Chapter 2
Curses, Blessings, and Surprises in High Dimensions

This chapter discusses the curse of dimensionality as well as the blessings of dimensionality. The first is caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space. The latter is a manifestation of the concentration of measure. We also discuss some surprising facts in high dimensions. Since several of the results discussed in this chapter require basic tools from probability, we begin this chapter by reviewing some fundamental probabilistic concepts.

2.1 Basic Concepts from Probability

We briefly review some fundamental concepts from probability theory, which are helpful or necessary to understand the blessings of dimensionality and some of the surprises encountered in high dimensions. More advanced probabilistic concepts will be presented in Chapter **??**. We assume that the reader is familiar with elementary probability as is covered in introductory probability courses.

The two most basic concepts in probability associated with a random variable X are *expectation* (or *mean*) and *variance*, denoted by

$$\mathbb{E}[X] \qquad \text{and} \qquad \text{Var}(X) := \mathbb{E}[X - \mathbb{E}[X]]^2,$$

respectively. An important tool to describe probability distributions is the *moment generating function* of X, defined by

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}.$$

The p-th moment of X is defined by $\mathbb{E}[X^p]$ for $p > 0$ and the p-th absolute moment is $\mathbb{E}[|X|^p]$.

We can introduce L^p-norms of random variables by taking the p-th root of moments, i.e.,

$$\|X\|_{L^p} := \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}}, \qquad p \in [0, \infty],$$

with the usual extension to $p = \infty$ by setting

$$\|X\|_\infty := \operatorname{ess\,sup} |X|.$$

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, where Σ denotes a σ-algebra on the sample space Ω and $\mathbb{P}$ is a probability measure on (Ω, Σ). For fixed p the vector space $L^p(\Omega, \Sigma, \mathbb{P})$ consists of all random variables X on Ω with finite L^p-norm, i.e.,

$$L^p(\Omega, \Sigma, \mathbb{P}) = \{X : \|X\|_{L^p} < \infty.$$

We will usually not mention the underlying probability space. For example, we will often simply write L^p for $L^p(\Omega, \Sigma, \mathbb{P})$.

The case $p = 2$ deserves special attention since L^2 is a Hilbert space with inner product and norm

$$\langle X, Y \rangle_{L^2} = \mathbb{E}[XY], \qquad \|X\|_{L^2} = \left(\mathbb{E}[X^2]\right)^{\frac{1}{2}},$$

respectively. Note that the *standard deviation* $\sigma(X) := \sqrt{\operatorname{Var}(X)}$ of X can be written as

$$\sigma(X) = \|X - \mathbb{E}[X]\|_{L^2}.$$

The *covariance* of the random variables X and Y is

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle_{L^2}. \tag{2.1}$$

We recall a few classical inequalities for random variables. *Hölder's inequality* states that for random variables X and Y on a common probability space and $p, q \geq 1$ with $1/p + 1/q = 1$, there holds

$$|\mathbb{E}[XY]| \leq \|X\|_{L^p} \|Y\|_{L^q}. \tag{2.2}$$

The special case $p = q = 2$ is the *Cauchy-Schwarz inequality*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[|X|^2]\mathbb{E}[|Y|^2]}. \tag{2.3}$$

Jenssen's inequality states that for any random variable X and a convex function $\varphi : \mathbb{R} \to \mathbb{R}$, we have

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]. \tag{2.4}$$

Since $\varphi(x) = x^{p/q}$ is a convex function it follows immediately from Jenssen's inequality that

$$\|X\|_{L^p} \leq \|X\|_{L_q} \qquad \text{for } 0 \leq p \leq q < \infty.$$

Minkovskii's inequality states that for any $p \in [0, \infty]$ and any random variables X, Y, we have

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}, \tag{2.5}$$

which can be viewed as the *triangle inequality*.

The *cumulative distribution function* of X is defined by

$$F_X(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}.$$

We have $\mathbb{P}\{X > t\} = 1 - F_X(t)$, where the function $t \mapsto \mathbb{P}\{|X| \geq t\}$ is called the *tail* of X. The following lemma establishes a close connection between expectation and tails.

