# Lecture 9. Curses, Blessings, and Surprises in High Dimensions

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
University of Utah

Math 5750/6880, Fall 2023

# Recap

We have studied:

- Statistic learning models: linear regression, logistic regression, SVM, kernel methods, regularization
- First-order optimization: gradient descent, subgradient, proximal gradient descent, heavy-ball, Nesterov acceleration, stochastic gradient

Next,

- Understanding high-dimensional spaces: high-dimensional geometry, concentration inequalities, clustering, dimension reduction, compressed sensing
- Deep learning.

- Curse of Dimensionality (CoD): Many algorithmic approaches to problems in $\mathbb{R}^d$ become exponentially more difficult as the dimension $d$ grows. E.g., finite difference solver for PDEs.

- Blessings of Dimensionality (BoD): concentration of measure. For instance, for a $d$-dimensional unit ball almost all of its volume is concentrated near the boundary sphere since $r^d \gg (r - \epsilon)^d$ when $d$ is large.

# Geometry of Spheres and balls in high dimension

Some notations:

- 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

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

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

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

We denote $B^d := B^d(1)$, $S^{d-1} := S^{d-1}(1)$, and $C^d := C^d(\frac{1}{2})$.

## Volume of the hyperball

**Theorem 1.** The volume of $B^d(R)$ is given by

$$Vol(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\Gamma(\frac{d}{2})}. \tag{1}$$

The Gamma function is defined by

$$\Gamma(n) = \int_0^\infty r^{n-1} e^{-r} dr = 2\int_0^\infty e^{-r^2} r^{2n-1} dr \underbrace{\sim}_{\text{Stirling's formula}} \sqrt{\frac{2\pi}{n}}\left(\frac{n}{e}\right)^n.$$

Therefore, we have an approximation for the volume of the $d$-unit ball for large $d$

$$Vol(B^d) \approx \frac{1}{\sqrt{d\pi}}\left(\frac{2\pi e}{d}\right)^{d/2}. \tag{2}$$

## Remarks

$$Vol(B^d) \approx \frac{1}{\sqrt{d\pi}}\left(\frac{2\pi e}{d}\right)^{d/2} \to 0 \ \text{ as } \ d \to \infty. \tag{3}$$

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 $\sqrt{\frac{d}{2\pi e}}$.

# Volume of the hyperball

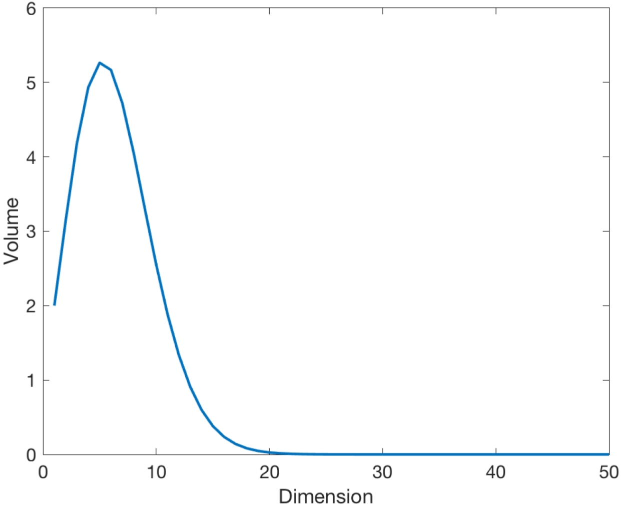

Figure: The volume of the unit $d$-ball. The volume reaches its maximum for $d = 5$ and decreases rapidly to zero with increasing dimension $d$.

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

$$Vol(B^d(R)) = \int_0^R s_d r^{d-1} dr = \frac{s_d R^d}{d}, \tag{4}$$

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

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

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

$$\frac{1}{2} s_d \Gamma(\frac{d}{2}) = \left[\Gamma(\frac{1}{2})\right]^d = (\pi^{1/2})^d, \Rightarrow s_d = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}.$$

Plugging this expression into (4) gives

$$Vol(B^d(R)) = \frac{\pi^{d/2} R^d}{\frac{d}{2}\Gamma(\frac{d}{2})}, \tag{5}$$

which concludes the proof.

## Concentration of the volume of a ball near its equator

Suppose we want to cut off a slab around the "equator" of the $d$-unit ball such that 99% of its volume is contained inside the slab. In 2D 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.

**Theorem 2.** Almost all the volume of $B^d(R)$ lies near its equator.

## Concentration of the volume of a ball near its equator – Result first

- Let $P = \{x : \|x\| \leq 1, x_1 \geq p_0\}$ be the "polar cap", i.e., part of the sphere above the slab of width $2p_0$ around the equator. Then the volume of the slab is $2Vol(P)$.

- We have

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

Proof. It suffices to prove the result for the unit $d$-ball. W.L.O.G. 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 $(d-1)$-D region $\{x : \|x\| \leq 1, x_1 = 0\}$. This intersection is a sphere of dimension $(d-1)$ with volume $Vol(B^{d-1})$ given by the $(d-1)$-analog of formula (5) with $R = 1$.

