CHAPTER ONE

Introduction

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

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

Early drafts of the book have been used for both undergraduate and graduate courses. Background material needed for an undergraduate course has been put into a background chapter with associated homework problems.

Modern data in diverse fields such as information processing, search, and machine learning is often advantageously represented as vectors with a large number of components. The vector representation is not just a book-keeping device to store many fields of a record. Indeed, the two salient aspects of vectors – geometric (length, dot products, orthogonality, etc.) and linear algebraic (independence, rank, singular values, etc.) – turn out to be relevant and useful. Chapters 2 and 3 lay the foundations of geometry and linear algebra, respectively. More specifically, our intuition from two- or three-dimensional space can be surprisingly off the mark when it comes to high dimensions. Chapter 2 works out the fundamentals needed to understand the differences. The emphasis of the chapter, as well as the book in general, is to get across the intellectual ideas and the mathematical foundations rather than focus

on particular applications, some of which are briefly described. Chapter 3 focuses on singular value decomposition (SVD) a central tool to deal with matrix data. We give a from-first-principles description of the mathematics and algorithms for SVD. Applications of singular value decomposition include principal component analysis, a widely used technique we touch on, as well as modern applications to statistical mixtures of probability densities, discrete optimization, etc., which are described in more detail.

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

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

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

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

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

Chapter 9 focuses on linear-algebraic problems of making sense from data, in particular topic modeling and nonnegative matrix factorization. In addition to discussing well-known models, we also describe some current research on models and

algorithms with provable guarantees on learning error and time. This is followed by graphical models and belief propagation.

Chapter 10 discusses ranking and social choice as well as problems of sparse representations such as compressed sensing. Additionally, Chapter 10 includes a brief discussion of linear programming and semidefinite programming. Wavelets, which are an important method for representing signals across a wide range of applications, are discussed in Chapter 11 along with some of their fundamental mathematical properties. Chapter 12 includes a range of background material.

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

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

CHAPTER TWO

High-Dimensional Space

2.1. Introduction

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

2.2. The Law of Large Numbers

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

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

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

Specifically, the Law of Large Numbers states that

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

The larger the variance of the random variable, the greater the probability that the error will exceed ϵ. Thus the variance of x is in the numerator. The number of samples n is in the denominator, since the more values that are averaged, the smaller the probability that the difference will exceed ϵ. Similarly the larger ϵ is, the smaller the

probability that the difference will exceed ϵ and hence ϵ is in the denominator. Notice that squaring ϵ makes the fraction a dimensionless quantity.

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

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

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

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

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

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

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

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

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

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

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

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

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

■

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

$$E(x + y) = E(x) + E(y),$$
$$Var(x - c) = Var(x),$$
$$Var(cx) = c^2 Var(x).$$

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

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

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

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

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

Proof $E(\frac{x_1+x_2+\cdots+x_n}{n}) = E(x)$ and thus

$$\text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) = \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)\right| \geq \epsilon\right)$$

By Chebyshev's inequality,

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

■

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

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

As an application of the Law of Large Numbers, let $\mathbf{z}$ be a d-dimensional random point whose coordinates are each selected from a zero mean, $\frac{1}{2\pi}$ variance Gaussian.

Table 2.1: Table of tail bounds. The Higher Moments bound is obtained by applying Markov to x^r. The Chernoff, Gaussian Annulus, and Power Law bounds follow from Theorem 2.5 which is proved in Chapter 12.

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

We set the variance to $\frac{1}{2\pi}$ so the Gaussian probability density equals one at the origin and is bounded below throughout the unit ball by a constant.[1] By the Law of Large Numbers, the square of the distance of $\mathbf{z}$ to the origin will be $\Theta(d)$ with high probability. In particular, there is vanishingly small probability that such a random point $\mathbf{z}$ would lie in the unit ball. This implies that the integral of the probability density over the unit ball must be vanishingly small. On the other hand, the probability density in the unit ball is bounded below by a constant. We thus conclude that the unit ball must have vanishingly small volume.

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

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

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

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

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

Theorem 2.5 (Master Tail Bounds Theorem) *Let $x = x_1 + x_2 + \cdots + x_n$, where $x_1, x_2, \ldots, x_n$ are mutually independent random variables with zero mean and variance at most σ^2. Let $0 \leq a \leq \sqrt{2}n\sigma^2$. Assume that $|E(x_i^s)| \leq \sigma^2 s!$ for $s = 3, 4, \ldots, \lfloor (a^2/4n\sigma^2) \rfloor$. Then,*

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

The proof of Theorem 2.5 is elementary. A slightly more general version, Theorem 12.5, is given in Chapter 12. For a brief intuition of the proof, consider applying Markov's inequality to the random variable x^r where r is a large even number. Since r is even, x^r is nonnegative, and thus $\text{Prob}(|x| \geq a) = \text{Prob}(x^r \geq a^r) \leq E(x^r)/a^r$. If $E(x^r)$ is not too large, we will get a good bound. To compute $E(x^r)$, write $E(x)$ as $E(x_1 + \cdots + x_n)^r$ and expand the polynomial into a sum of terms. Use the fact that by independence $E(x_i^{r_i} x_j^{r_j}) = E(x_i^{r_i})E(x_j^{r_j})$ to get a collection of simpler expectations that can be bounded using our assumption that $|E(x_i^s)| \leq \sigma^2 s!$. For the full proof, see Chapter 12.

2.3. The Geometry of High Dimensions

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

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

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

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

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

Let S denote the unit ball in d-dimensions, that is, the set of points within distance one of the origin. An immediate implication of the above observation is that at least a $1 - e^{-\epsilon d}$ fraction of the volume of the unit ball is concentrated in $S \setminus (1-\epsilon)S$, namely in a small annulus of width ϵ at the boundary. In particular, most of the volume of the d-dimensional unit ball is contained in an annulus of width $O(1/d)$ near the boundary. This is illustrated in Figure 2.1. If the ball is of radius r, then the annulus width is $O\left(\frac{r}{d}\right)$.

2.4. Properties of the Unit Ball

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

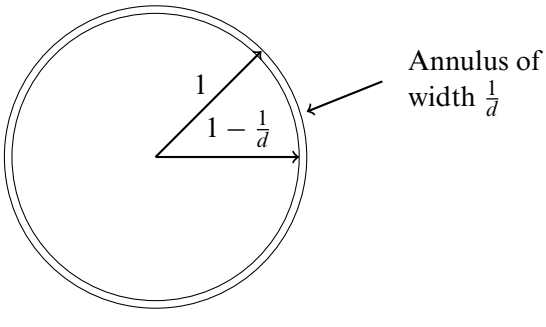

Figure 2.1: Most of the volume of the d-dimensional ball of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

the volume of the ball goes to zero. This result can be proven in several ways. Here we use integration.

2.4.1. Volume of the Unit Ball

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

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

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

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

Since the variables Ω and r do not interact,

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

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

Consider a different integral,

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

Including the exponential allows integration to infinity rather than stopping at the surface of the sphere. Thus, $I(d)$ can be computed by integrating in both Cartesian

and polar coordinates. Integrating in polar coordinates will relate $I(d)$ to the surface area $A(d)$. Equating the two results for $I(d)$ allows one to solve for $A(d)$.

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

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

Here, we have used the fact that $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$. For a proof of this, see Section 12.2 of Chapter 12. Next, calculate $I(d)$ by integrating in polar coordinates. The volume of the differential element is $r^{d-1} d\Omega dr$. Thus,

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

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

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

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

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

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

establishing the following lemma.

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

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

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