Proposition 2.1 (Integral identity). *Let X be a non-negative random variable. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}\{X > t\}\, dt.$$

The two sides of this identity are either finite or infinite simultaneously.

Markov's inequality is a fundamental tool to bound the tail in terms of expectation.

Proposition 2.2. *For any non-negative random variable $X : S \to \mathbb{R}$ we have*

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t} \qquad \text{for all } t > 0. \tag{2.6}$$

Proof. Let $\mathscr{I}$ denote the event $\{X \geq t\}$. Then

$$\mathbb{E}[X] = \sum_{s \in S} p(s)X(s) = \sum_{s \in \mathscr{I}} p(s)X(s) + \sum_{s \in \mathscr{I}^c} p(s)X(s).$$

Since X is non-negative, there holds $\sum_{s \in \mathscr{I}^c} p(s)X(s) \geq 0$ and

$$\mathbb{E}[X] \geq \sum_{s \in \mathscr{I}} p(s)X(s) \geq t \sum_{s \in \mathscr{I}} p(s) = t\mathbb{P}\{\mathscr{I}\}.$$

An important consequence of Markov's inequality is *Chebyshev's inequality*.

Corollary 2.1. *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$*

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}. \tag{2.7}$$

Chebyshev's inequality, which follows by applying Markov's inequality to the non-negative random variable $Y = (X - \mathbb{E}[X])^2$, is a form of concentration inequality, as it quantifies how close X is to its mean μ whenever the variance of X is small. Both, Markov's and Chebyshev's inequality are sharp, i.e., in general they cannot be improved.

Since we are not able to improve Markov's inequality and Chebyshev's inequality in general, the question arises whether we can give a stronger statement for a more restricted class of random variables. Of central importance in this context is the case of a random variable that is the *sum of a number of independent random variables*. This leads to the rich topic of *concentration inequalities* which is discussed in this chapter in Section **??** and as well in Chapter **??**.

2.2 The Curse of Dimensionality

The *curse of dimensionality* refers to the fact that many algorithmic approaches to problems in $\mathbb{R}^d$ become *exponentially* more difficult as the dimension d grows. The expression "curse of dimensionality" was coined by Bellman to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space [31].

For instance, if we want to sample the unit interval such that the distance between adjacent points is at most 0.01, 100 evenly-spaced sample points suffice; an equivalent sampling of a five-dimensional unit hypercube with a grid with a spacing of 0.01 between adjacent points would require 10^{10} sample points. Thus, a modest increase in dimensions results in a dramatic increase in required data points to cover the space at the same density.

Intimately connected to the curse of dimensionality is the problem of *overfitting* and *underfitting*. Here, overfitting refers to the problem that an algorithm shows performance on the training data, but poor generliazation to other data. Underfitting in turn, corresponds to poor performance on the training data and poor generalization to other data. This problem manifests itself in many machine learning algorithms.

We will discuss a toy example from image classification in more detail to illustrate the underlying issues. Assume we want to classify images into two groups, cars and bicycles. From the vast number of images depicting cars or bicycles, we are only able to obtain a small number of training images, say five images of cars and five images of bicycles. We want to train a simple linear classifier based on these ten labeled training images to correctly classify the remaining unlabeled car/bicycle images. We start with a simple feature, e.g. the amount or red pixels' in each image. However, this is unlikely to give a linear separation of the training data. We add more features and eventually the training images become linearly separable. This might suggest that increasing the number of features until perfect classification of the training data is achieved, is a sound strategy. However, as we *linearly increase* the dimension of the feature space, the density of our training data *decreases exponentially* with the feature dimension. In other words, the maintain a comparable density of our training data, we would need to increase the size of the datset exponentially – the curse of dimensionality. Thus, we risk producing a model that could be very good at predicting the target class on the training set, but it may fail miserably when faced with new data, that is, our model does not *generalize* from the training data to the test data.

