Contents

2 High-Dimensional Space

2.1 Introduction

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* [SWY75] (also called the *Bag of Words* model) is a good example. In this 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. That is, all linguistic structure in the text is ignored, and the document is just viewed as a "bag of words", arranged in a vector where component i is the number of occurences of the ith word. 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 2.1. 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 of two vectors, or the cosine of the angle between them (which is the dot product normalized by the lengths of 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. While it is by no means clear a priori 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 (perhaps normalized so that all columns have Euclidean length 1). 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 $\mathbf{u}$, for which the sum of squared perpendicular distances of all the vectors to $\mathbf{u}$ is minimized. See Figure 2.2. Then, one ranks the vectors

Figure 2.1: A document and its term-document vector along with a collection of documents represented by their term-document vectors.

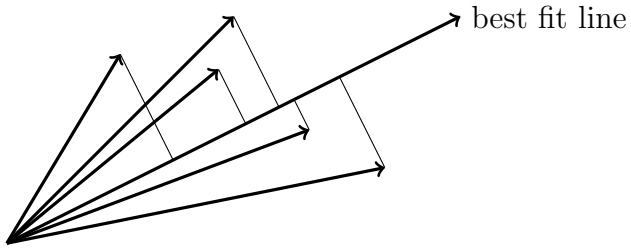

Figure 2.2: The best fit line is the line that minimizes the sum of the squared perpendicular distances.

according to their dot product with $\mathbf{u}$. 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 (so here we are not normalizing them to all have the same length) is the number of

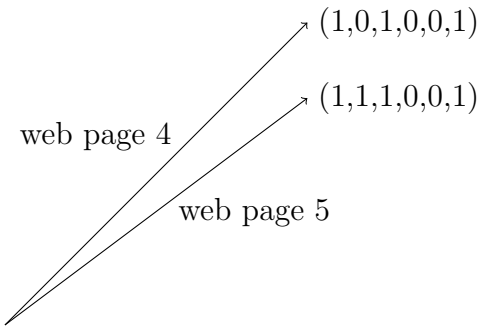

Figure 2.3: Two web pages as vectors. The squared distance between the two vectors is the number of web pages linked to by just one of the two web pages.

web pages linked to by only one of the pages. In Figure 2.3, 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.

2.1.1 A simple algorithm

[[A clean algo and intro to thinking in high dim space. A good ex to do on the board: ((1,0) +), ((1,1) +), ((0,1) -).]]

A common task to perform with data is what is called *binary classification*: given a new document, classify it as interesting or not interesting for a user, or given an email message, classify it as spam or not spam. We will discuss algorithms and basic principles for using data to find good classification rules in Chapter **??** on Machine Learning. Here, we present a simple classic algorithm for vector data called the *Perceptron Algorithm* [Blo62, Nov62, MP69].

Assume we have a collection of data points $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ (e.g., documents or email messages) in d-dimensional space, each labeled as *positive* (interesting) or *negative* (not interesting, spam). In machine learning, these would be called *training examples*. Our goal is to find a weight vector $\mathbf{w}$ such that $\mathbf{w} \cdot \mathbf{x_i} > 0$ for all the positive training examples $\mathbf{x_i}$ and $\mathbf{w} \cdot \mathbf{x_i} < 0$ for all the negative training examples $\mathbf{x_i}$, if such $\mathbf{w}$ exists. Let's assume such a weight vector $\mathbf{w}^*$ indeed exists, and without loss of generality (since we are using a threshold of 0) we may assume that $\mathbf{w}^*$ has Euclidean length 1. Let's also assume that all of our data points have been scaled to have Euclidean length 1 (which will also not affect the sign of the dot product) so they all live on the surface of the unit sphere. Define

$$\gamma = \min_i |\mathbf{w}^* \cdot \mathbf{x_i}|.$$

This is the minimum distance between any data point and the hyperplane $\mathbf{w}^* \cdot \mathbf{x} = 0$ (see Figure 2.4). Note that we are not given $\mathbf{w}^*$! This is what we are trying to find. In particular, we will show that the following algorithm finds a consistent weight vector $\mathbf{w}$ after at most $1/\gamma^2$ updates.

The Perceptron Algorithm:

1. Start with the all-zeroes weight vector $\mathbf{w}_0 = \mathbf{0}$, and initialize $t = 0$.
2. While there exists a training example $\mathbf{x_i}$ for which $\mathbf{w}_t$ is incorrect (i.e., either $\mathbf{x_i}$ is positive and yet $\mathbf{w}_t \cdot \mathbf{x_i} \leq 0$, or $\mathbf{x_i}$ is negative but $\mathbf{w}_t \cdot \mathbf{x_i} \geq 0$; if there are multiple such training examples, choose one arbitrarily) do:
   - If $\mathbf{x_i}$ is positive, let $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{x_i}$.
   - If $\mathbf{x_i}$ is negative, let $\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{x_i}$.
   - $t = t + 1$.

The algorithm should seem at least somewhat reasonable: when we make an update

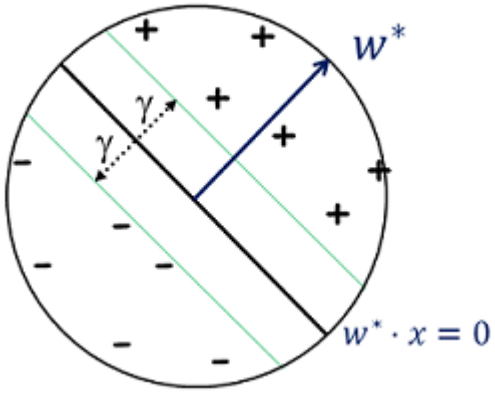

Figure 2.4: Positive and negative examples separated by a gap of γ (γ is often called the "margin" of the data). The vector w^* is orthogonal to the hyperplane $w^* \cdot x = 0$.

on some $\mathbf{x_i}$, the new vector $\mathbf{w}_{t+1}$ has a "better" dot product with $\mathbf{x_i}$ than $\mathbf{w}_t$ did. In particular, if $\mathbf{x_i}$ is positive then $\mathbf{w}_{t+1} \cdot \mathbf{x_i} = (\mathbf{w}_t + \mathbf{x_i}) \cdot \mathbf{x_i} = \mathbf{w}_t \cdot \mathbf{x_i} + 1$ and similarly if $\mathbf{x_i}$ is negative then $\mathbf{w}_{t+1} \cdot \mathbf{x_i} = (\mathbf{w}_t - \mathbf{x_i}) \cdot \mathbf{x_i} = \mathbf{w}_t \cdot \mathbf{x_i} - 1$. Of course, this doesn't by itself prove anything since we might have hurt our dot product with other training examples. The correctness of the algorithm is given by the following theorem.