We now compute the volume of $B^d$ that lies between $x_1 = 0$ and $x_1 = p_0$. Let $P = \{x : \|x\| \leq 1, x_1 \geq p_0\}$ 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 $p_0$ to 1. Each such slice will be a sphere of dimension $d-1$ and radius $\sqrt{1-p^2}$, hence its volume is $(1-p^2)^{\frac{d-1}{2}} Vol(B^{d-1})$. Therefore,

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

The polar cap has almost no volume when $d$ is large.

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

$$Vol(P) \leq Vol(B^{d-1}) \int_{p_0}^{\infty} e^{-\frac{d-1}{2}p^2} dp = Vol(B^{d-1}) \sqrt{\frac{2}{d-1}} \int_{p_0\sqrt{\frac{d-1}{2}}}^{\infty} e^{-u^2} du$$

$$= Vol(B^{d-1}) \sqrt{\frac{\pi}{2(d-1)}} erfc\Big(p_0 \sqrt{\frac{d-1}{2}}\Big),$$

where $erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2} du$ is the complementary error function. The upper bound $erfc(x) \leq e^{-x^2}/(\sqrt{\pi}x)$ gives

$$Vol(P) \leq Vol(B^{d-1}) \sqrt{\frac{\pi}{2(d-1)}} \frac{e^{-\frac{d-1}{2}p_0^2}}{\sqrt{\pi} p_0 \sqrt{\frac{d-1}{2}}} = \frac{Vol(B^{d-1})}{d-1} \frac{e^{-\frac{d-1}{2}p_0^2}}{p_0}.$$

Recall from (5) that $Vol(B^d) = \frac{\pi^{d/2}}{\frac{d}{2}\Gamma(\frac{d}{2})}$, so for $d$ large enough (since $\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{(d-1)}{2})} \approx \sqrt{d/2}$), we have

$$Vol(B^{d-1}) = \frac{\pi^{-1/2}}{\frac{d-1}{d}} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})} Vol(B^d) \leq \frac{d-1}{2} Vol(B^d).$$

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

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

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. This concludes the proof.

# Concentration of the volume of a ball on shells

Recall the volume of unit $d$-ball is $Vol(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\Gamma(\frac{d}{2})} \sim \frac{1}{\sqrt{d\pi}}\left(\frac{2\pi e}{d}\right)^{d/2} \cdot R^d$, thus the ratio of two concentric balls $B^d(1)$ and $B^d(1-\epsilon)$ is

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

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

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

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

# Geometry of the Hypercube

We have seen that most of the volume of the hypersphere is concentrated near its surface. A similar result 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.

**Proposition 3.** The hypercube $C^d$ has volume 1 and diameter $\sqrt{d}$.

**Remark.** Proposition 3 shows 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.

In 2D, the unit square is completely contained in the unit sphere. The distance from the center to a vertex (radius of the circumscribed sphere) is $\sqrt{2}/2$ and the apothem (radius of the inscribed sphere) is $1/2$.

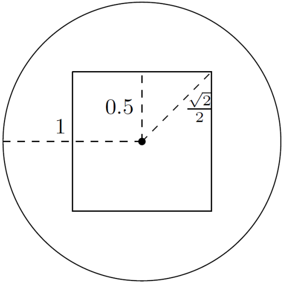

Figure: 2-dimensional unit sphere and unit cube, centered at the origin.

In 4D, the distance from the center to a vertex is 1 (note the diameter of the cube is 2 in 4D), so the vertices of the cube touch the surface of the sphere. However, the apothem is still 1/2. The result, when projected in 2D no longer appears convex, however all hypercubes are convex. This is part of the strangeness of higher dimensions – hypercubes are both convex and "pointy."

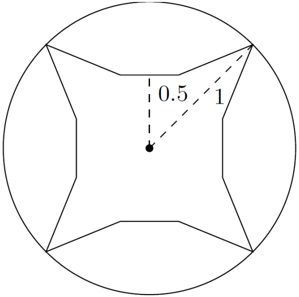

Figure: Projections of the 4D unit sphere and unit cube, centered at the origin (4 of the 16 vertices of the hypercube are shown).

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.

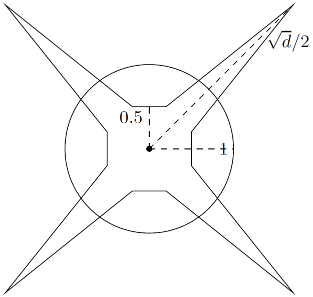

Figure: Projections of the $d$-dimensional unit sphere and unit cube, centered at the origin (4 of the $2^d$ vertices of the hypercube are shown).

**Most of the volume of the high-dimensional cube is located in its corners.**

Recall why $\ell_1$-regularization works?

# Basic Concepts from Probability

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

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

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 $\mathbb{E}[|X|^p]$. (Take the derivative of MGF w.r.t. $t$.)

