Math 710: Measure Concentration

Alexander Barvinok

These are day-by-day lecture notes, not really proofread or edited and with no references, for the course that I taught in Winter 2005 term. The course also included presentations given by the participants. The notes for the presentations are not included here, but the topics are listed on pages 113–115.

Contents

Lecture 1. Wednesday, January 5

1. Introduction

Measure concentration is a fairly general phenomenon, which asserts that a "reasonable" function $f : X \longrightarrow \mathbb{R}$ defined on a "large" probability space X "almost always" takes values that are "very close" to the average value of f on X. Here is an introductory example, which allows us to remove the quotation marks, although we don't prove anything yet.

(1.1) Example: concentration on the sphere. Let $\mathbb{R}^n$ be the n-dimensional Euclidean space of all n-tuples $x = (\xi_1, \ldots , \xi_n)$ with the scalar product

$$\langle x, y\rangle = \sum_{i=1}^{n} \xi_i \eta_i \quad \text{for} \quad x = (\xi_1, \ldots , \xi_n) \quad \text{and} \quad y = (\eta_1, \ldots , \eta_n),$$

and the norm

$$\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^{n} \xi_i^2} \quad \text{for} \quad x = (\xi_1, \ldots , \xi_n).$$

Let

$$\mathbb{S}^{n-1} = \Big\{ x \in \mathbb{R}^n : \quad \|x\| = 1 \Big\}$$

be the unit sphere. We introduce the geodesic metric on the sphere

$$\operatorname{dist}(x, y) = \arccos \langle x, y\rangle,$$

the distance between x and y being the angle between x and y (check that this is indeed a metric) and the rotation invariant probability measure μ (the existence and uniqueness of such a measure is pretty intuitive, although not so easy to prove formally).

Let $f : \mathbb{S}^{n-1} \longrightarrow \mathbb{R}$ be a function which is 1-Lipschitz:

$$|f(x) - f(y)| \leq \operatorname{dist}(x, y) \quad \text{for all} \quad x, y \in \mathbb{S}^{n-1}.$$

Let m_f be the *median* of f, that is the number such that

$$\mu\Big\{x : f(x) \geq m_f\Big\} \geq \frac{1}{2} \quad \text{and} \quad \mu\Big\{x : f(x) \leq m_f\Big\} \geq \frac{1}{2}$$

(prove that for a Lipschitz function f, the median m_f exists and unique). Then, for any $\epsilon \geq 0$,

$$\mu\Big\{x \in \mathbb{S}^{n-1} : \ |f(x) - m_f| \leq \epsilon\Big\} \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2(n-2)/2}.$$

In particular, the value of $f(x)$ in a "typical" point $x \in \mathbb{S}^{n-1}$ deviates from the median only by about $1/\sqrt{n}$: that is, if $\epsilon_n \sqrt{n} \longrightarrow +\infty$, however slowly, the probability that $f(x)$ does not deviate from m_f by more than ϵ_n tends to 1.

PROBLEM.

Check the assertion for a linear function $f : \mathbb{S}^{n-1} \longrightarrow \mathbb{R}$, $f(x) = \langle a, x \rangle$ for some $a \in \mathbb{R}^n$.

Although the sphere looks like a nice and intuitive object, it is not so easy to work with. The Gaussian measure, although not so intuitive, is much nicer to work with analytically.

2. "Condensing" the Sphere from the Gaussian Measure

As we all know

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} \ dx = 1.$$

The probability measure on $\mathbb{R}^1$ with the density

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is called the *standard Gaussian measure* on the line $\mathbb{R}^1$.

We define the *standard Gaussian measure* γ_n on $\mathbb{R}^n$ as the probability measure with the density

$$\frac{1}{(2\pi)^{n/2}} e^{-\|x\|^2/2},$$

so

$$\gamma_n(A) = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\|x\|^2/2} \ dx$$

for a measurable set A. We have

$$\gamma_n(\mathbb{R}^n) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-\|x\|^2/2} \ dx = 1.$$

Our immediate goal is to prove that for an overwhelming majority of vectors $x \in \mathbb{R}^n$ with respect to the Gaussian measure γ_n, we have $\|x\| \sim \sqrt{n}$. In other words, from the point of view of the standard Gaussian measure, the space $\mathbb{R}^n$ "looks like" the sphere of the radius $\sqrt{n}$. This will be our first rigorous concentration result and the proof will demonstrate some important technique.

(2.1) How to prove inequalities: the Laplace transform method. Let X be a space with a probability measure μ and let $f : X \longrightarrow \mathbb{R}$ be a function.

Suppose we want to estimate

$$\mu\Big\{x : f(x) \geq a\Big\} \quad \text{for some} \quad a \in \mathbb{R}.$$

Here is an amazingly efficient way to do it. Let us choose a $\lambda > 0$. Then

$$f(x) \geq a \Longrightarrow \lambda f \geq \lambda a \Longrightarrow e^{\lambda f} \geq e^{\lambda a}.$$

Now, $e^{\lambda f}$ is a positive function on X and $e^{\lambda a}$ is a positive number, hence

$$\int_X e^{\lambda f} \ d\mu \geq \int_{x: f(x) \geq a} e^{\lambda f} \ d\mu \geq \mu\Big\{x \in X: \ f(x) \geq a\Big\} e^{\lambda a},$$

from which

$$\mu\Big\{x : f(x) \geq a\Big\} \leq e^{-\lambda a} \int_X e^{\lambda f} \ d\mu.$$

Often, we can compute exactly or estimate the value of the integral $\int_X e^{\lambda f} \ d\mu$. At that point, we "tune up" λ so that we get the strongest inequality possible.

Suppose we want to estimate

$$\mu\Big\{x : f(x) \leq a\Big\} \quad \text{for some} \quad a \in \mathbb{R}.$$

Let us choose a $\lambda > 0$. Then

$$f(x) \leq a \Longrightarrow -\lambda f(x) \geq -\lambda a \Longrightarrow e^{-\lambda f} \geq e^{-\lambda a}.$$

Now, $e^{-\lambda f}$ is still a positive function and $e^{-\lambda a}$ is still a positive number, hence

$$\int_X e^{-\lambda f} \ d\mu \geq \int_{x: f(x) \leq a} e^{-\lambda f} \ d\mu \geq \mu\Big\{x \in X: \ f(x) \leq a\Big\} e^{-\lambda a},$$

from which

$$\mu\Big\{x : f(x) \leq a\Big\} \leq e^{\lambda a} \int_X e^{-\lambda f} \ d\mu.$$

Often, we can compute exactly or estimate the value of the integral $\int_X e^{-\lambda f} \ d\mu$. At that point, we "tune up" λ so that we get the strongest inequality possible.

Let us apply the method of Section 2.1 to the Gaussian measure.

(2.2) Proposition.

(1) *For any* $\delta \geq 0$

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x\|^2 \geq n + \delta\Big\} \leq \left(\frac{n}{n+\delta}\right)^{-n/2} e^{-\delta/2}.$$

(2) *For any* $0 < \delta \leq n$

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x\|^2 \leq n - \delta\Big\} \leq \left(\frac{n}{n-\delta}\right)^{-n/2} e^{\delta/2}.$$

A more or less straightforward corollary:

(2.3) Corollary. *For any* $0 < \epsilon < 1$

$$\gamma_n\Big\{x \in \mathbb{R}^n: \quad \|x\|^2 \geq \frac{n}{1-\epsilon}\Big\} \leq e^{-\epsilon^2 n/4} \quad \textit{and}$$
$$\gamma_n\Big\{x \in \mathbb{R}^n: \quad \|x\|^2 \leq (1-\epsilon)n\Big\} \leq e^{-\epsilon^2 n/4}.$$

Now to proofs.

Proof of Proposition 2.2. To prove Part 1, let us choose a $0 < \lambda < 1$ (to be adjusted later). Then

$$\|x\|^2 \geq n+\delta \Longrightarrow \lambda\|x\|^2/2 \geq \lambda(n+\delta)/2 \Longrightarrow e^{\lambda\|x\|^2/2} \geq e^{\lambda(n+\delta)/2}.$$

Reasoning as in Section 2.1, we get

$$\gamma_n\Big\{x \in \mathbb{R}^n: \quad \|x\|^2 \geq n+\delta\Big\} \leq e^{-\lambda(n+\delta)/2} \int_{\mathbb{R}^n} e^{\lambda\|x\|^2} \ d\gamma_n$$
$$= e^{-\lambda(n+\delta)/2}(2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{(\lambda-1)\|x\|^2/2} \ dx.$$

Quite handily, the integral can be computed exactly, because it factors into the product of 1-dimensional integrals:

$$(2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{(\lambda-1)\|x\|^2/2} \ dx = \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{(\lambda-1)x^2/2} \ dx\right)^n = (1-\lambda)^{-n/2}$$

(to compute the 1-dimensional integral, make the substitution $x = y/\sqrt{1-\lambda}$). Hence

$$\gamma_n\Big\{x \in \mathbb{R}^n: \quad \|x\|^2 \geq n+\delta\Big\} \leq (1-\lambda)^{-n/2} e^{-\lambda(n+\delta)/2}.$$

Now we choose $\lambda = \delta/(n+\delta)$ and Part 1 follows.

To prove Part 2, let us choose $\lambda > 0$ (to be adjusted later). Then

$$\|x\|^2 \leq n-\delta \Longrightarrow -\lambda\|x\|^2/2 \geq -\lambda(n-\delta)/2 \Longrightarrow e^{-\lambda\|x\|^2/2} \geq e^{-\lambda(n-\delta)/2}.$$

Reasoning as in Section 2.1, we get

$$\gamma_n\Big\{x \in \mathbb{R}^n: \quad \|x\|^2 \leq n-\delta\Big\} \leq e^{\lambda(n-\delta)/2} \int_{\mathbb{R}^n} e^{-\lambda\|x\|^2} \ d\gamma_n$$
$$= e^{\lambda(n-\delta)/2}(2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-(\lambda+1)\|x\|^2/2} \ dx.$$

Again, the integral is computed exactly:

$$(2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-(\lambda+1)\|x\|^2/2} \ dx = \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(\lambda+1)x^2/2} \ dx\right)^n = (1+\lambda)^{-n/2}$$

(to compute the 1-dimensional integral, make the substitution $x = y/\sqrt{1+\lambda}$). Hence

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x\|^2 \leq n - \delta\Big\} \leq (1+\lambda)^{-n/2} e^{\lambda(n-\delta)/2}.$$

Now we choose $\lambda = \delta/(n-\delta)$ and Part 2 follows. □

Lecture 2. Friday, January 7

---

2. "Condensing" the Sphere from the Gaussian Measure, Continued

Proof of Corollary 2.3. Let us choose $\delta = n\epsilon/(1-\epsilon)$ in Part 1 of Proposition 2.2. Then $n + \delta = n/(1-\epsilon)$ and

$$\begin{aligned}\gamma_n\Big\{x \in \mathbb{R}^n : \ \|x\|^2 \geq \frac{n}{1-\epsilon}\Big\} \leq &(1-\epsilon)^{-n/2} \exp\Big\{-\frac{n}{2}\frac{\epsilon}{1-\epsilon}\Big\} \\ = &\exp\Big\{-\frac{n}{2}\Big(\frac{\epsilon}{1-\epsilon} + \ln(1-\epsilon)\Big)\Big\}.\end{aligned}$$

Now,

$$\frac{\epsilon}{1-\epsilon} = \epsilon + \epsilon^2 + \epsilon^3 + \ldots \quad \text{and} \quad \ln(1-\epsilon) = -\epsilon - \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} - \ldots,$$

from which

$$\frac{\epsilon}{1-\epsilon} + \ln(1-\epsilon) \geq \frac{\epsilon^2}{2}$$

and

$$\gamma_n\Big\{x \in \mathbb{R}^n : \ \|x\|^2 \geq \frac{n}{1-\epsilon}\Big\} \leq e^{-\epsilon^2 n/4},$$

as promised.

Similarly, let us choose $\delta = n\epsilon$ in Part 2 of of Proposition 2.2. Then $n - \delta = (1-\epsilon)n$ and

$$\gamma_n\Big\{x \in \mathbb{R}^n : \ \|x\|^2 \leq (1-\epsilon)n\Big\} \leq (1-\epsilon)^{n/2} e^{\epsilon n/2} = \exp\Big\{\frac{n}{2}\Big(\ln(1-\epsilon) + \epsilon\Big)\Big\}.$$

Now,

$$\ln(1-\epsilon) = -\epsilon - \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} - \ldots,$$

from which

$$\ln(1-\epsilon) + \epsilon \leq -\frac{\epsilon^2}{2}$$

and

$$\gamma_n\Big\{x \in \mathbb{R}^n : \ \|x\|^2 \leq (1-\epsilon)n\Big\} \leq e^{-\epsilon^2 n/4},$$

as promised. □

One geometric interpretation of Corollary 2.3 is as follows: let us choose a sequence $\rho_n \longrightarrow +\infty$, however slowly, and consider a "fattening" of the sphere

$$\Big\{x \in \mathbb{R}^n : \quad \sqrt{n} - \rho_n \leq \|x\| \leq \sqrt{n} + \rho_n\Big\}.$$

Then the Gaussian measure γ_n of that "fattening" approaches 1 as n increases (let $\epsilon_n = \rho_n/\sqrt{n}$ in Corollary 2.3)

3. The Johnson-Lindenstrauss "Flattening" Lemma

The Gaussian measure behaves nicely with respect to orthogonal projections.

(3.1) Gaussian measures and projections. Let γ_n be the standard Gaussian measure in $\mathbb{R}^n$ and let $L \subset \mathbb{R}^n$ be a k-dimensional subspace. Let $p : \mathbb{R}^n \longrightarrow L$ be the orthogonal projection. Then the "push-forward" measure $p(\gamma_n)$ is just the standard Gaussian measure on L with the density $(2\pi)^{-k/2}e^{-\|x\|^2/2}$ for $x \in L$.

Recall (or just be aware) that if we have a map $\phi : X \longrightarrow Y$ and a measure μ on X then the push-forward measure $\nu = \phi(\mu)$ on Y is defined by $\nu(A) = \mu\big(\phi^{-1}(A)\big)$ for subsets $A \subset Y$. The reason why the push-forward of the Gaussian measure under orthogonal projections is the Gaussian measure on the image is as follows: since the Gaussian density is invariant under orthogonal transformations of $\mathbb{R}^n$, rotating the subspace, if necessary, we may assume that L is the coordinate subspace consisting of the points $(\xi_1, \ldots, \xi_k, 0, \ldots, 0)$, where the last $n-k$ coordinates are 0's. Suppose that $A \subset L$ is a measurable set and that $B = p^{-1}(A)$ is the preimage of A. Since the Gaussian density splits into the product of the coordinate Gaussian densities, we have

$$\begin{aligned}\gamma_n(B) =&(2\pi)^{-n/2} \int_B \exp\big\{-(\xi_1^2 + \ldots + \xi_n^2)/2\big\} \ d\xi_1 \cdots d\xi_n \\ =&(2\pi)^{-k/2} \int_A \exp\big\{-(\xi_1^2 + \ldots + \xi_k^2)/2\big\} \ d\xi_1 \cdots d\xi_k \\ &\times \prod_{i=k+1}^{n} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\xi_i^2/2} \ d\xi_i = \gamma_k(A).\end{aligned}$$

This fact that the projection of the standard Gaussian measure is the standard Gaussian measure on the image is very useful and important. Often, we will word it more informally as: "if vector x has the standard Gaussian distribution in $\mathbb{R}^n$, its orthogonal projection y onto a given subspace L has the standard Gaussian distribution in L".

(3.2) Lemma. *Let γ_n be the standard Gaussian measure in $\mathbb{R}^n$ and let $L \subset \mathbb{R}^n$ be a k-dimensional subspace. For a vector $x \in \mathbb{R}^n$, let x_L denote the orthogonal* *projection of x onto L. Then, for any $0 < \epsilon < 1$,*

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \sqrt{\frac{n}{k}}\|x_L\| \geq (1-\epsilon)^{-1}\|x\|\Big\} \leq e^{-\epsilon^2 k/4} + e^{-\epsilon^2 n/4} \quad \textit{and}$$

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \sqrt{\frac{n}{k}}\|x_L\| \leq (1-\epsilon)\|x\|\Big\} \leq e^{-\epsilon^2 k/4} + e^{-\epsilon^2 n/4}.$$

Proof. By Corollary 2.3,

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x\| \geq \sqrt{(1-\epsilon)n}\Big\} \geq 1 - e^{-\epsilon^2 n/4}.$$

If x has the standard Gaussian distribution in $\mathbb{R}^n$, then the projection x_L has the standard Gaussian distribution γ_k in L (cf. Section 3.1), and so by Corollary 2.3 applied to L we have

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x_L\| \leq \sqrt{\frac{k}{1-\epsilon}}\Big\} \geq 1 - e^{-\epsilon^2 k/4}.$$

Therefore,

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \sqrt{\frac{n}{k}}\|x_L\| \leq (1-\epsilon)^{-1}\|x\|\Big\} \geq 1 - e^{-\epsilon^2 n/4} - e^{-\epsilon^2 k/4}.$$

Similarly,

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x\| \leq \sqrt{\frac{n}{1-\epsilon}}\Big\} \geq 1 - e^{-\epsilon^2 n/4} \quad \text{and}$$

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \|x_L\| \geq \sqrt{(1-\epsilon)k}\Big\} \geq 1 - e^{-\epsilon^2 k/4},$$

from which

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \sqrt{\frac{n}{k}}\|x_L\| \geq (1-\epsilon)\|x\|\Big\} \geq 1 - e^{-\epsilon^2 n/4} - e^{-\epsilon^2 k/4}.$$

□

(3.3) The Gaussian measure in $\mathbb{R}^n$ and the uniform measure on the sphere. Let us consider the radial projection $\phi : \mathbb{R}^n \setminus \{0\} \longrightarrow \mathbb{S}^{n-1}$, $x \longmapsto x/\|x\|$. Not surprisingly, the push-forward of the standard Gaussian measure γ_n on $\mathbb{R}^n$ will be the uniform probability measure on the sphere $\mathbb{S}^{n-1}$. This is so because the the Gaussian measure is rotation-invariant, so the push-forward must be rotation invariant and there is a unique (Borel) rotation-invariant probability measure on $\mathbb{S}^{n-1}$. We don't prove the existence and uniqueness of the rotation-invariant Borel probability measure on $\mathbb{S}^{n-1}$, instead we describe a simple procedure to sample a point from this distribution: sample a random vector x from the Gaussian measure on $\mathbb{R}^n$ and project radially: $x \longmapsto x/\|x\|$ (note that $x \neq 0$ with probability 1).

(3.4) Corollary. *Let μ_n be the rotation-invariant probability measure on the unit sphere $\mathbb{S}^{n-1}$ and let $L \subset \mathbb{R}^n$ be a k-dimensional subspace. For a vector $x \in \mathbb{S}^{n-1}$, let x_L denote the orthogonal projection of x onto L. Then, for any $0 < \epsilon < 1$,*

$$\mu_n\Big\{x \in \mathbb{S}^{n-1} : \quad \sqrt{\frac{n}{k}}\|x_L\| \geq (1-\epsilon)^{-1}\|x\|\Big\} \leq e^{-\epsilon^2 k/4} + e^{-\epsilon^2 n/4} \quad \textit{and}$$

$$\mu_n\Big\{x \in \mathbb{S}^{n-1} : \quad \sqrt{\frac{n}{k}}\|x_L\| \leq (1-\epsilon)\|x\|\Big\} \leq e^{-\epsilon^2 k/4} + e^{-\epsilon^2 n/4}.$$

Proof. The ratio $\|x_L\|/\|x\|$ does not change under the radial projection $\phi : x \longmapsto x/\|x\|$. Hence the result follows from Lemma 3.2 and the discussion of Section 3.3 above. □

Now, instead of *fixing* a subspace $L \subset \mathbb{R}^n$ and choosing a *random* $x \in \mathbb{S}^{n-1}$, we *fix* a vector $x \in \mathbb{S}^{n-1}$ and choose a *random* k-dimensional subspace $L \subset \mathbb{R}^n$. It is intuitively obvious that "nothing changes".

(3.5) The orthogonal group, the Grassmannian, and the Gaussian measure. To explain rigorously why "nothing changes" would require us to prove some basic results about the existence and uniqueness of the Haar measure. This would lead us too far off the course. Instead, we present some informal discussion. Let $G_k(\mathbb{R}^n)$ be the *Grassmannian*, that is the set of all k-dimensional subspaces of $\mathbb{R}^n$. The set $G_k(\mathbb{R}^n)$ has a natural structure of a compact metric space: for example, one can define the distance between two subspaces as the Hausdorff distance between the unit balls in those subspaces. The metric is invariant under the action of the orthogonal group and the action is transitive on $G_k(\mathbb{R}^n)$. Therefore, there exists a unique (Borel) probability measure $\mu_{n,k}$ on $G_k(\mathbb{R}^n)$ invariant under the action of the orthogonal group. Again, we don't define this measure, instead we describe a procedure to sample a random subspace with respect to $\mu_{n,k}$: sample k vectors $x_1, \ldots, x_k \in \mathbb{R}^n$ independently from the standard Gaussian measure in $\mathbb{R}^n$ and let L be the subspace spanned by $x_1, \ldots, x_k$ (note that $x_1, \ldots, x_k$ are linearly independent with probability 1, so $\dim L = k$). In other words, we get the measure $\mu_{n,k}$ as the push-forward of the product $\gamma_n \times \ldots \times \gamma_n$ on $\mathbb{R}^n \oplus \ldots \oplus \mathbb{R}^n$ under the map $(x_1, \ldots, x_k) \longmapsto \text{span}\,\{x_1, \ldots, x_k\}$. Since the result is invariant under orthogonal transformations, the resulting measure should coincide with $\mu_{n,k}$.

Similarly, let O_n be the group of the orthogonal transformations of $\mathbb{R}^n$. Then O_n has an invariant metric: for example, for two orthogonal transformations A and B, let $\text{dist}(A, B) = \max_{x \in \mathbb{S}^{n-1}} \text{dist}(Ax, Bx)$. This makes O_n a compact metric space, so there is a unique Borel probability measure ν_n invariant under the right- and left- multiplications by orthogonal matrices. To sample an orthogonal matrix A from ν_n, we sample n vectors $x_1, \ldots, x_n$ independently from the standard Gaussian measure in $\mathbb{R}^n$, apply the Gram-Schmidt orthogonalization process $x_1, \ldots, x_n \longmapsto u_1, \ldots, u_n$ and let A be the matrix with the columns $u_1, \ldots, u_n$ (note that $x_1, \ldots, x_n$ are linearly independent with probability 1, so the process works). This allows us to understand ν_n as the push-forward of the product

$\gamma_n \times \ldots \times \gamma_n$ on $\mathbb{R}^n \oplus \ldots \oplus \mathbb{R}^n$ under that map $(x_1, \ldots, x_n) \longmapsto$ Gram-Schmidt orthogonalization of $x_1, \ldots, x_n$.

Now we can state a version of the famous Johnson-Lindenstrauss flattening Lemma.

(3.6) Lemma. *Let $x \in \mathbb{R}^n$ be a non-zero vector and let $\mu_{n,k}$ be the invariant probability measure on the Grassmannian $G_k(\mathbb{R}^n)$ of k-dimensional subspaces in $\mathbb{R}^n$. For $L \in G_k(\mathbb{R}^n)$, let x_L be the orthogonal projection of x onto L.*

Then, for any $0 < \epsilon < 1$,

$$\mu_{n,k}\Big\{L \in G_k(\mathbb{R}^n) : \quad \sqrt{\frac{n}{k}}\|x_L\| \geq (1-\epsilon)^{-1}\|x\|\Big\} \leq e^{-\epsilon^2 k/4} + e^{-\epsilon^2 n/4} \quad \textit{and}$$

$$\mu_{n,k}\Big\{L \in G_k(\mathbb{R}^n) : \quad \sqrt{\frac{n}{k}}\|x_L\| \leq (1-\epsilon)\|x\|\Big\} \leq e^{-\epsilon^2 k/4} + e^{-\epsilon^2 n/4}.$$

Lecture 3. Monday, January 10

---

3. The Johnson-Lindenstrauss "Flattening" Lemma, Continued

Proof of Lemma 3.6. Scaling, if necessary, we may assume that $\|x\| = 1$, so that $x \in \mathbb{S}^{n-1}$. Let us choose a k-dimensional subspace $L_0 \subset \mathbb{R}^n$. If $U \in O_n$ is an orthogonal transformation, then $L = U(L_0) = \Big\{Uy : \ y \in L_0\Big\}$ is a k-dimensional subspace of $\mathbb{R}^n$. Moreover, as U ranges over the orthogonal group O_n, the subspace L ranges over the Grassmannian $G_k(\mathbb{R}^n)$ and if we apply a "random" U, we get a "random" L, meaning that the invariant probability measure $\mu_{n,k}$ on $G_k(\mathbb{R}^n)$ is the push-forward of the invariant probability measure ν_n on O_n under the map $U \longmapsto U(L_0)$. Therefore,

$$\begin{aligned}
&\mu_{n,k}\Big\{L \in G_k(\mathbb{R}^n) : \quad \sqrt{\frac{n}{k}}\|x_L\| \geq (1-\epsilon)^{-1}\Big\} \\
=&\nu_n\Big\{U \in O_n : \quad \sqrt{\frac{n}{k}}\|x_{U(L_0)}\| \geq (1-\epsilon)^{-1}\Big\} \quad \text{and} \\
&\mu_{n,k}\Big\{L \in G_k(\mathbb{R}^n) : \quad \sqrt{\frac{n}{k}}\|x_L\| \leq (1-\epsilon)\Big\} \\
=&\nu_n\Big\{U \in O_n : \quad \sqrt{\frac{n}{k}}\|x_{U(L_0)}\| \leq (1-\epsilon)\Big\}.
\end{aligned}$$

Now, the length $\|x_{U(L_0)}\|$ of the projection of x onto $U(L_0)$ is equal to the length of the projection of $y = U^{-1}x$ onto L_0. As U ranges over the orthogonal group O_n, the vector $y = U^{-1}x$ ranges over the unit sphere $\mathbb{S}^{n-1}$ and if we apply a "random"

U, we get a "random" y, meaning that the invariant probability measure μ_n on $\mathbb{S}^{n-1}$ is the push-forward of the invariant probability measure ν_n on O_n under the map $U \longmapsto U^{-1}x$.

Therefore,

$$\begin{aligned}&\nu_n\Big\{U \in O_n : \quad \sqrt{\frac{n}{k}}\|x_{U(L_0)}\| \geq (1-\epsilon)^{-1}\Big\}\\ &= \mu_n\Big\{y \in \mathbb{S}^{n-1} : \quad \sqrt{\frac{n}{k}}\|y_{L_0}\| \geq (1-\epsilon)^{-1}\Big\} \quad \text{and}\\ &\nu_n\Big\{U \in O_n : \quad \sqrt{\frac{n}{k}}\|x_{U(L_0)}\| \leq (1-\epsilon)\Big\}\\ &= \mu_n\Big\{y \in \mathbb{S}^{n-1} : \quad \sqrt{\frac{n}{k}}\|y_{L_0}\| \leq (1-\epsilon)\Big\}.\end{aligned}$$

The proof follows by Corollary 3.4. □

Finally, we arrive to the following result, which is also referred to as a version of the Johnson-Lindenstrauss "Flattening" Lemma.

(3.7) Theorem. *Let $a_1, \ldots, a_N$ be points in $\mathbb{R}^n$. Given an $\epsilon > 0$, let us choose an integer k such that*

$$N(N-1)\left(e^{-k\epsilon^2/4} + e^{-n\epsilon^2/4}\right) \leq \frac{1}{3}.$$

Remark: for example,

$$k \geq 4\epsilon^{-2}\ln(6N^2)$$

will do. Assuming that $k \leq n$, let $L \subset \mathbb{R}^n$ be a k-dimensional subspace chosen at random with respect to the invariant probability measure $\mu_{n,k}$ on the Grassmannian $G_k(\mathbb{R}^n)$. Let $a'_1, \ldots, a'_N$ be the orthogonal projections of $a_1, \ldots, a_N$ onto L. Then,

$$(1-\epsilon)\|a_i - a_j\| \leq \sqrt{\frac{n}{k}}\|a'_i - a'_j\| \leq (1-\epsilon)^{-1}\|a_i - a_j\| \quad \textit{for all} \quad 1 \leq i,j \leq N$$

with probability at least $2/3$.

Proof. Let $c_{ij} = a_i - a_j$ for $i > j$. There are $\binom{N}{2} = N(N-1)/2$ vectors c_{ij} and the length of the orthogonal projection of c'_{ij} onto a subspace L is equal to $\|a'_i - a'_j\|$. Using Lemma 3.6, we conclude that the probability that for some pair i, j we have either

$$\sqrt{\frac{n}{k}}\|c'_{ij}\| \geq (1-\epsilon)^{-1}\|c_{ij}\| \quad \text{or} \quad \sqrt{\frac{n}{k}}\|c'_{ij}\| \leq (1-\epsilon)\|c_{ij}\|$$

(or both) is at most $1/3$. □

(3.8) Sharpening the estimates. Given a positive integer N and an $\epsilon > 0$, what is smallest possible dimension k so that for any N points $a_1, \ldots, a_N$ in Euclidean space $\mathbb{R}^n$, there are N points $b_1, \ldots, b_N$ in $\mathbb{R}^k$ with the property that

$$(1-\epsilon)\operatorname{dist}(a_i, a_j) \leq \operatorname{dist}(b_i, b_j) \leq (1-\epsilon)^{-1}\operatorname{dist}(a_i, a_j) \quad \text{for all} \quad i, j,$$

where dist is the Euclidean distance? It looks like a difficult question, but one thing is clear: the dimension k should not depend on n since we can always consider $\mathbb{R}^n$ as a subspace of a higher-dimensional space. Reasoning as in the proof of Theorem 3.7 and making $n \longrightarrow +\infty$ (which, in principle, provides us with more choice for a random subspace L), we can choose any positive integer k such that

$$k > 4\epsilon^{-2}\ln\Big(N(N-1)\Big).$$

For example, if $N = 10^9$ and $\epsilon = 0.1$ (so that we allow a 10% distortion), we can choose $k = 16,579$. It is not clear whether better estimates are possible, if they are, we need better methods.

Lecture 4. Wednesday, January 12

4. Concentration in The Boolean Cube

Let us switch the gears and consider something very discrete.

(4.1) The Boolean cube. Let $I_n = \{0,1\}^n$ be the Boolean cube, that is, the set of all 2^n sequences of length n of 0's and 1's. Combinatorially, we can think of I_n as of the set of all 2^n subsets of the set $\{1, \ldots, n\}$.

We make I_n a metric space by introducing the *Hamming distance*: the distance between two points $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$ is the number of the coordinates where x and y disagree:

$$\operatorname{dist}(x, y) = \big|i : \quad \xi_i \neq \eta_i\big|.$$

Furthermore, let us introduce the *counting probability measure* μ_n on I_n by $\mu_n\{x\} = 2^{-n}$.

QUESTIONS: what is the diameter of I_n? What is the distance between two "typical" points in I_n (whatever that means)?

The Boolean cube provides a very good model to observe various concentration phenomena. Since I_n is finite, all functions $f : I_n \longrightarrow \mathbb{R}$ are measurable, all integrals are sums, and we don't have to worry about subtleties. Later, the Boolean cube will serve as an inspiration to tackle more complicated spaces. In view of that, we will freely interchange between

$$\int_{I_n} f \, d\mu_n \quad \text{and} \quad \frac{1}{2^n}\sum_{x \in I_n} f(x).$$

Our next goal is to prove that sets $A \subset I_n$ with $\mu_n(A) = 1/2$ have large (measure-wise) ϵ-neighborhoods for moderate ϵ's of order $\sqrt{n}$. To do that, we prove a concentration inequality for the function $f(x) = \operatorname{dist}(x, A)$, where, as usual, $\operatorname{dist}(x, A) = \min_{y \in A} \operatorname{dist}(x, y)$. We use the Laplace transform method of Section 2.1. The result below is a particular case of a more general result by M. Talagrand. We also adopt Talagrand's proof (he used it in a more general situation).

(4.2) Theorem. *Let $A \subset I_n$ be a non-empty set and let $f : I_n \longrightarrow \mathbb{R}$, $f(x) = \operatorname{dist}(x, A)$ be the distance from x to A. Then, for any $t > 0$, we have*

$$\int_{I_n} e^{tf} \, d\mu_n \leq \frac{1}{\mu_n(A)} \left(\frac{1}{2} + \frac{e^t + e^{-t}}{4} \right)^n .$$

Proof. Let us denote

$$c(t) = \frac{1}{2} + \frac{e^t + e^{-t}}{4}.$$

We use the induction on the dimension n of the cube. If $n = 1$ then either A consists of 1 point, at which case $\mu_1(A) = 1/2$ and

$$\int_{I_1} e^{tf} \, d\mu_1 = \frac{1}{2} + \frac{1}{2} e^t \leq 2c(t),$$

so the result holds, or A consists of 2 points, at which case $\mu_1(A) = 1$ and

$$\int_{I_1} e^{tf} \, d\mu_1 = 1 \leq c(t),$$

in which case the result holds as well.

Suppose that $n > 1$. Let us split the cube I_n into two "facets", depending on the value of the last coordinate: $I_n^1 = \{x : \ \xi_n = 1\}$ and $I_n^0 = \{x : \ \xi_n = 0\}$. Note that I_n^1 and I_n^0 can be identified with the $(n-1)$-dimensional cube I_{n-1}. Consequently, let us define subsets $A_1, A_0 \subset I_{n-1}$ by

$$A_0 = \Big\{ x \in I_{n-1} : \quad (x, 0) \in A \Big\} \quad \text{and} \quad A_1 = \Big\{ x \in I_{n-1} : \quad (x, 1) \in A \Big\}.$$

Note that $|A| = |A_1| + |A_0|$, or, in the measure terms,

$$\mu_n(A) = \frac{\mu_{n-1}(A_1) + \mu_{n-1}(A_0)}{2}.$$

For a point $x \in I_n$, let $x' \in I_{n-1}$ be the point with the last coordinate omitted. Now, we have

$$\operatorname{dist}(x, A) = \min\Big\{ \operatorname{dist}(x', A_1), \quad \operatorname{dist}(x', A_0) + 1 \Big\} \quad \text{for every point} \quad x \in I_n^1 \quad \text{and}$$
$$\operatorname{dist}(x, A) = \min\Big\{ \operatorname{dist}(x', A_0), \quad \operatorname{dist}(x', A_1) + 1 \Big\} \quad \text{for every point} \quad x \in I_n^0,$$

Let us define functions f_0, f_1 by $f_0(x) = \operatorname{dist}(x, A_0)$ and $f_1(x) = \operatorname{dist}(x, A_1)$ for $x \in I_{n-1}$. Then

$$\begin{aligned}
\int_{I_n} e^{tf}\ d\mu_n =& \int_{I_n^1} e^{tf}\ d\mu_n + \int_{I_n^0} e^{tf}\ d\mu_n \\
=& \frac{1}{2^n} \sum_{x \in I_n^1} \exp\{t \operatorname{dist}(x, A)\} + \frac{1}{2^n} \sum_{x \in I_n^0} \exp\{t \operatorname{dist}(x, A)\} \\
=& \frac{1}{2^n} \sum_{x \in I_n^1} \exp\Big\{\min\{t \operatorname{dist}(x', A_1),\ t + t \operatorname{dist}(x', A_0)\}\Big\} \\
&+ \frac{1}{2^n} \sum_{x \in I_n^0} \exp\Big\{\min\{t \operatorname{dist}(x', A_0),\ t + t \operatorname{dist}(x', A_1)\}\Big\} \\
=& \frac{1}{2^n} \sum_{x \in I_n^1} \min\Big\{\exp\{t \operatorname{dist}(x', A_1)\},\ e^t \exp\{t \operatorname{dist}(x', A_0)\}\Big\} \\
&+ \frac{1}{2^n} \sum_{x \in I_n^0} \min\Big\{\exp\{t \operatorname{dist}(x', A_0)\},\ e^t \exp\{t \operatorname{dist}(x', A_1)\}\Big\} \\
=& \frac{1}{2} \int_{I_{n-1}} \min\{e^{tf_1},\ e^t e^{tf_0}\}\ d\mu_{n-1} + \frac{1}{2} \int_{I_{n-1}} \min\{e^{tf_0},\ e^t e^{tf_1}\}\ d\mu_{n-1}.
\end{aligned}$$

However, the integral of the minimum does not exceed the minimum of the integrals. We carry on, and use the induction hypothesis.

$$\begin{aligned}
\int_{I_n} e^{tf}\ d\mu_n \leq& \frac{1}{2} \min\Big\{\int_{I_{n-1}} e^{tf_1}\ d\mu_{n-1}, \quad e^t \int_{I_{n-1}} e^{tf_0}\ d\mu_{n-1}\Big\} \\
&+ \frac{1}{2} \min\Big\{\int_{I_{n-1}} e^{tf_0}\ d\mu_{n-1}, \quad e^t \int_{I_{n-1}} e^{tf_1}\ d\mu_{n-1}\Big\} \\
\leq& \frac{1}{2} \min\Big\{\frac{c(t)^{n-1}}{\mu_{n-1}(A_1)}, \quad e^t \frac{c^{n-1}(t)}{\mu_{n-1}(A_0)}\Big\} \\
&+ \frac{1}{2} \min\Big\{\frac{c^{n-1}(t)}{\mu_{n-1}(A_0)}, \quad e^t \frac{c^{n-1}(t)}{\mu_{n-1}(A_1)}\Big\} = \\
=& \frac{c^{n-1}(t)}{\mu_n(A)} \bigg(\frac{1}{2} \min\Big\{\frac{\mu_n(A)}{\mu_{n-1}(A_1)},\ e^t \frac{\mu_n(A)}{\mu_{n-1}(A_0)}\Big\} \\
&+ \frac{1}{2} \min\Big\{\frac{\mu_n(A)}{\mu_{n-1}(A_0)},\ e^t \frac{\mu_n(A)}{\mu_{n-1}(A_1)}\Big\}\bigg).
\end{aligned}$$

Let us denote

$$a_0 = \frac{\mu_{n-1}(A_0)}{\mu_n(A)} \quad \text{and} \quad a_1 = \frac{\mu_{n-1}(A_1)}{\mu_n(A)}, \quad \text{so that} \quad a_0 + a_1 = 2.$$

It all boils down to the question: what is the maximum possible value of

$$\frac{1}{2}\min\left\{a_1^{-1},\ e^t a_0^{-1}\right\} + \frac{1}{2}\min\left\{a_0^{-1},\ e^t a_1^{-1}\right\},$$

where a_0 and a_1 are non-negative numbers such that $a_0 + a_1 = 2$. If $a_0 = 0$ then $a_1 = 2$ and the value is

$$\frac{1}{4} + \frac{e^t}{4} \leq c(t).$$

The same value is attained if $a_1 = 0$ and $a_0 = 2$.