Theorem 2.1 *The Perceptron algorithm finds a consistent weight vector after at most* $1/\gamma^2$ *updates.*

Proof: We're going to look at the two quantities $\mathbf{w}_t \cdot \mathbf{w}^*$ and $|\mathbf{w}_t|$, and prove the theorem via two claims:

Claim 1: $\mathbf{w}_{t+1} \cdot \mathbf{w}^* \geq \mathbf{w}_t \cdot \mathbf{w}^* + \gamma$. That is, every time we make an update, the dot-product of our weight vector with the target increases by at least γ.

> Proof: if $\mathbf{x_i}$ was a positive example, then we get $\mathbf{w}_{t+1} \cdot \mathbf{w}^* = (\mathbf{w}_t + \mathbf{x_i}) \cdot \mathbf{w}^* = \mathbf{w}_t \cdot \mathbf{w}^* + \mathbf{x_i} \cdot \mathbf{w}^* \geq \mathbf{w}_t \cdot \mathbf{w}^* + \gamma$ (by definition of γ). Similarly, if $\mathbf{x_i}$ was a negative example, we get $(\mathbf{w}_t - \mathbf{x_i}) \cdot \mathbf{w}^* = \mathbf{w}_t \cdot \mathbf{w}^* - \mathbf{x_i} \cdot \mathbf{w}^* \geq \mathbf{w}_t \cdot \mathbf{w}^* + \gamma$.

Claim 2: $|\mathbf{w}_{t+1}|^2 \leq |\mathbf{w}_t|^2 + 1$. That is, every time we make an update, the length squared of our weight vector increases by at most 1.

> Proof: if $\mathbf{x_i}$ was a positive example, we get $|\mathbf{w}_t + \mathbf{x_i}|^2 = |\mathbf{w}_t|^2 + 2\mathbf{w}_t \cdot \mathbf{x_i} + |\mathbf{x_i}|^2$. This is less than $|\mathbf{w}_t|^2 + 1$ because $\mathbf{w}_t \cdot \mathbf{x_i}$ is negative (remember, we made a mistake on $\mathbf{x_i}$) and $|\mathbf{x_i}| = 1$. The exact same thing holds (flipping signs) if $\mathbf{x_i}$ was negative but we predicted positive.

Claim 1 implies that after T updates, $\mathbf{w}_{T+1} \cdot \mathbf{w}^* \geq \gamma T$. On the other hand, Claim 2 implies that after T updates, $|\mathbf{w}_{T+1}| \leq \sqrt{T}$. Now, all we need to do is use the fact that $\mathbf{w}_t \cdot \mathbf{w}^* \leq |\mathbf{w}_t|$, since $\mathbf{w}^*$ is a unit vector. So, this means we must have $\gamma T \leq \sqrt{T}$, and thus $T \leq 1/\gamma^2$. ∎

Of course, we have not explained why we should have any reason to believe that a vector $\mathbf{w}$ that correctly classifies a *sample* of data should necessarily do well on *new* data. For that and related topics, see Chapter **??**.

2.1.2 Structure of this chapter

Our aim in the rest of this chapter is to present the reader with some of the mathematical foundations to deal with high-dimensional data. There are two important parts of this foundation. The first is high-dimensional geometry, and the second more modern aspect is the combination with probability.

We begin by focusing on the *unit ball* in d dimensions, that is, the set of all points within distance 1 of the origin. The geometry of high-dimensional space is quite different from our intuitive understanding of two and three dimensions. For example, we will see that nearly all the volume of the ball is concentrated near its equator (no matter which direction we call "north"), and at the same time, nearly all its volume is in a narrow annulus near its boundary. Moreover, if you consider a ball and its enclosing cube, the volume of the ball is a vanishingly small fraction of the volume of the cube as the dimension becomes large.

Following that, we will study a fundamental probability distribution in d dimensions, the spherical Gaussian. We will use our understanding of the Gaussian distribution to analyze a powerful algorithmic tool for high dimensional problems: random projection and the Johnson-Lindenstrauss Lemma. For many applications, this technique can be used to convert a high-dimensional problem into a low-dimensional (and often therefore easier to solve) version of the same problem. In the process of this analysis, we will derive tail inequalities which are an important analytical tool for a number of problems we will study throughout this book. In the context of Gaussians, we will use tail inequalities to analyze Gaussian mixture models, distributions that often arise in analyzing "big data" problems.

Chapter **??** contains additional background on probability theory.

2.2 The Geometry of High Dimensions

We begin our discussion of high-dimensional geometry by discussing an important property of high-dimensional objects, that most of their volume is near their surface.

Specifically, consider any object A in R^d. If we shrink it by a factor γ to produce a new object γA (formally, $\gamma A = \{\gamma x : x \in A\}$) then volume$(\gamma A) = \gamma^d$volume$(A)$. We can see why this is true by partitioning A into infinitesimal cubes of side-length dx, and noticing that because this fact holds true for a cube, it also holds true for a union of disjoint cubes. Suppose we now set $\gamma = 1 - \epsilon$ for some small value ϵ. Using the fact that

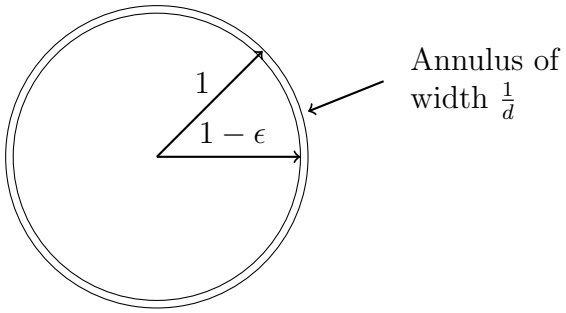

Figure 2.5: Most of the volume of the d-dimensional ball of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

$1 - x \leq e^{-x}$ we have that for any object A in R^d,

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

Fixing ϵ and letting $d \to \infty$, the above quantity rapidly approaches 0. This means that nearly all of the volume of A must be in points x such that $x \notin (1-\epsilon)A$.

Let S denote the unit ball in d dimensions, that is, the set of points within distance 1 of the origin. An immediate implication of the above 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 similarly the annulus width is $O\left(\frac{r}{d}\right)$.

2.3 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. We now will show a more subtle fact that additionally most of its volume is concentrated near its equator (no matter what direction we use to define "equator"). Specifically, letting x_1 (arbitrarily) denote "north", we will show that most of the volume of the unit ball has $|x_1| = O(1/\sqrt{d})$. Using this fact, we will then show that two random points in the unit ball are with high probability nearly orthogonal, and also that the volume of the unit ball goes to 0 as $d \to \infty$.

