Lecture Notes
Systems & Control Engineering, IIT Bombay

The Concentration Phenomenon: an Elementary Introduction

Debasish Chatterjee

Abstract

This is a set of telegraphic notes to accompany the course material of the Autumn 2022 avatar of the course SC629: Introductory Probability and Random Processes, Systems & Control Engineering, IIT Bombay. The contents consist of a *very* elementary treatment of the concentration phenomenon in probability theory. Most of the material has been collected from sources scattered in the literature and these notes have *not* been edited and/or proofread carefully; several errors appear in these notes with high probability. The recent textbooks [**Ver18**, **Wai19**] should be consulted in tandem.

The Concentration Phenomenon: an Elementary Introduction

Debasish Chatterjee
Systems & Control Engineering
IIT Bombay, Powai
Mumbai 400076
✉ dchatter@iitb.ac.in
⌂ https://www.sc.iitb.ac.in/~chatterjee

Contents

Prologue

> The first trick of the trade in concentration phenomena is to get rid of those two annoying things that scientists bring to the table — logic and common sense; concentration is a matter of the right brain only.
>
> Folklore.

> We are suspicious of “intuitive mathematical truth” and we do not trust *meta*mathematical rigor of formal logic.
>
> Misha Gromov.

« The concentration phenomenon in probability theory provides a gamut of tools for *non-asymptotic* estimates. It is a fair statement that from the perspective of engineers, probability theory consists of mathematical models of certain experiences or controlled experiments with the property that they are repeatable. Practical considerations, of course, dictate the number of possible repetitions.[1] There are only so many samples that can be drawn, only so many experiments repeated, and only so many computations that can possibly be carried out, before one must stop gathering data and get on with other tasks. In this scheme of things, devices that provide non-asymptotic guarantees and estimates are of crucial importance in engineering, and as such, concentration estimates are expected to play foundational roles in all engineering discplines that interact with data.

Concentration estimates have rarely been employed directly in control theory.[2] Apart from an early application in stochastic predictive control [**HCL13**], not much work appears to have focussed on exploiting concentration phenomena in this field. The more recent work [**MCB20**] explored the effect of concentration of high-dimensional random vectors in the so-called “scenario approach” (see, e.g., [**Ram18**] for a lucid introduction to the subject) to robust optimization, and the more recent work [**DACC22**] established an algorithm to mitigate the effects of concentration phenomena in a broad class of robust optimization problems.

Let us now jump to the probabilistic stuff.

> Knowledge I possess of the game of dice, in numbers thus am I skilled.
>
> Rituparna, King of Ayodhya;
> *nalopAkhyAnam*, Mahabharata.

All random variables appearing in these notes are defined on some fixed probability space that is sufficiently rich to carry a countable collection of independent random variables taking values on the unit interval $[0, 1]$.

[1] “Civilizations have finite lifetimes.” H. Witsenhausen.

[2] A search on Google Scholar with ‘concentration of measures’ and ‘control theory’ *anywhere* fetched articles mostly from the biological (and other softer) sciences on 29 Aug 2020.

§1. Volumes of Euclidean balls

«LET $d \in \mathbb{N}^*$ be fixed, and let

$$\mathsf{B}^d(y, r) := \left\{ x \in \mathbb{R}^d \;\middle|\; \|x - y\| < r \right\}$$

be the *Euclidean open ball* of radius r centered at $y \in \mathbb{R}^d$ (and let $\mathsf{B}^d[y, r]$ stand for the *Euclidean closed ball* of radius r centered at $y \in \mathbb{R}^d$). The **volume** of the Euclidean open/closed *unit* ball centered at $0 \in \mathbb{R}^d$ is given by

$$\operatorname{vol} \mathsf{B}^d(0, 1) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2} \cdot \Gamma(\frac{d}{2})}, \tag{1.1}$$

where $\Gamma(\cdot)$ is the standard Gamma function defined by

$$]0, +\infty[\ni s \mapsto \Gamma(s) := \int_0^{+\infty} t^{s-1} \mathrm{e}^{-t} \, \mathrm{d}t.$$

(A few features of the Gamma function are given in §B; the only property of the Gamma function needed here is that $\Gamma(s + 1) = s\Gamma(s)$ for $s > 0$, which immediately establishes the second equality in (1.1).) Since the d-dimensional volume is homogeneous of order d, the volume of the open ball of radius $r > 0$ centered at $0 \in \mathbb{R}^d$ is

$$\operatorname{vol} \mathsf{B}^d(0, r) = r^d \operatorname{vol} \mathsf{B}^d(0, 1) = r^d \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)}. \tag{1.2}$$

By translation independence of volumes,

$$\operatorname{vol} \mathsf{B}^d(y, r) = \operatorname{vol} \mathsf{B}^d(0, r) \quad \text{for all } y \in \mathbb{R}^d.$$

Of course, the aforementioned volumes remain unchanged if *closed* balls $\mathsf{B}^d[y, r]$ replace *open* balls $\mathsf{B}^d(y, r)$.

Here is an interesting short proof of (1.1) due to Svante Janson:[3]

§1.1. The proof of the formula for the volume of Euclidean balls.

The centerpiece of the proof is the formula

$$\int_{\mathbb{R}} \mathrm{e}^{-y^2} \, \mathrm{d}y = \sqrt{\pi} \tag{1.3}$$

coupled with the observation that

$$\int_a^{+\infty} \mathrm{e}^{-t} \, \mathrm{d}t = \mathrm{e}^{-a} \quad \text{for } a \geqslant 0.$$

It follows from the first formula that on the one hand,

$$\int_{\mathbb{R}^d} \mathrm{e}^{-\|y\|^2} \, \mathrm{d}y = \int_{\mathbb{R}^d} \mathrm{e}^{-\sum_{k=1}^d y_k^2} \, \mathrm{d}y_1 \cdots \mathrm{d}y_d = \prod_{k=1}^d \int_{\mathbb{R}} \mathrm{e}^{-y_k^2} \, \mathrm{d}y_k = \pi^{\frac{d}{2}}.$$

On the other hand, the second formula shows that

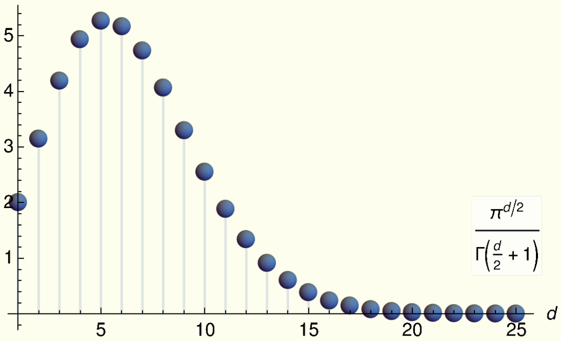
$$\begin{aligned} \int_{\mathbb{R}^d} \mathrm{e}^{-\|y\|^2} \, \mathrm{d}y &= \int_{\mathbb{R}^d} \Big(\int_{\|y\|^2}^{+\infty} \mathrm{e}^{-t} \, \mathrm{d}t \Big) \mathrm{d}y = \int_{\mathbb{R}^d} \Big(\int_0^{+\infty} \mathrm{e}^{-t} \mathbb{1}_{[\|y\|^2, +\infty[}(t) \, \mathrm{d}t \Big) \mathrm{d}y \\ &= \int_0^{+\infty} \mathrm{e}^{-t} \Big(\int_{\mathbb{R}^d} \mathbb{1}_{\mathsf{B}^d(0, \sqrt{t})}(y) \, \mathrm{d}y \Big) \mathrm{d}t = \int_0^{+\infty} \mathrm{e}^{-t} t^{\frac{d}{2}} \, \mathrm{d}t \cdot \operatorname{vol} \mathsf{B}^d(0, 1), \end{aligned}$$

[3] An alternative and more direct approach may be found in [Lan97, Exercise 3, p. 598].

where we have employed, respectively, Fubini's theorem [Zor16, §11.4.1] to interchange the integrals and homogeneity of order d of the Euclidean volume in $\mathbb{R}^d$ in the last two steps.[4] Since $\Gamma(\frac{d}{2}+1) = \int_0^{+\infty} \mathrm{e}^{-t} t^{\frac{d}{2}}\,\mathrm{d}t$, we have

$$\text{(1.4)} \qquad \operatorname{vol} \mathsf{B}^d(0,1) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}.$$

It follows at once from (1.1) that $\operatorname{vol} \mathsf{B}^d(0,1) \xrightarrow[d\to+\infty]{} 0$. Here is a figure describing the behavior of $\operatorname{vol} \mathsf{B}^d(0,1)$ against d:[5]



The proof above also provides an expression of the (hyper-)area of the $(d-1)$-dimensional unit sphere: Indeed, since $\mathbb{S}^{d-1} := \{y \in \mathbb{R}^d \mid \|y\| = 1\}$ is the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$, by homogeneity of degree $(d-1)$ of the $(d-1)$-dimensional area in $\mathbb{R}^d$,

$$\operatorname{vol} \mathsf{B}^d(0,1) = \int_0^1 \left(\operatorname{area} \mathbb{S}^{d-1}\right) r^{d-1}\,\mathrm{d}r = \frac{\operatorname{area} \mathbb{S}^{d-1}}{d},$$

which leads to the formula[6]

$$\text{(1.5)} \qquad \operatorname{area} \mathbb{S}^{d-1} = d \cdot \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}.$$

(1.6). Exercise. The preceding proof assumes the identity (1.3). However, if (1.3) is not taken for granted, then one can define the integral in (1.3) to be some quantity, say, $I > 0$ (the integral obviously converges), and arrive at the identity $\operatorname{vol} \mathsf{B}^d(0,1) = \frac{I^d}{\Gamma(\frac{d}{2}+1)}$ in place of (1.4). Substitute an appropriate value of d in the preceding identity to prove (1.3).

§1.2. Some consequences of concentration of volumes in $\mathbb{R}^d$.

(1.7) (Concentration of volume at the boundary). The first immediate consequence of (1.2) is that if ε is a small number (i.e., $|\varepsilon|$ is small), then

$$\frac{\operatorname{vol} \mathsf{B}^d(0, r+\varepsilon)}{\operatorname{vol} \mathsf{B}^d(0,r)} = \frac{(r+\varepsilon)^d}{r^d}.$$

In particular, if we measure in SI units, $d = 10^3$, and $r = 1$m, then the removal of a shell of thickness 10^{-2}m (i.e., 1cm) from $\mathsf{B}^d(0,1)$ leaves $\left(1-10^{-2}\right)^{10^3} = \left(\frac{99}{100}\right)^{10^3} \approx 0.4 \times 10^{-4}$ fraction of its original volume behind in the ball of radius 0.99m. To wit, an overwhelming fraction of the

[4] See [Lan97, Chapter XX] for a detailed and lucid treatment of multiple integration.

[5] See https://mathworld.wolfram.com/Ball.html for further details.

[6] See https://mathworld.wolfram.com/Sphere.html for further details.

volume of high-dimensional Euclidean balls is concentrated tightly around the boundary — an example of **concentration** of volumes in high dimensions.

(1.8). It is interesting to note that, in contrast to Euclidean balls, the volume of the open unit cube $]-1,1[^d$ (or the closed unit cube $[-1,1]^d$) grows (exponentially) as 2^d with d. Of course, $\mathsf{B}^d(0,1) \subset\,]-1,1[^d$ and $\mathsf{B}^d[0,1] \subset [-1,1]^d$ by definition, which gives an early clue that drawing independently from the uniform distribution on $]-1,1[^d$ or $[-1,1]^d$ will hardly ever fetch samples from the inscribed unit Euclidean ball as d becomes large. The ratio $\frac{\mathrm{vol}([-1+\varepsilon,1-\varepsilon]^d)}{\mathrm{vol}([-1,1]^d)}$ is $(1-\varepsilon)^d$ and this quantity goes to 0 exponentially fast with d for any fixed $\varepsilon \in\,]0,1[$, which shows that the volume of high-dimensional boxes also concentrate tightly around the boundary.

(1.9). Exercise. Recall that $\mathrm{vol}\,\mathsf{B}^d(0,r) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\cdot\Gamma(\frac{d}{2})} \cdot r^d$ for all $r > 0$ and $d \in \mathbb{N}^*$. Comment on the asymptotics, as d becomes large, of the function $d \mapsto r(d)$ that ensure $\mathrm{vol}\,\mathsf{B}^d\big(0, r(d)\big) = 1$. [*Hint*: Use Stirling's approximation $n! \approx \sqrt{2\pi n}\,\big(\frac{n}{\mathrm{e}}\big)^n$ for large values of $n \in \mathbb{N}^*$ (see §B and esp. (B.2)) to approximate the factorial for large values of d.]

(1.10) (Volumes of equatorial disks). Define the subset of $\mathsf{B}^d(0,1)$ that lies between the two hyperplanes

$$H_{-\varepsilon} := \big\{y \in \mathbb{R}^d \,\big|\, y_1 = -\varepsilon\big\} \quad \text{and} \quad H_{\varepsilon} := \big\{y \in \mathbb{R}^d \,\big|\, y_1 = \varepsilon\big\}$$

straddling the equator orthogonal to the y_1-direction. In other words, define the ε-strip $S(|\varepsilon|, d)$ of the unit ball

$$S(|\varepsilon|, d) := \big\{y \in \mathsf{B}^d(0,1) \,\big|\, |y_1| < \varepsilon\big\}.$$

Let us determine (tight bounds of) the ratio $\frac{\mathrm{vol}\,S(|\varepsilon|,d)}{\mathrm{vol}\,\mathsf{B}^d(0,1)}$ for large d. Of course,

$$\begin{aligned}
\mathrm{vol}\big(\mathsf{B}^d(0,1) \setminus S(|\varepsilon|,d)\big) &= 2 \cdot \int_{\varepsilon}^{1} \mathrm{vol}\,\mathsf{B}^{d-1}(0,1) \cdot \big((1-r^2)^{\frac{1}{2}}\big)^{d-1}\,\mathrm{d}r \\
&= 2 \cdot \frac{\pi^{\frac{d-1}{2}}}{\frac{d-1}{2}\cdot\Gamma(\frac{d-1}{2})} \cdot \int_{\varepsilon}^{1} (1-r^2)^{\frac{d-1}{2}}\,\mathrm{d}r \\
&= 2 \cdot \frac{\pi^{\frac{d-1}{2}}}{\frac{d-1}{2}\cdot\Gamma(\frac{d-1}{2})} \cdot \int_{\varepsilon}^{1} \mathrm{e}^{\frac{d-1}{2}\ln(1-r^2)}\,\mathrm{d}r.
\end{aligned}$$

Let $[\varepsilon, 1[\, \ni t \mapsto g(t) := \ln(1-t^2) \in \mathbb{R}$. Clearly, g is twice continuously differentiable, and attains its unique maximum at $t = \varepsilon$. An application of Theorem (A.1)-a) shows that

$$\int_{\varepsilon}^{1} \mathrm{e}^{\frac{d-1}{2}\ln(1-r^2)}\,\mathrm{d}r = \frac{1-\varepsilon^2}{2\varepsilon} \cdot \mathrm{e}^{\frac{d-1}{2}\ln(1-\varepsilon^2)} \cdot \Big(\frac{d-1}{2}\Big)^{-1} \cdot \Big(1 + O\Big(\frac{2}{d-1}\Big)\Big) \quad \text{as } d \to +\infty,$$

and consequently,

$$\mathrm{vol}\big(\mathsf{B}^d(0,1) \setminus S(|\varepsilon|,d)\big) = \frac{\pi^{\frac{d-1}{2}}}{(\frac{d-1}{2})^2\cdot\Gamma(\frac{d-1}{2})} \cdot \frac{(1-\varepsilon^2)^{\frac{d+1}{2}}}{\varepsilon} \cdot \Big(1 + O\Big(\frac{2}{d-1}\Big)\Big) \quad \text{as } d \to +\infty.$$

From this stage it follows at once that irrespective of the value of $\varepsilon \in\,]0,1[$,

$$\frac{\mathrm{vol}\,S(|\varepsilon|,d)}{\mathrm{vol}\,\mathsf{B}^d(0,1)} = 1 - \frac{d}{\sqrt{\pi}\cdot(\frac{d-1}{2})^2} \cdot \frac{(1-\varepsilon^2)^{\frac{d+1}{2}}}{\varepsilon} \cdot \Big(1 + O\Big(\frac{2}{d-1}\Big)\Big) \xrightarrow[d\to+\infty]{} 1.$$

Of course, symmetry considerations immediately show that the particular orientation of the two parallel hyperplanes $H_{-\varepsilon}$ and H_{ε} play *no* role the preceding estimates.

(1.11). Exercise. Let $V_{\mathrm{cyl}}(\varepsilon, d)$ denote the volume of the cylinder with cross-section $\mathbb{S}^{d-2}$, height 2ε, with its axis aligned alone the y_1-direction, and centered at $0 \in \mathbb{R}^d$. How does the ratio

$\frac{\operatorname{vol} S(|\varepsilon|,d)}{V_{\text{cyl}}(\varepsilon,d)}$ behave for large d.[7] [*Hint*: We note that $V_{\text{cyl}}(\varepsilon, d) = 2\varepsilon \cdot \frac{\pi^{\frac{d-1}{2}}}{\frac{d-1}{2}\cdot\Gamma(\frac{d-1}{2})}$, and observe that

$$\frac{\operatorname{vol} S(|\varepsilon|, d)}{V_{\text{cyl}}(\varepsilon, d)} = \frac{\frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\cdot\Gamma(\frac{d}{2})} - \frac{\pi^{\frac{d-1}{2}}}{(\frac{d-1}{2})^2\cdot\Gamma(\frac{d-1}{2})} \cdot (1-\varepsilon^2)^{\frac{d+1}{2}} \cdot \frac{1}{\varepsilon} \cdot \left(1 + O(\frac{2}{d-1})\right)}{2\varepsilon \cdot \frac{\pi^{\frac{d-1}{2}}}{\frac{d-1}{2}\cdot\Gamma(\frac{d-1}{2})}},$$

and it remains to consider the asymptotics as $d \to +\infty$. (Stirling's formula §B may be useful.)]

(1.12). Fix $\delta > 0$. If two closed unit balls $\mathsf{B}^d[z', 1]$ and $\mathsf{B}^d[z'', 1]$ intersect, where $z' = (\delta, 0, \ldots, 0)$ and $z'' = (-\delta, 0, \ldots, 0)$ in $\mathbb{R}^d$ are the centers of the two balls, then their common region is contained in the ball $\mathsf{B}^d\big[0, (1-\delta^2)^{\frac{1}{2}}\big]$. The volume of this ball goes to 0 as d becomes large, and consequently, the volume of the overlap between the two balls is vanishingly small in high dimensions irrespective of δ:

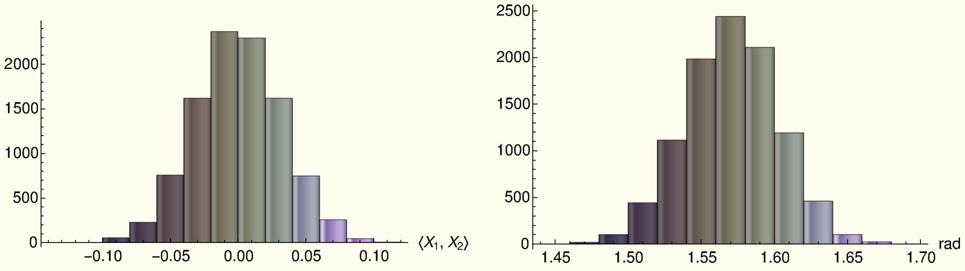
$$\operatorname{vol} \mathsf{B}^d\big[0, (1-\delta^2)^{\frac{1}{2}}\big] = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\cdot\Gamma(\frac{d}{2})} \cdot (1-\delta^2)^{\frac{d}{2}} \xrightarrow[d\to+\infty]{} 0.$$

This fact lies at the core of Shannon information theory; see [Zor16, Appendix C] for pointers.

