0, 1)$,*

$$Prob\left(\Big||f(\mathbf{v})| \; - \; \sqrt{k}|\mathbf{v}|\Big| \geq \varepsilon\sqrt{k}|\mathbf{v}|\right) \leq 3e^{-ck\varepsilon^2},$$

where the probability is taken over the random draws of vectors $\mathbf{u_i}$ used to construct f.

Proof: By scaling both sides of the inner inequality by $|\mathbf{v}|$, we may assume that $|\mathbf{v}| = 1$. The sum of independent normally distributed real variables is also normally distributed where the mean and variance are the sums of the individual means and variances. Since $\mathbf{u_i} \cdot \mathbf{v} = \sum_{j=1}^{d} u_{ij} v_j$, the random variable $\mathbf{u_i} \cdot \mathbf{v}$ has Gaussian density with zero mean and unit variance, in particular,

$$Var(\mathbf{u_i} \cdot \mathbf{v}) = Var\left(\sum_{j=1}^{d} v_{ij} v_j\right) = \sum_{j=1}^{d} v_j^2 Var(u_{ij}) = \sum_{j=1}^{d} v_j^2 = 1$$

Since $\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}$ are independent Gaussian random variables, $f(\mathbf{v})$ is a random vector from a k-dimensional spherical Gaussian with unit variance in each coordinate, and so the theorem follows from the Gaussian Annulus Theorem (Theorem 2.9) with d replaced by k. ■

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 $|\mathbf{v_i} - \mathbf{v_j}|$ among n vectors $\mathbf{v_1}, \ldots, \mathbf{v_n}$ differs significantly from their expected values is small, provided $k \geq \frac{3}{c\varepsilon^2} \ln n$. Thus, this random projection preserves all relative pairwise distances between points in a set of n points with high probability. This is the content of the Johnson-Lindenstrauss Lemma.

Theorem 2.11 (Johnson-Lindenstrauss Lemma) *For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ with c as in Theorem 2.9. For any set of n points in R^d, the random projection $f : R^d \to R^k$ defined above has the property that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $1 - 3/2n$,*

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

Proof: Applying the Random Projection Theorem (Theorem 2.10), for any fixed $\mathbf{v_i}$ and $\mathbf{v_j}$, the probability that $|f(\mathbf{v_i} - \mathbf{v_j})|$ is outside the range

$$\left[(1 - \varepsilon)\sqrt{k}|\mathbf{v_i} - \mathbf{v_j}|, (1 + \varepsilon)\sqrt{k}|\mathbf{v_i} - \mathbf{v_j}|\right]$$

is at most $3e^{-ck\varepsilon^2} \leq 3/n^3$ for $k \geq \frac{3 \ln n}{c\varepsilon^2}$. Since there are $\binom{n}{2} < n^2/2$ pairs of points, by the union bound, the probability that any pair has a large distortion is less than $\frac{3}{2n}$. ■

Remark: It is important to note that the conclusion of Theorem 2.11 asserts for all $\mathbf{v_i}$ and $\mathbf{v_j}$, not just for most of them. The weaker assertion for most $\mathbf{v_i}$ and $\mathbf{v_j}$ is typically less useful, since our algorithm for a problem such as nearest-neighbor search might return one of the bad pairs of points. 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. In applications, the dominant term is typically the $1/\varepsilon^2$ term.

For the nearest neighbor problem, if the database has n_1 points and n_2 queries are expected during the lifetime of the algorithm, take $n = n_1 + n_2$ and project the database to a random k-dimensional space, for k as in Theorem 2.11. On receiving a query, project the query to the same subspace and compute nearby database points. The Johnson Lindenstrauss Lemma says that with high probability this will yield the right answer whatever the query. Note that the exponentially small in k probability was useful here in making k only dependent on $\ln n$, rather than n.

2.8 Separating Gaussians

Mixtures of Gaussians are often used to model heterogeneous data coming from multiple sources. For example, suppose we are recording the heights of individuals age 20-30 in a city. We know that on average, men tend to be taller than women, so a natural model would be a Gaussian mixture model $p(x) = w_1p_1(x) + w_2p_2(x)$, where $p_1(x)$ is a Gaussian density representing the typical heights of women, $p_2(x)$ is a Gaussian density representing the typical heights of men, and w_1 and w_2 are the *mixture weights* representing the proportion of women and men in the city. The *parameter estimation problem* for a mixture model is the problem: given access to samples from the overall density p (e.g., heights of people in the city, but without being told whether the person with that height is male or female), reconstruct the parameters for the distribution (e.g., good approximations to the means and variances of p_1 and p_2, as well as the mixture weights).

