

## Chapter 3

# Measure Concentration

Suppose the surface of the earth is completely covered with grass, and your task is to mow it. You have a giant lawn mower able to mow a strip that spans a spherical angle of  $\alpha$ , say (where  $\alpha$  is small in order not to make your task too easy). What percentage of the grass have you mowed after you have gone around the equator once? See Figure 3.1 (left) for an illustration of the situation.

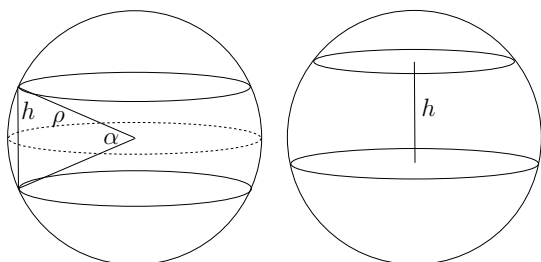


Figure 3.1: The giant lawn mower covers a strip of width  $h$

The lawn mower question is easy to solve using the following:

**Fact 3.1.** *On the 2-dimensional sphere of radius  $\rho$ , area covered by a strip of height  $h$  equals*

$$2\pi\rho h.$$

(In the case of the surface of the earth,  $\rho \approx 6,378\text{km}$ ) Interestingly, for the 2-dimensional sphere, the area of a strip does not depend on the strip being centered around the equator—any strip of height  $h$  has the same area  $2\pi\rho h$ , see Figure 3.1 (right). This is easy to prove using polar coordinates, and a straightforward computation.

For  $h = 2\rho$ , the strip covers the whole surface, and the area is  $4\pi\rho^2$ , which follows from the above formula for the surface area of  $\mathbb{S}^2$  by scaling by a factor of  $\rho$ .

Because the strip spans spherical angle  $\alpha$ , we get  $h = 2\rho \sin(\alpha/2)$ , meaning that the fraction of the earth's surface you have mowed is

$$\frac{2\pi\rho h}{4\pi\rho^2} = \frac{h}{2\rho} = \sin(\alpha/2).$$

If  $\alpha = 10^\circ$ , for example (a pretty big mower, the strip is more than 1,000km wide), the fraction covered is 8.7%.

### 3.1 Measure Concentration on the Sphere

The counterintuitive phenomenon is that your task would be much simpler if the earth were of higher dimension. For sufficiently large dimension, one round with your  $10^\circ$ -mower (or any  $\alpha$ -mower, for fixed  $\alpha$ ) covers 90% (or any desired percentage) of the surface. This means, the surface area of  $B_d$  is concentrated around the equator for large  $d$ . Not only that: by symmetry of  $B_d$ , the surface area is concentrated around *any* equator.

In fact, there is a much more general result of this kind. Since we are interested in relative surface area, we might as well renormalize the surface area measure so that the whole sphere gets measure 1. Thus, we define, for  $E \subseteq \mathbb{S}^{d-1}$ ,

$$P(E) := \frac{\sigma_{d-1}(E)}{\sigma_{d-1}(\mathbb{S}^{d-1})}.$$

In other words,  $P$  is the uniform probability measure on  $\mathbb{S}^{d-1}$ .

Furthermore, we need the following notion: For  $A \subseteq \mathbb{S}^{d-1}$  and a real number  $t > 0$ , let  $A_t := \{x \in \mathbb{S}^d : \text{dist}(x, A) \leq t\}$  be the set of points at (Euclidean) distance at most  $t$  from  $A$ , where, formally,  $\text{dist}(x, A) := \inf_{y \in A} \|x - y\|$ . With this notation, we can state the theorem about measure concentration on the sphere:

**Theorem 3.2.** *Let  $A \subseteq \mathbb{S}^{d-1}$ ,  $P(A) \geq 1/2$ . Then for any real  $t > 0$ ,*

$$1 - P(A_t) \leq 2e^{-t^2 d/2}.$$

For the aforementioned question of the relative surface measure of a strip around the equator, we can apply the theorem twice, once taking  $A$  to be the northern hemisphere, and once the southern hemisphere. In fact, there is an even stronger statement (which is referred to as the *isoperimetric inequality for the sphere*) which states that among all sets  $A$  with  $P(A) = 1/2$ ,  $P(A_t)$  is minimized, simultaneously for all  $t$ , if  $A$  is a hemisphere.

Figure 3.2 shows the (width of the) strip around the equator that contains 90% of the area, for three values of  $d$ , see Matoušek's book [Mat02]).