(1.13) (Almost orthogonality of independently sampled unit vectors). A third interesting consequence of the preceding observations is that if two unit vectors X_1 and X_2 in high dimensional Euclidean space are sampled independently and uniformly randomly, then they turn out to be almost orthogonal with high probability.

$$\mathsf{P}\big(|\langle X_1, X_2\rangle| > \varepsilon\big) < \sqrt{\frac{\pi}{2}}\, \mathrm{e}^{-\varepsilon^2 \frac{d}{2}} \quad \text{for } \varepsilon > 0. \tag{1.14}$$

Since the random vectors X_1, X_2 are drawn independently and uniformly randomly from $\mathbb{S}^{d-1}$, the preceding estimate follows quite naturally from the discussion in §1.3. Here are histograms of, respectively, the inner products of, and the angles between, 10^4 pairs of random vectors drawn uniformly and independently from the unit sphere in dimension $d = 10^3$; we note that $\frac{\pi}{2}$rad ≈ 1.57rad:



(1.15). Example (Expected length of a uniform random vector in the unit ball)**.** Fix an integer $d > 2$ and consider the change of coordinates

$$]0, +\infty[\times]0, \pi[^{d-2} \times]0, 2\pi[\ni (r, \theta_1, \ldots, \theta_{d-2}, \theta_{d-1}) \mapsto (x_1, \ldots, x_d) \in \mathbb{R}^d \smallsetminus \{0\}$$

given by

$$\begin{aligned} x_1 &\coloneqq r\cos\theta_1, \\ x_2 &\coloneqq r\sin\theta_1\cos\theta_2, \\ &\vdots \end{aligned}$$

[7] See `https://mathworld.wolfram.com/Cylinder-SphereIntersection.html` and the references therein for several interesting details.

$$\begin{aligned} x_{d-1} &:= r \sin\theta_1 \sin\theta_2 \cdots \cos\theta_{d-1}, \\ x_d &:= r \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{d-1}, \end{aligned}$$

The Jacobian of this coordinate change is

$$\begin{pmatrix} \cos\theta_1 & -r\sin\theta_1 & \cdots & & 0 \\ \sin\theta_1 \cos\theta_2 & r\cos\theta_1\cos\theta_2 & \cdots & & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ \sin\theta_1 \cdots \sin\theta_{d-1} & r\cos\theta_1 \cdots \sin\theta_{d-1} & \cdots & & -r\sin\theta_1 \cdots \sin\theta_{d-2}\sin\theta_{d-1} \\ \sin\theta_1 \cdots \sin\theta_{d-1} & r\cos\theta_1 \cdots \sin\theta_{d-1} & \cdots & & r\sin\theta_1 \cdots \sin\theta_{d-2}\cos\theta_{d-1} \end{pmatrix},$$

and its determinant is given by $r^{d-1} \sin^{d-2}\theta_1 \, \sin^{d-3}\theta_2 \, \cdots \, \sin\theta_{d-2}$. Since this determinant does not vanish on the (open) domain of the coordinate change, it follows from the theory of integration in multiple dimensions that

$$\text{(1.16)} \qquad \mathrm{d}x_1 \cdots \mathrm{d}x_d = r^{d-1} \sin^{d-2}\theta_1 \, \sin^{d-3}\theta_2 \, \cdots \, \sin\theta_{d-2} \, \mathrm{d}r \, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{d-1}.$$

If a random vector X is picked uniformly randomly from the unit ball $\mathsf{B}^d[0,1]$, then its expected length is given by

$$\begin{aligned} \mathsf{E}[\|X\|] &= \int_{\mathbb{R}^d} \|x\| \cdot \frac{1}{\operatorname{vol} \mathsf{B}^d[0,1]} 1_{\mathsf{B}^d[0,1]}(x) \, \mathrm{d}x \\ &= \int_0^{2\pi} \int_{]0,\pi[^{d-2}} \int_0^1 \frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}} \cdot r^d \sin^{d-2}\theta_1 \, \sin^{d-3}\theta_2 \, \cdots \, \sin\theta_{d-2} \, \mathrm{d}r \, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{d-1} \\ &= \frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}} \cdot (d+1)} \int_0^{2\pi} \int_{]0,\pi[^{d-2}} \sin^{d-2}\theta_1 \, \sin^{d-3}\theta_2 \, \cdots \, \sin\theta_{d-2} \, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{d-1} \\ &= \frac{2\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}-1} \cdot (d+1)} \int_0^{\pi} \sin^{d-2}\theta_1 \, \mathrm{d}\theta_1 \int_0^{\pi} \sin^{d-3}\theta_2 \, \mathrm{d}\theta_2 \cdots \int_0^{\pi} \sin\theta_{d-2} \, \mathrm{d}\theta_{d-2}. \end{aligned}$$

Recalling that $\int_0^\pi \sin^k\theta \, \mathrm{d}\theta = \sqrt{\pi} \cdot \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2}+1)}$ for $k \geqslant 0$, we get

$$\begin{aligned} \mathsf{E}[\|X\|] &= \frac{2\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}-1} \cdot (d+1)} \cdot \pi^{\frac{d-2}{2}} \cdot \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d-2}{2}+1)} \cdot \frac{\Gamma(\frac{d-2}{2})}{\Gamma(\frac{d-3}{2}+1)} \cdots \frac{\Gamma(\frac{1+1}{2})}{\Gamma(\frac{1}{2}+1)} \\ &= \frac{d \cdot \Gamma(\frac{d}{2})}{d+1} \cdot \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} \cdot \frac{\Gamma(\frac{d-2}{2})}{\Gamma(\frac{d-1}{2})} \cdots \frac{\Gamma(1)}{\Gamma(\frac{1}{2}+1)} \\ &= \frac{d}{d+1}. \end{aligned}$$

See (1.17) for an alternate approach to calculating $\mathsf{E}[\|X\|]$.

(1.17). Exercise (Sampling uniformly from the unit ball)**.** Consider the task of sampling uniformly from the (closed) unit ball $\mathsf{B}^d[0,1]$. In low dimensions such as $d = 2, 3$, it is sometimes ‘roughly okay’ to sample uniformly from the unit cube $[-1,1]^d$ (which is easy since the components of such a random vector are independent) and accepting a sample if it lies in the unit ball and rejecting it otherwise. However, even for moderate dimensions d, the aforementioned strategy performs abysmally poorly for reasons that should be clear at this stage. How, then, does one sample uniformly from the unit ball? One approach could be to sample a random vector Y from a d-dimensional Gaussian and then defining $Z := \frac{Y}{\|Y\|}$ to get a uniformly distributed random vector Z on the unit sphere, followed by multiplying Z by a scalar random variable R (a priori, dependent on Z) taking values in $[0,1]$ that scales Z appropriately. What is the relevant distribution of R so that $R \cdot Z \overset{\text{dist}}{\sim} \operatorname{Uniform}\bigl(\mathsf{B}^d[0,1]\bigr)$?[8] [*Hint*: Pay close attention to the formula

[8] This exercise was suggested by Niranjan Balachandran.

(1.16); alternatively, recall that $\operatorname{vol} \mathsf{B}^d[0, r] = r^d \cdot \operatorname{vol} \mathsf{B}^d[0, 1]$, which means that the density of the volume at radius r is $dr^{d-1} \operatorname{vol} \mathsf{B}^d[0, 1]$.]

§1.3. Some consequences of concentration of areas on $\mathbb{S}^{d-1}$.

> The shape of the heaven is of necessity spherical.
>
> Aristotle.

As before, the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$ is $\mathbb{S}^{d-1} := \{y \in \mathbb{R}^d \mid \|y\| = 1\}$. We are interested in the $(d-1)$-area of *bands* on the unit sphere of the type

$$S_\varepsilon := \{(y_1, \ldots, y_d) \in \mathbb{S}^{d-1} \mid y_1 \in]-\varepsilon, \varepsilon[\}.$$

Of course, area S_ε is precisely equal to area $\mathbb{S}^{d-1}$ minus the $(d-1)$-area of the union of two *caps*

$$\{(y_1, \ldots, y_d) \in \mathbb{S}^{d-1} \mid y_1 \in [-1, -\varepsilon] \cup [\varepsilon, 1]\}.$$

These two caps are the regions (subsets) of the sphere that lie beyond (away from $0 \in \mathbb{R}^d$) the two hyperplanes $H_{-\varepsilon}$ and H_ε defined in (1.10).

(1.18). Exercise (Concentration of area around the equator)**.** Establish that

$$\operatorname{area}\left(\mathbb{S}^{d-1} \setminus S_\varepsilon\right) \approx \frac{2}{\sqrt{2\pi d \varepsilon^2}} \mathrm{e}^{-\varepsilon^2 \frac{d}{2}} \quad \text{as } d \to +\infty.$$

[*Hint*: Employ the techniques in (1.10), use Theorem (A.1) appropriately.]

It follows at once from Exercise (1.18) that an overwhelming proportion of the area of the unit sphere is concentrated on a thin band around *any* equator.

(1.19). Exercise. Establish the estimate (1.14) under the hypotheses in (1.13).

(1.20). Exercise. A student attempts to establish (1.14) under the conditions in (1.13). Accordingly, he picks a unit vector from $\mathbb{S}^{d-1}$ uniformly randomly and calls it X_1. Then he considers the subspace L of $\mathbb{R}^d$ orthogonal to the span of X_1, i.e., $L := (\operatorname{span} X_1)^\perp$, and looks at the intersection $L \cap \mathbb{S}^{d-1}$. Since $L \cap \mathbb{S}^{d-1}$ is an equator of $\mathbb{S}^{d-1}$, he argues that since X_2 is independent of X_1, the conditional probability given X_1 of X_2 being sampled from an ε-neighborhood of $L \cap \mathbb{S}^{d-1}$ is simply the probability of X_2 being sampled from the said neighborhood, and it is high for large d due to (1.18). Is his reasoning correct?

§2. Key inequalities

> One must always begin from Markov's inequality.
>
> Folklore.

« We study a few elementary but key inequalities to be employed in the later sections. The treatment here is brief and geared towards exposing a few of the *most elementary* inequalities. In particular, we do not study the multitude of *general techniques* for arriving at concentration inequalities; the reader is referred to the excellent textbooks [**Led01**, **BLM13**] for detailed treatments of such topics.

(2.1). Theorem (Markov's inequality)**.** *If X is a random variable taking values in $[0, +\infty[$ and its mathematical expectation exists, then*

$$\lambda \cdot \mathsf{P}(X \geqslant \lambda) \leqslant \mathsf{E}[X] \quad \textit{for all } \lambda \geqslant 0.$$

Proof. Fix $\lambda \geqslant 0$ and observe that since X takes values in the non-negative real numbers, $\lambda 1_{\{X \geqslant \lambda\}} \leqslant X 1_{\{X \geqslant \lambda\}} \leqslant X$. Taking expectations yields Markov's inequality. □

We need **Jensen's inequality** in the sequel. Recall that for $n \in \mathbb{N}^*$ a function $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ is **convex** if for all $t \in [0,1]$ and all $x, y \in \mathbb{R}^n$ we have $\phi\big((1-t)x + ty\big) \leqslant (1-t)\phi(x) + t\phi(y)$.

(2.2). Theorem. *Let $n \in \mathbb{N}^*$ and let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. If X is a random vector taking values in $\mathbb{R}^n$ such that $\mathsf{E}[X]$ and $\mathsf{E}\big[\phi(X)\big]$ exist,*[9] *then $\phi\big(\mathsf{E}[X]\big) \leqslant \mathsf{E}\big[\phi(X)\big]$.*

Proof. Due to convexity of ϕ, it admits a support function at each point of $\mathbb{R}^n$, i.e., at $x' \in \mathbb{R}^n$ there exists $\ell' \in \mathbb{R}^n$ such that

$$\phi(x) \geqslant \phi(x') + \langle \ell', x - x' \rangle \qquad \text{for all } x \in \mathbb{R}^n.$$

Since $\mathsf{E}[X]$ exists, we substitute $x' = \mathsf{E}[X]$ and an element ℓ' corresponding to $\mathsf{E}[X]$ above, and observe that for almost every outcome,

$$\phi(X) \geqslant \phi\big(\mathsf{E}[X]\big) + \langle \ell', X - \mathsf{E}[X] \rangle.$$

Taking expectations on both sides of the preceding inequality yields the assertion. □

For sums of bounded independent random variables we have an elegant inequality due to Hoeffding:

(2.3). Theorem (Hoeffding's inequality). *Let $n \in \mathbb{N}^*$ and suppose that $(\alpha_i)_{i=1}^n$ and $(\beta_i)_{i=1}^n$ are two sequences of real numbers satisfying $\alpha_i < \beta_i$ for each i. Let $(X_i)_{i=1}^n$ be a sequence of independent random variables with $(X_i - \mathsf{E}[X_i]) \in [\alpha_i, \beta_i]$ for each i. If $S_n := \sum_{i=1}^n X_i$, then*

$$\left.\begin{aligned} \mathsf{P}\big(S_n - \mathsf{E}[S_n] > t\big) &\leqslant \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n (\beta_i - \alpha_i)^2}} \\ \mathsf{P}\big(S_n - \mathsf{E}[S_n] < -t\big) &\leqslant \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n (\beta_i - \alpha_i)^2}} \end{aligned}\right\} \quad \textit{for all } t > 0.$$

(2.4). Remarks. Two remarks are in order here:

- Hoeffding's inequalities (2.3) feature the sum of a finite family of independent random variables and assert that the sum S_n is sharply concentrated around the mean $\mathsf{E}[S_n]$. It is clear from the inequalities that the smaller the bounds $(\beta_i - \alpha_i)$ of the random variables X_i, the sharper the concentration of S_n around $\mathsf{E}[S_n]$.
- The statement that 'the sum of a large family of independent random variables sharply concentrates around the mean of the sum' works as a weak *principle* — a statement that holds under a fair bit of generalization to families of dependent random variables (such as martingales, cf. Azuma's inequality [**BLM13**, Chapter 1], etc.).

The driving engine behind the proof of Hoeffding's inequalities (2.3), which we shall discuss momentarily, is Hoeffding's lemma:

(2.5). Lemma. *Let $\alpha, \beta \in \mathbb{R}$ satisfy $\alpha < 0 < \beta$. If X is a random variable with $\mathsf{E}[X] = 0$ and such that $X \in [\alpha, \beta]$, then for every $s \geqslant 0$ we have*

$$\mathsf{E}\big[\mathrm{e}^{sX}\big] \leqslant \mathrm{e}^{s^2 \cdot \frac{(\beta - \alpha)^2}{8}}.$$

Proof. The convexity of the exponential function plays a central role here. We fix $s \geqslant 0$ and express $x \in [\alpha, \beta]$ as

$$x = \big(1 - \eta(x)\big)\alpha + \eta(x)\beta \qquad \text{for } \eta(x) := \frac{x - \alpha}{\beta - \alpha},$$

[9] While it is possible to weaken these assumptions, we shall ignore such technicalities here.

and then appeal to convexity of $y \mapsto \mathrm{e}^{sy}$ to arrive at

$$\mathrm{e}^{sx} \leqslant \bigl(1 - \eta(x)\bigr)\mathrm{e}^{s\alpha} + \eta(x)\mathrm{e}^{s\beta}.$$

We observe that $\mathsf{E}[X] = 0$ leads to $\langle\eta\rangle := \mathsf{E}[\eta(X)] = \frac{-\alpha}{\beta-\alpha}$. (We note that $\langle\eta\rangle \in \,]0, 1[$ due to our assumption $\alpha < 0 < \beta$.) It follows that

$$\mathsf{E}\bigl[\mathrm{e}^{sX}\bigr] \leqslant (1 - \langle\eta\rangle)\mathrm{e}^{s\alpha} + \langle\eta\rangle\,\mathrm{e}^{s\beta} = \bigl(1 - \langle\eta\rangle + \langle\eta\rangle\,\mathrm{e}^{s(\beta-\alpha)}\bigr)\mathrm{e}^{-s\langle\eta\rangle(\beta-\alpha)},$$

and we would like to express the right-hand side of the equality as the exponential of some function of $s(\beta - \alpha)$ for the sake of a convenient calculus. With the aforementioned objective in mind, we define

$$\phi(t) := -\langle\eta\rangle\, t + \ln\bigl(1 - \langle\eta\rangle + \langle\eta\rangle\,\mathrm{e}^{t}\bigr) \quad \text{for } t \geqslant 0, \tag{2.6}$$

and write

$$\mathsf{E}\bigl[\mathrm{e}^{sX}\bigr] \leqslant \mathrm{e}^{\phi(s(\beta-\alpha))}. \tag{2.7}$$

A quick analysis of the function ϕ defined in (2.6) is needed here. We start by noting that ϕ is smooth, and its derivative $\frac{\mathrm{d}\phi}{\mathrm{d}t}$ vanishes at t_* satisfying

$$-\langle\eta\rangle + \frac{1}{1 - \langle\eta\rangle + \langle\eta\rangle\,\mathrm{e}^{t_*}} \cdot \langle\eta\rangle\,\mathrm{e}^{t_*} = 0.$$

This equation yields the unique solution $t_* = 0$ (in the light of $\langle\eta\rangle \in \,]0, 1[$) with the understanding that only the right-hand derivative is involved at 0. The double derivative $\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2}(t)$ for any $t > 0$ admits the bound

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2}(t) = \frac{(1 - \langle\eta\rangle)\cdot\langle\eta\rangle\,\mathrm{e}^{t}}{\bigl((1 - \langle\eta\rangle) + \langle\eta\rangle\,\mathrm{e}^{t}\bigr)^2} \leqslant \frac{1}{4} \qquad \text{since } 4ab \leqslant (a+b)^2 \text{ for } a, b > 0. \tag{2.8}$$