2.3 Blessings of Dimensionality

Suppose we wish to predict the outcome of an event of interest. One natural approach would be to compute the expected value of the object. However, how can we tell how good the expected value is to the actual outcome of the event? Without further information of how well the actual outcome concentrates around its expec-

tation, the expected value is of little use. We would like to have an estimate for the probability that the actual outcome deviates from its expectation by a certain amount. This is exactly the role that *concentration inequalities* play in probability and statistics.

The concentration of measure phenomenon was put forward by Vitali Milman in the asymptotic geometry of Banach spaces regarding probabilities on product spaces in high dimensions [**?**, **?**].

The celebrated laws of large numbers of classical probability theory is the most well known form of *concentration of measure*; it states that sums of independent random variables are, under very mild conditions, close to their expectation with a large probability. We will see various quantitative versions of such concentration inequalities throughout this course. Some deal with sums of scalar random variables, others with sums of random vectors or sums of random matrices. Such concentration inequalities are instances of what is sometimes called *Blessings of dimensionality* (cf. [61]). This expression refers to the fact that certain random fluctuations can be well controlled in high dimensions, while it would be very complicated to make such predictive statements in moderate dimensions.

In this section we will review some concentration inequalities that deal with scalar random variables. One example is Chernoff's bound. There are many different forms of Chernoff bounds, each associated with different assumptions. Some forms assume a certain distribution, others apply to bounded random variables, regardless of their distribution. The version we state here is of the latter type:

Theorem 2.1. *Let $X_1, X_2, \ldots, X_n$ be random variables such that $a \leq X_i \leq b$ for all i. Let $X = \sum_{i=1}^{n} X_i$ and set $\mu = \mathbb{E}[X]$. Then, for all $\delta > 0$:*

(i) Upper tail: $\mathbb{P}\{X \geq (1+\delta)\mu\} \leq e^{-\frac{2\delta^2\mu^2}{n(b-a)^2}}$.

(ii) Lower tail: $\mathbb{P}\{X \leq (1-\delta)\mu\} \leq e^{-\frac{2\delta^2\mu^2}{n(b-a)^2}}$.

For more concentration inequalities, such as Hoeffding's inequality and Bernstein's inequality see Chapter 6.

2.4 Surprises in High Dimensions

When we peel an orange, then after having removed the rind we are still left with the majority of the orange. Suppose now we peel a d-dimensional orange for large d, then after removing the orange peel we would be left with essentially nothing. The reason for this from a healthy nutrition viewpoint discouraging fact is that for a d-dimensional unit ball almost all of its volume is concentrated near the boundary sphere. This is just one of many surprising phenomena in high dimensions. Many of these surprises are actually a manifestation of some form of concentration of measure that we encountered in the previous section (and thus they are not so surprising anymore ...).

When introducing data analysis concepts, we typically use few dimensions in order to facilitate visualization. However, our intuition about space, which is naturally based on two and three dimensions, can often be misleading in high dimensions. Many properties of even very basic objects become counterintuitive in higher dimensions. Understanding these paradoxical properties is essential in data analysis as it allows us to avoid pitfalls in the design of algorithms and statistical methods for high-dimensional data. It is instructive to analyze the shape and properties of some basic geometric forms, which we understand very well in dimensions two and three, in high dimensions.

To that end, we will look at some of the properties of the sphere and the cube as the dimension increases. The d-dimensional hyperball of radius R is defined by

$$B^d(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 \leq R^2\},$$

the d-dimensional hypersphere (or d-sphere) of radius R is given by

$$S^{d-1}(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 = R^2\},$$

and the d-dimensional hypercube with side length $2R$ is the subset of $\mathbb{R}^d$ defined as the d-fold product of intervals $[-R,R]$:

$$C^d(R) = \underbrace{[-R,R] \times \cdots \times [-R,R]}_{d \text{ times}}.$$

If there is no danger of confusion, we may write B^d for $B^d(1)$, S^{d-1} for $S^{d-1}(1)$, and C^d for $C^d(\frac{1}{2})$.