- The $L^p$-norms of random variables is defined by taking the $p$-th root of moments:

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

with

$$\|X\|_\infty := ess \sup |X|.$$

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

For $p = 2$, $L^2$ is a Hilbert space with inner product and inner product and norm

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

respectively.

- The *standard deviation* $\sigma(X) := \sqrt{Var(X)}$ of $X$ can be written as

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

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

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle_{L^2}. \qquad (6)$$

## Classical inequalities for random variables

- *Holder's inequality:* for random variables $X$ and $Y$ on a common probability space and $p, q \geq 1$ with $1/p + 1/q = 1$, there holds

$$|\mathbb{E}[XY]| \leq \|X\|_{L^p} \|Y\|_{L^q}. \tag{7}$$

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

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[|X|^2]\mathbb{E}[|Y|^2]}. \tag{8}$$

- *Jensen's inequality:* for any random variable $X$ and a convex function $\phi : \mathbb{R} \to \mathbb{R}$:

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]. \tag{9}$$

Since $\phi(x) = x^{q/p}$ is convex for $q \geq p \geq 0$, it follows from Jensen's inequality that

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

## Classical inequalities for random variables

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

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}, \tag{10}$$

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 $\mathbb{P}\{X > t\} = 1 - F_X(t)$.
- The function $t \to \mathbb{P}\{|X| \geq t\}$ is called the *tail* of $X$.

The following lemma establishes a close connection between expectation and tails.

**Proposition 4.** (Integral identity). Let $X$ be a non-negative random variable. Then

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

Given an event $E$ with non-zero probability, $\mathbb{P}(\cdot|E)$ denotes conditional probability, furthermore for a random variable $X$ we use $\mathbb{E}[X|E]$ to denote the conditional expectation.

# Markov's inequality

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

**Proposition 5.** For any non-negative random variable $X : S \to \mathbb{R}$ we have

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all} \quad t > 0. \tag{11}$$

Proof.

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

## Chebyshev's inequality

**Corollary 6.** Let $X$ be a random variable with mean $\mu$ and variance $\sigma^2$. Then, for any $t > 0$

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}. \tag{12}$$

**Remark.** 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 guarantees that $X$ must be close to its mean $\mu$ whenever the variance of $X$ is small. Both, Markov's and Chebyshev's inequality are sharp, i.e., in general they cannot be improved.

# Chernoff bound

Markov's inequality only requires the existence of the first moment. We can say a bit more if in addition the random variable $X$ has a moment generating function in a neighborhood around zero, that is, there is a constant $b > 0$ such that $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists for all $\lambda \in [0, b]$. In this case we can apply Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$ and obtain the generic *chernoff bound*

$$\mathbb{P}\{X - \mu \geq t\} = \mathbb{P}\{e^{\lambda(X-\mu)} \geq e^{\lambda t}\} \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}. \tag{13}$$

Optimizing over $\lambda$ in order to obtain the tightest bound in (13) gives

$$\log \mathbb{P}\{X - \mu \geq t\} \leq - \sup_{\lambda \in [0,b]} \{\lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}]\}.$$

A very useful trick!

# Gaussian tail bound

A Gaussian random variable $X$ with mean $\mu$ and standard deviation $\sigma$ has a probability density function given by

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big( - \frac{(t-\mu)^2}{2\sigma^2} \Big). \tag{14}$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$. We call a Gaussian random variable $X$ with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$ a *standard Gaussian* or *standard normal* (random variable).

**Proposition 7.** [Gaussian tail bounds] Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then for all $t > 0$

$$\mathbb{P}(X \geq \mu + t) \leq e^{-t^2/2\sigma^2}. \tag{15}$$

## Gaussian tail bound

Proof. We use the moment-generating function $\lambda \to \mathbb{E}[e^{\lambda X}]$. A simple calculation gives

$$\mathbb{E}[e^{\lambda X}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x - x^2/2} dx = \frac{1}{\sqrt{2\pi}} e^{\lambda^2/2} \int_{-\infty}^{\infty} e^{-(x-\lambda)^2/2} dx = e^{\lambda^2/2},$$

where we have used the fact that $\int_{-\infty}^{\infty} e^{-(x-\lambda)^2/2} dx$ is just the entire Gaussian integral shifted and therefore its value is $\sqrt{2\pi}$. We now apply Chernoff's bound (13) and obtain $\mathbb{P}(X > t) \leq \mathbb{E}[e^{\lambda X}] e^{-\lambda t}$. Minimizing this expression over $\lambda$ gives $\lambda = t$ and thus $\mathbb{P}(X > t) \leq e^{-t^2/2}$.

# Sub-Gaussian random variables

**Definition 8.** A random variable $X$ with mean $\mu = \mathbb{E}[X]$ is called sub-Gaussian if there is a positive number $\sigma$ such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2}, \quad \text{for all } \lambda \in \mathbb{R}.$$

If $X$ satisfies the above definition, we also say that $X$ is sub-Gaussian with parameter $\sigma$, or $X$ is $(\mu, \sigma)$ sub-Gaussian.