Since Taylor's theorem now shows that there exists $t' \in [0, t]$ such that

$$\phi(t) = \phi(0) + t \cdot \frac{\mathrm{d}\phi}{\mathrm{d}t}(0) + \frac{t^2}{2} \cdot \frac{\mathrm{d}^2\phi}{\mathrm{d}t^2}(t') = \frac{t^2}{2} \cdot \frac{\mathrm{d}^2\phi}{\mathrm{d}t^2}(t'),$$

in the light of the inequality in (2.8) we arrive at

$$\phi(t) \leqslant \frac{t^2}{8}.$$

Reverting back to the variable $s \geqslant 0$ and substituting in (2.7) leads to

$$\mathsf{E}\bigl[\mathrm{e}^{sX}\bigr] \leqslant \mathrm{e}^{s^2 \cdot \frac{(\beta-\alpha)^2}{8}},$$

which completes the proof of the lemma. □

Proof of Theorem (2.3). The so-called *Chernoff bounding method* applies to the random variable S_n: we observe that for $t > 0$ and $s \geqslant 0$ we have

$$S_n - \mathsf{E}[S_n] \geqslant t \quad \Leftrightarrow \quad \mathrm{e}^{s(S_n - \mathsf{E}[S_n])} \geqslant \mathrm{e}^{st},$$

which leads to

$$\begin{aligned} \mathsf{P}\bigl(S_n - \mathsf{E}[S_n] \geqslant t\bigr) &= \mathsf{P}\bigl(\mathrm{e}^{s(S_n - \mathsf{E}[S_n])} \geqslant \mathrm{e}^{st}\bigr) \\ &\leqslant \mathrm{e}^{-st} \cdot \mathsf{E}\bigl[\mathrm{e}^{s(S_n - \mathsf{E}[S_n])}\bigr] && \text{by Markov's inequality (2.1)} \\ &= \mathrm{e}^{-st} \prod_{i=1}^{n} \mathsf{E}\bigl[\mathrm{e}^{s(X_i - \mathsf{E}[X_i])}\bigr] && \text{by independence of } (X_i)_{i=1}^{n}. \end{aligned} \tag{2.9}$$

Of course, since $X_i - \mathsf{E}[X_i] \in [\alpha_i, \beta_i]$ and $X_i - \mathsf{E}[X_i]$ is a mean zero random variable, Hoeffding's Lemma (2.5) applies to X_i, and

$$\mathsf{E}\bigl[\mathrm{e}^{s(X_i - \mathsf{E}[X_i])}\bigr] \leqslant \mathrm{e}^{s^2 \cdot \frac{(\beta_i - \alpha_i)^2}{8}}.$$

Substituting in the preceding expression we arrive at

$$\mathsf{P}\big(S_n - \mathsf{E}[S_n] \geqslant t\big) \leqslant \mathrm{e}^{-st}\mathrm{e}^{\frac{s^2}{8}\sum_{i=1}^n(\beta_i-\alpha_i)^2}.$$

Since $s > 0$ is arbitrary, we may minimize the right-hand side over s. This manoeuvre yields the unique optimizer $s_* = \frac{4t}{\sum_{i=1}^n(\beta_i-\alpha_i)^2}$, at which the optimal right-hand side is

$$\mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n(\beta_i-\alpha_i)^2}}$$

as asserted in the first inequality. The proof of the second inequality is similar. □

Remark. The Chernoff bounding technique (2.9) is a central theme in the art of probabilistic inequalities: on the left-hand side is the quantity whose estimate we seek, and the right-hand side features a parameter-dependent family of quantities, each of which provides such an estimate. Typically, all estimates are not created equal (there is little need for democracy in the world of inequalities), and the parameters that offer the tightest inequality are the ones we keep.

(2.10). Example. We provide a simple and important example illustrating the Chernoff bounding technique. Consider a standard normal random variable $X \overset{\text{dist}}{\sim} \mathfrak{N}(0,1)$. We shall establish that

$$\mathsf{P}(X > x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t \leqslant \mathrm{e}^{-\frac{x^2}{2}} \quad \text{for } x \geqslant 0.$$

Indeed, if $x > 0$, then $\mathsf{P}(X > x) = \int_x^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$ and

$$X > x \quad\Leftrightarrow\quad sX > sx \text{ for } s > 0 \quad\Leftrightarrow\quad \mathrm{e}^{sX} > \mathrm{e}^{sx} \text{ for } s > 0,$$

and Markov's inequality (2.1) shows that

$$\mathsf{P}(X > x) = \mathsf{P}\big(\mathrm{e}^{sX} > \mathrm{e}^{sx}\big) \leqslant \mathrm{e}^{-sx}\cdot\mathsf{E}\big[\mathrm{e}^{sX}\big] \quad \text{for } s > 0.$$

Of course, for all $s \in \mathbb{R}$ we have

$$\mathsf{E}\big[\mathrm{e}^{sX}\big] = \int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{st-\frac{t^2}{2}}\,\mathrm{d}t = \mathrm{e}^{\frac{s^2}{2}}\cdot\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-s)^2}{2}}\,\mathrm{d}t = \mathrm{e}^{\frac{s^2}{2}},$$

and substituting back into the preceding inequality leads to

$$\mathsf{P}(X > x) \leqslant \mathrm{e}^{-sx+\frac{s^2}{2}} \quad \text{for } s > 0.$$

Minimizing the right-hand side over $s \in \left]0, +\infty\right[$ leads to

$$\mathsf{P}(X > x) \leqslant \mathrm{e}^{-\frac{x^2}{2}} \quad \text{for } x \geqslant 0. \tag{2.11}$$

The bound (2.11) is, however, not too sharp; in fact, sharper bounds are provided by Mitrinovic's inequalities:[10]

$$\sqrt{\frac{2}{\pi}}\cdot\frac{\mathrm{e}^{-\frac{x^2}{2}}}{x+\sqrt{x^2+4}} < \mathsf{P}(X > x) \leqslant \sqrt{\frac{2}{\pi}}\cdot\frac{\mathrm{e}^{-\frac{x^2}{2}}}{x+\sqrt{x^2+\frac{8}{\pi}}} \quad \text{for all } x \geqslant 0.$$

The concentration inequalities we shall study below will conform to the type (2.11) featuring square exponential decay.

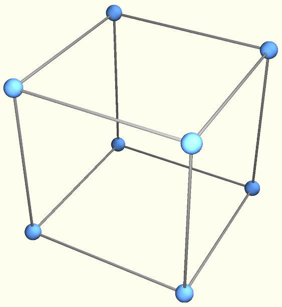
[10] See [AS64, p. 298] for more information.

§3. Concentration of the uniform distribution on the discrete cube

« The discrete d-dimensional cube (also known as the *Boolean cube*) is the set

$$\mathsf{b}^d := \{-1, 1\}^d.$$

By definition, b^d is the family of all functions from a d-element set into the set $\{-1, 1\}$, or equivalently, it is the set of all sequences $y = (y_n)_{n=1}^d$ of length d such that each $y_n \in \{-1, 1\}$. Of course, there are 2^d elements in b^d, i.e., $\left|\mathsf{b}^d\right| = 2^d$. A pictorial representation of the discrete cube in dimension $d = 3$ is the following (the spheres represent the triplets $(\pm 1, \pm 1, \pm 1)$, the skeleton is irrelevant albeit convenient for visualization):



The Boolean cube b^d is one of the simplest spaces on which one can observe the concentration phenomenon; for the purposes of our discussion, it will be equipped with the following two structures:

- The *Hamming distance* between two elements in b^d is defined by the number of indices at which the two elements differ:

(3.1)
$$\rho_0(y, y') := \left|\{n = 1, \ldots, d \mid y_n \neq y'_n\}\right| \quad \text{for } y, y' \in \mathsf{b}^d.$$

- The *uniform distribution* on b^d: $Y \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{b}^d)$ if

$$\mathsf{P}(Y \in A) = 2^{-d} \cdot |A| \quad \text{for all } A \subset \mathsf{b}^d.$$

  Consequently,

  ▻ if $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{b}^d)$, then symmetry considerations immediately imply that

$$\mathsf{E}[X] = 2^{-d} \cdot \sum_{y \in \mathsf{b}^d} y = 0 \in \mathbb{R}^d;$$

  ▻ more generally, if $f : \mathsf{b}^d \longrightarrow \mathbb{R}$ is any function, then

$$\mathsf{E}\big[f(X)\big] = 2^{-d} \cdot \sum_{y \in \mathsf{b}^d} f(y) \in \mathbb{R}.$$

A magnificent treatment of concentration phenomena on product spaces may be found in the seminal article [Tal95]; we employ the simplest of ideas and techniques therein for our purposes here.

(3.2). Exercise. Demonstrate that the Hamming distance satisfies the properties of a metric. What do unit balls in the Hamming distance look like? What do Hamming balls of radius $r > 0$ look like, and how many elements of b^d are present in the Hamming ball of radius r centered at $0 \in \mathbb{R}^d$?

(3.3). Exercise. Recall that the ℓ_1-, ℓ_2-, and ℓ_∞-distances between two vectors $x, y \in \mathbb{R}^d$ are

$$\rho_1(x, y) = \sum_{n=1}^{d} \left|x_n - y_n\right|, \;\; \rho_2(x, y) = \left(\sum_{n=1}^{d} \left|x_n - y_n\right|^2\right)^{\frac{1}{2}}, \;\; \rho_\infty(x, y) = \max_n \left|x_n - y_n\right|.$$

If we regard b^d as a subset of $\mathbb{R}^d$ in a natural way, then are ρ_0, ρ_1, ρ_2, and ρ_∞ related to each other on b^d? Justify.

(3.4). Exercise. Let $d \in \mathbb{N}^*$ be at least 2, and suppose that $(X_1, \ldots, X_d) \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{b}^d)$.

- Find the (marginal) distributions of X_1 and X_d.
- Justify whether the random variables $X_1, \ldots, X_d$ are mutually independent.

Distances between points on b^d will be measured in terms of the Hamming distance; naturally, the Hamming distance on b^d is bounded above by d. The Hamming distance between a point $y \in \mathsf{b}^d$ and a set $A \subset \mathsf{b}^d$ is the standard one:

$$\rho_0(y, A) := \min_{y' \in A} \rho_0(y, y'); \tag{3.5}$$

it is the minimum number of sign flips of the components of y that are needed to bring y to the set A is given by $\rho_0(y, A)$.

We shall prove the following theorem in a short while:

(3.6). Theorem. *Let $d \in \mathbb{N}^*$ and fix a non-empty set $A \subset \mathsf{b}^d$. Let $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{b}^d)$, and define the random variable*

$$X_A := \rho_0(X, A)$$

describing the Hamming distance of X from A. Then for $s > 0$ we have

$$\mathsf{E}\left[\mathrm{e}^{sX_A}\right] \cdot \mathsf{P}(X \in A) \leqslant \left(\cosh\left(\frac{s}{2}\right)\right)^{2d}.$$

Theorem (3.6) is a special case of **Talagrand's inequality**.[11] Here is an immediate consequence of Theorem (3.6):

(3.7). Corollary. *Let $d \in \mathbb{N}^*$ and let $A \subset \mathsf{b}^d$ be a non-empty set. If $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{b}^d)$ and $X_A := \rho_0(X, A)$ is the Hamming distance of X from A, then for $s > 0$ we have*

$$\mathsf{E}\left[\mathrm{e}^{sX_A}\right] \cdot \mathsf{P}(X \in A) \leqslant \mathrm{e}^{s^2 \frac{d}{4}}.$$

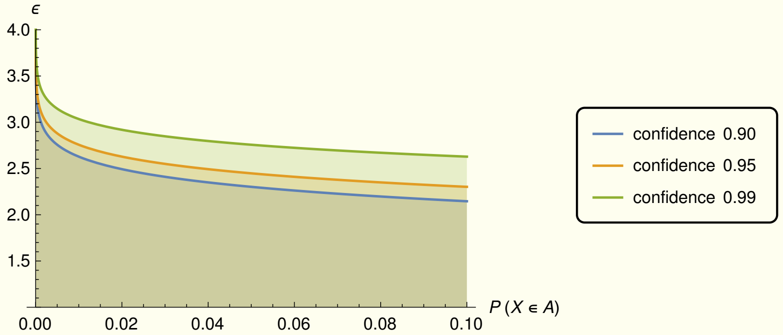
In particular, for $\varepsilon > 0$ we have

$$\mathsf{P}\left(X_A > \varepsilon\sqrt{d}\right) \cdot \mathsf{P}(X \in A) \leqslant \mathrm{e}^{-\varepsilon^2}.$$

(3.8). Remark. Let us demonstrate a few immediate implications of these inequalities. Consider any non-empty set $A \subset \mathsf{b}^d$ such that $\mathsf{P}(X \in A) =$ constant (does not change with d). Then the inequality $\mathsf{P}\left(\rho_0(X, A) > \varepsilon\sqrt{d}\right) \leqslant \frac{\mathrm{e}^{-\varepsilon^2}}{\mathsf{P}(X \in A)}$ permits us to provide probabilistic guarantees of finding samples of X beyond ρ_0-distance $\varepsilon\sqrt{d}$ away from A: it says that with probability at least $1 - \frac{\mathrm{e}^{-\varepsilon^2}}{\mathsf{P}(X \in A)}$ a vector X sampled uniformly randomly from b^d will lie within $\varepsilon\sqrt{d}$ distance (measured in terms of ρ_0) of A. Here are pictorial representations of the variation of ε with $\mathsf{P}(X \in A)$ for confidence levels 0.99, 0.95, 0.90; only the lower 10% range of $\mathsf{P}(X \in A)$ has been

[11] See https://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/ for a brief treatment of this inequality.

depicted for the sake of clarity.



Fix $d = 10^3$. If $A \subset \mathsf{b}^d$ is picked such that A contains 1% of b^d, then $\mathsf{P}(X \in A) = 10^{-2}$. Any uniformly sampled random vector from b^d is, with probability at least 0.99, at most $3.03 \times \sqrt{10^3} \approx 96$ in ρ_0-distance away from A. That is, with probability at least 0.99, at most 96 sign flips are needed to bring a uniformly sampled random vector to a set $A \subset \mathsf{b}^d$ containing 1% of b^d. If A contains 5% of the elements of b^{10^3}, then with probability at least 0.99 the distance from A of a sample uniformly randomly extracted from b^d is within $2.76 \times \sqrt{10^3} \approx 87$. The point here is that the numbers 96 and 87 are *small* compared to the number of dimensions $d = 10^3$ and they are *incomparably small* compared to the total number of elements $\left|\mathsf{b}^{10^3}\right| = 2^{10^3}$ in b^{10^3}.[12] Consider a somewhat more realistic example of $d = 40$. There are more than 10^{12} points in b^{40}, 1% of this figure is a little over ten billion 10^{10}, and a reasonably large supercomputer can store these many floating points today. Talagrand's inequality asserts that with probability at least 0.99, a uniformly randomly sampled point in b^{40} is within $3.03 \times \sqrt{40} \approx 19$ sign flips away from any fixed $A \subset \mathsf{b}^{40}$ that contains around 10^{10} points of b^{40}. Observe that 19 accounts for about half as many coordinates as in the elements of b^{40}.

(3.9). Preparatory to the proofs of Theorem (3.6) and Corollary (3.7), we collect several elementary definitions and observations in this paragraph. Introduce the notation

$$a \wedge b := \min\{a, b\} \quad \text{for } a, b \in \mathbb{R}.$$

For a vector $v \in \mathsf{b}^d$,

$$\begin{cases} v_d & \text{is the } d\text{-th component of } v\text{, of course, and we define} \\ v_{\widehat{d}} & \text{to be the vector } (v_1, \ldots, v_{d-1}) \text{ of the first } (d-1) \text{ components of } v. \end{cases}$$

Of course, $v_d \in \mathsf{b}^1$ and $v_{\widehat{d}} \in \mathsf{b}^{d-1}$. If $A \subset \mathsf{b}^d$ is a non-empty set, then we define its positive and negative *slices*

$$A_+ := \left\{v \in A \,\middle|\, v_d = 1\right\} \quad \text{and} \quad A_- := \left\{v \in A \,\middle|\, v_d = -1\right\}; \tag{3.10}$$

naturally,

$$A = A_+ \sqcup A_-.$$

In particular, if $A = \mathsf{b}^d$, then b^d_+ and b^d_- are the *slices* of b^d with the last component equal to 1 and -1, respectively, and $\mathsf{b}^d = \mathsf{b}^d_+ \sqcup \mathsf{b}^d_-$; moreover, each of b^d_+ and b^d_- is identifiable in a natural way with a copy of b^{d-1} and we write $\mathsf{b}^d_\pm \equiv \mathsf{b}^{d-1}$. The metric property of the Hamming distance (3.5) shows that

$$\rho_0(v, A) = \rho_0(v, A_+) \wedge \rho_0(v, A_-) \quad \text{for any } v \in \mathsf{b}^d. \tag{3.11}$$

Fix a non-empty set $A \subset \mathsf{b}^d$.

[12] We shall *not* be addressing the issue of realizing the set b^{10^3} on a physical device; experiments are inexpensive in mathematics.

- ○ If $v \in \mathsf{b}^d_+$, then v is of the form $(v_{\widehat{d}}, 1)$ and (3.11) shows that

$$\rho_0(v, A) = \rho_0\big((v_{\widehat{d}}, 1), A_+\big) \wedge \rho_0\big((v_{\widehat{d}}, 1), A_-\big);$$

identifying b^d_+ with b^{d-1} and regarding A_+ as a subset of $\mathsf{b}^d_+ \equiv \mathsf{b}^{d-1}$, the definition of the Hamming distance (3.1) leads to

$$\text{if } v \in \mathsf{b}^d_+, \text{ then} \rho_0(v, A) = \rho_0(v_{\widehat{d}}, A_+) \wedge \big(1 + \rho_0(v_{\widehat{d}}, A_-)\big). \tag{3.12}$$

- ○ Similarly, if $v \in \mathsf{b}^d_-$, then v is of the form $(v_{\widehat{d}}, -1)$, and (3.11) it shows that

$$\rho_0(v, A) = \rho_0\big((v_{\widehat{d}}, -1), A_-\big) \wedge \rho_0\big((v_{\widehat{d}}, -1), A_+\big);$$

identifying b^d_- with b^{d-1} and regarding A_- as a subset of $\mathsf{b}^d_- \equiv \mathsf{b}^{d-1}$, the definition of the Hamming distance (3.1) yields

$$\text{if } v \in \mathsf{b}^d_-, \text{ then} \rho_0(v, A) = \rho_0(v_{\widehat{d}}, A_-) \wedge \big(1 + \rho_0(v_{\widehat{d}}, A_+)\big). \tag{3.13}$$

We need two auxiliary technical lemmas:

(3.14). Lemma. *The function* $\mathbb{R} \times \mathbb{R} \ni (x, y) \longmapsto x \wedge y \in \mathbb{R}$ *is Lipschitz continuous and concave.*

Proof. The assertions follow at once from the facts that $\mathbb{R} \ni z \longmapsto |z| \in [0, +\infty[$ is convex and Lipschitz continuous, $\mathbb{R}^2 \ni (x, y) \longmapsto x - y \in \mathbb{R}$ and $\mathbb{R}^2 \ni (x, y) \longmapsto x + y \in \mathbb{R}$ are linear and therefore convex, the identity

$$x \wedge y = \frac{1}{2}(x + y) - \frac{1}{2}\left|x - y\right|$$

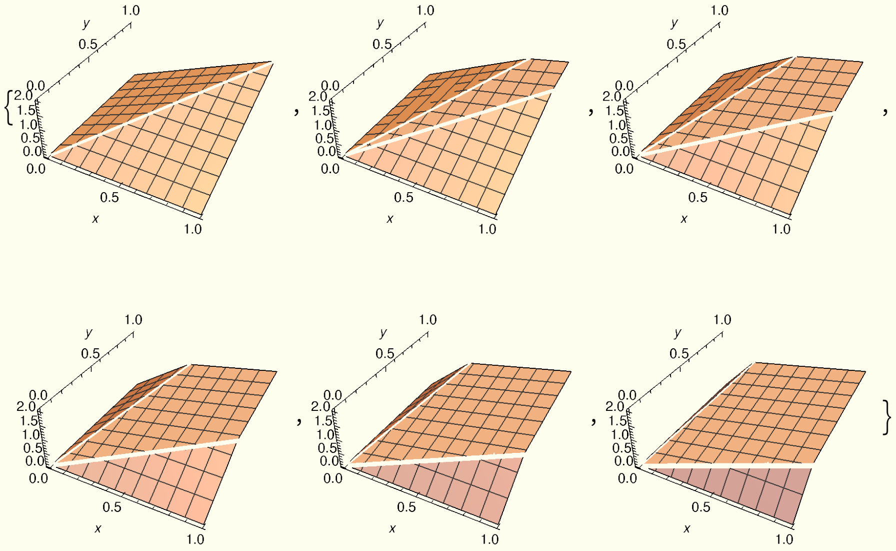
holds for all $x, y \in \mathbb{R}$, and that the sum of two concave functions is concave. □

(3.15). Lemma. *Let* $\alpha \geqslant 1$ *and consider the equality-constrained optimization problem*

$$\begin{aligned} &\underset{x,y}{\text{maximize}} && \big(x \wedge (\alpha y)\big) + \big(y \wedge (\alpha x)\big) \\ &\text{subject to} && \begin{cases} x, y \geqslant 0, \\ x^{-1} + y^{-1} = 2. \end{cases} \end{aligned} \tag{3.16}$$

(3.16) *admits (at most) two maximizers* $\left(\frac{1+\alpha}{2}, \frac{1+\alpha}{2\alpha}\right)$ *and* $\left(\frac{1+\alpha}{2\alpha}, \frac{1+\alpha}{2}\right)$, *and the maximum value is* $\frac{(1+\alpha)^2}{2\alpha}$.