There are taller women and shorter men, so even if one solved the parameter estimation problem for heights perfectly, given a data point, one couldn't necessarily tell which population it came from. That is, given a height, one couldn't necessarily tell if it came from a man or a woman. In this section, we will look at a problem that is in some ways easier and some ways harder than this problem of heights. It will be harder in that we will be interested in a mixture of two Gaussians in high-dimensions as opposed to the $d = 1$ case of heights. But it will be easier in that we will assume the means are quite well-separated compared to the variances. Specifically, our focus will be on a mixture of two spherical unit-variance Gaussians whose means are separated by a distance $\Omega(d^{1/4})$. We will show that at this level of separation, we can with high probability uniquely determine which Gaussian each data point came from. The algorithm to do so will actually be quite simple. Calculate the distance between all pairs of points. Points whose distance apart is smaller are from the same Gaussian, whereas 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.

First, consider just one spherical unit-variance Gaussian centered at the origin. From Theorem 2.9, most of its probability mass lies on an annulus of width $O(1)$ at radius $\sqrt{d}$. Also $e^{-|\mathbf{x}|^2/2} = \prod_i e^{-x_i^2/2}$ and almost all of the mass is within the slab $\{\ \mathbf{x} \mid -c \leq x_1 \leq c\ \}$, for $c \in O(1)$. Pick a point $\mathbf{x}$ from this Gaussian. After picking $\mathbf{x}$, rotate the coordinate system to make the first axis align with $\mathbf{x}$. Independently pick a second point $\mathbf{y}$ from

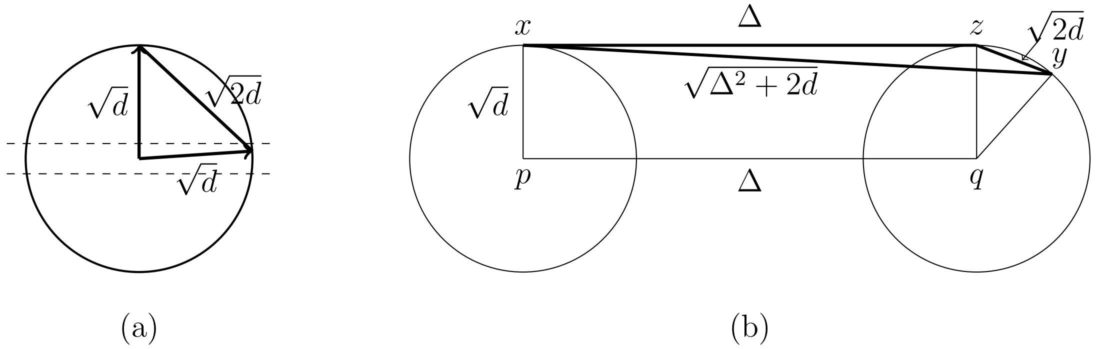


Figure 2.5: (a) indicates that two randomly chosen points in high dimension are surely almost nearly orthogonal. (b) indicates the distance between a pair of random points from two different unit balls approximating the annuli of two Gaussians.

this Gaussian. The fact that almost all of the probability mass of the Gaussian is within the slab $\{\mathbf{x} \mid -c \leq x_1 \leq c,\ c \in O(1)\}$ at the equator implies that $\mathbf{y}$'s component along $\mathbf{x}$'s direction is $O(1)$ with high probability. Thus, $\mathbf{y}$ is nearly perpendicular to $\mathbf{x}$. So, $|\mathbf{x}-\mathbf{y}| \approx \sqrt{|\mathbf{x}|^2 + |\mathbf{y}|^2}$. See Figure 2.5(a). More precisely, since the coordinate system has been rotated so that $\mathbf{x}$ is at the North Pole, $\mathbf{x} = (\sqrt{d} \pm O(1), 0, \ldots, 0)$. Since $\mathbf{y}$ is almost on the equator, further rotate the coordinate system so that the component of $\mathbf{y}$ that is perpendicular to the axis of the North Pole is in the second coordinate. Then $\mathbf{y} = (O(1), \sqrt{d} \pm O(1), 0, \ldots, 0)$. Thus,