Owing to the symmetry in the definition, $-X$ is sub-Gaussian if and only of $X$ is sub-Gaussian.

Any Gaussian random variable with variance $\sigma^2$ is sub-Gaussian with parameter $\sigma$.

## Sub-Gaussian tail bounds

**Proposition 9.** [Sub-Gaussian tail bounds] Assume $X$ is sub-Gaussian with parameter $\sigma$. Then for all $t > 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq e^{-t^2/2\sigma^2} \quad \text{for all } t \in \mathbb{R}. \tag{16}$$

Proof. Combining the moment condition in Def. 8 with calculations similar to those that lead us to the Gaussian tail bounds in Proposition. 7.

Example. An important example of non-Gaussian, but sub-Gaussian random variables are *Rademacher random variables*. A Rademacher random variable $\epsilon$ takes on the values $\pm 1$ with equal probability and is sub-Gaussian with parameter $\sigma$. Indeed, any bounded random variable is sub-Gaussian.

Many important random variables have a sub-Gaussian distribution, this class of random variables does not include several frequently occurring distributions with heavier tails. A classical example is the $\chi^2$ *distribution*.

Relaxing slightly the condition on the moment-generating function in Def 8 leads to the class of *sub-exponential* random variables.

**Definition 10.** A random variable $X$ with mean $\mu = \mathbb{E}[X]$ is called sub-exponential if there are parameters $\nu, b$ such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}, \quad \text{for all } |\lambda| \leq \frac{1}{b}.$$

Clearly, a sub-Gaussian random variable is sub-exponential (set $\nu = \sigma$ and $b = 0$, where $1/b$ is interpreted as $+\infty$). However, the converse is not true. Take for example $X \sim \mathcal{N}(0,1)$ and consider the random variable $Z = X^2$. For $\lambda < \frac{1}{2}$ it holds that

$$\mathbb{E}[e^{\lambda(Z-1)}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(x^2-1)} e^{-x^2/2} dx = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}. \tag{17}$$

However, for $\lambda \geq \frac{1}{2}$ the moment-generating function does not exist, which implies that $X^2$ is not sub-Gaussian/ But $X^2$ is sub-exponential. Indeed, a brief computation shows that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}, \quad \text{for all } |\lambda| \leq 1/4,$$

which in turn implies that $X^2$ is sub-exponential with parameters $(\nu, b) = (2, 4)$.

## Sub-exponential tail bounds

**Proposition 11.** [Sub-exponential tail bounds] Assume $X$ is sub-exponential with parameters $(\nu, b)$. Then

$$\mathbb{P}(X \geq \mu + t) \leq \begin{cases} e^{-t^2/2\nu^2} & \text{if } 0 \leq t \leq \frac{\nu^2}{b}, \\ e^{-t/2b} & \text{if } t > \frac{\nu^2}{b}. \end{cases} \tag{18}$$

## Comments on sub-Gaussian and sub-exponential random variables

Both sub-Gaussian and sub-exponential properties are preserved under summation for independent random variables and the associated parameters transform in a simple manner.

A collection $X_1, \cdots, X_n$ of mutually independent random variables that all have the same distribution is called independent identically distributed (i.i.d.). A random variable $X'$ is called an independent copy of $X$ if $X$ and $X'$ are independent and have the same distribution.

In general, we cannot improve Markov's inequality and Chebyshev's inequality, 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 random variable that is the *sum of a number of independent random variables*. This leads to the rich topic of *concentration inequalities*.

## An example

If $X_1, \cdots, X_n$ are independent, standard normal random varables, then the sum of their squares, $Z = \sum_{k=1}^n X_k^2$ is distributed according to the $\chi^2$ distribution with $n$ degrees of freedom. We denote this by $Z \sim \xi^2(n)$. Its probability density function is

$$\phi(t) = \begin{cases} \frac{t^{\frac{n}{2}-1}e^{-\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}, & t > 0. \\ 0 & \text{else.} \end{cases}$$

Since the $X_k^2, k = 1, \cdots, n$ are sub-exponential with parameters $(2, 4)$ and independent, $Z = \sum_{k=1}^n X_k^2$ is sub-exponential with parameters $(2\sqrt{n}, 4)$. Therefpre, using (18), we obtain the $\chi^2$ tail bound

$$\mathbb{P}\left(\frac{1}{n}\Big|\sum_{k=1}^n X_k^2 - 1\Big| \geq t\right) \leq \begin{cases} 2e^{-nt^2/8} & \text{for } \ t \in (0,1). \\ 2e^{-nt/8} & \text{if } t \geq 1. \end{cases} \qquad (19)$$

# 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. We would also 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.

Concentration inequalities are instances of what is sometimes called *Blessings of dimensionality*. 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.

## Large Deviation Inequalities

Concentration and large deviations inequalities are among the most useful tools when understanding the performance of some algorithms. We start with two of the most fundamental results in probability.