Theorem 2.2 *A $1 - O(e^{-\gamma^2/2})$ fraction of the volume of the unit ball has $|x_1| \leq \frac{\gamma}{\sqrt{d-1}}$.*

Proof: If we consider a slice of the ball at $x_1 = a$ of width δ, the volume of this slice is δ times the $(d-1)$-dimensional volume, or "area", of the cross-section, in the limit as $\delta \to 0$ (so we can think of all cross-sections in the slice as having the same radius). This cross-section is just a $(d-1)$-dimensional ball of radius $\sqrt{1-a^2}$. Therefore, applying

the basic facts about volume discussed in Section 2.2, its $(d-1)$-dimensional volume is exactly $(\sqrt{1-a^2})^{d-1}V_{d-1}$, where V_{d-1} denotes the $(d-1)$-dimensional volume of the $(d-1)$-dimensional ball of radius 1. Using the fact that $1-x \leq e^{-x}$ we get:

$$(\sqrt{1-a^2})^{d-1}V_{d-1} \;=\; (1-a^2)^{\frac{d-1}{2}}V_{d-1} \;\leq\; e^{-(a^2/2)(d-1)}V_{d-1}.$$

In the other direction, using the fact that $(1-\frac{1}{d-1})^{\frac{d-1}{2}} \geq 1/e$ for $d \geq 3$, we get:

$$\left(\sqrt{1-(\tfrac{1}{\sqrt{d-1}})^2}\right)^{d-1} V_{d-1} \;\geq\; e^{-1}V_{d-1}.$$

Now, we can make a few observations. First, the cross-sections for $x_1 \in [\frac{-1}{\sqrt{d-1}}, \frac{1}{\sqrt{d-1}}]$ have "area" between V_{d-1} and $e^{-1}V_{d-1}$, so this region has volume at least $\frac{2e^{-1}}{\sqrt{d-1}}V_{d-1}$. On the other hand, for $\gamma \geq 1$, the cross-sections for $x_1 \in [\frac{\gamma}{\sqrt{d-1}}, \frac{\gamma+1}{\sqrt{d-1}}]$ have "area" at most $e^{-\gamma^2/2}V_{d-1}$, so this region has volume *at most* $\frac{1}{\sqrt{d-1}}e^{-\gamma^2/2}V_{d-1}$. Furthermore, if we add 1 to γ, this quantity drops by at least a factor $1/e$, meaning that if we sum up the volumes for regions defined by $\gamma, \gamma+1, \gamma+2, \ldots$ we get a telescoping series whose total volume is $O(\frac{1}{\sqrt{d-1}}e^{-\gamma^2/2}V_{d-1})$. Thus, the *fraction* of volume of the unit ball with $|x_1| \geq \frac{\gamma}{\sqrt{d-1}}$ is only $O(e^{-\gamma^2/2})$. ∎

One immediate implication of the above analysis is that if we draw two points at random from the unit ball, with high probability they (their vectors) will be nearly orthogonal to each other. Specifically, from our previous analysis in Section 2.2, we know with high probability both will have length $1-O(1/d)$. From our analysis above, we know that 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. 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 theorem:

Theorem 2.3 *Consider drawing n points $\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_n}$ at random from the unit ball. With probability $1-O(1/n)$ we have both:*

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

Proof: For any fixed i, the bound on $|\mathbf{z_i}|$ holds with probability at least $1-1/n^2$ by the analysis of Section 2.2 and so it holds for all i with probability at least $1-1/n$. For the second part, we have $\binom{n}{2}$ pairs i, j, and for each such pair, if we define $\mathbf{z_i}$ as "north", the probability that the projection of $\mathbf{z_j}$ onto that direction is more than $\frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ (a necessary condition for the dot-product to be large) is at most $O(e^{-\frac{6\ln n}{2}}) = O(n^{-3})$ by Theorem 2.2. Thus, this condition is violated with probability at most $O(\binom{n}{2}n^{-3}) = O(1/n)$ as well. ∎

Another immediate implication of the above analysis is that as $d \to \infty$, the volume of the ball approaches 0. Specifically, setting $\gamma = 2\sqrt{\ln d}$ above we have that at most an $O(1/d^2)$ fraction of the volume of the ball has $|x_1| \geq \frac{\gamma}{\sqrt{d-1}}$. Since this is true for each of the d dimensions, we have that at least a $1 - O(\frac{1}{d}) \geq \frac{1}{2}$ fraction of the volume of the ball lies in a cube of side-length $2\frac{\gamma}{\sqrt{d-1}}$. This cube has volume of the form $(\frac{c \ln d}{d-1})^{d/2}$ for constant c, and this quantity goes to 0 as $d \to \infty$. Since the ball has volume at most twice that of this cube, its volume goes to 0 as well.

2.4 Generating Points Uniformly at Random from a Ball

How can we generate points uniformly at random from the unit ball? First, let's 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, here is one approach. 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, unfortunately only an exponentially small fraction of the cube lies inside the unit ball, rapidly making this process impractical. 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, $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ on the real line.[1] This is called $N(0, 1)$. So the probability density of $\mathbf{x}$ is

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

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

Now to generate a point $\mathbf{y}$ uniformly over the ball (surface and interior), we have to scale the point $\frac{\mathbf{x}}{|\mathbf{x}|}$ generated on the surface by a scalar $\rho \in [0, 1]$. What is the distribution

[1] One might naturally ask: "how do you generate a random number from a 1-dimensional Gaussian?" A general method to generate a number from any distribution given its CDF P is to first select a uniform random number $u \in [0, 1]$ and then choose $x = P^{-1}(u)$; that is because the probability this generates a number between x and $x + \delta$ is $P(x + \delta) - P(x) = \int_x^{x+\delta} p(x')dx'$ 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.

of ρ ? It is certainly not uniform, even in 2 dimensions. Indeed, the density of ρ 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 (**??**). By similar reasoning, it is easy to see that in d dimensions, the density of ρ at distance r is proportional to r^{d-1}. Solving $\int_{r=0}^{r=1} cr^{d-1}dr = 1$ (the integral of density must equal 1) we see we should set $c = d$. Another way to see this formally is that we know the volume of the radius-r ball in d dimensions is $r^d V_d$, where V_d is the volume of the unit ball. The density *at* radius r is exactly $\frac{d}{dr}(r^d V_d) = dr^{d-1}V_d$. So, pick ρ with density (of $\rho = r$) equal to dr^{d-1} over $[0, 1]$.

Now we have succeeded in generating a point

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

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

2.5 The Law of Large Numbers

If we draw points at random from the d-dimensional spherical Gaussian for large d, we will find that they are all essentially the same distance apart. The reason is that if one averages n independent samples $x_1, x_2, \ldots, x_n$ of a random variable x of bounded variance, the result will be close to the expected value of x. In our case, we can think of x_i as representing the squared distance between two points in coordinate i, and $n = d$, so that the sum of the x_i is the overall squared distance between the two points. Later in Theorem 2.11 we will give tight concentration bounds of this form. For now, we will give a less tight but more general bound called the Law of Large Numbers. Specifically, the Law of Large Numbers states:

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

Here the σ^2 in the numerator is the variance of x. The larger the variance of the random variable, the greater the probability that the error will exceed ϵ. The number of points 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 dimensionalless quantity.

To prove the law of large numbers we use two inequalities. The first is Markov's inequality. One can bound the probability that a nonnegative random variable exceeds a by the expected value of the variable divided by a.