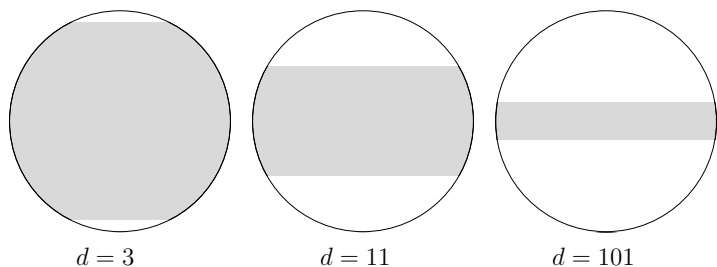


Figure 3.2: strip around the equator containing 90% of the area

## 3.2 Measure Concentration for the Discrete Cube: Chernov Bounds

In this section, we prove some technical results concerning random variables that are strongly concentrated around their expectation. All random variables are real-valued.

Some readers may find it convenient to think about random variables that take only finitely many distinct values, i.e., that there are finitely many real numbers  $a_1, \dots, a_n \in \mathbb{R}$  and  $p_1, \dots, p_n \in [0, 1]$  such that  $\Pr[X = a_i] = p_i$  and  $\sum_{i=1}^n p_i = 1$ . However, we try to formulate the theorems and lemmas so that they remain valid in general.

The most basic estimate for the probability that a random variable differs from its expectation<sup>1</sup>

**Fact 3.3 (Markov's Inequality).** *Let  $X$  be a nonnegative random variable, i.e.,  $X \geq 0$ . Then*

$$\Pr[X \geq \lambda] \leq \frac{\mathbf{E}[X]}{\lambda}$$

for all  $\lambda > 0$ .

Note that we allow here the possibility that  $\mathbf{E}[X] = +\infty$  (something we do not have to worry about if  $X$  takes only finitely many values), in which case the statement remains true but has absolutely no strength.

Next, consider a random variable  $X$  that is not necessarily nonnegative. Recall that the expectation of  $X$  is said to *exist* if  $\mathbf{E}[|X|] < \infty$ . (This implies

<sup>1</sup>Recall that for a discrete random variable  $X$  as above, the expectation is defined as  $\mathbf{E}[X] := \sum_i p_i a_i$ . The same definition works if the random variable takes countably many values. For absolutely continuous random variables, the sum is replaced by an integral  $\int_{\mathbb{R}} f_X(x) dx$ , where  $f_X$  is the *density* of the random variable. In general, random variables need to be neither discrete nor absolutely continuous and one needs a moderate amount of measure theory to define the expectation rigorously.

that  $\mathbf{E}[X]$  is also finite.) In this case, we can also estimate the probability that  $X$  deviates from its expectation in terms of the *variance*

$$\text{Var}[X] := \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

**Fact 3.4 (Chebyshev's Inequality).** *If  $X$  is a random variable for which  $\mathbf{E}[X]$  exists, then*

$$\Pr[|X - \mathbf{E}[X]| \geq \lambda] \leq \frac{\text{Var}[X]}{\lambda^2}$$

for all  $\lambda > 0$ .

In Markov's inequality, the estimate for the probability decreases linearly as  $\lambda \rightarrow \infty$ , and in Chebyshev's Inequality, it decreases quadratically. Generally, one can consider higher moments of a random variable to obtain bounds that decrease polynomially.

In this section, we consider random variables for which there is an *exponential* decrease.

**The normal distribution.** For the purposes of illustration, we mention a particularly well-known case of this behavior, the *standard normal* or *Gaussian distribution*. Recall that a random variable  $X$  is said to have the standard normal distribution if it has the distribution function

$$\Pr[X \leq \lambda] = \Phi(\lambda) := \int_{-\infty}^{\lambda} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-t^2/2}}_{=: \varphi(t)} dt$$

for all  $\lambda \in \mathbb{R}$ , see Figure 3.3.

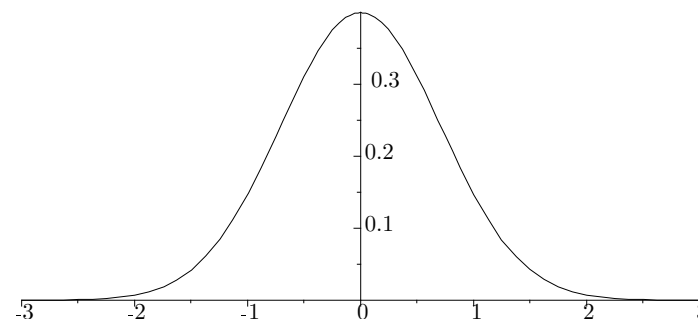


Figure 3.3: The density  $\varphi(t)$  of the standard normal distribution.