Here are six figures describing the way that the objection function, restricted to the square $[0, 1]^2 \subset \mathbb{R}^2$, in (3.16) changes with increasing values of α, starting with $\alpha = 1$ (which corresponds to the function $(x, y) \mapsto 2(x \wedge y))$.



Proof of Lemma (3.15). It is not difficult to see that a solution of (3.24) exists for every fixed $\alpha \geqslant 1$. Indeed, fix $\alpha \geqslant 1$ and observe that if x is large, then the equality constraint forces y to decrease. Beyond a point the variable x ceases to matter in the objective function (due to the two minimums) and only the y terms matter; moreover, as x increases beyond a certain threshold, since the equality constraint forces y to decrease, the objective function monotonically decreases. A maximizer, therefore, cannot involve large values of x. Symmetry arguments show that an identical conclusion holds for the variable y. Thus, a large enough compact square of the form $[0, a]^2$ (with a depending on α) must contain the maximizer. In view of continuity of the objective function, an appeal to Weierstrass's theorem suffices to conclude the existence of a maximizer in the aforementioned square. With the existence question settled, we turn to finding the maximizer(s). It is also clear that the equality constraint forces the inequality constraints $x \geqslant 0, y \geqslant 0$ to be *inactive* at any maximizer (x_*, y_*);[13] consequently, we shall ignore the inequality constraints in the remainder of the proof. The standard non-smooth *Lagrange multiplier rule* Theorem (D.2) applies to the problem (3.24) because the functions

$$\mathbb{R}^2 \ni (x, y) \mapsto f(x, y) := (x \wedge \alpha y) + (y \wedge \alpha x) \in \mathbb{R}$$

and

$$]0, +\infty[^2 \ni (x, y) \mapsto g(x, y) := x^{-1} + y^{-1} - 2 \in \mathbb{R}$$

describing the objective and the equality constraints, respectively, are both locally Lipschitz continuous. Accordingly, if (x_*, y_*) is a local maximizer solving (3.24), then there exist $\eta \in \{0, 1\}$ and $\lambda \in \mathbb{R}$ such that

- the *nontriviality condition*

(3.17) $$(\eta, \lambda) \neq (0, 0)$$

holds, and

- the *stationarity condition*

(3.18) $$0 \in \partial_C\big(\eta \cdot f + \langle \lambda, g \rangle\big)(x_*, y_*) \subset \mathbb{R}^2$$

holds, where ∂_C denotes the (Clarke) generalized gradient.

[13] This statement means that $(x_*, y_*) \in \,]0, +\infty[^2$.

Theorem (D.1) permits us to employ the standard gradient formula to construct the set on the right-hand side of (3.18), and performing the necessary computations we arrive at the following observations:

- If $\eta = 0$, then the objective function f becomes irrelevant and since g is smooth, the stationarity condition takes the form $0 \in \left\{\lambda \cdot \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix}\right\}$, and since the set on the right-hand side is a singleton, we must have $\lambda = 0$, contradicting the nontriviality condition (3.17). Thus, we may take $\eta = 1$.
- For $\eta = 1$, we note that for every $\alpha \geqslant 1$ the region $]0, +\infty[^2$ is the union of the following three subsets:

$$S_1 := \left\{(x, y) \in]0, +\infty[^2 \;\middle|\; y \leqslant \tfrac{1}{\alpha} x\right\},$$
$$S_2 := \left\{(x, y) \in]0, +\infty[^2 \;\middle|\; \tfrac{1}{\alpha} x \leqslant y \leqslant \alpha x\right\},$$
$$S_3 := \left\{(x, y) \in]0, +\infty[^2 \;\middle|\; y \geqslant \alpha x\right\}.$$

  The boundaries of the sets S_1, S_2, S_3 are of measure 0. Assuming that $\alpha > 1$, we deduce the following:

  ▻ On S_1 the objective function is $f(x, y) = (1 + \alpha)y$ and f is differentiable on $\mathrm{int}(S_1)$; consequently, $\nabla f(x, y) = \begin{pmatrix} 0 \\ 1+\alpha \end{pmatrix}$ for $(x, y) \in \mathrm{int}(S_1)$. If $(x_*, y_*) \in \mathrm{int}(S_1)$, then

$$0 = \begin{pmatrix} 0 \\ 1+\alpha \end{pmatrix} + \lambda \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix} \quad \text{for some } \lambda \in \mathbb{R},$$

  which is impossible because $x_* > 0$ and $\alpha > 1$. We conclude that $(x_*, y_*) \notin \mathrm{int}(S_1)$.

  ▻ On S_2 the objective function is $f(x, y) = x + y$ and f is differentiable on $\mathrm{int}(S_2)$; consequently, $\nabla f(x, y) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ for $(x, y) \in \mathrm{int}(S_2)$. If $(x_*, y_*) \in \mathrm{int}(S_2)$, then the equation

$$0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix}$$

  is solvable for some $\lambda \in \mathbb{R}$, and it leads to a stationary point (candidate maximizer)

(3.19) $$(x_*, y_*) = (1, 1) \text{ with the corresponding cost equal to } 2$$

  in view of the equality constraint $x_*^{-1} + y_*^{-1} = 2$.

  ▻ On S_3 the objective function is $f(x, y) = (1 + \alpha)x$ and f is differentiable on $\mathrm{int}(S_3)$; consequently, $\nabla f(x, y) = \begin{pmatrix} 1+\alpha \\ 0 \end{pmatrix}$ for $(x, y) \in \mathrm{int}(S_3)$. If $(x_*, y_*) \in \mathrm{int}(S_3)$, then

$$0 = \begin{pmatrix} 1+\alpha \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix} \quad \text{for some } \lambda \in \mathbb{R},$$

  which is impossible because $y_* > 0$ and $\alpha > 1$. We conclude that $(x_*, y_*) \notin \mathrm{int}(S_3)$.

  ▻ On $S_1 \cap S_2$ the stationarity condition yields

$$0 \in \mathrm{co}\left\{\begin{pmatrix} 0 \\ 1+\alpha \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\} + \lambda \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix}.$$

  If for some $c \in [0, 1]$ we have

$$0 = (1 - c)\begin{pmatrix} 0 \\ 1+\alpha \end{pmatrix} + c\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix},$$

  then in the light of the constraint $x_*^{-1} + y_*^{-1} = 2$ and the fact that $y_* = \alpha^{-1} x_*$ on $S_1 \cap S_2$, after the necessary algebraic manipulations we arrive at the solution $c = \alpha^{-1}$ (satisfying the condition $c \in [0, 1]$) of this system of equations. This leads to the point $(x_*, y_*) = \left(\frac{1+\alpha}{2}, \frac{1+\alpha}{2\alpha}\right)$ as the solution satisfying $x_*^{-1} + y_*^{-1} = 2$ and $y_* = \alpha^{-1} x_*$. The corresponding value of the objective function is $\frac{(1+\alpha)^2}{2\alpha}$, which (since $(1-\alpha)^2 \geqslant 0$) is at least 2 — the value of the objective function at the point $(1,1)$ obtained in (3.19). We conclude that $(x_*, y_*) = \left(\frac{1+\alpha}{2}, \frac{1+\alpha}{2\alpha}\right)$ is a maximizer.

▻ On $S_2 \cap S_3$ the necessary condition yields

$$0 \in \operatorname{co}\left\{ \begin{pmatrix} 1+\alpha \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} + \lambda \begin{pmatrix} -x_*^{-2} \\ -y_*^{-2} \end{pmatrix}.$$

The analysis proceeds in a similar fashion as the preceding case, leading to the maximizer $(x_*, y_*) = \left(\frac{1+\alpha}{2\alpha}, \frac{1+\alpha}{2}\right)$ with the same value $\frac{(1+\alpha)^2}{2\alpha}$ of the objective function as in the preceding case.

An examination of the remaining case of $\alpha = 1$ (and $\eta = 1$) and its proof is relegated to Exercise (3.20). We merely note that for $\alpha = 1$ the set S_2 is the diagonal of the open first quadrant and $f(x, y) = 2(x \wedge y)$.

Since the preceding cases are exhaustive, our proof is complete. □

(3.20). Exercise. Complete the proof of the case $\eta = 1$ and $\alpha = 1$ in the Proof of Lemma (3.15).

We are ready for:

Proof of Theorem (3.6). The "rectangular" nature of the set b^d permits the employment of mathematical induction. We begin with the *induction base* for $d = 1$. Since $\mathsf{b}^1 = \{-1, 1\}$ and $A \subset \mathsf{b}^1$ is non-empty by hypothesis, three mutually exclusive cases arise, namely, $A = \{-1\}$, $A = \{1\}$, and $A = \{-1, 1\}$. Fix $s \geqslant 0$. If $A = \{-1\}$, then $\mathsf{E}\left[e^{s\rho_0(X,A)}\right] = e^{s\cdot 0} \cdot \mathsf{P}(X = -1) + e^{s\cdot 1} \cdot \mathsf{P}(X = 1) = \frac{1}{2}(1 + e^s)$, and similarly if $A = \{1\}$, then $\mathsf{E}\left[e^{s\rho_0(X,A)}\right] = \frac{1}{2}(1 + e^s)$; if $A = \{-1, 1\}$, then $\rho_0(X, A) = 0$, which means that $\mathsf{E}\left[e^{s\rho_0(X,A)}\right] = 1$. Since $\cosh(\frac{s}{2}) \geqslant 1$ for all $s \geqslant 0$, the purported inequality is immediately verified for $A = \{-1, 1\}$ for which $\mathsf{P}(A) = 1$. For either of the cases $A = \{-1\}$ and $A = \{1\}$, it suffices to observe that $\mathsf{P}(A) = \frac{1}{2}$ and $2\cosh(\frac{s}{2})^2 = 2\left(\frac{1}{2} + \frac{1}{4}(e^s + e^{-s})\right) \geqslant \frac{1}{2}(1 + e^s)$ for all $s \geqslant 0$. This completes the induction base.

Assuming that $d \geqslant 2$ and that the purported inequality holds for $d - 1$ as our *induction hypothesis*, we move to the *induction step*. Observe that for $s \geqslant 0$,

$$\begin{aligned} \mathsf{E}\left[e^{sX_A}\right] &= \mathsf{E}\left[e^{sX_A}\left(1_{\{1\}}(X_d) + 1_{\{-1\}}(X_d)\right)\right] \\ &= \mathsf{E}\left[e^{sX_A} \,\middle|\, X_d = 1\right]\mathsf{P}(X_d = 1) + \mathsf{E}\left[e^{sX_A} \,\middle|\, X_d = -1\right]\mathsf{P}(X_d = -1). \end{aligned} \tag{3.21}$$

Since $X_A = \rho_0(X, A)$, $X = (X_{\widehat{d}}, X_d)$, and $A = A_+ \sqcup A_-$ as defined in (3.10), conditional on $X_d = 1$ we see from (3.12) that

$$\begin{aligned} \mathsf{E}\left[e^{sX_A} \,\middle|\, X_d = 1\right] &= \mathsf{E}\left[e^{s\rho_0(X,A)} \,\middle|\, X_d = 1\right] \\ &= \mathsf{E}\left[\exp\Bigl(s\bigl(\rho_0(X_{\widehat{d}}, A_+) \wedge (\rho_0(X_{\widehat{d}}, A_-) + 1)\bigr)\Bigr) \,\middle|\, X_d = 1\right]. \end{aligned}$$

Since $X_{\widehat{d}}$ is independent of X_d and $\exp(\cdot)$ is monotone increasing, the right-hand side is equal to

$$\mathsf{E}\left[\exp\bigl(s\rho_0(X_{\widehat{d}}, A_+)\bigr) \wedge \exp\Bigl(s\bigl(\rho_0(X_{\widehat{d}}, A_-) + 1\bigr)\Bigr)\right],$$

and in view of Lemma (3.14) and Jensen's inequality (2.2), the last expression is bounded above by

$$\mathsf{E}\left[\exp\bigl(s\rho_0(X_{\widehat{d}}, A_+)\bigr)\right] \wedge \mathsf{E}\left[\exp\Bigl(s\bigl(\rho_0(X_{\widehat{d}}, A_-) + 1\bigr)\Bigr)\right]. \tag{3.22}$$

By our induction hypothesis (3.22) is bounded above by

$$\frac{1}{\mathsf{P}(X_{\widehat{d}} \in A_+)} \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^{2(d-1)} \wedge \frac{1}{\mathsf{P}(X_{\widehat{d}} \in A_-)} \cdot e^s \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^{2(d-1)}$$

$$= \frac{1}{\mathsf{P}(X \in A)} \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^{2(d-1)} \cdot \left(\frac{\mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_+)} \wedge \frac{\mathrm{e}^s \cdot \mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_-)} \right).$$

Similar arguments show that $\mathsf{E}\big[\mathrm{e}^{sX_A} \bigm| X_d = -1\big]$ is bounded above by

$$\frac{1}{\mathsf{P}(X \in A)} \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^{2(d-1)} \cdot \left(\frac{\mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_-)} \wedge \frac{\mathrm{e}^s \cdot \mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_+)} \right).$$

Substituting into (3.21) we arrive at the inequality

$$\begin{aligned} \mathsf{E}\big[\mathrm{e}^{sX_A}\big] \cdot \mathsf{P}(X \in A) \leqslant \left(\cosh\left(\frac{s}{2}\right)\right)^{2d} \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^{-2} \times \\ \frac{1}{2} \cdot \left(\left(\frac{\mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_+)} \wedge \frac{\mathrm{e}^s \cdot \mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_-)} \right) + \left(\frac{\mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_-)} \wedge \frac{\mathrm{e}^s \cdot \mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_+)} \right) \right) \end{aligned}$$

Define $a_+ := \frac{\mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_+)}$ and $a_- := \frac{\mathsf{P}(X \in A)}{\mathsf{P}(X_{\hat{d}} \in A_-)}$. Of course, $a_+, a_- \geqslant 0$ and since

$$\begin{aligned} \mathsf{P}(X \in A) &= \mathsf{P}\big(X_{\hat{d}} \in A_+, X_d = 1\big) + \mathsf{P}\big(X_{\hat{d}} \in A_+, X_d = -1\big) \\ &\qquad + \mathsf{P}\big(X_{\hat{d}} \in A_-, X_d = 1\big) + \mathsf{P}\big(X_{\hat{d}} \in A_-, X_d = -1\big) \\ &= \mathsf{P}\big(X_{\hat{d}} \in A_+, X_d = 1\big) + \mathsf{P}\big(X_{\hat{d}} \in A_-, X_d = -1\big) \\ &= \frac{1}{2}\big(\mathsf{P}(X_{\hat{d}} \in A_+) + \mathsf{P}(X_{\hat{d}} \in A_-)\big), \end{aligned}$$

we have $a_+^{-1} + a_-^{-1} = 2$. At this stage, we observe that *if* for all $s \geqslant 0$ and all $a_+, a_- \geqslant 0$ satisfying $a_+^{-1} + a_-^{-1} = 2$, we have

$$\text{(3.23)} \qquad \Big(\big(a_+ \wedge (\mathrm{e}^s a_-)\big) + \big(a_- \wedge (\mathrm{e}^s a_+)\big) \Big) \leqslant 2 \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^2,$$

then the asserted inequality will hold and our proof will be complete, so we turn to establishing (3.23). Fix $s \geqslant 0$, and consider the equality-constrained optimization problem

$$\text{(3.24)} \qquad \begin{aligned} &\underset{a_+, a_-}{\text{maximize}} \quad \big(a_+ \wedge (\mathrm{e}^s a_-)\big) + \big(a_- \wedge (\mathrm{e}^s a_+)\big) \\ &\text{subject to} \quad \begin{cases} a_+, a_- \geqslant 0, \\ a_+^{-1} + a_-^{-1} = 2. \end{cases} \end{aligned}$$

Since (3.24) is identical to (3.16), Lemma (3.15) shows that there are (at most) two solutions of (3.24), given by $\left(\frac{1+\mathrm{e}^s}{2}, \frac{1+\mathrm{e}^s}{2\mathrm{e}^s}\right)$ and $\left(\frac{1+\mathrm{e}^s}{2\mathrm{e}^s}, \frac{1+\mathrm{e}^s}{2}\right)$, the corresponding (maximum) value of the objective function being

$$\frac{(1+\mathrm{e}^s)^2}{2\mathrm{e}^s} = \frac{1}{2} \cdot \left(\mathrm{e}^{-s} + 2 + \mathrm{e}^s\right) = 2 \cdot \left(\frac{\mathrm{e}^{\frac{s}{2}} + \mathrm{e}^{-\frac{s}{2}}}{2} \right)^2 = 2 \cdot \left(\cosh\left(\frac{s}{2}\right)\right)^2.$$

In other words, (3.23) indeed holds, and this completes the proof. □

(3.25). Exercise. In the Proof of Theorem (3.6) we employed the induction hypothesis by tacitly assuming that $\mathsf{P}(X \in A_+), \mathsf{P}(X \in A_-) > 0$. What happens if one of them is 0?