Theorem 2.4 (Markov's inequality) *Let x be a nonnegative random variable. Then for $a > 0$,*

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

Proof: The proof is easiest to see if multiply both sides of the inequality by a, producing the statement $E(x) \geq a \cdot \text{Prob}(x \geq a)$. This now follows directly from the definition of expectation. Specifically, for a continuous random variable x with density p, we have:

$$E(x) = \int_0^\infty xp(x)dx \geq \int_a^\infty xp(x)dx \geq a \int_a^\infty p(x)dx = a\text{Prob}(x \geq a)$$

For a discrete nonnegative random variable, the same proof applies:

$$E(x) = \sum_{v \geq 0} v\text{Prob}(x = v) \geq a\text{Prob}(x \geq a). \quad \blacksquare$$

Corollary 2.5 *Prob* $(x \geq cE(x)) \leq \frac{1}{c}$

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.

Theorem 2.6 (Chebyshev's inequality) *Let x be a random variable with mean m and variance σ^2. Then*

$$Prob(|x - m| \geq a\sigma) \leq \frac{1}{a^2}.$$

Proof: $\text{Prob}(|x - m| \geq a\sigma) = \text{Prob}\left((x - m)^2 \geq a^2\sigma^2\right)$. Note that $(x - m)^2$ is a nonnegative random variable, so Markov's inequality can be applied giving:

$$\text{Prob}\left((x - m)^2 \geq a^2\sigma^2\right) \leq \frac{E\left((x - m)^2\right)}{a^2\sigma^2} = \frac{\sigma^2}{a^2\sigma^2} = \frac{1}{a^2}.$$

Thus, $\text{Prob}\left(|x - m| \geq a\sigma\right) \leq \frac{1}{a^2}$. $\blacksquare$

The law of large numbers follows from Chebyshev's inequality. Recall that $E(x+y) = E(x) + E(y)$, $\sigma^2(cx) = c^2\sigma^2(x)$, $\sigma^2(x - m) = \sigma^2(x)$, and if x and y are independent, then $E(xy) = E(x)E(y)$ and $\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y)$. To prove $\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y)$ when x and y are independent, since $\sigma^2(x - m) = \sigma^2(x)$, one can assume $E(x) = 0$ and $E(y) = 0$. Thus,

$$\begin{aligned}\sigma^2(x + y) &= E\left((x + y)^2\right) = E(x^2) + E(y^2) + 2E(xy) \\ &= E(x^2) + E(y^2) + 2E(x)E(y) = \sigma^2(x) + \sigma^2(y).\end{aligned}$$

Replacing $E(xy)$ by $E(x)E(y)$ required independence.

Theorem 2.7 (Law of large numbers) *Let $x_1, x_2, \ldots, x_n$ be n samples of a random variable x. Then*

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

Proof: By Chebychev's inequality

$$\begin{aligned}\text{Prob}\left(\left|\frac{x_1+x_2+\cdots+x_n}{n}-E(x)\right|>\epsilon\right) &\le \frac{\sigma^2\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)}{\epsilon^2}\\ &\le \frac{1}{n^2\epsilon^2}\sigma^2(x_1+x_2+\cdots+x_n)\\ &\le \frac{1}{n^2\epsilon^2}\left(\sigma^2(x_1)+\sigma^2(x_2)+\cdots+\sigma^2(x_n)\right)\\ &\le \frac{\sigma^2(x)}{n\epsilon^2}.\end{aligned}$$

■

The law of large numbers is quite general. In the sections below we will look at tighter concentration bounds for spherical Gaussians and sums of 0-1 valued random variables.

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

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

The value of the density is maximum at the origin, but there is very little volume there. When $\sigma = 1$, integrating the probability density over a unit ball centered at the origin yields nearly zero mass since the volume of such a ball is negligible. In fact, one needs to increase the radius of the ball 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 (2.9). But we will first use it in the next section. First, note that

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

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

[replace this with figure file]

Figure 2.6: Most of the probability mass of d dimensional Gaussian of radius r is contained in an annulus of width $O(r/\sqrt{d})$.

2.7 Random Projection and Johnson-Lindenstrauss Lemma

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 $\mathbf{R}^d$, 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 $\mathbf{R}^d$ 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 $k \ll d$ (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 (in fact, many related projections are known to work as well) is the following. Pick k vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$, independently from the Gaussian distribution $\frac{1}{(2\pi)^{d/2}} \exp(-|\mathbf{x}|^2/2)$. We then define for any vector $\mathbf{v}$, the projection $f(\mathbf{v})$ by:

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

So, $f(\mathbf{v})$ is just a vector of dot products of $\mathbf{v}$ with the $\mathbf{u_i}$. We will show that $|f(\mathbf{v})| \approx \sqrt{k}|\mathbf{v}|$, so if we have to find the distance $|\mathbf{v_1} - \mathbf{v_2}|$ between two vectors $\mathbf{v_1}, \mathbf{v_2}$ in $\mathbf{R}^d$, it will suffice instead 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 we can just divide by it).

Theorem 2.9 (The Random Projection Theorem) *Let $\mathbf{v}$ be a fixed vector in $\mathbf{R}^d$ and let f be defined as above. Then, for $\varepsilon \in (0, 1)$,*

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

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

Proof: By scaling both sides by $|\mathbf{v}|$, we may assume that $|\mathbf{v}| = 1$. The sum of independent normally distributed real variables is also normally distributed; the means and variances just sum up. Since $\mathbf{u_i} \cdot \mathbf{v} = \sum_{j=1}^{d} u_{ij} v_j$, we see that the random variable $\mathbf{u_i} \cdot \mathbf{v}$ has Gaussian density with mean 0 and variance equal to $\sum_{j=1}^{d} v_j^2 = |\mathbf{v}|^2 = 1$. Further, $\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}$ are independent. So the current theorem follows from the Gaussian annulus theorem (2.8). ∎

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, the projection to a random subspace 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 2.10 (Johnson-Lindenstrauss Lemma) *For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ for c as in Theorem 2.8. For any set P of n points in R^d, the random projection f $f : R^d \to R^k$ defined above has the property that for all $\mathbf{v_i}$, $\mathbf{v_j}$ in P, with probability at least $1 - (1.5/n)$,*

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

Proof: Applying the random projection theorem (Theorem 2.9), for any fixed $\mathbf{v_i}$ and $\mathbf{v_j}$, the probability that $|f(\mathbf{v_i}) - f(\mathbf{v_j})| = |f(\mathbf{v_i} - \mathbf{v_j})|$ is outside the range

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

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, by the union bound, the probability that some pair has a large distortion is less than $\frac{3}{2n}$. ∎

Remark: It is important to note that the conclusion of Theorem 2.10 asserts for all $\mathbf{v_i}$ and $\mathbf{v_j}$ in P, 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 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.

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, for k as in Theorem 2.10. 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 was useful here in making k only dependent on $\ln n$, rather than n.