**Theorem 12.** [Strong law of large numbers] Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with mean $\mu$. Denote

$$S_n := X_1 + \cdots + X_n.$$

Then, as $n \to \infty$

$$\frac{S_n}{n} \to \mu \;\text{ almost surely.} \tag{20}$$

**Theorem 13.** [Lindeberg-Levy Central limit theorem] Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with mean $\mu$ and variance $\sigma^2$. Denote

$$S_n := X_1 + \cdots + X_n,$$

and consider the normalized random variable $Z_n$ with mean zero and variance one, given by

$$Z_n := \frac{S_n - \mathbb{E}[S_n]}{\sqrt{VarS_n}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Then, as $n \to \infty$

$$Z_n \to \mathcal{N}(0,1) \text{ in distribution.} \tag{21}$$

## Large Deviation Inequalities

The strong law of large numbers and the central limit theorem give us qualitative statements about the behavior of a sum of i.i.d. random variables. In many applications it is desirable to be able to quantify how such a sum deviates around its mean. This is where concentration inequalities come into play.

The intuitive idea is that if we have a sum of independent random variables

$$X = X_1 + \cdots + X_n,$$

where $X_i$ are i.i.d. centered random variables, then while the value of $X$ can be of order $\mathcal{O}(n)$ it will very likely be of order $\mathcal{O}(\sqrt{n})$ (the order of its standard deviation). The inequalities that follow are ways of very precisely controlling the probability of $X$ being larger (or smaller) than $\mathcal{O}(\sqrt{n})$ While we could use, for example, Chebyshev's inequality for this, in the inequalities that follow the probabilities will be exponentially small, rather than just quadratically small, which will be crucial in many applications to come. Moreover, classical central limit theorem, the concentration inequalities below are non-asymptotic in the sense that they hold for all fixed $n$ and not just for $n \to \infty$.

# Large Deviation Inequalities

**Theorem 14.** [Hoeffding's Inequality] Let $X_1, \cdots, X_n$ be independent bounded random variables, i.e. $|X_i| \leq a_i$ and $\mathbb{E}[X_i] = 0$. Then,

$$\mathbb{P}\left\{\Big|\sum_{i=1}^{n} X_i\Big| > t\right\} \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\right).$$

Remark. The inequality implies that fluctuations larger than $\mathcal{O}(\sqrt{n})$ have small probability. For example, if $a_i = a$ for all $i$, setting $t = a\sqrt{2n\log n}$ yields that the probability is at most $\frac{2}{n}$.

Proof. We prove the result for the case $|X_i| \leq a$, the extension to the case $|X_i| \leq a_i$ is straightforward. We first get a probability bound for the event $\sum_{i=1}^{n} X_i > t$. The proof, again, will follow from Markov. Since we want an exponentially small probability, we use a classical trick that involves exponentiating with any $\lambda > 0$ and then choosing the optimal $\lambda$.

$$\begin{aligned} \mathbb{P}\Big\{ \sum_{i=1}^{n} X_i > t \Big\} &= \mathbb{P}\Big\{ \sum_{i=1}^{n} X_i > t \Big\} && (21) \\ &= \mathbb{P}\Big\{ e^{\lambda \sum_{i=1}^{n} X_i} > e^{\lambda t} \Big\} \leq \frac{\mathbb{E}[e^{\lambda \sum_{i=1}^{n} X_i}]}{e^{t\lambda}} \\ &= e^{-t\lambda} \prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}] && (22), \end{aligned}$$

where the penultimate step follows from Markov's inequality and the last equality follows from independence of the $X_i$'s.

We now use the fact that $|X_i| \leq a$ to bound $\mathbb{E}[e^{\lambda X_i}]$. Because the function $f(x) = e^{\lambda x}$ is convex,

$$e^{\lambda x} \leq \frac{a+x}{2a} e^{\lambda a} + \frac{a-x}{2a} e^{-\lambda a}, \quad \text{for all } x \in [-a, a].$$

Since, for all $i$, $\mathbb{E}[X_i] = 0$ we get

$$\mathbb{E}[e^{\lambda X_i}] \leq \mathbb{E}\Big[\frac{a+X_i}{2a} e^{\lambda a} + \frac{a-X_i}{2a} e^{-\lambda a}\Big] \leq \frac{1}{2}(e^{\lambda a} + e^{-\lambda a}) = cosh(\lambda a).$$

Since $\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}$, $e^{x^2/2} = \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n n!}$, and $(2n)! \geq 2^n n!$, we have

$$\cosh(x) \leq e^{x^2/2}, \quad \text{for all} \ \ x \in \mathbb{R}.$$

Hence,

$$\mathbb{E}[e^{\lambda X_i}] \leq e^{(\lambda a)^2/2}.$$

Together with (21), this gives

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} \leq e^{-t\lambda}\prod_{i=1}^{n} e^{(\lambda a)^2/2} = e^{-t\lambda}e^{n(\lambda a)^2/2}.$$

