PSTAT 262MC. Monte Carlo Methods

Lecture 2. *Concentration of measure, Stein's paradox, the isoperimatric problem, and the Johnstone-Lindenstrauss lemma.*

Alex Shkolnik
shkolnik@ucsb.edu

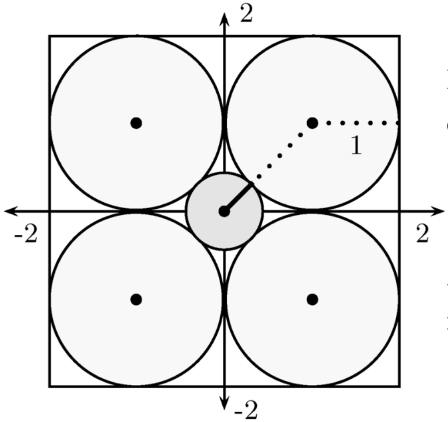
January 17, 2025.

Department of Statistics & Applied Probability
University of California, Santa Barbara

Warm up

Problem on the hypercube $[-2, 2]^d$ (taken from Steele (2004)).

52 *On Geometry and Sums of Squares*



In $\mathbb{R}^2$, one places a unit circle in each quadrant of the square $[-2, 2]^2$.

A non-overlapping circle of maximal radius is then centered at the origin.

Problem 4.1 (Thinking Outside the Box)

Is the central sphere $\mathcal{S}(d)$ contained in the cube $[-2, 2]^d$ for all $d \geq 2$?

Compute the radius of the central sphere for any $d \geq 2$.

The radius of the inner sphere is compute via,

$$\frac{\sqrt{d\,2^2}}{2} - 1 = \sqrt{d} - 1$$

so that the sphere exists the box for dimensions $d > 9$.

The fraction of the volume the central sphere occupies in the cube $[-2, 2]^d$ tends to zero exponentially in d (Steele 2004).

High dimensional space oddity (`https://arxiv.org/abs/2409.13046`).

- *What is the fraction of light emanating from the origin that leaves the cube (none for $d = 2$)?*
- *"We could not answer this question using geometry. ... Our solution is probabilistic in nature."*

Collection of lectures on convex geometry (Ball 1997).

An Elementary Introduction to Modern Convex Geometry

KEITH BALL

Contents

Notation.

d denotes the dimension.

For $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$ and $v = (v_1, \ldots, v_m) \in \mathbb{R}^m$ we let $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_j$ denote the standard inner product.

$|x| = \sqrt{\langle x, x \rangle}$, the Eucledian length.

Geometric LLN and Stein's paradox

INADMISSIBILITY OF THE USUAL ESTIMATOR FOR THE MEAN OF A MULTIVARIATE NORMAL DISTRIBUTION

CHARLES STEIN
STANFORD UNIVERSITY

1. Introduction

If one observes the real random variables $X_1, \cdots, X_n$ independently normally distributed with unknown means $\xi_1, \cdots, \xi_n$ and variance 1, it is customary to estimate ξ_i

– Stein (1955),

Let $X = \xi + \epsilon$ be a noisy observation of vector $\xi \in \mathbb{R}^d$.

- *This is the classical setting of observing $X \sim N(\xi, I_d)$ to estimate the population mean ξ (in the above image $n = d$).*
- *X is the maximum likelyhood estimator of the unknown ξ.*
- *Stein's paradox concerns the existence of $X' \in \mathbb{R}^d$ such that*

$$E\,|X' - \xi| < E\,|X - \xi|$$

This is possible for $d \geq 3$ (dimension n is above image)
(https://en.wikipedia.org/wiki/Stein'sexample)

Example from Stigler's 1988 Neyman Memorial Lecture.

Let X be the price of apples in Washington.

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i \in \mathbb{R}$$

Let Y be the price of oranges in Florida.

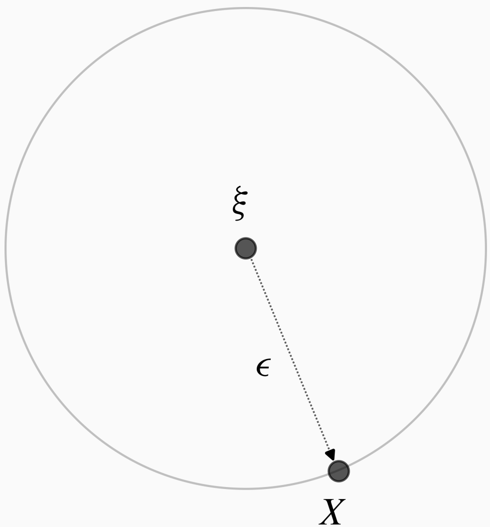
$$\begin{pmatrix} X \\ Y \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \in \mathbb{R}^2$$

Finally, let Z be the price of wine in France.

$$Q = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \in \mathbb{R}^3$$

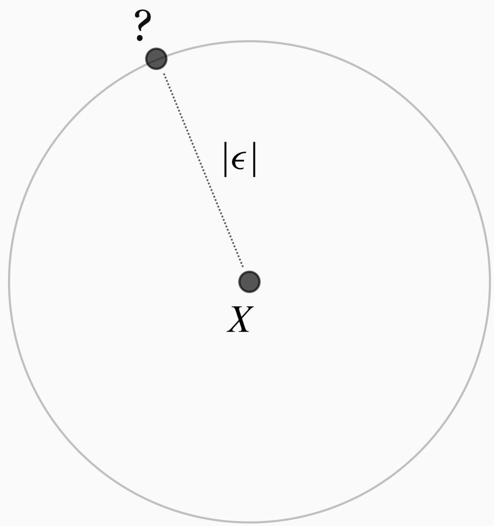
There is an estimator better than Q (in mean-squared error).