2.8 Bounds on Tail Probability

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

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

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 $1/a^2$:

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

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 **Central Limit Theorem** for independent, identically distributed random variables $x_1, x_2, \cdots, x_n$ with zero mean and $\text{Var}(x_i) = \sigma^2$ 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 σ^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 σ^2. But this theorem is only in the limit, whereas we want (and will 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 σ^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 2.11 *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. Assume for $s = 3, 4, \ldots, \lfloor (a^2/4n\sigma^2) \rfloor$, $|E(x_i^s)| \leq \sigma^2 s!$, then for $0 \leq a \leq \sqrt{2}n\sigma^2$,*

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

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 $(x_1 + x_2 + \cdots + x_n)^r$.

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

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 $\binom{r-2t+t-1}{t-1} = \binom{r-t-1}{t-1}$. So,

$$|J_t| \leq \binom{n}{t}\binom{r-t-1}{t-1}$$

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

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 $s \leq \frac{a^2}{4n\sigma^2}$. Thus, r is at most $n\sigma^2/2$. For $t \leq r/2$, increasing t by one, increases $h(t)$ by at least $n\sigma^2/(2t)$, which is at least two. This gives

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

Applying Markov inequality,

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

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 $e \cdot e^{-a^2/(12n\sigma^2)} \leq 3 \cdot e^{-a^2/(12n\sigma^2)}$, proving the theorem. ∎

2.9 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 ball 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 2.11.

Theorem 2.11 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 $y^2 = 2z$,

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

The last inequality is from the Gamma integral.
$Ex_i = 0$ and so $\mathrm{Var}(x_i) = E(x_i^2) \leq 2^2 2 = 8$. But to make $|E(x_i^s)| \leq 8s!$ as required in theorem (2.11), we use $w_i = x_i/2$. Then, $\mathrm{Var}(w_i) = 2$ and $|E(w_i^s)| \leq 2s!$.

Proof: (of Theorem (2.8)) 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 $|y_1^2 + y_2^2 + \cdots + y_d^2 - d| \geq \beta\sqrt{d}$ and hence $|x_1 + x_2 + \cdots + x_d| \geq \beta\sqrt{d}$ or $|w_1 + w_2 + \cdots + w_d| \geq \frac{\beta\sqrt{d}}{2}$. Applying Theorem 2.11 where $\sigma^2 = 2$ and $n = d$, this occurs with probability less than or equal to $3e^{-\frac{\beta^2}{200}}$. ∎

Chernoff Bounds

Chernoff bounds deal with sums of Bernoulli random variables. Here we apply Theorem 2.11 to derive similar bounds. For this direct application, we will require $p < 1 - 1/\sqrt{2}$.

Theorem 2.12 *Suppose $y_1, y_2, \ldots, y_n$ are independent 0-1 random variables with $E(y_i) = p$ for all i, where $p < 1 - 1/\sqrt{2}$. Let $y = y_1 + y_2 + \cdots + y_n$. Then for any $c \in [0, 1]$,*

$$\mathit{Prob}\left(|y - E(y)| \geq cnp\right) \leq 3e^{-npc^2/8}.$$

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 $s \geq 3$,

$$\begin{aligned} |E(x_i^s)| &= |E(y_i - p)^s| \\ &= |p(1-p)^s + (1-p)(0-p)^s| \\ &= \left|p(1-p)\left((1-p)^{s-1} + p^{s-1}\right)\right| \\ &\leq p(1-p). \end{aligned}$$

Apply Theorem 2.11 with $a = cnp$. Noting that $1 - p > 1/\sqrt{2}$ so $a < \sqrt{2}\ np(1-p)$, 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

[replace this with figure file]

Figure 2.7: Zipf's Law: Number of words versus frequency.

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 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 2.13 *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 $a \leq \frac{1}{10}\sqrt{\frac{n}{k}}$,*

$$\textit{Prob}\left(|x| \geq a\right) \leq e^{-\frac{a^2}{8var(x)}}.$$

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 $s \leq k-2$,

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

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

$$\begin{aligned} E(x_i^s) =& (k-1)\int_1^\infty \frac{(y-\mu)^s}{y^k} dy \\ =& \frac{k-1}{\mu^{k-s-1}}\int_0^1 (1-z)^s z^{k-s-2} dz + \frac{k-1}{\mu^{k-s-1}}\int_1^\mu (1-z)^s z^{k-s-2} dz. \end{aligned}$$

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 $z \in [1, \mu]$, $|z-1| \leq \frac{1}{k-2}$ and

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

Apply Theorem 2.11 requires bounding $|E(x_i^s)|$ for $3 \leq s \leq \left\lfloor \frac{a^2}{4n\mathrm{Var}(x_i)} \right\rfloor$. Since $a \leq \frac{1}{10}\sqrt{\frac{n}{k}}$; it follows that

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

So it suffices to prove that $|E(x_i^s)| \leq s!\mathrm{Var}(x)$ 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 $\mu > 1$,

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

Now, the theorem follows from Theorem 2.11. ∎

2.10 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 $p(\mathbf{x}) = w_1 p_1(\mathbf{x}) + w_2 p_2(\mathbf{x})$, where $p_1(\mathbf{x})$ is a Gaussian density representing the typical heights of women, $p_2(\mathbf{x})$ 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).

Now of course there are taller women and shorter men, so even if one solved the parameter estimation problem for heights perfectly, given a data point (a height) one couldn't necessarily tell which population it came from (male or female). 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 in fact 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.

Consider two spherical unit-variance Gaussians. From Theorem 2.11, most of the probability mass of each Gaussian lies on an annulus of width $O(1)$ at radius $\sqrt{d-1}$. Also

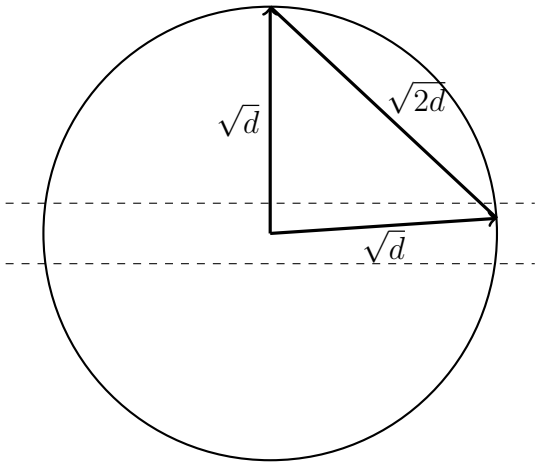

Figure 2.8: Two randomly chosen points in high dimension are almost surely nearly orthogonal.

$e^{-|\mathbf{x}|^2/2} = \prod_i e^{-x_i^2/2}$ 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 the first Gaussian. After picking $\mathbf{x}$, rotate the coordinate system to make the first axis point towards $\mathbf{x}$. Independently pick a second point $\mathbf{y}$ also from the first Gaussian. The fact that almost all of the mass of the Gaussian is within the slab $\{\mathbf{x} \mid -c \leq x_1 \leq c,\ c \in O(1)\}$ 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 2.8. 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 $|\mathbf{x} - \mathbf{y}| = \sqrt{2d} \pm O(1)$.