Suppose that at a maximum point we have $a_0, a_1 > 0$. If $a_1 = a_0 = 1$, the value is

$$\frac{1}{2} + \frac{1}{2} \leq c(t).$$

If $a_0 \neq a_1$, without loss of generality, we can assume that $a_0 > a_1$, so that $a_0 = 1+b$ and $a_1 = 1 - b$ for some $0 < b < 1$. Then $a_0^{-1} < e^t a_1^{-1}$. If $a_1^{-1} < e^t a_0^{-1}$, the value is

$$\frac{1}{2}(1-b)^{-1} + \frac{1}{2}(1+b)^{-1} = \frac{1}{1-b^2}.$$

Increasing b slightly, we increase the value, which is a contradiction. If $a_1^{-1} > e^t a_0^{-1}$, then the value is

$$\frac{1}{2}e^t a_0^{-1} + \frac{1}{2}a_0^{-1} = \frac{e^t}{2(1+b)}.$$

Decreasing b slightly, we increase the value, which is contradiction. Hence at the maximum point we must have

$$a_1^{-1} = e^t a_0^{-1}, \quad \text{that is,} \quad a_0 = \frac{2e^t}{1+e^t} \quad \text{and} \quad a_1 = \frac{2}{1+e^t},$$

with the value

$$\frac{1}{2}a_1^{-1} + \frac{1}{2}a_0^{-1} = \frac{1}{2} + \frac{e^t + e^{-t}}{4} = c(t).$$

This allows us to complete the proof:

$$\int_{I_n} e^{tf}\ d\mu_n \leq \frac{c^{n-1}(t)}{\mu_n(A)}c(t) = \frac{c^n(t)}{\mu_n(A)}.$$

□

Here is a useful corollary.

(4.3) Corollary. *Let $A \subset I_n$ be a non-empty set and let $f : I_n \longrightarrow \mathbb{R}$, $f(x) = \operatorname{dist}(x, A)$ be the distance from x to A. Then, for any $t > 0$, we have*

$$\int_{I_n} e^{tf} \ d\mu_n \leq \frac{1}{\mu_n(A)} e^{t^2 n/4}.$$

Proof. Follows from Theorem 4.2, since

$$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots \quad \text{and} \quad e^{-t} = 1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \ldots \ ,$$

so

$$\begin{aligned}\frac{1}{2} + \frac{e^t + e^{-t}}{4} =& 1 + \frac{t^2}{4} + \frac{t^4}{2 \cdot 4!} \ldots + \frac{t^{2k}}{4(2k)!} + \ldots \\ \leq& 1 + \frac{t^2}{4} + \ldots + \frac{t^{2k}}{4^k k!} + \ldots = e^{t^2/4}.\end{aligned}$$

□

Lecture 6. Wednesday, January 19

Lecture 5 on Friday, January 14, covered the proof of Theorem 4.2 from the previous handout.

4. Concentration in The Boolean Cube, Continued

With Corollary 4.3 in hand, we can show that "almost all" points in the Boolean cube I_n are "very close" to any "sufficiently large" set $A \subset I_n$.

(4.4) Corollary. *Let $A \subset I_n$ be a non-empty set. Then, for any $\epsilon > 0$, we have*

$$\mu_n \Big\{ x \in I_n : \quad \operatorname{dist}(x, A) \geq \epsilon\sqrt{n} \Big\} \leq \frac{e^{-\epsilon^2}}{\mu_n(A)}.$$

Proof. We use the inequality of the Laplace transform method of Section 2.1. For any $t \geq 0$, we have

$$\begin{aligned}\mu_n \Big\{ x \in I_n : \quad \operatorname{dist}(x, A) \geq \epsilon\sqrt{n} \Big\} \leq& e^{-t\epsilon\sqrt{n}} \int_{I_n} \exp\big\{ t \operatorname{dist}(x, A) \big\} \ d\mu_n \\ \leq& \frac{e^{-t\epsilon\sqrt{n}}}{\mu_n(A)} e^{t^2 n/4},\end{aligned}$$

where the last inequality follows by Lemma 4.3. Now, we optimize on t by substituting $t = 2\epsilon/\sqrt{n}$. □

For example, if $\mu_n(A) = 0.01$, that is, the set A contains 1% of the points in the cube and we choose $\epsilon = 2\sqrt{\ln 10} \approx 3.035$ in Corollary 4.4, we conclude that 99% of the points of the cube I_n are within distance $3.035\sqrt{n}$ from A.

Let us consider a "reasonable" function $f : I_n \longrightarrow \mathbb{R}$. By "reasonable" we mean reasonably slowly varying, for example 1-Lipschitz:

$$|f(x) - f(y)| \leq \operatorname{dist}(x, y) \quad \text{for all} \quad x, y \in I_n.$$

We can show that f "almost always" remains very close to its median, that is the number m_f such that

$$\mu_n\Big\{x \in I_n : \quad f(x) \geq m_f\Big\} \geq \frac{1}{2} \quad \text{and} \quad \mu_n\Big\{x \in I_n : \quad f(x) \leq m_f\Big\} \geq \frac{1}{2}.$$

(4.5) Theorem. *Let $f : I_n \longrightarrow \mathbb{R}$ be a function such that $|f(x) - f(y)| \leq \operatorname{dist}(x, y)$ for all $x, y \in I_n$. Let m_f be the median of f. Then for all $\epsilon > 0$*

$$\mu_n\Big\{x \in I_n : \quad |f(x) - m_f| \geq \epsilon\sqrt{n}\Big\} \leq 4e^{-\epsilon^2}.$$

Proof. Let us consider two sets

$$A_+ = \Big\{x \in I_n : \quad f(x) \geq m_f\Big\} \quad \text{and} \quad A_- = \Big\{x \in I_n : \quad f(x) \leq m_f\Big\}.$$

Then $\mu_n(A_+), \mu_n(A_-) \geq 1/2$. Let $A^+(\epsilon)$ be the $\epsilon\sqrt{n}$-neighborhood of A_+ and $A_-(\epsilon)$ be the $\epsilon\sqrt{n}$-neighborhood of A_-.

$$\begin{aligned} A_+(\epsilon) &= \Big\{x \in I_n : \quad \operatorname{dist}(x, A_+) \leq \epsilon\sqrt{n}\Big\} \quad \text{and} \\ A_-(\epsilon) &= \Big\{x \in I_n : \quad \operatorname{dist}(x, A_-) \leq \epsilon\sqrt{n}\Big\}. \end{aligned}$$

By Corollary 4.4,

$$\mu_n\left(A_+(\epsilon)\right) \geq 1 - 2e^{-\epsilon^2} \quad \text{and} \quad \mu_n\left(A_-(\epsilon)\right) \geq 1 - 2e^{-\epsilon^2}.$$

Therefore, for

$$A(\epsilon) = A_+(\epsilon) \cap A_-(\epsilon) \quad \text{we have} \quad \mu_n\left(A(\epsilon)\right) \geq 1 - 4e^{-\epsilon^2}.$$

Now, for every $x \in A(\epsilon)$, we have $|f(x) - m_f| \leq \epsilon\sqrt{n}$, which completes the proof. □

For example, if we choose $\epsilon = \sqrt{\ln(400)} \approx 2.45$, we conclude that for 99% of the points of I_n, the function f does not deviate from the median value by more than $2.45\sqrt{n}$.

5. Isoperimetric Inequalities in the Boolean Cube: A Soft Version

Looking at Corollary 4.4 we can ask ourselves, if we can sharpen the bound for small sets A, where "small" means that $\mu_n(A)$ is exponentially small in n, such as $(1.1)^{-n}$, say (it cannot be smaller than 2^{-n}). That is, what can be the *minimum measure* of the ϵ-neighborhood

$$A(\epsilon) = \Big\{ x \in I_n : \quad \operatorname{dist}(x, A) \leq \epsilon\sqrt{n} \Big\}$$

of a set $A \subset I_n$ of a given measure $\mu_n(A)$. The answer to this question is known, but the construction is particular to the Boolean cube. Instead of discussing it now, let us discuss a softer version first: for a set $A \subset I_n$, let us consider the function $f(x) = \operatorname{dist}(x, A)$. Since f is 1-Lipschitz, by Theorem 4.5 it concentrates around its median. Just as well, it should concentrate around its mean value.

PROBLEMS.

1. Let $A \subset I_n$ be a non-empty set and let us consider $f : I_n \longrightarrow \mathbb{R}$ defined by $f(x) = \operatorname{dist}(x, A)$. Let m_f be the median of f and let

$$a_f = \int_{I_n} \operatorname{dist}(x, A) \ d\mu_n$$

be the average value of f. Deduce from Theorem 4.5 that

$$|m_f - a_f| \leq \sqrt{\frac{n \ln n}{2}} + 4\sqrt{n}.$$

2. Let $A \subset I_n$ be a set consisting of a single point $A = \{a\}$. Prove that

$$\int_{I_n} \operatorname{dist}(x, A) \ d\mu_n = \frac{n}{2}.$$

Deduce that for all non-empty $A \subset I_n$,

$$\int_{I_n} \operatorname{dist}(x, A) \ d\mu_n \leq \frac{n}{2}.$$

Thus we ask ourselves, given $\mu_n(A) > 0$, how small can the average value a of $\operatorname{dist}(x, A)$ be? Our next goal is to prove the following result.

(5.1) Theorem. *Let $A \subset I_n$ be a non-empty set and let*

$$\rho = \frac{1}{2} - \frac{1}{n} \int_{I_n} \operatorname{dist}(x, A) \ d\mu_n, \quad \textit{so} \quad 0 \leq \rho \leq \frac{1}{2}.$$

Then

$$\frac{\ln |A|}{n} \leq \rho \ln \frac{1}{\rho} + (1 - \rho) \ln \frac{1}{1 - \rho}.$$

Remark. The function

$$H(\rho) = \rho \ln \frac{1}{\rho} + (1 - \rho) \ln \frac{1}{1 - \rho} \quad \text{for} \quad 0 \leq \rho \leq 1.$$

is called the *entropy* of ρ. Note that $H(\rho) = H(1 - \rho)$ and that $H(0) = H(1) = 0$. The maximum value of H is attained at $\rho = 1/2$ and is equal to $H(1/2) = \ln 2$. For $0 \leq \rho \leq 1/2$, $H(\rho)$ is increasing and for $1/2 \leq \rho \leq 1$, $H(\rho)$ is decreasing.

The inequality of Theorem 5.1 should be understood as follows: if the set A is reasonably large, then the average distance $\operatorname{dist}(x, A)$ from $x \in I_n$ to A cannot be very large. For example, if $|A| = 2^n$, the left hand of the inequality is equal to $\ln 2$, so that ρ must be equal to $1/2$ and the average distance from $x \in I_n$ to A is 0, which is indeed the case since A is the whole cube. On the opposite end, the minimum possible value of the right hand side is 0 at $\rho = 0$. In this case, the average distance from x to A is $n/2$ and $|A|$ should be equal to 1, so A must be a point. If $|A| \geq (1.3)^n$, then we must have $H(\rho) \geq 0.26$, so $\rho \geq 0.07$ and the average distance from $x \in I_n$ to A should be at most $0.43n$. If we try to apply Corollary 4.4, to get the right hand side of the inequality less than 1, we should choose $\epsilon \geq \sqrt{-n \ln 0.65} \approx 0.66\sqrt{n}$ and we get a vacuous estimate that there are points $x \in I_n$ with $\operatorname{dist}(x, A) \leq 0.66n$. The moral of the story is that Corollary 3.3 is not optimal for sets having exponentially small measures.

We will see that as $n \longrightarrow +\infty$, the estimate of Theorem 5.1 is optimal.

We will deduce Theorem 5.1 from the following lemma.

(5.2) Lemma. *Let $A \subset I_n$ be a non-empty set. Then, for every $t \geq 0$,*

$$\ln |A| + t \int_{I_n} \operatorname{dist}(x, A) \ d\mu_n \leq n \ln \left(e^{t/2} + e^{-t/2} \right).$$

Proof. We prove the inequality by induction on n and in doing so, we mimic the proof of Theorem 4.2. Let $c(t) = \ln \left(e^{t/2} + e^{-t/2} \right)$.

For $n = 1$, there are two cases: if $|A| = 1$ then the inequality reads $t/2 \leq c(t)$, which is indeed the case. If $|A| = 2$, the inequality reads $\ln 2 \leq c(t)$, which is also the case.

Suppose that $n > 1$. Let us define the facets $I_n^0, I_n^1 \subset I_n$ the sets $A_0, A_1 \subset I_{n-1}$ and $x' \in I_{n-1}$ for $x \in I_n$ as in the proof of Theorem 4.2. Let us also denote $f(x) = \operatorname{dist}(x, A)$, $f_0(x) = \operatorname{dist}(x, A_0)$, and $f_1(x) = \operatorname{dist}(x, A_1)$. So

$$\begin{aligned} &\operatorname{dist}(x, A) = \min\Bigl\{\operatorname{dist}(x', A_1), \quad \operatorname{dist}(x', A_0) + 1\Bigr\} \quad \text{for every point} \quad x \in I_n^1 \quad \text{and} \\ &\operatorname{dist}(x, A) = \min\Bigl\{\operatorname{dist}(x', A_0), \quad \operatorname{dist}(x', A_1) + 1\Bigr\} \quad \text{for every point} \quad x \in I_n^0. \end{aligned}$$

Then

$$\int_{I_n} f(x)\ d\mu_n = \int_{I_n^0} f(x)\ d\mu_n + \int_{I_n^1} f(x)\ d\mu_n.$$

Besides,

$$\begin{aligned} &\int_{I_n^0} f(x)\ d\mu_n \leq \frac{1}{2}\int_{I_{n-1}} f_0\ d\mu_{n-1}, \quad \frac{1}{2} + \frac{1}{2}\int_{I_{n-1}} f_1\ d\mu_{n-1} \quad \text{and} \\ &\int_{I_n^1} f(x)\ d\mu_n \leq \frac{1}{2}\int_{I_{n-1}} f_1\ d\mu_{n-1}, \quad \frac{1}{2} + \frac{1}{2}\int_{I_{n-1}} f_0\ d\mu_{n-1}. \end{aligned}$$

Summarizing,

$$\begin{aligned} \int_{I_n} f\ d\mu_n \leq \min\Biggl\{&\frac{1}{2}\int_{I_{n-1}} f_0\ d\mu_{n-1} + \frac{1}{2}\int_{I_{n-1}} f_1\ d\mu_{n-1}, \\ &\frac{1}{2} + \int_{I_{n-1}} f_0\ d\mu_{n-1}, \quad \frac{1}{2} + \int_{I_{n-1}} f_1\ d\mu_{n-1}\Biggr\}. \end{aligned}$$

Since $|A| = |A_0| + |A_1|$, we can write $|A_0| = \lambda|A|$, $|A_1| = (1-\lambda)|A|$ for some $0 \leq \lambda \leq 1$. In other words,

$$\begin{aligned} &\ln|A| = \ln|A_0| + \ln\frac{1}{\lambda}, \quad \ln|A| = \ln|A_1| + \ln\frac{1}{1-\lambda}, \quad \text{and} \\ &\ln|A| = \frac{1}{2}\ln|A_0| + \frac{1}{2}\ln|A_1| + \frac{1}{2}\ln\frac{1}{\lambda} + \frac{1}{2}\ln\frac{1}{1-\lambda}. \end{aligned}$$

Using the induction hypothesis, we conclude that

$$\begin{aligned} &\ln|A| + \frac{t}{2}\int_{I_{n-1}} f_0\ d\mu_{n-1} + \frac{t}{2}\int_{I_{n-1}} f_1\ d\mu_{n-1} \leq (n-1)c(t) + \frac{1}{2}\ln\frac{1}{\lambda} + \frac{1}{2}\ln\frac{1}{1-\lambda} \\ &\ln|A| + \frac{t}{2} + t\int_{I_{n-1}} f_0\ d\mu_{n-1} \leq (n-1)c(t) + \frac{t}{2} + \ln\frac{1}{\lambda} \quad \text{and} \\ &\ln|A| + \frac{t}{2} + t\int_{I_{n-1}} f_1\ d\mu_{n-1} \leq (n-1)c(t) + \frac{t}{2} + \ln\frac{1}{1-\lambda}, \end{aligned}$$

from which

$$\begin{aligned} \ln|A| + t\int_{I_n} f\ d\mu_n \leq &(n-1)c(t) \\ &+ \min\Biggl\{\frac{1}{2}\ln\frac{1}{\lambda} + \frac{1}{2}\ln\frac{1}{1-\lambda}, \quad \frac{t}{2} + \ln\frac{1}{\lambda}, \quad \frac{t}{2} + \ln\frac{1}{1-\lambda}\Biggr\}. \end{aligned}$$

Hence it all boils down to proving that

$$\min\left\{\frac{1}{2}\ln\frac{1}{\lambda}+\frac{1}{2}\ln\frac{1}{1-\lambda},\quad \frac{t}{2}+\ln\frac{1}{\lambda},\quad \frac{t}{2}+\ln\frac{1}{1-\lambda}\right\}\leq c(t)$$

for all $0 \leq \lambda \leq 1$.

If $\lambda = 0$ or $\lambda = 1$, the minimum is $t/2 \leq c(t)$. If $\lambda = 1/2$, the minimum is $\ln 2 \leq c(t)$. Hence without loss of generality, we may assume that $\lambda > 1/2$. Next, we say that the maximum of the minimum is attained when

$$\frac{1}{2}\ln\frac{1}{\lambda}+\frac{1}{2}\ln\frac{1}{1-\lambda}=\frac{t}{2}+\ln\frac{1}{\lambda}.$$

Indeed, if the left hand side is greater, we can increase the minimum by decreasing λ slightly. If the right hand side is greater, we can increase the minimum by increasing λ slightly. Thus at the maximum point we have the equality, from which $\lambda = e^t/(1+e^t)$ and the value of the minimum is exactly $c(t)$. □

Now we are ready to prove Theorem 5.1.

Proof of Theorem 5.1. Using Lemma 5.2, we conclude that for any $t \geq 0$, we have

$$\begin{aligned}\frac{\ln|A|}{n} &\leq \ln\left(e^{t/2}+e^{-t/2}\right)-\frac{t}{n}\int_{I_n}\operatorname{dist}(x,A)\,d\mu_n\\ &=\ln\left(e^{t/2}+e^{-t/2}\right)+t\left(\rho-\frac{1}{2}\right).\end{aligned}$$

Now we optimize on t and substitute

$$t=\ln(1-\rho)-\ln\rho,$$

which completes the proof. □

PROBLEM. Prove that under the conditions of Theorem 5.1,

$$\begin{gathered}\frac{\ln|A|}{n}\geq \ln 2 - H\left(\frac{1}{2}-\rho\right),\quad\text{where}\\ H(x)=x\ln\frac{1}{x}+(1-x)\ln\frac{1}{1-x}\end{gathered}$$

is the entropy function. The bound is asymptotically sharp as $n \longrightarrow +\infty$.

This problem is harder than our problems have been so far.

Lecture 8. Monday, January 24

Lecture 7 on Friday, January 21, covered the proof of Theorem 5.1 from the previous handout.

6. The Hamming Ball

Let us consider a ball in the Hamming metric in the Boolean cube I_n. All such balls look the same, so we consider one particular, centered at the point $\mathbf{0} = (0, \dots, 0)$.

(6.1) Definition. The *Hamming ball* $B(r)$ of radius r is the set of points in the Boolean cube within Hamming distance r from $\mathbf{0}$, that is,

$$B(r) = \Big\{x : \quad \operatorname{dist}(x, \mathbf{0}) \leq r\Big\}.$$

Equivalently, $B(r)$ consists of the points $(\xi_1, \dots, \xi_n)$, such that

$$\sum_{i=1}^{n} \xi_i \leq r \quad \text{and} \quad \xi_i \in \{0, 1\} \quad \text{for all} \quad i = 1, \dots, n.$$

How many points are there in the Hamming ball? We have

$$|B(r)| = \sum_{k=0}^{r} \binom{n}{k},$$

since to choose a point $x \in B(r)$ we have to choose $k \leq r$ positions, fill them by 1's and fill the remaining $n - k$ positions by 0's.

Here is a useful asymptotic estimate.

(6.2) Lemma. *Let*

$$H(x) = x \ln \frac{1}{x} + (1-x) \ln \frac{1}{1-x} \quad \text{for} \quad 0 \leq x \leq 1$$

be the entropy function. Let us choose a $0 \leq \lambda \leq 1/2$ *and let* B_n *be the Hamming ball in* I_n *of radius* $\lfloor \lambda n \rfloor$*. Then*

(1)
$$\ln |B_n| \leq n H(\lambda);$$

(2)
$$\lim_{n \longrightarrow +\infty} n^{-1} \ln |B_n| = H(\lambda).$$

Proof. Clearly, we may assume that $\lambda > 0$.

To prove (1), we use the Binomial Theorem:

$$1 = (1 + (1 - \lambda))^n = \sum_{k=0}^{n} \binom{n}{k} \lambda^k (1-\lambda)^{n-k} \geq \sum_{0 \leq k \leq \lambda n} \binom{n}{k} \lambda^k (1-\lambda)^{n-k}.$$

Now, since $\lambda \leq 1/2$, we have

$$\frac{\lambda}{1-\lambda} \leq 1, \quad \text{and, therefore,} \quad \lambda^i(1-\lambda)^{n-i} \geq \lambda^j(1-\lambda)^{n-j} \quad \text{for} \quad j \geq i.$$

Continuing the chain of inequalities above, we have

$$1 \geq \sum_{0 \leq k \leq \lambda n} \binom{n}{k} \lambda^{\lambda n}(1-\lambda)^{n-\lambda n} = |B_n| e^{-nH(\lambda)}.$$

Thus $|B_n| \leq e^{nH(\lambda)}$ as desired.

To prove (2), we use (1) and Stirling's formula

$$\ln n! = n \ln n - n + O(\ln n) \quad \text{as} \quad n \longrightarrow +\infty.$$

Letting $m = \lfloor \lambda n \rfloor = \lambda n + O(1)$ as $n \longrightarrow +\infty$, we get

$$\begin{aligned} n^{-1} \ln |B_n| \geq & n^{-1} \ln \binom{n}{m} = n^{-1} \left(\ln n! - \ln m! - \ln(n-m)! \right) \\ = & n^{-1} \left(n \ln n - n - m \ln m + m - (n-m) \ln(n-m) + O(\ln n) \right) \\ = & n^{-1} \big(n \ln n - n - \lambda n \ln(\lambda n) + \lambda n \\ & \quad - (1-\lambda) n \ln(1-\lambda) n + (1-\lambda) n + o(n) \big) \\ = & n^{-1} \left(nH(\lambda) + o(n) \right) = H(\lambda) + o(1). \end{aligned}$$

Together with (1), this concludes the proof. □

(6.3) The Hamming ball as an asymptotic solution to the isoperimetric inequality. Let us fix a $0 \leq \lambda \leq 1/2$. In Theorem 5.1, let us choose $A = B(r)$ with $r = \lfloor \lambda n \rfloor$ as $n \longrightarrow +\infty$.

What is the average distance from a point $x \in I_n$ to $A = B(r)$? If the coordinates of $x = (\xi_1, \ldots, \xi_n)$ contain k ones and $n-k$ zeroes, then $\operatorname{dist}(x, A) = \min\{0, k-r\}$. As follows from Corollary 4.4, for example, a "typical" point $x \in I_n$ contains $n/2 + O(\sqrt{n})$ ones, so the distance $\operatorname{dist}(x, A)$ should be roughly $0.5n - r = n(0.5 - \lambda)$. This is indeed so.

PROBLEM. Let $A = B(r)$ with $r = \lfloor \lambda n \rfloor$ for some $0 \leq \lambda \leq 1/2$. Prove that

$$\lim_{n \longrightarrow +\infty} \frac{1}{n} \int_{I_n} \operatorname{dist}(x, A) \, d\mu_n = \frac{1}{2} - \lambda.$$

On the other hand, as follows by Lemma 6.2,

$$\lim_{n \longrightarrow +\infty} \frac{\ln |A|}{n} = H(\lambda) = \lambda \ln \frac{1}{\lambda} + (1-\lambda) \ln \frac{1}{1-\lambda}.$$

This shows that the inequality of Theorem 5.1 is sharp on Hamming balls.

(6.3) The exact solution to the isoperimetric inequality. As we mentioned, the Hamming ball is also the exact solution to the isoperimetric inequality in the Boolean cube: among all sets of a given cardinality it has the smallest cardinality of the ϵ-neighborhood for any given $\epsilon > 0$, not necessarily small. There is, of course, a catch here: not every positive integer can be the cardinality of a Hamming ball, because $|B(r)|$ is always a certain sum of binomial coefficients. Thus to cover truly all possibilities, we should consider partially filled Hamming balls. This calls for a definition.

(6.3.1) Definition. Let us define the *simplicial order* in the Boolean cube I_n as follows: for $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$, we say that $x < y$ if either $\xi_1 + \ldots + \xi_n < \eta_1 + \ldots + \eta_n$ or $\xi_1 + \ldots + \xi_n = \eta_1 + \ldots + \eta_n$ and for the smallest i such that $\xi_i \neq \eta_i$, we have $\xi_i = 1$ and $\eta_i = 0$.

Given an integer $0 \leq k \leq 2^n$, we can consider the first k elements of the simplicial order in I_n. For example, if $n = 5$ and $k = 12$, then the first 12 elements in the simplicial order are

$$\begin{gathered}(0,0,0,0,0), (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0),\\ (0,0,0,0,1), (1,1,0,0,0), (1,0,1,0,0), (1,0,0,1,0), (1,0,0,0,1),\\ (0,1,1,0,0), (0,1,0,1,0).\end{gathered}$$

For a non-empty set $A \subset I_n$ and a positive integer t, let us define

$$A(t) = \Big\{x \in I_n : \quad \operatorname{dist}(x, A) \leq t\Big\}.$$

PROBLEMS.

1. Suppose that $A = B(r)$ is the Hamming ball of radius r. Prove that $A(t) = B(r+t)$.

2. Check that the Hamming ball is an interval in the simplicial order, that is, the set of all points not exceeding a particular $x \in I_n$ in the simplicial order.

3. Let $B \subset I_n$ be an interval in the simplicial order. Prove that $B(t)$ is an interval in the simplicial order for any positive integer t.

Now, Harper's Theorem.

(6.3.1) Theorem. *Given a non-empty set $A \subset I_n$, let $B \subset I_n$ be the first $|A|$ elements of I_n in the simplicial order. Then*

$$|A(1)| \geq |B(1)|.$$

Although we don't prove Theorem 6.3.1, here a useful corollary.

(6.3.2) Corollary. *Let $A \subset I_n$ be a set such that*

$$|A| \geq \sum_{k=0}^{r} \binom{n}{k}.$$

Then

$$|A(t)| \geq \sum_{k=0}^{r+t} \binom{n}{k} \quad \text{for all} \quad t > 0.$$

Here is another useful corollary.

(6.3.3) Corollary. *Suppose that n is even and let $A \subset I_n$ be a set such that $|A| \geq 2^{n-1}$ (that is, $\mu_n(A) \geq 1/2$). Then*

$$|A(t)| \geq 2^n - \exp\Bigl\{nH\left(\frac{1}{2} - \frac{t}{n}\right)\Bigr\}$$

that is,

$$\mu_n\left(A(t)\right) \geq 1 - \exp\Bigl\{-n\left(\ln 2 - H\left(\frac{1}{2} - \frac{t}{n}\right)\right)\Bigr\}, \quad \text{for} \quad t = 1, \ldots, n/2,$$

where

$$H(x) = x \ln \frac{1}{x} + (1-x) \ln \frac{1}{1-x}, \quad 0 \leq x \leq \frac{1}{2}$$

is the entropy function.

PROBLEM.

1. Deduce Corollaries 6.3.2 and 6.3.3 from Theorem 6.3.1.

Lecture 9. Wednesday, January 26

---

7. The Martingale Method and Its Application to the Boolean Cube

Having the Boolean cube I_n as an example, we describe a general method of obtaining concentration inequalities via *martingales*. Although the inequalities we obtain through the martingale approach are not as sharp as the ones we get via isoperimetric inequalities (exact or asymptotic), the martingale approach is very general, very simple, and allows us to get concentration results in a variety of instances. Besides, it produces estimates for the concentration about the average value of the function (as opposed to the median value), which comes in handy under a variety of circumstances. We start with an abstract definition.

(7.1) Conditional expectation. Let X be a space with a probability measure μ, let $f : X \longrightarrow \mathbb{R}$ be an integrable function and let $\mathcal{F}$ be a σ-algebra of some measurable subsets of X. In other words, $\mathcal{F}$ is a collection of subsets which contains $\emptyset$ and X, if $A \in \mathcal{F}$ then $X \setminus A \in \mathcal{F}$ and $\mathcal{F}$ is closed under operations of taking countable unions and intersections.

A function $h : X \longrightarrow \mathbb{R}$ is called the *conditional expectation* of f with respect to $\mathcal{F}$ and denoted $h = \mathbf{E}(f|\mathcal{F})$ if h is measurable with respect to $\mathcal{F}$ (that is, $h^{-1}(A) \in \mathcal{F}$ for every Borel set $A \subset \mathbb{R}$) and

$$\int_Y h \ d\mu = \int_Y f \ d\mu \quad \text{for all} \quad Y \in \mathcal{F}.$$

Although conditional expectations exist (and unique) under widest assumptions (the Radon-Nikodym Theorem), we will use them in the following simple (but sufficiently general) situation. The space X will be finite (but large, keep the Boolean cube in mind), so that μ assigns some positive real weights to the elements $x \in X$ and the integral is just the sum

$$\int_Y f \ d\mu = \sum_{x \in Y} f(x)\mu(x).$$

The family $\mathcal{F}$ will be a collection of pairwise disjoint subsets, called *blocks*, whose union is X. In this case, the conditional expectation $h = \mathbf{E}(f|\mathcal{F})$ is defined as follows: the value of $h(y)$ is just the average value of f on the block containing y:

$$h(y) = \frac{1}{\mu(Y)} \int_Y f(x) \ d\mu \quad \text{provided} \quad y \in Y \quad \text{and} \quad Y \in \mathcal{F}.$$

To be consistent, let us denote

$$\mathbf{E}(f) = \int_X f \ d\mu.$$

If $\mathcal{F}$ consist of a single block $\{X\}$, the conditional expectation $\mathbf{E}(f|\mathcal{F})$ is just the constant function on X whose value at every point is equal to the average value of f on X. Here are some trivial, yet useful, properties of the conditional expectation:

- We have

$$\mathbf{E}(f) = \mathbf{E}\left(\mathbf{E}(f|\mathcal{F})\right).$$

In words: to compute the average of f, one can first compute the averages of f over each block of the partition $\mathcal{F}$ and then average those averages;

- Similarly, if a partition $\mathcal{F}_2$ refines a partition $\mathcal{F}_1$, that is, if every block of $\mathcal{F}_1$ is a union of some blocks of $\mathcal{F}_2$, then

$$\mathbf{E}(\mathbf{E}(f|\mathcal{F}_2)|\mathcal{F}_1) = \mathbf{E}(f|\mathcal{F}_1).$$

In words: if we average over the smaller blocks and then over the larger blocks, the result is the same as if we average over the larger blocks;

- Suppose that $f(x) \leq g(x)$ for all $x \in X$. Then

$$\mathbf{E}(f|\mathcal{F}) \leq \mathbf{E}(g|\mathcal{F}) \quad \text{pointwise on} \quad X.$$

In words: if f does not exceed g at every point of x, then the average of f does not exceed the average of g over every block of the partition $\mathcal{F}$;

- Suppose that g is constant on every block of the partition $\mathcal{F}$. Then

$$\mathbf{E}(gf|\mathcal{F}) = g\mathbf{E}(f|\mathcal{F}).$$

In words: if every value of the function on a block of the partition $\mathcal{F}$ is multiplied by a constant, the average value of the function on that block gets multiplied by a constant.

(7.2) The idea of the martingale approach. Suppose we have a space X as above and a function $f : X \longrightarrow \mathbb{R}$. Suppose that we have a sequence of partitions $\mathcal{F}_0, \ldots, \mathcal{F}_n$ of X into blocks. That is, each $\mathcal{F}_i$ is just a collection of some pairwise disjoint blocks $Y \subset X$ whose union is X. We also assume that the following holds:

The first partition $\mathcal{F}_0$ consists of a single block that is the set X itself;

The blocks of the last partition $\mathcal{F}_n$ are the singletons $\{x\}$ for $x \in X$;

Each $\mathcal{F}_{i+1}$ refines $\mathcal{F}_i$ for each i, that every block $Y \in \mathcal{F}_i$ is a union of some blocks Y_j from $\mathcal{F}_{i+1}$.

Then, for each $i = 0, \ldots, n$, we have a function $f_i : X \longrightarrow \mathbb{R}$ which is obtained by averaging f on the blocks of $\mathcal{F}_i$. That is, $f_i = \mathbf{E}(f|\mathcal{F}_i)$. Note that $\mathbf{E}(f|\mathcal{F}_0)$ is the constant function equal to the average value of f on X, while $f_n = f$ is the function f itself. Hence the functions f_i kind of interpolate between f and its average. The collection of function $f_0, \ldots, f_n$ is called a *martingale.* The name comes from games of chance, where it refers to a specific strategy of playing roulette. The idea of using the term martingale in a more general context is that it refers to a strategy of playing (averaging) that does not change the final average (expected win).

Our main result is as follows.

(7.3) Theorem. *Let $f_i : X \longrightarrow \mathbb{R}$, $i = 0, \ldots, n$, be a martingale. Suppose that for some numbers $d_1, \ldots, d_n$, we have*

$$|f_i(x) - f_{i-1}(x)| \leq d_i \quad \textit{for} \quad i = 1, \ldots, n \quad \textit{and all} \quad x \in X.$$

Let

$$a = \mathbf{E}(f) = \int_X f \ d\mu,$$

so that $f_0(x) = a$ for all $x \in X$. Let

$$D = \sum_{i=1}^n d_i^2.$$

Then, for any $t \geq 0$, we have

$$\mu\Big\{x : \quad f(x) \geq a + t\Big\} \leq e^{-t^2/2D}$$

and

$$\mu\Big\{x : \quad f(x) \leq a - t\Big\} \leq e^{-t^2/2D}.$$

To prove Theorem 7.3, we need a little technical lemma.

(7.4) Lemma. *Let $f : X \longrightarrow \mathbb{R}$ be a function such that*

$$\int_X f \ d\mu = 0 \quad \textit{and} \quad |f(x)| \leq d \quad \textit{for all} \quad x \in X.$$

Then, for any $\lambda \geq 0$,

$$\int_X e^{\lambda f} \ d\mu \leq \frac{e^{\lambda d} + e^{-\lambda d}}{2} \leq e^{\lambda^2 d^2/2}.$$

Proof. Without loss of generality we assume that $d = 1$. Next, we note that $e^{\lambda t}$ is a convex function and hence its graph lies beneath the chord connecting the points $(-1, e^{-\lambda})$ and $(1, \lambda)$. This gives us the inequality

$$e^{\lambda t} \leq \frac{e^{\lambda} + e^{-\lambda}}{2} + \frac{e^{\lambda} - e^{-\lambda}}{2} t \quad \text{for} \quad -1 \leq t \leq 1.$$

The substitution $t = f(x)$ and integration with respect to x produces

$$\int_X e^{\lambda f} \ d\mu \leq \frac{e^{\lambda} + e^{-\lambda}}{2} \leq e^{\lambda^2/2},$$

where the last inequality follows by comparing the Taylor series expansions. □

Proof of Theorem 7.3. Clearly, it suffices to prove the first inequality as the second follows by switching f to $-f$.

We use the Laplace transform method of Section 2.1. That is, for every $\lambda \geq 0$, we can write

$$\mu\Big\{x \in X : \quad f(x) - a \geq t\Big\} \leq e^{-\lambda t} \int_X e^{\lambda(f-a)} \ d\mu,$$

and our goal is to estimate the integral.

We can write

$$f - a = f_n - a = \sum_{i=1}^{n} (f_i - f_{i-1}).$$

Letting $g_i = f_i - f_{i-1}$, we have

$$\int_X e^{\lambda(f-a)}\ d\mu = \int_X \prod_{i=1}^{n} e^{\lambda g_i}\ d\mu = \mathbf{E}\left(e^{\lambda g_1}\cdots e^{\lambda g_n}\right).$$

Let us denote

$$\mathbf{E}_k(h) = \mathbf{E}(h|\mathcal{F}_k) \quad \text{for any function} \quad h : X \longrightarrow \mathbb{R}.$$

Hence we can write

$$\mathbf{E}\left(e^{\lambda g_1}\cdots e^{\lambda g_n}\right) = \mathbf{E}\left(\mathbf{E}_0\mathbf{E}_1\ldots\mathbf{E}_{n-1}\left(e^{\lambda g_1}\cdots e^{\lambda g_n}\right)\right).$$

In words: we compute the average first by averaging over the finest partition, then over the less fine partition and so on, till we finally average over the partition consisting of the whole space.

Let us take a closer look at

$$\mathbf{E}_0\ldots\mathbf{E}_{k-1}\left(e^{\lambda g_1}\cdots e^{\lambda g_k}\right).$$

We observe that g_i are constants on the blocks of the partition $\mathcal{F}_i$ and hence on the blocks of each coarser partition $\mathcal{F}_j$ with $j \leq i$. In particular, $g_1, \ldots, g_{k-1}$ are constants on the blocks of $\mathcal{F}_{k-1}$, which enables us to write

$$\mathbf{E}_0\ldots\mathbf{E}_{k-1}\left(e^{\lambda g_1}\cdots e^{\lambda g_k}\right) = \mathbf{E}_0\cdots\mathbf{E}_{k-2}\left(e^{\lambda g_1}\cdots e^{\lambda g_{k-1}}\right)\mathbf{E}_{k-1}(e^{\lambda g_k}).$$

Now, what do we know about g_k? First, $|g_k(x)| \leq d_k$ for all x, and second,

$$\mathbf{E}_{k-1}g_k = \mathbf{E}(f_k|\mathcal{F}_{k-1}) - \mathbf{E}(f_{k-1}|\mathcal{F}_{k-1}) = f_{k-1} - f_{k-1} = 0$$

(the function that is identically zero on X). Therefore, by Lemma 7.4,

$$\mathbf{E}_{k-1}e^{\lambda g_k} \leq e^{\lambda^2 d_k^2/2}$$

(the pointwise inequality on X). Hence

$$\mathbf{E}_0\ldots\mathbf{E}_{k-1}\left(e^{\lambda g_1}\cdots e^{\lambda g_k}\right) \leq e^{\lambda^2 d_k^2/2}\mathbf{E}_0\cdots\mathbf{E}_{k-2}\left(e^{\lambda g_1}\cdots e^{\lambda g_{k-1}}\right).$$

Proceeding as above, we conclude that

$$\begin{aligned}\mathbf{E}\left(e^{\lambda g_1}\cdots e^{\lambda g_n}\right) =& \mathbf{E}\left(\mathbf{E}_0\mathbf{E}_1\ldots\mathbf{E}_{n-1}\left(e^{\lambda g_1}\cdots e^{\lambda g_n}\right)\right)\\ \leq& \exp\left\{\frac{\lambda^2}{2}\sum_{k=1}^{n} d_k^2\right\} = e^{\lambda^2 D/2}.\end{aligned}$$

Now we choose $\lambda = t/D$ and conclude that

$$\mu\left\{x \in X : \quad f(x) - a \geq t\right\} \leq e^{-\lambda t}\int_X e^{\lambda(f-a)}\ d\mu \leq e^{-t^2/D}e^{-t^2/2D} = e^{-t^2/2D}$$

as claimed. □

As an application, let us see what kind of a concentration result can we get for the Boolean cube.

