Lecture 1

Introduction to Fourier Analysis

Jan 7, 2005
Lecturer: Nati Linial
Notes: Atri Rudra & Ashish Sabharwal

1.1 Text

The main text for the first part of this course would be

- T. W. Körner, *Fourier Analysis*

The following textbooks are also "fun"

- H. Dym and H. P. Mckean, *Fourier Series and Integrals*.
- A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications, Vols. I, II*.

The following text follows a more terse exposition

- Y. Katznelson, *An Introduction to Harmonic Analysis*.

1.2 Introduction and Motivation

Consider a vector space V (which can be of finite dimension). From linear algebra we know that at least in the finite-dimension case V has a basis. Moreover, there are more than one basis and in general different bases are the same. However, in some cases when the vector space has some additional structure, some basis might be preferable over others. To give a more concrete example consider the vector space $V = \{f : X \to \mathbb{R} \text{ or } \mathbb{C}\}$ where X is some universe. If $X = \{1, \cdots, n\}$ then one can see that V is indeed the space $\mathbb{R}^n$ or $\mathbb{C}^n$ respectively in which case we have no reason to prefer any particular basis. However, if X is an abelian group[1] there may be a reason to prefer a basis $\mathcal{B}$ over others. As an example, let us consider $X = \mathbb{Z}/n\mathbb{Z}$, $V = \{(y_0, \cdots, y_n) | y_i \in \mathbb{R}\} = \mathbb{R}^n$. We now give some scenarios (mostly inspired by the engineering applications of Fourier Transforms) where we want some properties for $\mathcal{B}$ aka our "wish list":

[1] An abelian group is given by $\langle S, + \rangle$ where S is the set of elements which is closed under the commutative operation $+$. Further there exists an identity element 0 and every element in S has an inverse.

1. Think of the elements of $\mathbb{Z}/n\mathbb{Z}$ as time units and let the vector $\boldsymbol{y} = (y_1, \cdots, y_n)$ denote some measurements taken at the different time units. Now consider another vector $\boldsymbol{z} = \{z_1, \cdots, z_n\}$ such that for all i, $z_i = y_{i+1 \mod (n)}$. Note that the vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ are different though from the measurement point of view they are not much different– they correspond to the same physical situation when our clock was shifted by one unit of time. Thus, with this application in mind, we might want to look for a basis for V such that the representation of $\boldsymbol{z}$ is "closely related" to that of $\boldsymbol{y}$.

2. In a setting more general than the previous one, if for $f : X \to \mathbb{R}$, a given member of V and $a \in X$, $g : X \to \mathbb{R}$ be such that $g(x) = f(x + a)$ then we would like to have representations of f and g being "close" in $\mathcal{B}$ for any f and a. Note that in the previous example x corresponds to the index i and $a = 1$.

3. In situations where derivatives (or discrete analogues thereof are well-defined), we would like f' to have a representation similar to that of f.

4. In real life, signals are never nice and smooth but suffer from noise. One way to reduce the noise is to "average-out" the signal. As a concrete example let the signal samples be given by $f_0, \cdots, f_{n-1}$ then the smoothened out signal could be given by the samples $g_i = \frac{1}{4}f_{i-1} + \frac{1}{2}f_i + \frac{1}{4}f_{i+1}$. Define $g_{-1} = \frac{1}{4}, g_0 = \frac{1}{2}, g_1 = \frac{1}{4}$. We now look at an new operator: *convolution* which is defined as follows. Let g be f convolved where $h = f * g$ and $h(x) = \sum_y f(y)g(x - y)$. So another natural property of $\mathcal{B}$ is that the representation of $f * g$ should be related to that of f and g.

Before we go ahead, here are some frequently used instantiations of X:

- $X = \mathbb{Z}/n\mathbb{Z}$. This is the Discrete Fourier Transform (DFT).
- $X = \mathbb{T} = \{$Real Numbers mod 1, addition mod 1$\} \cong \{e^{i\theta}$,multiplication$\}$. The isomorphism exists because multiplication of elements in the second group is the same as addition mod 2π of the angles θ.

  This is the most classical case of the theory, covered by the book *Trigonometric Polynomials* by Zygmund.
- $X = (\mathbb{R}, +)$. This is the Real Fourier Transform. In this case, in order to get meaningful analysis, one has to restrict the family of functions $f : X \to \mathbb{R}$ under consideration e.g. ones with converging integrals or those with compact support. The more general framework is that of Locally compact Abelian groups.
- $X = \{0, 1\}^n$ where the operations are done mod 2. Note that $\{f : X \to \{0, 1\}\}$ are simply the boolean functions.

1.3 A good basis

As before let X be an abelian group and define the vector space $V = \{f : X \to \mathbb{R}\}$.

Definition 1.1. The *characters* of X is the set $\{\chi : X \to \mathbb{C} | \chi$ is a homomorphism$\}$.