Given two spherical unit variance Gaussians with centers $\mathbf{p}$ and $\mathbf{q}$ separated by a distance δ, the distance between a randomly chosen point $\mathbf{x}$ from the first Gaussian and a randomly chosen point $\mathbf{y}$ from the second is close to $\sqrt{\delta^2 + 2d}$, since $\mathbf{x} - \mathbf{p}, \mathbf{p} - \mathbf{q}$, and $\mathbf{q} - \mathbf{y}$ 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 $\mathbf{q} - \mathbf{z}$. Also, most of the mass of each Gaussian is within distance $O(1)$ of the respective equators perpendicular to the line $\mathbf{q} - \mathbf{p}$. See Figure 2.9. Thus,

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

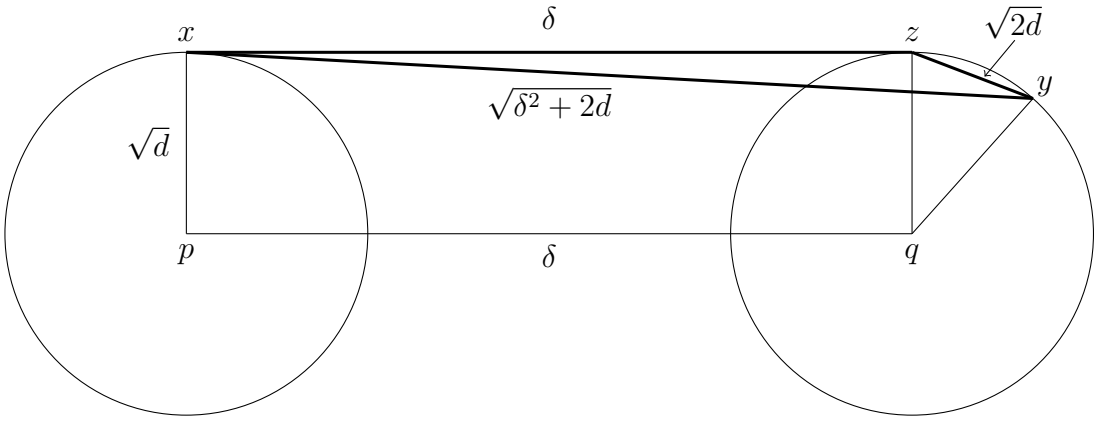

Figure 2.9: Distance between a pair of random points from two different unit balls approximating the annuli of two Gaussians.

To ensure that the distance between two points picked from the same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of distance between points from different Gaussians. This requires that $\sqrt{2d} + O(1) \leq \sqrt{2d + \delta^2} - O(1)$ or $2d + O(\sqrt{d}) \leq 2d + \delta^2$, which holds when $\delta \in \Omega(d^{1/4})$. Thus, mixtures of spherical Gaussians can be separated, provided their centers are separated by more than $d^{\frac{1}{4}}$. One can actually separate Gaussians where the centers are much closer. 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, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, 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 could 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 2.14 *Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n points in d-space. Then $(\mathbf{x}_1 - \boldsymbol{\mu})^2 + (\mathbf{x}_2 - \boldsymbol{\mu})^2 + \cdots + (\mathbf{x}_n - \boldsymbol{\mu})^2$ is minimized when $\boldsymbol{\mu}$ is the centroid of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, namely $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$.*

Proof: Setting the gradient of $(\mathbf{x}_1 - \boldsymbol{\mu})^2 + (\mathbf{x}_2 - \boldsymbol{\mu})^2 + \cdots + (\mathbf{x}_n - \boldsymbol{\mu})^2$ with respect $\boldsymbol{\mu}$ 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 $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)$. ∎

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

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

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

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

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

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

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

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

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

2.11 Bibliographic Notes

The word vector model was introduced by Salton [SWY75]. 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 [MU05] or [MR95b]. The proof here and the application to heavy-tailed distributions is simplified from [Kan09]. The original proof of the random projection theorem by Johnson and Lindenstrauss was complicated. Several authors used Gaussians to simplify the proof. See [Vem04] for details and applications of the theorem. The proof here is due to Dasgupta and Gupta [DG99].

2.12 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 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.*

Exercise 2.2 *Consider two random vectors in $\{0,1\}^d$ for large d. The angle between them will be concentrated around what value?*

Exercise 2.3 *The distance of a point to the center of a ball in d-dimensions is likely to be between $1-\frac{c}{d}$ and 1. Additionally, 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.*

Exercise 2.4 *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 Prob $\big(x \geq aE(x)\big) = \frac{1}{a}$.*

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

Exercise 2.5 *In what sense is Chebyshev's inequality tight?*

Exercise 2.6 *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?*

Exercise 2.7 *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 2.8 *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 2.9 *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.*

Exercise 2.10 *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?*

Exercise 2.11 *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.*

Exercise 2.12 *For $i \leq i \leq d$, how many i-dimensional faces does a d-dimensional tetrahedron have?*

Exercise 2.13 *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 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 $\{(x_1, x_2, x_3, x_4) | x_2^2 + x_3^2 + x_4^2 \leq 1\}$ where the coordinate x_1 is the axis.*

Exercise 2.14 *What is the surface area of a d-dimensional cylinder of radius two and height one in terms of $V(d)$ and $A(d)$?*

Exercise 2.15 *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 ball at each vertex. Each ball fits in a cube of width two and thus no two balls intersect. Show that the probability that a point of the cube picked at random will fall into one of the 2^d unit-radius balls, centered at the vertices of the cube, goes to 0 as d tends to infinity.*

Exercise 2.16 *Place two unit-radius balls in d-dimensions, one at (-2,0,0,. . . ,0) and the other at (2,0,0,. . . ,0). Give an upper bound on the probability that a random line through the origin will intersect the balls.*

Exercise 2.17 *Let* $\mathbf{x}$ *be a random sample from the unit ball* $\{\mathbf{x}|\ |\mathbf{x}| \leq 1\}$ *in d-dimensions with the origin as center.*

1. *What is the mean of the random variable* $\mathbf{x}$*? The mean, denoted* $E(\mathbf{x})$*, is the vector, whose* i^{th} *component is* $E(x_i)$*.*

2. *What is the component-wise variance of* $\mathbf{x}$*?*

3. *For any unit length vector* $\mathbf{u}$*, the variance of the real-valued random variable* $\mathbf{u}^T\mathbf{x}$ *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* $\mathbf{x}$*.*

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

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

times the volume of each ball. Hint: Relate the volume of the intersection to the volume of a cap; then, use Lemma **??***.*

5. *From (4), conclude that if the inter-center separation of the two balls 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 ball, it is possible to tell which ball contains which point, i.e., classify them into two clusters so that each cluster is exactly the set of points generated from one ball. The actual classification requires an efficient algorithm to achive this. Note that the inter-center separation required here goes to zero as d gets larger, provided the radius of the balls remains the same. So, it is easier to tell apart balls (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 balls as* $\int_{\mathbf{x}\in \text{ space}} \min(f(x), g(x))dx$*, where* f *and* g *are just the uniform densities in the two balls 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* σ *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.*