(7.5) Theorem. *Let $f : I_n \longrightarrow \mathbb{R}$ be a function such that*

$$|f(x) - f(y)| \leq \operatorname{dist}(x, y) \quad \textit{for all} \quad x, y \in I_n.$$

Let

$$a_f = \int_{I_n} f \ d\mu_n$$

be the average value of f on the Boolean cube. Then, for all $t \geq 0$,

$$\mu_n\Big\{x : \quad f(x) \geq a_f + t\Big\} \leq e^{-2t^2/n} \quad \textit{and} \quad \mu_n\Big\{x : \quad f(x) \leq a_f - t\Big\} \leq e^{-2t^2/n}.$$

Proof. To apply Theorem 7.3, we need to associate a martingale $f_0, \ldots , f_n$ to f. Let $\mathcal{F}_0$ be the partition consisting of a single block, the cube I_n itself, and let $\mathcal{F}_n$ be the partition where the blocks are the points of I_n. To interpolate between $\mathcal{F}_0$ and $\mathcal{F}_n$, let us define $\mathcal{F}_1$ as the partition consisting of the two blocks, the "upper facet"

$$I_n^1 = \Big\{(\xi_1, \ldots , \xi_n) : \quad \xi_n = 1\Big\}$$

and the "lower facet"

$$I_n^0 = \Big\{(\xi_1, \ldots , \xi_n) : \quad \xi_n = 0\Big\}.$$

Since each of the facets looks like the Boolean cube I_{n-1}, we can proceed subdividing them further, finally arriving to $\mathcal{F}_n$. This defines the martingale $f_0, \ldots , f_n$, where $f_i : I_n \longrightarrow \mathbb{R}$ is the function obtained by averaging of f on the blocks of $\mathcal{F}_i$ (where each of the 2^i blocks of $\mathcal{F}_i$ is obtained by fixing the last i coordinates of a point $x \in I_n$). To apply Theorem 7.3, we need to estimate $|f_i - f_{i-1}|$.

Let us estimate $|f_1 - f_0|$. We observe that f_0 is a constant on the whole cube equal to the average value of f on I_n. Next, we observe that f_1 takes two values: if $x \in I_n^1$ then $f(x)$ is the average value of f on the upper facet I_n^1 and if $x \in I_n^0$ then $f(x)$ is the average value of f on the lower facet I_n^0. These two averages cannot be too different: if we switch the last coordinate of $x \in I_n$, we switch between the upper and the lower facets and while we are doing that, the value of f can change by at most 1. This proves that the difference between the average values of f on I_n^1 and on I_n^0 does not exceed 1. Since the average value of f over the whole cube is the average of the averages over the facets, we conclude that we can choose $d_1 = 1/2$ in Theorem 7.3. Just the same, we can choose $d_i = 1/2$ for $i = 1, \ldots , n$, which completes the proof. □

Rescaling, we obtain

$$\mu_n\Big\{x \in I_n : \quad |f(x) - a_f| \geq \epsilon\sqrt{n}\Big\} \leq 2e^{-2\epsilon^2},$$

which is similar in spirit to Theorem 4.5, only instead estimating the deviation from the median m_f, we estimate the deviation from the average a_f.

PROBLEMS.

1. Let S_n be the symmetric group, that is, the group of all bijections $\sigma : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$. Let us make S_n a metric space by introducing the Hamming distance

$$\operatorname{dist}(\sigma, \tau) = |i : \ \sigma(i) \neq \tau(i)|$$

and a probability space by introducing the uniform probability measure $\mu_n(\sigma) = 1/n!$ for all $\sigma \in S_n$. Using the martingale approach, prove that for a 1-Lipschitz function $f : S_n \longrightarrow \mathbb{R}$ and any $t \geq 0$,

$$\mu_n\Big\{x : \quad f(x) \geq a_f + t\Big\} \leq e^{-t^2/2n} \quad \text{and} \quad \mu_n\Big\{x : \quad f(x) \leq a_f - t\Big\} \leq e^{-t^2/2n}.$$

2. Let us modify the measure μ_n on the Boolean cube as follows. Let $p_1, \ldots, p_n$ be numbers such that $0 < p_i < 1$ for $i = 1, \ldots, n$. For $x \in I_n$, $x = (\xi_1, \ldots, \xi_n)$, let us define $\mu_n(\{x\})$ as the product of n factors, where the ith factor is p_i if $\xi_i = 1$ and the ith factor is $1 - p_i$ if $\xi_i = 0$. Prove that under conditions of Theorem 7.5, we have

$$\mu_n\Big\{x : \quad f(x) \geq a_f + t\Big\} \leq e^{-t^2/2n} \quad \text{and} \quad \mu_n\Big\{x : \quad f(x) \leq a_f - t\Big\} \leq e^{-t^2/2n}.$$

Lecture 11. Monday, January 31

Lecture 10 on Friday, January 28, covered the proof of Theorem 7.3 from the previous handout.

8. Concentration in the product spaces. An application: the law of large numbers for bounded functions

Let us apply the martingale approach in the following situation. Suppose that for $i = 1, \ldots, n$ we have a space X_i with a probability measure μ_i and let

$$X = X_1 \times \cdots \times X_n$$

be the direct product of the spaces X_i, that is, the set of all n-tuples $(x_1, \ldots, x_n)$ with $x_i \in X_i$. Let us introduce the product measure $\mu = \mu_1 \times \cdots \times \mu_n$ on X. Thus the measurable sets in X are countable unions of the sets $A = A_1 \times \cdots \times A_n$, where $A_i \subset X_i$ is measurable, and

$$\mu(A) = \mu_1(A_1) \cdots \mu_n(A_n).$$

We also consider the Hamming distance on X:

$$\operatorname{dist}(x, y) = |i : \quad x_i \neq y_i| \quad \text{for} \quad x = (x_1, \ldots, x_n) \quad \text{and} \quad y = (y_1, \ldots, y_n).$$

(8.1) Theorem. *Let $f : X \longrightarrow \mathbb{R}$ be an integrable function and let $d_1, \dots, d_n$ be numbers such that $|f(x) - f(y)| \leq d_i$ if x and y differ in the ith coordinate only. Let*

$$a = \int_X f \ d\mu$$

be the average value of f and let

$$D = \sum_{i=1}^n d_i^2.$$

Then, for any $t \geq 0$

$$\mu\Big\{x : \quad f(x) \geq a + t\Big\} \leq e^{-t^2/2D} \quad and \quad \mu\Big\{x : \quad f(x) \leq a - t\Big\} \leq e^{-t^2/2D}.$$

Proof. Of course, we construct a martingale associated with f. Namely, we define $f_i, i = 0, \dots, n$ as follows. We have $f_0 : X \longrightarrow \mathbb{R}$ is a function which is constant on X, $f_n = f$ and f_i is obtained by integrating f with respect to the last $n - i$ coordinates:

$$f_i(x_1, \dots, x_n) = \int_{X_{i+1} \times \dots \times X_n} f(x_1, \dots, x_n) \ d\mu_{i+1} \cdots d\mu_n.$$

In words: $f_i(x_1, \dots, x_n)$ depends on the first i coordinates only and obtained from f by averaging over the last $n - i$ coordinates. One can view f_i as the conditional expectation $\mathbf{E}(f|\mathcal{F}_i)$, where $\mathcal{F}_i$ is the σ-algebra consisting of the countable unions of the sets of the type

$$A_1 \times \dots \times A_i \times X_{i+1} \times \dots \times X_n \quad \text{where} \quad A_j \subset X_j \quad \text{are measurable} \quad \text{for} \quad j \leq i.$$

If we think of $\mathcal{F}_i$ in terms of partitions and blocks, then the blocks of $\mathcal{F}_i$ consist of singletons of $X_1 \times \dots \times X_i$ multiplied by the whole spaces $X_{i+1} \times \dots \times X_n$. Letting $g_i = f_i - f_{i-1}$, it is pretty clear that

$$|g_i(x)| \leq d_i \quad \text{for} \quad i = 1, \dots, n$$

and that

$$\mathbf{E}(g_i|\mathcal{F}_{i-1}) = \mathbf{E}(f_i|\mathcal{F}_{i-1}) - \mathbf{E}(f_{i-1}|\mathcal{F}_{i-1}) = \int_{X_i} f_i \ d\mu_i - f_{i-1} = f_{i-1} - f_{i-1} = 0,$$

so the whole proof of Theorem 7.3 carries over. □

(8.2) A typical application: the law of large numbers. Typically, Theorem 8.1 is applied under the following circumstances. We have a probability space Y, say, with a measure ν, say, and an integrable function $h : Y \longrightarrow \mathbb{R}$ such that

$$\mathbf{E}h = \int_Y h \, d\nu = a,$$

say. We sample n points $y_1, \ldots , y_n \in Y$ independently at random, compute the sample average

$$\frac{h(y_1) + \ldots + h(y_n)}{n},$$

and ask ourselves how far and how often can it deviate from the average a? We suppose that $0 \leq h(y) \leq d$ for some d and all $y \in Y$.

Let us make n copies $X_1, \ldots , X_n$ of Y and let μ_i be the copy of ν on X_i. To sample n points $y_1, \ldots , y_n \in Y$ "independently at random" is the same as to sample a single point $x = (y_1, \ldots , y_n) \in X$ for $X = X_1 \times \cdots \times X_n$ with respect to the product measure $\mu = \mu_1 \times \cdots \times \mu_n$. Now,

$$f(y_1, \ldots , y_n) = \frac{h(y_1) + \ldots + h(y_n)}{n},$$

so if we change only one coordinate, the value of f changes by not more than d/n. Hence we can apply Theorem 8.1 with

$$D = \sum_{i=1}^n \left(\frac{d}{n}\right)^2 = \frac{d^2}{n}.$$

This gives us that

$$\mu_n\Big\{x : \quad f(x) \geq a + t\Big\} \leq e^{-nt^2/2d^2} \quad \text{and} \quad \mu_n\Big\{x : \quad f(x) \leq a - t\Big\} \leq e^{-nt^2/2d^2}.$$

Rescaling $t = \epsilon d/\sqrt{n}$, we get that the probability that

$$\frac{h(y_1) + \ldots + h(y_n)}{n}$$

deviates from a by more than $\epsilon d/\sqrt{n}$ does not exceed $2e^{-\epsilon^2/2}$. Taking $\epsilon = 3$, for example, we conclude that the probability that the sample average of h over n random samples does not deviate from the average by $3d/\sqrt{n}$ is at least 97%. Hence, if we take $n = 10,000$, say, the probability that the sample average will not deviate from the average by more than $0.03d$ will be at least 97%.

(8.3) Another typical application: serving many functions at the same time. Let us adjust the situation of Section 8.2 a bit. Suppose that instead of one function h we have N functions $h_1, \ldots, h_N$, having averages $a_1, \ldots, a_N$, and such that $0 \leq h_i(y) \leq d$ for all $y \in Y$ and all $i = 1, \ldots, N$. We sample n points $y_1, \ldots, y_n$ and compute the sample average for every function h_i:

$$\frac{h_i(y_1) + \ldots + h_i(y_n)}{n} \quad \text{for} \quad i = 1, \ldots, N$$

on the same set of samples. We ask ourselves, how many points should we sample so that with high probability *each* of the N sample averages will be reasonably close to the corresponding average a_i. Applying the estimate, we conclude that the probability to get within t from the expectation a_i for any particular sample average is $1-2e^{-nt^2/2d^2}$. Therefore, the probability that every single sample average is within ϵd from the expectation a_i is at least $1 - 2Ne^{-n\epsilon^2/2}$. To make that reasonably high, we should choose $n \sim \epsilon^{-2} \ln N$. In other words, given ϵ and d, the number of samples should be only logarithmic in the number N of functions.

For example, to get the probability at least $2/3$, we can take any $n > 2\epsilon^{-2} \ln(6N)$. Thus if $N = 10^9$ and $\epsilon = 0.1$ (so that we stay within 10% of d from each of the 10^9 averages), we can choose $n = 4,504$.

PROBLEMS

1. Suppose that we require $|h(y)| \leq d$ in Sections 8.2 and 8.3 above (that is, we allow h to be negative). Check that the inequalities read

$$\mu_n\Bigl\{x : \quad f(x) \geq a + t\Bigr\} \leq e^{-nt^2/8d^2} \quad \text{and} \quad \mu_n\Bigl\{x : \quad f(x) \leq a - t\Bigr\} \leq e^{-nt^2/8d^2}.$$

2. Suppose that $h_1, \ldots, h_n : Y \longrightarrow \mathbb{R}$ are functions such that $0 \leq h_i \leq 1$ for $i = 1, \ldots, n$ and let $a_i = \mathbf{E} h_i$. For n points $y_1, \ldots, y_n$ sampled independently at random, let

$$f(y_1, \ldots y_n) = \frac{h_1(y_1) + \ldots + h_n(y_n)}{n}$$

and let $a = (a_1 + \ldots + a_n)/n$. Prove that

$$\mu_n\Bigl\{x : \quad f(x) \geq a + t\Bigr\} \leq e^{-nt^2/2d^2} \quad \text{and} \quad \mu_n\Bigl\{x : \quad f(x) \leq a - t\Bigr\} \leq e^{-nt^2/2d^2}.$$

9. How to sharpen martingale inequalities?

Although often quite helpful, martingale inequalities are rarely sharp. One of the rare examples where we can figure out optimal estimates is the Boolean cube.

(9.1) Example: tossing a fair coin. Let $I_n = \{0, 1\}^n$ be the Boolean cube with the uniform probability measure $\mu_n\{x\} = 2^{-n}$. Let us consider $f : I_n \longrightarrow \mathbb{R}$ defined by $f(\xi_1, \ldots, \xi_n) = \xi_1 + \ldots + \xi_n$. That is, we toss a fair coin n times and count the number of heads, say. Then the average value a_f of f is $n/2$ (why?) and f is 1-Lipschitz. Hence Theorem 7.5 gives us

$$\mu_n\Big\{x: \quad f(x) - n/2 \leq -t\Big\} \leq e^{-2t^2/n}.$$

Suppose that t grows linearly with n, $t = \lambda n$ for some $0 < \lambda < 1/2$. Then the estimate becomes

$$\mu_n\Big\{x: \quad f(x) - n/2 \leq -\lambda n\Big\} \leq e^{-2n\lambda^2}.$$

Suppose that n is even, $n = 2m$. Suppose that t is integer. Then the points $x \in I_n$ where $f(x) - m \leq -t$ are the points with x having $0, 1, \ldots, m-t$ zero coordinates. Hence

$$\mu_n\Big\{x: \quad f(x) - m \leq -t\Big\} = 2^{-n} \sum_{k=0}^{m-t} \binom{n}{k} = 2^{-n}|B(m-t)|,$$

where $B(m-t)$ is the Hamming ball of radius $m-t$. If $t \approx \lambda n$, by Part (2) of Lemma 6.2, we can estimate the right hand side by

$$2^{-n} \exp\Big\{nH\left(\frac{1}{2} - \lambda\right)\Big\},$$

where H is the entropy function. Assuming that $\lambda \approx 0$, we get

$$H\left(\frac{1}{2} - \lambda\right) = \left(\frac{1}{2} - \lambda\right) \ln \frac{1}{1-2\lambda} + \left(\frac{1}{2} + \lambda\right) \ln \frac{1}{1+2\lambda} + \ln 2.$$

Using that

$$\ln(1+x) = x - \frac{x^2}{2} + O(x^3),$$

we get

$$H\left(\frac{1}{2} - \lambda\right) = \ln 2 - 2\lambda^2 + O(\lambda^3).$$

Hence, for $\lambda \approx 0$, we true estimate

$$\mu_n\Big\{x: \quad f(x) - m \leq -\lambda n\Big\} \approx e^{-2n\lambda^2}$$

agrees with the one given by Theorem 7.5.

However, as λ grows, the margin between the martingale bound and the true bound starts to widen since

$$H\left(\frac{1}{2} - \lambda\right) - \ln 2$$

starts to deviate from $-2\lambda^2$. But even for the endpoint $\lambda = 1/2$, the difference is not too great: the martingale bound estimates the probabilities by $e^{-0.5n}$, while the true value is $e^{-(\ln 2)n} \approx e^{-0.69n}$.

Lecture 12. Wednesday, February 2

---

9. How to sharpen martingale inequalities? Continued

(9.2) Example: tossing an unfair coin. Let us choose a number $0 < p < 1$ and let $q = 1 - p$. Let $I_n = \{0, 1\}^n$ be the Boolean cube and let us define the measure μ_n on I_n by $\mu_n\{x\} = p^k q^{n-k}$, where k is the number of 1's among the coordinates of I_n (check that this indeed defines a probability measure on I_n). Let us consider $f : I_n \longrightarrow \mathbb{R}$ defined by $f(\xi_1, \dots , \xi_n) = \xi_1 + \ldots + \xi_n$. That is, we have a coin which turns up heads with probability p, we toss it n times and count heads. The average value a_f of f is np (why?) and f is 1-Lipschitz. It follows from Theorem 7.3 (cf. also Problem 2 after Theorem 7.5) that

$$\mu_n\Bigl\{x : \quad f(x) - np \leq -t\Bigr\} \leq e^{-t^2/2n} \quad \text{and} \quad \mu_n\Bigl\{x : \quad f(x) - np \geq t\Bigr\} \leq e^{-t^2/2n}.$$

After some thought, we conclude that it is natural to scale $t = \alpha np$ for some $\alpha \geq 0$, which gives us

$$\mu_n\Bigl\{x : \quad f(x) - np \leq -\alpha np\Bigr\} \leq e^{-np^2\alpha^2/2} \quad \text{and}$$
$$\mu_n\Bigl\{x : \quad f(x) - np \geq \alpha np\Bigr\} \leq e^{-np^2\alpha^2/2}.$$

These estimates do not look too good for small p. We would rather have something of the order of $e^{-np\alpha^2/2}$. Can we make it?

In fact, we can, if the examine the proof of Theorem 7.3 and its application to the Boolean cube. Let us construct the martingale $f_0, \dots , f_n : I_n \longrightarrow \mathbb{R}$ as in the proof of Theorem 7.5, repeatedly cutting the cube into the "upper" and "lower" facets I_n^1 and I_n^0 and averaging f over the facets. Thus, f_0 is the constant function on I_n, which at every point $x \in I_n$ is equal to the average value of f on I_n. Also, $f_n = f$. Furthermore, for every k, the function f_k depends on the first k coordinates only. That is, $f_k(\xi_1, \dots , \xi_n)$ is equal to the average value of f on the set of points of the face of the Boolean cube consisting of the points where the first k coordinates have the prescribed values $\xi_1, \dots , \xi_k$. In terms of the coin, $f_k(\xi_1, \dots , \xi_n)$ is the expected number of heads on n tosses given that the first k tosses resulted in $\xi_1, \dots , \xi_k$, where $\xi_i = 1$ if we got heads at the ith toss and $\xi_i = 0$ if we got tails. Hence we conclude that

$$f_k(\xi_1, \dots , \xi_n) = (n - k)p + \sum_{i=1}^{k} \xi_i.$$

Indeed, after k tosses, we got $\xi_1 + \ldots + \xi_k$ heads and $n - k$ tosses left. Since the coin has no memory, the expected number of heads obtained in the course of those $n - k$ tosses is $(n - k)p$. Going back to the proof of Theorem 7.3, we should refrain from using Lemma 7.4 as too crude and compute $\mathbf{E}_{k-1}\left(e^{\lambda g_k}\right)$ directly. Recall that $g_k = f_k - f_{k-1}$. Hence g_k is defined by a very simple formula:

$$g_k(\xi_1, \dots , \xi_n) = \xi_k - p.$$

Consequently,

$$e^{\lambda g_k(x)} = \begin{cases} e^{\lambda q} & \text{if } \xi_k = 1 \\ e^{-\lambda p} & \text{if } \ \xi_k = 0 \end{cases} \quad \text{where} \quad x = (\xi_1, \ldots, \xi_n).$$

Since $\mathbf{E}_{k-1}\left(e^{\lambda g_k}\right)$ is obtained by averaging over the kth coordinate, we conclude that $\mathbf{E}_{k-1}\left(e^{\lambda g_k}\right)$ is the constant function equal to $pe^{\lambda q} + qe^{-\lambda p}$. Therefore, in the proof of Theorem 7.3, we have

$$\int_{I_n} e^{\lambda(f-a)} \, d\mu_n = \left(pe^{\lambda q} + qe^{-\lambda p}\right)^n,$$

which gives us the bound:

$$\mu_n\Bigl\{x \in I_n : \quad f(x) - np \geq t\Bigr\} \leq e^{-\lambda t}\left(pe^{\lambda q} + qe^{-\lambda p}\right)^n.$$

To obtain an inequality in another direction, by the Laplace transform method, we have

$$\mu_n\Bigl\{x \in I_n : \quad f(x) - np \leq -t\Bigr\} \leq e^{-\lambda t} \int_{I_n} e^{\lambda(a-f)} \, d\mu_n.$$

It remains to notice that $n-a$ and $n-f$ are obtained from a and f respectively by switching p and q. Therefore,

$$\mu_n\Bigl\{x \in I_n : \quad f(x) - np \leq -t\Bigr\} \leq e^{-\lambda t}\left(qe^{\lambda p} + pe^{-\lambda q}\right)^n.$$

Optimizing on λ, we get in the first inequality and the second inequality

$$\frac{t}{npq} = 1 - e^{-\lambda}, \quad \text{so} \quad \lambda = -\ln\left(1 - \frac{t}{npq}\right).$$

Note that we should assume that $t < npq$, which we may or may not want to do.

Now, if $t = \alpha npq$ for sufficiently small $\alpha > 0$ then the optimal $\lambda \approx \alpha$. Thus it makes sense to substitute $\lambda = \alpha$, optimal or not. This gives us the bounds

$$\mu_n\Bigl\{x \in I_n : \quad f(x) - np \geq \alpha npq\Bigr\} \leq \Bigl(pe^{\alpha q(1-\alpha p)} + qe^{-\alpha p(1+\alpha q)}\Bigr)^n \quad \text{and}$$
$$\mu_n\Bigl\{x \in I_n : \quad f(x) - np \leq -\alpha npq\Bigr\} \leq \Bigl(qe^{\alpha p(1-\alpha q)} + pe^{-\alpha q(1+\alpha p)}\Bigr)^n.$$

If $p = o(1)$ and $\alpha \approx 0$ then both bounds are of about $(1 - \alpha^2 p/2)^n \approx e^{-\alpha^2 pn/2}$ as we wanted.

Lecture 13. Friday, February 4

---

10. Concentration and Isoperimetry

We saw in Section 4 how to deduce concentration from isoperimetric inequalities. We worked with the Boolean cube as an example, but the construction is fairly general. Namely, if X is a metric space with a probability measure μ such that t-neighborhoods

$$A(t) = \Big\{x : \quad \operatorname{dist}(x, A) \leq t\Big\}$$

of sets $A \subset X$ with $\mu(A) \geq 1/2$ have a large measure, say,

$$\mu\left(A(t)\right) \geq 1 - e^{-ct^2}$$

for some constant $c > 0$, then 1-Lipschitz functions $f : X \longrightarrow \mathbb{R}$ concentrate around their median m_f. To show that, we consider the sets

$$A_+ = \Big\{x : \quad f(x) \geq m_f\Big\} \quad \text{and} \quad A_- = \Big\{x : \quad f(x) \leq m_f\Big\}.$$

Then $\mu(A_-), \mu(A_+) \geq 1/2$ and so $\mu\left(A_-(t)\right) \geq 1 - e^{-ct^2}$ and $\mu\left(A_+(t)\right) \geq 1 - e^{-ct^2}$. Therefore,

$$\mu\left(A_-(t) \cap A_+(t)\right) \geq 1 - 2e^{-ct^2}.$$

Furthermore, we have

$$|f(x) - m_f| \leq t \quad \text{for} \quad x \in A_-(t) \cap A_+(t).$$

Hence

$$\mu\Big\{x : \quad |f(x) - m_f| \leq t\Big\} \geq 1 - 2e^{-ct^2}.$$

It works in the other direction too. Suppose that we have a concentration result for 1-Lipschitz functions f. That is, we have

$$\mu\Big\{x : \quad f(x) \geq a_f + t\Big\} \leq e^{-ct^2} \quad \text{and} \quad \mu\Big\{x : \quad f(x) \leq a_f - t\Big\} \leq e^{-ct^2}$$

for some constant $c > 0$, where a_f is the average value of f. Then sets A with $\mu(A) \geq 1/2$ have large t-neighborhoods. To see that, let us consider the function

$$f(x) = \operatorname{dist}(x, A).$$

Let us choose $\delta > 0$ such that $e^{-c\delta^2} < 1/2$. Then there exists an $x \in A$ such that $f(x) > a_f - \delta$. But $f(x) = 0$ for $x \in A$, which implies that $a_f < \delta$. Therefore,

$$\mu\Big\{x : \quad \operatorname{dist}(x, A) \geq \delta + t\Big\} \leq 1 - e^{-ct^2},$$

that is, $\mu\left(A(\delta + t)\right) \geq 1 - e^{-ct^2}$.

Therefore, martingale concentration results imply some kind of isoperimetric inequalities. Consider, for example, Theorem 8.1. It describes the following situation. For $i = 1, \ldots, n$ we have a space X_i with a probability measure μ_i. We let

$$X = X_1 \times \cdots \times X_n \quad \text{and} \quad \mu = \mu_1 \times \cdots \times \mu_n.$$

We consider the Hamming distance on X:

$$\operatorname{dist}(x, y) = |i : \quad x_i \neq y_i| \quad \text{for} \quad x = (x_1, \ldots, x_n) \quad \text{and} \quad y = (y_1, \ldots, y_n).$$

Given a set $A \subset X$ such that $\mu(A) \geq 1/2$, we consider the function $f : X \longrightarrow \mathbb{R}$ defined by $f(x) = \operatorname{dist}(x, A)$ (we assume it is measurable). Then we can choose $d_i = 1$ in Theorem 8.1, so $D = n$. Hence we define δ from the inequality $e^{-\delta^2/2n} > 1/2$, which gives us $\delta < \sqrt{n 2 \ln 2}$. Therefore, $a_f \leq \sqrt{n 2 \ln 2}$. This gives us the isoperimetric inequality

$$\mu \left(A(t + \sqrt{n 2 \ln 2}) \right) \geq 1 - e^{-t^2/2n}.$$

Rescaling $t = (\epsilon - \sqrt{2 \ln 2})\sqrt{n}$ for $\epsilon > \sqrt{2 \ln 2}$, we have

$$\mu \Bigl\{ x : \quad \operatorname{dist}(x, A) \leq \epsilon \sqrt{n} \Bigr\} \geq 1 - \exp \Bigl\{ -\frac{(\epsilon - \sqrt{2 \ln 2})^2}{2} \Bigr\}.$$

PROBLEM. Let $f : I_n \longrightarrow \mathbb{R}$ be a 1-Lipschitz function on the Boolean cube I_n endowed with the standard probability measure $\mu_n\{x\} = 2^{-n}$. Let a_f be the average value of f and let m_f be the median of f. Prove that

$$|a_f - m_f| \leq \sqrt{\frac{n \ln 2}{2}},$$

which improves the bound of Problem 1 of Section 5.

In view of what's been said, an isoperimetric inequality is a natural way to supplement a martingale bound. A counterpart to the martingale concentration of Theorem 8.1 is an isoperimetric inequality for general product spaces proved by M. Talagrand.

It turns out that the estimate of Theorem 4.2 holds in this general situation.

(10.1) Theorem. *Let (X_i, μ_i) be probability spaces and let $X = X_1 \times \cdots \times X_n$ be the product space with the product measure $\mu = \mu_1 \times \cdots \times \mu_n$. Let $A \subset X$ be a non-empty set. Then, for $f(x) = \operatorname{dist}(x, A)$ and $t > 0$ we have*

$$\int_X e^{tf} \, d\mu \leq \frac{1}{\mu(A)} \left(\frac{1}{2} + \frac{e^t + e^{-t}}{4} \right)^n \leq \frac{e^{t^2 n/4}}{\mu(A)}.$$

Disclaimer: all sets and functions encountered in the course of the proof are assumed to be measurable. This is indeed so in all interesting cases (X finite or A compact, and so on).

Let us denote

$$c(t) = \frac{1}{2} + \frac{e^t + e^{-t}}{4}.$$

The proof (due to M. Talagrand) is by induction on n and is based on a lemma (also due to M. Talagrand), which simultaneously takes care of the case $n = 1$.

(10.2) Lemma. *Let Y be a space with a probability measure ν and let $g : Y \longrightarrow \mathbb{R}$ be an integrable function such that $0 \leq g(y) \leq 1$ for all y. Then, for any $t \geq 0$*

$$\left(\int_Y \min\left\{ e^t, \frac{1}{g} \right\} \, d\nu \right) \left(\int_Y g \, d\nu \right) \leq \frac{1}{2} + \frac{e^t + e^{-t}}{4} = c(t).$$

Proof. First, we note that if we replace g by $\max\{g, e^{-t}\}$, the first integral does not change, while the second can only increase. Thus we have to prove that if $e^{-t} \leq g(y) \leq 1$ for all y then

$$\left(\int_Y \frac{1}{g} \, d\nu \right) \left(\int_Y g \, d\nu \right) \leq c(t). \tag{10.2.1}$$

Let us assume ν does not have atoms, that, is, for every measurable $B \subset Y$ there is a measurable $C \subset B$ with $\nu(C) = \nu(B)/2$. Let us consider $g \in L^\infty(Y, \nu)$, let us fix $e^{-t} \leq b \leq 1$ and let us consider the set $G_b \subset L^\infty(Y, \nu)$ of functions g such that $e^{-t} \leq g \leq 1$ and

$$\int_Y g \, d\nu = b.$$

Thus G_b is a weak compact* set, so the function

$$\phi : \quad g \longmapsto \int_Y \frac{1}{g} \, d\nu$$

attains its maximum on G_b. The next observation is that G_b is convex, that ϕ is convex, that G_b is the closed convex hull of the set of its extreme points (in the weak* topology) and that the extreme points of G_b consist of the functions g such that $g = e^{-t}$ or $g = 1$ almost everywhere (this is the point where we use that ν has no atoms: if $e^{-t} + \epsilon < g < 1 - \epsilon$ on a set of positive measure, we can perturb g a bit, thus representing as a midpoint of two functions $g_1, g_2 \in G_b$). Hence the maximum of ϕ is attained at a function g with $g = e^{-t}$ or $g = 1$ almost everywhere. Hence is suffices to check (10.2.1) for such functions g. We have

$$g(y) \in \left\{ e^{-t}, 1 \right\} \quad \text{and} \quad \int_Y g \, d\nu = b \quad \text{implies} \quad \int_Y \frac{1}{g} \, d\nu = e^t - be^t + 1.$$

Hence (10.2.1) reduces to the inequality

$$b(e^t - be^t + 1) \leq c(t) \quad \text{for} \quad e^{-t} \leq b \leq 1.$$

But the left hand side is a quadratic function of b. The maximum of the left hand side occurs at $b = (1 + e^{-t})/2$ and is equal to $c(t)$.

This completes the proof under the additional assumption that ν has no atoms. For a general ν, we note that the probability space Y is not so important. Introducing the cumulative distribution function F of g, we can write (10.2.1) as

$$\left(\int_{\mathbb{R}} \frac{1}{x} \, dF\right)\left(\int_{\mathbb{R}} x \, dF\right).$$

It remains to notice that F can be arbitrarily well approximated by a continuous cumulative distribution functions and that such functions correspond to measures without atoms. □

Now we can prove Theorem 10.1.

Proof of Theorem 10.1. We proceed by induction on n. For $n = 1$, let $g(x) = 1 - f(x)$. Thus $g(x) = 1$ if $x \in A$ and $g(x) = 0$ if $x \notin A$.

$$\left(\int_{X_1} e^{tf} \, d\mu_1\right)\mu_1(A) = \left(\int_{X_1} \min\left\{e^t, \frac{1}{g}\right\} \, d\mu_1\right)\left(\int_{X_1} g \, d\mu_1\right)$$

and the result follows by Lemma 10.2.

Suppose that $n > 1$. For $z \in X_n$, let us define

$$A_z = \Big\{(x_1, \ldots, x_{n-1}) : \quad (x_1, \ldots, x_{n-1}, z) \in A\Big\}$$

and let B be the projection of A onto the first $(n-1)$ coordinates:

$$B = \Big\{(x_1, \ldots, x_{n-1}) : \quad (x_1, \ldots, x_{n-1}, z) \in A \quad \text{for some} \quad z \in X_n\Big\}.$$

Hence $A_z, B \subset X_1 \times \cdots \times X_{n-1}$. Let us denote $Y = X_1 \times \cdots \times X_{n-1}$ and $\nu = \mu_1 \times \cdots \times \mu_{n-1}$. We denote $y = (x_1, \ldots, x_{n-1}) \in Y$. By Fubini's theorem,

$$\int_X e^{tf} \, d\mu = \int_{X_n} \Big(\int_Y e^{tf(y,z)} \, d\nu(y)\Big) \, d\mu_n(z).$$

Now,

$$\operatorname{dist}\big((y,z), A\big) \leq \operatorname{dist}(y, A_z) \quad \text{and} \quad \operatorname{dist}\big((y,z), A\big) \leq \operatorname{dist}(y, B) + 1.$$

Therefore, by the induction hypothesis,

$$\int_Y e^{tf(y,z)}\ d\nu(y) \le \int_Y e^{t\,\mathrm{dist}(y,A_z)}\ d\nu(y) \le \frac{1}{\nu(A_z)} c^{n-1}(t)$$

and

$$\int_Y e^{tf(y,z)}\ d\nu(y) \le \int_Y e^{t\,\mathrm{dist}(y,B)}\ d\nu \le \frac{e^t}{\nu(B)} c^{n-1}(t).$$

Besides, by Fubini's Theorem

$$\int_{X_n} \nu(A_z)\ d\mu_n(z) = \mu(A).$$

Let us define a function $g : X_n \longrightarrow X$ by $g(z) = \nu(A_z)/\nu(B)$. Since $A_z \subset B$ for all z, we have $\nu(A_z) \le \nu(B)$ and $0 \le g(z) \le 1$ for all z. Moreover,

$$\int_{X_n} g\ d\mu_n = \frac{\mu(A)}{\nu(B)}.$$

Applying Lemma 10.2, we have

$$\begin{aligned}\int_X e^{tf}\ d\mu \le & c^{n-1}(t) \int_{X_n} \min\Big\{\frac{1}{\nu(A_z)},\ \frac{e^t}{\nu(B)}\Big\}\ d\mu_n(z) \\ = & \frac{c^{n-1}(t)}{\nu(B)} \int_{X_n} \min\Big\{\frac{\nu(B)}{\nu(A_z)},\ e^t\Big\}\ d\mu_n(z) \\ = & \frac{c^{n-1}(t)}{\nu(B)} \int_{X_n} \min\Big\{\frac{1}{g},\ e^t\Big\}\ d\mu_n \\ \le & \frac{c^n(t)}{\nu(B)} \left(\int_{X_n} g\ d\mu_n\right)^{-1} = \frac{c^n(t)}{\mu(A)},\end{aligned}$$

which completes the proof of the first inequality. The second inequality is proved in Corollary 4.3. □

Thus as in Section 4 (cf. Corollary 4.4 and Theorem 4.5) we get the isoperimetric inequality in the product space.

(10.3) Corollary. *Let $A \subset X_1 \times \cdots \times X_n$ be a non-empty set. Then, for any $\epsilon > 0$, we have*

$$\mu\Big\{x \in I_n : \quad \mathrm{dist}(x, A) \ge \epsilon\sqrt{n}\Big\} \le \frac{e^{-\epsilon^2}}{\mu(A)}.$$

Similarly, we get concentration about the median for Lipschitz functions on the product space.

(10.4) Theorem. *Let (X_i, μ_i) be probability spaces, let $X = X_1 \times \cdots \times X_n$ and $\mu = \mu_1 \times \cdots \times$ be the product probability space. Let $f : X \longrightarrow \mathbb{R}$ be a function such that $|f(x) - f(y)| \leq 1$ if x and y differ in at most 1 coordinate and let m_f be the median of f. Then for all $\epsilon > 0$,*

$$\mu\Big\{x \in X : \quad |f(x) - m_f| \geq \epsilon\sqrt{n}\Big\} \leq 4e^{-\epsilon^2}.$$

Lecture 15. Wednesday, February 9

Lecture 14 on Monday, February 7, covered the proof of Theorem 10.1 from the previous handout.

11. Why do we care: some examples leading to discrete measure concentration questions

A rich source of questions regarding concentration inequalities is the theory of random graphs. Suppose we have n points labeled $1, \ldots, n$ and we connect each pair (i, j) by an edge at random with some probability p, independently of others. Thus we get a random graph G and we may start asking about its various characteristics. For example, how many connected components may it have? What is the chromatic number of G, that is, the minimum number of colors we need to color the vertices of G so that the endpoints of every edge have different colors? What is the clique number of G, that is the maximum cardinality of a subset S of vertices of G such that every two vertices from S are connected by an edge?

Let $m = \binom{n}{2}$ and let $I_m = \{0, 1\}^m$ be the corresponding Boolean cube. We interpret a point $x = (\xi_1, \ldots, \xi_m)$ as a graph G, where $\xi_k = 1$ if the corresponding pair of vertices is connected by an edge and $\xi_k = 0$ otherwise. All the above characteristics (the number of connected components, the chromatic number, the clique number) are represented by a function $f : I_m \longrightarrow \mathbb{N}$. Moreover, in all of the above examples, the function is 1-Lipschitz (why?), so it must concentrate for large n.

An example of an interesting function that is not 1-Lipschitz is the number of triangles in the graph G.

12. Large Deviations and Isoperimetry

A natural question is to ask whether bounds in Corollary 10.3 and Theorem 10.4 can be sharpened, for instance, when ϵ is large. That is, given a set $A \subset X$ with, say, $\mu(A) \approx 1/2$, we want to estimate more accurately

$$\mu\Big\{x : \quad \operatorname{dist}(x, A) \geq t\Big\},$$

when t is large. Since this measure is small anyway, the correct scale is logarithmic, so we will be estimating

$$\ln \mu\Big\{x : \quad \operatorname{dist}(x, A) \geq t\Big\}.$$

Equivalently, we may think of a small set

$$B = \Big\{x : \quad \operatorname{dist}(x, A) \geq t\Big\}$$

and ask ourselves how large a neighborhood of B should we take so that it will take approximately half of X:

$$\mu\Big\{x : \quad \operatorname{dist}(x, B) \leq s\Big\} \approx \frac{1}{2}.$$

Questions of this kind are called the *large deviation questions*.