By homomorphism we mean that the following relationship holds: $\chi(x+y) = \chi(x) \cdot \chi(y)$ for any $x, y \in X$. As a concrete example, $X = \mathbb{Z}/n\mathbb{Z}$ has n distinct characters and the jth character is given by $\chi_j(x) = \omega^{jx}$ for any $x \in X$ where $\omega = e^{2\pi i/n}$.

We now state a general theorem without proof (we will soon prove a special case):

Theorem 1.1. *Distinct characters (considered as functions from $X \to \mathbb{C}$) are orthogonal*[2].

We now have the following fact:

Fact 1.1. If X is a finite abelian group of n elements, then X has n distinct characters which form an orthogonal basis for $V = \{f : X \to \mathbb{R}\}$.

Consider the special case of $X = \mathbb{Z}/n\mathbb{Z}$. We will show that the characters are orthogonal. Recall that in this case, $\chi_j(x) = \omega^{jx}$. By definition, $\langle \chi_j, \chi_k \rangle = \sum_{x=0}^{n-1} \omega^{jx}\omega^{-kx} = \sum_{x=0}^{n-1} \omega^{(j-k)x}$. If $j = k$ then each term is one and the inner product evaluates to n. If $j \neq k$, then summing up the geometric series, we have $\langle \chi_j, \chi_k \rangle = \frac{(\omega^{j-k})^n - 1}{\omega^{j-k}} = 0$. The last equality follows from the fact that $\omega^n = 1$.

We will take a quick detour and mention some applications where Fourier Analysis has had some measure of success:

- Coding Theory. A code $\mathcal{C}$ is a subset of $\{0,1\}^n$ where we want each element to be as far as possible from each other (where far is measured in terms of the hamming distance). We would like $\mathcal{C}$ to be as large as possible while keeping the distance as large as possible. Note that these are two opposing goals.
- Influence of variables on boolean functions. Say you have an array of sensors and there is some function which computes an answer. If a few of the sensors fail then answers should not change: in other words we need to find functions that are not too influenced by their variables.
- Numerical Integration/ Discrepancy. Say you want to integrate over some domain Ω. Of course one cannot find the exact integral if one does not have have an analytical expression of the function. So one would sample measurements at some discrete points and try and approximate the integral. Suppose that we further know that certain subdomains of Ω are significant for the computation. The main question is how to spread n points in Ω such that every "significant" region is sampled with the "correct" number of points.

1.4 A Rush Course in Classical Fourier Analysis

Let $X = \mathbb{T} = \left(\left\{e^{i\theta} \mid 0 \leq \theta < 2\pi\right\}, \text{multiplication}\right)$. Let $f : \mathbb{T} \to \mathbb{C}$, which can alternatively be thought of as a periodic function $f : \mathbb{R} \to \mathbb{C}$. What do characters of X look like?

There are infinitely many characters and each is a periodic function from $\mathbb{R}$ to $\mathbb{C}$. In fact, every character of X is a function $\chi : \mathbb{X} \to \mathbb{T}$, i.e. $\chi : \mathbb{T} \to \mathbb{T}$. Being a homomorphism, it must also satisfy $\chi(x.y) =$

[2]There is natural notion of inner-product among functions $f, g : X \to \mathbb{R}$. $\langle f, g \rangle = \sum_{x \in X} f(x)g(x)$ in the discrete case and $\langle f, g \rangle = \int f(x)g(x)\, dx$ in the continuous case. If the functions maps into $\mathbb{C}$, then $g(x)$ is replaced by its conjugate $\overline{g}(x)$ in the expressions. Finally f and g are *orthogonal* if $\langle f, g \rangle = 0$

$\chi(x).\chi(y)$. This implies that the only continuous characters of X are $\chi_k(x) = x^k, k \in \mathbb{Z}$. Note that if $k \notin \mathbb{Z}$, then x^k can have multiple values, discontinuities, etc. It is an easy check to see that $\langle \chi_k, \chi_l \rangle = \delta_{kl}$:

$$\begin{aligned}
\langle \chi_k, \chi_l \rangle &= \frac{1}{2\pi} \int_{\mathbb{T}} \chi_k(x)\, \overline{\chi_l(x)}\, dx \\
&= \frac{1}{2\pi} \int_{\mathbb{T}} x^k x^{-l}\, dx \\
&= \frac{1}{2\pi} \int_{\mathbb{T}} x^{k-l} dx \\
&= \frac{1}{2\pi} \int_0^{2\pi} e^{i\theta(k-l)}\, d\theta \\
&= \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases} \\
&= \delta_{kl}
\end{aligned}$$

How do we express a given $f : \mathbb{T} \to \mathbb{C}$ in the basis of the characters of X? Recall that if V is a finite dimentional vector space over a field $\mathbb{F}$ with an inner product and $u_1, \ldots, u_n$ is an orthonormal basis for V, then every $f \in V$ can be expressed as

$$f = \sum_{j=1}^{n} a_j u_j, \quad a_j \in \mathbb{F}, a_j = \langle f, u_j \rangle \tag{1.1}$$

We would like to obtain a similar representation of f in the basis of the characters $\chi_k, k \in \mathbb{Z}$.