Exercise 2.18 *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?*

Exercise 2.19 *Derive an upper bound on* $\int_{x=a}^{\infty} e^{\frac{-x^2}{2}}\, dx$ *where a is a positive real. Discuss for what values of a this is a good bound.*
Hint: Use $e^{\frac{-x^2}{2}} \leq \frac{x}{a}e^{\frac{-x^2}{2}}$ *for* $x \geq a$*.*

Exercise 2.20 *Verify the formula* $V(d) = 2\int_0^1 V(d-1)(1-x_1^2)^{\frac{d-1}{2}}dx_1$ *for* $d = 1$ *and* $d = 2$ *by integrating and comparing with* $V(2) = \pi$ *and* $V(3) = \frac{4}{3}\pi$

Exercise 2.21 *What is the volume of a radius* r *cylinder of height* h *in* d*-dimensions?*

Exercise 2.22 *Consider the upper half 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 2.23 *What is the volume of the maximum size* d*-dimensional hypercube that can be placed entirely inside a unit radius* d*-dimensional ball?*

Exercise 2.24 *In showing that the volume of a unit ball 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* $\frac{r^{d+1}}{\epsilon(d-1)}e^{\frac{d-1}{2}\left(\frac{\epsilon}{r}\right)^2}V(d-1)$ *and* $\frac{r^d}{2\sqrt{d-1}}V(d-1)$ *and that the ratio is* $\frac{2r}{\epsilon\sqrt{d-1}}e^{\frac{d-1}{2}\left(\frac{\epsilon}{r}\right)^2}$*.*

Exercise 2.25 *For a 1,000-dimensional unit-radius ball centered at the origin, what fraction of the volume of the upper hemisphere is above the plane* $x_1 = 0.1$*? Above the plane* $x_1 = 0.01$*?*

Exercise 2.26 *Let* $\left\{\mathbf{x}\,\middle|\,|\mathbf{x}| \leq 1\right\}$ *be a* d*-dimensional, unit radius ball centered at the origin. What fraction of the volume is the set* $\{(x_1, x_2, \ldots, x_d)|\forall i\ |x_i| \leq \frac{1}{\sqrt{d}}\}$*?*

Exercise 2.27 *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 2.28 *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 2.29 *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.*

Exercise 2.30

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.*

Exercise 2.31 *Define the equator of a d-dimensional unit cube to be the hyperplane* $\left\{\mathbf{x} | \sum_{i=1}^{d} x_i = \frac{d}{2}\right\}$.

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?*

Exercise 2.32 *How large must ε be for 99% of the volume of a d-dimensional unit-radius ball to lie in the shell of ε-thickness at the surface of the ball?*

Exercise 2.33 *Calculate the ratio of area above the plane $x_1 = \epsilon$ of a unit radius ball 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 $d = 1,000$.*

Exercise 2.34

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.*

Exercise 2.35 *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 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 2.36 *We have claimed that a randomly generated point on a ball lies near the equator of the ball, 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 ± 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 2.37 *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 ball. Assume one of the vectors is the North pole. Prove that the ratio of the area of a cone, with axis at the North Pole of fixed angle say* 45° *to the area of a hemisphere, goes to zero as the dimension increases. Thus, the probability that the angle between two random vectors is at most* 45° *goes to zero. How does this relate to the result that most of the volume is near the equator?*

Exercise 2.38 *Consider a slice of a 100-dimensional ball 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 ball to the poles must the planes be to contain 95% of the surface area?*

Exercise 2.39 *Place* n *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 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 2.40 *Project the surface area of a* d*-dimensions ball 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 2.41 *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*.*

Exercise 2.42 *How would you sample uniformly at random from the parallelepiped*

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

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

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

Your algorithms must run in polynomial time.

Exercise 2.43 *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.*

Exercise 2.44

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 2.45 *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?*

Exercise 2.46 *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.*

Exercise 2.47 *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?*

Exercise 2.48 *Consider two unit-radius spheres in d-dimensions whose centers are distance* δ *apart where* $\delta < 1$ *is a constant independent of d. Let* $\mathbf{x}$ *be a random point on the surface of the first sphere and* $\mathbf{y}$ *a random point on the surface of the second sphere. Prove that as d goes to infinity, the probability that* $|\mathbf{x} - \mathbf{y}|^2$ *is more than* $2 + \delta^2 + s$, *falls off exponentially with s.*

Exercise 2.49 *Pick a point* $\mathbf{x}$ *uniformly at random from the following set in d-space:*

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

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

2. *Show that with high probability,* $x_1^4 + x_2^4 + \cdots + x_d^4 \geq 1 - O(1/d)$.

3. *Show that with high probability,* $|x_1| \leq O(1/d^{1/4})$.

Exercise 2.50 *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?*

Exercise 2.51 *Let $x_1, x_2, \ldots, x_n$ be independent samples of a random variable* $\mathbf{x}$ *with mean m and variance σ^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 $var(m_s) = \frac{1}{n}var(\mathbf{x})$. 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 $(x_i - m) - (m_s - m)$.

Exercise 2.52 *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?*

Exercise 2.53 *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 μ, so that with probability at least 99/100,*

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

How many samples are sufficient to ensure that

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

Exercise 2.54 *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 fomula $\sigma^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - 5)^2$. How close is the estimate of σ^2 to the true variance of 9?*

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

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

Exercise 2.55 *The Cauchy distribution in one dimension is* $Prob(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. *What would happen if one tried to extend the distribution to higher dimensions by the formula* $Prob(r) = c\frac{1}{1+r^2}$, *where* r *is the distance from the origin? What happens when you try to determine a normalization constant* c*?*

Exercise 2.56 *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* $E(x^2)$*?*

Exercise 2.57 *Consider* d*-space and the following density over the positive orthant:*

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

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* $\mathbf{x}$ *with this density to have an expected value* $E(|\mathbf{x}|)$. *What condition do you need if we want* $E(|\mathbf{x}|^2)$ *to exist?*

Exercise 2.58 *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* $p(x)$?

Exercise 2.59 *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* $Prob(|x| > a) = Prob(x^2 > a^2)$ *to get a tighter bound.*
3. *What is the bound using* $Prob(|x| > a) = Prob(x^r > a^r)$?

Exercise 2.60 *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.

Exercise 2.61 *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *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.*

Exercise 2.62 *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 $\sqrt{\frac{k}{d}}$.*

Exercise 2.63 *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* $\mathbf{x}$ *in P and* $\mathbf{y}$ *in Q such that* $|\mathbf{x} - \mathbf{y}|$ *is minimum.*

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

Exercise 2.64 *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%.

Exercise 2.65 *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 2.66 *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 say* $O(\ln n/\varepsilon^2)$ *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?*

Exercise 2.67 *Create a list of the five most important things that you learned about high dimensions.*

Exercise 2.68 *Write a short essay whose purpose is to excite a college freshman to learn about high dimensions.*

References

[ABC+08] Reid Andersen, Christian Borgs, Jennifer T. Chayes, John E. Hopcroft, Vahab S. Mirrokni, and Shang-Hua Teng. Local computation of pagerank contributions. *Internet Mathematics*, 5(1):23–45, 2008.