To do that, we would need sharper isoperimetric inequalities in product spaces. Unlike in the case of the standard Boolean cube where such exact isoperimetric inequalities are known (cf. Section 6), in most product spaces no exact isoperimetric inequalities are known. However, *asymptotic* isoperimetric inequalities in the product spaces have been recently discovered. To be able to discuss them, we need some useful notion.

(12.1) The log-moment function. Let X, μ be a probability space and let $f :\longrightarrow \mathbb{R}$ be a function. The *log-moment* function $L_f : \mathbb{R} \longrightarrow \mathbb{R}$ associated to f is defined as follows:

$$L_f(\lambda) = \ln \int_X e^{\lambda f} \ d\mu = \ln \mathbf{E} e^{\lambda f}.$$

If the integral diverges, we let $L_f(\lambda) = +\infty$. The function L_f is also known as the *cumulant generating function* of f.

Here are some useful properties of L_f. We have

$$L_f(0) = \ln 1 = 0.$$

Since $e^{\lambda x}$ is a convex function, by Jensen's inequality

$$L_f(\lambda) \geq \ln e^{\mathbf{E}\lambda f} = \lambda \mathbf{E} f.$$

In particular, if $\mathbf{E} f = 0$ (which we will soon assume), then L_f is non-negative. Next,

$$L_f' = \frac{\mathbf{E}\left(f e^{\lambda f}\right)}{\mathbf{E}\left(e^{\lambda f}\right)} \quad \text{and} \quad L_f'' = \frac{\mathbf{E}\left(f^2 e^{\lambda f}\right)\mathbf{E}\left(e^{\lambda f}\right) - \mathbf{E}^2\left(f e^{\lambda f}\right)}{\mathbf{E}^2\left(e^{\lambda f}\right)},$$

provided all integrals converge, which we assume. By Cauchy-Schwarz inequality,

$$\mathbf{E}^2\left(f e^{\lambda f}\right) = \mathbf{E}\left(f e^{\lambda f/2} e^{\lambda f/2}\right) \leq \mathbf{E}\left(f^2 e^{\lambda f}\right)\mathbf{E}\left(e^{\lambda f}\right),$$

from which $L_f'' \geq 0$ and L_f is convex. Similarly, this can be extracted directly from the definition. Likewise, for any fixed λ, the function $f \longmapsto L_f(\lambda)$ is convex.

(12.2) The rate function. Let $g : \mathbb{R} \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a convex function. The function

$$g^*(t) = \sup_{\lambda \in \mathbb{R}} \Big(t\lambda - g(\lambda) \Big)$$

is called the *Legendre-Fenchel* transform of g. Thus $g^* : \mathbb{R} \longrightarrow \mathbb{R} \cup \{+\infty\}$. Function g^* is also called the *conjugate* to g. If g and g^* are proper, that is, do not acquire the value of $+\infty$, there is a remarkable duality $(g^*)^* = g$. If g is finite in a neighborhood of $\lambda = 0$, then g^* is finite in a neighborhood of $t = 0$.

For example, if $g(\lambda) = \lambda^2$, then $g^*(t) = t^2/4$ and if $g(\lambda) = \lambda^2/4$ then $g^*(t) = t^2$.

The Legendre-Fenchel transform $R_f(t) = L_f(\lambda)^*$ of the log-moment function L_f is called the *rate function* of f.

The importance of the rate function is given by the following result. It is known as the *Large Deviations* Theorem, and, apparently, goes back to Cramér.

(12.3) Theorem. *Let X, μ be a probability space, let $f : X \longrightarrow \mathbb{R}$ be a function and let $t \in \mathbb{R}$ be a number such that*

(1)

$$\mathbf{E} f = \int_X f \ d\mu = a_f \quad \textit{exists}$$

(2)

$$L_f(\lambda) = \int_X e^{\lambda f} \ d\mu \quad \textit{is finite in a neighborhood of} \quad \lambda = 0;$$

(3)

$$t > a_f \quad \textit{and} \quad \mu \Big\{ x : \quad f(x) > t \Big\} > 0.$$

For a positive integer n, let us consider the product space $X_n = X \times \ldots \times X$ with the product measure $\mu_n = \mu \times \cdots \times \mu$. Let us define $F : X_n \longrightarrow \mathbb{R}$ by

$$F(x) = f(x_1) + \ldots + f(x_n) \quad \textit{for} \quad x = (x_1, \ldots, x_n).$$

Then $R_f(t) > 0$ and

$$\lim_{n \longrightarrow +\infty} n^{-1} \ln \mu_n \Big\{ x \in X_n : \quad F(x) > nt \Big\} = -R_f(t),$$

where R_f is the rate function.

Besides,

$$\mu_n \Big\{ x \in X_n : \quad F(x) > nt \Big\} \leq e^{-nR_f(t)} \quad \textit{for all} \quad n.$$

In other words, if we sample n points $x_1, \ldots, x_n \in X$ independently and at random, the probability that the average $\left(f(x_1) + \ldots + f(x_n) \right)/n$ exceeds t is roughly of the order of $\exp\big\{-nR_f(t)\big\}$. This should be contrasted with the estimates of Section 8.2.

(12.4) Tossing a fair and an unfair coin again. Suppose that $X = \{0, 1\}$ and $\mu\{0\} = \mu\{1\} = 1/2$. Let $f(x) = x - 1/2$ for $x \in X$. Then

$$L_f(\lambda) = \ln\left(\frac{e^{\lambda/2} + e^{-\lambda/2}}{2}\right).$$

The maximum of $t\lambda - L_f(\lambda)$ is attained at

$$\lambda = \ln \frac{1+2t}{1-2t} \quad \text{if} \quad -1/2 < t < 1/2,$$

so

$$R_f(t) = -\left(\frac{1}{2} + t\right) \ln \frac{1}{1+2t} - \left(\frac{1}{2} - t\right) \ln \frac{1}{1-2t} = -H\left(\frac{1}{2} - t\right) + \ln 2,$$

where

$$H(x) = x \ln \frac{1}{x} + (1-x) \ln \frac{1}{1-x}$$

is the entropy function.

Let us define the product space X_n, μ_n and the function $F : X_n \longrightarrow \mathbb{R}$ as in Theorem 12.3. Then

$$F(x_1, \ldots, x_n) = x_1 + \ldots + x_n - n/2$$

is interpreted as the deviation of the number of heads shown by a fair coin in n tosses from the expected number $n/2$ of heads. Thus Theorem 12.3 tells us that the probability that the number of heads exceeds $n(t + 0.5)$ is of the order of

$$2^{-n} \exp\left\{nH\left(\frac{1}{2} - t\right)\right\} \quad \text{for} \quad 0 < t < \frac{1}{2},$$

which we established already in Example 9.1.

Let us modify the above example. We still have $X = \{0, 1\}$, but now $\mu\{1\} = p$ and $\mu\{0\} = 1 - p = q$ for some $0 < p < 1$. Also, $f(x) = x - p$. In this case,

$$L_f(\lambda) = \ln\left(pe^{\lambda q} + qe^{-\lambda p}\right).$$

The maximum of $t\lambda - L_f(\lambda)$ is attained at

$$\lambda = \ln \frac{q(t+p)}{p(q-t)} \quad \text{for} \quad -p < t < q$$

Consequently,

$$R_f(t) = (t+p) \ln \left(\frac{q(p+t)}{p(q-t)}\right) - \ln\left(\frac{q}{q-t}\right).$$

Let us define the product space X_n, μ_n and let

$$F(x_1, \ldots, x_n) = x_1 + \ldots + x_n - np.$$

Thus $F(x)$ is the deviation of the number of heads shown by an unfair coin from the expected number np of heads. Theorem 12.3 tells us that the probability that the number of heads exceeds nt is of the order of

$$\exp\{-nR_f(b)\} = \left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right)^n \quad \text{for} \quad 0 < t < q.$$

We don't prove Theorem 12.3, but note that the upper bound for

$$\mu_n\Big\{x \in X_n : \quad F(x) > nt\Big\}$$

by now is a simple exercise in the Laplace transform method:

$$\begin{aligned}\mu_n\Big\{x \in X_n : \quad F(x) > nt\Big\} \leq & e^{-nt\lambda}\mathbf{E}e^{\lambda F(x)} = \left(e^{-t\lambda}\mathbf{E}e^{\lambda f}\right)^n \\ = & \exp\Big\{-t\lambda + \ln \mathbf{E}^{\lambda f}\Big\}^n = \exp\Big\{-\left(t\lambda - L_f(\lambda)\right)\Big\}^n \\ \leq & e^{-R_f(t)n}.\end{aligned}$$

(12.5) Large deviations in the product spaces. Let (X, μ) be a finite probability metric space with the metric dist. For a positive integer n, let us consider the product

$$X_n = X \times \cdots \times X$$

with the product measure $\mu_n = \mu \times \cdots \times \mu$ and the metric

$$\text{dist}_n(x, y) = \sum_{i=1}^n \text{dist}(x_i, y_i) \quad \text{for} \quad x = (x_1, \ldots, x_n) \quad \text{and} \quad y = (y_1, \ldots, y_n).$$

Let us choose a $t > 0$ and consider the following quantity

$$\max_{\substack{A \subset X \\ \mu_n(A) \geq 1/2}} \quad \mu_n\Big\{x : \quad \text{dist}_n(x, A) \geq t\Big\}.$$

In words: we are interested in the maximum possible measure of the complement of the t-neighborhood of a set of measure at least 1/2. Furthermore, we will be interested in the case when t grows proportionately with n, $t \approx \alpha n$ for some $\alpha > 0$. Clearly, we shouldn't try to choose α too large. Let us define

$$\kappa = \max_{\substack{x \in X \\ \mu\{x\} > 0}} \mathbf{E}\,\text{dist}(x, \cdot)$$

In words: we pick a point $x \in X$ such that $\mu\{x\} > 0$, compute the average distance to x and take the maximum over all such points $x \in X$. It is more or less clear that we should choose $t < \kappa n$, since otherwise the measure of the set we care about is way too small even by our standards.

The following remarkable result was obtained by N. Alon, R. Boppana, and J. Spencer (in fact, they proved a more general result).

(12.6) Theorem. *For a real λ, let us define $L(\lambda)$ as the maximum of $L_f(\lambda) = \ln \mathbf{E} e^{\lambda f}$ taken over all 1-Lipschitz functions $f : X \longrightarrow \mathbb{R}$ such that $\mathbf{E} f = 0$. For a $t > 0$, let us define*

$$R(t) = \sup_{\lambda \in \mathbb{R}} \lambda t - L(\lambda).$$

Let us choose $0 < \alpha < \kappa$. Then

$$\lim_{n \longrightarrow +\infty} n^{-1} \ln \max_{\substack{A \subset X_n \\ \mu_n(A) \geq 1/2}} \mu_n \Big\{ x : \quad \operatorname{dist}_n(x, A) \geq \alpha n \Big\} = -R(\alpha).$$

(12.7) Example: weighted Boolean cube. Let $X = \{0, 1\}$ with the metric $\operatorname{dist}(0, 1) = 1$ and the probability measure $\mu\{1\} = p$, $\mu\{0\} = 1 - p = q$ for some $0 < p < 1$. We assume that $p \leq 1/2$. Then $X_n = \{0, 1\}^n$ is the Boolean cube, dist_n is the familiar Hamming distance, and μ_n is the familiar measure of the "unfair coin":

$$\mu_n\{x\} = p^k (1 - q)^{n-k} \quad \text{provided} \quad x = (\xi_1, \ldots, \xi_n) \quad \text{and} \quad \sum_{i=1}^{n} \xi_i = k.$$

If $x = 0$ then the average distance to x is p and if $x = 1$ then the average distance to x is q. Since we assumed that $p \leq q$, we must choose $\kappa = q$ in Theorem 12.6.

Moreover, it is not hard to see that the 1-Lipschitz function $f : \{0, 1\} \longrightarrow \mathbb{R}$ such that $\mathbf{E} f = 0$ and $\mathbf{E} e^{\lambda f}$ is the largest possible (for $\lambda \geq 0$) is defined by $f(1) = q$ and $f(0) = -p$ (again, we used that $p \leq q$). Thus $L(\lambda) = \ln \big(p e^{\lambda q} + q e^{-\lambda p} \big)$ for $\lambda \geq 0$, just as in Example 12.4. Consequently,

$$R(t) = (t + p) \ln \left(\frac{q(p + t)}{p(q - t)} \right) - \ln \frac{q}{q - t} \quad \text{for} \quad 0 < t < q.$$

PROBLEM. In the above example, let $A_n \subset \{0, 1\}^n$ be a Hamming ball centered at $(0, \ldots, 0)$ such that $\mu_n(A_n) = 1/2 + o(1)$. Prove that

$$\lim_{n \longrightarrow +\infty} n^{-1} \ln \mu_n \Big\{ x : \quad \operatorname{dist}_n(x, A) \geq \alpha n \Big\} = -R(\alpha).$$

Lecture 16. Friday, February 11

We don't prove Theorem 12.6, but discuss some ideas of the proof by Alon, Boppana, and Spencer. This is a martingale proof in spirit.

(12.8) The additive property of $L(\lambda)$. Let us look closely at the quantity $L(\lambda) = L_X(\lambda)$ introduced in Theorem 12.6, that is, the maximum of $\ln \mathbf{E} e^{\lambda f}$ taken over all 1-Lipschitz functions $f : X \longrightarrow \mathbb{R}$ such that $\mathbf{E} f = 0$. The crucial observation is that $L(\lambda)$ is additive with respect to the direct product of spaces. Namely, suppose that (X_1, μ_1) and (X_2, μ_2) are probability metric spaces with metrics dist_1 and dist_2 and let $X = X_1 \times X_2$ be the space with the product measure $\mu = \mu_1 \times \mu_2$ and the distance $\operatorname{dist}(x, y) = \operatorname{dist}_1(x_1, y_1) + \operatorname{dist}_2(x_2, y_2)$ for $x = (x_1, x_2)$ and $y = (y_1, y_2)$. We claim that

$$L_X(\lambda) = L_{X_1}(\lambda) + L_{X_2}(\lambda).$$

It is easy to see that $L_X(\lambda) \geq L_{X_1}(\lambda) + L_{X_2}(\lambda)$. Indeed, if $f_i : X_i \longrightarrow \mathbb{R}$ are 1-Lipschitz functions with $\mathbf{E} f_1 = \mathbf{E} f_2 = 0$ and we define $f : X \longrightarrow \mathbb{R}$ by $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$ then f is 1-Lipschitz, $\mathbf{E} f = 0$ and

$$\int_X e^{\lambda f} \, d\mu = \left(\int_{X_1} e^{\lambda f_1} \, d\mu_1 \right) \left(\int_{X_2} e^{\lambda f_2} \, d\mu_2 \right),$$

from which $L_f(\lambda) = L_{f_1}(\lambda) + L_{f_2}(\lambda)$, so we have

$$\begin{aligned} L_{X_1}(\lambda) + L_{X_2}(\lambda) =& \max_{f_1} L_{f_1}(\lambda) + \max_{f_2} L_{f_2}(\lambda) \\ =& \max_{f_1, f_2} \Big(L_{f_1}(\lambda) + L_{f_2}(\lambda) \Big) = \max_{f = f_1 + f_2} L_f(\lambda) \leq L_X(\lambda). \end{aligned}$$

The heart of the argument is the inequality in the opposite direction: $L_X(\lambda) \leq L_{X_1}(\lambda) + L_{X_2}(\lambda)$. Let us choose a 1-Lipschitz function $f : X \longrightarrow \mathbb{R}$ such that $\mathbf{E} f = 0$. Let us define $g : X_1 \longrightarrow \mathbb{R}$ by

$$g(x_1) = \int_{X_2} f(x_1, x_2) \, d\mu_2(x_2).$$

Thus g is the conditional expectation of f with respect to the first variable x_1. Clearly, $\mathbf{E}(g) = 0$. Moreover, g is 1-Lipschitz since

$$\begin{aligned} |g(x_1) - g(y_1)| =& \Big| \int_{X_2} f(x_1, x_2) - f(y_1, x_2) \, d\mu_2(x_2) \Big| \\ \leq& \int_{X_2} |f(x_1, x_2) - f(y_1, x_2)| \, d\mu_2(x_2) \\ \leq& \int_{X_2} \operatorname{dist}_1(x_1, y_1) \, d\mu_2 = \operatorname{dist}_1(x_1, y_1). \end{aligned}$$

Now, for each $x_1 \in X_1$, let us define a function $h_{x_1} : X_1 \longrightarrow \mathbb{R}$ by $h_{x_1}(x_2) = f(x_1, x_2) - g(x_1)$. Again, $\mathbf{E}h_{x_1}(x_2) = 0$ for all $x_1 \in X_1$ since

$$\mathbf{E}h_{x_1} = \int_{X_2} h_{x_1}(x_2)\ d\mu_2(x_2) = \int_{X_2} f(x_1, x_2) - g(x_1)\ d\mu_2(x_2) = g(x_1) - g(x_1) = 0.$$

Moreover, h_{x_1} is 1-Lipschitz, since

$$\begin{aligned}|h_{x_1}(x_2) - h_{x_1}(y_2)| =&|f(x_1, x_2) - g(x_1) - f(x_1, y_2) + g(x_1)| \\ =&|f(x_1, x_2) - f(x_1, y_2)| \leq \operatorname{dist}(y_1, y_2).\end{aligned}$$

Hence we conclude that

$$\ln \mathbf{E}e^{\lambda g} \leq L_{X_1}(\lambda) \quad \text{and} \quad \ln \mathbf{E}e^{\lambda h_{x_1}} \leq L_{X_2}(\lambda) \quad \text{for all} \quad \lambda.$$

Besides, $f(x_1, x_2) = g(x_1) + h_{x_1}(x_2)$. Therefore,

$$\begin{aligned}\int_X e^{\lambda f}\ d\mu =& \int_{X_1} \left(\int_{X_2} e^{\lambda f}\ d\mu_2 \right)\ d\mu_1 = \int_{X_1} \left(\int_{X_2} e^{\lambda g} e^{\lambda h_{x_1}}\ d\mu_2 \right)\ d\mu_1 \\ =& \int_{X_1} e^{\lambda g} \left(\int_{X_2} e^{\lambda h_{x_1}}\ d\mu_2 \right)\ d\mu_1 \leq \int_{X_1} e^{\lambda g} e^{L_{X_2}(\lambda)}\ d\mu_1 \\ =& e^{L_{X_2}(\lambda)} \int_{X_1} e^{\lambda g}\ d\mu_1 \leq e^{L_{X_2}(\lambda)} e^{L_{X_1}(\lambda)}.\end{aligned}$$

Taking the logarithm, we happily conclude that

$$\ln \mathbf{E}e^{\lambda f} \leq L_{X_1}(\lambda) + L_{X_2}(\lambda),$$

and since the function f was arbitrary, $L_X(\lambda) \leq L_{X_1}(\lambda) + L_{X_2}(\lambda)$, which finally proves that $L_X(\lambda) = L_{X_1}(\lambda) + L_{X_2}(\lambda)$.

PROBLEMS.

1. Suppose that X consists of 3 points, $X = \{1, 2, 3\}$ and that $\mu\{1\} = \mu\{2\} = \mu\{3\} = 1/3$. Prove that

$$L_X(\lambda) = \ln \left(\frac{1}{3} e^{2\lambda/3} + \frac{2}{3} e^{-\lambda/3} \right) \quad \text{for} \quad \lambda \geq 0.$$

2. Let X, μ be a finite probability space with the Hamming metric

$$\operatorname{dist}(x, y) = \begin{cases} 1 & \text{if} \quad x \neq y \\ 0 & \text{if} \quad x = y. \end{cases}$$

Let $f : X \longrightarrow \mathbb{R}$ be a 1-Lipschitz function with $\mathbf{E}f = 0$ which achieves the maximum of $\mathbf{E}^{\lambda f}$ for some λ. Prove that f takes at most two values.

Now we go back to Theorem 12.6.

(12.9) Proving the upper bound in Theorem 12.6. Let us choose a set $A \subset X_n$ with $\mu_n(A) \geq 1/2$. Let us define $f : X \longrightarrow \mathbb{R}$ by $f(x) = \operatorname{dist}(x, A)$ and let $a = \mathbf{E}f$. It follows from almost any result we proved so far (see Corollary 10.3, Theorem 8.1, etc.) that $a = O(\sqrt{n}) = o(n)$. Next, we note that $f - a$ is 1-Lipschitz and that $\mathbf{E}(f - a) = 0$. Therefore,

$$\ln \mathbf{E} e^{\lambda(f-a)} \leq L_{X_n}(\lambda) = nL_X(\lambda),$$

as we proved above.

Now, applying the Laplace transform bound, let us choose $\lambda \geq 0$. Then

$$\begin{aligned}\mu_n\Big\{x : \quad f(x) \geq \alpha n\Big\} \leq & e^{-\lambda \alpha n} \mathbf{E} e^{\lambda f} = e^{-\lambda \alpha n} e^{\lambda a} \mathbf{E} e^{\lambda(f-a)} \\ \leq & e^{-\lambda(\alpha n + o(1))} e^{L_{X_n}(\lambda)} = \exp\Big\{\big(-\lambda(\alpha + o(1)) + L_X(\lambda)\big)n\Big\}.\end{aligned}$$

Hence

$$n^{-1} \ln \mu_n\Big\{x : \quad f(x) \geq \alpha n\Big\} \leq -\lambda\big(\alpha + o(1)\big) + L_X(\lambda) = -\Big(\lambda(\alpha + o(1)) - L_X(\lambda\Big).$$

Optimizing on λ (note that since $L_X(\lambda) \geq 0$, the optimal λ is non-negative), we get that

$$n^{-1} \ln \mu_n\Big\{x : \quad f(x) \geq \alpha n\Big\} \leq -R\big(\alpha + o(1)\big).$$

(12.10) The spread constant of a space. In the same paper, Alon, Boppana, and Spencer study another interesting parameter associated with a probability metric space, the *spread constant*. The spread constant $c(X)$ is the maximum value of $\mathbf{E}(f^2)$ taken over all 1-Lipschitz functions $f : X \longrightarrow \mathbb{R}$ such that $\mathbf{E}f = 0$. They prove that in the situation of Theorem 12.6, for $n^{1/2} \ll t \ll n$, we have

$$\max_{\substack{A \subset X_n \\ \mu_n(A) \geq 1/2}} \mu_n\Big\{x : \quad \operatorname{dist}_n(x, A) \geq t\Big\} = \exp\Big\{-\frac{t^2}{2cn}\big(1 + o(1)\big)\Big\},$$

which sharpens the estimates of Corollary 10.3 and Theorem 8.1.

PROBLEMS.

1. Let X be a finite probability measure space. Prove that

$$\lim_{\lambda \longrightarrow 0} \frac{L_X(\lambda)}{\lambda^2} = \frac{c(X)}{2}.$$

2. Deduce that for $X = X_1 \times X_2$ with the product measure $\mu = \mu_1 \times \mu_2$ and the distance $\operatorname{dist}(x, y) = \operatorname{dist}(x_1, y_1) + \operatorname{dist}(x_2, y_2)$ for $x = (x_1, y_1)$ and $y = (y_1, y_2)$, one has $c(X) = c(X_1) + c(X_2)$.

3. Let $I_n = \{0,1\}^n$ be the Boolean cube. Let us choose a number $0 < p \leq 1/2$ and let $q = 1-p$. Let us introduce the product measure μ_n by $\mu_n(x) = p^k q^{n-k}$ if there are exactly k 1's among the coordinates of x. Prove the following asymptotic isoperimetric inequality. Let $A_n \subset I_n$ be a sequence of sets and let $B_n \subset I_n$ be a ball centered at $(1, \ldots, 1)$ and such that $|\ln \mu_n(B_n) - \ln \mu_n(A_n)| = o(n)$. Let us choose a number $0 < \alpha < 1$. Prove that

$$\ln \mu_n \Bigl\{ x : \quad \operatorname{dist}(x, A_n) \leq \alpha n \Bigr\} \geq \ln \mu_n \Bigl\{ x : \quad \operatorname{dist}(x, B_n) \leq \alpha n \Bigr\} + o(n).$$

4. Let $X = \{1,2,3\}$ with pairwise distances equal to 1 and the uniform probability measure $\mu\{1\} = \mu\{2\} = \mu\{3\} = 1/3$. Let $X_n = X \times \ldots \times X$ be the product space with the product measure and the Hamming metric. Prove the following asymptotic isoperimetric inequality. Let $A_n \subset X_n$ be a sequence of sets and let $B_n \subset X_n$ be a ball such that $|\ln \mu_n(A_n) - \ln \mu_n(B_n)| = o(n)$. Let us choose a number $0 < \alpha < 1$. Prove that

$$\ln \mu_n \Bigl\{ x : \quad \operatorname{dist}(x, A_n) \leq \alpha n \Bigr\} \geq \ln \mu_n \Bigl\{ x : \quad \operatorname{dist}(x, B_n) \leq \alpha n \Bigr\} + o(n).$$

Lecture 18. Wednesday, February 16

Lecture 17 on Monday, February 14, covered the material in the previous handout.

13. Gaussian measure on Euclidean Space as a limit of the projection of the uniform measure on the sphere

Having studied product spaces for a while, we go back now with the newly acquired wisdom to where we started, to the Gaussian measure on Euclidean space and the uniform measure on the sphere.

We will need to compute some things on the unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$. First, we will need to compute the "surface area" (which we just call "volume") $|\mathbb{S}^{n-1}|$ of the unit sphere.

(13.1) Lemma. *For the unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, we have*

$$|\mathbb{S}^{n-1}| = \frac{2\pi^{n/2}}{\Gamma(n/2)},$$

where

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} \, dx \quad \textit{for} \quad t > 0$$

is the Gamma-function.

Proof. Let

$$p_n(x) = (2\pi)^{-n/2} e^{-\|x\|/2} \quad \text{for} \quad x \in \mathbb{R}^n$$

be the standard Gaussian density in $\mathbb{R}^n$ and let

$$\mathbb{S}^{n-1}(r) = \Big\{ x : \quad \|x\| = r \Big\}$$

be the sphere of radius r. We integrate p_n over $\mathbb{R}^n$ using the polar coordinates:

$$\begin{aligned} 1 = \int_{\mathbb{R}^n} p_n(x) \ dx = &(2\pi)^{-n/2} \int_0^{+\infty} e^{-r^2/2} |\mathbb{S}^{n-1}(r)| \ dr \\ =&(2\pi)^{-n/2} |\mathbb{S}^{n-1}| \int_0^{+\infty} r^{n-1} e^{-r^2/2} \ dr. \end{aligned}$$

The integral reduced to the Gamma-function by the substitution $r^2/2 = t \Leftrightarrow r = (2t)^{1/2}$.

$$\int_0^{+\infty} r^{n-1} e^{-r^2/2} \ dr = 2^{(n-2)/2} \int_0^{+\infty} t^{(n-2)/2} e^{-t} \ dt = 2^{(n-2)/2} \Gamma(n/2).$$

Therefore,

$$|\mathbb{S}^{n-1}| = \frac{(2\pi)^{n/2} 2^{(2-n)/2}}{\Gamma(n/2)} = \frac{2\pi^{n/2}}{\Gamma(n/2)},$$

as claimed. □

We recall that $\Gamma(t+1) = t\Gamma(t)$, so that $\Gamma(t) = (t-1)!$ for positive integer t and that $\Gamma(1/2) = \sqrt{\pi}$. Stirling's formula says that

$$\Gamma(t+1) = \sqrt{2\pi t} \left(\frac{t}{e} \right)^t \Big(1 + O(t^{-1}) \Big) \quad \text{as} \quad t \longrightarrow +\infty.$$

We have already shown how to "condense" the uniform probability measure on the sphere of radius $\sim \sqrt{n}$ in $\mathbb{R}^n$ (Section 2) and how to obtain the uniform probability measure on the unit sphere $\mathbb{S}^{n-1}$ from the Gaussian measure on $\mathbb{R}^n$ via the radial projection $\mathbb{R}^n \setminus \{0\} \longrightarrow \mathbb{S}^{n-1}$, Section 3.3. However, the relation between the Gaussian measure on Euclidean space and the uniform measure on the sphere is more interesting. We show how to obtain the Gaussian measure on Euclidean space as a limit of the push-forward of the uniform measure on the sphere.

(13.2) Theorem. *Let $\Sigma_n \subset \mathbb{R}^n$ be the sphere centered at the origin and of radius $\sqrt{n}$,*

$$\Sigma_n = \Big\{ x \in \mathbb{R}^n : \quad \|x\| = \sqrt{n} \Big\},$$

and let μ_n be the uniform (that is, rotationally invariant, Borel) probability measure on Σ_n. Let us consider the projection $\phi : \Sigma_n \longrightarrow \mathbb{R}$

$$(\xi_1, \dots, \xi_n) \longmapsto \xi_1$$

and let ν_n be the push-forward measure on the line $\mathbb{R}^1$, that is,

$$\nu_n(A) = \mu_n\left(\phi^{-1}(A)\right)$$

for any Borel set $A \subset \mathbb{R}$. Let γ_1 be the standard Gaussian measure on $\mathbb{R}$ with the standard normal density

$$\frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}.$$

Then, for any Borel set $A \subset \mathbb{R}$, we have

$$\lim_{n \longrightarrow +\infty} \nu_n(A) = \gamma_1(A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-\xi^2/2} \, d\xi.$$

Moreover, the density of ν_n converges to the standard normal density uniformly on compact subsets of $\mathbb{R}$.

Proof. Let us choose an interval $A = (a, b) \subset \mathbb{R}$ and compute $\nu_n(A)$. Let $B = \phi^{-1}(A) \subset \Sigma_n$. Then B is the subset of the sphere Σ_n consisting of the points such that $a < \xi_1 < b$. If we fix the value of ξ_1, we obtain the set of points that is the $(n-2)$-dimensional sphere of radius $\sqrt{n - \xi_1^2}$. Note that since a and b are fixed and n grows, for $a < \xi_1 < b$, we have $d\xi_1$ almost tangent to the sphere, so

$$\begin{aligned}\Big(1 + o(1)\Big)\mu_n(B) =& \frac{|\mathbb{S}^{n-2}|}{n^{(n-1)/2}|\mathbb{S}^{n-1}|} \int_a^b (n - \xi_1^2)^{(n-2)/2} \, d\xi_1 \\ =& \frac{|\mathbb{S}^{n-2}| n^{(n-2)/2}}{n^{(n-1)/2}|\mathbb{S}^{n-1}|} \int_a^b \left(1 - \frac{\xi_1^2}{n}\right)^{(n-2)/2} d\xi_1 \\ =& \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{n}\Gamma\left(\frac{n-1}{2}\right)} \int_a^b \left(1 - \frac{\xi_1^2}{n}\right)^{(n-2)/2} d\xi_1.\end{aligned}$$

It follows from Stirling's formula that

$$\lim_{n \longrightarrow +\infty} \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{n}\Gamma\left(\frac{n-1}{2}\right)} = \frac{1}{\sqrt{2}}.$$

Next,

$$\left(1 - \frac{\xi^2}{n}\right)^{\frac{n-2}{2}} = \exp\left\{\frac{n-2}{2} \ln\left(1 - \frac{\xi^2}{n}\right)\right\}.$$

Using that

$$\ln\left(1-\frac{\xi^2}{n}\right) = -\frac{\xi^2}{n} + O\left(n^{-2}\right),$$

we conclude that

$$\left(1-\frac{\xi^2}{n}\right)^{\frac{n-2}{2}} \longrightarrow e^{-\xi^2/2}$$

uniformly on the interval (a, b), which completes the proof. □

Similarly, the standard Gaussian measure on Euclidean space of any dimension can be obtained as the limit of the push-forward of the uniform probability measure on the sphere under the projection onto a subset of coordinates.

(13.3) Corollary. *Let $\Sigma_n \subset \mathbb{R}^n$ be the sphere centered at the origin and of radius $\sqrt{n}$ and let μ_n be the uniform probability measure on Σ_n. For a fixed k, let us consider the projection $\Sigma_n \longrightarrow \mathbb{R}^k$*

$$(\xi_1, \dots, \xi_n) \longmapsto (\xi_1, \dots, \xi_k)$$

and let ν_n be the push-forward measure on $\mathbb{R}^k$. Let γ_k be the standard Gaussian measure on $\mathbb{R}^k$ with the standard normal density

$$(2\pi)^{-k/2} e^{-\|x\|^2/2}.$$

Then, for any Borel set $A \subset \mathbb{R}^k$, we have

$$\lim_{n \longrightarrow +\infty} \nu_n(A) = \gamma_k(A) = (2\pi)^{-k/2} \int_A e^{-\|x\|^2/2} \, dx.$$

Moreover, the density of ν_n converges to the standard normal density uniformly on compact subsets of $\mathbb{R}^k$.

Proof. Perhaps the easiest way to obtain the corollary is via the Fourier transform (characteristic functions) method. Let us consider the characteristic function $F_n : \mathbb{R}^k \longrightarrow \mathbb{C}$ of the measure ν_n: for a vector $c \in \mathbb{R}^k$, we have

$$F_n(c) = \int_{\mathbb{R}^k} e^{i\langle c, x\rangle} \, d\nu_n(x).$$

Similarly, let $G : \mathbb{R}^k \longrightarrow \mathbb{R}$ be the characteristic function of the standard Gaussian measure on $\mathbb{R}^k$:

$$G(c) = (2\pi)^{-k/2} \int_{\mathbb{R}^k} e^{i\langle c, x\rangle} e^{-\|x\|^2/2} \, dx = e^{-\|c\|^2/2}$$

(check the formula). The result would follow if we can prove that $F_n(c) \longrightarrow G(c)$ uniformly on compact subsets of $\mathbb{R}^n$ (we can recover densities from $F_n(c)$ and $G(c)$ by the inverse Fourier transform).

We can write

$$F_n(c) = \int_{\Sigma_n} e^{i\langle c,x\rangle} \ d\mu_n(x).$$

Since the measure μ_n on Σ_n is rotation-invariant, we can assume that $c = (\|c\|, 0, \ldots , 0)$. Therefore,

$$F_n(c) = \int_{\Sigma_n} e^{i\|c\|\xi_1} \ d\mu_n(x).$$

Now, using Theorem 13.3, we conclude that as $n \longrightarrow +\infty$ we have

$$F_n(c) \longrightarrow \int_{\mathbb{R}} e^{i\|c\|\xi} \ d\gamma_1(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\|c\|\xi} e^{-\xi^2/2} \ d\xi = e^{-\|c\|^2/2}$$

and that the convergence is uniform on compact subsets. This is what is needed. □

Lecture 19. Friday, February 18

14. Isoperimetric inequalities for the sphere and for the Gaussian measure

Let $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ be the unit sphere

$$\mathbb{S}^{n-1} = \Big\{ x : \quad \|x\| = 1 \Big\}$$

with the rotation invariant (Haar) probability measure $\mu = \mu_n$ and the geodesic metric

$$\operatorname{dist}(x, y) = \arccos\langle x, y\rangle,$$

cf. Section 1.1. We define the *spherical cap* as a ball in the metric of $\mathbb{S}^{n-1}$:

$$B_a(r) = \Big\{ x \in \mathbb{S}^{n-1} : \quad \operatorname{dist}(x, a) \leq r \Big\}.$$

The famous result of P. Levy states that among all subset $A \subset \mathbb{S}^{n-1}$ of a given measure, the spherical cap has the smallest measure of the neighborhood.

(14.1) Theorem. *Let $A \subset \mathbb{S}^{n-1}$ be a closed set and let $t \geq 0$ be a number. Let $B = B(a, r) \subset \mathbb{S}^{n-1}$ be a spherical cap such that*

$$\mu(A) = \mu(B).$$

Then

$$\mu\Big\{ x : \quad \operatorname{dist}(x, A) \leq t \Big\} \geq \mu\Big\{ x : \quad \operatorname{dist}(x, B) \leq t \Big\} = \mu\left(B_a(r + t)\right).$$

We don't prove this nice theorem here, but extract some corollaries instead. Clearly, if $B \subset \mathbb{S}^{n-1}$ is a spherical cap such that $\mu(B) = 1/2$ then the radius of B is $\pi/2$, so $B = B_a(\pi/2)$ for some a. In this case,

$$B_a(\pi/2 + t) = \mathbb{S}^{n-1} \setminus B_{-a}(\pi/2 - t)$$

(why?), so

$$\mu\left(B_a(\pi/2 + t)\right) = 1 - \mu\left(B_{-a}(\pi/2 - t)\right).$$

We want to estimate the measure of the spherical cap of radius $\pi/2 - t$ for $0 \leq t \leq \pi/2$. One way to do it is to note that the cap consists of the vectors x whose orthogonal projection onto the hyperplane $a^{\perp}$ has length at most $\cos(t) \approx 1 - t^2/2$ for small t and use Corollary 3.4, say. This would be good enough, though it does not give the optimal constant. One can compute the measure directly. Note that it is convenient to "shift" the dimension and to prove the result for the $(n+1)$-dimensional sphere $\mathbb{S}^{n+2} \subset \mathbb{R}^{n+2}$.

(14.2) Lemma. *For the spherical cap $B \subset \mathbb{S}^{n+1}$ of radius $\pi/2 - t$, we have*

$$\mu_{n+2}(B) \leq \sqrt{\frac{\pi}{8}} e^{-t^2 n/2}.$$

Besides, for any $t > 0$,

$$\mu_{n+2}(B) \leq \frac{1}{2} e^{-t^2 n/2}\Big(1 + o(1)\Big) \quad \textit{as} \quad n \longrightarrow +\infty.$$

Proof. We introduce a coordinate system so that $(1, 0, \ldots, 0)$ becomes the center of the cap. Let us slice the cap by the hyperplanes $\xi_1 = \cos\phi$ onto n-dimensional spheres of radius $\sin\phi$. Then

$$\begin{aligned}\mu(B) =& \frac{|\mathbb{S}^n|}{|\mathbb{S}^{n+1}|} \int_0^{\pi/2 - t} \sin^n \phi \; d\phi = \frac{2\pi^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \frac{\Gamma\left(\frac{n+2}{2}\right)}{2\pi^{\frac{n+2}{2}}} \int_t^{\pi/2} \cos^n \phi \; d\phi \\ =& \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi}} \int_t^{\pi/2} \cos^n \phi \; d\phi = \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi}\sqrt{n}} \int_{t\sqrt{n}}^{\pi\sqrt{n}/2} \cos^n\left(\frac{\psi}{\sqrt{n}}\right) \; d\psi\end{aligned}$$

Now we use the inequality

$$\cos\alpha \leq e^{-\alpha^2/2} \quad \text{for} \quad 0 \leq \alpha \leq \frac{\pi}{2}$$

to bound

$$\begin{aligned}\mu(B) \leq& \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi n}} \int_{t\sqrt{n}}^{\pi\sqrt{n}/2} e^{-\psi^2/2} \; d\psi \\ \leq& \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi n}} e^{-t^2 n/2} \int_0^{(\pi/2 - t)\sqrt{n}} e^{-\psi^2/2} \; d\psi \\ \leq& \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi n}} e^{-t^2 n/2} \int_0^{+\infty} e^{-\psi^2/2} \; d\psi = \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{2n}} e^{-t^2 n/2}\end{aligned}$$

Now it remains to show that

$$\frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{2n}} \leq \sqrt{\frac{\pi}{8}},$$

which is done by observing that replacing n by $n+2$ gets the fraction multiplied by

$$\frac{(n+4)}{(n+3)}\sqrt{\frac{n}{n+2}},$$

that is makes is slightly smaller. Hence the maximum value is attained at $n=2$ or at $n=1$, which is indeed the case and is equal to $\sqrt{\pi/8}$.