2.4.1 Geometry of spheres and balls in high dimension

2.4.1.1 Volume of the hyperball

Theorem 2.2. *The volume of $B^d(R)$ is given by*

$$\mathrm{Vol}(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\,\Gamma(\frac{d}{2})}. \tag{2.8}$$

Proof. The volume of $B^d(R)$ is given by

$$\mathrm{Vol}(B^d(R)) = \int_0^R s_d r^{d-1} dr = \frac{s_d R^d}{d}, \tag{2.9}$$

where s_d denotes the (hyper-)surface area of a unit d-sphere. A unit d-sphere must satisfy

$$s_d \int_0^\infty e^{-r^2} r^{d-1} dr = \underbrace{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty}}_{d \text{ times}} e^{-(x_1^2+\dots+x_d^2)} dx_1 \dots dx_d = \Big(\int_{-\infty}^{\infty} e^{-x^2} dx \Big)^d.$$

Recall that the Gamma function is given by

$$\Gamma(n) = \int_0^\infty r^{n-1} e^{-r} dr = 2 \int_0^\infty e^{-r^2} r^{2n-1} dr,$$

hence

$$\frac{1}{2} s_d \Gamma(\frac{d}{2}) = \Big[\Gamma(\frac{1}{2}) \Big]^d = \big(\pi^{\frac{1}{2}} \big)^d,$$

and thus

$$s_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}.$$

Plugging this expression into (2.9) gives

$$\mathrm{Vol}(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\, \Gamma(\frac{d}{2})}. \tag{2.10}$$

□

For positive integers n there holds $\Gamma(n) = (n-1)!$. Using Stirling's Formula,

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}} (\frac{n}{e})^n$$

we obtain as approximation for the volume of the unit d-ball for large d

$$\mathrm{Vol}(B^d) \approx \frac{1}{\sqrt{2\pi}} \Big(\frac{2\pi e}{d} \Big)^{\frac{d}{2}}. \tag{2.11}$$

Since the denominator in the parenthesis of equation (2.11) goes to infinity much faster than the numerator, the volume of the unit d-sphere goes rapidly to 0 as the dimension d increases to infinity, see also Figure 2.1.

Thus, unit spheres in high dimensions have almost no volume—compare this to the unit cube, which has volume 1 in any dimension. For $B^d(R)$ to have volume equal to 1, its radius R must be approximately (asymptotically) equal to $\sqrt{\frac{d}{2\pi e}}$.

2.4.1.2 Concentration of the volume of a ball near its equator

If we take an orange and cut it into slices, then the slices near the center are larger since the sphere is wider there. This effect increases dramatically (exponentially with the dimension) with increasing dimension. Assume we want to cut off a slab

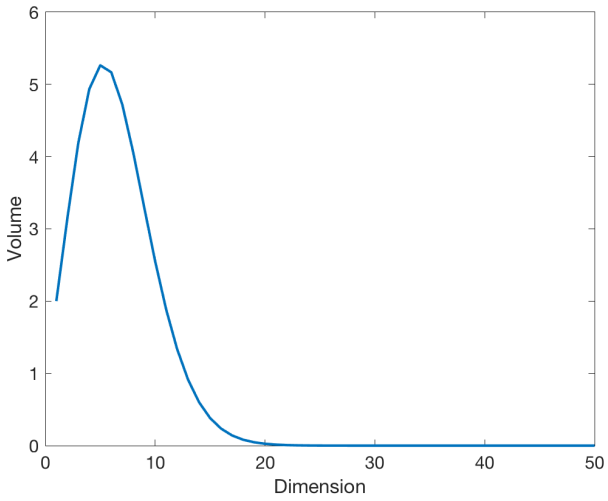

Fig. 2.1: The volume of the unit d-ball using the exact formula in equation (2.10). The volume reaches its maximum for $d = 5$ and decreases rapidly to zero with increasing dimension d.

around the "equator[1]" of the d-unit ball such that 99% of its volume is contained inside the slab. In two dimensions the width of the slab has to be almost 2, so that 99% of the volume are captured by the slab. But as the dimension increases the width of the slab gets rapidly smaller. Indeed, in high dimensions only a very thin slab is required, since nearly all the volume of the unit ball lies a very small distance away from the equator. The following theorem makes the considerations above precise.