We begin with $X = \xi + \epsilon$ but instead of assuming that the noise $\epsilon \sim N(0, I_d)$ we use geometrical arguments.



Pictured are $X, \xi \in \mathbb{R}^d$ and $\epsilon = X - \xi$ (the noise vector).

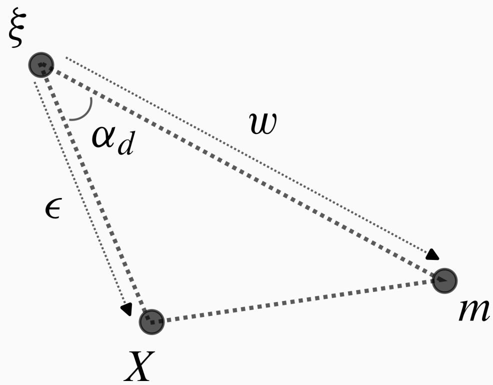
The $\xi \in \mathbb{R}^d$ is unknown (we only see the observation X).



To make things easier assume the length $|\epsilon|$ is known. But even when given $|\epsilon|$ it is not clear how to improve the estimate X.

- *The unknown ξ could be on any point on the circle pictured.*

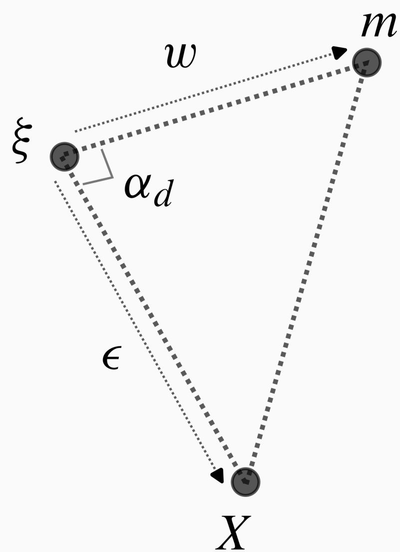
Add any point $m \in \mathbb{R}^d$. We now have the vectors $\epsilon = \eta - \beta$ and $w = m - \xi$ with α_d denoting the angle between them.



The vector $m \in \mathbb{R}^p$ is treated as nonrandom (or independent).

– *For a low dimension d we cannot say more, but ...*

As the dimension d becomes large, the angle α_d tends to 90° with high probability (given mild assumptions on the noise).



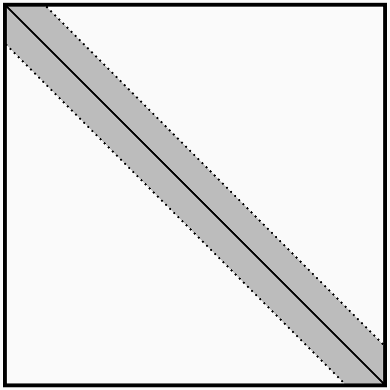
Geometric LLN.

$$\langle X - \xi, m - \xi \rangle / d = \langle \epsilon, w \rangle / d = \frac{1}{d} \sum_{i=1}^{d} w_i \epsilon_i \to 0$$

As $d \uparrow \infty$ the angle α_p between ϵ and $w = m - \xi$ tends to 90°.

- *The convergence above may be viewed as a law of large number for non-identical random variables.*
  (e.g., `https://arxiv.org/pdf/1701.02234`*)*
- *Noise (pure randomness) is orthogonal to (or independent of) any high dimensional vector not corrupted by it.*
- *Very similar to the concentration of measure result due to Borel 1914 (mass of hypercube concentrates on equator).*

Borel 1914 – Concentration of mass of hypercube on its equator.

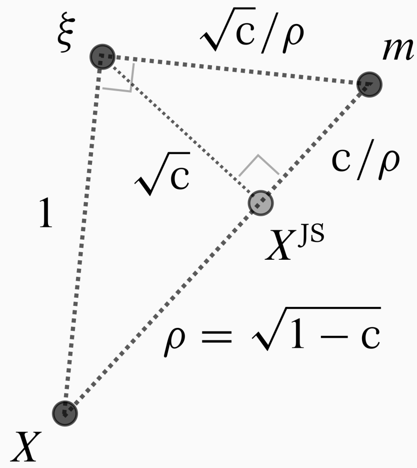


A point u is on the equator of the hypercube $[0, 1]^d$ whenever,

$$\langle u - v, v \rangle = 0$$

where $v = (\frac{1}{2}, \ldots, \frac{1}{2})$ is the center of the hypercube.

Define X^{JS} as the closest point to ξ on the line segment between X and m (this is the James-Stein estimator).



- *See James & Stein (1961) for* $m = 0_d$ *and finite* $d > 2$.
- *We took* $X^{\text{JS}} = X + c(X - m)$ *for any* $m \in \mathbb{R}^d$.
- *We will be able to estimate* $c \in [0, 1]$ *in the limit* $d \uparrow \infty$.

Once we reach dimension ∞ the geometry above may be computed via the Pythagorean theorem (i.e. $\alpha_\infty = 90°$).