As follows from Stirling's formula (cf. also the proof of Theorem 13.2),

$$\lim_{n \longrightarrow +\infty} \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\sqrt{2n}} = \frac{1}{2}.$$

□

An immediate corollary of Theorem 14.1 and Lemma 14.2 is the concentration result for Lipschitz functions on the unit sphere.

(14.3) Theorem. *Let $f: \mathbb{S}^{n+1} \longrightarrow \mathbb{R}$ be a 1-Lipschitz function and let m_f be its median, that is, the number such that*

$$\mu\Big\{x: \quad f(x) \geq m_f\Big\} \geq \frac{1}{2} \quad \textit{and} \quad \mu\Big\{x: \quad f(x) \leq m_f\Big\} \geq \frac{1}{2}.$$

Then, for any $\epsilon > 0$,

$$\mu\Big\{x: \quad |f(x) - m_f| \geq \epsilon\Big\} \leq \sqrt{\frac{\pi}{2}} e^{-\epsilon^2 n/2}.$$

Besides, for any $\epsilon > 0$,

$$\mu\Big\{x: \quad |f(x) - m_f| \geq \epsilon\Big\} \leq e^{-\epsilon^2 n/2}\Big(1 + o(1)\Big) \quad \textit{as} \quad n \longrightarrow +\infty.$$

Proof. The proof follows the same lines as the proof of Theorem 4.5 (see also Section 10). We define

$$A_+ = \Big\{x: \quad f(x) \geq m_f\Big\} \quad \text{and} \quad A_- = \Big\{x: \quad f(x) \leq m_f\Big\}$$

so that $\mu(A_+), \mu(A_-) \geq 1/2$. Therefore, by Theorem 14.1, the measure of the ϵ-neighborhoods of A_- and A_+ is at least as large as the measure of the ϵ-neighborhood of the hemisphere, which, by Lemma 14.2 is at least

$1-\sqrt{\pi/8}e^{-\epsilon^2 n/2}$. This, in turn, implies that the measure of the intersection of the ϵ-neighborhoods of A_+ and A_- is at least $1-\sqrt{\pi/2}e^{-\epsilon^2 n/2}$. However, for all $x \in A_+(\epsilon) \cap A_-(\epsilon)$ we have $|f(x)-m_f| \leq \epsilon$ which completes the proof. □

Let us turn now to the space $\mathbb{R}^n$ with the standard Euclidean distance

$$\operatorname{dist}(x,y)=\|x-y\|$$

and the standard Gaussian measure $\gamma=\gamma_n$ with the density

$$(2\pi)^{-n/2}e^{-\|x\|^2/2}.$$

We define a *halfspace* as a set of the type:

$$H(\alpha)=\Big\{(\xi_1,\ldots,\xi_n): \quad \xi_1 \leq \alpha\Big\}$$

for some $\alpha \in \mathbb{R}$. In general, we call a halfspace the set of points whose scalar product with a given non-zero vector does not exceed a given number.

The following remarkable result due to C. Borell states that among all subsets $A \subset \mathbb{R}^n$ of a given measure γ, halfspaces have the smallest measure of the neighborhood.

(14.4) Theorem. *Let $A \subset \mathbb{R}^n$ be a closed set and let $t \geq 0$ be number. Let $H=H(\alpha) \subset \mathbb{R}^n$ be a halfspace such that*

$$\gamma(A)=\gamma(H).$$

Then

$$\gamma\Big\{x: \quad \operatorname{dist}(x,A) \leq t\Big\} \geq \gamma\Big\{x: \quad \operatorname{dist}(x,H) \leq t\Big\} = \gamma\big(H(\alpha+t)\big).$$

We don't prove this nice theorem (at least, not now), but extract some corollaries instead. Clearly, if $H=H(\alpha)$ is a halfspace such that $\mu(H)=1/2$ then $\alpha=0$. In this case,

$$\gamma\big(H(t)\big)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\xi^2/2}\, d\xi = 1-\frac{1}{\sqrt{2\pi}}\int_{t}^{+\infty} e^{-\xi^2/2}\, d\xi,$$

so we want to estimate the integral.

(14.5) Lemma. *For the halfspace $H=H(t)$ with $t \geq 0$, we have*

$$\gamma(H) \geq 1-e^{-t^2/2}.$$

Besides,

$$\gamma(H) \geq 1 - \frac{e^{-t^2/2}}{t}.$$

Proof. We use the by now standard Laplace transform method. For any $\lambda \geq 0$,

$$\gamma\Big\{x: \quad \xi_1 \geq t\Big\} \leq e^{-\lambda t}\mathbf{E}e^{\lambda x} = e^{-\lambda t}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\lambda\xi}e^{-\xi^2/2}\ d\xi = e^{\lambda t - \lambda^2/2}.$$

Optimizing on λ, we substitute $\lambda = t$, from which we get the desired estimate. To get the second estimate, we note that

$$\int_t^{+\infty} e^{-\xi^2/2}\ d\xi \leq t^{-1}\int_t^{+\infty} \xi e^{-\xi^2/2} = \frac{e^{-t^2/2}}{t}.$$

□

Just as before, we obtain a concentration result.

(14.6) Theorem. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a 1-Lipschitz function and let m_f be its median. Then, for any $\epsilon > 0$,*

$$\gamma\Big\{x: \quad |f(x) - m_f| \geq \epsilon\Big\} \leq 2e^{-\epsilon^2/2}.$$

Besides,

$$\gamma\Big\{x: \quad |f(x) - m_f| \geq \epsilon\Big\} \leq \frac{2}{\epsilon}e^{-\epsilon^2/2}.$$

Hence we observe that as long as $\epsilon \longrightarrow +\infty$, the probability that $f(x)$ deviates from the median by more than ϵ tends to 0.

(14.7) Relations between the isoperimetric inequalities on the sphere and for the Gaussian measure in $\mathbb{R}^n$.

One can deduce Theorem 14.4 as a "limit case" of Theorem 14.1. Let us choose a large number N, let us consider the unit sphere $\mathbb{S}^{N-1} \subset \mathbb{R}^n$ with the uniform probability measure μ and the scaled projection $\mathbb{S}^{N-1} \longrightarrow \mathbb{R}^n$.

$$\phi : (\xi_1, \ldots, \xi_N) \longmapsto \sqrt{N}(\xi_1, \ldots, \xi_n).$$

Corollary 13.3 tells us that the push-forward ν converges to the Gaussian measure γ on $\mathbb{R}^n$ as N grows. Let $A \subset \mathbb{R}^n$ be a closed bounded set. Then $C = \phi^{-1}(A) \subset \mathbb{S}^{N-1}$ is a closed set and since we assumed that A is bounded, for every $x = (\xi_1, \ldots, \xi_N) \in C$ we have $|\xi_i| = O(N^{-1/2})$ for $i = 1, \ldots, n$. Besides,

$$\mu(C) = \gamma(A) + o(1).$$

Let us choose a halfspace $H \subset \mathbb{R}^n$, $H = \Big\{x : \quad \xi_1 \leq \alpha\Big\}$ such that $\gamma(H) = \gamma(A)$. Then $B = \phi^{-1}(H)$ is the spherical cap in $\mathbb{S}^{N-1}$ centered at $(-1, 0, \dots, 0)$ and defined by the inequality $\xi_1 \leq N^{-1/2}\alpha$. Similarly, we have

$$\mu(B) = \gamma(H) + o(1).$$

Let us compare the t-neighborhoods of A_t and H_t of A and H respectively:

$$A_t = \Big\{x \in \mathbb{R}^n : \quad \operatorname{dist}(x, A) \leq t\Big\} \quad \text{and} \quad H_t = \Big\{x \in \mathbb{R}^n : \quad \xi_1 \leq \alpha + t\Big\}.$$

Then

$$\gamma(A_t) = \mu\left(\phi^{-1}(A_t)\right) + o(1) \quad \text{and} \quad \gamma(H_t) = \mu\left(\phi^{-1}(H_t)\right) + o(1).$$

Now, $\phi^{-1}(H_t)$ is the spherical cap in $\mathbb{S}^{N-1}$ defined by the inequality $\xi_1 \leq N^{-1/2}(\alpha + t)$. This is *not* the $N^{-1/2}t$-neighborhood of the spherical cap B but something very close to it. On the sphere $\mathbb{S}^{N-1}$ we measure distances by the length of the geodesic arc. However, if the points on the sphere are close enough, the length of the arc approaches the Euclidean distance between the points. That is, if the geodesic length is β, say, then the Euclidean distance is $2\sin\beta/2 = \beta + O(\beta^3)$ for small β. This allows us to prove that if we choose $\epsilon = N^{-1/2}t$ and consider the ϵ-neighborhood B_ϵ of B, we get

$$\mu(B_\epsilon) = \mu\left(\phi^{-1}(H_t)\right) + o(1).$$

To claim that, we need the estimate of Lemma 13.1 for the volume of the sphere. Namely, we need that $|\mathbb{S}^{N-1}|/|\mathbb{S}^{N-2}| = O(N^{1/2})$, so that by modifying distances by something of the order of $O(N^{-3/2})$ we don't lose any substantial volume.

Similarly, $\phi^{-1}(A_t)$ is *not* the ϵ-neighborhood C_ϵ of C but something very close to it, so

$$\mu(C_\epsilon) = \mu\left(\phi^{-1}(A_t)\right) + o(1).$$

By the isoperimetric inequality on the sphere, $\mu(C_\epsilon) \geq \mu(B_\epsilon)$ and hence

$$\gamma(A_t) \geq \gamma(H_t) + o(1).$$

Taking the limit as $N \longrightarrow +\infty$, we get

$$\gamma(A_t) \geq \gamma(H_t),$$

that is, the isoperimetric inequality for the Gaussian measure.

We assumed that A is bounded, but the inequality extends to unbounded sets by a limit argument (approximate A by bounded subsets).

(14.8) Looking at the unit sphere through a microscope. What we did in Section 14.7 provokes the following thought. Let us looks at a higher-dimensional unit sphere $\mathbb{S}^{N-1}$. Let us choose some constant number n of coordinates $\xi_1, \ldots, \xi_n$ and let us consider the t-neighborhood of the section $\xi_1 = \ldots = \xi_n = 0$ for $t = \epsilon_N N^{-1/2}$ where $\epsilon_N \longrightarrow +\infty$, however slowly. Then the measure of the sphere outside of the neighborhood tends to 0 as N grows. In the "normal direction" the neighborhood of any particular point of the section $\xi_1 = \ldots = \xi_n = 0$ looks something akin to $\mathbb{R}^n$ endowed with the Gaussian measure.

Lecture 20. Monday, February 21

---

15. Another concentration inequality for the Gaussian measure

We discuss a relatively short, though far from straightforward, proof of a concentration result for the Gaussian measure in $\mathbb{R}^n$. The result concerns concentration with respect to the average, not the median, and extends to a somewhat larger class of functions $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ than Lipschitz functions. The proof is due to B. Maurey and G. Pisier.

The proof is based on several new (to us) ideas.

The first idea: function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ does not deviate much from its average if and only if f does not deviate much from itself. Namely, let $\mathbb{R}^n$ be another copy of $\mathbb{R}^n$ and let us consider the direct sum $\mathbb{R}^{2n} = \mathbb{R}^n \oplus \mathbb{R}^n$ endowed with the product Gaussian measure $\gamma_{2n} = \gamma_n \times \gamma_n$. Let us define $F : \mathbb{R}^{2n} \longrightarrow \mathbb{R}$ by $F(x, y) = f(x) - f(y)$. Then f concentrates around somewhere if and only if F concentrates around 0.

The second idea: the Gaussian measure is invariant under orthogonal transformations, that is, for any measurable $A \subset \mathbb{R}^n$, any orthogonal transformation $U : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, for $B = \big\{Ux : x \in A\big\}$ we have $\gamma_n(B) = \gamma_n(A)$. We will also state this as follows: if $x = (\xi_1, \ldots, \xi_n)$ is a vector of independent standard Gaussian random variables and U is an orthogonal transformation, then Ux is a vector of independent standard Gaussian random variables. In particular, we will use the following orthogonal transformation of $\mathbb{R}^{2n} = \mathbb{R}^n \oplus \mathbb{R}^n$:

$$(x, y) \longmapsto \big((\sin\theta)x + (\cos\theta)y,\ (\cos\theta)x - (\sin\theta)y\big).$$

Next, we will work with differentiable functions $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. Let

$$\nabla f = \left(\frac{\partial f}{\partial \xi_1}, \ldots, \frac{\partial f}{\partial \xi_n}\right)$$

denote the gradient of f. As usual,

$$\|\nabla f\|^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2$$

is the squared length of the gradient. If $\|\nabla f\| \leq 1$ then f is 1-Lipschitz. If f is 1-Lipschitz, it can be approximated by smooth functions with $\|\nabla f\| \leq 1+\epsilon$ (we will not discuss how and why).

For a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ we denote by $\mathbf{E}f$ its average with respect to the standard Gaussian measure in $\mathbb{R}^n$:

$$\mathbf{E}f = (2\pi)^{-n/2} \int_{\mathbb{R}^n} f(x) e^{-\|x\|^2/2} \ dx.$$

If f is the exponentiation of a linear function, the expectation can be computed exactly:

$$\mathbf{E}f = e^{\|a\|^2/2} \quad \text{for} \quad f(x) = e^{\langle a, x\rangle}.$$

We will repeatedly (namely, twice) use the following particular version of Jensen's inequality: if $f : X \longrightarrow \mathbb{R}$ is a function on a probability space (X, μ), then

$$\int_X e^f \ d\mu \geq \exp\Bigl\{\int_X f \ d\mu\Bigr\},$$

which follows since e^x is a convex function.

One, by now standard, feature of the result is that we use the Laplace transform estimate.

(15.1) Theorem. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a smooth function such that $\mathbf{E}(f) = 0$. Then*

$$\mathbf{E}\exp\{f\} \leq \mathbf{E}\exp\Bigl\{\frac{\pi^2}{8}\|\nabla f\|^2\Bigr\},$$

assuming that both integrals are finite.

Proof. Let us consider the second copy of $\mathbb{R}^n$ endowed with the standard Gaussian measure γ_n and let $\mathbb{R}^{2n} = \mathbb{R}^n \oplus \mathbb{R}^n$ endowed with the product measure γ_{2n}.

Let us define $F : \mathbb{R}^{2n} \longrightarrow \mathbb{R}$ by $F(x, y) = f(x) - f(y)$. We claim that

$$\mathbf{E}\exp\{f\} \leq \mathbf{E}\exp\{F\}$$

Indeed,

$$\begin{aligned}\mathbf{E}\exp\{F\} &= \int_{\mathbb{R}^n \oplus \mathbb{R}^n} e^{f(x)} e^{-f(y)} \ d\gamma_{2n}(x, y) \\ &= \int_{\mathbb{R}^n} e^{f(x)} \left(\int_{\mathbb{R}^n} e^{-f(y)} \ d\gamma_n(y)\right) \ d\gamma_n(x).\end{aligned}$$

Now, $\mathbf{E}(-f) = 0$ and hence by Jensen's inequality (first use) the inner integral is at least 1. Therefore, $\mathbf{E}e^F \geq \mathbf{E}e^f$ as claimed.

We will be proving that

$$\mathbf{E}\exp\{F\} \leq \mathbf{E}\exp\Bigl\{\frac{\pi^2}{8}\|\nabla f\|^2\Bigr\}.$$

Let us consider the following function

$$G : \mathbb{R}^n \times \mathbb{R}^n \times [0, \pi/2] \longrightarrow \mathbb{R}, \quad G(x, y, \theta) = f\Big((\sin\theta)x + (\cos\theta)y\Big).$$

Then

$$\begin{aligned} &G(x, y, 0) = f(y), \quad G(x, y, \pi/2) = f(x), \quad \text{and} \\ &F(x, y) = G(x, y, \pi/2) - G(x, y, 0) = \int_0^{\pi/2} \frac{\partial}{\partial\theta} G(x, y, \theta)\ d\theta. \end{aligned}$$

Therefore,

$$\mathbf{E} \exp\{F\} = \mathbf{E} \exp\Big\{ \int_0^{\pi/2} \frac{\partial}{\partial\theta} G(x, y, \theta)\ d\theta \Big\}.$$

Of course, we would like to apply Jensen's inequality once again and switch exp and the integral against $d\theta$. This can be done but with some care since $d\theta$ is not a probability measure on the interval $[0, \pi/2]$. To make it a probability measure, we have to replace $d\theta$ by $(2/\pi)\ d\theta$ thus multiplying the integrand by $\pi/2$. This leads to

$$\exp\Big\{ \int_0^{\pi/2} \frac{\partial}{\partial\theta} G(x, y, \theta)\ d\theta \Big\} \leq \frac{2}{\pi} \int_0^{\pi/2} \exp\Big\{ \frac{\pi}{2} \frac{\partial}{\partial\theta} G(x, y, \theta) \Big\}\ d\theta.$$

Hence

$$\mathbf{E} \exp\{F\} \leq \frac{2}{\pi} \int_0^{\pi/2} \mathbf{E} \exp\Big\{ \frac{\pi}{2} \frac{\partial}{\partial\theta} G(x, y, \theta) \Big\}\ d\theta.$$

Now (note that we switched two integrals in the process).

Now,

$$\begin{aligned} \frac{\partial}{\partial\theta} G(x, y, \theta) =& \langle \nabla f(x'), y' \rangle \quad \text{where} \\ & x' = (\sin\theta)x + (\cos\theta)y \quad \text{and} \quad y' = (\cos\theta)x - (\sin\theta)y. \end{aligned}$$

In words: we compute the gradient of f at the point $x' = (\sin\theta)x + (\cos\theta)y$ and take its scalar product with the vector $y' = (\cos\theta)x - (\sin\theta)y$.

Here is the punch-line: we want to compute the expectation of

$$\exp\Big\{ \frac{\pi}{2} \langle \nabla f(x'), y' \rangle \Big\}.$$

However, the vector (x', y') is obtained from the vector (x, y) by a rotation. Since (x, y) has the standard Gaussian distribution, (x', y') also has the standard Gaussian distribution. In particular, x' and y' are independent. Therefore, we can first take the expectation with respect to y' and then with respect to x'. But with respect to y', we are dealing with the exponentiation of a linear function. Therefore,

$$\mathbf{E}_{y'} \exp\Big\{ \frac{\pi}{2} \langle \nabla f(x'), y' \rangle \Big\} = \exp\Big\{ \frac{\pi^2}{8} \|\nabla f(x')\|^2 \Big\}.$$

In the end, we get

$$\mathbf{E}\exp\{F\} \leq \mathbf{E}_{x'} \frac{2}{\pi}\int_0^{\pi/2} \exp\Big\{\frac{\pi^2}{8}\|\nabla f(x')\|^2\Big\}\, d\theta = \mathbf{E}\exp\Big\{\frac{\pi^2}{8}\|\nabla f\|^2\Big\},$$

as claimed □

(15.2) Corollary. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a 1-Lipschitz function and let $a = \mathbf{E}f$ be the average value of f. Then*

$$\gamma_n\Big\{x : \quad f(x) - a \geq t\Big\} \leq \exp\Big\{-\frac{2t^2}{\pi^2}\Big\} \quad \textit{and}$$
$$\gamma_n\Big\{x : \quad f(x) - a \leq -t\Big\} \leq \exp\Big\{-\frac{2t^2}{\pi^2}\Big\}.$$

Proof. Without loss of generality, we assume that f is smooth and that $\|\nabla f\| \leq 1$.

We apply the Laplace transform method. To prove the first inequality, we choose a $\lambda \geq 0$ and claim that

$$\gamma_n\Big\{x : \quad f(x) - a \geq t\Big\} \leq e^{-\lambda t}\mathbf{E}e^{\lambda(f-a)} \leq e^{-\lambda t}\exp\Big\{\frac{\pi^2}{8}\lambda^2\Big\},$$

where the last inequality follows from Theorem 15.1 applied to $\lambda(f - a)$. Now we optimize on λ by substituting

$$\lambda = \frac{4t}{\pi^2}.$$

The second inequality is proved in a similar way. □

If we ask ourselves, what are the "internal reasons" why the proof of Theorem 15.1 worked, the following picture seems to emerge. Instead of proving that $f : X \longrightarrow \mathbb{R}$ concentrates about any particular number, say, its expectation or median, we prove that f concentrates "by itself", that is, that $|f(x)-f(y)|$ is small typically. For that, we double the space X to its direct product $X \times X$ with itself. This gives us more freedom to choose a path connecting x and y. We show that if we choose a path Γ the right way, the gradient of f will be more or less independent on the direction of Γ, so for typical x and y the change $|f(x) - f(y)|$ will not be that big.

(15.3) Concentration on the sphere as a corollary. One can obtain a concentration result for 1-Lipschitz functions on the unit sphere $\mathbb{S}^{n-1}$ by thinking of the invariant probability measure μ_n on $\mathbb{S}^{n-1}$ as the push-forward of the Gaussian measure γ_n via the radial projection $\mathbb{R}^n \setminus \{0\} \longrightarrow \mathbb{S}^{n-1}$. If $f : \mathbb{S}^{n-1} \longrightarrow \mathbb{R}$ is a 1-Lipschitz function with $\mathbf{E}f = 0$ then $\tilde{f} : \mathbb{R}^n \longrightarrow \mathbb{R}$ defined by $\tilde{f}(x) = \|x\|f(x/\|x\|)$ is a 3-Lipschitz on $\mathbb{R}^n$.

Lecture 21. Wednesday, February 23

---

16. Smoothing Lipschitz functions

In Section 15, we mentioned that any 1-Lipschitz function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ can be approximated by a smooth 1-Lipschitz function. In this section, we describe a simple way to do it.

(16.1) Theorem. *Let $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a 1-Lipschitz function. For $\epsilon > 0$ let*

$$B_\epsilon(x) = \Big\{ y : \quad \|y - x\| \leq \epsilon \Big\}$$

be the ball of radius ϵ centered at x. Let us define $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ by

$$f(x) = \frac{1}{\operatorname{vol} B_\epsilon(x)} \int_{B_\epsilon(x)} g(y) \ dy$$

(in words: $f(x)$ is the average of g over the ball of radius ϵ centered at x). Then

(1) *Function f is smooth and*

$$\|\nabla f(x)\| \leq 1 \quad \textit{for all} \quad x \in \mathbb{R}^n;$$

(2) *We have*

$$|f(x) - g(x)| \leq \frac{\epsilon n}{n+1} \leq \epsilon \quad \textit{for all} \quad x \in \mathbb{R}^n.$$

Proof. We check the case of $n = 1$ first when computations are especially simple. In this case, $g : \mathbb{R} \longrightarrow \mathbb{R}$ is a 1-Lipschitz function and $f : \mathbb{R} \longrightarrow \mathbb{R}$ is defined by

$$f(x) = \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} g(y) \ dy.$$

Therefore, we have

$$f'(x) = \frac{g(x+\epsilon) - g(x-\epsilon)}{2\epsilon}.$$

Since g is 1-Lipschitz, we have $|f'(x)| \leq 1$ for all $x \in \mathbb{R}$ and

$$\begin{aligned}
|f(x) - g(x)| =& \Big| \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} g(y) \ dy - \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} g(x) \ dy \Big| \\
\leq& \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} |g(y) - g(x)| \ dy \leq \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} |y - x| \ dy \\
=& \frac{\epsilon^2}{2\epsilon} = \frac{\epsilon}{2}.
\end{aligned}$$

Suppose now that $n > 1$. Let us choose a unit vector $u \in \mathbb{R}^n$. Strictly speaking, we will not prove that f is smooth, although it can be deduced from our construction. Instead, we are going to prove that f is differentiable in the direction of u and that the derivative of f in the direction of u does not exceed 1 in the absolute value, that is,

$$|\langle \nabla f, u \rangle| \leq 1.$$

Without loss of generality, we assume that $x = 0$ and that $u = (0, \dots, 0, 1)$. Then the orthogonal complement $u^{\perp}$ can be identified with $\mathbb{R}^{n-1}$.

Let

$$D_\epsilon = \Big\{ z \in \mathbb{R}^{n-1} : \quad \|z\| < \epsilon \Big\}$$

be the open ball centered at 0 of radius ϵ in $\mathbb{R}^{n-1}$. Then, for any $z \in D_\epsilon$, the intersection of the line in the direction of u through z and $B_\epsilon = B_\epsilon(0)$ is the interval with the endpoints $\big(z, -a(z)\big)$ and $\big(z, a(z)\big)$, where $a(z) > 0$.

We note that

$$\operatorname{vol} B_\epsilon = \int_{D_\epsilon} 2a(z) \; dz. \tag{16.1.1}$$

Let us estimate the derivative of f in the direction of u, that is $\langle \nabla f, \; u \rangle$. For any $\tau \in \mathbb{R}$, we have

$$f(\tau u) = \frac{1}{\operatorname{vol} B_\epsilon} \int_{B_\epsilon + \tau u} g(z) \; dz = \frac{1}{\operatorname{vol} B_\epsilon} \int_{D_\epsilon} \left(\int_{-a(z)+\tau}^{a(z)+\tau} g\big(z, \xi\big) \; d\xi \right) \; dz$$

and

$$\frac{\partial f(\tau u)}{\partial \tau}\Big|_{\tau=0} = \frac{1}{\operatorname{vol} B_\epsilon} \int_{D_\epsilon} g\big(z, a(z)\big) - g\big(z, -a(z)\big) \; dz.$$

Since g is 1-Lipschitz, we have $|g\big(z, a(z)\big) - g\big(z, -a(z)\big)| \leq 2a(z)$. Therefore, in view of (16.1.1),

$$\frac{\partial f(\tau u)}{\partial \tau}\Big|_{\tau=0} = \langle \nabla f, u \rangle \leq 1$$

and since u was arbitrary, we get Part (1).

To prove Part (2), we write

$$\begin{aligned} |f(0) - g(0)| =& \Big| \frac{1}{\operatorname{vol} B_\epsilon} \int_{B_\epsilon} g(y) - g(0) \; dy \Big| \leq \frac{1}{\operatorname{vol} B_\epsilon} \int_{B_\epsilon} |g(y) - g(0)| \; dy \\ \leq& \frac{1}{\operatorname{vol} B_\epsilon} \int_{B_\epsilon} \|y\| \; dy = \frac{|\mathbb{S}^{n-1}|}{\operatorname{vol} B_\epsilon} \int_0^\epsilon r^n \; dr = \frac{\epsilon |\mathbb{S}^{n-1}|}{(n+1) \operatorname{vol} B_1}. \end{aligned}$$

It is not hard to show that

$$|\mathbb{S}^{n-1}| = n \operatorname{vol} B_1,$$

from which Part (2) follows. □

PROBLEM. Prove that

$$\|\nabla f(x) - \nabla f(y)\| \leq \frac{c\sqrt{n}}{\epsilon}\|x-y\|$$

for some absolute constant $c > 0$.

17. Concentration on the sphere and functions almost constant on a subspace

We discuss one notable corollary of the concentration on the sphere. It claims that for any fixed $\epsilon > 0$ and any 1-Lipschitz function $f : \mathbb{S}^{n-1} \longrightarrow \mathbb{R}$ there exists a subspace $L \subset \mathbb{R}^n$ of dimension linear in n such that the restriction of f onto L is ϵ-close to a constant on the unit sphere $L \cap \mathbb{S}^{n-1}$ in L.

(17.1) Theorem. *There exists an absolute constant $\kappa > 0$ with the following property. For any $\epsilon > 0$, for any n, for any 1-Lipschitz function $f : \mathbb{S}^{n-1} \longrightarrow \mathbb{R}$, there exists a number c (which can be chosen to be the median of f or the average value of f) and a subspace $L \subset \mathbb{R}^n$ such that*

$$|f(x) - c| \leq \epsilon \quad \textit{for all} \quad x \in L \cap \mathbb{S}^{n-1}$$

and

$$\dim L \geq \frac{\kappa \epsilon^2}{\ln(1/\epsilon)} n.$$

One crucial argument in the proof is the existence of moderately sized δ-nets. It is interesting in its own right.

(17.2) Lemma. *Let V be an n-dimensional normed space with norm $\|\cdot\|$ and let*

$$\Sigma = \Big\{x \in V : \quad \|x\| = 1\Big\}$$

be the unit sphere in V. Then, for any $\delta > 0$ there exists a set $S \subset \Sigma$ such that

1. *For every $x \in \Sigma$ there is a $y \in S$ such that $\|x - y\| \leq \delta$, so S is a δ-net in Σ;*
2. *We have*

$$|S| \leq \left(1 + \frac{2}{\delta}\right)^n$$

for the cardinality $|S|$ of S.

Similarly, for the unit ball

$$B = \Big\{x \in V : \quad \|x\| \leq 1\Big\}$$

there exists a δ-net $S \subset B$ of cardinality $|S| \leq \left(1 + \frac{2}{\delta}\right)^n$.

Proof. Let us choose $S \subset \Sigma$ to be the maximal subset with the property that $\|y_1 - y_2\| > \delta$ for any two points $y_1, y_2 \in S$. In other words, we cannot include into S any additional point $x \in \Sigma$ without this property being violated. Clearly, S must be a δ-net for Σ. For every $y \in S$, let us consider a ball $B(y, \delta/2)$ of radius $\delta/2$ centered at y. Then the balls are pairwise disjoint:

$$B(y_1, \delta/2) \cap B(y_2, \delta/2) = \emptyset \quad \text{provided} \quad y_1 \neq y_2.$$

Moreover, each ball $B(y, \delta/2)$ lies within the ball centered at the origin and of radius $1 + \delta/2 = (2 + \delta)/2$, so

$$\bigcup_{y \in S} B(y, \delta/2) \subset B\left(0, 1 + \frac{\delta}{2}\right) = B\left(0, \frac{2+\delta}{2}\right).$$

Estimating the volume of the union, we get that

$$|S| \operatorname{vol} B(0, \delta/2) \leq \operatorname{vol} B\left(0, \frac{2+\delta}{2}\right).$$

Since the volume of the n-dimensional ball is proportional to the nth power of the radius, we get

$$|S| \left(\frac{\delta}{2}\right)^n \leq \left(\frac{2+\delta}{2}\right)^n,$$

from which (2) follows. □

Now we are ready to prove Theorem 17.1.

Proof. Let us choose a k-dimensional subspace $A \subset \mathbb{R}^n$ (k will be adjusted later). The goal is to prove that for a random orthogonal transformation $U \in O_n$, the "rotated" subspace $L = U(A) = \Big\{Ux : \ x \in A\Big\}$ will satisfy the desired properties. For that, let us choose an $\epsilon/2$-net $S \subset A \cap \mathbb{S}^{n-1}$. As follows from Lemma 17.2, we can choose S such that

$$|S| \leq \left(1 + 4\epsilon^{-1}\right)^k = \exp\Big\{k \ln(1 + 4\epsilon^{-1})\Big\}.$$

Let

$$X = \Big\{x \in \mathbb{S}^{n-1} : \quad |f(x) - c| \leq \epsilon/2\Big\}.$$

As follows from the concentration results (cf. Theorem 14.3 or Section 15.3),

$$\mu(X) \geq 1 - e^{-\alpha n \epsilon^2}$$

for some absolute constant $\alpha > 0$.

If we manage to find an orthogonal transformation U such that $U(S) \subset X$ then for $L = U(A)$ the restriction of f onto $L \cap \mathbb{S}^{n-1}$ does not deviate from c by more than ϵ. Indeed, for every $x \in L \cap \mathbb{S}^{n-1}$ there is a $y \in U(S)$ such that $\|y - x\| \leq \epsilon/2$ (since $U(S)$ is an $\epsilon/2$-net for $L \cap \mathbb{S}^{n-1}$) and $|f(y) - c| \leq \epsilon/2$ (since $y \in X$). Since f is 1-Lipschitz, we get that $|f(x) - c| \leq \epsilon$.

Now, since X is pretty big and S is reasonably small, one can hope that a random orthogonal transformation U will do.

Let $\nu = \nu_n$ be the Haar probability measure on the orthogonal group O_n (cf. Section 3.5). Let us pick a particular $x \in S$. As U ranges over the group O_n, the point Ux ranges over the unit sphere $\mathbb{S}^{n-1}$. Therefore,

$$\nu\Big\{U : \quad Ux \notin X\Big\} = \mu\left(\mathbb{S}^{n-1} \setminus X\right) \leq e^{-\alpha n \epsilon^2}$$

(we used a similar reasoning in Section 3). Therefore,

$$\nu\Big\{U : \quad Ux \notin X \quad \text{for some} \quad x \in S\Big\} \leq |S| e^{-\alpha n \epsilon^2} \leq \exp\Big\{k \ln(1 + 4\epsilon^{-1}) - \alpha n \epsilon^2\Big\}.$$

We can make the upper bound less than 1 by choosing

$$k = O\left(\frac{\epsilon^2 n}{\ln(1/\epsilon)}\right).$$

□

Lecture 22. Friday, February 25

18. Dvoretzky's Theorem

We discuss one of the first and most famous applications of measure concentration, Dvoretzky's Theorem, conjectured by A. Grothendieck in 1956, proved by A. Dvoretzky in 1961, reproved by V. Milman in 1971 using the concentration of measure on the unit sphere, and by T. Figiel, J. Lindenstrauss, and V. Milman with better constants and broad extensions and ramifications in 1977.

(18.1) Definition. Let V be an n-dimensional normed space with norm $\| \cdot \|$. We say that V is ϵ-close to Euclidean space $\mathbb{R}^n$ if there exists a vector spaces isomorphism $\phi : V \longrightarrow \mathbb{R}^n$ such that

$$(1 - \epsilon)\|\phi(x)\|_{\mathbb{R}^n} \leq \|x\|_V \leq (1 + \epsilon)\|\phi(x)\|_{\mathbb{R}^n}, \quad \text{for all} \quad x \in V$$

where $\|x\|_V$ is measured with respect to the norm in V and $\|\phi(x)\|_{\mathbb{R}^n}$ is measured with respect to the Euclidean norm in $\mathbb{R}^n$.

First we state an infinite-dimensional version of the theorem.

(18.2) Theorem. *Let W be an infinite-dimensional normed space with norm $\| \cdot \|$ and let $\epsilon > 0$ be a number. Then for any positive integer n there exists a subspace $V \subset W$ with $\dim V = n$ and such that the space V with the norm $\| \cdot \|$ inherited from W is ϵ-close to Euclidean space $\mathbb{R}^n$.*

This does sound counterintuitive: for example, choose W to be the space of continuous functions $f : [0,1] \longrightarrow \mathbb{R}$ with the norm

$$\|f\| = \max_{0 \leq x \leq 1} |f(x)|$$

choose $\epsilon = 0.1$ and try to present an n-dimensional subspace V of W ϵ-close to Euclidean space.

Theorem 18.2 is deduced from its finite-dimensional version.

(18.3) Theorem. *For any $\epsilon > 0$ and any positive integer k there exists a positive integer $N = N(k, \epsilon)$ such that for any normed space W with $\dim W \geq N$ there exists a k-dimensional subspace V of W which is ϵ-close to Euclidean space $\mathbb{R}^k$.*

Currently, the best bound for N is $N = \exp\Big\{O\left(\epsilon^{-2}k\right)\Big\}$. We will not prove Dvoretzky's Theorem but explain how one could get a slightly weaker bound $N = \exp\{O\left(\epsilon^{-2}\ln(1/\epsilon)k\right)\Big\}$. The exponential dependence on k is optimal, the corresponding example is given by $W = \ell_N^\infty$ consisting of all N-tuples $x = (\xi_1, \dots, \xi_N)$ of real numbers with the norm

$$\|x\|_\infty = \max_{i=1,\dots,N} |\xi_i|.$$

The main thrust of the proof is Theorem 17.1.

(18.4) A plan of the proof of Dvoretzky's Theorem. Let W be an n-dimensional normed space with norm $\| \cdot \|_W$. Our goal is to find a k-dimensional subspace $V \subset W$ which is ϵ-close to Euclidean space.

Step 1. *Let us round.* We may think of W as of $\mathbb{R}^n$ endowed with the norm $\| \cdot \|_W$ as opposed to the standard Euclidean norm $\| \cdot \|$. This way, we can freely operate with the standard Euclidean scalar product $\langle x, y \rangle$ in W.

An *ellipsoid* E centered at $a \in \mathbb{R}^n$ is a set of the type

$$E = \Big\{x \in \mathbb{R}^n : \quad q(x-a) \leq 1\Big\},$$

where $q : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a positive definite quadratic form. A famous result due to F. John states that every convex body K contains a *unique* ellipsoid of the maximum volume and is contained in a *unique* ellipsoid of the minimum volume. If the convex body is symmetric about the origin, that is, $K = -K$, then the situation is especially attractive: in this case, both ellipsoids have to be symmetric and so have to be centered at the origin. If $E \subset K$ is the maximum volume ellipsoid then $K \subset \sqrt{n}E$ and if $E \supset K$ is the minimum volume ellipsoid then $n^{-1/2}E \subset K$ (if K is not symmetric, we should dilate by a factor of n in the worst case).

Let

$$K = \Big\{ x \in \mathbb{R}^n : \quad \|x\|_W \leq 1 \Big\}$$

be the unit ball of the norm $\| \cdot \|_W$. We find the maximum volume ellipsoid $E \subset K$ and choose a new scalar product in $\mathbb{R}^n$ in which E is the standard unit ball B:

$$B = \Big\{ x \in \mathbb{R}^n : \quad \|x\| \leq 1 \Big\}.$$

Hence we have $B \subset K \subset \sqrt{n}B$. After this step, we can be more specific: we will look for a k-dimensional subspace $V \subset \mathbb{R}^n$ for which the restriction of $\| \cdot \|_W$ onto V is ϵ-close to the restriction of the Euclidean norm $\| \cdot \|$ onto V.

Step 2. *Let us dualize.* Let

$$K^\circ = \Big\{ x \in \mathbb{R}^n : \quad \langle x, y \rangle \leq 1 \quad \text{for all} \quad y \in D \Big\}$$

be the polar dual of D. The standard duality argument implies that

$$\|x\|_W = \max_{y \in K^\circ} \langle x, y \rangle.$$

Besides,

$$n^{-1/2}B \subset K^\circ \subset B.$$

Step 3. *Let us estimate the average or the median.* Let us consider the norm $\|x\|_W$ as a function on the unit sphere $\mathbb{S}^{n-1}$. We want to obtain a lower bound for its median m or average a on the sphere. Since K° contains the ball of radius $n^{-1/2}$, the median is at least $n^{-1/2}$, but this bound is too weak for our purposes. A stronger (though still correct) bound is

$$m, a \geq \alpha \sqrt{\frac{\ln n}{n}}$$

for some absolute constant $\alpha > 0$. This is deduced from Dvoretzky-Rogers Lemma, which we don't discuss here. Instead, we discuss some intuitive reasons where an extra $\sqrt{\ln n}$ appears from.