This inequality holds for any choice of $\lambda \geq 0$, so we choose the value of $\lambda$ that minimizes

$$\min_{\lambda}\left\{n\frac{(\lambda a)^2}{2} - t\lambda\right\}.$$

Differentiating readily shows that the minimizer is given by

$$\lambda = \frac{t}{na^2},$$

which satisfies $\lambda > 0$. For this choice of $\lambda$,

$$n(\lambda a)^2/2 - t\lambda = \frac{1}{n}\Big(\frac{t^2}{2a^2} - \frac{t^2}{a^2}\Big) = -\frac{t^2}{2na^2}.$$

Thus,

$$\mathbb{P}\Big\{ \sum_{i=1}^{n} X_i > t \Big\} \leq e^{-\frac{t^2}{2na^2}} .$$

By using the same argument on $\sum_{i=1}^{n}(-X_i)$, and union bounding over the two events we get,

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^{n} X_i \Big| > t \Big\} \leq 2e^{-\frac{t^2}{2na^2}} ,$$

which concludes the proof.

Remark. Hoeffding's inequality is suboptimal in a sense we now describe. Let us say that we have random variables $r_1, \cdots, r_n$ i.i.d. distributed as

$$r_i = \begin{cases} -1 & \text{with probability } p/2 \\ 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p/2. \end{cases}$$

Then, $\mathbb{E}(r_i) = 0$ and $|r_i| \leq 1$ so Hoeffding's inequality gives:

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n} r_i\Big| > t\Big\} \leq 2\exp\Big(-\frac{t^2}{2n}\Big).$$

Intuitively, the smaller $p$ is, the more concentrated $|\sum_{i=1}^{n} r_i|$ should be, however HOeffding's inequality does not capture this behaviour.

A natural way to capture this behaviour is by noting that the variance of $\sum_{i=1}^{n} r_i$ depends on $p$ as $Var(r_i) = p$. The inequality that follows, Bernstein's inequality, uses the variance of the summands to improve over Hoeffding's inequality.

The way this is going to be achieved is by strengthening the proof above, more specifically in step (22) we will use the bound on the variance to get a better estimate on $\mathbb{E}[e^{\lambda X_i}]$ essentially by realizing that if $X_i$ is centered, $\mathbb{E}X_i^2 = \sigma^2$, and $|X_i| \leq a$ then, for $k \geq 2$,

$$\mathbb{E}X_i^k \leq \mathbb{E}|X_i|^k \leq \sigma^2 \mathbb{E}|X_i|^{k-2} \leq \sigma^2 a^{k-2} = \left(\frac{\sigma^2}{a^2}\right) a^k.$$

## Large Deviation Inequalities

**Theorem 16.** [Bernstein's inequality] Let $X_1, \cdots, X_n$ be independent centered bounded random variables satisfying $|X_i| \leq a$ and $\mathbb{E}[X_i^2] = \sigma^2$. Then,

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n} X_i\Big| > t\Big\} \leq 2\exp\Big(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\Big).$$

Remark. For the random variables $r_1, \cdots, r_n$ i.i.d. distributed as

$$r_i = \begin{cases} -1 & \text{with probability } p/2 \\ 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p/2. \end{cases}$$

we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n} r_i\Big| > t\Big\} \leq 2\exp\Big(-\frac{t^2}{2np + \frac{2}{3}t}\Big),$$

which depends on $p$; for small $p$, it is considerably smaller than what Hoeffding's inequality gives.

## Bernstein's inequality

Proof. As before, we will prove

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} \leq \exp\Big(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\Big),$$

and then union bound with the same result for $-\sum_{i=1}^{n} X_i$, to prove the Theorem. For any $\lambda > 0$ we have

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} = \mathbb{P}\{e^{\lambda \sum X_i} > e^{\lambda t}\} \leq \frac{\mathbb{E}[e^{\lambda \sum X_i}]}{e^{\lambda t}} = e^{-\lambda t} \prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}].$$

The following calculation reveals the source of the improvement over Hoeffding's inequality:

$$\mathbb{E}[e^{\lambda X_i}] = \mathbb{E}\Big[1 + \lambda X_i + \sum_{m=2}^{\infty} \frac{\lambda^m X_i^m}{m!}\Big] \leq 1 + \sum_{m=2}^{\infty} \frac{\lambda^m a^{m-2}\sigma^2}{m!}$$

$$= 1 + \frac{\sigma^2}{a^2} \sum_{m=2}^{\infty} \frac{(\lambda a)^m}{m!} = 1 + \frac{\sigma^2}{a^2}\big(e^{\lambda a} - 1 - \lambda a\big).$$

Therefore,

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} \leq e^{-\lambda t}\Big[1 + \frac{\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a)\Big]^n.$$

We will use a few simple inequalities (that can be easily proved with calculus) such as $1 + x \leq e^x$, for all $x \in \mathbb{R}$. This means that

$$1 + \frac{\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a) \leq e^{\frac{\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a)},$$

which readily implies

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} \leq e^{-\lambda t} e^{\frac{n\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda)}.$$