We also recall the basic facts that  $\mathbf{E}[X] = 0$  (in fact,  $X$  is symmetric about the origin) and  $\text{Var}[X] = 1$

**Fact 3.5 (Gaussian Tails).** Let  $X$  be a standard Gaussian random variable. There is a constant  $a > 0$  (in fact,  $a = \ln 2/(2\pi)$  works) such that <sup>2</sup>

$$\Pr[X > \lambda] \leq e^{-a\lambda^2} \quad (3.1)$$

for all  $\lambda > 0$ . By symmetry, we also have the symmetric inequality  $\Pr[X < -\lambda] \leq e^{-a\lambda^2}$ .

This motivates the following definition that gives a name to random variables which exhibit a similar behavior.<sup>3</sup>

**Definition 3.6 (Subgaussian tails).** Let  $X$  be a random variable with  $\mathbf{E}[X] = 0$ . We say that  $X$  has a subgaussian upper tail if there exists a constant  $a > 0$  such that

$$\Pr[X > \lambda] \leq e^{-a\lambda^2} \quad (3.2)$$

for all  $\lambda > 0$ . We say that  $X$  has a subgaussian upper tail up to  $\lambda_0$  if (3.2) holds for all  $0 < \lambda \leq \lambda_0$ . We say that  $X$  has a subgaussian tail if both  $X$  and  $-X$  have subgaussian upper tails.

An important example of strong concentrated is provided by the following.

**Theorem 3.7 (Chernov Bound).** Let  $X_1, \dots, X_n$  be independent random variables with  $\Pr[X_i = +1] = \Pr[X_i = -1] = 1/2$ . Let  $X := \sum_i X_i$ , so  $\mathbf{E}[X] = 0$  and  $\text{Var}[X] = n$ . Then the “normalized sum”  $Z := \frac{1}{\sqrt{n}}X$  has a subgaussian tail: there is a constant  $a > 0$  independent of  $n$  (in fact,  $a = 1/2$  works) such that

$$\Pr[X > \lambda\sqrt{n}] = \Pr[X < -\lambda\sqrt{n}] \leq e^{-a\lambda^2}.$$

Recall that the normalization factor  $\sqrt{n} = \sqrt{\text{Var}[X]}$  is called the standard variation of  $X$ .

<sup>2</sup>This is not hard to prove. First note that  $-\varphi'(t) = t\varphi(t)$ . Thus,

$$\Pr[X > \lambda] = 1 - \Phi(\lambda) = \int_{\lambda}^{\infty} \varphi(t) dt \leq \int_{\lambda}^{\infty} \underbrace{\frac{t}{\lambda}}_{\geq 1} \varphi(t) dt = -\frac{1}{\lambda} \int_{\lambda}^{\infty} \varphi'(t) dt = \frac{\varphi(\lambda)}{\lambda}.$$

Thus,  $\Pr[X > \lambda] \leq \frac{1}{\sqrt{2\pi}\lambda} e^{-\lambda^2/2}$ . This is almost, but not quite, what we claim. To remedy this, note first of all that by symmetry of the normal distribution,  $\Pr[X > \lambda] < 1/2$  for all  $\lambda > 0$ . For  $\lambda$  close to 0, more precisely, for  $0 < \lambda < \frac{1}{\sqrt{2\pi}}$ , we thus have  $\Pr[X > \lambda] < 1/2 < e^{-a_1\lambda^2}$  if we choose  $a_1 := \frac{\ln 2}{2\pi}$ , and for  $\lambda \geq 1/\sqrt{2\pi}$ , we have  $\Pr[X > \lambda] \leq e^{-a_2\lambda^2}$  with  $a_2 = 1/2$ , so the smaller of these two  $a$ 's works for all  $\lambda$ .

<sup>3</sup>The normal distribution, and random variables that are “approximately” normally distributed, are ubiquitous, because by the *Central Limit Theorem* from probability theory, if  $X$  is the sum of “many” independent random variables, none of which has “too large” variance compared to the others, then the normalized random variable  $\frac{X - \mathbf{E}[X]}{\sqrt{\text{Var}[X]}}$  has “approximately” the standard normal distribution. (We refer to any probability theory textbook for a precise formulation of the theorem.) This gives some intuition why sums of independent random variables might exhibit a similar concentration behavior as the normal distribution.

Before we give a proof, we mention the following “geometric” interpretation.

**A geometric picture: the discrete cube.** The Chernov bound is a discrete analogue of the phenomenon of measure concentration on the sphere.

Consider the *discrete cube*  $\{-1, +1\}^n$ , i.e., the set of vertices of the box  $Q_n(-1, 1)$ . The role of the equator of the sphere is played by the hyperplane  $\{x : \sum_i x_i = 0\}$  (there may or may not be any vertices that lie on the equator, depending on whether  $n$  is even or odd), see Figure 3.4 (left) for a picture in dimension 3.