One can argue that B is the smallest volume ellipsoid containing K°. Therefore, there must be some points of K° on the boundary of B, and, in some sense, they should be spread sufficiently evenly on the surface of B. If, for example, $\partial B \cap K^\circ$

contains an orthonormal basis (which happens, for example, if K is a cube so K° is an octahedron), then

$$\|x\|_W \geq \max_{i=1,\dots,n} |\xi_i| \quad \text{for} \quad x = (\xi_1, \dots, \xi_n)$$

and the average value of $\|x_W\|$ is at least as large as

$$\int_{\mathbb{S}^{n-1}} \max_{i=1,\dots,n} |\xi_i| \; d\mu_n(x).$$

It is not hard to argue that the integral is of the order of

$$c\sqrt{\frac{\ln}{n}}$$

for some positive c. Perhaps the easiest way to deduce this is by passing to the Gaussian measure and proving that

$$\int_{\mathbb{R}^n} \max_{i=1,\dots,n} |\xi_i| \; d\gamma_n(x) \sim \sqrt{\ln n} \quad \text{as} \quad n \longrightarrow +\infty.$$

In general, it is not true that K° contains n orthonormal vectors on the boundary of B, but something close to it is true. F. John's criterion for the optimality of the minimum volume ellipsoid states that there are vectors $x_i \in \partial B \cap K^\circ, i \in I$ such that

$$\sum_{i \in I} \lambda_i x_i \otimes x_i = I,$$

where $\lambda_i \geq 0$, $x \otimes x$ denotes the square matrix with the entries $\xi_i \xi_j$ for $x = (\xi_1, \dots, \xi_n)$ and I is the identity matrix. This turns out to be enough to produce enough (in fact, about $n/2$) orthonormal vectors $x_i \in K_i^\circ$ that are sufficiently long ($\|x_i\| \geq 1/2$). This is what Dvoretzky-Rogers Lemma is about, and this is enough to establish the lower bound.

Step 4. *Let us apply (modified) Theorem 17.1.* Considering $\|x\|_W$ as a function on the unit sphere $\mathbb{S}^{n-1}$, we would like to use Theorem 17.1 to claim that there exists a section of the sphere of a moderately high dimension such that the restriction of $\|x\|_W$ onto that section is almost a constant. This would do the job. However, if we just view $\|x\|_W$ as a 1-Lipschitz function (which it is) without any special properties (which it does have) and just apply Theorem 17.1, we'll get nothing.

We will go back to the proof of Theorem 17.1 and use that $\|x\|_W$ is homogeneous of degree 1 and convex with the goal of replacing the additive error by the multiplicative error.

Let a be the average or the median of $\|x\|_W$ on $\mathbb{S}^{n-1}$, so

$$a = \Omega\left(\sqrt{\frac{\ln}{n}}\right).$$

We are given an $\epsilon > 0$ and an integer k. We would like to prove that if n is large enough, there will be a k-dimensional subspace L for which the restriction of $\| \cdot \|_W$ onto L is ϵ-close to the average. Let us choose a k-dimensional subspace $A \subset \mathbb{R}^n$ and consider the intersection $\mathbb{S}^{k-1} = A \cap \mathbb{S}^{n-1}$. Let us choose a (small) $\delta = \delta(\epsilon)$ to be adjusted later and let us construct a net $S \subset \mathbb{S}^{k-1}$ such that

$$\operatorname{conv}(S) \subset \mathbb{S}^{k-1} \subset (1+\delta) \operatorname{conv}(S),$$

where conv is the convex hull. If we manage to rotate $A \longmapsto L = U(A)$ by an orthogonal transformation U in such a way that

$$(1-\delta)a \leq \|U(x)\|_W \leq (1+\delta)a \quad \text{for all} \quad x \in S,$$

we'll have

$$(1-\delta)a \leq \|x\|_W \leq (1+\delta)^2 a \quad \text{for all} \quad x \in L.$$

Thus we should choose $\delta \leq \epsilon$ such that $(1+\delta)^2 < (1+\epsilon)$.

The net S, however large, is finite. Going back to the proof of Theorem 17.1, we need to estimate the probability, that for every particular $x \in X$, for a random orthogonal transformation U we don't have

$$(1-\delta)a \leq \|U(x)\|_W \leq (1+\delta)a.$$

From the concentration inequality on the sphere, this probability is at most

$$\exp\Big\{-\Omega(n\delta^2 a^2)\Big\} = \exp\Big\{-\Omega\left(n\delta^2 \frac{\ln n}{n}\right)\Big\} = \exp\Big\{-\Omega(\ln n)\Big\},$$

that is, tends to 0 as n grows. This implies that for a sufficiently large n, we will be able to produce such a transformation U.

Lecture 23. Monday, March 7

---

18. Dvoretzky's Theorem, continued

The following simple estimate shows that the dimension of an "almost Euclidean section" $V \subset W$ cannot grow faster than $\log \dim W$ in the worst case.

(18.5) Lemma. *Let $B \subset \mathbb{R}^{n+2}$ be the unit ball and let $P \subset \mathbb{R}^{n+2}$ be a polyhedron defined by m inequalities*

$$P = \Big\{x \in \mathbb{R}^{n+2} : \quad \langle u_i, x \rangle \leq \alpha_i \quad \textit{for} \quad i = 1, \ldots, m\Big\}.$$

Suppose that $\|u_i\| = 1$ for $i = 1, \ldots, m$ and that $\alpha_i \geq 1$, so $B \subset P$. Suppose further, that $P \subset \rho B$ for some $\rho \geq 1$. Then

$$2\rho^2 \ln m \geq n.$$

Proof. Let us choose a $t > 0$ such that $me^{-t^2 n/2} < 1$ and let us consider the spherical caps of radius $\pi/2 - t$ centered at u_i for $i = 1, \ldots, m$. As follows by Lemma 14.2, these caps fail to cover the whole sphere. Therefore, there is a point $v \in \mathbb{S}^{n+1}$ such that

$$\langle v, u_i \rangle < \cos\left(\frac{\pi}{2} - t\right) = \sin t < t \quad \text{for} \quad i = 1, \ldots, m.$$

Therefore, $t^{-1}v \in P$ and so $t^{-1} \leq \rho$.

Now, we can choose any

$$t \geq \sqrt{\frac{2 \ln m}{n}}$$

and so

$$\sqrt{\frac{n}{2 \ln m}} \leq \rho,$$

which completes the proof. □

(18.6) Almost Euclidean sections of ℓ^∞. Let W be the space $\mathbb{R}^n$ albeit with the norm

$$\|x\|_W = \max_{i=1,\ldots,n} |\xi_i| \quad \text{for} \quad x = (\xi_1, \ldots, \xi_n).$$

Such a space W is called the ℓ^∞_n space. The unit ball

$$P = \Big\{ x \in W : \quad \|x\|_W \leq 1 \Big\}$$

is the cube $|\xi_i| \leq 1$ defined by $2n$ inequalities. If we manage to find an m-dimensional subspace $V \subset W$ which is ϵ-close to Euclidean space, then, after a linear transformation, the polyhedron $P \cap V$ contains the Euclidean unit ball B and is contained in ρB for $\rho = (1+\epsilon)/(1-\epsilon) \approx 1 + 2\epsilon$ for $\epsilon \approx 0$. Applying Lemma 18.5, we get

$$2\rho^2 \ln(2n) \geq m + 2.$$

Thus we must have

$$m = O(\ln n).$$

Thus the dimension of an almost Euclidean section cannot grow faster than the logarithm of the dimension of the ambient space in the worst case. However, somewhat disappointingly, we are unable to recover the worst possible dependence on ϵ.

PROBLEM. Let W be the space $\mathbb{R}^n$ with the norm

$$\|x\|_W = \sum_{i=1}^n |\xi_i| \quad \text{for} \quad x = (\xi_1, \ldots , \xi_n).$$

Such a space W is called the ℓ_n^1 space. Let $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ be the Euclidean unit sphere in $\mathbb{R}^n$. Let c be the average value of $\|x\|_W$ on $\mathbb{S}^{n-1}$. Prove that $c = \sqrt{2n/\pi}$ and deduce from Theorem 17.1 that for any $\epsilon > 0$ space W has a subspace $V \subset W$, which is ϵ-close to Euclidean space and such that $\dim V = \Omega(n)$.

In fact, there is a subspace of dimension $n/2$ which is constant-close to Euclidean space. This cannot be deduced from Theorem 17.1 and requires a different technique.

Lecture 24. Wednesday, March 9

---

19. The Prékopa-Leindler Inequality

We prove a very useful inequality, called the Prékopa-Leindler inequality, which allows us to establish concentration in a wide variety of situations.

(19.1) Theorem. *Let $f, g, h : \mathbb{R}^n \longrightarrow \mathbb{R}$ be non-negative integrable functions and let $\alpha, \beta > 0$ be numbers such that $\alpha + \beta = 1$ and*

$$h(\alpha x + \beta y) \geq f^{\alpha}(x) g^{\beta}(y) \quad \textit{for all} \quad x, y \in \mathbb{R}^n.$$

Then

$$\int_{\mathbb{R}^n} h \ dx \geq \left(\int_{\mathbb{R}^n} f(x) \ dx \right)^{\alpha} \left(\int_{\mathbb{R}^n} g(x) \ dx \right)^{\beta}.$$

Proof. The proof is by induction on the dimension n.

Suppose that $n = 1$. We may assume that $f(x) > 0$ and $g(x) > 0$ for all $x \in \mathbb{R}$. Scaling, if needed, we assume that

$$\int_{-\infty}^{+\infty} f(x) \ dx = \int_{-\infty}^{+\infty} g(x) \ dx = 1.$$

This allows us to think of f and g as *densities* of probability distributions with strictly increasing cumulative distribution functions

$$F(t) = \int_{-\infty}^{t} f(x) \ dx \quad \text{and} \quad G(t) = \int_{-\infty}^{t} g(x) \ dx.$$

Thus, $F, G : \mathbb{R} \longrightarrow (0,1)$. Let $u(t)$ be the inverse of F and $v(t)$ be the inverse of G, so $u, v : (0,1) \longrightarrow \mathbb{R}$ and

$$\int_{-\infty}^{u(t)} f(x)\ dx = t \quad \text{and} \quad \int_{-\infty}^{v(t)} g(x)\ dx = t \quad \text{for} \quad t \in (0,1).$$

Besides, $u(t)$ and $v(t)$ are both smooth and increasing. Differentiating both integrals, we get

$$u'(t) f\big(u(t)\big) = v'(t) g\big(v(t)\big) = 1. \tag{19.1.1}$$

Let

$$w(t) = \alpha u(t) + \beta v(t).$$

Then $w(t)$ is smooth and increasing. Furthermore,

$$w'(t) = \alpha u'(t) + \beta v'(t) \geq \big(u'(t)\big)^{\alpha} \big(v'(t)\big)^{\beta},$$

since ln is a concave function. In particular, $w : (0,1) \longrightarrow \mathbb{R}$ is a smooth increasing function. Let us make the substitution $x = w(t)$ in the integral of h over $\mathbb{R}$. We can write

$$\begin{aligned}\int_{-\infty}^{+\infty} h(x)\ dx = \int_0^1 h\big(w(t)\big) w'(t)\ dt \geq \int_0^1 f\big(u(t)\big)^{\alpha} g\big(v(t)\big)^{\beta} \big(u'(t)\big)^{\alpha} \big(v'(t)\big)^{\beta}\ dt \\ = \int_0^1 \big(f\big(u(t)\big) u'(t)\big)^{\alpha} \big(g\big(v(t)\big) v'(t)\big)^{\beta}\ dt = 1,\end{aligned}$$

where we used (19.1.1) in the last equality. This proves the inequality in the case of $n = 1$.

Suppose that $n > 1$. Let us slice $\mathbb{R}^n$ into flats $\xi_n = \tau$. Each such a flat can be identified with $\mathbb{R}^{n-1}$. Let us define

$$\begin{aligned} f_1(\tau) &= \int_{\mathbb{R}^{n-1}} f(y, \tau)\ dy, \quad g_1(\tau) = \int_{\mathbb{R}^{n-1}} g(y, \tau)\ dy, \quad \text{and} \\ h_1(\tau) &= \int_{\mathbb{R}^{n-1}} h(y, \tau)\ dy, \end{aligned}$$

where dy is the Lebesgue measure on $\mathbb{R}^{n-1}$. Then f_1, g_1, and h_1 are univariate non-negative integrable functions. Let us fix some $\tau_1, \tau_2 \in \mathbb{R}$. Then for any $y_1, y_2 \in \mathbb{R}^{n-1}$ we have

$$h(\alpha y_1 + \beta y_2, \alpha \tau_1 + \beta \tau_2) \geq f(y_1, \tau_1)^{\alpha} g(y_2, \tau_2)^{\beta}.$$

Applying the induction hypothesis to the $(n-1)$-variate functions $h(\cdot, \alpha\tau_1 + \beta\tau_2)$, $f(\cdot, \tau_1)$, and $g(\cdot, \tau_2)$, we get

$$h_1(\alpha\tau_1 + \beta\tau_2) \geq \left(f_1(\tau_1)\right)^{\alpha} \left(g_1(\tau_2)\right)^{\beta}.$$

Therefore, by the already proven univariate case,

$$\int_{-\infty}^{+\infty} h_1(\tau)\ d\tau \geq \left(\int_{-\infty}^{\infty} f_1(\tau)\ d\tau\right)^{\alpha} \left(\int_{-\infty}^{+\infty} g_1(\tau)\ d\tau\right)^{\beta}.$$

However, by Fubini's Theorem,

$$\int_{-\infty}^{+\infty} h_1(\tau)\ d\tau = \int_{\mathbb{R}^n} h(x)\ dx, \quad \int_{-\infty}^{\infty} f_1(\tau)\ d\tau = \int_{\mathbb{R}^n} f(x)\ dx, \quad \text{and}$$
$$\int_{-\infty}^{\infty} g_1(\tau)\ d\tau = \int_{\mathbb{R}^n} g(x)\ dx.$$

This completes the proof. □

20. Logarithmically concave measures. The Brunn-Minkowski inequality

We start with some definitions.

(20.1) Definitions. A function $p : \mathbb{R}^n \longrightarrow \mathbb{R}$ is called *logarithmically concave* or *log-concave* if it is non-negative: $p(x) \geq 0$ for all $x \in \mathbb{R}^n$ and

$$p(\alpha x + \beta y) \geq p^{\alpha}(x) p^{\beta}(y) \quad \text{for all} \quad \alpha, \beta \geq 0 \quad \text{such that} \quad \alpha + \beta = 1$$
$$\text{and all} \quad x, y \in \mathbb{R}^n.$$

We agree that $0^0 = 1$. Equivalently, p is log-concave if $p(x) \geq 0$ and $\ln f$ is concave.

A measure μ on $\mathbb{R}^n$ is called *log-concave* if μ has a density which is a measurable log-concave function f. Examples of log-concave measures include the Lebesgue measure dx and the standard Gaussian measure γ_n.

Let $A, B \subset \mathbb{R}^n$ be sets. The set

$$A + B = \Big\{ x + y : \quad x \in A \quad \text{and} \quad y \in B \Big\}$$

is called the *Minkowski sum* of A and B.

For a $X \subset \mathbb{R}^n$, let $[X] : \mathbb{R}^n \longrightarrow \mathbb{R}$ be the indicator function of X:

$$[X](x) = \begin{cases} 1 & \text{if } x \in X \\ 0 & \text{if } x \notin X. \end{cases}$$

PROBLEMS.

1. Prove that the product of log-concave functions is log-concave.

2. Let $A \subset \mathbb{R}^n$ be a convex body (a compact convex set with non-empty interior). Let us define a measure ν on $\mathbb{R}^n$ by $\nu(X) = \text{vol}(X \cap A)$. Prove that ν is a log-concave measure.

3. Prove that the Minkowski sum of convex sets is a convex set.

4. Let $A, B \subset \mathbb{R}^n$ be sets and let α be a number. Prove that $\alpha(A+B) = \alpha A+\alpha B$.

5. Let A be a convex set and let $\alpha, \beta \geq 0$ be numbers. Prove that $(\alpha+\beta)A = \alpha A + \beta A$. Prove that in general the convexity and non-negativity assumptions cannot be dropped.

The following is the famous Brunn-Minkowski inequality.

(20.2) Theorem. *Let μ be a log-concave measure on $\mathbb{R}^n$, let $\alpha, \beta \geq 0$ be numbers such that $\alpha+\beta = 1$, and let $A, B \subset \mathbb{R}^n$ be measurable sets such that the set $\alpha A+\beta B$ is measurable. Then*

$$\mu(\alpha A + \beta B) \geq \mu^{\alpha}(A)\mu^{\beta}(B) \quad \textit{for all} \quad \alpha, \beta \geq 0 \quad \textit{such that} \quad \alpha+\beta = 1.$$

Proof. Let p be the log-concave density of μ. Let us define functions $f, g, h : \mathbb{R}^n \longrightarrow \mathbb{R}$ by $f = p[A]$, $g = p[B]$, and $h = p[\alpha A + \beta B]$ (we mean the product of the density and indicator functions). Then

$$\begin{aligned} h(\alpha x + \beta y) \geq f^{\alpha}(x) g^{\beta}(y) \quad &\text{for all} \quad \alpha, \beta \geq 0 \quad \text{such that} \quad \alpha+\beta = 1 \\ &\text{and all} \quad x, y \in \mathbb{R}^n. \end{aligned}$$

Now we apply the Prékopa-Leindler Inequality (Theorem 19.1) and notice that

$$\int_{\mathbb{R}^n} f(x)\ dx = \mu(A), \quad \int_{\mathbb{R}^n} g(x)\ dx = \mu(B), \quad \text{and} \quad \int_{\mathbb{R}^n} h(x)\ dx = \mu(\alpha A + \beta B).$$

□

In the classical case of the Lebesgue measure μ, the Brunn-Minkowski inequality is often stated in the equivalent additive form:

(20.3) Corollary. *Let μ be the Lebesgue meausre on $\mathbb{R}^n$ and let $A, B \subset \mathbb{R}^n$ be measurable sets such that $\alpha A + \beta B$ is measurable for all $\alpha, \beta \geq 0$. Then*

$$\mu^{1/n}(A+B) \geq \mu^{1/n}(A) + \mu^{1/n}(B).$$

Proof. Without loss of generality, we assume that $\mu(A), \mu(B) > 0$. Let

$$\alpha = \frac{\mu^{1/n}(A)}{\mu^{1/n}(A) + \mu^{1/n}(B)} \quad \text{and} \quad \beta = \frac{\mu^{1/n}(B)}{\mu^{1/n}(A) + \mu^{1/n}(B)},$$

so $\alpha, \beta \geq 0$ and $\alpha+\beta = 1$. Let $A_1 = \alpha^{-1}A$ and $B_1 = \beta^{-1}B$. Then

$$\begin{aligned} \mu(A_1) = \alpha^{-n}\mu(A) = \Big(\mu^{1/n}(A) + \mu^{1/n}(B)\Big)^n \quad &\text{and} \\ \mu(B_1) = \beta^{-n}\mu(B) = \Big(\mu^{1/n}(A) + \mu^{1/n}(B)\Big)^n. \end{aligned}$$

Here we used that the Lebesgue measure is homogenous of degree n, that is, $\mu(tA) = t^n \mu(A)$. Applying Theorem 20.2, we get

$$\mu(A+B) = \mu(\alpha A_1 + \beta B_1) \geq \mu^{\alpha}(A_1)\mu^{\beta}(B_1) = \left(\mu^{1/n}(A) + \mu^{1/n}(B)\right)^n .$$

The proof now follows. □

PROBLEM. Let μ be a measure on $\mathbb{R}^n$ such that $\mu(tA) = t^n\mu(A)$ for all measurable $A \subset \mathbb{R}^n$ and all $t \geq 0$. Let $\alpha, \beta \geq 0$ be numbers such that $\alpha + \beta = 1$. Suppose further that $\mu^{1/n}(\alpha A + \beta B) \geq \mu^{1/n}(\alpha A) + \mu^{1/n}(\beta B)$ for some measurable A and B such that $\alpha A + \beta B$ is measurable. Prove that $\mu(\alpha A + \beta B) \geq \mu^{\alpha}(A)\mu^{\beta}(B)$.

Lecture 25. Friday, March 11

21. The isoperimetric inequality for the Lebesgue measure in $\mathbb{R}^n$

The Brunn-Minkowski inequality provides a simple solution for the isoperimetric problem for the Lebesgue measure in Euclidean space $\mathbb{R}^n$. For a closed set $A \subset \mathbb{R}^n$ and a number $\rho \geq 0$, let us define the ρ-neighborhood A_ρ of A by

$$A(\rho) = \Big\{x : \quad \operatorname{dist}(x,y) \leq \rho \quad \text{for some} \quad y \in A\Big\}.$$

Using Minkowski addition, we can write

$$A(\rho) = A + B_\rho,$$

where

$$B_\rho = \Big\{x : \quad \|x\| \leq \rho\Big\}$$

is the ball of radius ρ centered at the origin.

In this section, μ denotes the standard Lebesgue measure in $\mathbb{R}^n$. The isoperimetric inequality for μ states that among all sets of a given measure, the Euclidean ball has the smallest measure of any ρ-neighborhood.

(21.1) Theorem. *For a compact set $A \subset \mathbb{R}^n$ let $B_r \subset \mathbb{R}^n$ be a ball such that*

$$\mu\left(B_r\right) = \mu(A).$$

Then

$$\mu\big(A(\rho)\big) \geq \mu\left(B_{r+\rho}\right) \quad \text{for any} \quad \rho \geq 0.$$

Proof. Applying the Brunn-Minkowski inequality (Corollary 2.3), we get

$$\mu^{1/n}\big(A(\rho)\big) = \mu^{1/n}\big(A + B_\rho\big) \geq \mu^{1/n}(A) + \mu^{1/n}\big(B_\rho\big) = \mu^{1/n}\big(B_r\big) + \mu^{1/n}\big(B_\rho\big).$$

However, the volume of an n-dimensional ball is proportional to the nth power of its radius. Therefore the volume$^{1/n}$ is a linear function of the radius. Hence

$$\mu^{1/n}\big(B_r\big) + \mu^{1/n}\big(B_\rho\big) = \mu^{1/n}\big(B_{\rho+r}\big),$$

and the proof follows. □

If A is "reasonable", so that the limit

$$\lim_{\rho \longrightarrow 0+} \frac{\mu\big(A(\rho)\big) - \mu(A)}{\rho}$$

exists and can be interpreted as the surface area $|\partial A|$ of A, Theorem 21.1 implies that among all sets of a given volume, the ball has the smallest surface area.

22. The concentration on the sphere and other strictly convex surfaces

We apply the Brunn-Minkowski inequality to re-establish the measure concentration on the unit sphere (see Section 14), though with weaker constants. On the other hand, we'll obtain concentration for more general surfaces than the sphere, namely *strictly convex* surfaces. Originally, the concentration on strictly convex surfaces was obtained by M. Gromov and V. Milman via a different approach. Here we present a simple proof due to J. Arias-de-Reyna, K. Ball, and R. Villa.

(22.1) Definition. Let $K \subset \mathbb{R}^n$ be a convex body (that is, a convex compact set with a non-empty interior). Suppose that K contains the origin in its interior and let $S = \partial K$ be the surface of K. We say that K is *strictly convex* if for any $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that whenever $x, y \in S$ and $\operatorname{dist}(x, y) \geq \epsilon$, we have $(x+y)/2 \in (1-\delta)K$. The function $\epsilon \longmapsto \delta(\epsilon)$ is called a *modulus of convexity* of K.

PROBLEM. Let $B = \Big\{x \in \mathbb{R}^n : \quad \|x\| \leq 1\Big\}$ be a unit ball. Prove that B is strictly convex and that one can choose

$$\delta(\epsilon) = 1 - \sqrt{1 - \frac{\epsilon^2}{4}} \geq \frac{\epsilon^2}{8} \quad \text{for} \quad 0 \leq \epsilon \leq 2.$$

Let $S = \partial K$ be the surface of K. We introduce a probability measure ν on S as follows: for a subset $A \subset S$, let

$$\overline{A} = \Big\{\alpha x : \quad x \in A \quad \text{and} \quad 0 \leq \alpha \leq 1\Big\}$$

be the "pyramid" over A centered at the origin. We let

$$\nu(A) = \frac{\mu(\overline{A})}{\mu(K)},$$

where μ is the Lebesgue measure. Check that if K is the unit ball and $S = \partial K$ is the unit sphere then ν is the rotation invariant probability measure on S. In general, however, ν is *not* the (normalized) surface area of S induced from $\mathbb{R}^n$. In a small neighborhood of a point $a \in S$, the measure ν is roughly proportional to the surface area times the distance from the tangent plane to the origin.

We claim that for strictly convex surfaces, the measure of an ϵ-neighborhood of a set of measure $1/2$ is almost 1. Note that we measure distances in the ambient space $\mathbb{R}^n$ and not intrinsically via a geodesic on the surface.

(22.2) Theorem. *Let K be a strictly convex body with a modulus of convexity $\delta = \delta(\epsilon)$. Let $S = \partial K$ be the surface of K and let $A \subset S$ be a set such that $\nu(A) \geq 1/2$. Then, for any $\epsilon > 0$ such that $\delta(\epsilon) \leq 1/2$, we have*

$$\nu\Big\{x \in S : \quad \operatorname{dist}(x, A) \geq \epsilon\Big\} \leq 2\big(1 - \delta(\epsilon)\big)^{2n} \leq 2e^{-2n\delta(\epsilon)}.$$

Proof. Let

$$B = \Big\{x \in S : \quad \operatorname{dist}(x, A) \geq \epsilon\Big\}.$$

Then for all pairs $x \in A$ and $y \in B$ we have $(x + y)/2 \in (1 - \delta)K$. Let $\overline{A}$ and $\overline{B}$ be the pyramids over A and B respectively, as defined above. We claim that for all $\overline{x} \in \overline{A}$ and $\overline{y} \in \overline{B}$ we have $(\overline{x} + \overline{y})/2 \in (1 - \delta)K$. Indeed, $\overline{x} = \alpha x$ and $\overline{y} = \beta y$ for some $x \in A$, $y \in B$ and $0 \leq \alpha, \beta \leq 1$. Without loss of generality, we can assume that $\alpha \geq \beta$ and $\alpha > 0$, so $\beta/\alpha = \gamma \leq 1$. Then

$$\begin{aligned}\frac{\overline{x} + \overline{y}}{2} =& \frac{\alpha x + \beta y}{2} = \alpha\left(\frac{x + \gamma y}{2}\right) = \alpha\left(\gamma\frac{x + y}{2} + (1 - \gamma)\frac{x}{2}\right) \\ =& \alpha\gamma\left(\frac{x + y}{2}\right) + \alpha(1 - \gamma)\frac{x}{2}.\end{aligned}$$

Therefore,

$$\frac{\overline{x} + \overline{y}}{2} \in \alpha\gamma(1 - \delta)K + \alpha(1 - \gamma)(1 - \delta)K = \alpha(1 - \delta)K \subset (1 - \delta)K,$$

cf. Problem 5 of Section 20.1.

Hence

$$\frac{1}{2}\overline{A} + \frac{1}{2}\overline{B} \subset (1 - \delta)K, \quad \text{and, therefore,} \quad \mu\left(\frac{1}{2}\overline{A} + \frac{1}{2}\overline{B}\right) \leq (1 - \delta)^n \mu(K).$$

Now we apply the multiplicative form of the Brunn-Minkowski inequality (Theorem 20.2) to claim that

$$\mu\left(\frac{1}{2}\overline{A}+\frac{1}{2}\overline{B}\right) \geq \mu^{1/2}(\overline{A})\mu^{1/2}(\overline{B}).$$

Hence

$$\nu(B)=\frac{\mu(\overline{B})}{\mu(K)} \leq (1-\delta)^{2n}\frac{\mu(K)}{\mu(A)} = \frac{(1-\delta)^{2n}}{\nu(A)} \leq 2(1-\delta)^{2n} \leq 2e^{-2n\delta(\epsilon)}.$$

□

In particular, if $K=B$ is the unit ball, we get the inequality

$$\nu\Big\{x \in S: \quad \operatorname{dist}(x,A) \geq \epsilon\Big\} \leq 2e^{-\epsilon^2 n/4},$$

which should be compared with Theorem 14.3. Note that in Section 14 we measure distances intrinsically, via a geodesic in $\mathbb{S}^{n-1}$. Here we measure distances extrinsically, via a chord in the space $\mathbb{R}^n$. We have

chord distance $\leq$ geodesic distance $\leq$ π chord distance, and

chord distance $\approx$ geodesic distance for close points

so the results are very much alike in spirit, although the constants we get here are worse than those in Section 14. However, the result applies to more general surfaces.

One can modify the strict convexity definition as follows: let us fix a norm $p(x)$ on $\mathbb{R}^n$, so $p(x) \geq 0$ for all $x \in \mathbb{R}^n$ and $p(x)=0$ implies $x=0$, $p(\alpha x)=|\alpha|p(x)$ for all $x \in \mathbb{R}^n$, and $p(x+y) \leq p(x)+p(y)$ for all $x,y \in \mathbb{R}^n$.

Let

$$K=\Big\{x \in \mathbb{R}^n: \quad p(x) \leq 1\Big\}$$

be the unit ball of this norm. We say that K (and p) are *strictly convex* if for any $\epsilon>0$ there is a $\delta=\delta(\epsilon)>0$ such that if $p(x)=p(y)=1$ and $p(x-y) \geq \epsilon$ then $p\big((x+y)/2\big) \leq (1-\delta)$.

Let $S=\partial K$ be the surface of K as before, and let ν be the measure on S defined as before. Let us measure the distance between $x,y \in K$ as $p(x-y)$. Then Theorem 22.2 holds in this modified situation.

Lecture 26. Monday, March 14

---

23. Concentration for the Gaussian measure and other strictly log-concave measures

In Section 22, we applied the Brunn-Minkowski inequality for the Lebesgue measure in $\mathbb{R}^n$ to recover concentration on the unit sphere and strictly convex surfaces. Now we apply the Prékopa-Leindler inequality of Section 19 to recover concentration for the Gaussian measure with some extensions. The inequalities we prove have non-optimal constants, but they are easy to deduce and they extend to a wide class of measures.

The proof is by the Laplace transform method, for which we need a (by now standard) estimate.

(23.1) Theorem. *Let μ be a probability measure in $\mathbb{R}^n$ with the density e^{-u}, where $u : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a function satisfying*

$$u(x) + u(y) - 2u\left(\frac{x+y}{2}\right) \geq c\|x-y\|^2$$

for some absolute constant $c > 0$. Let $A \subset \mathbb{R}^n$ be a closed set. Then

$$\int_{\mathbb{R}^n} \exp\left\{c \operatorname{dist}^2(x, A)\right\} \ d\mu \leq \frac{1}{\mu(A)}.$$

Proof. Let us define three functions $f, g, h : \mathbb{R}^n \longrightarrow \mathbb{R}$ as follows:

$$f(x) = \exp\left\{c \operatorname{dist}^2(x, A) - u(x)\right\}, \quad g(x) = [A]e^{-u(x)}, \quad \text{and} \quad h(x) = e^{-u(x)}.$$

We are going to apply the Prékopa-Leindler inequality (Theorem 19.1) with $\alpha = \beta = 1/2$. We have to check that

$$h\left(\frac{x+y}{2}\right) \geq f^{1/2}(x)g^{1/2}(y).$$

Clearly, it suffices to check the inequality in the case when $y \in A$ since otherwise $g(y) = 0$. If $y \in A$ then $\operatorname{dist}(x, A) \leq \|x-y\|$ for all $x \in \mathbb{R}^n$. Therefore, for $y \in A$,

$$\begin{aligned} f(x)g(y) =& \exp\left\{c \operatorname{dist}^2(x, A) - u(x) - u(x)\right\} \leq \exp\left\{c\|x-y\|^2 - u(x) - u(y)\right\} \\ \leq& \exp\left\{-2u\left(\frac{x+y}{2}\right)\right\} = h^2\left(\frac{x+y}{2}\right), \end{aligned}$$

which is what we need. Therefore, by Theorem 19.1,

$$\int_{\mathbb{R}^n} h(x) \ dx \geq \left(\int_{\mathbb{R}^n} f(x) \ dx\right)^{1/2} \left(\int_{\mathbb{R}^n} g(x) \ dx\right)^{1/2}.$$

However,

$$\begin{aligned} &\int_{\mathbb{R}^n} h(x) \ dx = \mu\left(\mathbb{R}^n\right) = 1, \quad \int_{\mathbb{R}^n} g(x) \ dx = \mu(A), \quad \text{and} \\ &\int_{\mathbb{R}^n} f(x) \ dx = \int_{\mathbb{R}^n} \exp\left\{c \operatorname{dist}^2(x, A)\right\} \ d\mu, \end{aligned}$$

and the proof follows. □

An immediate corollary is a concentration inequality.

(23.2) Corollary. *Let μ be a probability measure as in Theorem 23.1 and let $A \subset \mathbb{R}^n$ be a set such that $\mu(A) \geq 1/2$. Then, for any $t \geq 0$, we have*

$$\mu\Big\{x \in \mathbb{R}^n : \quad \operatorname{dist}(x, A) \geq t\Big\} \leq 2e^{-ct^2}.$$

Proof. Using the Laplace transform estimate (Section 2.1), we write

$$\begin{aligned}\mu\Big\{x \in \mathbb{R}^n : \quad c\operatorname{dist}^2(x, A) \geq ct^2\Big\} \leq & e^{-ct^2} \int_{\mathbb{R}^n} \exp\left\{c\operatorname{dist}^2(x, A)\right\} \, d\mu \\ \leq & 2e^{-ct^2}.\end{aligned}$$

□

The concentration for the Gaussian measure γ_n follows instantly. We choose

$$u(x) = \|x\|^2/2 + (n/2)\ln(2\pi).$$

In this case,

$$\begin{aligned}u(x) + u(y) - 2u\left(\frac{x+y}{2}\right) = & \frac{2\|x\|^2 + 2\|y\|^2 - \|x+y\|^2}{4} = \frac{\|x\|^2 + \|y\|^2 - 2\langle x, y\rangle}{4} \\ = & \frac{\|x-y\|^2}{4},\end{aligned}$$

so we can choose $c = 1/4$ in Theorem 23.1. This gives us the inequality

$$\gamma_n\Big\{x \in \mathbb{R}^n : \quad \operatorname{dist}(x, A) \geq t\Big\} \leq 2e^{-t^2/4} \quad \text{provided} \quad \gamma_n(A) \geq 1/2,$$

which is a bit weaker than the inequalities of Section 14, but of the same spirit.

24. Concentration for general log-concave measures

Let us consider a log-concave probability measure μ on $\mathbb{R}^n$, see Section 20. In fact, the only thing we need from μ is the Brunn-Minkowski inequality

$$\mu(\alpha A + \beta B) \geq \mu^{\alpha}(A)\mu^{\beta}(B) \quad \text{for} \quad \alpha, \beta \geq 0 \quad \text{such that} \quad \alpha + \beta = 1.$$

This is not quite the type of concentration we were dealing with so far. We take a sufficiently large convex body $A \subset \mathbb{R}^n$, where "sufficiently large" means that $\mu(A) > 1/2$, "inflate" it $A \longmapsto tA$ for some $t > 1$ and see how much is left, that is, what is $\mu\left(\mathbb{R}^n \setminus tA\right)$. We claim that what is left decreases exponentially with t.

(24.1) Theorem. *Let μ be a log-concave probability measure on $\mathbb{R}^n$ and let $A \subset \mathbb{R}^n$ be a convex symmetric body (that is, $-A = A$). Then, for all $t > 1$,*

$$\mu\left(\mathbb{R}^n \setminus tA\right) \leq \mu(A) \left(\frac{1-\mu(A)}{\mu(A)}\right)^{(t+1)/2}.$$

Proof. Let

$$B = \mathbb{R}^n \setminus (tA) \quad \text{and let} \quad \alpha = \frac{t-1}{t+1} \quad \text{and} \quad \beta = \frac{2}{t+1}.$$

Hence $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$. Applying the Brunn-Minkowski inequality, we get

$$\mu(\alpha A + \beta B) \geq \mu^{\alpha}(A)\mu^{\beta}(B).$$

On the other hand, we claim that

$$\alpha A + \beta B \subset \mathbb{R}^n \setminus A.$$

Indeed, if there is an $a \in A$, a $b \in B$, and a $c \in A$ such that $\alpha a + \beta b = c$, then

$$b = \beta^{-1}(c - \alpha a) = \frac{t+1}{2}c - \frac{t-1}{2}a = \frac{t+1}{2}c + \frac{t-1}{2}(-a) \in tA,$$

which is a contradiction. This gives us

$$1 - \mu(A) \geq \mu^{\alpha}(A)\mu^{\beta}(B),$$

that is,

$$\mu(B) \leq \left(\frac{1-\mu(A)}{\mu^{\alpha}(A)}\right)^{1/\beta} = \left(\frac{1-\mu(A)}{\mu^{(t-1)/(t+1)}(A)}\right)^{(t+1)/2} = \mu(A)\left(\frac{1-\mu(A)}{\mu(A)}\right)^{(t+1)/2}.$$

□

Suppose, for example, that $\mu(A) = 2/3$. Then Theorem 24.1 states that

$$\mu\left(\mathbb{R}^n \setminus tA\right) \leq \frac{2}{3}2^{-(t+1)/2},$$

so the measure of complement of tA indeed decreases exponentially. This is not quite the sort of decreasing we are used to, we'd prefer something of the type e^{-ct^2}. However, this is the best we can get. The estimate of Theorem 24.1 is "dimension-free". Let us choose $n = 1$ and let μ be the symmetric exponential measure with the density $0.5e^{-|x|}$. This is certainly a log-concave measure. Let us choose, $A = [-2, 2]$, say, so that $\mu(A) = 1 - e^{-2} \approx 0.865 > 2/3$. Then $tA = [-2t, 2t]$ and $\mu(A) = 1 - e^{-2t}$ and $\mu\left(\mathbb{R} \setminus A\right) = e^{-2t}$, so the measure outside of tA decreases exponentially with t, as promised. This is due to the fact that μ is not *strictly* log-concave. A typical application of Theorem 24.1 concerns norms.