Proof of Corollary (3.7). Fix $s \geqslant 0$. For the first inequality it suffices to verify that $\left(\cosh\left(\frac{s}{2}\right)\right)^{2d} \leqslant \mathrm{e}^{s^2 \cdot \frac{d}{4}}$. Since considerable attention in [Tal95, p. 84] is devoted to proving this fact, we reproduce the author's proof almost verbatim by noting first that

$$\cosh\left(\frac{s}{2}\right) = 1 + \sum_{n=1}^{+\infty} \frac{s^{2n}}{2(2n)!},$$

and then that $2(2n)! \geqslant 4^n \cdot n!$ because the last inequality holds for $n = 1, 2$, and if $n \geqslant 3$, then

$$\frac{(2n)!}{n!} = (n+1) \cdots (2n) \geqslant 4^n.$$

Therefore,

$$\cosh\left(\frac{s}{2}\right) \leqslant 1 + \sum_{n=1}^{+\infty} \frac{s^{2n}}{4^n \cdot n!} = \mathrm{e}^{\frac{s^2}{4}},$$

which completes our verification. For the second inequality we employ Markov's inequality in the following computations: for $s > 0$,

$$\begin{aligned}\mathsf{P}\big(X_A > \varepsilon\sqrt{d}\big)\cdot\mathsf{P}(X \in A) &\leqslant \mathsf{P}\big(\mathrm{e}^{sX_A} > \mathrm{e}^{s\varepsilon\sqrt{d}}\big)\cdot\mathsf{P}(X\in A)\\ &\leqslant \mathrm{e}^{-s\varepsilon\sqrt{d}}\cdot\mathsf{E}\big[\mathrm{e}^{sX_A}\big]\cdot\mathsf{P}(X\in A) \leqslant \mathrm{e}^{-s\varepsilon\sqrt{d}}\mathrm{e}^{s^2\frac{d}{4}}.\end{aligned}$$

Minimizing the right-hand side with respect to $s \in \,]0, +\infty[$ leads to the unique minimizer $s_* = \frac{2\varepsilon}{\sqrt{d}}$, and the corresponding right-hand side for $s = s_*$ is $\mathrm{e}^{-\varepsilon^2}$. The second inequality follows. □

(3.26). Exercise. What sort of concentration bounds does one get by employing Hoeffding's inequalities on b^d for large d? Does it help to take a simpler notion of the metric on b^d compared to ρ_0?

Remark. A significant refinement of Talagrand's concentration inequality is possible if b^d is realized as a subset of $\mathbb{R}^d$ in a natural way and the set $A \subset \mathbb{R}^d$ in (3.6) is assumed to be *convex*: If $A \subset \mathbb{R}^d$ is convex and $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{b}^d)$, then there exists an absolute constant $c > 0$ such that[14]

(3.27) $$\mathsf{P}\big(\rho_2(X, A) > t\big)\cdot\mathsf{P}(X \in A) \leqslant \mathrm{e}^{-ct^2} \quad \text{for all } t > 0.$$

On the one hand, the estimate in Theorem (3.6) applies to *arbitrary* non-empty sets $A \subset \mathsf{b}^d$, while on the other hand, convexity of A is crucial in (3.27). Moreover, (3.27) features the standard Euclidean distance ρ_2 (for which $\rho_2(y, A) := \inf_{z\in A}\|y - z\|$ for all $y \in \mathbb{R}^d$) and not the (combinatorial) Hamming distance ρ_0 employed in (3.6) (cf. Exercise (3.3)).

§4. Concentration of Gaussian random vectors

« Let X denote a d-dimensional *standard normal* random vector.[15] Recall that this means X is a Gaussian random vector $X \overset{\text{dist}}{\sim} \mathfrak{N}(0, I_d)$, i.e., its probability density function is[16]

$$\mathbb{R}^d \ni x \longmapsto f_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}}}\mathrm{e}^{-\frac{1}{2}\|x\|^2} \in \,]0, +\infty[.$$

We will establish:

(4.1). Theorem. *If $d \in \mathbb{N}^*$ and $X \overset{\text{dist}}{\sim} \mathfrak{N}(0, I_d)$, then for all $\varepsilon \in \,]0, 1[$,*

$$\begin{cases}\mathsf{P}\left(\|X\| > \sqrt{\dfrac{d}{1-\varepsilon}}\right) < \mathrm{e}^{-\varepsilon^2\frac{d}{4}}, \quad \textit{and}\\[2ex] \mathsf{P}\left(\|X\| < \sqrt{d(1-\varepsilon)}\right) < \mathrm{e}^{-\varepsilon^2\frac{d}{4}}.\end{cases}$$

Remark. An immediate consequence of the two estimates in Theorem (4.1) is that

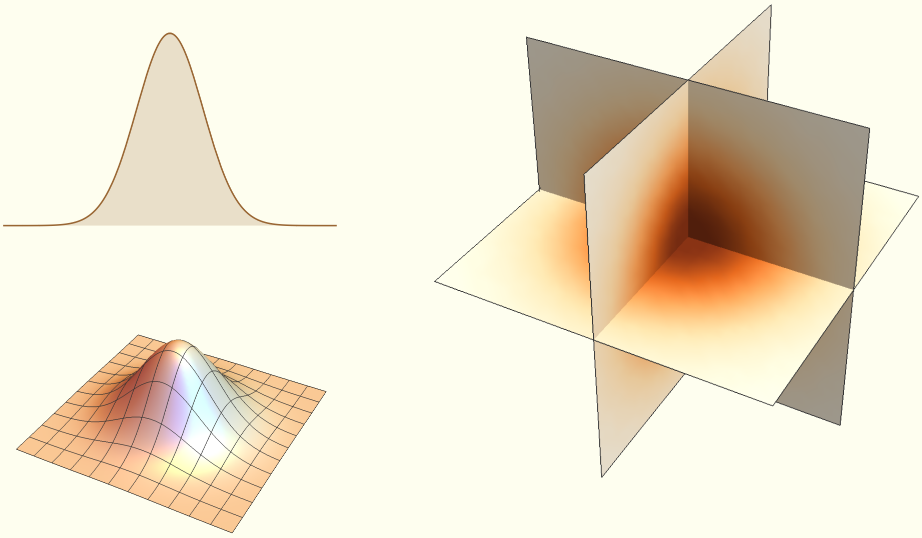
(4.2) $$\mathsf{P}\left(\sqrt{d(1-\varepsilon)} \leqslant \|X\| \leqslant \sqrt{\frac{d}{1-\varepsilon}}\right) \geqslant 1 - 2\mathrm{e}^{-\varepsilon^2\frac{d}{4}}.$$

[14] A proof of (3.27) may be found in the reference mentioned in footnote 11. See [Tal95, Theorem 4.1.1] for a more general version of this inequality.

[15] See `https://mathworld.wolfram.com/NormalDistribution.html` for details.

[16] See `https://mathworld.wolfram.com/GaussianFunction.html` for several interesting calculations.

This means that for large d, the samples drawn from $\mathfrak{N}(0, I_d)$ are to be found with high probability in a thin annulus around the spherical shell of radius $\sqrt{d}$. We immediately encounter some difficulty reconciling the preceding fact with our intuition derived from the standard normal densities in low dimensions, such as $d = 1$, $d = 2$, and $d = 3$ shown below:

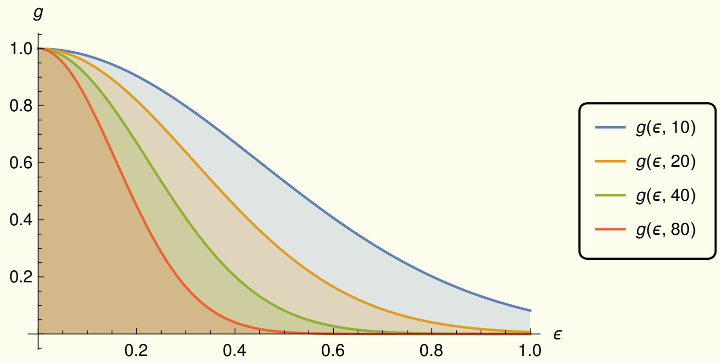


(The figure for $d = 3$ represents the probability density in terms of a color-map on three orthogonal planes passing through the origin: the deeper the shade, the higher the value of the probability density function.) These pictures strongly suggest that there ought to be a large concentration of samples close to the mean $0 \in \mathbb{R}^d$ of the distribution because the probability density peaks at 0 irrespective of the dimension. The estimate (4.2), however, directly contradicts that intuition: it says that as d increases, *very few* (independently drawn) samples will arise from the region immediately around the mean, and further that most such samples will lie in a thin spherical shell around the radius $\sqrt{d}$; see Remark (4.3) for further details.[17] Observe that the value of the probability density at 0 is $(2\pi)^{-\frac{d}{2}}$, while the value of the probability density on the $\sqrt{d}$-sphere is $(2\pi)^{-\frac{d}{2}} \mathrm{e}^{-\frac{d}{2}}$; the ratio of the latter to the former vanishes as $d \to +\infty$.

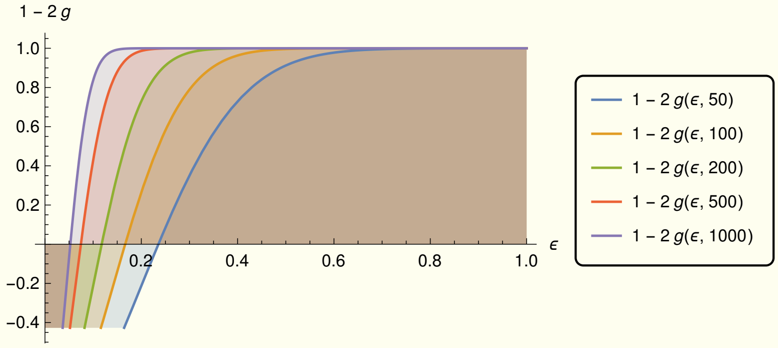
(4.3). Remark. The mapping $]0, 1[\times \mathbb{N}^* \ni (\varepsilon, d) \mapsto g(\varepsilon, d) := \mathrm{e}^{-\varepsilon^2 \frac{d}{4}}$ appearing on the right-hand sides of the estimates in Theorem (4.1) is worth studying in detail; here is a pictorial

[17] There is an urban legend about sampling from regions where the probability density function attains the peak values... One should not pay too much attention to such legends in high dimensions.

representation of g for four representative values of d:



The functions on the right-hand side of (4.2) for various moderately large values of d are given below:



(Of course, the domain of the curves on which it takes negative values are of little importance for they merely say that a lower bound of a certain probability is a negative number.) It follows from the preceding figure that $d = 500$ dimensional standard normal random vectors are to be found with probability at least 0.9 within the shell

$$\left\{y \in \mathbb{R}^{500} \;\middle|\; 20.62 \leqslant \|y\| \leqslant 24.25\right\}.$$

Drawing 10^4 independent samples from a $d = 10^3$-dimensional standard Gaussian and evaluating their norms yields histograms of the following type (note that $\sqrt{10^3} \approx 31.62$):

Proof of Theorem (4.1). Observe first that

$$\|X\| > \sqrt{\frac{d}{1-\varepsilon}} \quad \text{if and only if} \quad \|X\|^2 > \frac{d}{1-\varepsilon},$$

and if $t > 0$ is a parameter, then

$$\|X\| > \sqrt{\frac{d}{1-\varepsilon}} \quad \Leftrightarrow \quad \frac{t}{2}\|X\|^2 > \frac{td}{2(1-\varepsilon)} \quad \Leftrightarrow \quad \mathrm{e}^{\frac{t\|X\|^2}{2}} > \mathrm{e}^{\frac{td}{2(1-\varepsilon)}}.$$

At this stage we restrict $t \in]0,1[$ to ensure that $\mathsf{E}\big[\mathrm{e}^{\frac{t\|X\|^2}{2}}\big]$ is well-defined, and proceed with applying the Markov's inequality (2.1) to arrive at

$$\begin{aligned}\mathsf{P}\left(\|X\| > \sqrt{\frac{d}{1-\varepsilon}}\right) &= \mathsf{P}\left(\mathrm{e}^{\frac{t\|X\|^2}{2}} > \mathrm{e}^{\frac{td}{2(1-\varepsilon)}}\right)\\ &\leqslant \mathrm{e}^{-\frac{td}{2(1-\varepsilon)}} \cdot \mathsf{E}\big[\mathrm{e}^{\frac{t\|X\|^2}{2}}\big] \quad \text{for every } t \in]0,1[. \end{aligned} \tag{4.4}$$

It is clear that for each fixed $t \in]0,1[$,

$$\begin{aligned}\mathsf{E}\big[\mathrm{e}^{\frac{t\|X\|^2}{2}}\big] &= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{\frac{d}{2}}} \mathrm{e}^{-\frac{(1-t)}{2}\|x\|^2}\,\mathrm{d}x\\ &= \frac{1}{(1-t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{(1-t)^{\frac{-d}{2}}} \mathrm{e}^{-\frac{(1-t)}{2}\|x\|^2}\,\mathrm{d}x\\ &= (1-t)^{-\frac{d}{2}}.\end{aligned}$$

Substituting back into (4.4) we arrive at the inequality

$$\mathsf{P}\left(\|X\| > \sqrt{\frac{d}{1-\varepsilon}}\right) \leqslant (1-t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{td}{2(1-\varepsilon)}} = \exp\left(-\frac{d}{2}\left(\ln(1-t) + \frac{t}{1-\varepsilon}\right)\right).$$

Since the preceding inequality is valid for all $t \in]0,1[$, we minimize the right-hand side with respect to $t \in]0,1[$ to find that it admits a unique *minimizer* at $t_\star = \varepsilon$, which is a point in the domain $]0,1[$. (Check!) But then

$$\ln(1-\varepsilon) + \frac{\varepsilon}{1-\varepsilon} = \left(-\varepsilon - \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} - \cdots\right) + \left(\varepsilon + \varepsilon^2 + \varepsilon^3 + \cdots\right) \geqslant \frac{\varepsilon^2}{2},$$

which implies that

$$\exp\left(-\frac{d}{2}\left(\ln(1-t) + \frac{t}{1-\varepsilon}\right)\right) \leqslant \mathrm{e}^{-\varepsilon^2\frac{d}{4}}.$$

This proves the first inequality.

To prove the second inequality we proceed by noting that

$$\|X\| < \sqrt{d(1-\varepsilon)} \quad \text{if and only if} \quad -\|X\|^2 > -d(1-\varepsilon),$$

and therefore, if $t > 0$ is a parameter, then

$$\|X\| < \sqrt{d(1-\varepsilon)} \quad \Leftrightarrow \quad -\frac{t\|X\|^2}{2} > -\frac{td(1-\varepsilon)}{2} \quad \Leftrightarrow \quad \mathrm{e}^{-\frac{t\|X\|^2}{2}} > \mathrm{e}^{-\frac{td(1-\varepsilon)}{2}}.$$

Applying the Markov's inequality (2.1) leads to

$$\begin{aligned}\mathsf{P}\left(\|X\| < \sqrt{d(1-\varepsilon)}\right) &= \mathsf{P}\left(\mathrm{e}^{-\frac{t\|X\|^2}{2}} > \mathrm{e}^{-\frac{td(1-\varepsilon)}{2}}\right)\\ &\leqslant \mathrm{e}^{\frac{td(1-\varepsilon)}{2}} \mathsf{E}\big[\mathrm{e}^{-\frac{t\|X\|^2}{2}}\big] \quad \text{for all } t > 0.\end{aligned} \tag{4.5}$$

Of course, $\mathsf{E}\big[\mathrm{e}^{-\frac{t\|X\|^2}{2}}\big] = (1+t)^{-\frac{d}{2}}$ for each $t > 0$, and substituting in (4.5) gives us

$$\mathsf{P}\left(\|X\| < \sqrt{d(1-\varepsilon)}\right) \leqslant (1+t)^{-\frac{d}{2}} \mathrm{e}^{\frac{td(1-\varepsilon)}{2}} = \exp\left(-\frac{d}{2}\big(\ln(1+t) - t(1-\varepsilon)\big)\right).$$

Since this inequality is valid for all $t > 0$, we minimize the right-hand side with respect to t to find out that there exists a unique *minimizer* $t_\star = \frac{\varepsilon}{1-\varepsilon}$. (Check!) But

$$\begin{aligned}\ln(1+t_\star) - t_\star(1-\varepsilon) = \ln\left(1 + \frac{\varepsilon}{1-\varepsilon}\right) - \varepsilon &= -\ln(1-\varepsilon) - \varepsilon\\ &= -\left(-\varepsilon - \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} - \cdots\right) - \varepsilon \geqslant \frac{\varepsilon^2}{2},\end{aligned}$$

which gives

$$\exp\Bigl(-\frac{d}{2}\bigl(\ln(1+t)-t(1-\varepsilon)\bigr)\Bigr) \leqslant \mathrm{e}^{-\varepsilon^2\frac{d}{4}}.$$

The second inequality follows immediately from here. □

(4.6). Exercise. Fill out the details omitted in the preceding proof (— the two "Check!" points).

(4.7). Exercise. If $X \overset{\text{dist}}{\sim} \mathfrak{N}(0, I_d)$, then show by direct calculations (i.e., without relying on Theorem (4.1)) that,

$$\mathsf{P}\bigl(\lVert X\rVert > \sqrt{d+\varepsilon}\bigr) < \Bigl(\frac{d}{d+\varepsilon}\Bigr)^{-\frac{d}{2}}\mathrm{e}^{-\frac{\varepsilon}{2}} \quad \text{for } \varepsilon \geqslant 0,$$
$$\mathsf{P}\bigl(\lVert X\rVert < \sqrt{d-\varepsilon}\bigr) < \Bigl(\frac{d}{d-\varepsilon}\Bigr)^{-\frac{d}{2}}\mathrm{e}^{\frac{\varepsilon}{2}} \quad \text{for } \varepsilon \in \left]0, d\right[.$$

The preceding two inequalities provide estimates of probabilities pertaining to a symmetric shell with the $\sqrt{d}$-sphere at the center as opposed to estimates involving an asymmetric shell in Theorem (4.1).

(4.8). Exercise. Let $\alpha \in \mathbb{R}^d$ and let $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric and positive definite matrix. Suppose that $Y \overset{\text{dist}}{\sim} \mathfrak{N}(\alpha, \Sigma)$, i.e., the probability density function of Y is

$$\mathbb{R}^d \ni z \longmapsto f_Y(z) := \frac{1}{(2\pi)^{\frac{d}{2}}\det(\Sigma)^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}\left\langle z-\alpha, \Sigma^{-1}(z-\alpha)\right\rangle} \in \left]0, +\infty\right[.$$

Derive natural analogs of the inequalities in Theorem (4.1) that Y satisfies.