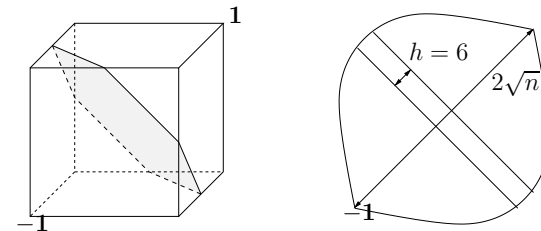


Figure 3.4: The “equator” of discrete cube (left); symbolic drawing of a strip of width  $h$  around the equator that contains 96.3% of all vertices of the cube (right)

Note that if  $X_1, \dots, X_n$  are independent balanced  $\pm 1$ -variables as in the Chernov bound, then the vector  $(X_1, \dots, X_n) \in \{-1, +1\}^n$  is a uniformly random vertex of the discrete cube, and  $X = \sum_i X_i$  measures the distance of this vertex from the equator, and in complete analogy with the spherical case, the Chernov bound says that most of the vertices are concentrated in a strip of small height around the equator (the normalization by  $1/\sqrt{n}$  corresponds to the fact that the cube has diameter  $2\sqrt{n}$ , while the sphere has diameter 2).

Setting the parameter in the Chernov bound to  $\lambda = 3$ , for example, yields

$$\Pr[|X| > 2\sqrt{n}] < 2e^{-9/2} \approx 0.022.$$

Thus, only a fraction of 2.2% of all vertices lie outside the strip  $\{x \in \mathbb{R}^n : |\langle \mathbf{1}, x \rangle| \leq 3\sqrt{n}\}$  around the equator. Note that the height of this strip, i.e., the distance between the two bounding hyperplanes  $\{x : \langle \mathbf{1}, x \rangle = -3\sqrt{n}\}$  and  $\{x : \langle \mathbf{1}, x \rangle = +3\sqrt{n}\}$  equals  $6\sqrt{n}/\|\mathbf{1}\| = 6$ , a constant, as opposed to the diameter of the cube, which equals  $2\sqrt{n}$ . See Figure 3.4 (right) for a symbolic picture.

**Moment generating functions.** We now proceed with the proof of the Chernov bound. First, we recall some standard inequalities for the exponential function.

**Lemma 3.8.**

$$1 + x \leq e^x \quad \text{for all } x \in \mathbb{R} \quad (3.3)$$

$$\frac{e^x + e^{-x}}{2} \leq e^{x^2/2} \quad \text{for all } x \in \mathbb{R} \quad (3.4)$$

All of these are easy to prove using the power series expansion  $e^x = \sum_{k=0}^{\infty} x^k/k!$  for the exponential function.

**Lemma 3.9 (Moment generating function and subgaussian tails).** *Let  $X$  be a random variable with  $\mathbf{E}[X] = 0$ . If  $\mathbf{E}[e^{tX}] \leq e^{Ct^2}$  for some constant  $C$  and all  $t > 0$  then  $X$  has a subgaussian upper tail; more precisely,*

$$\Pr[X \geq \lambda] \leq e^{-t^2/4C}$$

for all  $\lambda > 0$ . If  $\mathbf{E}[e^{tX}] \leq e^{Ct^2}$  for all  $t \in (0, t_0]$  then  $X$  has a subgaussian upper tail up to  $2Ct_0$ . Similarly, if  $\mathbf{E}[e^{-tX}] \leq e^{Ct^2}$  for all  $t \in (0, t_0]$ , then  $X$  has a subgaussian lower tail up to  $-2Ct_0$ .

We leave the proof as an exercise.

We recall the following basic fact about random variables:

**Lemma 3.10.** *If  $X$  and  $Y$  are independent random variables whose expectations exist. Then the expectation of  $XY$  exists and*

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

*Proof of the Theorem 3.7.* We use the preceding lemma. Let now  $X = \sum_{i=1}^n X_i$ , where the  $X_i$  are independent random variables with  $\Pr[X_i = 1] = \Pr[X_i = -1] = 1/2$ , and let  $Z = \frac{t}{\sqrt{n}}X$ . Let  $t > 0$ . Then

$$\mathbf{E}[e^{tZ}] = \mathbf{E}[e^{\frac{t}{\sqrt{n}}\sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbf{E}[e^{\frac{t}{\sqrt{n}}X_i}] = \left( \frac{e^{\frac{t}{\sqrt{n}}} + e^{-\frac{t}{\sqrt{n}}}}{2} \right)^n \leq e^{t^2}$$

Here, we use that the  $X_i$  and hence the random variables  $\frac{t}{\sqrt{n}}X_i$  are independent. For the next step, we simply use the definition of expectation, and for the last step, we apply estimate (3.4). The bound for the upper tail follows by Lemma 3.9 with  $C = 1$ . The proof for the lower tail is completely analogous.  $\square$