(24.2) Corollary. *Let $p : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a norm and let μ be a log-concave measure on $\mathbb{R}^n$. Let us choose a number $r > 1/2$ and let ρ be a number such that*

$$\mu\Big\{x \in \mathbb{R}^n : \quad p(x) \leq \rho\Big\} = r.$$

Then, for all $t > 1$,

$$\mu\Big\{x \in \mathbb{R}^n : \quad p(x) > t\rho\Big\} \leq r\left(\frac{1-r}{r}\right)^{(t+1)/2}.$$

Proof. We apply Theorem 24.1 to

$$A = \Big\{x : \quad p(x) \leq \rho\Big\}.$$

□

Lecture 27. Wednesday, March 16

---

25. An application: reverse Hölder inequalities for norms

Let μ be a probability measure on $\mathbb{R}^n$ and let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a non-negative integrable function. Hölder inequalities state that

$$\left(\int_{\mathbb{R}^n} f^p \ d\mu\right)^{1/p} \leq \left(\int_{\mathbb{R}^n} f^q(x) \ d\mu\right)^{1/q} \quad \text{provided} \quad q \geq p > 0.$$

Furthermore, for $p = 0$, the above inequality reads

$$\exp\left\{\int_{\mathbb{R}^n} \ln f \ d\mu\right\} \leq \left(\int_{\mathbb{R}^n} f^q(x) \ d\mu\right)^{1/q} \quad \text{provided} \quad q > 0.$$

Indeed,

$$f^p = \exp\{p \ln f\} = 1 + p \ln f + O\left(p^2 \ln^2 f\right).$$

Assuming that $\ln^2 f$ is integrable, we get

$$\begin{aligned}\left(\int_{\mathbb{R}^n} f^p(x) \ d\mu\right)^{1/p} =& \left(1 + p\int_{\mathbb{R}^n} (\ln f) \ d\mu + O(p^2)\right)^{1/p} \\ =& \exp\left\{\frac{1}{p} \ln\left(1 + p\int_{\mathbb{R}^n} \ln f \ d\mu + O(p^2)\right)\right\} \\ =& \exp\left\{\int_{\mathbb{R}^n} \ln \ d\mu\right\}\Big(1 + O(p)\Big) \quad \text{for} \quad p \approx 0.\end{aligned}$$

Suppose now that μ is log-concave and that f is a norm. In this case, the inequalities can be reversed up to some constants.

(25.1) Theorem. *Let μ be a log-concave probability measure on $\mathbb{R}^n$ and let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a norm. Then*

$$\left(\int_{\mathbb{R}^n} f^p \ d\mu\right)^{1/p} \leq cp \int_{\mathbb{R}^n} f \ d\mu$$

for all $p \geq 1$ and some absolute constant $c > 0$.

Proof. Without loss of generality, we assume that

$$\int_{\mathbb{R}^n} f \ d\mu = 1.$$

Let us choose a $\rho > 2$. Since f is non-negative,

$$\mu\Big\{x \in \mathbb{R}^n : \quad f(x) \geq \rho\Big\} \leq \frac{1}{\rho}.$$

Then

$$\mu\Big\{x \in \mathbb{R}^n : \quad f(x) \leq \rho\Big\} \geq \frac{\rho - 1}{\rho}$$

and as in Corollary 24.2, we get

$$\mu\Big\{x \in \mathbb{R}^n : \quad f(x) \geq \rho t\Big\} \leq \frac{\rho-1}{\rho}\left(\frac{1}{\rho-1}\right)^{(t+1)/2} \leq \left(\frac{1}{\rho-1}\right)^{t/2}$$

for $t > 1$.

To make the computations simpler, let us choose $\rho = 1 + e < 4$. Thus

$$\mu\Big\{x \in \mathbb{R}^n : \quad f(x) \geq 4t\Big\} \leq e^{-t/2} \quad \text{for} \quad t > 1.$$

For $t \geq 0$, let

$$F(t) = \mu\Big\{x \in \mathbb{R}^n : \quad f(x) \leq t\Big\}$$

be the cumulative distribution function of f. In particular,

$$1 - F(t) \leq e^{-t/8} \quad \text{for} \quad t > 4.$$

Then

$$\begin{aligned}\int_{\mathbb{R}^n} f^p \ d\mu =& \int_0^{+\infty} t^p \ dF(t) = -\int_0^{+\infty} t^p \ d\big(1 - F(t)\big) \\ =& t^p\big(1 - F(t)\big)\Big|_{t=0}^{t=+\infty} + \int_0^{+\infty} pt^{p-1}\big(1 - F(t)\big) \ dt.\end{aligned}$$

Since $1 - F(t)$ is exponentially decreasing as t grows, the first term is 0. Therefore,

$$\int_{\mathbb{R}^n} f^p \ d\mu = \int_0^{+\infty} pt^{p-1}\big(1 - F(t)\big) \ dt \leq \int_0^4 pt^{p-1} \ dt + \int_4^{+\infty} pt^{p-1}e^{-t/8} \ dt$$
$$\leq 4^p + p\int_0^{+\infty} t^{p-1}e^{-t/8} \ dt = 4^p + p8^p\Gamma(p).$$

Since $\Gamma(p) \leq p^p$, the proof follows. □

The proof of Theorem 25.1 is due to C. Borell.

PROBLEM. Deduce from Theorem 25.1 that for any $p > q > 0$ and some constant $c(p,q) > 0$, we have

$$\left(\int_{\mathbb{R}^n} f^p \ d\mu\right)^{1/p} \leq c(p,q)\left(\int_{\mathbb{R}^n} f^q \ d\mu\right)^{1/q}$$

for any norm f.

The following extension of Theorem 25.1 was obtained by R. Latała.

(25.2) Theorem. *There exists an absolute constant $c > 0$ such that*

$$\ln\left(\int_{\mathbb{R}^n} f \ d\mu\right) \leq c + \int_{\mathbb{R}^n} \big(\ln f\big) \ d\mu$$

for any log-concave probability measure μ on $\mathbb{R}^n$ and any norm $f : \mathbb{R}^n \longrightarrow \mathbb{R}$.

Note that Jensen's inequality implies that

$$\int_{\mathbb{R}^n} \big(\ln f\big) \ d\mu \leq \ln\left(\int_{\mathbb{R}^n} f \ d\mu\right).$$

Theorem 25.2 can be recast as

$$\left(\int_{\mathbb{R}^n} f^p \ d\mu\right)^{1/p} \geq c\int_{\mathbb{R}^n} f \ d\mu$$

for some other absolute constant c and all $p > 0$. As we remarked before, as $p \longrightarrow 0+$, the left hand side approaches

$$\exp\left\{\int_{\mathbb{R}^n} \big(\ln f\big) \ d\mu\right).$$

If the main idea of the proof of Theorem 25.1 is to bound the measure outside of a large ball, the main idea of the proof of Theorem 25.2 is to bound the measure inside a small ball in the norm f. More precisely, we want to prove that

$$\mu\Big\{x \in \mathbb{R}^n : \quad f(x) \leq t\Big\} \leq ct$$

for some absolute constant c and all sufficiently small t.

The core of the argument is the following lemma.

(25.3) Lemma. *For a norm $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, let*

$$B = \Big\{ x \in \mathbb{R}^n : \quad f(x) \leq 1 \Big\}$$

be the unit ball. Let us choose a $\delta > 0$ and a $\rho > 1$ such that

$$\mu(\rho B) \geq (1+\delta)\mu(B).$$

Then for some $c = c(\rho/\delta) > 0$, we have

$$\mu(tB) \leq ct\mu(B) \quad \text{for all} \quad 0 < t < 1.$$

Proof. It suffices to prove Lemma 25.3 for any t of the form $t = (2m)^{-1}$, where m is a positive integer.

Let us pick a particular m and let

$$\mu\left(\frac{1}{2m}B\right) = \kappa(m)\frac{\mu(B)}{m}.$$

We must prove that $\kappa = \kappa(m) \geq 0$ is bounded from above by a universal constant depending only on the ratio ρ/δ.

The idea is to slice the ball B onto m concentric "rings" with the "core" being the ball of radius $1/2m$. Using the conditions of the Lemma, we prove that the measure of the outside ring is large enough. If the measure of the core is large, then by the Brunn-Minkowski inequality the measure of each ring should be large enough. This would contradict to the fact that all those measures sum up to the measure of the ball.

Without loss of generality, we can assume that

$$\kappa(m) \geq \frac{2\delta}{\rho} \quad \text{and hence} \quad \mu\left(\frac{1}{2m}B\right) \geq \frac{2\delta\mu(B)}{\rho m} > \frac{\delta\mu(B)}{\rho m}. \tag{25.3.1}$$

We have

$$\mu(\rho B \setminus B) \geq \delta\mu(B).$$

For $\tau \geq 0$ let

$$A_\tau = \Big\{ x \in \mathbb{R}^n : \quad \tau - \frac{1}{2m} < f(x) < \tau + \frac{1}{2m} \Big\}.$$

The interval $[1, \rho]$ can be covered by a disjoint union of at most ρm non-overlapping intervals of length $1/m$ centered at points $\tau \in [1, \rho]$. Therefore, there there is a $\tau' \geq 1$ such that

$$\mu\left(A_{\tau'}\right) \geq \frac{\delta\mu(B)}{\rho m}.$$

Now,

$$\lambda A_\tau + \frac{(1-\lambda)}{2m} B \subset A_{\lambda\tau} \quad \text{for} \quad 0 < \lambda < 1,$$

from which by the Brunn-Minkowski inequality

$$\mu(A_{\lambda\tau}) \geq \mu^{\lambda}(A_\tau)\mu^{1-\lambda}\left(\frac{1}{2m}B\right). \tag{25.3.2}$$

Choosing $\tau = \tau'$ and $\lambda = 1/\tau'$ in (25.3.2) and using (25.3.1), we get

$$\mu\left(A_1\right) \geq \frac{\delta\mu(B)}{\rho m}. \tag{25.3.3}$$

Let us look at the sets $A_{(m-1)/m}, A_{(m-2)/m}, \dots, A_{1/m}$ (these are our "rings") and $A_0 = (2m)^{-1}B$ (the core). These sets are disjoint and lie in B. Combining (25.3.1)–(25.3.3), we get

$$\mu\left(A_{i/m}\right) \geq \frac{\mu(B)}{m}\left(\frac{\delta}{\rho}\right)^{(m-i)/m}\kappa^{i/m}.$$

Summing up over $i = 1, \dots, m$, we get

$$\mu(B) \geq \frac{\kappa\mu(B)}{m}\frac{1-\delta/\rho\kappa}{1-(\delta/\rho\kappa)^{1/m}}.$$

Now, $\kappa = \kappa(m) \geq 2\delta/\rho$, from which

$$\mu(B) \geq \frac{\kappa}{2m}\mu(B)\frac{1}{1-(\delta/\rho\kappa)^{1/m}}.$$

This gives us the inequality

$$\kappa \leq 2m\left(1-\left(\frac{\delta}{\rho\kappa}\right)^{1/m}\right).$$

It remains to notice that for $a < 1$ and $x > 0$ we have

$$x\left(1-a^{1/x}\right) = x\left(1-\exp\left\{x^{-1}\ln a\right\}\right) \leq x\left(1-1-x^{-1}\ln a\right) = \ln a^{-1},$$

from which we get

$$\kappa(m) \leq 2\ln\left(\frac{\rho\kappa(m)}{\delta}\right).$$

It follows now that $\kappa(m)$ must be bounded by some constant depending on ρ/δ only. □

Now we can prove Theorem 25.2.

Proof of Theorem 25.2. Without loss of generality, we assume that

$$\int_{\mathbb{R}^n} f \ d\mu = 1,$$

so our goal is to prove that

$$\int_{\mathbb{R}^n} (\ln f) \ d\mu \geq c$$

for some absolute constant c.

For $t > 0$, let

$$B_t = \Big\{ x \in \mathbb{R}^n : \quad f(x) \leq t \Big\}$$

denote the ball of radius t in the norm f.

Let us choose a ρ such that

$$\mu\left(B_\rho\right) = 2/3.$$

Recall that by our Definition 20.1, μ has a density, so such a ρ exists. This density assumption is not crucial though.

We observe that by Markov inequality

$$\mu \Big\{ x \in \mathbb{R}^n : \quad f(x) \geq 4 \Big\} \leq \frac{1}{4},$$

so $\rho \leq 4$.

Next, by Corollary 24.2 with $r = 2/3$ and $t = 3$,

$$\mu\left(B_{3\rho}\right) \geq 1 - \frac{2}{3} \cdot \frac{1}{4} = \frac{5}{6}.$$

Rescaling the norm $f \longmapsto \rho f$ and applying Lemma 25.3, we conclude that for some absolute constant $\omega > 0$ and some absolute constant $c > 0$, we have

$$\mu\left(B_t\right) \leq ct \quad \text{for} \quad t \leq \omega.$$

Without loss of generality, we assume that $\omega \leq 1$. This is enough to complete the proof. Denoting $F(t) = \mu\left(B_t\right)$, we have

$$\int_{\mathbb{R}^n} (\ln f) \ d\mu = \int_0^{+\infty} (\ln t) \ dF(t) \geq \int_0^1 (\ln t) \ dF(t) = (\ln t) F(t) \Big|_{t=0}^{1} - \int_0^1 t^{-1} F(t) \ dt.$$

Now, the first term is 0 since $F(t) \leq ct$ in a neighborhood of $t = 0$. Hence we have to estimate the second term.

$$\begin{aligned} \int_0^1 t^{-1} F(t) \ dt &= \int_0^\omega t^{-1} F(t) \ dt + \int_\omega^1 t^{-1} F(t) \ dt \\ &\leq \int_0^\omega t^{-1} (ct) \ dt + \int_0^1 \omega^{-1} \ dt = c\omega + \omega^{-1}, \end{aligned}$$

and the proof follows. □

PROBLEM. Let γ_n be the standard Gaussian measure in $\mathbb{R}^n$ and let $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ be a positive definite quadratic form such that

$$\int_{\mathbb{R}^n} f \ d\gamma_n = 1.$$

Prove that

$$\int_{\mathbb{R}^n} (\ln f) \ d\gamma_n \geq c,$$

where $c = -\lambda - \ln 2 \approx -1.27$, where $\lambda \approx 0.577$ is the Euler constant, that is,

$$\lambda = \lim_{n \longrightarrow +\infty} \ln n - \sum_{k=1}^{n} \frac{1}{k}.$$

The bound is the best possible.

Lecture 28. Friday, March 18

---

26. Log-concave measures as projections of the Lebesgue measure

Generally, we call a Borel probability measure μ on $\mathbb{R}^n$ (with density or not) *log-concave* if it satisfies the Brunn-Minkowski inequality

$$\mu(\alpha A + \beta B) \geq \mu^{\alpha}(A) \mu^{\beta}(B) \quad \text{where} \quad \alpha, \beta \geq 0 \quad \text{and} \quad \alpha + \beta = 1$$

and $A, B \subset \mathbb{R}^n$ are reasonable (say, closed) subsets, so that A, B, and $A + B$ are measurable.

As we established in Section 20, measure with densities $e^{-u(x)}$, where $u: \mathbb{R}^n \longrightarrow \mathbb{R}$ are convex functions, are log-concave. This can be rephrased as "locally log-concave measures are globally log-concave", since the condition of having a log-concave density is more or less equivalent to the Brunn-Minkowski inequality for small neighborhoods A and B of points.

The goal of this section is to present a number of simple geometric constructions showing that the class of all log-concave measures is a very natural object.

We start with a simple observation: the projection of a log-concave measure is log-concave.

(26.1) Theorem. *Let μ be a log-concave measure on $\mathbb{R}^n$ and let $T : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a linear transformation. Then the push-forward $\nu = T(\mu)$ defined by*

$$\nu(A) = \mu\left(T^{-1}(A)\right)$$

is a log-concave measure on $\mathbb{R}^m$.

Proof. Let us choose $A, B \subset \mathbb{R}^m$ and let $A_1 = T^{-1}(A)$ and $B = T^{-1}(B)$, so $A_1, B_1 \subset \mathbb{R}^n$. Let us choose $\alpha, \beta \geq 0$ with $\alpha + \beta = 1$. We claim that

$$\alpha A_1 + \beta B_1 \subset T^{-1}(\alpha A + \beta B).$$

Indeed, if $a \in A_1$ and $b \in B_1$ then $Ta \in A$ and $Tb \in B$ so $T(\alpha a + \beta b) = \alpha Ta + \beta Tb \in \alpha A + \beta B$. This implies that $\alpha a + \beta b \in T^{-1}(\alpha A + \beta B)$.

Therefore,

$$\begin{aligned}\nu(\alpha A + \beta B) =&\mu\left(T^{-1}(\alpha A + \beta B)\right) \geq \mu(\alpha A_1 + \beta B_1) \geq \mu^{\alpha}(A_1)\mu^{\beta}(B_1)\\ =&\nu^{\alpha}(A)\nu^{\beta}(B),\end{aligned}$$

as claimed. □

An important corollary of Theorem 26.1 is that the convolution of log-concave densities is a log-concave density.

PROBLEMS.

1. Let $f, g : \mathbb{R}^n \longrightarrow \mathbb{R}$ be log-concave probability densities. Let $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ be their *convolution*

$$h(x) = \int_{\mathbb{R}^n} f(y)g(x-y) \ dy.$$

Prove that h is a log-concave density.

2. Check that $T^{-1}\left(\alpha A + \beta B\right) = \alpha T^{-1}(A) + \beta T^{-1} B$.

One can think of h as of the density of the push-forward of the probability measure on $\mathbb{R}^{2n}$ with the density $f(x)g(y)$ under a linear transformation. Another corollary of Theorem 26.1 is that the sum of independent random variables with log-concave distributions has a log-concave distribution.

The standard example of a log-concave measure is the Lebesgue measure restricted to a convex set. Namely, we fix a convex body $K \subset \mathbb{R}^n$ and let

$$\mu(A) = \operatorname{vol}(A \cap K) \quad \text{for} \quad A \subset \mathbb{R}^n.$$

Theorem 26.1 tells us that by projecting such measures we get log-concave measures. A natural question is: what kind of a measure can we get by projecting the Lebesgue measure restricted to a convex set. The answer is: pretty much any log-concave measure, if we allow taking limits.

(26.2) Theorem. *Let $u : \mathbb{R}^m \longrightarrow \mathbb{R}$ be a convex function, let $K \subset \mathbb{R}^m$ be a convex body and let us consider the measure ν with the density $e^{-u}[K]$ (that is, with the density e^{-u} restricted to K). Then there exists a sequence of convex bodies $K_n \subset \mathbb{R}^{d_n}$ and projections $T_n : \mathbb{R}^{d_n} \longrightarrow \mathbb{R}^m$ such that*

$$T(\mu_n) \longrightarrow \nu \quad as \quad n \longrightarrow +\infty,$$

where μ_n is the Lebesgue measure restricted to K_n and the density of $T(\mu_n)$ uniformly converges to the density of ν.

Proof. Let $d_n = m + n$. We define the projection

$$T_n : \mathbb{R}^{m+n} \longrightarrow \mathbb{R}^m, \quad (\xi_1, \dots, \xi_{m+n}) \longmapsto (\xi_1, \dots, \xi_m).$$

We represent vectors $x \in \mathbb{R}^{m+n}$ in the form $x = (y, \xi_{m+1}, \dots, \xi_{m+n})$, where $y \in \mathbb{R}^m$ and define $K_n \subset \mathbb{R}^{m+n}$ by

$$K_n = \Big\{ (y, \xi_{m+1}, \dots, \xi_{m+n}) : \quad y \in K \quad \text{and} \\ 0 \leq \xi_i \leq 1 - \frac{u(y)}{n} \quad \text{for} \quad i = m+1, \dots, m+n \Big\}.$$

Since $u(y)$ is bounded on K, for all sufficiently large n we have $u(y) < n$ for all $y \in K$ and K_n are not empty for all sufficiently large n.

Next, we claim that K_n are convex, since u is convex. Finally, let us consider the push-forward $T_n(\mu_n)$. Clearly, $T(K_n) = K$ and the preimage of every point $y \in K$ is an n-dimensional cube with the side $1 - u(y)/n$ and the n-dimensional volume $(1 - u(y)/n)^n$. Since

$$\left(1 - \frac{u(y)}{n}\right)^n \longrightarrow e^{-u(y)}$$

uniformly on K, the proof follows. □

In view of Theorems 26.1 and 26.2, the Brunn-Minkowski inequality for any log-concave density follows from the Brunn-Minkowski inequality for the Lebesgue measure (though in a higher dimension). Wouldn't it be nice to complete the circle by providing a fairly elementary proof of the Brunn-Minkowski inequality for the Lebesgue measure?

(26.3) An elementary proof of the Brunn-Minkowski inequality for the Lebesgue measure. As before, our main tool is the concavity of $\ln x$:

$$\ln(\alpha x + \beta y) \geq \alpha \ln x + \beta \ln y \quad \text{for all} \quad \alpha, \beta \geq 0 \quad \text{such that} \quad \alpha + \beta = 1$$

and all $x, y > 0$. In the multiplicative form:

$$\alpha x + \beta y \geq x^\alpha y^\beta.$$

Our first observation is that both sides of the inequality

$$\operatorname{vol}(\alpha A + \beta B) \geq \operatorname{vol}^{\alpha}(A) \operatorname{vol}^{\beta}(B)$$

do not change if we apply translations $A \longmapsto A + a$, $B \longmapsto B + b$. Indeed, $\alpha A + \beta B$ gets translated by $\alpha a + \beta b$ and volumes do not change under translations.

Next, we establish the inequality when both A and B are axis-parallel parallelepipeds, called "bricks":

$$\begin{aligned} A &= \Big\{(\xi_1, \dots, \xi_n) : \quad s_i' \leq \xi_i \leq s_i \quad \text{for} \quad i = 1, \dots, n\Big\} \quad \text{and} \\ B &= \Big\{(\xi_1, \dots, \xi_n) : \quad t_i' \leq \xi_i \leq t_i \quad \text{for} \quad i = 1, \dots, n\Big\}. \end{aligned}$$

Translating, if necessary, we assume that $s_i' = t_i' = 0$. Then $A + B$ is also a brick

$$\alpha A + \beta B = \Big\{(\xi_1, \dots, \xi_n) : \quad 0 \leq \xi_i \leq \alpha s_i + \beta t_i \quad \text{for} \quad i = 1, \dots, n\Big\}.$$

In this case, the Brunn-Minkowski inequality is equivalent to the concavity of the logarithm:

$$\operatorname{vol}(\alpha A + \beta B) = \prod_{i=1}^{n} (\alpha s_i + \beta t_i) \geq \prod_{i=1}^{n} s_i^{\alpha} t_i^{\beta} = \operatorname{vol}^{\alpha}(A) \operatorname{vol}^{\beta}(B).$$

Next, we prove the inequality for sets that are finite unions of non-overlapping bricks. We use induction on the *total* number of bricks in A and B.

Let k be the total number of bricks, so $k \geq 2$. If $k = 2$ then each set consists of a single brick, the case we have already handled. Suppose that $k > 2$. Then one of the sets, say A, contains at least two bricks. These two bricks differ in at least one coordinate, say ξ_1. The hyperplane ("wall") $\xi_1 = 0$ cuts $\mathbb{R}^n$ into two closed halfspaces and cuts each brick which it cuts into two bricks. Translating A, if necessary, we can assure that at least one brick lies exactly inside each of the halfspaces. Let A_1 and A_2 be the intersections of A with the halfspaces. Then both A_1 and A_2 contain *fewer* bricks than A. Now, we start translating B. We translate B in such a way that the corresponding sets B_1 and B_2 (the intersections of B with the halfspaces) satisfy

$$\frac{\operatorname{vol} B_1}{\operatorname{vol} B} = \frac{\operatorname{vol} A_1}{\operatorname{vol} A} = \gamma,$$

say. We observe that both B_1 and B_2 are made of at most as many bricks as B. Hence the total number of bricks in the pair (A_1, B_1) is at most $k - 1$ and the total number of bricks in the pair (A_2, B_2) is at most $k - 1$. We apply the induction hypothesis to each pair and get

$$\begin{aligned} &\operatorname{vol}(\alpha A_1 + \beta B_1) \geq \operatorname{vol}^{\alpha}(A_1) \ln \operatorname{vol}^{\beta}(B_1) \quad \text{and} \\ &\operatorname{vol}(\alpha A_2 + \beta B_2) \geq \operatorname{vol}^{\alpha}(A_2) \operatorname{vol}^{\beta}(B_2). \end{aligned}$$

However, $\alpha A_1 + \beta B_1 \subset \alpha A + \beta B$ and $\alpha A_2 + \beta B_2 \subset \alpha A + \beta B$ lie on the opposite sides of the wall $\xi_1 = 0$ and so do not overlap.

Therefore,

$$\begin{aligned}\operatorname{vol}(\alpha A + \beta B) \geq &\operatorname{vol}(\alpha A_1 + \beta B_1) + \operatorname{vol}(\alpha A_2 + \beta B_2)\\ \geq &\operatorname{vol}^{\alpha}(A_1)\operatorname{vol}^{\beta}(B_1) + \operatorname{vol}^{\alpha}(A_2)\operatorname{vol}^{\beta}(B_2)\\ =&\gamma^{\alpha}\operatorname{vol}^{\alpha}(A)\gamma^{\beta}\operatorname{vol}^{\beta}(B) + (1-\gamma)^{\alpha}\operatorname{vol}^{\alpha}(A)(1-\gamma)^{\beta}\operatorname{vol}^{\beta}(B)\\ =&\gamma\operatorname{vol}^{\alpha}(A)\operatorname{vol}^{\beta}(B) + (1-\gamma)\operatorname{vol}^{\alpha}\operatorname{vol}^{\beta}(B) = \operatorname{vol}^{\alpha}(A)\operatorname{vol}^{\beta}(B).\end{aligned}$$

This completes the proof for sets that are finite unions of bricks.

The final step consists of approximating reasonable sets $A, B \subset \mathbb{R}^n$ (say, Jordan measurable) by finite unions of bricks.

Lecture 29. Monday, March 21

27. The needle decomposition technique for proving dimension-free inequalities

In this section we discuss a remarkable technique, called "the needle decomposition" technique or the "localization technique" which allows one to prove inequalities in $\mathbb{R}^n$ by reducing them to an inequality in $\mathbb{R}^1$. This technique was developed and used for a variety of purposes by S. Bobkov, M. Gromov, V. Milman, L. Lovász, M. Simonovits, R. Kannan, F. Nazarov, M. Sodin, A. Volberg, M. Fradelizi, and O. Guédon among others. While it is not easy to describe the method in all generality, we state one particular application as a theorem below in the hope that it captures the spirit of the method.

(27.1) Theorem. *Let $K \subset \mathbb{R}^n$ be a convex body, let $f : K \longrightarrow \mathbb{R}$ be a continuous function, let μ be a log-concave measure on K, and let $\Phi, \Psi : \mathbb{R} \longrightarrow \mathbb{R}$ be continuous functions.*

Suppose that for every interval $I \subset K$ and for every log-concave measure ν on I the inequality

$$\Phi\left(\frac{1}{\nu(I)}\int_I f\ d\nu\right) \leq \frac{1}{\nu(I)}\int_I \Psi(f)\ d\nu \tag{27.1.1}$$

holds.

Then the inequality

$$\Phi\left(\frac{1}{\mu(K)}\int_K f\ d\mu\right) \leq \frac{1}{\mu(K)}\int_K \Psi(f)\ d\mu \tag{27.1.2}$$

holds.

For example, suppose that $\Phi(x) = \ln x$ and that $\Psi(x) = c + \ln x$, where c is a constant. Hence to establish the inequality

$$\ln\left(\frac{1}{\mu(K)}\int_K f\ d\mu\right) \leq c + \frac{1}{\mu(K)}\int_K (\ln f)\ d\mu,$$

for every log-concave measure μ on K, it suffices to establish the one-dimensional version of the inequality for every interval $I \subset K$.

Similarly, suppose that $\Phi(x) = x^p$ for $x > 0$ and $\Psi(x) = c^p x^p$ for $x > 0$, where c is a constant. Hence to establish the inequality

$$\frac{1}{\mu(K)}\int_K f\ d\mu \leq c\left(\frac{1}{\mu(K)}\int_K f^p d\mu\right)^{1/p},$$

for every log-concave measure μ on K, it suffices to establish the one-dimensional version of the inequality for every interval $I \subset K$ and any log-concave measure ν on I.

One more example: let us choose $\Phi(x) = -ce^x$, where $c > 0$ is a constant and $\Psi(x) = -e^x$. Thus to prove the inequality

$$\frac{1}{\mu(K)}\int_K e^f\ d\mu \leq c\exp\left\{\frac{1}{\mu(K)}\int_K f\ d\mu\right\},$$

it suffices to prove the one-dimensional version of the inequality.

The idea of the proof is to iterate a certain construction, which we call "halving", which reduces the inequality over K to the inequalities over smaller convex bodies. These bodies become thinner and thinner, needle-like, and, in the limit look like one-dimensional intervals.

(27.2) Halving a convex body.

Let $K \subset \mathbb{R}^n$ be a convex body, let μ be any (not necessarily log-concave) measure on K with a positive continuous density function and let $f : K \longrightarrow \mathbb{R}$ be a continuous function on K.

Our goal is to cut the convex body K by an affine hyperplane H into two non-overlapping convex bodies K_1 and K_2 in such a way that

$$\frac{1}{\mu(K_1)}\int_{K_1} f\ d\mu = \frac{1}{\mu(K_2)}\int_{K_2} f\ d\mu. \tag{27.2.1}$$

Note that then we necessarily have

$$\frac{1}{\mu(K)}\int_K f\ d\mu = \frac{1}{\mu(K_i)}\int_{K_i} f\ d\mu \quad \text{for} \quad i = 1, 2. \tag{27.2.2}$$

Indeed,

$$\begin{aligned}\frac{1}{\mu(K)}\int_K f\ d\mu =&\frac{1}{\mu(K)}\left(\int_{K_1} f\ d\mu + \int_{K_2} f\ d\mu\right)\\ =&\frac{\mu(K_1)}{\mu(K)}\left(\frac{1}{\mu(K_1)}\int_{K_1} f\ d\mu\right) + \frac{\mu(K_2)}{\mu(K)}\left(\frac{1}{\mu(K_2)}\int_{K_2} f d\mu\right)\\ =&\frac{\mu(K_1)+\mu(K_2)}{\mu(K)}\left(\frac{1}{\mu(K_i)}\int_{K_1} f\ d\mu\right) = \frac{1}{\mu(K_i)}\int_{K_i} f\ d\mu.\end{aligned}$$

Suppose for a moment that we managed to do that. Suppose further that the inequalities

$$\Phi\left(\frac{1}{\mu(K_i)}\int_{K_i} f\ d\mu_i\right) \leq \frac{1}{\mu(K_i)}\int_{K_i} \Psi(f)\ d\mu_i \quad \text{for} \quad i=1,2$$

hold. Then, in view of (27.2.2),

$$\begin{aligned}\Phi\left(\frac{1}{\mu(K)}\int_K f\ d\mu\right) =&\Phi\left(\frac{1}{\mu(K_i)}\int_{K_i} f\ d\mu_i\right)\\ \leq&\frac{1}{\mu(K_i)}\int_{K_i} \Psi(f)\ d\mu_i \quad \text{for} \quad i=1,2.\end{aligned}$$

Taking the convex combination of the above inequalities for $i = 1, 2$ with the coefficients $\mu(K_i)/\mu(K)$, we get the inequality

$$\Phi\left(\frac{1}{\mu(K)}\int_K f\ d\mu\right) \leq \frac{1}{\mu(K)}\int_K \Psi(f)\ d\mu.$$

Thus we reduced the inequality for a larger body K to the inequalities for smaller bodies K_1 and K_2.

There are plenty of ways to cut K into K_1 and K_2 so that (27.2.1) is satisfied. We impose one extra condition, after which there is essentially one way to do it.

Let us fix an $(n-2)$-dimensional affine subspace $L \subset \mathbb{R}^n$ passing through an interior point of K. Let $H \subset \mathbb{R}^n$ be a hyperplane passing through L. We have one degree of freedom in choosing H. Indeed, let us choose a 2-dimensional plane A orthogonal to L such that $L \cap A = a$, say. Let us consider the orthogonal projection $pr : \mathbb{R}^n \longrightarrow A$ along L. Then every hyperplane $H \supset L$ is projected onto a line passing through a, and to choose a line through a in A is to choose a hyperplane H containing L. Hence we got a one-parametric continuous family of hyperplanes $H(t)$, where $0 \leq t \leq \pi$ and $H(0) = H(\pi)$. In other words, $H(t)$ is obtained from $H(0)$ by rotating through an angle of t about L.

Each hyperplane $H(t)$ defines two (closed) halfspaces $H_+(t)$ and $H_-(t)$ and, consequently, two convex bodies

$$K_+(t) = K \cap H_+(t) \quad \text{and} \quad K_-(t) = K \cap H_-(t).$$

It really doesn't matter which one is which, but if we make L oriented, we can choose $K_+(t)$ and $K_-(t)$ consistently, that is, continuously in t and in such a way that

$$K_+(0) = K_-(\pi) \quad \text{and} \quad K_-(0) = K_+(\pi). \tag{27.2.3}$$

Here is the moment of truth: as t changes from 0 to π, the two integrals

$$\frac{1}{\mu\left(K_+(t)\right)} \int_{K_+(t)} f \ d\mu \quad \text{and} \quad \frac{1}{\mu\left(K_-(t)\right)} \int_{K_-(t)} f \ d\mu$$

interchange because of (27.2.3). Therefore, by continuity, there is $t' \in [0, \pi]$ such that

$$\frac{1}{\mu\left(K_+(t')\right)} \int_{K_+(t')} f \ d\mu = \frac{1}{\mu\left(K_-(t')\right)} \int_{K_-(t')} f \ d\mu.$$

Hence we let

$$K_1 = K_+(t') \quad \text{and} \quad K_2 = K_-(t')$$

and (27.2.1) is satisfied.

This construction allows us to reduce proving the inequality of Theorem 27.1 for a convex body K to proving the inequality for smaller convex bodies. We are going to iterate the construction making bodies smaller and smaller. In the limit, we would like to make them 1-*dimensional*.

(27.3) Definition. A convex body $K \subset \mathbb{R}^n$ is called an ϵ-*needle* if there exists a line $l \subset \mathbb{R}^n$ such that all points of K are within distance ϵ from l.

(27.4) Constructing a needle decomposition. Let us choose an $\epsilon > 0$. Let us choose many, but finitely many, affine $(n-2)$-dimensional subspaces $L_1, \ldots, L_N$ piercing K in various directions. What we want is the following: for every affine plane A, the set of intersections $A \cap L_1, \ldots, A \cap L_N$ is an $(\epsilon/4)$-net for the intersection $A \cap K$. More precisely, we want to choose $L_1, \ldots, L_N$ so dense that for every affine two-dimensional plane A the intersection $A \cap K$ contains no disc of radius $\epsilon/4$ without a point of the form $A \cap L_i$ inside. A standard compactness argument implies the existence of such $L_1, \ldots, L_N$. For example, choose a two-dimensional plane A such that $A \cap K \neq \emptyset$. Choose a sufficiently dense net in $K \cap A$ and construct affine subspaces L_i orthogonal to A and passing through the points of the net. The set $L_i \cap \tilde{A}$ will provide a sufficiently dense net for $\tilde{A} \cap K$ if a two-dimensional plane $\tilde{A}$ is sufficiently close to A. Since the set of two-dimensional affine subspaces A intersecting K is compact, by repeating this construction with finitely many subspaces A, we construct $L_1, \ldots, L_N$.

Let us take the subspaces $L_1, \ldots, L_N$ one after another and apply the halving procedure. More precisely, we apply the halving procedure with L_1, thus obtaining K_1 and K_2. Taking the subspace L_i, we find all previously constructed convex bodies K_j, check for which of them L_i passes through an interior point and halve those using the procedure of (27.2). In the end, we represent K as a finite union of non-overlapping convex bodies K_i, $i \in I$, such that

$$\frac{1}{\mu(K_i)} \int_{K_i} f \; d\mu = \frac{1}{\mu(K)} \int_K f \; d\mu \quad \text{for all} \quad i \in I.$$

We claim that every K_i is an ϵ-needle.

Indeed, let us pick a particular $C = K_i$. One thing is certain: none of the subspaces $L_1, \dots, L_N$ passes through an interior point of C. Because if it did, we would have halved C along that subspace some time ago. Let $a, b \in C$ be two points with the maximum value of $\operatorname{dist}(a, b)$ for $a, b \in C$. Let us draw a line through l through a and b. Let $c \in C$ be any point and consider the triangle (abc). The interior of (abc) does not intersect any of the L_i (otherwise, we would have halved C). Therefore, the triangle (abc) does not contain a disc of radius greater than $\epsilon/4$. Since a and b are the maximum distance apart, the angles at a an b are acute and a picture shows that the distance from c to l cannot be larger than ϵ.

Proof of Theorem 27.1. Since f, Φ, and Ψ are uniformly continuous on K, there exist functions ω_f, ω_Φ, and ω_Ψ such that

$$\begin{aligned} \operatorname{dist}(x, y) \leq \epsilon &\Longrightarrow |f(x) - f(y)| \leq \omega_f(\epsilon), \\ |x - y| \leq \epsilon &\Longrightarrow |\Phi(x) - \Phi(y)| \leq \omega_\Phi(\epsilon), \quad \text{and} \\ |x - y| \leq \epsilon &\Longrightarrow |\Psi(x) - \Psi(y)| \leq \omega_\Psi(\epsilon) \end{aligned}$$

for all $\epsilon > 0$ and

$$\omega_f(\epsilon), \omega_\Phi(\epsilon), \omega_\Psi(\epsilon) \longrightarrow 0 \quad \text{as} \quad \epsilon \longrightarrow 0.$$

Let us choose an $\epsilon > 0$ and let us construct a decomposition $K = \bigcup_{i \in I} K_i$, where K_i are non-overlapping ϵ-needles and

$$\frac{1}{\mu(K_i)} \int_{K_i} f \; d\mu = \frac{1}{\mu(K)} \int_K f \; d\mu \quad \text{for} \quad i \in I.$$

For each needle K_i let us pick an interval $I = I_i \subset K_i$ with the endpoints maximum distance apart in K_i. Let $pr : K \longrightarrow I$ be the orthogonal projection onto I, cf. Section 27.4. Let $\nu = \nu_i$ be the push-forward measure on I. Then, by Theorem 26.1, ν is log-concave.