§5. Concentration of the uniform distribution on the unit cube

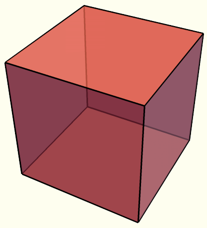
« The (closed) *unit cube* in d-dimension is the set

(5.1) $$\mathsf{C}^d := [-1, 1]^d$$

centered at $0 \in \mathbb{R}^d$, and the *uniform distribution* $\mathrm{Uniform}(\mathsf{C}^d)$ on C^d has the probability density function

(5.2) $$\mathbb{R}^d \ni y \longmapsto 2^{-d} \cdot 1_{\mathsf{C}^d}(y) \in \{0, 2^{-d}\},$$

that takes only two possible values. By definition, the higher the dimension d, the lower is the maximum value of the probability density. Here is the unit cube in dimension $d = 3$:

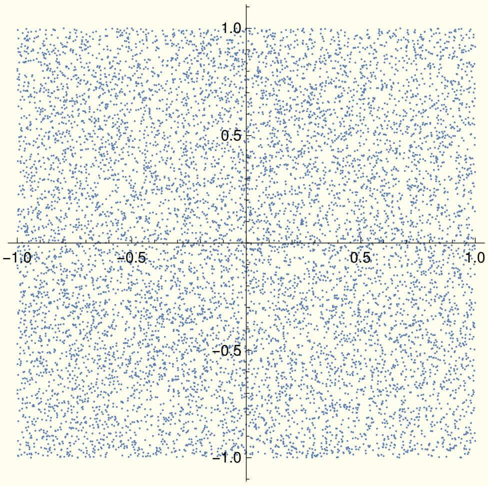


It may appear strange that, despite the name "uniform", sampling from the uniform distribution in high dimension begets samples from the vicinity of 0 quite rarely. In the sequel we shall investigate this phenomenon in some detail.

Let us begin with an exercise concerning the components of a uniform random vector:

(5.3). Exercise. If $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{C}^d)$, then show that the random variables in the family $(X_n)_{n=1}^d$ are independent of each other and that every $X_n \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{C}^1)$.

Here is a picture of 10^4 samples drawn independently at random from $\mathrm{Uniform}(\mathsf{C}^2)$:



While the samples appear to be somewhat uniformly spread on C^2, we shall demonstrate below that drawing independently and uniformly at random from the unit cube C^d does *not* produce samples that are "uniformly" scattered in the cube as d becomes large; instead, the samples congregate towards the corners of the cube C^d with high probability as d grows.

Remark. The aforementioned fact about nonuniform clumps of independently sampled uniform random vectors in high dimensions was exploited in [**MCB20**] in the context of robust optimization. The numerical experiments therein indicate that non-i.i.d. sampling may be needed for performance enhancement in certain classes of robust optimization problems.

Let $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{C}^d)$, and let us compute $\mathsf{E}[\|X\|^2]$. We quickly recall that the mean of $Y \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{C}^1)$ is 0 and its variance is

$$\mathrm{var}(Y) = \mathsf{E}[Y^2] - \mathsf{E}[Y]^2 = \int_{-1}^{1} \frac{1}{2} \cdot t^2 \, \mathrm{d}t = \frac{1}{3}.$$

Due to independence of the components of X in view of Exercise (5.3), it follows that

$$\mathsf{E}[\|X\|^2] = \mathsf{E}\left[\sum_{k=1}^{d} \|X_k\|^2\right] = \sum_{k=1}^{d} \mathsf{E}[\|X_k\|^2] = \frac{d}{3}.$$

Since $\mathsf{E}[X] = 0 \in \mathbb{R}^d$ due to symmetry about $0 \in \mathbb{R}^d$ of the uniform distribution, it follows that the scalar variance of X is

$$\mathsf{E}[\|X\|^2] = \frac{d}{3}.$$

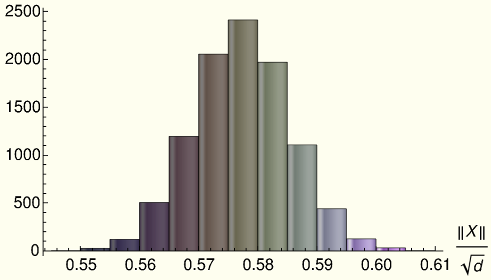
- Since the Euclidean distance between any two opposite faces of the cube C^d is 2, the preceding equality suggests that relatively few samples drawn from the uniform distribution on C^d would be found near the center of the cube C^d for large values of d compared to near its boundary.
- In view of the general principle mentioned in Remarks (2.4) about sums of independent random variables, the formula $\|X\|^2 = \sum_{n=1}^{d} X_n^2$ expressing $\|X\|^2$ as the sum of d independent and bounded variables suggests that one can expect sharp concentration of $\|X\|^2$ around its mathematical expectation $\frac{d}{3}$.

That the preceding description is indeed correct is the subject of:

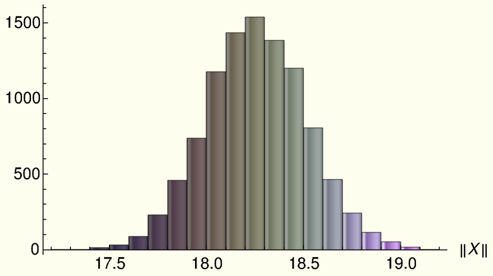
(5.4). Theorem. *If* $d \in \mathbb{N}^*$ *and* $X \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{C}^d)$*, then for every* $\varepsilon > 0$*,*

$$\mathsf{P}\left(\left|\left\|\frac{X}{\sqrt{d}}\right\|^2 - \frac{1}{3}\right| > \varepsilon\right) < 2\mathrm{e}^{-2\varepsilon^2 d}.$$

Remark. Here is a histogram depicting the sharp concentration around $\frac{1}{\sqrt{3}} \approx 0.5773$, of the ratio $\frac{\|X\|}{\sqrt{d}}$ (of the Euclidean norms to $\sqrt{d}$) for 10^4 samples drawn independently and uniformly randomly from the unit cube (5.1) in dimension $d = 10^3$.



Without the scaling by $\frac{1}{\sqrt{d}}$, the histogram of $\|X\|$ corresponding to the preceding data becomes sharply concentrated around $\sqrt{\frac{10^3}{3}} \approx 18.257$ as shown below:



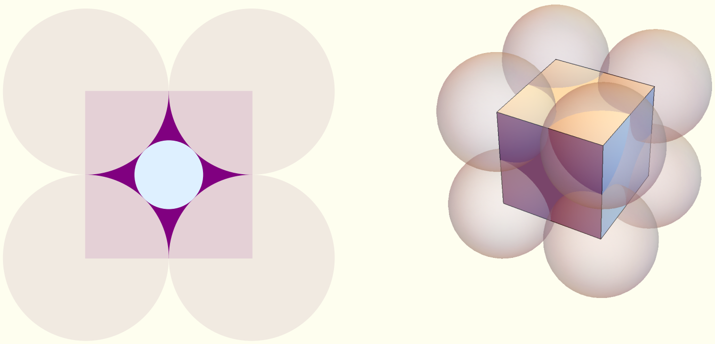
It should be evident from the latter figure that *very few* i.i.d. samples drawn from C^{10^3} lie anywhere 'near' the origin in $\mathbb{R}^{10^3}$.

Proof of Theorem (5.4). Fix $\varepsilon > 0$. Since Exercise (5.3) shows that the components $\{X_n \mid n = 1, \ldots, d\}$ are independent and identically distributed random variables with $X_1 \overset{\text{dist}}{\sim} \mathrm{Uniform}(\mathsf{C}^1)$. Consequently, $\frac{1}{d}X_n^2 \in [0, \frac{1}{d}]$ for each n. Hoeffding's inequality (2.3) now shows that

$$\mathsf{P}\left(\left|\frac{\|X\|^2}{d} - \frac{1}{3}\right| > \varepsilon\right) = \mathsf{P}\left(\left|\sum_{n=1}^{d} \frac{X_n^2}{d} - \frac{1}{3}\right| > \varepsilon\right) \leqslant 2\mathrm{e}^{-\frac{2\varepsilon^2}{\sum_{n=1}^{d} d^{-2}}} = 2\mathrm{e}^{-2\varepsilon^2 d}. \qquad \square$$

(5.5). Exercise. Let $d \in \mathbb{N}^*$, and consider the unit cube C^d. At each corner (i.e., a point with coordinates equal to 1 in magnitude) of the unit cube place a unit ball. The situation in $d = 2, 3$

are depicted below:



Place a ball centered at $0 \in \mathbb{R}^d$ such that it tangentially grazes each of the 2^d balls in the aforementioned array of balls (see e.g., the light cyan disk at the center of the left-hand figure). How does the radius of this ball change with d?

Fix $n \in \mathbb{N}^*$, and draw n samples independently from the unit cube C^d. On the one hand, since the uniform distribution on the unit cube C^d is non-zero everywhere inside the cube, it is quite natural to expect that, for a given $\varepsilon > 0$, every point in the cube is within ε distance of at least one of the samples (i.e., ε-dense), provided, of course, that n is *sufficiently large*. On the other hand, observe that Theorem (5.4) asserts that i.i.d. samples from $\mathrm{Uniform}(\mathsf{C}^d)$ exhibit tight concentration at the corners of C^d for all large d. In view of these two opposing facts, a natural question is how large would n have to be in order for the preceding "density" statement to hold? The next result provides a probabilistic estimate of the number n in terms of the threshold ε; we shall work with the ℓ_∞ distance for convenience, and for this purpose recall that for vectors $x, x' \in \mathbb{R}^d$ the ℓ_∞-distance between x and x' is given by $\rho_\infty(x, x') := \max_{i=1,\ldots,d} \lvert x_i - x'_i \rvert$.

(5.6). Theorem ([BG05, Theorem 4.2])**.** *Let $d \in \mathbb{N}^*$. Fix a small number $\varepsilon > 0$ such that $2 \cdot \varepsilon^{-1} \in \mathbb{N}^*$ and pick $n \in \mathbb{N}^*$. Suppose that $\mathcal{X}_n$ is a collection of n i.i.d.* $\mathrm{Uniform}(\mathsf{C}^d)$ *random vectors, and define the event*

$$A_{\varepsilon,n} := \bigl\{\textit{every point of } \mathsf{C}^d \textit{ is within } \rho_\infty\textit{-distance of } \varepsilon \textit{ from } \mathcal{X}_n\bigr\}.$$

Then

$$\mathsf{P}(A_{\varepsilon,n}) \geqslant 1 - \left(\tfrac{\varepsilon}{2}\right)^{-d}\left(1 - \left(\tfrac{\varepsilon}{2}\right)^{d}\right)^{n} \geqslant 1 - \left(\tfrac{2}{\varepsilon}\right)^{d} \mathrm{e}^{-n\cdot\left(\frac{2}{\varepsilon}\right)^{-d}}.$$

Remark. It is instructive to note from the last inequality that the regime when it is meaningful is when the estimate $\left(\tfrac{2}{\varepsilon}\right)^{d} \mathrm{e}^{-n\cdot\left(\frac{2}{\varepsilon}\right)^{-d}} < 1$ holds, or equivalently, whenever $n > d \cdot \left(\tfrac{2}{\varepsilon}\right)^{d} \cdot \ln\left(\tfrac{2}{\varepsilon}\right)$; the right-hand side grows faster than exponentially in the number of dimensions d! The proof of Theorem (5.6) involves the so-called 'metric entropy' argument, a mechanism that is extremely versatile but provides tight estimates only asymptotically. The procedure begins by defining a partition of the set under consideration in a way that the 'diameters' of the smaller subsets are within a certain desired threshold. In view of the topological structure of C^d, smaller sub-cubes become a natural choice (which is also the reason for picking the ρ_∞-metric in the theorem) for the partitioning.

Proof of Theorem (5.6). We subdivide the unit cube C^d into $\left(\tfrac{2}{\varepsilon}\right)^{d}$ many small cubes of edge length ε such that the small cubes $\bigl\{C'_k \bigm| k = 1, \ldots, \left(\tfrac{2}{\varepsilon}\right)^{d}\bigr\}$ are either mutually disjoint or overlap only along their faces; consequently, the set of overlaps is of zero volume in C^d. Observe that if each of these small cubes C'_k-s contains at least one element of $\mathcal{X}_n$, then the event $A_{\varepsilon,n}$ holds. The probability that a particular small cube C'_i contains a uniformly sampled vector from C^d is $\left(\tfrac{\varepsilon}{2}\right)^{d}$ in view of (5.2), which means that *none* of the samples in $\mathcal{X}_n$ lies in this particular small cube with probability $\mathsf{P}(\mathcal{X}_n \cap C_i' = \varnothing) = \big(1 - (\frac{\varepsilon}{2})^d\big)^n$ due to independence. There are $(\frac{2}{\varepsilon})^d$ many small cubes, which means that the probability that at least one of the small cubes does not intersect with $\mathcal{X}_n$ is

$$\mathsf{P}\left(\bigcup_{i=1}^{(\frac{2}{\varepsilon})^d} \{\mathsf{X}_n \cap C_i' = \varnothing\}\right) \leqslant \sum_{i=1}^{(\frac{2}{\varepsilon})^d} \mathsf{P}(\mathcal{X}_n \cap C_i' = \varnothing) = (\tfrac{\varepsilon}{2})^{-d}\big(1 - (\tfrac{\varepsilon}{2})^d\big)^n.$$

To wit, the probability of the complement of the event $A_{\varepsilon,n}$ is at most $(\frac{\varepsilon}{2})^{-d}\big(1 - (\frac{\varepsilon}{2})^d\big)^n$, and the first inequality follows. The second one follows immediately from the first in view of the fact that $(1-a)^n = \mathrm{e}^{n\ln(1-a)} \leqslant \mathrm{e}^{-an}$ for $a \in \,]0, 1[$. □

§6. Two applications

«We study two important applications of the concentration phenomenon in data science and optimization. The first application in §6.1 — the Johnson-Lindenstrauss lemma — is a story of positivity, while the second application in §6.2 is a caveat.

§6.1. The Johnson-Lindenstrauss lemma. The Johnson-Lindenstrauss lemma has become an extremely useful tool in data science today. We shall derive a basic version of this lemma using the results we have developed in §4.

The Johnson-Lindenstrauss lemma concerns the matter of orthogonally projecting a finite set of points from a high-dimensional Euclidean space into reasonably low dimensional subspaces while maintaining the mutual separation between the points within given tolerance margins. This is the subject of the following:

(6.1). Theorem (Johnson-Lindenstrauss Lemma). *Fix $d \in \mathbb{N}^*$, let $N \in \mathbb{N}^*$, and pick $\varepsilon, \eta \in \,]0, 1[$. Fix an integer $k \geqslant \frac{4}{\varepsilon^2}\ln\big(\frac{2N^2}{\eta}\big)$. Suppose that $x_1, \ldots, x_N \in \mathbb{R}^d$. If $k \leqslant d$, then choose a k-dimensional subspace L of $\mathbb{R}^d$ uniformly randomly from the family of all k-dimensional subspaces of $\mathbb{R}^d$, and let $x_i^L := \pi_L(x_i)$ denote the orthoprojection of x_i on L for each i. Then*

$$\mathsf{P}\left((1-\eta)\|x_m - x_n\| \leqslant \sqrt{\frac{d}{k}}\,\|x_m^L - x_n^L\| \leqslant (1+\eta)\|x_m - x_n\| \;\textit{ for all } m, n\right) \geqslant 1 - \varepsilon.$$

(6.2). Remarks. Several points need commentary:

- Theorem (6.1) asserts, in rough and broad strokes, that if L is a subspace of a high-dimensional ambient Euclidean space and if $\{x_i \mid i = 1, \ldots, N\}$ is a family of vectors in the ambient space, then the mutual distances between the vectors are roughly preserved under the orthogonal projection to the subspace. The probability P concerns the joint event of all the mutual distances — the qualifier 'for all m, n' is *inside* the probability.
- The numbers ε, η are 'tolerances'. Of course, the orthogonal projection operation introduces errors in the mutual distances; for one, vectors are liable to shrink under orthoprojection. The number η measures the error between the mutual distances $\sqrt{\frac{d}{k}}\|x_n^L - x_m^L\|$ and $\|x_n - x_m\|$ that we are happy to accept. The number ε denotes the probability of violation of the (joint) event under consideration, and $(1-\varepsilon)$ stands for the confidence with which the mutual distances are preserved.
- The dimension d of the ambient Euclidean space is *irrelevant* insofar as the dimension of the subspace L is concerned — it does not influence the lower bound of k in any way.
- Theorem (6.1) is most spectacular in the large d regime. Let us take a look at some numerics. Suppose that $d = 10^6$, that we are happy with $\eta = 10^{-1}$ and $\varepsilon = 0.05$, and that there are $N = 50$ points in the d-dimensional space. Then k should be at least, roughly, 17350.

Next suppose that $\dim = 10^9$ while maintaining the rest of the preceding parameters. Then $k \geqslant 17350$ once again.

A (sketch of a) proof of Theorem (6.1) will be given after extracting the basic inequality that lies at the heart of it, and it is captured by the following:

(6.3). Lemma. *Consider a random vector* $X \overset{\text{dist}}{\sim} \mathfrak{N}(0, I_d)$ *and let* $L \subset \mathbb{R}^d$ *be a subspace of dimension* k*. If* $\pi_L(X)$ *is the orthogonal projection of* X *on* L*, then*

$$\text{(6.4)} \qquad \textit{for } \varepsilon \in \,]0,1[, \qquad \begin{cases} \mathsf{P}\left(\sqrt{\dfrac{d}{k}}\,\|\pi_L(X)\| > \dfrac{\|X\|}{1-\varepsilon}\right) \leqslant \mathrm{e}^{-\varepsilon^2\frac{k}{4}} + \mathrm{e}^{-\varepsilon^2\frac{d}{4}} & \textit{and} \\ \mathsf{P}\left(\sqrt{\dfrac{d}{k}}\,\|\pi_L(X)\| < \|X\|(1-\varepsilon)\right) \leqslant \mathrm{e}^{-\varepsilon^2\frac{k}{4}} + \mathrm{e}^{-\varepsilon^2\frac{d}{4}}. \end{cases}$$

Consequently, for $\varepsilon \in \,]0,1[$,

$$\text{(6.5)} \qquad \mathsf{P}\left((1-\varepsilon)\sqrt{\frac{k}{d}}\,\|X\| \leqslant \|\pi_L(X)\| \leqslant \sqrt{\frac{k}{d}}\,\frac{1}{(1-\varepsilon)}\|X\|\right) \geqslant 1 - 2\left(\mathrm{e}^{-\varepsilon^2\frac{k}{4}} + \mathrm{e}^{-\varepsilon^2\frac{d}{4}}\right).$$

(6.5) is the key probabilistic inequality that drives the Johnson-Lindenstrauss Lemma (Theorem (6.1)) and the engine behind establishing Lemma (6.3) will be the probability estimates in Theorem (4.1):

Proof of Lemma (6.3). First pick an orthonormal basis on L and then extend the basis to $\mathbb{R}^d$ by picking orthonormal vectors by means of the Gram-Schmidt technique. The transformation from the original orthonormal basis on $\mathbb{R}^d$ to this new basis is via an orthogonal matrix; since $\mathfrak{N}(0, I_d)$ is spherically symmetric, the distribution of X as represented in the new basis remains unchanged, and we continue to denote this random vector by X.