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

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

Consider two spherical unit variance Gaussians with centers $\mathbf{p}$ and $\mathbf{q}$ separated by a distance Δ. The distance between a randomly chosen point $\mathbf{x}$ from the first Gaussian and a randomly chosen point $\mathbf{y}$ from the second is close to $\sqrt{\Delta^2 + 2d}$, since $\mathbf{x}-\mathbf{p}, \mathbf{p}-\mathbf{q}$, and $\mathbf{q}-\mathbf{y}$ are nearly mutually perpendicular. Pick $\mathbf{x}$ and rotate the coordinate system so that $\mathbf{x}$ is at the North Pole. Let $\mathbf{z}$ be the North Pole of the ball approximating the second Gaussian. Now pick $\mathbf{y}$. Most of the mass of the second Gaussian is within $O(1)$ of the equator perpendicular to $\mathbf{z}-\mathbf{q}$. Also, most of the mass of each Gaussian is within distance $O(1)$ of the respective equators perpendicular to the line $\mathbf{q}-\mathbf{p}$. See Figure 2.5 (b). Thus,

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

To ensure that the distance between two points picked from the same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of distance between points from different Gaussians. This requires that $\sqrt{2d}+O(1) \leq \sqrt{2d+\Delta^2}-O(1)$ or $2d+O(\sqrt{d}) \leq 2d+\Delta^2$, which holds when $\Delta \in \omega(d^{1/4})$. Thus, mixtures of spherical Gaussians can be separated in this way, provided their centers are separated by $\omega(d^{1/4})$. If we have n points and want to correctly separate all of them with high probability, we need our individual high-probability statements to hold with probability $1-1/poly(n)$,[3] which means our $O(1)$ terms from Theorem 2.9 become $O(\sqrt{\log n})$. So we need to include an extra $O(\sqrt{\log n})$ term in the separation distance.

> **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.

One can actually separate Gaussians where the centers are much closer. In the next chapter we will use singular value decomposition to separate points from a mixture of two Gaussians when their centers are separated by a distance $O(1)$.

2.9 Fitting a Spherical Gaussian to Data

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

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

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

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

Lemma 2.12 *Let* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ *be a set of* n *d-dimensional points. Then* $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ *is minimized when* $\boldsymbol{\mu}$ *is the centroid of the points* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$*, namely* $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$*.*

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

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

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

[3] poly(n) means bounded by a polynomial in n.

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

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

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

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

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

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

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

$$-a + \frac{n}{\nu}\,\frac{\int\limits_y y^2 e^{-y^2}dy}{\int\limits_y e^{-y^2}dy}.$$

Since the ratio of the two integrals is the expected distance squared of a d-dimensional spherical Gaussian of standard deviation $\frac{1}{\sqrt{2}}$ to its center, and this is known to be $\frac{d}{2}$, we get $-a + \frac{nd}{2\nu}$. Substituting σ^2 for $\frac{1}{2\nu}$ gives $-a + nd\sigma^2$. Setting $-a + nd\sigma^2 = 0$ shows that the maximum occurs when $\sigma = \frac{\sqrt{a}}{\sqrt{nd}}$. Note that this quantity is the square root of the average coordinate distance squared of the samples to their mean, which is the standard deviation of the sample. Thus, we get the following lemma.

Lemma 2.13 *The maximum likelihood spherical Gaussian for a set of samples is the Gaussian with center equal to the sample mean and standard deviation equal to the standard deviation of the sample from the true mean.*

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample of points generated by a Gaussian probability distribution. Then $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$ is an unbiased estimator of the expected value of the distribution. However, if in estimating the variance from the sample set, we use the estimate of the expected value rather than the true expected value, we will not get an unbiased estimate of the variance, since the sample mean is not independent of the sample set. One should use $\tilde{\boldsymbol{\mu}} = \frac{1}{n-1}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$ when estimating the variance. See Section 12.5.10 of the appendix.

2.10 Bibliographic Notes

The word vector model was introduced by Salton [SWY75]. 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. The Master Tail Bounds theorem and the derivation of Chernoff and other inequalities from it are from [Kan09]. The original proof of the Random Projection Theorem by Johnson and Lindenstrauss was complicated. Several authors used Gaussians to simplify the proof. The proof here is due to Dasgupta and Gupta [DG99]. See [Vem04] for details and applications of the theorem. [MU05] and [MR95b] are text books covering much of the material touched upon here.