[AF] David Aldous and James Fill. *Reversible Markov Chains and Random Walks on Graphs*. http://www.stat.berkeley.edu/ aldous/RWG/book.html.

[AK] Sanjeev Arora and Ravindran Kannan. Learning mixtures of separated nonspherical gaussians. *Annals of Applied Probability*, 15(1A):6992.

[Alo86] Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6:83–96, 1986.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *COLT*, pages 458–469, 2005.

[AN72] Krishna Athreya and P. E. Ney. *Branching Processes*, volume 107. Springer, Berlin, 1972.

[AP03] Dimitris Achlioptas and Yuval Peres. The threshold for random k-sat is 2^k (ln 2 - o(k)). In *STOC*, pages 223–231, 2003.

[Aro11] Multiplicative weights method: a meta-algorithm and its applications. *Theory of Computing journal - to appear*, 2011.

[AS08] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., Hoboken, NJ, third edition, 2008. With an appendix on the life and work of Paul Erdős.

[BA] Albert-Lszl Barabsi and Rka Albert. Emergence of scaling in random networks. *Science*, 286(5439).

[BEHW] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinary*.

[BGG97] C Sidney Burrus, Ramesh A Gopinath, and Haitao Guo. Introduction to wavelets and wavelet transforms: a primer. 1997.

[Ble12] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[Blo62] H.D. Block. The perceptron: A model for brain functioning. *Reviews of Modern Physics*, 34:123–135, 1962. Reprinted in *Neurocomputing*, Anderson and Rosenfeld.

[BMPW98] Sergey Brin, Rajeev Motwani, Lawrence Page, and Terry Winograd. What can you do with a web in your pocket? *Data Engineering Bulletin*, 21:37–47, 1998.

[Bol01] Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[BT87] Béla Bollobás and Andrew Thomason. Threshold functions. *Combinatorica*, 7(1):35–38, 1987.

[CF86] Ming-Te Chao and John V. Franco. Probabilistic analysis of two heuristics for the 3-satisfiability problem. *SIAM J. Comput.*, 15(4):1106–1118, 1986.

[CGTS99] Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem (extended abstract). In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, STOC '99, pages 1–10, New York, NY, USA, 1999. ACM.

[CHK+] Duncan S. Callaway, John E. Hopcroft, Jon M. Kleinberg, M. E. J. Newman, and Steven H. Strogatz. Are randomly grown graphs really random?

[Chv92] *33rd Annual Symposium on Foundations of Computer Science, 24-27 October 1992, Pittsburgh, Pennsylvania, USA*. IEEE, 1992.

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[DFK91] Martin Dyer, Alan Frieze, and Ravindran Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. *Journal of the Association for Computing Machinary*, 1991.

[DFK+99] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA*, pages 291–299, 1999.

[DG99] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the johnson-lindenstrauss lemma. 99(006), 1999.

[DS84] Peter G. Doyle and J. Laurie Snell. *Random walks and electric networks*, volume 22 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC, 1984.

[DS07] Sanjoy Dasgupta and Leonard J. Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[ER60] Paul Erdös and Alfred Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[Fel68] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968.

[FK99] Alan M. Frieze and Ravindan Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[Fri99] Friedgut. Sharp thresholds of graph properties and the k-sat problem. *Journal of the American Math. Soc.*, 12, no 4:1017–1054, 1999.

[FS96] Alan M. Frieze and Stephen Suen. Analysis of two simple heuristics on a random instance of k-sat. *J. Algorithms*, 20(2):312–355, 1996.

[GKP94] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics - a foundation for computer science (2. ed.)*. Addison-Wesley, 1994.

[GvL96] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996.

[HBB10] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online learning for latent dirichlet allocation. In *NIPS*, pages 856–864, 2010.

[Jer98] Mark Jerrum. Mathematical foundations of the markov chain monte carlo method. In Dorit Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, 1998.

[JKLP93] Svante Janson, Donald E. Knuth, Tomasz Luczak, and Boris Pittel. The birth of the giant component. *Random Struct. Algorithms*, 4(3):233–359, 1993.

[JLR00] Svante Janson, Tomasz Ĺuczak, and Andrzej Ruciński. *Random Graphs*. John Wiley and Sons, Inc, 2000.

[Kan09] Ravindran Kannan. A new probability inequality using typical moments and concentration results. In *FOCS*, pages 211–220, 2009.

[Kar90] Richard M. Karp. The transitive closure of a random digraph. *Random Structures and Algorithms*, 1(1):73–94, 1990.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *JOURNAL OF THE ACM*, 46(5):604–632, 1999.

[Kle00] Jon M. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC*, pages 163–170, 2000.

[Kle02] Jon M. Kleinberg. An impossibility theorem for clustering. In *NIPS*, pages 446–453, 2002.

[KV95] Michael Kearns and Umesh Vazirani. *An introduction to Computational Learning Theory*. MIT Press, 1995.

[KV09] Ravi Kannan and Santosh Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3-4):157–288, 2009.

[Liu01] Jun Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

[Mat10] Jiří Matoušek. *Geometric discrepancy*, volume 18 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2010. An illustrated guide, Revised paperback reprint of the 1999 original.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[MP69] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969.

[MR95a] Michael Molloy and Bruce A. Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms*, 6(2/3):161–180, 1995.

[MR95b] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[MR99] Rajeev Motwani and Prabhakar Raghavan. Randomized algorithms. In *Algorithms and theory of computation handbook*, pages 15–1–15–23. CRC, Boca Raton, FL, 1999.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102, 2010.

[Nov62] A.B.J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata, Vol. XII*, pages 615–622, 1962.

[Pal85] Edgar M. Palmer. *Graphical evolution*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1985. An introduction to the theory of random graphs, A Wiley-Interscience Publication.

[Par98] Beresford N. Parlett. *The symmetric eigenvalue problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[per10] *Markov Chains and Mixing Times*. American Mathematical Society, 2010.

[Sch90] Rob Schapire. Strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[SJ] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing markov chains. *Information and Computation*.

[Sly10] Allan Sly. Computational transition at the uniqueness threshold. In *FOCS*, pages 287–296, 2010.

[SS01] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[Val84] Leslie G. Valiant. A theory of the learnable. In *STOC*, pages 436–445, 1984.

[Val13] L. Valiant. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books, 2013.

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[Vem04] Santosh Vempala. *The Random Projection Method*. DIMACS, 2004.

[VW02] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. *Journal of Computer and System Sciences*, pages 113–123, 2002.

[Wil06] H.S. Wilf. *Generatingfunctionology*. Ak Peters Series. A K Peters, 2006.

[WS98a] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684), 1998.

[WS98b] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[WW96] E. T. Whittaker and G. N. Watson. *A course of modern analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1996. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions, Reprint of the fourth (1927) edition.