- *Each length pictured is scaled by* $|\epsilon|$ *(the only left unknown).*

Letting ν be a d-consistent estimate[1] of $|\epsilon| = |X - \xi|$, let

$$X^{\mathrm{JS}} = m + \mathrm{c}(X - m), \qquad \left(\mathrm{c} = 1 - \frac{\nu^2}{|X - m|^2}\right)$$

Theorem. If both $|\xi|^2/p$ and $|\epsilon|^2/p$ are bounded in $(0, \infty)$ as the dimension $d \uparrow \infty$ and the Geometric LLN holds,

$$|X^{\mathrm{JS}} - \xi| \sim \sqrt{\mathrm{c}}\,|X - \xi|.$$

with $\sqrt{\mathrm{c}}$ is eventually in the interval $(0, 1)$.

- *We write* $a_d \sim b_d$ *provided* $\lim_{d\uparrow\infty} a_d/b_d = 1$.
- *For large dimension* X^{JS} *is clearly preferred to* X *and this works for any choice of* m *(hence Stein's paradox).*

[1] For this we require two observations of ξ and putting them in a $d \times 2$ matrix M it may be shown that ν^2 defined as the second eigenvalue of $M^\top M$ satisfies $\nu \sim |\epsilon|$ as $d \uparrow \infty$.

Proof sketch.

- *Start with* $X(\mathrm{c}) = m + \mathrm{c}(X - m)$ *and minimize the error function* $\mathrm{c} \mapsto |X(\mathrm{c}) - \xi|^2$ *to obtain the minimizer,*

$$\mathrm{c}^* = \frac{\langle X - m, \xi - m\rangle}{|X - m|^2} = 1 - \frac{\langle X - m, \epsilon\rangle}{|X - m|^2}.$$

- *Show that* $|\mathrm{c} - \mathrm{c}^*|$ *with* c *as stated in the theorem vanishes as* d *tends to infinity assuming the geometric* LLN.
- *Deduce that* $\langle \xi - X(\mathrm{c}^*), m - X(\mathrm{c}^*)\rangle = 0$ *which establishes the right angle at* X^{JS} *in the limit as shown in the figure.*
- *Using that* $\nu \sim |\epsilon|$ *compute all the lenghts between the points* $(X, \xi, m, X^{\mathrm{JS}})$ *in the infinite dimensional limit.*

An application. We saw the goal of Monte Carlo integration is to estimate intergrals such as $\int_{\mathcal{D}} f(x)\,\mathrm{d}x$ for some domain $\mathcal{D}$.

Let $\theta \in \mathbb{R}^d$ be some high dimensional set of parameters, and

$$F(\theta) = \int_{\mathcal{D}} f(x, \theta)\,\mathrm{d}x\,.$$

In many application we wish to estimate the gradient of F,

$$\nabla_\theta F(\theta) = \left(\frac{\mathrm{d}F(\theta)}{\mathrm{d}\theta_1}, \ldots, \frac{\mathrm{d}F(\theta)}{\mathrm{d}\theta_d}\right) \in \mathbb{R}^d$$

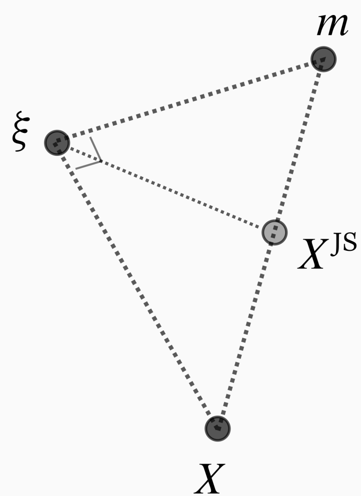
Mohamed et al. (2020) apply Monte Carlo integration (given $\nabla_\theta f(x, \theta)$ may be computed) to estimate $\nabla_\theta F(\theta)$.

- *The discuss many applications to inference, reinforcement learning, queuing theory, sensitivity analysis, etc.*

Taking $X = \nabla_\theta F(\theta) + \epsilon$ we could use X^{JS} to possibly improve upon Monte Carlo estimates when d is large.

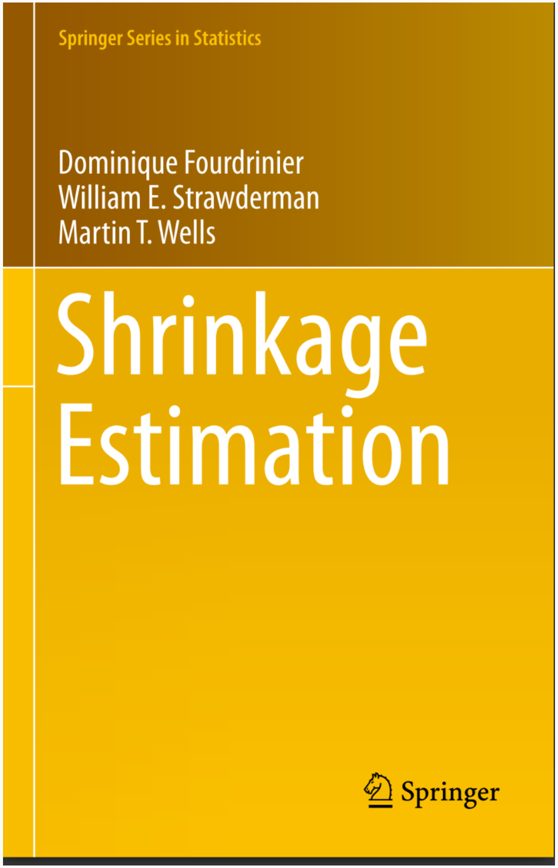
Related is also the area of "derivative free optimization" which does not assume $\nabla_\theta f(x, \theta)$ is known (see Scheinberg (2022)).