Let us establish the first inequality in (6.4). It is clear that, by construction, $\pi_L(X) \overset{\text{dist}}{\sim} \mathfrak{N}(0, I_k)$. Theorem (4.1), therefore, applies to the random vector $\pi_L(X)$, giving the estimates

$$\text{(6.6)} \qquad \begin{cases} \mathsf{P}\left(\|\pi_L(X)\| > \sqrt{\dfrac{k}{1-\varepsilon}}\right) \leqslant \mathrm{e}^{-\varepsilon^2\frac{k}{4}}, & \text{and} \\ \mathsf{P}\left(\|\pi_L(X)\| < \sqrt{k(1-\varepsilon)}\right) \leqslant \mathrm{e}^{-\varepsilon^2\frac{k}{4}}. \end{cases}$$

Since both the first estimate of Theorem (4.1) and the second estimate in (6.6) hold, we have

$$\mathsf{P}\left(\frac{\|X\|}{\sqrt{d(1-\varepsilon)}} \geqslant 1\right) \geqslant 1 - \mathrm{e}^{-\varepsilon^2\frac{d}{4}}$$

and

$$\mathsf{P}\left(\sqrt{\frac{1-\varepsilon}{k}}\,\|\pi_L(X)\| \leqslant 1\right) \geqslant 1 - \mathrm{e}^{-\varepsilon^2\frac{k}{4}}.$$

If two events have large probabilities, then their intersection has large probability as well. Indeed, if $A_1, A_2 \subset \Omega$ are events satisfying $\mathsf{P}(A_1) \geqslant 1 - p_1$ and $\mathsf{P}(A_2) \geqslant 1 - p_2$ for $p_1, p_2 \in \,]0,1[$, then obviously $\mathsf{P}(\Omega \setminus A_1) \leqslant p_1$ and $\mathsf{P}(\Omega \setminus A_2) \leqslant p_2$, and this leads to the (standard) estimate (a variant of the so-called union bound)

$$\text{(6.7)} \qquad \begin{aligned} \mathsf{P}(A_2 \cap A_2) = 1 - \mathsf{P}\big(\Omega \setminus (A_1 \cap A_2)\big) &= 1 - \mathsf{P}\big((\Omega \setminus A_1) \cup (\Omega \setminus A_2)\big) \\ &\geqslant 1 - \mathsf{P}(\Omega \setminus A_1) - \mathsf{P}(\Omega \setminus A_2) \geqslant 1 - p_1 - p_2. \end{aligned}$$

In the present case an application of the preceding inequality gives us

$$\mathsf{P}\left(\frac{\|X\|}{\sqrt{d(1-\varepsilon)}} \geqslant \sqrt{\frac{1-\varepsilon}{k}}\,\|\pi_L(X)\|\right) = \mathsf{P}\left(\sqrt{\frac{d}{k}}\,\|\pi_L(X)\| \leqslant \frac{\|X\|}{1-\varepsilon}\right) \geqslant 1 - \left(\mathrm{e}^{-\varepsilon^2\frac{d}{4}} + \mathrm{e}^{-\varepsilon^2\frac{k}{4}}\right).$$

The other inequality follows from similar considerations and is left as Exercise (6.8). Applying (6.7) on these two inequalities gives us (6.5) at once. □

(6.8). Exercise. Establish the second inequality in Lemma (6.3) in complete detail.

Sketch of a proof of Theorem (6.1). It should be clear that one needs to get a sense of the family $\mathrm{RP}(k;d)$ of all k-dimensional subspaces of $\mathbb{R}^d$, the **real projective space** because the 'probability' in Theorem (6.1) is on this family.

Step 1. To this end, let n, k be positive integers satisfying $k < n$. Consider the set of all k-dimensional subspaces of $\mathbb{R}^n$. From a fixed such subspace we can pick k linearly independent vectors in $\mathbb{R}^n$ that span the subspace, and concatenate them to form an $n \times k$ matrix. It is readily observed that a suitable permutation of the rows of this matrix brings the k linearly independent rows to the top of the matrix, and, since the upper $k \times k$ submatrix has rank k, a multiplication on the right by a suitable non-singular $k \times k$ matrix brings it to the form in which the top $k \times k$ submatrix is precisely the k-dimensional identity, i.e., to the form $\begin{pmatrix} I_k \\ Z \end{pmatrix}$ for some $Z \in \mathbb{R}^{(n-k)\times k}$. More formally, let α be a subset of $\{1, \ldots, n\}$ containing k elements, and let the set of $n \times k$ matrices with the k independent rows corresponding to the elements in α be denoted by U_α. If A is an arbitrary element of U_α, then we permute the rows of A to bring all k linearly independent rows to the top, multiply a suitable $k \times k$ matrix (depending on A) on the right of the resulting matrix to arrive at the form $\begin{pmatrix} I \\ Z_A \end{pmatrix}$ for some $Z_A \in \mathbb{R}^{(n-k)\times k}$, and finally stack the columns of Z_A on top of each other in a predefined way to map Z_A into $\mathbb{R}^{k(n-k)}$. The composite, denoted by φ_α, is a smooth bijection from the set U_α into $\mathbb{R}^{k(n-k)}$. For each k-element subset of $\{1, \ldots, n\}$, therefore, this mapping $\varphi_\alpha : U_\alpha \longrightarrow \mathbb{R}^{k(n-k)}$ is a *chart map* in the language of smooth manifolds, and the union of the U_αs as α varies over all k-element subsets of $\{1, \ldots, n\}$ covers the set $\mathrm{RP}(k;n)$. Of course, the dimension of $\mathrm{RP}(k;n)$ is $k(n-k)$.

Step 2. If $\mathbb{R}^d$ is the ambient space, then we need a mechanism to equip the family $\mathrm{RP}(k;d)$ with a probability measure. It turns out that $\mathrm{RP}(k;d)$ is a compact metric space under the natural Hausdorff metric between the unit balls of two such subspaces, and this particular metric is invariant under the group of orthogonal transformations on $\mathbb{R}^d$. These facts lead to, after appealing to the appropriate (and somewhat abstract) results, the fact that there exists a unique probability measure on the Borel subsets of $\mathrm{RP}(k;d)$ that is invariant relative to the action of the aforementioned orthogonal transformations.[18]

Step 3. It is not important to describe this probability measure as long as one can sample from it. To this end, observe that if $k < n$ (which is our premise), then sampling k vectors from the standard normal distribution on $\mathbb{R}^d$ produces k linearly independent vectors with probability 1, and their linear span is a k dimensional subspace. Moreover, the standard normal distribution is invariant under orthogonal transformations on $\mathbb{R}^d$. Consequently, the uniqueness statement in Step 2 implies that k-dimensional subspaces produced in this way is effectively sampling from the distribution mentioned in Step 2.

The remainder of the proof is delegated to the following (slightly nontrivial) Exercise. □

(6.9). Exercise. Complete the remainder of the proof of Theorem (6.1) starting from the final statement of Step 2.

[18] This is the sketch part…

§6.2. The scenario approach in robust optimization. Consider the **robust optimization** problem

$$
\begin{aligned}
&\underset{x}{\text{minimize}} && c(x)\\
&\text{subject to} && \begin{cases} g(x,\xi) \leqslant 0 \text{ for each } \xi \in \Xi,\\ x \in S,\\ S \times \Xi \subset \mathbb{R}^n \times \mathbb{R}^m \text{ closed and non-empty,}\end{cases}
\end{aligned}
\tag{6.10}
$$

where $c : \mathbb{R}^n \longrightarrow [0, +\infty[$ is a continuous objective function and the mapping g is the stacked vector function of continuous functions $g_i : \mathbb{R}^n \times \Xi \longrightarrow \mathbb{R}$ for $i = 1, \ldots, p$. We shall deliberately refrain from generalizing to the setting of (6.10) beyond the current stage, referring the reader to [MCB20] for a discussion. The parameter ξ plays the role of uncertainty, and Ξ is the set of possible uncertainties.

A standard optimization problem consists of the minimization of a certain objective function subject to certain constraints; the point of departure of (6.10) from such a standard optimization problem is the presence of the family Ξ of uncertainties in the constraints: *each* constraint $g(x,\xi) \leqslant 0$ corresponding to $\xi \in \Xi$ must be satisfied by a solution of (6.10). Problems such as (6.10) arises in a plethora of situations in engineering; we refer the reader to [BTEN09] for a comprehensive discussion and a treatment of such problems and robust optimization in general.

If, on the one hand, Ξ is a finite set, then we have a finite family of constraints in (6.10). On the other hand, if Ξ is uncountable, e.g., Ξ is compact interval with non-empty interior, then the corresponding family of constraints in (6.10) is infinite. Such optimization problems are known to be hard, and are known as **semi-infinite programs**. Observe that the larger the set Ξ is, the smaller is the feasible set of (6.10); indeed, if $\Xi', \Xi'' \subset \Xi$ are finite sets and $\Xi' \subset \Xi''$, then the value of the problem with Ξ' as the set of uncertainties is smaller than that with Ξ'' as the set of uncertainties; in this sense the behavior of the value is *monotone non-decreasing*.

(6.11). Remark. Standard minmax problems of the form

$$\inf_{x\in S} \sup_{\theta\in\Theta} C(x,\theta)$$

where $C : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow [0, +\infty[$ is a continuous objective function, $S \subset \mathbb{R}^n$ is the (closed) set of optimization variables, $\Theta \subset \mathbb{R}^m$ is a (closed) set of uncertainties, can be readily recast in the language of (6.10) by means of the introduction of a *slack variable*. Indeed, it is an easy exercise to note that the value of the optimization problem

$$
\begin{aligned}
&\underset{t,x}{\text{minimize}} && t\\
&\text{subject to} && \begin{cases} C(x,\theta) - t \leqslant 0 \text{ for each } \theta \in \Theta,\\ (t,x) \in [0, +\infty[\times S,\end{cases}
\end{aligned}
$$

which is of the form (6.10), is identical (in the sense of having equal values) to that of the minmax problem indicated above.

(6.12). Exercise. Argue that the objective function c in (6.10) being independent of ξ is no loss of generality, and in fact, it suffices to consider c to be *linear*.

A popular technique in robust optimization is known as the **scenario approach**. It consists of selecting randomly and independently a finite subset $\xi_1, \ldots, \xi_N$ from the set Ξ of uncertainties, and adjoining the corresponding constraints to the original robust optimization problem. For instance, in the context of (6.10) one constructs the optimization problem

$$\begin{aligned}&\underset{x}{\text{minimize}} && c(x)\\ &\text{subject to} && \begin{cases} g(x,\xi_k)\leqslant 0 \text{ for each } k=1,\ldots,N,\\ x\in S,\\ S\times\Xi\subset\mathbb{R}^n\times\mathbb{R}^m \text{ closed and non-empty,}\end{cases}\end{aligned}\tag{6.13}$$

after obtaining N samples $\xi_1,\ldots,\xi_N$ from Ξ. Clearly, (6.13) serves as a surrogate of (6.10) in some sense although the value of (6.13) is a random variable.[19] Indeed, if all the elements of Ξ *could* be sampled in the aforementioned fashion, then there would be no difference between (6.10) and (6.13). Even if sampling all the elements of Ξ is not possible, as long as N is sufficiently large, it is natural to expect that the value of (6.13) would be close to that of (6.10) in some probabilistic sense.

While we shall not concern ourselves with the precise sense in which the two aforementioned values would be close, being content to refer the reader to [**Ram18**, **MCB20**] instead, the interesting observation at this stage consists of the behavior of the value of (6.13) as the dimension m of the set Ξ of uncertainties increases. We have, in the preceding sections, studied several situations in high dimensions in which independent and identical distributed samples tend to 'congregate' on particular regions of the space. In particular, we have witnessed in Chapter 5 that independent samples generated uniformly randomly from the unit cube C^m tend to 'congregate' towards the corners of the cube as m becomes large. Such behavior of the samples may lead to adverse effects in the values of the robust optimization problem (6.13) even when N appears to be 'large'.

As an illustration, consider the following numerical example:

$$\begin{aligned}&\underset{t,x}{\text{minimize}} && t\\ &\text{subject to} && \begin{cases} x\|\xi\|_\infty-\|\xi\|_\infty^2-t\leqslant 0 \text{ for all } \xi,\\ (t,x)\in\mathbb{R}\times[0,1],\ \ \xi\in[-1,1]^m,\end{cases}\end{aligned}\tag{6.14}$$

This problem is so simple that it can be solved by hand. Indeed, it is not difficult to see (cf. Remark (6.11)) that the **value** of the problem (6.14) is identical to the value of the minmax problem

$$V_\star := \min_{x\in[0,1]}\ \sup_{\xi\in[-1,1]^m}\ \left(x\|\xi\|_\infty-\|\xi\|_\infty^2\right).$$

It follows at once that for each fixed $x\in[0,1]$, any ξ with $\|\xi\|_\infty=\frac{x}{2}$ maximizes $\left(x\|\xi\|_\infty-\|\xi\|_\infty^2\right)$, the corresponding maximum value being $\frac{x^2}{4}$; therefore, $\min_{x\in[0,1]}\frac{x^2}{4}=0$ is the value of this minmax problem and of (6.14). Moreover, we note that

$$\check{V} := \min_{x\in[0,1]}\ \min_{\xi\in[-1,1]^m}\left(x\|\xi\|_\infty-\|\xi\|_\infty^2\right)=-1$$

is the *worst* possible (in a reasonable sense) estimate of the value $V_\star$ because the first term in the objective is always non-negative.

Let us see what the scenario approach tells us in the context of (6.14). The approach itself consists of

- fixing a positive integer N,
- sampling N times identically and uniformly randomly from $\Xi:=[-1,1]^m$ to generate the samples $\xi_1,\ldots,\xi_N$, and

[19] We shall ignore technicalities concerning *why* the value is a bona fide random variable.

- in analogy with (6.13), solving the problem

$$\begin{aligned} &\underset{t,x}{\text{minimize}} && t \\ &\text{subject to} && \begin{cases} x\|\xi_k\|_\infty - \|\xi_k\|_\infty^2 - t \leqslant 0 \text{ for all } k = 1, \ldots, N, \\ (t, x) \in \mathbb{R} \times [0, 1],\ \ \xi_k \in [-1, 1]^m. \end{cases} \end{aligned} \tag{6.15}$$

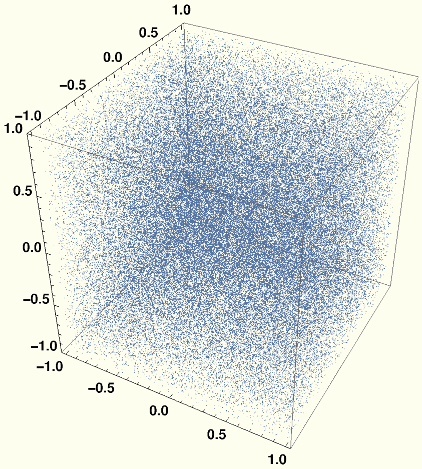
We denote the value of the optimization problem (6.15) by $V_*(N, m)$, thereby making explicit the dependence of the value on both the number of samples N and the dimension m of the set Ξ of uncertainties; it should be clear that $V_*(N, m)$ is a random variable.

The following data-set for $V_*(N, m)$ was recorded from one numerical experiment, where we recall that the true value of the problem is $V_* = 0$ and the worst possible estimate is $\check{V} = -1$:[20]

$V_*(N, m)$

	Number of samples (N)				
dim (m)	10	10^2	10^3	10^4	10^5
2	-0.0808521	-0.0027655	-0.0024113	-0.00006952	-5.8517×10^{-6}
3	-0.514566	-0.0155123	-0.0166664	-0.00071971	-0.000470249
5	-0.479553	-0.18534	-0.0740244	-0.0262403	-0.00899445
10	-0.443152	-0.127072	-0.291861	-0.147158	-0.0690109
20	-0.827773	-0.567134	-0.499702	-0.436	-0.264414
50	-0.920135	-0.850023	-0.782451	-0.646217	-0.671429
100	-0.924552	-0.933968	-0.862346	-0.819696	-0.763204
200	-0.946027	-0.95713	-0.938177	-0.896738	-0.887372

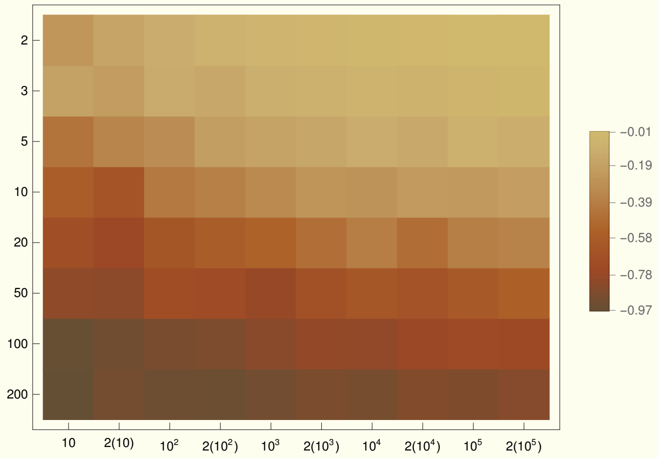
Observe that the best results are (naturally!) obtained when the dimension m is 2 and the number of samples N is large; e.g., $V_*(10^5, 2)$ almost equal to the true value 0. Here is a figure depicting $\Xi = [-1, 1]^3$ with 10^5 points sampled independently and uniformly randomly from it; the density of the samples is noteworthy:



While the preceding numbers correspond to one particular experiment, the behavior illustrated above is typical: As the number of samples N decreases for a fixed m, the corresponding value deviates progressively more from the true value 0; if N is held fixed, then the corresponding value falls away from 0 as m increases (i.e., the error increases). A matrix plot of the values computed for a different (and larger) set of experiments in which the vertical axis depicts the dimensions

[20] Bob Hanlon provided the seed code that was employed in this experiment on Mathematica; the relevant page is https://mathematica.stackexchange.com/questions/246128/list-of-constraints-in-minimize.

$m = 2, 3, 5, 10, 20, 50, 100, 200$ from top to bottom, and the horizontal axis depicts the number of samples $N = 10, 2(10), 10^2, 2(10^2), 10^3, 2(10^3), 10^4, 2(10^4), 10^5, 2(10^5)$ from left to right, is shown below:



In this context, we remind the reader that the optimal value of (6.14) is $V_\star = 0$ and the worst possible estimate of (6.14) is $\check{V} = -1$.

Whether such behavior of the values of semi-infinite programs under the scenario approach are satisfactory or not is difficult to assess unilaterally and uniformly across the spectrum of robust optimization problems (6.10), and such conclusions are best left to the judgment of the practitioners concerned. However, it is undeniable that the main culprit in the preceding example is our reliance on i.i.d. samples despite with the fact that i.i.d. samples of high dimensional random vectors tend to concentrate with high probability around certain regions of the space leaving the rest of the space unexplored; this feature leads to a preference for certain (typically thin) regions of the sample space of the algorithm, and unless the optimizers are in these thin sets, the quality of approximation may be low. The preceding observations clearly point to the fact that there is still scope to develop general, computationally feasible, and tight approximation schemes in robust optimization problems, especially in high dimensions; one such approximation method [**DACC22**] involving better (non-i.i.d.) sampling techniques has recently been reported.