2.11 Exercises

Exercise 2.1

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?*

Exercise 2.2 *Randomly generate 30 points inside the cube $[-\frac{1}{2},\frac{1}{2}]^{100}$ and plot distance between points and the angle between the vectors from the origin to the points for all pairs of points.*

Exercise 2.3 *Show that for any $a \geq 1$ there exist distributions for which Markov's inequality is tight by showing the following:*

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

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

Exercise 2.4 *Show that for any $c \geq 1$ there exist distributions for which Chebyshev's inequality is tight, in other words, $Prob(|x - E(x)| \geq c) = Var(x)/c^2$.*

Exercise 2.5 *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 $x \geq 3$.*

2. *Make use of $Prob(|x| \geq a) = Prob(x^2 \geq a^2)$ to get a tighter bound.*

3. *What is the bound using $Prob(|x| \geq a) = Prob(x^r \geq a^r)$?*

Exercise 2.6 *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 a as a function of a for the bound given by Markov's inequality and by Markov's inequality applied to x^2 and x^4.*

Exercise 2.7 *Consider the probability density function $p(x) = 0$ for $x < 1$ and $p(x) = c\frac{1}{x^4}$ for $x \geq 1$.*

1. *What should c be to make p a legal probability density function?*

2. *Generate 100 random samples from this distribution. How close is the average of the samples to the expected value of x?*

Exercise 2.8 *Let G be a d-dimensional spherical Gaussian with variance $\frac{1}{2}$ in each direction, centered at the origin. Derive the expected squared distance to the origin.*

Exercise 2.9 *Consider drawing a random point $\mathbf{x}$ on the surface of the unit sphere in R^d. What is the variance of x_1 (the first coordinate of $\mathbf{x}$)? See if you can give an argument without doing any integrals.*

Exercise 2.10 *How large must ε be for 99% of the volume of a 1000-dimensional unit-radius ball to lie in the shell of ε-thickness at the surface of the ball?*

Exercise 2.11 *Prove that $1 + x \le e^x$ for all real x. For what values of x is the approximation $1 + x \approx e^x$ within 0.01?*

Exercise 2.12 *For what value of d does the volume, $V(d)$, of a d-dimensional unit ball take on its maximum? Hint: Consider the ratio $\frac{V(d)}{V(d-1)}$.*

Exercise 2.13 *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.*

1. *For $0 \le k \le d$, how many k-dimensional faces does a d-dimensional cube have?*
2. *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.*
3. *What is the surface area of a unit cube in d-dimensions (a unit cube has side-length one in each dimension)?*
4. *What is the surface area of the cube if the length of each side was 2?*
5. *Prove that the volume of a unit cube is close to its surface.*

Exercise 2.14 *Consider the portion of the surface area of a unit radius, 3-dimensional ball 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 ball 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 $36°$? The angle of the cone is measured from the axis of the cone to a ray on the surface of the cone.*

Exercise 2.15 *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 ball. 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 ball 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 ball and $A(d)$ for the surface area of a unit radius, d-dimensional ball. An infinite length, unit radius, circular cylinder in 4-dimensions would be the set $\{(x_1, x_2, x_3, x_4) | x_2^2 + x_3^2 + x_4^2 \le 1\}$ where the coordinate x_1 is the axis.*

Exercise 2.16 *Given a d-dimensional circular cylinder of radius r and height h*

1. *What is the surface area in terms of $V(d)$ and $A(d)$?*

2. *What is the volume?*

Exercise 2.17 *How does the volume of a ball 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 ball of radius r to have approximately constant volume as the dimension increases? Hint: you may want to use Stirling's approximation, $n! \approx \left(\frac{n}{e}\right)^n$, for factorial.*

Exercise 2.18 *If $\lim\limits_{d\to\infty} V(d) = 0$, the volume of a d-dimensional ball for sufficiently large d must be less than $V(3)$. How can this be if the d-dimensional ball contains the three dimensional ball?*

Exercise 2.19

1. *Write a recurrence relation for $V(d)$ in terms of $V(d-1)$ by integrating over x_1. Hint: At $x_1 = t$, the $(d-1)$-dimensional volume of the slice is the volume of a $(d-1)$-dimensional sphere of radius $\sqrt{1-t^2}$. Express this in terms of $V(d-1)$ and write down the integral. You need not evaluate the integral.*

2. *Verify the formula for $d = 2$ and $d = 3$ by integrating and comparing with $V(2) = \pi$ and $V(3) = \frac{4}{3}\pi$*