The estimator X is called inadmissible because, for large enough d, the James-Stein estimator X^{JS} improves upon it.



The vector m is called the “shrinkage target”.

The James-Stein estimator has inspired the area of shrinkage estimation in statistics (Fourdrinier et al. 2018).



Isoperimetric problem

Let $(\mathbb{X}, \mathrm{d})$ be a metric space.

- $\mathrm{d}(x, y)$ *is the distance between* $x, y \in \mathbb{X}$.
- $\mathrm{d}(x, A) = \inf_{y \in A} \mathrm{d}(x, y)$.

Consider the ϵ-neighborhood of A, i.e. for $\epsilon > 0$,

$$A_\epsilon = \{x \in \mathbb{X} : \mathrm{d}(x, A) \leq \epsilon\}$$

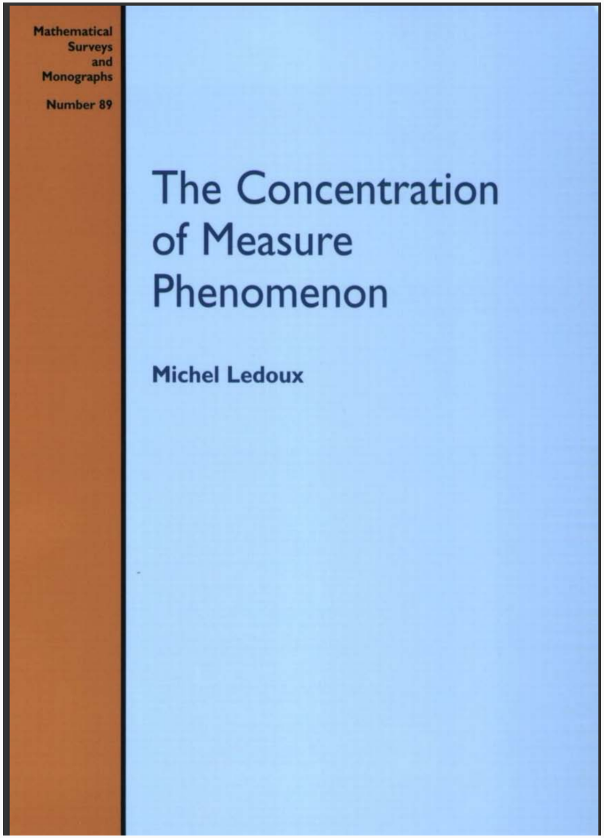
For fixed $\epsilon > 0$, we define the function ϕ_v as follows.

$$\phi_v(\epsilon) = \inf_{A \subseteq \mathbb{X}} \{\mu(A_\epsilon) : \mu(A) \geq v\}$$

Isoperimetric problem. *"Of all shapes A with volume V, what is the one that minimizes the volume of the ϵ-neighborhood of A."*

- *Few explicit solutions are known (e.g., most notably the constant curvature spaces; see Section 2.1 of Ledoux (2001)).*
- *Lower bounds for the function ϕ_v are good enough (see concentration of measure below).*

Geometric/functional analysis perspective (with mix of probability where it is helpful) – Ledoux (2001).



The isoperimetric problem on $\mathbb{R}^d$

Let μ measure volume in $\mathbb{R}^d$ uniformly, e.g., the μ-measure of a box $[a_1, b_1] \times \cdots \times [a_d, b_d] \subseteq \mathbb{R}^d$ is given by,

$$\mu([a_1, b_1] \times \cdots \times [a_d, b_d]) = \prod_{i=1}^{d} (b_i - a_i).$$

Let d be the Euclidean metric (e.g., $\mathrm{d}(x, y) = |x - y|$).

Consider all $A \subseteq \mathbb{R}^d$ of fixed volume $v = \mu(A)$.

The body A that minimizes $\mu(A_\epsilon)$ is the Euclidean ball B, i.e.,

$$B = B(x, r) = \{v \in \mathbb{R}^d \,|\, \mathrm{d}(x, v) \leq r\}$$

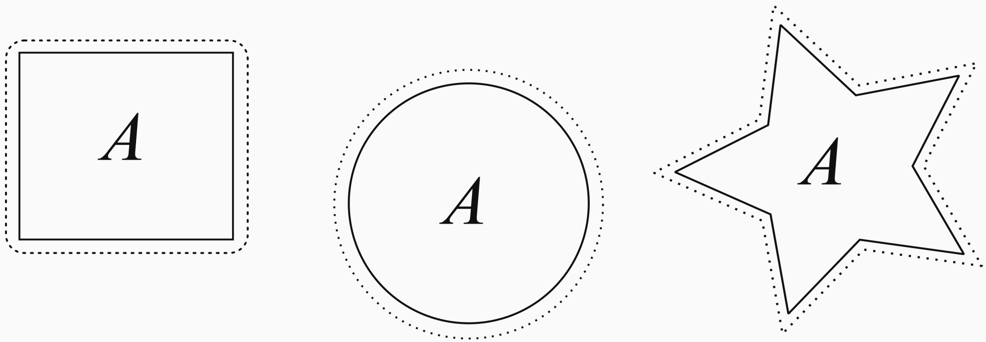
where the radius of the ball r is set to achieve $\mu(B) = v$. The center x of B does not change the volume v.