Thus we have $\mu(K_i) = \nu_i(I_i)$. Moreover, since $\operatorname{dist}(x, pr(x)) \leq \delta$,

$$\begin{aligned} \Big| \frac{1}{\mu(K_i)} \int_{K_i} f \; d\mu - & \frac{1}{\nu_i(I_i)} \int_{I_i} f \; d\nu_i \Big| \\ &= \Big| \frac{1}{\mu(K_i)} \int_{K_i} f(x) \; d\mu - \frac{1}{\mu(K_i)} \int_{K_i} f\big(pr(x)\big) \; d\mu \Big| \\ &\leq \frac{1}{\mu(K_i)} \int_{K_i} \big| f(x) - f\big(pr(x)\big) \big| \; d\mu \leq \omega(\epsilon) \end{aligned}$$

Therefore,

$$\left|\Phi\left(\frac{1}{\mu(K_i)}\int_K f\ d\mu\right) - \Phi\left(\frac{1}{\nu_i(I_i)}\int_{I_i} f\ d\nu_i\right)\right| \leq \omega_\Phi\left(\omega_f(\epsilon)\right).$$

Similarly,

$$\left|\frac{1}{\mu(K_i)}\int_{K_i} \Psi(f)\ d\mu - \frac{1}{\nu_i(I_i)}\int_{I_i} \Psi(f)\ d\nu_i\right| \leq \omega_\Psi(\epsilon).$$

Since

$$\Phi\left(\frac{1}{\nu_i(I_i)}\int_{I_i} f\ d\nu_i\right) \leq \frac{1}{\nu_i(I_i)}\int_{I_i} \Psi(f)\ d\nu_i,$$

we conclude that

$$\begin{aligned}\Phi\left(\frac{1}{\mu(K)}\int_K f\ d\mu\right) =&\Phi\left(\frac{1}{\mu(K_i)}\int_{K_i} f\ d\mu\right)\\ \leq&\frac{1}{\mu(K_i)}\int_{K_i} \Psi(f)\ d\mu + \omega_\Phi\left(\omega_f(\epsilon)\right) + \omega_\Psi(\epsilon) \quad \text{for all} \quad i \in I.\end{aligned}$$

Taking the convex combination of the above inequalities with the coefficients $\mu(K_i)/\mu(K)$, we get

$$\Phi\left(\frac{1}{\mu(K)}\int_K f\ d\mu\right) \leq \frac{1}{\mu(K)}\int_K \Psi(f)\ d\mu + \omega_\Phi\left(\omega_f(\epsilon)\right) + \omega_\Psi(\epsilon).$$

Taking the limit as $\epsilon \longrightarrow 0$, we complete the proof. □

Theorem 27.1 admits generalizations. Here is an interesting one. Let $k < n$ be a positive integer and let $\Phi, \Psi : \mathbb{R}^k \longrightarrow \mathbb{R}$ be continuous functions. Let $f_1, \ldots, f_k : K \longrightarrow \mathbb{R}$ be functions. Suppose we want to establish the inequality

$$\Phi\left(\frac{1}{\mu(K)}\int_K f_1\ d\mu, \ldots, \frac{1}{\mu(K)}\int_K f_k\ d\mu\right) \leq \frac{1}{\mu(K)}\int_K \Psi\left(f_1, \ldots, f_k\right)\ d\mu$$

for any log-concave measure μ. Then it suffices to establish the inequality for all k-dimensional convex compact subsets of K with any log-concave measure ν on them. The "halving procedure" 27.2 still works, but it has to be modified as follows.

Lecture 31. Friday, March 25

Lecture 30 on Wednesday, March 23, covered the material from the previous handout.

27. The needle decomposition technique for proving dimension-free inequalities, continued

(27.5) Definition. Let us call a convex body $K \subset \mathbb{R}^n$ an (k, ϵ)-*pancake* if there exists a k-dimensional affine subspace $A \subset \mathbb{R}^n$ such that all points of K are within distance ϵ from A.

Thus we get the following version (extension) of Theorem 27.1.

(27.6) Theorem. *Let $K \subset \mathbb{R}^n$ be a convex body, let μ be a measure on K with a positive continuous density, and let $f_1, \dots, f_k : K \longrightarrow \mathbb{R}$ be continuous functions. Then for any $\epsilon > 0$ there is decomposition*

$$K = \bigcup_{i \in I} K_i$$

of K into a finite union of non-overlapping (that is, with pairwise disjoint interiors) convex bodies $K_i \subset K$, $i \in I$, such that

(1) *Each K_i is a (k, ϵ)-pancake;*
(2) *The average value of every function f_j on every piece K_i is equal to the average value of f_j on the whole convex body K:*

$$\frac{1}{\mu(K_i)} \int_{K_i} f_j \, d\mu = \frac{1}{\mu(K)} \int_K f_j \, d\mu$$

for all $i \in I$ and $j = 1, \dots, k$.

The proof is very similar to that of Theorem 27.1. We have the following modification of the "halving" procedure of Section 27.4.

(27.7) Lemma. *Let $K \subset \mathbb{R}^n$ be a convex body, let μ be a measure on K with a positive continuous density, and let $f_1, \dots, f_k : K \longrightarrow \mathbb{R}$ be continuous functions. Let $L \subset \mathbb{R}^n$ be an affine subspace passing through an interior point of K and such that $\dim L = n - k - 1$. Then there exists an affine hyperplane $H \supset L$ which cuts K into two bodies K_1 and K_2 such that*

$$\frac{1}{\mu(K_1)} \int_{K_1} f_j \, d\mu = \frac{1}{\mu(K_2)} \int_{K_2} f_j \, d\mu \quad \textit{for} \quad j = 1, \dots, k.$$

Proof. We parameterize oriented hyperplanes $H \supset L$ by the points of the sphere $\mathbb{S}^k$. To do that, we choose the origin $0 \in L$ and project $\mathbb{R}^n$ along L onto $L^{\perp}$ which we identify with $\mathbb{R}^{k+1}$. Then each unit vector $u \in \mathbb{S}^k$ defines an oriented hyperplane $H \supset L$ with the normal vector u. Note that u and $-u$ define the same non-oriented hyperplane but different oriented hyperplanes. Let $K_+(u)$ be the part of the convex body lying in the same halfspace as u and let $K_-(u)$ be the part of the convex body lying in the opposite halfspace. Note that $K_+(u) = K_-(-u)$ and $K_-(u) = K_+(-u)$. Let us consider the map

$$\phi : \quad \mathbb{S}^k \longrightarrow \mathbb{R}^k$$

defined by

$$\begin{aligned}\phi(u) = \Bigg(&\frac{1}{\mu\left(K_+(u)\right)} \int_{K_+(u)} f_1 \ d\mu - \frac{1}{\mu\left(K_-(u)\right)} \int_{K_-(u)} f_1 \ d\mu, \quad \ldots, \\ &\frac{1}{\mu\left(K_+(u)\right)} \int_{K_+(u)} f_k \ d\mu - \frac{1}{\mu\left(K_-(u)\right)} \int_{K_-(u)} f_k \ d\mu \Bigg).\end{aligned}$$

Then ϕ is continuous and $\phi(-u) = -\phi(u)$. Hence by the Borsuk-Ulam Theorem, there is a point $u \in \mathbb{S}^k$ such that $\phi(u) = 0$. This point defines the desired affine hyperplane. □

Note that we necessarily have

$$\frac{1}{\mu(K_i)} \int_{K_i} f_j \ d\mu = \frac{1}{\mu(K)} \int_K f_j \ d\mu \quad \text{for} \quad j = 1, \ldots, k \quad \text{and} \quad i = 1, 2.$$

Proof of Theorem 27.6. We choose many but finitely many $(n-k-1)$-dimensional affine subspaces $L_1, \ldots, L_N$, each intersecting the interior of K and such that for every $(k+1)$-dimensional affine subspace A, if the intersection $A \cap K$ contains a ball of radius $\delta = \delta(\epsilon)$, the ball contains a point of the type $A \cap L_i$ for some i. The existence of such a set of subspaces follows by the standard compactness argument. We construct the bodies K_i as follows: starting with K, for each of the subspaces $L_1, \ldots, L_N$, one after another, we find all previously constructed convex bodies for which the subspace passes through an interior point, and halve all such bodies using Lemma 27.7. Eventually, we get the set of bodies K_i and Part (2) is clear. Moreover, for each K_i and any $(k+1)$-dimensional affine subspace A, the intersection $K_i \cap A$ does not contain a ball of radius δ, since otherwise there would have been a subspace L_j intersecting K_i in its interior point and we would have halved K_i along the way. We claim that we can choose δ (it may depend on the body K as well) so that every K_i is a (k, ϵ)-pancake. One possible way to prove it goes via the Blaschke selection principle. Suppose that there is an $\epsilon > 0$ such that for any positive integer m there is a convex body $C_m \subset K$ which does not contain a $(k+1)$-dimensional ball of radius $1/m$ but not a (k, ϵ)-pancake. Then there is a convex body $C \subset K$ which is the limit point of C_m in the Hausdorff metric. Then C contains no $(k+1)$-dimensional ball at all and hence must lie in a k-dimensional affine subspace $A \subset \mathbb{R}^n$. But then all C_m with sufficiently large m must be ϵ-close to A, which is a contradiction. □

Lecture 32. Monday, March 28

(28.1) Reducing inequalities for general functions to inequalities for particular functions. Suppose we want to prove reverse Hölder inequalities for some class of functions, say, polynomials of a given degree, which are valid for all convex bodies. Namely, we want to prove that for any positive integer d there exists a constant $c(d)$ such that for all convex bodies $K \subset \mathbb{R}^n$, for all polynomials $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that $\deg f \leq d$ and $f(x) > 0$ for all $x \in K$, one has

$$\ln\left(\frac{1}{\operatorname{vol} K}\int_K f \ dx\right) \leq c(d) + \frac{1}{\operatorname{vol} K}\int_K \ln f \ dx.$$

Validity of such an inequality follows from results of J. Bourgain. It follows from Theorem 27.1 that its suffices to check the above inequality only for polynomials that are essentially univariate, that is, have the structure

$$f(x) = g\bigl(T(x)\bigr),$$

where g is a univariate polynomial and $T : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a projection. Indeed, choosing $\Phi(x) = \ln x$, $\Psi(x) = c(d) + \ln x$, by Theorem 27.1, we conclude that it suffices to check the inequality

$$\ln\left(\frac{1}{\nu(I)}\int_I f \ d\nu\right) \leq c(d) + \frac{1}{\nu(I)}\int_I \ln f \ d\nu$$

For any interval $I \subset K$ and any log-concave measure ν on I. By changing coordinates, we can reduce the inequality to the following one: Given a univariate polynomial g with $\deg g \leq d$, which is non-negative on the interval $[0,1]$ and a log-concave measure ν on $[0,1]$, prove that

$$\ln\left(\frac{1}{\nu\bigl([0,1]\bigr)}\int_0^1 g \ d\nu\right) \leq c(d) + \frac{1}{\nu\bigl([0,1]\bigr)}\int_0^1 \ln g \ d\nu.$$

The structure of polynomials g non-negative on $[0,1]$ is well-known. Every such a polynomial is a convex combination of polynomials with all roots real and in the interval $[0,1]$.

On the other hand, every log-concave measure is the limit of a projection of the Lebesgue measure restricted to a convex body. Therefore, it suffices to prove the inequality

$$\ln\left(\frac{1}{\operatorname{vol} K}\int_K g\bigl(T(x)\bigr) \ dx\right) \leq c(d) + \frac{1}{\operatorname{vol} K}\int_K \ln g\bigl(T(x)\bigr) \ dx,$$

where $K \subset \mathbb{R}^n$ is a convex body, $T : K \longrightarrow \mathbb{R}$ is a projection such that $T(K) \subset [0,1]$ and g is a univariate polynomial non-negative on the interval $[0,1]$. Hence Theorem 27.1 allows us to restrict the class of polynomials considerably.

Moreover, suppose we *do* care about the ambient dimension n. Analyzing the proof of Theorem 27.1, we observe that ν is the limit of the projections of the Lebesgue measure restricted to convex bodies $K \subset \mathbb{R}^n$. Then ν is not only log-concave but a bit more is true. One can deduce from the Brunn-Minkowski Theorem for the Lebesgue measure (Corollary 20.3) that if ϕ is the density of ν, then $\phi^{1/(n-1)}$ is concave. Let

$$K = \Big\{(\xi, \eta_1, \ldots, \eta_{n-1}) : \quad 0 \leq \xi \leq 1 \quad \text{and} \\ 0 \leq \eta_i \leq \phi^{1/(n-1)}(\xi) \quad \text{for} \quad i = 1, \ldots, n-1\Big\}.$$

Then $K \subset \mathbb{R}^n$ is a convex body and ν is the push-forward of the Lebesgue measure restricted to K under the projection $\mathbb{R}^n \longrightarrow \mathbb{R}$, $(\xi, \eta_1, \ldots, \eta_{n-1}) \longmapsto \xi$. This shows that to establish the inequality

$$\ln\left(\frac{1}{\operatorname{vol} K} \int_K f \ dx\right) \leq c(d,n) + \frac{1}{\operatorname{vol} K} \int_K \ln f \ dx$$

for any convex body $K \subset \mathbb{R}^n$ and any polynomial f with $\deg f \leq d$ which is non-negative on K, it suffices to check the inequality for polynomials $f = g\big(T(x)\big)$, where $T : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the projection, $T(K) = [0,1]$ and g is a univariate polynomial $\deg g \leq d$, non-negative on $[0,1]$.

(28.2) Reducing an arbitrary log-concave measure to a log-affine measure. In view of what's been said, it would be helpful to understand the structure of log-concave measures on the interval. An important example of a log-concave measure on an interval is provided by a measure with the density e^{ax+b} for some numbers a and b. We call such measures *log-affine.*

The following construction is a particular case of a more general result by M. Fradelizi and O. Guédon. Let us fix the interval $[0,1]$, a continuous function $h : [0,1] \longrightarrow \mathbb{R}$, a number a, and consider the set $X_a(h)$ of all log-concave probability measures ν on the interval $[0,1]$ that satisfy the condition

$$\int_0^1 h \ d\nu = a.$$

Suppose that $\nu \in X_a(h)$ is a measure with a continuous density. Then, unless ν is a log-affine measure, it can be expressed as a convex combination

$$\nu = \alpha_1 \nu_1 + \alpha_2 \nu_2 : \quad \alpha_1 + \alpha_2 = 1 \quad \text{and} \quad \alpha_1, \alpha_2 \geq 0$$

of two distinct measures $\nu_1, \nu_2 \in X_a(h)$.

We sketch a proof below.

Without loss of generality, we may assume that $a = 0$ (otherwise, we replace h with $h - a$). Next, we may assume that

$$\int_0^x h(t) \ d\nu(t) \quad \text{for} \quad 0 < x < 1$$

does not change sign on the interval. Indeed, if it does, then there is a point $0 < c < 1$ such that

$$\int_0^c h\ d\nu = \int_c^1 h\ d\nu = 0.$$

We define ν_1 as the normalized restriction of ν onto $[0, c]$ and ν_2 as the normalized restriction of ν onto $[c, 1]$.

Hence we assume that

$$\int_0^x h(t)\ d\nu(t) \geq 0 \quad \text{for all} \quad 0 \leq x \leq 1.$$

Let $e^{\phi(x)}$ be the density of ν, so ϕ is a concave function on $[0, 1]$. We are going to modify ϕ as follows. Let us pick a $c \in (0, 1)$ and let ℓ be an affine function on the interval $[0, 1]$ such that $\ell(c) < \phi(c)$ and the slope of ℓ is some number α (to be adjusted later). Let

$$\psi(x) = \min\bigl\{\phi(x),\ \ell(x)\bigr\} \quad \text{for} \quad 0 \leq x \leq 1\bigr\}$$

and let us consider the measure ν_1 with the density e^{ψ}. Note that ψ is a concave function. Now, if $\alpha = +\infty$ (or close), so $\ell(x)$ steeply increases, then $e^{\psi(x)} = e^{\phi(x)}$ for $x > c$ and $e^{\psi(x)}$ is 0 for $x < c$. Thus we have

$$\int_0^1 h\ d\nu_1 = \int_c^1 h\ d\nu \leq 0.$$

Similarly, if $\alpha = -\infty$ (or close) then $\ell(x)$ steeply decreases and $e^{\psi(x)} = e^{\phi(x)}$ for $x < c$ and $e^{\psi(x)}$ is 0 for $x > c$. Thus we have

$$\int_0^1 h\ d\nu_1 = \int_0^c h\ d\nu \geq 0.$$

This proves that there exists a slope α such that

$$\int_0^1 h\ d\nu_1 = 0.$$

Let $\nu_2 = \nu - \nu_1$. Then ν_2 is supported on the interval where $\phi(x) > \ell(x)$ and its density there is $e^{\phi(x)} - e^{\ell(x)}$. We claim that ν_2 is log-concave. Factoring the density of ν_2 into the product $e^{\ell(x)}\left(e^{\phi(x)-\ell(x)} - 1\right)$, we reduce the proof to the following statement: if $\rho(x) = \phi(x) - \ell(x)$ is a non-negative concave function, then $\ln\left(e^{\rho(x)} - 1\right)$ is concave. Since a concave non-negative function is a pointwise minimum of a family of linear (affine) non-negative functions, it suffices to prove that

$$\ln\left(e^{\alpha x + \beta} - 1\right)$$

is concave whenever defined. This is obtained by checking that the second derivative, equal to

$$-\frac{\alpha^2}{\left(e^{\alpha x+\beta}-1\right)^2}$$

is non-positive.

Hence we obtain a decomposition $\nu = \nu_1 + \nu_2$, where ν_1 and ν_2 are log-concave, and

$$\int_0^1 h \ d\nu_1 = \int_0^1 h \ d\nu_2 = 0.$$

Normalizing ν_1 and ν_2, we complete the proof.

Suppose now that we want to check the inequality

$$\ln\left(\int_0^1 g \ d\nu\right) \leq c(d) + \int_0^1 \ln g \ d\nu$$

for all log-concave probability measures ν on the interval $[0,1]$. Let us fix the value of

$$\int_0^1 g \ d\nu = a, \tag{28.2.1}$$

say. Then the infimum of

$$\int_0^1 \ln g \ d\nu$$

is attained on an extreme point of the set of log-concave measures ν satisfying (28.2.1), which must be a log-affine measure or a δ-measure.

Hence in proving reverse Hölder inequalities on an interval, we can always restrict ourselves to log-affine measures on the interval.

Similarly, M. Fradelizi and O. Guédon show that if X_a is the set of probability densities ϕ such that

$$\int_0^1 h\phi(x) \ dx = a$$

and $\phi^{1/n}$ is concave, then the extreme points of X_a are the densities $\ell^n(x)$, where $\ell : [0,1] \longrightarrow \mathbb{R}$ is an affine function, non-negative on $[0,1]$.

Lecture 33. Wednesday, March 30

29. Graphs, eigenvalues, and isoperimetry

We consider some global invariants in charge of the isoperimetric properties of a metric space. We discuss the discrete situation first.

Let $G=(V,E)$ be a finite undirected graph without loops or multiple edges, with the set V of vertices and the set E of edges. With this, V becomes a metric space: the distance $\operatorname{dist}(u,v)$ between two vertices $u,v \in V$ is the length (the number of edges) of a shortest path in G connecting u and v. Thus the diameter of this metric space is finite if and only if G is connected. We consider connected graphs only. Given a vertex $v \in V$, the number of edges incident to v is called the *degree* of v and denoted $\deg v$. With a graph G, we associate a square matrix $A(G)$, called the *Laplacian* of G.

(29.1) Definition. Let $G=(V,E)$ be a graph. Let us consider the $|V| \times |V|$ matrix $A=A(G)$, $A=a_{u,v}$, where

$$a_{u,v}=\begin{cases} \deg u & \text{if } u=v \\ -1 & \text{if } \{u,v\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

We think of $A(G)$ as of an operator on the space $\mathbb{R}^V$ of functions $f: V \longrightarrow \mathbb{R}$. Namely, $g=Af$ if

$$g(v)=(\deg v)f(v)-\sum_{\substack{u\in V\\ \{u,v\}\in E}} f(u).$$

We consider the scalar product on $\mathbb{R}^V$ defined by

$$\langle f,g\rangle=\langle f,g\rangle_V=\sum_{v\in V} f(v)g(v).$$

Here is the first result.

(29.2) Lemma. *Let $G=(V,E)$ be a graph. Let us orient every edge $e \in E$, so that one vertex of e becomes the beginning (denoted e_+) and the other becomes the end (denoted e_-). Let $L=L(\tilde{G})$ (where $\tilde{}$ stands for an orientation), $L=l_{e,v}$, be the $|E| \times |V|$ incidence matrix of G:*

$$l_{e,v}=\begin{cases} 1 & \textit{if } v=e_+ \\ -1 & \textit{if } v=e_- \\ 0 & \textit{otherwise.} \end{cases}$$

Then

$$A(G)=\tilde{L}^T(G)\tilde{L}(G).$$

Proof. The (u,v)th entry of $L^T L$ is

$$\sum_{e \in E} l_{e,u} l_{e,v}.$$

If $u = v$ then the eth term is 1 if and only if e is incident to v and 0 otherwise, so the entry is $\deg v$, If $u \neq v$ and $\{u,v\}$ is an edge, then the only one non-zero term corresponds to the edge $e = \{u,v\}$ and is equal to -1. If $u \neq v$ and $\{u,v\} \notin E$, then all the terms are 0. □

We can think of L as of the linear transformation $\mathbb{R}^V \longrightarrow \mathbb{R}^E$, which, for each function $f : V \longrightarrow \mathbb{R}$ computes the function $g = Lf$, $g : E \longrightarrow \mathbb{R}$ defined by $g(e) = f(e_+) - f(e_-)$. We introduce the scalar product on $\mathbb{R}^E$ by

$$\langle f, g \rangle = \langle f, g \rangle_E = \sum_{e \in E} f(e) g(e).$$

Then L^T is the matrix of the conjugate linear transformation $\mathbb{R}^E \longrightarrow \mathbb{R}^V$. For a function $g : E \longrightarrow \mathbb{R}$, the function $f = L^T g$, $f : V \longrightarrow \mathbb{R}$ is defined by

$$f(v) = \sum_{\substack{e \in E \\ e_+ = v}} g(e) - \sum_{\substack{e \in E \\ e_- = v}} g(e).$$

We have

$$\langle Lf, g \rangle_E = \langle f, L^T g \rangle_V \quad \text{for all} \quad f \in \mathbb{R}^V \quad \text{and all} \quad g \in \mathbb{R}^E.$$

(29.3) Corollary. *The matrix $A(G)$ is positive semidefinite. If G is connected, then the eigenspace of $A(G)$ corresponding to the eigenvalue $\lambda = 0$ consists of the functions $f : V \longrightarrow \mathbb{R}$ that are constant on all vertices:*

$$f(v) = c \quad \textit{for some} \quad c \in \mathbb{R} \quad \textit{and all} \quad v \in V.$$

Proof. Since $A = L^T L$, we have

$$\langle Af, f \rangle = \langle L^T L f, f \rangle = \langle Lf, Lf \rangle \geq 0.$$

If f is a constant on the vertices of G, then $Lf = 0$ and $\langle Af, f \rangle = 0$. On the other hand, suppose that $\langle Af, f \rangle = 0$. Let us pick two different vertices $u, v \in V$. Since G is connected, there is an orientation $\tilde{G}$ in which u and v are connected by a directed path (say, u is the beginning of the path and v is an end). Let $L = L(\tilde{G})$ be the corresponding operator. We must have $Lf = 0$, which means that $f(e_+) = f(e_-)$ for every edge e of G. Since u and v are connected by a directed path, we must have $f(u) = f(v)$. Since u and v were arbitrary, f must be constant on V. □

PROBLEM. Check that in general, the 0th eigenspace of A consists of the functions that are constant on connected components of G.

In what follows, the crucial role is played the *smallest positive eigenvalue* $\lambda = \lambda(G)$ of $A(G)$. We will rely on the following simple consideration: if $f \in \mathbb{R}^V$ is a function orthogonal to the 0th eigenspace, then, of course,

$$\langle Af, f \rangle \geq \lambda \langle f, f \rangle = \lambda \|f\|^2.$$

This is called sometimes the *Poincaré inequality*, sometimes the *Rayleigh principle.* If G is connected, being orthogonal to the 0th eigenspace means that

$$\sum_{v \in V} f(v) = 0.$$

We are going to relate $\lambda(G)$ with isoperimetric properties of V.

(29.4) Theorem. *Let $G = (V, E)$ be a connected graph and let $X \subset V$ be a set of vertices. Let $E(X, V \setminus X)$ be the set of all edges $e \in E$ with one endpoint in X and the other in $V \setminus X$.*

Then

$$|E(X, V \setminus X)| \geq \lambda(G) \frac{|X||V \setminus X|}{|V|}$$

Proof. Let us consider the indicator $[X] : \mathbb{R}^V \longrightarrow \mathbb{R}$, that is, the function equal to 1 on the points from X and 0 otherwise. To be able to apply the Poincaré inequality, we modify $[X]$ it to make it orthogonal to the 0th eigenspace, for which purpose we subtract from $[X]$ the average value of $[X]$ on V, which is equal to $|X|/|V| = p$, say. Thus we consider the function $f : V \longrightarrow \mathbb{R}$, $f = [X] - p[V]$,

$$f(v) = \begin{cases} 1 - p & \text{if } v \in X \\ -p & \text{if } v \notin X, \end{cases}$$

where $p = |X|/|V|$. We have

$$\begin{aligned} \|f\|^2 =& \langle f, f \rangle = \big\langle [X] - p[V],\ [X] - p[V] \big\rangle \\ =& \langle [X], [X] \rangle - 2p \langle [X], [V] \rangle + p^2 \langle [V], [V] \rangle = |X| - 2p|X| + p^2|V| \\ =& |X| - 2p|X| + p|X| = (1 - p)|X| \\ =& \frac{|X||V \setminus X|}{|V|}. \end{aligned}$$

We have

$$\langle Af, f \rangle \geq \lambda \|f\|^2 = \lambda \frac{|X||V \setminus X|}{|V|}.$$

Now we compute $\langle Af, f\rangle$ directly. We have

$$Af = A[X] - pA[V] = A[X].$$

So

$$\langle Af, f\rangle = \langle Af, [X] - p[V]\rangle = \langle Af, [X]\rangle - p\langle Af, [V]\rangle = \langle A[X], X\rangle,$$

where $\langle Af, [V]\rangle = 0$ since $[V]$ lies in the 0th eigenspace, f is orthogonal to the 0th eigenspace and so is Af.

Let us denote $g = A[X]$. Then

$$g(v) = (\deg v)[X] - \sum_{\substack{u \in V \\ \{u,v\} \in E}} [X](u).$$

If $v \in X$ then $g(v)$ is equal to the number of edges with one endpoint at v and the other outside of X. Therefore,

$$\langle A[X], [X]\rangle = \sum_{v \in V} g(v) = |E(X, V \setminus X)|,$$

from which the proof follows. □

(29.5) Eigenvalues and products. Given two graphs $G_i = (V_i, E_i)$, $i = 1, 2$, we define their *product* $G = (V, E)$ as the graph with the set $V = V_1 \times V_2$ of vertices and the edges between vertices (u_1, u_2) and (w_1, w_2), if either $u_1 = w_1 \in V_1$ and $\{u_2, w_2\} \in E_2$ or $u_2 = w_2 \in V_2$ and $\{u_1, w_1\} \in E_1$. Then

$$A(G) = A(G_1) \otimes I_{V_2} + I_{V_1} \otimes A(G_2).$$

From this, the eigenvalues of $A(G)$ are the pairwise sums of the eigenvalues of $A(G_1)$ and $A(G_2)$ (the corresponding eigenvectors are the tensor products of the eigenvectors for $A(G_1)$ and $A(G_2)$). Hence

$$\lambda(G) = \min\Big\{\lambda(G_1),\ \lambda(G_2)\Big\}.$$

For example, the 1-skeleton I_n of the n-dimensional cube is the nth power of the graph consisting of two vertices and the edge connecting them. Hence $\lambda(I_n) = 2$. If we choose X to the the set of vertices on one facet of the cube, then the left hand side in the formula of Theorem 29.4 is equal to 2^{n-1}, and this is exactly the right hand side.

Friday, April 1

Student's presentation: Bourgain's Theorem on embedding of finite metric spaces

---

Monday, April 4

Student's presentation: Talagrand's convex hull inequality

Wednesday, April 6

Student's presentation: Talagrand's convex hull inequality, continued

---

Friday, April 8

Student's presentation: the isoperimetric inequality on the sphere

---

Monday, April 11

Student's presentation: the isoperimetric inequality on the sphere, continued

---

Wednesday, April 13

Visitor's presentation: concentration for the singular values of a random matrix

---

Friday, April 15

Visitor's presentation: concentration for the singular values of a random matrix, continued

---

Monday, April 18

Student's presentation: a proof of the isoperimetric inequality in the Boolean cube

The last lecture: prepared but not delivered (end of term)

29. Graphs, eigenvalues, and isoperimetry, continued

Theorem 29.4 implies the following bound.

(29.6) Corollary. *Let $G=(V,E)$ be a connected graph and let d be the minimum degree of a vertex $v \in V$. Then*

$$\lambda(G) \leq \frac{|V|}{|V|-1} d.$$

Proof. Let $X=\{v\}$ with $\deg v = d$ in Theorem 29.4. Then $E(X, V \setminus X) = d$. □

Theorem 29.4 can be extended to a more general situation. The following result is due to N. Alon and V. Milman.

(29.7) Theorem. *Let $G=(V,E)$ be a connected graph and let $A, B \subset V$ be two disjoint subsets. Let $E(A)$ be the set of all edges with both endpoints in A and let $E(B)$ be the set of edges with both endpoints in B. Let ρ be the distance between A and B, that is, the minimum of* $\operatorname{dist}(u,v)$ *with $u \in A$ and $v \in B$. Then*

$$|E| - |E(A)| - |E(B)| \geq \lambda(G) \rho^2 \frac{|A||B|}{|A|+|B|}.$$

Indeed, taking $A = X$, $B = V \setminus X$ and $\rho = 1$, we get Theorem 29.4.

Proof. The proof consists of applying the inequality $\langle Aff, f \rangle \geq \lambda \|f\|^2$ to a specially constructed function.

Let

$$a = \frac{|A|}{|V|} \quad \text{and} \quad b = \frac{|B|}{|V|}$$

and let

$$g(v) = \frac{1}{a} - \frac{1}{\rho}\left(\frac{1}{a} + \frac{1}{b}\right) \min\bigl(\operatorname{dist}(v, A),\ \rho\bigr).$$

Furthermore, let

$$p = \frac{1}{|V|} \sum_{v \in V} g(v)$$

and let us define $f : V \longrightarrow \mathbb{R}$ by

$$f(v) = g(v) - p.$$

Now,

$$\sum_{v \in V} f(v) = 0,$$

so we may (and will) apply the inequality

$$\langle Af, f \rangle \geq \lambda \langle f, f \rangle$$

to this particular function f.

First, we note that $g(v) = 1/a$ for $v \in A$ and $g(v) = -1/b$ for $v \in B$. Then

$$\begin{aligned}
\langle f, f \rangle =& \sum_{v \in V} f^2(v) \geq \sum_{v \in A \cup B} f^2(v) = \sum_{v \in A \cup B} \big(g(v) - p\big)^2 \\
=& \sum_{v \in A} \big(g(v) - p\big)^2 + \sum_{v \in B} \big(g(v) - p\big)^2 = \sum_{v \in A} \left(\frac{1}{a} - p\right)^2 + \sum_{v \in B} \left(-\frac{1}{b} - p\right)^2 \\
=& \sum_{v \in A} \left(\frac{1}{a^2} - \frac{2p}{a} + p^2\right) + \sum_{v \in B} \left(\frac{1}{b^2} + \frac{2p}{a} + p^2\right) \\
=& \frac{|V|^2}{|A|} + \frac{|V|^2}{|B|} + p^2\left(|A| + |B|\right) \geq |V| \left(\frac{1}{a} + \frac{1}{b}\right).
\end{aligned}$$

Next, we observe that if $\{u, v\} \in E$ then

$$|g(u) - g(v)| \leq \frac{1}{\rho}\left(\frac{1}{a} + \frac{1}{b}\right), \quad \text{and, consequently} \quad |f(u) - f(v)| \leq \frac{1}{\rho}\left(\frac{1}{a} + \frac{1}{b}\right).$$

Let us orient G in an arbitrary way, and let L be the corresponding operator, so that $A = L^T L$. Then

$$\begin{aligned}
\langle Af, f \rangle =& \langle Lf, Lf \rangle = \langle Lg, Lg \rangle = \sum_{e \in E} \big(g(e_+) - g(e_-)\big)^2 \\
=& \sum_{e \in E \setminus (E(A) \cup E(B))} \big(g(e_+) - g(e_-)\big)^2 \\
\leq& \Big(|E| - |E(A)| - |E(B)|\Big) \frac{1}{\rho^2} \left(\frac{1}{a} + \frac{1}{b}\right)^2.
\end{aligned}$$

Therefore, $\langle Af, f \rangle \geq \lambda \langle f, f \rangle$ implies that

$$\Big(|E| - |E(A)| - E(B)|\Big) \geq \lambda \rho^2 |V| \left(\frac{1}{a} + \frac{1}{b}\right)^{-1} = \lambda \rho^2 \frac{|A||B|}{|A| + |B|}.$$

□

Theorem 29.7 admits an extension to the continuous case.

30. Expansion and concentration

Let $G=(V,E)$ be a graph and let $X\subset V$ be a set of vertices. We would like to estimate how fast the ϵ-neighborhood of X grows. Theorem 29.4 provides a way to do it. Suppose that $\deg v\le d$ for all $v\in V$. For an $X\subset V$ and a positive r, let

$$X(r)=\Big\{v\in V:\quad \operatorname{dist}(v,X)\le r\Big\}$$

be the r-neighborhood of X. Let us introduce the counting probability measure μ on V:

$$\mu(A)=\frac{|A|}{|V|}\quad\text{for}\quad A\subset V.$$

We note that every edge $e\in E(X,V\setminus X)$ is incident to one vertex in X and one vertex in $X(1)\setminus X$. Therefore,

$$|X(1)\setminus X|\ge\frac{1}{d}E(X,V\setminus X)\ge\frac{\lambda(G)}{d}\frac{|X||V\setminus X|}{|V|}.$$

Therefore,

$$\mu\Big(X(1)\setminus X\Big)\ge\frac{\lambda}{d}\mu(X)\big(1-\mu(X)\big)$$

and

$$\mu\big(X(1)\big)\ge\left(1+\frac{\lambda}{d}\big(1-\mu(X)\big)\right)\mu(X).$$

In particular,

$$\mu\big(X(1)\big)\ge\left(1+\frac{\lambda}{2d}\right)\mu(X)\quad\text{if}\quad\mu(X)\le\frac{1}{2},\tag{30.1}$$

Which demonstrates a sustained grows of the neighborhood as long as the measure of the set stays below 1/2. Iterating, we get for positive integer r

$$\mu\big(X(r)\big)\ge\left(1+\frac{\lambda}{2d}\right)^r\mu(X)\quad\text{provided}\quad\mu\big(X(r-1)\big)\le\frac{1}{2}.$$

One can deduce a concentration result from this estimate. It follows that

$$\mu\big(X(r)\big)\le\frac{1}{2}\Longrightarrow\mu(X)\le\frac{1}{2}\left(1+\frac{\lambda}{2d}\right)^{-r}.$$

Let $Y\subset V$ be a set such that $\mu(Y)\ge 1/2$ and let $X=V\setminus Y(r)$. Then $X(r)\subset V\setminus Y$ and so $\mu\big(X(r)\big)\le 1/2$. Thus we obtain

$$\mu(Y)\ge\frac{1}{2}\Longrightarrow\mu\Big\{v\in V:\quad\operatorname{dist}(v,Y)>r\Big\}\le\frac{1}{2}\left(1+\frac{\lambda}{2d}\right)^{-r},$$

which is a concentration result of some sort. There are examples of graphs for which the estimate thus obtained is quite reasonable, but often it is too weak. For example, for the 1-skeleton of the n-dimensional cube with $\lambda/2d=1/n$ (see Section 29.5) it is nearly vacuous.

One can get stronger estimates if instead of Theorem 29.4 we use Theorem 29.7.

(30.2) Theorem. *Let $G=(V,E)$ be a connected graph and let $X \subset V$ be a set. Suppose that* $\deg v \leq d$ *for all $v \in G$. Then, for $r \geq 1$, we have*

$$\mu\bigl(X(r)\bigr) \geq \left(1+\frac{\lambda(G)r^2}{2d}\right)\mu(X) \quad \textit{provided} \quad \mu\bigl(X(r)\bigr) \leq \frac{1}{2}.$$

Proof. Let us apply Theorem 29.7 to $X=A$ and $B=V \setminus X(r)$. Then the set $E \setminus \bigl(E(A) \cup E(B)\bigr)$ consists of the edges with one endpoint in $X(r) \setminus X$ and the other outside of $X(r) \setminus X$. Therefore,

$$|X(r) \setminus X| \geq \frac{\lambda(G)r^2}{d}\frac{|X||V \setminus X(r)|}{|X|+|V \setminus X(r)|}.$$

Since

$$\frac{|V \setminus X(r)|}{|X|+|V \setminus X(r)|} = \left(1+\frac{|X|}{|V \setminus X(r)|}\right)^{-1} \geq \frac{1}{2},$$

the result follows. □

Note that the estimate holds for not necessarily integer r.

Comparing the estimate of Theorem 30.2 with that of (30.1), we observe that the former is stronger if the ratio λ/d is small, since the iteration of (30.1) produces the lower bound

$$\left(1+\frac{\lambda}{2d}\right)^r \approx 1+\frac{\lambda r}{2d} < 1+\frac{\lambda r^2}{2d}.$$

We obtain the following corollary.

(30.3) Corollary. *Let $G=(V,E)$ be a connected graph and let $A \subset V$ be a set such that $\mu(A) \geq 1/2$. Suppose that* $\deg v \leq d$ *for all $v \in V$. Then, for $t \geq 1$, we have*

$$\mu\Bigl\{v \in V: \quad \operatorname{dist}(v,A) > t\Bigr\} \leq \frac{1}{2}\inf_{\substack{r:\\ t>r>1}}\left(1+\frac{\lambda r^2}{2d}\right)^{-\lfloor t/r \rfloor}.$$

Proof. We consider $B=V \setminus A(t)$. Thus $B(t) \subset V \setminus A$ and hence $\mu\bigl(B(t)\bigr) \leq 1/2$. Let $s=\lfloor t/r \rfloor$ and let us construct a sequence of subsets $X_0=B, X_1=X_0(r), \ldots, X_k=X_{k-1}(r)$ for $k \leq s$. Thus $X_s=B_{rs} \subset B_t$, so $\mu\bigl(X_k\bigr) \leq 1/2$ for $k=0,1\ldots,s$. Applying Theorem 30.2 to $X_0, X_1, \ldots, X_s$, we conclude that

$$\frac{1}{2} \geq \mu(X_s) \geq \left(1+\frac{\lambda r^2}{2d}\right)^s \mu(B)$$

and hence

$$\mu(B) \leq \frac{1}{2}\left(1+\frac{\lambda r^2}{2d}\right)^{-s}.$$

Now we optimize on r. □

As follows from Corollary 29.6, $\lambda/2d \leq 1$, so we can choose

$$r = \sqrt{\frac{2d}{\lambda}}.$$

in Corollary 30.3. This gives us the estimate

$$\mu\Big\{v \in V : \quad \operatorname{dist}(v, A) > t\Big\} \leq 2^{-\lfloor t\sqrt{\lambda/2d}\rfloor - 1}$$

In the case of the 1-skeleton of the cube I_n, we have $\lambda = 2$ and $2d = 2n$, we get an apper bound of the type $\exp\big\{-ct/\sqrt{n}\big\}$, which is, although substantially weaker than the bound of Corollary 4.4, say, is not obvious and provides, at least, the right scale for t so that $A(t)$ is almost everything.