Exercise 2.20 *Consider a unit ball A centered at the origin and a unit ball B whose center is at distance s from the origin. Suppose that a random point x is drawn from the mixture distribution: "with probability $1/2$, draw at random from A; with probability $1/2$, draw at random from B". Show that a separation $s \gg 1/\sqrt{d-1}$ is sufficient so that Prob$(x \in A \cap B) = o(1)$; i.e., for any $\epsilon > 0$ there exists c such that if $s \geq c/\sqrt{d-1}$, then Prob$(x \in A \cap B) < \epsilon$. In other words, this extent of separation means that nearly all of the mixture distribution is identifiable.*

Exercise 2.21 *Consider the upper hemisphere of a unit-radius ball in d-dimensions. What is the height of the maximum volume cylinder that can be placed entirely inside the hemisphere? As you increase the height of the cylinder, you need to reduce the cylinder's radius so that it will lie entirely within the hemisphere.*

Exercise 2.22 *What is the volume of the maximum size d-dimensional hypercube that can be placed entirely inside a unit radius d-dimensional ball?*

Exercise 2.23 *Calculate the ratio of area above the plane $x_1 = \epsilon$ to the area of the upper hemisphere of a unit radius ball in d-dimensions for $\epsilon = 0.001, 0.01, 0.02, 0.03, 0.04, 0.05$ and for $d = 100$ and $d = 1,000$.*

Exercise 2.24 *Almost all of the volume of a ball in high dimensions lies in a narrow slice of the ball at the equator. However, the narrow slice is determined by the point on the surface of the ball 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.*

Exercise 2.25 *Explain how the volume of a ball 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 ball.*

Exercise 2.26 *Generate 500 points uniformly at random on the surface of a unit-radius ball in 50 dimensions. Then randomly generate five additional points. For each of the five new points, calculate a narrow band of width* $\frac{2}{\sqrt{50}}$ *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?*

Exercise 2.27 *Place* 100 *points at random on a* d*-dimensional unit-radius ball. 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 close can the hyperplanes be moved and still have at least a .99 probability that all of the* 100 *points land between them?*

Exercise 2.28 *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be* d*-dimensional zero mean, unit variance Gaussian vectors. Prove that* $\mathbf{x}$ *and* $\mathbf{y}$ *are almost orthogonal by considering their dot product.*

Exercise 2.29 *Prove that with high probability, the angle between two random vectors in a high-dimensional space is at least* 45°. *Hint: use Theorem 2.8.*

Exercise 2.30 *Project the volume of a* d*-dimensional ball of radius* $\sqrt{d}$ *onto a line through the center. For large* d*, give an intuitive argument that the projected volume should behave like a Gaussian.*

Exercise 2.31

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

Exercise 2.32 *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?*

Exercise 2.33 *Randomly generate a 100 points on the surface of a sphere in 3-dimensions and in 100-dimensions. Create a histogram of all distances between the pairs of points in both cases.*

Exercise 2.34 *We have claimed that a randomly generated point on a ball lies near the equator of the ball, independent of the point picked to be the North Pole. Is the same claim true for a randomly generated point on a cube? To test this claim, randomly generate ten* ±1 *valued vectors in 128 dimensions. Think of these ten vectors as ten choices for the North Pole. Then generate some additional* ±1 *valued vectors. To how many of the original vectors is each of the new vectors close to being perpendicular; that is, how many of the equators is each new vector close to?*

Exercise 2.35 *Define the equator of a d-dimensional unit cube to be the hyperplane* $\left\{\mathbf{x}\middle|\sum\limits_{i=1}^{d} x_i = \frac{d}{2}\right\}$.

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?*

Exercise 2.36 *Consider a nonorthogonal basis* $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_d}$. *The* $\mathbf{e_i}$ *are a set of linearly independent unit vectors that span the space.*

1. *Prove that the representation of any vector in this basis is unique.*

2. *Calculate the squared length of* $\mathbf{z} = (\frac{\sqrt{2}}{2}, 1)_e$ *where* $\mathbf{z}$ *is expressed in the basis* $\mathbf{e_1} = (1, 0)$ *and* $\mathbf{e_2} = (-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$