It is important to ponder whether it is indeed relevant to consider the uncertainty to take values in some high-dimensional space. Here is an example from control theory that should strike a chord immediately: Consider a finite horizon **robust optimal control problem**

$$\tag{6.16} \begin{aligned} &\inf_{u}\ \sup_{w} && \sum_{t=0}^{N-1} c\big(x(t), u(t)\big) + c_F\big(x(N)\big) \\ &\text{subject to} && \begin{cases} x(t+1) = f\big(x(t), u(t), w(t)\big), \\ x(0) = \bar{x}, \\ u(t) \in \mathbb{U} \text{ for each } t, \\ w(t) \in \mathbb{W} \text{ for each } t, \\ u \coloneqq \big(u(0), u(1), \ldots, u(N-1)\big), \\ w \coloneqq \big(w(0), w(1), \ldots, w(N-1)\big), \end{cases} \end{aligned}$$

where N is a pre-specified positive integer playing the role of the *control horizon*, $\mathbb{U} \subset \mathbb{R}^m$ is a non-empty set of *admissible control actions*, $\mathbb{W} \subset \mathbb{R}^p$ is a non-empty set of *uncertainties*, the mapping $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \ni (x, u, w) \mapsto f(x, u, w) \in \mathbb{R}^d$ describing the *dynamics* of the process is continuous, as are the *cost-per-stage function* $c : \mathbb{R}^d \times \mathbb{R}^m \longrightarrow [0, +\infty[$ and the *final cost function* $c_F : \mathbb{R}^d \longrightarrow [0, +\infty[$. The task of the *control sequence* u in (6.16) is to minimize the *worst case* cost accrued due to the sequence w of uncertainties and/or disturbances that affect the dynamical system described by f. Such uncertainties may arise due to modeling inaccuracies, exogenous noise, etc.[21] The epithet *robust* refers to the fact that the control guards against the *worst case performance* due to such uncertainties.

The optimal control problem (6.16) is, in general, difficult to solve analytically except in the simplest of cases, and even numerical algorithms for solving it are hard to arrive at. It is not difficult to see that (6.16) is a robust optimization problem with the uncertainty variable w and the sequence u of control actions playing the role of the decision variables, and one often attempts to solve (6.16) via the scenario approach when the set $\mathbb{W}$ of uncertainties admits a reasonably concrete description, such as polytopes, ellipsoids, etc. Samples from $\mathbb{W}^N$ are drawn in an i.i.d. fashion and the resulting (random) sampled minmax problem is solved as a surrogate of (6.16). One realizes immediately that whenever N is large, the dimension of the set $\mathbb{W}^N$ is pN, and the concentration phenomenon is liable to influence the resulting programs in dramatic ways.

[21] There is a somewhat subtle point at play here. While guarding against uncertainties in the model — parametric or otherwise — is a good idea in general (and is the central tenet of robust control) and is generally accepted as standard practice, treating the exogenous noise from the worst case view point typically leads to conservative designs. Unless the applications are safety-critical, it may be better to avoid conservative designs and opt instead for a stochastic model of the noise. In such a situation, the objective function is usually replaced by the expected value (with respect to the probability distribution of the underlying model for the random noise process) of a finite sum of cost functions over the random part of each w_t; semi-infinite programs arise naturally once again when worst case effects with respect to the modeling uncertainty part of the w_t-s are considered in this setting.

Appendix A. Asymptotics of Laplace integrals: the principal term

This appendix contains a fundamental result on the asymptotics of Laplace integrals. A proof may be found in the cited reference.

(A.1). Theorem ([Zor16, §19.2.4, p. 612, Theorem 1]). *Consider the integral*

$$F(\lambda) := \int_a^b f(t) \mathrm{e}^{\lambda g(t)} \, \mathrm{d}t, \tag{A.2}$$

where $-\infty < a < b < +\infty$, *the functions* $f, g : [a, b] \longrightarrow \mathbb{R}$ *are continuous, and* $\lambda > 0$ *is a parameter. Suppose that* $\max_{t \in [a,b]} g(t)$ *is attained at a unique point* $t_* \in [a, b]$. *Moreover, assume that*

- $f(t_*) \neq 0$ *and*
- $f(t) = f(t_*) + O(t - t_*)$ *as* $t \to t_*$ *on* $[a, b]$.

(A.1)-a) *If* g *is twice continuously differentiable on a neighborhood of* t_*, *and* $t_* = a$ *with* $\frac{\mathrm{d}g}{\mathrm{d}t}(t_*) \neq 0$, *then*

$$F(\lambda) = \frac{f(t_*)}{-\frac{\mathrm{d}g}{\mathrm{d}t}(t_*)} \cdot \mathrm{e}^{\lambda g(t_*)} \cdot \lambda^{-1}\big(1 + O(\lambda^{-1})\big) \quad \text{as } \lambda \to +\infty.$$

(A.1)-b) *If* g *is thrice continuously differentiable on a neighborhood of* t_*, *and* $t_* \in \,]a, b[$ *with* $\frac{\mathrm{d}^2 g}{\mathrm{d}t^2}(t_*) \neq 0$, *then*

$$F(\lambda) = \sqrt{\frac{2\pi}{-\frac{\mathrm{d}^2 g}{\mathrm{d}t^2}(t_*)}} \cdot f(t_*) \cdot \mathrm{e}^{\lambda g(t_*)} \cdot \lambda^{-\frac{1}{2}}\big(1 + O(\lambda^{-\frac{1}{2}})\big) \quad \text{as } \lambda \to +\infty.$$

(A.1)-c) *If* g *is thrice continuously differentiable on a neighborhood of* t_*, *and* $t_* = a$ *with* $\frac{\mathrm{d}g}{\mathrm{d}t}(t_*) = 0$ *and* $\frac{\mathrm{d}^2 g}{\mathrm{d}t^2}(t_*) \neq 0$, *then*

$$F(\lambda) = \sqrt{\frac{\pi}{-2\frac{\mathrm{d}^2 g}{\mathrm{d}t^2}(t_*)}} \cdot f(t_*) \cdot \mathrm{e}^{\lambda g(t_*)} \cdot \lambda^{-\frac{1}{2}}\big(1 + O(\lambda^{-\frac{1}{2}})\big) \quad \text{as } \lambda \to +\infty.$$

Appendix B. The Gamma function and Stirling's formula

Recall that the Gamma function is[22]

$$\Gamma(\lambda) := \int_0^{+\infty} s^{\lambda - 1} \mathrm{e}^{-s} \, \mathrm{d}s \quad \text{for } \lambda > 0. \tag{B.1}$$

It is representable as a Laplace integral (A.2) in a natural way:

$$\Gamma(\lambda + 1) = \int_0^{+\infty} \mathrm{e}^{-s} \mathrm{e}^{\lambda \ln s} \, \mathrm{d}s \quad \text{for } \lambda > 0,$$

and introducing the new variable $s = \lambda t$, we arrive at

$$\Gamma(\lambda + 1) = \int_0^{+\infty} \mathrm{e}^{-\lambda t} \cdot \mathrm{e}^{\lambda \ln \lambda + \lambda \ln t} \cdot \lambda \, \mathrm{d}t = \lambda^{\lambda + 1} \int_0^{+\infty} \mathrm{e}^{\lambda(\ln t - t)} \, \mathrm{d}t.$$

Defining $]0, +\infty[\ni t \mapsto g(t) := \ln t - t \in \mathbb{R}$, we find that g has a unique maximizer $t_* = 1$ on the interval $]0, +\infty[$, and $\frac{\mathrm{d}^2 g}{\mathrm{d}t^2}(1) = -1$. Then Theorem (A.1)-b) shows that

$$\Gamma(\lambda + 1) = \sqrt{2\pi\lambda} \cdot \left(\frac{\lambda}{\mathrm{e}}\right)^{\lambda} \cdot \big(1 + O(\lambda^{-\frac{1}{2}})\big) \quad \text{as } \lambda \to +\infty.$$

[22] See https://mathworld.wolfram.com/GammaFunction.html for details.

Since $\Gamma(n+1) = n!$ for $n \in \mathbb{N}^*$, we get Stirling's formula[23]

$$n! = \sqrt{2\pi n} \cdot \left(\frac{n}{\mathrm{e}}\right)^n \cdot \left(1 + O(n^{-\frac{1}{2}})\right) \quad \text{as } n \longrightarrow +\infty. \tag{B.2}$$

Appendix C. The Hamming metric: sundry facts

The Hamming metric ρ_0 encountered in (3.1) gives rise to metric balls and spheres of interesting shapes in a natural way, and in this appendix we take a quick look at a few of them. Let $d \in \mathbb{N}^*$ and observe that the definition of the Hamming distance between two points has the following natural analog on $\mathbb{R}^d$: for $x, x' \in \mathbb{R}^d$,

$$\rho_0(x, x') = \left|\{n = 1, \ldots, d \mid x_n \neq x'_n\}\right|,$$

i.e., it is the number of disagreements in the entries of x and x'. Naturally, ρ_0 takes values in $\{0, 1, \ldots, d\}$. The closed Hamming ball of radius $r \geqslant 0$ centered at $y \in \mathbb{R}^d$ is the set

$$\mathsf{B}^d_0[y, r] := \left\{x \in \mathbb{R}^d \,\middle|\, \rho_0(x, y) \leqslant r\right\}$$

and the Hamming sphere of radius $r' \in \mathbb{N}$ centered at $y \in \mathbb{R}^d$ is the set

$$\mathbb{S}^d_0[y, r'] := \left\{x \in \mathbb{R}^d \,\middle|\, \rho_0(x, y) = r'\right\}.$$

It does not make sense to talk about the Hamming sphere of radius $r' > d$ in $\mathbb{R}^d$, although the closed balls do not suffer from this issue with the definition. Note that $\mathsf{B}^d_0[y, r] = y + \mathsf{B}^d_0[0, r]$ and $\mathbb{S}^d_0[y, r'] = y + \mathbb{S}^d_0[0, r']$.

Let us look at some examples of the Hamming ball and sphere.

- If $d = 1$, then
  - $\mathsf{B}^1_0[0, 0] = \{0\}$, $\mathsf{B}^1_0[0, r] = \mathbb{R}$ for all $r \geqslant 1$;
  - $\mathbb{S}^1_0[0, 0] = \{0\}$, $\mathbb{S}^1_0[0, 1] = \mathbb{R}^1 \smallsetminus \{0\}$.
- If $d = 2$, then
  - $\mathsf{B}^2_0[0, 0] = \{0\}$, for every $r \in [1, 2[$ we have $\mathsf{B}^2_0[0, r] = \{(x_1, x_2) \in \mathbb{R}^2 | x_1 = 0 \text{ or } x_2 = 0\}$, i.e., the union of the two axes, and $\mathsf{B}^2_0[0, r] = \mathbb{R}^2$ for each $r \geqslant 2$;
  - $\mathbb{S}^2_0[0, 0] = \{0\}$, $\mathbb{S}^2_0[0, 1] = \mathsf{B}^2_0[0, 1] \smallsetminus \{0\}$, $\mathbb{S}^2_0[0, 2] = \mathbb{R}^2 \smallsetminus \mathsf{B}^2_0[0, 1]$.
- If $d = 3$, then
  - $\mathsf{B}^3_0[0, 0] = \{0\}$, for every $r \in [1, 2[$ the set $\mathsf{B}^3_0[0, r]$ is the union of the three axes, for each $r \in [2, 3[$ the set $\mathsf{B}^3_0[0, r]$ is the *disjoint union* of the pairs x_1-x_2, x_1-x_3, and x_2-x_3 of axes, and $\mathsf{B}^3_0[0, r] = \mathbb{R}^3$ for every $r \geqslant 3$;
  - $\mathbb{S}^3_0[0, 1] = \mathsf{B}^3_0[0, 1] \smallsetminus \{0\}$, $\mathbb{S}^3_0[0, 2]$ is the *disjoint union* of the x_1-x_2, x_1-x_3, and x_2-x_3 planes with the corresponding pairs of axes removed from them, $\mathbb{S}^3_0[0, 3]$ is the set $\mathbb{R}^3$ with the unions of the three x_1-x_2, x_1-x_3, and x_2-x_3 planes removed.

Appendix D. Nonsmooth Lagrange multiplier rule

Let $d \in \mathbb{N}^*$. We recall two definitions central to nonsmooth calculus. Recall first that a mapping $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is Lipschitz continuous of rank L near a given point $x \in \mathbb{R}^d$ if for some $\varepsilon > 0$ we have

$$\left|f(x') - f(x'')\right| \leqslant L\|x' - x''\| \quad \text{for all } x', x'' \in \mathsf{B}^d(x, \varepsilon).$$

[23] See https://mathworld.wolfram.com/StirlingsApproximation.html for details.

The **generalized derivative** of a Lipschitz continuous function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ in the direction $v \in \mathbb{R}^d$ is

$$f^\circ(x; v) := \limsup_{\substack{\mathbb{R}^d \ni y \to x \\ t \downarrow 0}} \frac{f(y + tv) - f(y)}{t}.$$

The (Clarke) **generalized gradient** $(\partial_C f)(x)$ of f at $x \in \mathbb{R}^d$ is the unique compact convex subset of the dual $(\mathbb{R}^d)^\star$ of $\mathbb{R}^d$ (and canonically identified with $\mathbb{R}^d$ in view of the Riesz representation theorem) whose *support function* is the generalized derivative $f^\circ(x; \cdot)$. Accordingly,

$$\xi \in (\partial_C f)(x) \quad \Leftrightarrow \quad f^\circ(x; v) \geqslant \langle \xi, v \rangle \ \text{ for all } v \in \mathbb{R}^d,$$

$$f^\circ(x; v) = \max_{\xi \in (\partial_C f)(x)} \langle \xi, v \rangle \quad \text{for all } v \in \mathbb{R}^d.$$

Naturally, the generalized gradient is equal to the ordinary gradient at points of differentiability. We denote the convex hull of a non-empty set $S \subset \mathbb{R}^d$ by $\mathrm{co}(S)$.

The following theorem is essential to the computation of (Clarke) generalized gradients in finite dimensions:

(D.1). Theorem ([Cla13, Theorem 10.27])**.** *Let $x \in \mathbb{R}^d$ and let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a function Lipschitz continuous near x. Let E be any subset of zero measure in $\mathbb{R}^d$, and let E_f denote the set of points at which f is non-differentiable. Then*

$$(\partial_C f)(x) = \mathrm{co}\Big\{ \lim_{n \to +\infty} \nabla f(x_n) \ \Big| \ x_n \notin E \cup E_f \text{ and } x_n \xrightarrow[n \to +\infty]{} x \Big\}.$$

We need the following basic nonsmooth Lagrange multiplier rule for equality constrained optimization problems; it appears as [Cla13, Theorem 10.47] in the context of a minimization problem.[24]

(D.2). Theorem. *Let $d, d', d'' \in \mathbb{N}^*$. Consider the optimization problem*

$$\begin{aligned} &\underset{x}{\text{maximize}} \quad f(x) \\ &\text{subject to} \quad \begin{cases} g(x) = 0, \\ h(x) \leqslant 0, \\ x \in \mathbb{R}^d, \end{cases} \end{aligned}$$

where $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, $g : \mathbb{R}^d \longrightarrow \mathbb{R}^{d'}$, and $h : \mathbb{R}^d \longrightarrow \mathbb{R}^{d''}$ are Lipschitz continuous maps. Suppose that $x_\star$ solves the preceding optimization problem. Then there exists a triplet $(\eta, \lambda', \lambda'') \in \{0, 1\} \times \mathbb{R}^{d'} \times \mathbb{R}^{d''}$ satisfying the nontriviality condition $(\eta, \lambda', \lambda'') \neq (0, 0, 0)$, *the* positivity *and* complementary slackness conditions

$$\lambda'' \geqslant 0, \quad \langle \lambda'', h(x_\star) \rangle = 0,$$

and the stationarity condition

$$0 \in \partial_C\big(\eta \cdot f + \langle \lambda', g \rangle + \langle \lambda'', h \rangle\big)(x_\star).$$

The number $\eta \in \{0, 1\}$ is the *abnormal multiplier*; it attains the value 0 when the constraints of the optimization problem in (D.2) are so tight that the cost does not play any role in the determination of the optimizers.

[24] Since this result deals with first-order necessary conditions, it does not matter whether a maximization is involved or a minimization.

References

[AS64] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, vol. 55, For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[BG05] R. F. Bass and K. Gröchenig, *Random sampling of multivariate trigonometric polynomials*, SIAM Journal on Mathematical Analysis **36** (2004/05), no. 3, 773–795, doi: https://doi.org/10.1137/S0036141003432316.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities*, Oxford University Press, Oxford, 2013, A nonasymptotic theory of independence, With a foreword by Michel Ledoux. doi: https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

[BTEN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2009, doi: https://doi.org/10.1515/9781400831050.

[Cla13] F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics, vol. 264, Springer, London, 2013, doi: https://doi.org/10.1007/978-1-4471-4820-3.

[DACC22] S. Das, A. Aravind, A. Cherukuri, and D. Chatterjee, *Near-optimal solutions of convex semi-infintie programs via targeted sampling*, Annals of Operations Research (2022), doi: https://doi.org/10.1007/s10479-022-04810-4.

[HCL13] P. Hokayem, D. Chatterjee, and J. Lygeros, *Chance-constrained LQG with bounded control policies*, Proceedings of the 52nd IEEE Conference on Decision & Control, 2013, doi: https://doi.org/10.1109/CDC.2013.6760251, pp. 2471–2476.

[Lan97] S. Lang, *Undergraduate Analysis*, 2nd ed., Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1997, doi: https://doi.org/10.1007/978-1-4757-2698-5.

[Led01] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, vol. 89, American Mathematical Society, Providence, RI, 2001.

[MCB20] P. K. Mishal Assif, D. Chatterjee, and R. Banavar, *Scenario approach for minmax optimization with emphasis on the nonconvex case: positive results and caveats*, SIAM Journal on Optimization **30** (2020), no. 2, 1119–1143, doi: https://doi.org/10.1137/19M1271026.

[Ram18] F. A. Ramponi, *Consistency of the scenario approach*, SIAM Journal on Optimization **28** (2018), no. 1, 135–162, doi: https://doi.org/10.1137/16M109819X.

[Tal95] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Institut des Hautes Études Scientifiques. Publications Mathématiques (1995), no. 81, 73–205, doi: http://www.numdam.org/item?id=PMIHES_1995__81__73_0.

[Ver18] R. Vershynin, *High-Dimensional Probability*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 47, Cambridge University Press, Cambridge, 2018, doi: https://doi.org/10.1017/9781108231596; author's draft copy available at https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

[Wai19] M. J. Wainwright, *High-Dimensional Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 48, Cambridge University Press, Cambridge, 2019, doi: https://doi.org/10.1017/9781108627771.

[Zor16] V. A. Zorich, *Mathematical Analysis: Vol. II*, 2nd ed., Universitext, Springer, Heidelberg, 2016, doi: https://doi.org/10.1007/978-3-662-48993-2.

Index