Then, $\phi_v(\epsilon) = \mu(B_\epsilon)$ and $B_\epsilon = B(x, r + \epsilon)$ is also a ball just like $B = B(x, r)$ but with the larger radius $r + \epsilon$.

This is a corollary of Brunn-Minkowski inequality (see Theorem 5.2 and Lecture 8 in Ball (1997)).

Illustration of the isoperimetric problem in $\mathbb{R}^d$.

The box, the ball and the star A all have volume $v = \mu(A)$.
Their ϵ-neighborhoods are pictured with a dashed boundary.



Out of the three A (and every set of volume v) the ball $A = B$ (middle) minimizes the volume of its ϵ-neighborhood $\mu(A_\epsilon)$

The ball B also minimizes the surface area of any body $A \subseteq \mathbb{R}^d$ provided that $\mu(A) = \mu(B)$ (Theorem 5.3 in Ball (1997).

The isoperimetric inequality on $\mathbb{R}^d$

Let $B = \{y \in \mathbb{R}^d : \mathrm{d}(x, y) \leq r\}$ be a Euclidean ball of radius r centered at x. Let $\mu(B) = v$ be its volume (unchanged by x).

We have the following **isoperimetric inequality** on $\mathbb{R}^d$. For all measurable $A \subseteq \mathbb{R}^d$ of volume $v = \mu(A) = \mu(B)$ and $\epsilon > 0$,

$$\mu(A_\epsilon) \geq \mu(B_\epsilon) = \phi_v(\epsilon).$$

This says that "growing" a body A to its ϵ-neighborhood takes up the least volume for $A = B$, the Euclidean ball.

The advantage is that $\mu(B_\epsilon)$ is much easier to calculate than would $\mu(A_\epsilon)$ be for many useful A encountered in applications.

We can apply this result to see where the uniform measure μ concentrates inside the hypercube $[0, 1]^d \subseteq \mathbb{R}^d$.

Let $A \subseteq [0,1]^d$ with $\mu(A) \geq 1/2$ (i.e., A is a body in the hypercube occupying at least half its volume; $\mu([0,1]^d) = 1$).

Let $v(r)$ denote the volume $\mu(B)$ of a ball B of radius $r > 0$. Fix a ball B to have volume $v(r_d) = 1/2$. It's radius has $r_d \sim \sqrt{d}$ (i.e., $\frac{r_d}{\sqrt{d}} \to 1$). [2] By the isoperimetric inequality,

$$\mu(A_\epsilon) \geq \mu(B_\epsilon) = v(r_d + \epsilon)$$

Some further calculations yield $v(r_d + \epsilon) \geq 1 - e^{-\pi\epsilon^2}$.

- *For any A in $[0,1]^d$ of more than half the volume of $[0,1]^d$, the ϵ-neighborhood volume grows exponentially in ϵ^2.*

The largest distance between any two corner of the hypercube is $\sqrt{d}$, so provided that $A_{\delta\sqrt{d}} \subseteq [0,1]^d$ for some $\delta > 0$,

$$\mu(A^c_{\delta\sqrt{d}}) \leq e^{-\pi\delta^2 d}$$

which tends to zero exponentially fast in the dimension.

[2] see (www.wikipedia.org/wiki/Volume_of_an_n-ball)

Another classical domain is $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : |x| = 1\}$, the surface of a sphere of radius 1 in d-dimensions.

Let μ be the uniform area measure on $\mathbb{S}^{d-1}$ with $\mu(\mathbb{S}^{d-1}) = 1$.

- *This means the $\mu(A) = \mu(RA)$ where RA is any rotation of $A \subseteq \mathbb{S}^{d-1}$ (www.wikipedia.org/wiki/Rotation_matrix).*
- *Statistically, a vector $u \in \mathbb{S}^{d-1}$ from the distribution μ will be "equally likely to be at any point in $\mathbb{R}^{d-1}$".[3]*

We can let d be the Euclidean metric or geodesic distance.

- $\mathrm{d}(x, y) = |x - y|$ *for $x \in \mathbb{S}^{d-1}$ for Euclidean metric.*
- $\mathrm{d}(x, y) = \arccos\langle x, y\rangle$ *for $x \in \mathbb{S}^{d-1}$ for geodesic metric (this is just the length of the arc between x and y on $\mathbb{S}^{d-1}$).*

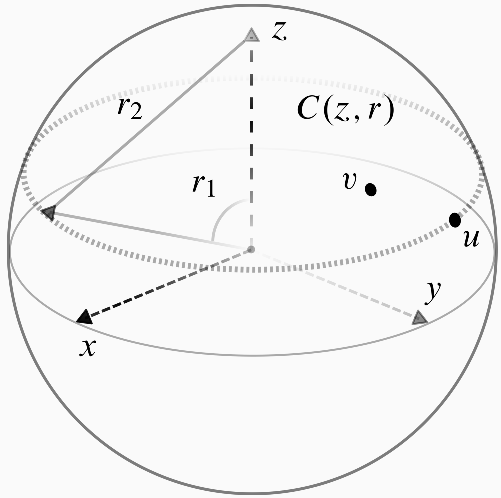
Either metric may be used to define spherical cap,

$$C = C(z, r) = \{v \in \mathbb{S}^{d-1} : \mathrm{d}(z, v) \leq r\},$$

i.e., all points within distance/angle r of the cap center $z \in \mathbb{S}^{d-1}$.