Theorem 2.3. *Almost all the volume of $B^d(R)$ lies near its equator.*

Proof. It suffices to prove the result for the unit d-ball. Without loss of generality we pick as "north" the direction x_1. The intersection of the sphere with the plane $x_1 = 0$ forms our equator, which is formally given by the $d-1$-dimensional region $\{x : \|x\| \leq 1, x_1 = 0\}$. This intersection is a sphere of dimension $d-1$ with volume $\mathrm{Vol}(B^{d-1})$ given by the $(d-1)$-analog of formula (2.10) with $R = 1$.

We now compute the volume of B^d that lies between $x_1 = 0$ and $x_1 = p_0$. Let $P_0 = \{x : \|x\| \leq 1, x_1 \geq p_0\}$ be the "polar cap", i.e., part of the sphere above the slab of width $2p_0$ around the equator. To compute the volume of the cap P we will integrate over all slices of the cap from 0 to p_0. Each such slice will be a sphere of dimension $d-1$ and radius $\sqrt{1-p^2}$, hence its volume is $(1-p^2)^{\frac{d-1}{2}} \mathrm{Vol}(B^{d-1})$. Therefore

[1] To define the "equator" of a the d-dimensional ball, we need to pick a "north pole" as reference. Without loss of generality we could pick the unit vector in the x_1-direction as defining "north".

$$\text{Vol}(P) = \int_{p_0}^{1} (1-p^2)^{\frac{d-1}{2}} \text{Vol}(B^{d-1})\, dp = \text{Vol}(B^{d-1}) \int_{p_0}^{1} (1-p^2)^{\frac{d-1}{2}}\, dp.$$

Using $e^x \geq 1+x$ for all x we can upper bound this integral by

$$\text{Vol}(P) \leq \text{Vol}(B^{d-1}) \int_{p_0}^{\infty} e^{-\frac{d-1}{2}p^2}\, dp \leq \frac{\text{Vol}(B^{d-1})}{d-1} e^{-\frac{(d-1)p_0^2}{2}},$$

where we have bounded the integral via the complementary error function $\text{erfc}(x)$ and used the fact that $\text{erfc}(x) \leq e^{-x^2}$.

Finally, a simple calculation shows that the ratio between the volume of the polar caps and the entire hypersphere is bounded by

$$\frac{2\,\text{Vol}(P)}{\text{Vol}(B^d)} \leq \frac{2\,\text{Vol}(P)}{\text{Vol}(B^{d-1})} \leq \frac{2}{d-1} e^{-\frac{d-1}{2}p_0^2}.$$

The expression above shows that this ratio decreases exponentially as both d and p increase, proving our claim that the volume of the sphere concentrates strongly around its equator. □

2.4.1.3 Concentration of the volume of a ball on shells

We consider two concentric balls $B^d(1)$ and $B^d(1-\varepsilon)$. Using equation (2.10), the ratio of their volumes is

$$\frac{\text{Vol}(B^d(1-\varepsilon))}{\text{Vol}(B^d(1))} = (1-\varepsilon)^d.$$

Clearly, for every ε this ratio tends to zero as $d \to \infty$. This implies that the spherical shell given by the region between $B^d(1)$ and $B^d(1-\varepsilon)$ will contain most of the volume of $B^d(1)$ for large enough d even if ε is very small. How quickly does the volume concentrate at the surface of $B^d(1)$? We choose ε as a function of d, e.g. $\varepsilon = \frac{t}{d}$, then

$$\frac{\text{Vol}(B^d(1-\varepsilon))}{\text{Vol}(B^d(1))} = (1-\frac{t}{d})^d \to e^{-t}.$$

Thus, almost all the volume of $B^d(R)$ is contained in an annulus of width R/d.

Therefore, if we peel a d-dimensional orange and even if we peel it very carefully so that we remove only a very thin layer of its peel, we will have removed most of the orange and are left with almost nothing.