As before, we try to find the value of $\lambda > 0$ that minimizes

$$\min_{\lambda} \Big\{ -\lambda t + \frac{n\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a) \Big\}.$$

Differentiation gives

$$-t + \frac{n\sigma^2}{a^2}(a e^{\lambda a} - a) = 0,$$

which implies that the optimal choice of $\lambda$ is given by

$$\lambda^* = \frac{1}{a} \log\Big(1 + \frac{at}{n\sigma^2}\Big).$$

If we set

$$u = \frac{at}{n\sigma^2}, \tag{22}$$

then $\lambda^* = \frac{1}{a}\log(1+u)$.

Now, the value of the minimum is given by

$$-\lambda^* t + \frac{n\sigma^2}{a^2}(e^{\lambda^* a} - 1 - \lambda^* a) = -\frac{n\sigma^2}{a^2}[(1+u)\log(1+u) - u].$$

This means that

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} \le \exp\Big(-\frac{n\sigma^2}{a^2}\{(1+u)\log(1+u) - u\}\Big).$$

The rest of the proof follows by noting that, for every $u > 0$,

$$(1+u)\log(1+u) - u \ge \frac{u}{\frac{2}{u} + \frac{2}{3}}, \tag{23}$$

which implies

$$\mathbb{P}\Big\{\sum_{i=1}^{n} X_i > t\Big\} \le \exp\Big(-\frac{n\sigma^2}{a^2}\frac{u}{\frac{2}{u} + \frac{2}{3}}\Big) = \exp\Big(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\Big).$$

# The Geometry of the Hypercube Revisited

**Theorem 18.** Almost all the volume of the high-dimensional cube is located in its corners.

Remark. The proof of this statement will be based on a probabilistic argument, thereby illustrating (again) the nice and fruitful connection between geometry and probability 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 $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$ 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

$$\mathbb{E}[z_i] = \frac{1}{2}\int_{-1}^{1} t^2 dt = \frac{1}{3} \Rightarrow \mathbb{E}[\|x\|_2^2] = \frac{d}{3}; \quad Var(z_i) = \frac{1}{2}\int_{-1}^{1} t^4 dt - (\frac{1}{3})^2 = \frac{1}{5} - \frac{1}{9} \leq \frac{1}{10}.$$

Using Hoeffding's inequality,

$$\mathbb{P}(\|x\|_2^2 \leq 1) = \mathbb{P}(\sum_{i=1}^{d} x_i^2 \leq 1) = \mathbb{P}(\sum_{i=1}^{d}(z_i - \mathbb{E}[z_i]) \leq 1 - \frac{d}{3}) \leq \exp\left[-\frac{(\frac{d}{3} - 1)^2}{2d(\frac{2}{3})^2}\right] \leq \exp(-\frac{d}{9}),$$

for sufficiently large $d$.

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.

Remark. 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 “pointines” of the hypercube (in form of the $\ell_1$-ball) turns out to very useful in the areas of compressive sensing and sparse recovery.

# 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 slides 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.

Picking a point $x$ uniformly at random on the sphere $S^{d-1}$ is not too different from picking a vector at random with entries of the form $(\pm\frac{1}{\sqrt{d}}, \cdots, \pm\frac{1}{\sqrt{d}})$, since every point on the sphere has to fulfill $x_1^2 + \cdots + x_d^2 = 1$, hence the "average magnitude" of $x_i$ will be $\frac{1}{\sqrt{d}}$.

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. W.L.O.G. we pick an arbitrary unit vector $x_1$ which represents the "north pole", and the intersection of the sphere with the plane $x_1 = 0$ forms our equator. We extend $x_1$ to an orthonormal basis $x_1, \cdots, x_d$.

We create a random vector by sampling $(Z_1, \cdots, Z_d) \sim \mathcal{N}(0, I_d)$ and normalize the vector to get $X = (X_1, \cdots, X_d) = \frac{1}{\sum_{k=1}^d Z_k^2}(Z_1, \cdots, Z_d)$. 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 $\mathbb{E}[\sum_{k=1}^d \langle X, x_k \rangle^2] = \mathbb{E}[1] = 1$. Thus, by symmetry, $\mathbb{E}[\langle X, x_1 \rangle^2] = \frac{1}{d}$. Applying Markov's inequality (11) gives

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

For fixed $\epsilon$ 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.

# Random Vectors in High Dimensions

Two basic geometric questions from a probabilistic point of view are:

- What length do we expect a random vector $x \in \mathbb{R}^n$ to have?
- What angle do we expect two random vectors $x, y \in \mathbb{R}^n$ to have?