[3] Of course this event has measure zero.

$C(z, r)$ is in fact a ball on $\mathbb{S}^{d-1}$ centered at z and with radius r.



- *The cap $C(z, r)$ contains all area above the dashed line.*
- *The illustrated cap $C(z, r)$ uses angle r_1 with the geodesic metric and r_2 with the Euclidean metric (set to match area).*
- *$r_1 \neq r_2$ when describing the same cap with different metrics.*
- *Above $u \in \partial C(z, r)$, the boundary of the cap (at maximal angle r_1), and $v \in C(z, r)$ at some angle smaller than r_1.*

The isoperimetric problem on $\mathbb{S}^{d-1}$.

Let μ be the uniform surface area measure on $\mathbb{S}^{d-1}$.

Let d be the Euclidean metric or geodesic distance.

Consider all $A \subseteq \mathbb{R}^d$ of fixed area $a = \mu(A)$.

The body A that minimizes $\mu(A_\epsilon)$ is the spherical cap C, i.e.,

$$C = C(z, r) = \{v \in \mathbb{S}^{d-1} : \mathrm{d}(z, v) \geq r\}$$

where the radius r of the cap is set to achieve area $\mu(C) = a$. The center z of C does not change the area a.

Then, $\phi_v(\epsilon) = \mu(C_\epsilon)$ and $C_\epsilon = C(z, r + \epsilon)$ is also a cap just like $C = B(z, r)$ but with the larger radius $r + \epsilon$.

The proof is difficult to find/read (Chapter IV of Part 3 of Lévy & Pellegrino (1951) credited to Lévy and dating to 1919).

- *A self-contained proof is in Section 8 of Figiel et al. (1977).*
- *Gromov (1980) generalizes to Riemannian manifolds (exact answers known only for constant curvature).*

The isoperimetric inequality on $\mathbb{S}^{d-1}$

Let $C = \{v \in \mathbb{S}^{d-1} : \mathrm{d}(z, v) \leq r\}$ be a Euclidean ball of radius r centered at z. Let $\mu(C) = a$ be its area (unchanged by z).

We have the following **isoperimetric inequality** on $\mathbb{S}^{d-1}$. For all measurable $A \subseteq \mathbb{S}^{d-1}$ with $a = \mu(A) = \mu(C)$ and $\epsilon > 0$,

$$\mu(A_\epsilon) \geq \mu(C_\epsilon) = \phi_a(\epsilon).$$

This says that "growing" an area A to its ϵ-neighborhood takes up the least the area of $A = C$, the spherical cap.

The advantage is that $\mu(C_\epsilon)$ is much easier to calculate than would $\mu(A_\epsilon)$ be for many useful A encountered in applications.

For calculations of areas of spherical caps of $\mathbb{S}^{d-1}$ see Li (2010).

Bounds are derived in Lecture 2 of Ball (1997). These give the famous concentration on the equation result (next slide).

Consider any area A on $\mathbb{S}^{-1}$ with measure $\mu(A) \geq 1/2$ (i.e., A occupies at least half the sphere).

The spherical cap C with area $\mu(C) = 1/2$ extends to the "equator" of the sphere, i.e. $C = C(z, 90°)$ (geodesic metric).

For any A of area at least half of the sphere and $\epsilon > 0$,

$$\mu(A_\epsilon) \geq \phi_{1/2}(\epsilon) \geq 1 - e^{-(d-1)\epsilon^2/2}$$

where $0 < \epsilon < \pi = 180°$ the exponential bound is derived for the geodesic metric in Section 2.1 of Ledoux (2001).

Eucledian d has $1 - e^{-d\epsilon^2/2}$ (Lemma 2.2 of Ball (1997)).

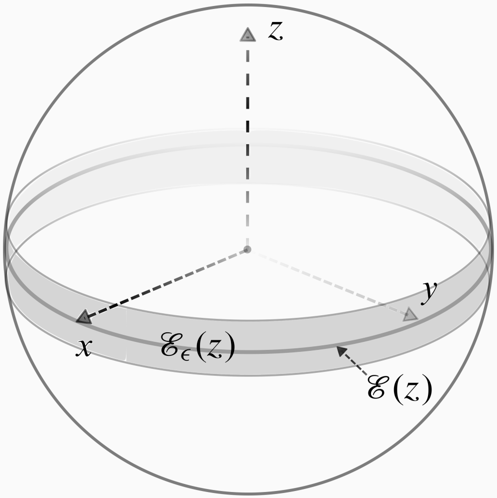
The above may be used to justify that the area of the sphere concentrates on the equator of the center z of C (next slide).

Since z is arbitrary, μ concentrates on every equator!

The equator $\mathscr{E}(z) = \{v \in \mathbb{S}^{d-1} : \langle v, z \rangle = 0\}$ with respect to $z \in \mathbb{S}^{d-1}$ is all points perpendicular to z, e.g., $x, y \in \mathscr{E}(z)$.

Note, $\mathscr{E}(z) = \partial C(z, 90°)$, the boundary of the half-sphere.