2.4.2 Geometry of the d-dimensional cube

We have seen that most of the volume of the hypersphere is concentrated near its surface. A similar result also holds for the hypercube, and in general for high-dimensional geometric objects. Yet, the hypercube exhibits an even more interesting volume concentration behavior, which we will establish below.

We start with a basic observation.

Proposition 2.3. *The hypercube C^d has volume 1 and diameter $\sqrt{d}$.*

The above proposition, although mathematically trivial, hints already at a somewhat counterintuitive behavior of the cube in high dimensions. Its corners seem to get "stretched out" more and more, while the rest of the cube must "shrink" to keep the volume constant. This property becomes even more striking when we compare the cube with the sphere as the dimension increases.

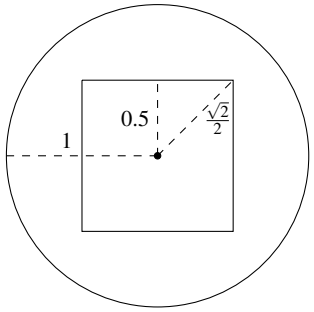

Fig. 2.2: 2-dimensional unit sphere and unit cube, centered at the origin.

In two dimensions (Figure 2.2), the unit square is completely contained in the unit sphere. The distance from the center to a vertex (radius of the circumscribed sphere) is $\frac{\sqrt{2}}{2}$ and the apothem (radius of the inscribed sphere) is $\frac{1}{2}$. In four dimensions (Figure 2.3), the distance from the center to a vertex is 1, so the vertices of the cube touch the surface of the sphere. However, the apothem is still $\frac{1}{2}$. The result, when projected in two dimensions no longer appears convex, however all hypercubes are convex. This is part of the strangeness of higher dimensions - hypercubes are both convex and "pointy." In dimensions greater than 4 the distance from the center to a vertex is $\frac{\sqrt{d}}{2} > 1$, and thus the vertices of the hypercube extend far outside the sphere, cf. Figure 2.4.

The considerations above suggest the following observation.

Theorem 2.4. *Almost all the volume of the high-dimensional cube is located in its corners.*

The proof of this statement will be based on a probabilistic argument, thereby illustrating (again) the nice and fruitful connection between geometry and probability

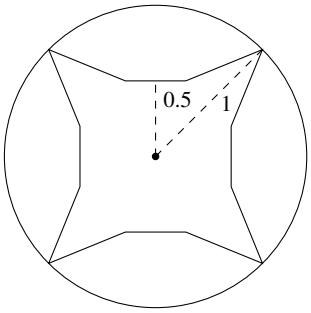

Fig. 2.3: Projections of the 4-dimensional unit sphere and unit cube, centered at the origin (4 of the 16 vertices of the hypercube are shown).

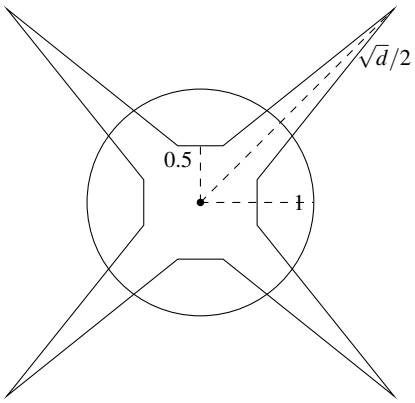

Fig. 2.4: Projections of the d-dimensional unit sphere and unit cube, centered at the origin (4 of the 2^d vertices of the hypercube are shown).

in high dimension. Pick a point at random in the box $[-1,1]^d$. We want to calculate the probability that the point is also in the sphere.