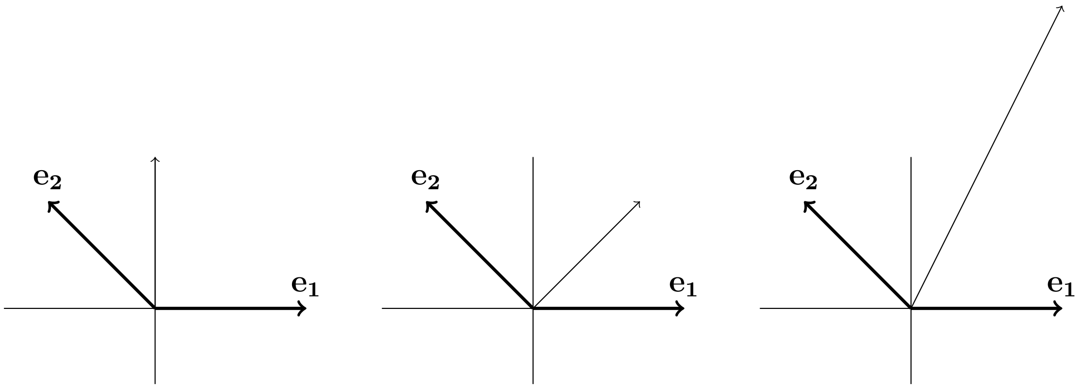
3. *If* $\mathbf{y} = \sum_i a_i\mathbf{e_i}$ *and* $\mathbf{z} = \sum_i b_i\mathbf{e_i}$, *with* $0 < a_i < b_i$, *is it necessarily true that the length of* $\mathbf{z}$ *is greater than the length of* $\mathbf{y}$*? Why or why not?*

4. *Consider the basis* $\mathbf{e_1} = (1, 0)$ *and* $\mathbf{e_2} = (-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$.

   (a) *What is the representation of the vector (0,1) in the basis* $(\mathbf{e_1}, \mathbf{e_2})$.

   (b) *What is the representation of the vector* $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$?

   (c) *What is the representation of the vector* $(1, 2)$?



Exercise 2.37 *Generate 20 points uniformly at random on a 900-dimensional sphere of radius 30. Calculate the distance between each pair of points. Then, select a method of projection and project the data onto subspaces of dimension k=100, 50, 10, 5, 4, 3, 2, 1 and calculate the difference between $\sqrt{k}$ times the original distances and the new pair-wise distances. For each value of k what is the maximum difference as a percent of $\sqrt{k}$.*

Exercise 2.38 *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 1000 almost orthogonal vectors in 100 dimensions. Here are two ways you could do it:*

1. *Begin with 1,000 orthonormal 1,000-dimensional vectors, and then project them to a random 100-dimensional space.*

2. *Generate 1000 100-dimensional random Gaussian vectors.*

Implement both ideas and compare them to see which does a better job.

Exercise 2.39 *Suppose there is an object moving at constant velocity along a straight line. You receive the gps coordinates corrupted by Gaussian noise every minute. How do you estimate the current position?*

Exercise 2.40

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

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

Exercise 2.41 *Let $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ be independent samples of a random variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and variance σ^2. Let $\mathbf{m_s} = \frac{1}{n}\sum\limits_{i=1}^{n} \mathbf{x_i}$ be the sample mean. Suppose one estimates the variance using the sample mean rather than the true mean, that is,*

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

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 $var(\mathbf{m_s}) = \frac{1}{n}var(\mathbf{x})$. Then calculate $E(\sigma_s^2) = E[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i}-\mathbf{m_s})^2]$ by replacing $\mathbf{x_i}-\mathbf{m_s}$ with $(\mathbf{x_i}-\mathbf{m})-(\mathbf{m_s}-\mathbf{m})$.

Exercise 2.42 *Generate ten values by a Gaussian probability distribution with zero mean and variance one. What is the center determined by averaging the points? What is the variance? In estimating the variance, use both the real center and the estimated center. When using the estimated center to estimate the variance, use both $n = 10$ and $n = 9$. How do the three estimates compare?*

Exercise 2.43 *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 $\mathbf{m}_s$ of the true center $\boldsymbol{\mu}$, so that with probability at least 99%,*

$$\|\boldsymbol{\mu} - \mathbf{m}_s\|_\infty \leq \boldsymbol{\varepsilon}.$$

How many samples are sufficient to ensure that with probability at least 99%

$$\|\boldsymbol{\mu} - \mathbf{m}_s\|_2 \leq \boldsymbol{\varepsilon}?$$

Exercise 2.44 *Use the probability distribution $\frac{1}{3\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-5)^2}{9}}$ to generate ten points.*

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

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

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

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

Exercise 2.45 *Create a list of the five most important things that you learned about high dimensions.*

Exercise 2.46 *Write a short essay whose purpose is to excite a college freshman to learn about high dimensions.*