Suppose that the coordinates $x_1, \cdots x_n$ of $x$ are independent random variables with zero mean and unit variances (and similarly for $y$). It holds that

$$\mathbb{E}\|x\|_2^2 = \mathbb{E}\Big[\sum_{k=1}^{n} |x_k|^2\Big] = \sum_{k=1}^{n} \mathbb{E}[|x_k|^2] = n.$$

Hence, we expect the typical length $\|x\|_2$ of $x$ to be approximately $\sqrt{n}$. But how well does the length of a random vector concentrate around its "typical length"?

Assume for instance the entries $x_k \sim \mathcal{N}(0,1)$. In this case we can use the $\chi^2$-concentration bound (19), which gives

$$\mathbb{P}\Big(\Big|\frac{1}{n}\|x\|_2^2 - 1\Big| \geq t\Big) \leq 2\exp\Big(-\frac{n}{8}\min(t, t^2)\Big). \tag{24}$$

This represents a concentration inequality for $\|x\|_2^2$, but we aim for a concentration inequality for the length $\|x\|$. To do this, we use the following elementary observation that holds for all $z \geq 0$:

$$|z-1| \geq \delta \;\text{ implies }\; |z^2 - 1| \geq \max(\delta, \delta^2).$$

Using this observation we obtain for any $\delta > 0$ that

$$\mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\|x\|_2^2 - 1\Big| \geq \delta\Big) \leq \mathbb{P}\Big(\Big|\frac{1}{n}\|x\|_2^2 - 1\Big| \geq \max(\delta, \delta^2)\Big) \leq 2e^{-nt^2/8}, \tag{25}$$

where we have used $t = \max(\delta, \delta^2)$ in (24).

With some minor modifications of these steps (and a slightly different constant) one can extend this result to random vectors with sub-Gaussian coordinates.

We now turn our attention to the expected angle between two random vectors. We will show that two randomly drawn vectors in high dimensions are almost perpendicular. The following theorem quantifies this statements. We denote the angle $\theta_d$ between two vectors $x, y$ by $\theta_{x,y}$ and recall that $\cos\theta_{x,y} = \frac{\langle x,y\rangle}{\|x\|_2\|y\|_2}$.

**Theorem 19.** Let $x, y \in \mathbb{R}^d$ be two random vectors with i.i.d. Rademacher variables, i.e. the entries $x_i, y_i$ take values $\pm 1$ with equal probability. Then

$$\mathbb{P}\Big(|\cos\theta_{x,y}| \geq \sqrt{\frac{2\log d}{d}}\Big) \leq \frac{2}{d}. \tag{26}$$

Proof. Note that $\langle x, y\rangle = \sum_i x_i y_i$ is the sum of i.i.d. Rademacher variables. Hence, $\mathbb{E}[\langle x, y\rangle] = \sum_i \mathbb{E}[x_i y_i] = 0$. Therefore, we can apply Hoeffding's inequality. For any given $t > 0$

$$\mathbb{P}(|\langle x, y\rangle| \geq t) = \mathbb{P}\Big(\frac{|\langle x, y\rangle|}{\|x\|_2\|y\|_2} \geq \frac{t}{d}\Big) \leq 2\exp\Big(-\frac{t^2}{2d}\Big).$$

To establish the bound (26), we set $t = \sqrt{2d\log d}$ and obtain

$$\mathbb{P}\Big(|\cos\theta_{x,y}| > \sqrt{\frac{2\log d}{d}}\Big) = \mathbb{P}\Big(\frac{|\langle x, y\rangle|}{d} \geq \sqrt{\frac{2\log d}{d}}\Big) \leq 2\exp(-\log d) = \frac{2}{d}.$$

Remark. It is not surprising that a similar result holds for Gaussian random vectors in $\mathbb{R}^d$ or random vectors chosen from the sphere $S^{d-1}$. Indeed, even more is true. While we can have only $d$ vectors that are exactly orthogonal in $\mathbb{R}^d$, for large $d$ we can have exponentially many vectors that are almost orthogonal in $\mathbb{R}^d$. To see this we return to the setting of Theorem 19, choosing $m$ random vectors $x_1, \cdots, x_m$ with i.i.d. Rademacher variables as their entries. We proceed as in the proof of Theorem 19 but let $t = \sqrt{2d \log c}$ where $c > 0$ is a constant. This yields

$$\mathbb{P}\left(|\cos \theta_{x_i,x_j}| \geq \sqrt{\frac{2 \log c}{d}}\right) \leq \frac{2}{c}.$$

Note that we need to consider $\theta_{x_i,x_j}$ for $(m^2 - m)/2$ such pairs $(x_i, x_j)$. To make things concrete, we can set for instance $m = \sqrt{c}/4$. Using the union bound we obtain that with probability at least $7/8$ it holds that

$$\max_{i,j,i \neq j} |\cos \theta_{x_i,x_j}| \leq \sqrt{\frac{2 \log c}{d}}.$$

Choose e.g. $c = e^{d/200}$ and obtain that we have exponentially many (w.r.t. $d$) vectors in $\mathbb{R}^d$ that are almost orthogonal in the sense that the cosine of their pairwise angle is at most $1/200$.