Let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and each $x_i \in [-1,1]$ is chosen uniformly at random. The event that x also lies in the sphere means

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \leq 1.$$

Let $z_i = x_i^2$ and note that

$$\mathbb{E}[z_i] = \frac{1}{2}\int_{-1}^{1} t^2\,dt = \frac{1}{3} \implies \mathbb{E}[\|x\|_2^2] = \frac{d}{3}$$

and

$$\mathrm{Var}(z_i) = \frac{1}{2}\int_{-1}^{1} t^4\,dt - (\frac{1}{3})^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45} \leq \frac{1}{10}$$

Using Chernoff's inequality,

$$\begin{aligned}
\mathbb{P}(\|x\|_2^2 \leq 1) &= \mathbb{P}(\sum_{i=1}^{d} x_i^2 \leq 1) \\
&= \mathbb{P}(\sum_{i=1}^{d} (z_i - \mathbb{E}[z_i)] \leq 1 - \frac{d}{3}) \\
&\leq \exp[-\frac{(\frac{d}{3}-1)^2}{\frac{4d}{10}}] \\
&\leq \exp[-\frac{d}{10}].
\end{aligned}$$

Since this value converges to 0 as the dimension d goes to infinity, this shows random points in high cubes are most likely outside the sphere. In other words, almost all the volume of a hypercube concentrates in its corners.

Since we now have gained a better understanding of the properties of the cube in high dimensions, we can use this knowledge to our advantage. For instance, this "pointiness" of the hypercube (in form of the ℓ_1-ball) turns out to very useful in the areas of compressive sensing and sparse recovery, see Chapter **??**.

2.4.3 How to generate random points on a sphere

How can we sample a point uniformly at random from S^{d-1}? The first approach that may come to mind is the following method to generate random points on a unit circle. Independently generate each coordinate uniformly at random from the interval $[-1,1]$. This yields points that are distributed uniformly at random in a square that contains the unit circle. We could now project all points onto the unit circle. However, the resulting distribution will not be 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 due to the difference in length of the diagonal of the square to its side length. To remedy this problem, we could discard all points outside the unit circle and project the remaining points onto the circle. However, if we generalize this technique to higher dimensions, the analysis in the previous section has shown that the ratio of the volume of $S^{d-1}(1)$ to the volume of $C^d(1)$ decreases rapidly. This makes this process not practical, since almost all the generated points will be discarded in this process and we end up with essentially no points inside (and thus, after projection, on) the sphere.

Instead we can proceed as follows. Recall that the multivariate Gaussian distribution is symmetric about the origin. This rotation invariance is exactly what we need.

We simply construct a vector in $\mathbb{R}^d$ whose entries are independently drawn from a univariate Gaussian distribution. We then normalize the resulting vector to lie on the sphere. This gives a distribution of points that is uniform over the sphere.

Having a method of generating points uniformly at random on S^{d-1} at our disposal, we can now give a probabilistic proof that points on S^{d-1} concentrate near its equator. Without loss of generality we pick an arbitrary unit vector x_1 which represents the "northpole", and the intersection of the sphere with the plane $x_1 = 0$ forms our equator. We extend x_1 to an orthonormal basis $x_1, \ldots, x_d$. We create a random vector by sampling $(Z_1, \ldots, Z_d) \sim \mathcal{N}(0, I_d)$ and normalize the vector to get $X = (X_1, \ldots, X_d) = \frac{1}{\sum_{k=1}^d Z_k^2}(Z_1, \ldots, Z_d)$. Because X is on the sphere, it holds that $\sum_{k=1}^d \langle X, x_k \rangle^2 = 1$. Note that we also have $\mathbb{E}[\sum_{k=1}^d \langle X, x_1 \rangle^2] = \mathbb{E}[1] = 1$. Thus, by symmetry, $\mathbb{E}[\langle X, x_1 \rangle^2] = \frac{1}{d}$. Applying Markov's inequality (2.6) yields

$$\mathbb{P}(|\langle X, x_1 \rangle| > \varepsilon) = \mathbb{P}(\langle X, x_1 \rangle^2 > \varepsilon^2) \leq \frac{\mathbb{E}(\langle X, x_1 \rangle^2)}{\varepsilon^2} = \frac{1}{d\varepsilon^2}.$$

For fixed ε we can make this probability arbitrarily small by increasing the dimension d. This proves our claim that points on the high-dimensional sphere concentrate near its equator.

2.4.4 Random vectors in high dimensions

What is the typical length of a random vector in high dimensions? And why are two randomly drawn vectors in high dimensions almost perpendicular?

Theorem 2.5. *Two random vectors in high dimensions are almost orthogonal. ....*