- *The upper cap* $C(z, 90° + \epsilon)$ *contains* $1 - e^{-(d-1)\epsilon^2/2}$ *fraction of the total area* $\mu(\mathbb{S}^{d-1}) = 1$.
- *The lower cap* $C(-z, 90° + \epsilon)$ *contains* $1 - e^{-(d-1)\epsilon^2/2}$.



- *It follows that* $\mathscr{E}_\epsilon(z) = C(z, 90° + \epsilon) \cap C(-z, 90° + \epsilon)$ *has area at least* $1 - 2e^{-(d-1)\epsilon^2/2}$ *(close to* 1 *for large* d *).*

Concentration of measure

Let $(\mathbb{X}, \mathrm{d})$ be a metric space and μ denote a Borel probability (i.e. open sets are with respect to d) measure with $\mu(\mathbb{X}) = 1$.

Theorem. If $f : \mathbb{X} \to \mathbb{R}$ is Lipschitz with constant L, then there is a constant $M \in \mathbb{R}$ such that for any $\epsilon > 0$,

$$\mu\Big(x \in \mathbb{X} : |f(x) - M| \geq \epsilon\Big) \leq 2(1 - \phi_{1/2}(\epsilon/L))$$

Remarks.

- *f Lipschitz means there exists a constant L such that*

$$|f(x) - f(y)| \leq L\mathrm{d}(x, y), \quad \forall x, y \in \mathbb{X}.$$

  (i.e., the function f is relatively "smooth")
- $M = \inf\{t \in \mathbb{R} : \mu(x \in \mathbb{X} : f(x) > t) \geq 1/2\}$ *(the median) but also holds for* $M = \int_{\mathbb{X}} f(x)\mathrm{d}\mu(x)$ *(mean).*
- *When $\phi_{1/2}(\epsilon/L)$ is close to 1, this states that "smooth functions in high dimensional spaces are nearly constant".*
- *In the context of Monte Carlo integration, the function f is easy to integrate if $\phi_{1/2}(\epsilon/L)$ is close to 1.*

Let $(\mathbb{X}, \mathrm{d})$ be a metric space and μ denote a Borel probability (i.e. open sets are with respect to d) measure with $\mu(\mathbb{X}) = 1$.

Theorem. If $f : \mathbb{X} \to \mathbb{R}$ is Lipschitz with constant L, then there is a constant $M \in \mathbb{R}$ such that for any $\epsilon > 0$,

$$\mu\Big(x \in \mathbb{X} : |f(x) - M| \geq \epsilon\Big) \leq 2(1 - \phi_{1/2}(\epsilon/L))$$

Proof outline.

- *Letting* $\{x \in \mathbb{X} : |f(x) - M| \geq \epsilon\} = U \cup V$.

$$U = \Big\{x \in \mathbb{X} : f(x) \leq M - \epsilon\Big\}$$
$$V = \Big\{x \in \mathbb{X} : f(x) \geq M + \epsilon\Big\}$$

- *We will let* $A = \{x \in \mathbb{X} : f(x) > M\}$ *and show that* $U \subseteq A_{\epsilon/L}$ *which has* $\mu(A) \geq 1/2$ *(by definition of* M*).*
- *Then* $\mu(U) \leq \mu(A^c_{\epsilon/L}) = 1 - \mu(A_{\epsilon/L}) \leq 1 - \phi_{1/2}(\epsilon/L)$.
- *Similarly,* $\mu(V) \leq \mu(A^c_{\epsilon/L}) \leq 1 - \phi_{1/2}(\epsilon/L)$ *but now with* $A = \{x \in \mathbb{X} : f(x) \leq M\}$ *which also has* $\mu(A) \geq 1/2$.

Proof.

- *Let* $A = \{z \in \mathbb{X} : f(z) > M\}$ *and note that* $\mu(A) \geq 1/2$ *by the definition of the median* M*. We prove that,*

$$\mu(U) = \mu\big(\{x \in \mathbb{X} : f(x) \leq M - \epsilon\}\big) \leq 1 - \mu(A_{\epsilon/L})$$

  where $A_{\epsilon/L}$ *is the* ϵ*-neighborhood of* A*.*

- *The isoperimetric inequality* $\mu(A_{\epsilon/L}) \geq \phi_{1/2}(\epsilon/L)$ *then completes the* U *part of the theorem (the* V *part is similar).*
- *This reduces the calculation of the measure of a set* $A_{\epsilon/L}$ *to an isoperimetric problem/inequality.*
- *Let* $x \in U = \{x \in \mathbb{X} : f(x) \geq M + \epsilon\}$ *to show* $x \in A^c_{\epsilon/L}$ *which implies* $\mu(U) \leq 1 - \mu(A_{\epsilon/L})$ *as required. Note,*

$$f(x) \leq M - \epsilon < f(y) - \epsilon$$

  for all $y \in A$ *(because* $f(y) > M$*)*

- *This violates the Lipschitz property of* f *unless out point* x *satisfies* $\mathrm{d}(x, y) > \epsilon/L$*. It follows that* $x \in A^c_{\epsilon/L}$*.*

Johnstone-Lindenstrauss lemma

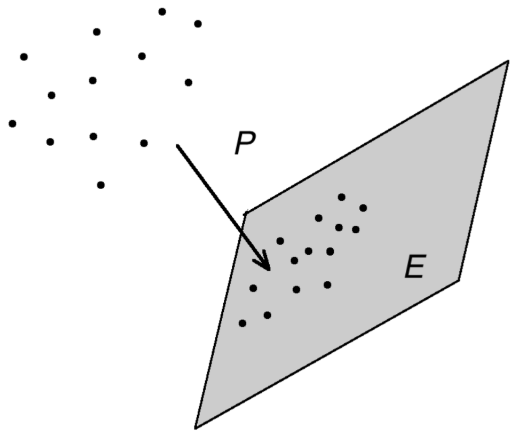


Figure 5.2 In Johnson-Lindenstrauss Lemma, the dimension of the data is reduced by projection onto a random low-dimensional subspace.

See Chapter 5 Section 5.3 in Vershynin (2018).

Project N points in $\mathbb{R}^d$ into an m-dimensional space ($m \ll d$).

- *Let $\mathcal{D}_N$ be a set of N points in $\mathbb{R}^d$.*
- *Let E be a (uniformly) random m-dim. subspace of R^d.*
- *Let P denote the orthogonal projection onto E.*
  e.g., for a $m \times d$ matrix X with i.i.d. entries $X_{ij} \sim N(0, 1)$,

$$P = X(X^\top X)^{-1} X^\top$$

- *Let μ be the probability on the space on which X is defined.*

Johnstone-Lindenstrass. There are absolute constants $C, c > 0$ such that for all $\epsilon > 0$ and $m \geq (C/\epsilon^2) \log N$,

$$\mu(A_\epsilon) \geq 1 - 2e^{-c\epsilon^2 m} \geq 1 - 2N^{-c \times C}$$

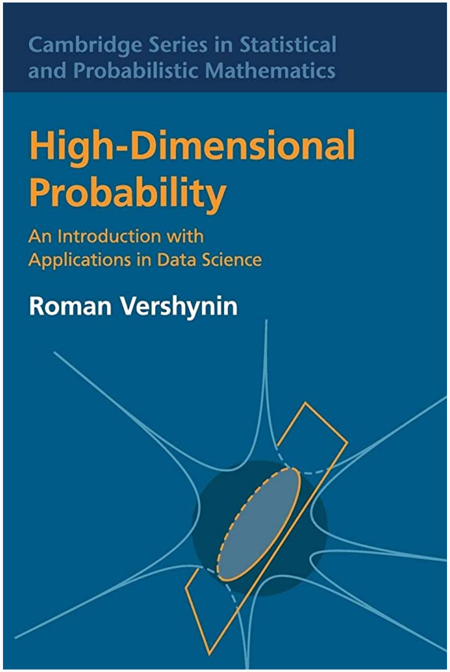
(i.e. with high probability for m sufficiently large) where

$$A_\epsilon = \left\{ (1+\epsilon) \leq \frac{|Px - Py|}{|x - y|} \sqrt{\frac{d}{m}} \leq (1+\epsilon),\ \forall x, y \in \mathcal{D}_N \right\}.$$

i.e., all distances are preserved upto a scaling by $\sqrt{d/m}$.

The proof of the Johnstone-Lindenstrauss lemma is based on Lévy's concentration of measure on the sphere.

The proof is given below Theorem 5.3.1 of Vershynin (2018).



Dvoretzky's theorem

Theorem (Dvoretzky). *A random projection of a convex body in* $\mathbb{R}^d$ *onto a subspace of dimension* $k \leq C\epsilon^2 \log d$ *is approximately a ball in* $\mathbb{R}^k$ *with probability at least* $1 - e^{-ck}$.

Proofs of statements similar to this one may be found in Ball (1997), Vershynin (2018) and Meckes (2019).

References

Ball, K. (1997), 'An elementary introduction to modern convex geometry', *Flavors of geometry* **31**(1-58), 26.

Figiel, T., Lindenstrauss, J. & Milman, V. (1977), 'The dimension of almost spherical sections of convex bodies', *Séminaire Maurey-Schwartz* pp. 1–13.

Fourdrinier, D., Strawderman, W. E. & Wells, M. T. (2018), *Shrinkage estimation*, Springer.

Gromov, M. (1980), 'Paul levy's isoperimetric inequality', *preprint IHES* **2**, 398.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', The Regents of the University of California.

Ledoux, M. (2001), *The concentration of measure phenomenon*, number 89, American Mathematical Soc.

Lévy, P. & Pellegrino, F. (1951), 'Problèmes concrets d'analyse fonctionnelle', *(No Title)* .

Li, S. (2010), 'Concise formulas for the area and volume of a hyperspherical cap', *Asian Journal of Mathematics & Statistics* 4(1), 66–70.

Meckes, E. S. (2019), *The random matrix theory of the classical compact groups*, Vol. 218, Cambridge University Press.

Mohamed, S., Rosca, M., Figurnov, M. & Mnih, A. (2020), 'Monte carlo gradient estimation in machine learning', *Journal of Machine Learning Research* **21**(132), 1–62.

Scheinberg, K. (2022), 'Finite difference gradient approximation: To randomize or not?', *INFORMS Journal on Computing* **34**(5), 2384–2388.

Steele, J. M. (2004), *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*, Cambridge University Press.

Stein, C. (1955), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', The Regents of the University of California.

Vershynin, R. (2018), *High-dimensional probability: An introduction with applications in data science*, Vol. 47, Cambridge university press.