Definition 1.2 (Fourier Coefficients). For $r \in \mathbb{Z}$, the r^{th} Fourier coefficient of f is

$$\hat{f}(r) = \frac{1}{2\pi} \int_{\mathbb{T}} f(t)\, e^{-irt}\, dt$$

The analogue of Equation 1.1 now becomes

$$S_n(f,t) = \sum_{r=-n}^{n} \hat{f}(r)\, e^{irt}, \quad \text{does } \lim_{n\to\infty} S_n(f,t) = f(t)? \tag{1.2}$$

Here $\hat{f}(r)$ replaces a_j and $\chi_r(e^{it})$ replaces u_j in Equation 1.1. In a dream world, we would like to ask whether $\sum_{r=-\infty}^{\infty} \hat{f}(r)\, e^{irt} \stackrel{?}{=} f(t)$ holds. We are, however, being more careful and asking the question by making this sum go from $-n$ to n and considering the limit as $n \to \infty$.

1.4.1 Notions of Convergence

Before attempting to answer the question of representation of f in terms of its Fourier coefficients, we must formalize what it means for two functions defined over a domain A to be "close". Three commonly studied notions of distance between functions (and hence of convergence of functions) are as follows.

L_∞ **Distance:** $||f - g||_\infty = \sup_{x \in A} |f(x) - g(x)|$. Recall that convergence in the sense of L_∞ is called *uniform convergence.*

L_1 **Distance:** $||f-g||_1 = \int_A |f(x)-g(x)|\,dx$

L_2 **Distance:** $||f-g||_2 = \sqrt{\int_A |f(x)-g(x)|^2\,dx}$

In Fourier Analysis, all three measures of proximity are used at different times and in different contexts.

1.4.2 Fourier Expansion and Fejer's Theorem

The first correct proof (under appropriate assumptions) of the validity of Equation 1.2 was given by Dirichlet:

Theorem 1.2 (Dirichlet). *Let $f : \mathbb{T} \to \mathbb{C}$ be a continuous function whose first derivative is continuous with the possible exception of finitely many points. Then Equation 1.2 holds for every $t \in \mathbb{T}$ at which f is continuous.*

Even before Dirichlet proved this theorem, DuBois Reymond gave an example of a continuous f for which $\limsup_{n\to\infty} S_n(f,0) = \infty$. This ruled out the possibility that continuity is sufficient for Equation 1.2 to hold. The difficulty in answering the question affirmatively came in proving convergence of $S_n(f,t)$ as $n \to \infty$. Fejer answered a more relaxed version of the problem, namely, when can f be *reconstructed* from $\hat{f}(r)$ in possibly other ways? He showed that if f satisfies certain conditions even weaker than continuity, then it can be reconstructed from $\hat{f}(r)$ by taking averages.

Definition 1.3 (Cesaro Means). Let $a_1, a_2, \ldots$ be a sequence of real numbers. Their k^{th} *Cesaro mean* is $b_k = (1/k)\sum_{j=1}^{k} a_j$.

Proposition 1.3. *Let $a_1, a_2, \ldots$ be a sequence of real numbers that converges to a. Then the sequence $b_1, b_2, \ldots$ of its Cesaro means converges to a as well. Moreover, the sequence $\{b_i\}$ can converge even when the sequence $\{a_i\}$ does not (e.g. $a_{2j} = 1, a_{2j+1} = 0$).*

Let us apply the idea of Cesaro means to S_n. Define

$$\begin{aligned}
\sigma_n(f,t) &= \frac{1}{n+1}\sum_{k=0}^{n} S_k(f,t) \\
&= \frac{1}{n+1}\sum_{k=0}^{n}\sum_{r=-k}^{k} \hat{f}(r)\, e^{irt} \\
&= \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1}\,\hat{f}(r)\, e^{irt}
\end{aligned}$$

Theorem 1.4 (Fejer). *Let $f : \mathbb{T} \to \mathbb{C}$ be Riemann integrable. If f is continuous at $t \in \mathbb{T}$, then $\lim_{n\to\infty} \sigma_n(f,t) = f(t)$. Further, if f is continuous then the above holds uniformly.*

Proof. Note that $\lim_{n\to\infty} \sigma_n(f,t) = f(t)$ means that $\forall \epsilon > 0, \exists n_0 : n > n_0 \Rightarrow |\sigma_n(f,t) - f(t)| < \epsilon$. The convergence is uniform if the same n_0 works for all t simultaneously. The proof of the Theorem uses Fejer's

kernels K_n that behave as continuous approximations to the Dirac delta function.

$$\begin{aligned}
\sigma_n(f,t) &= \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1}\,\hat{f}(r)\,e^{irt} \\
&= \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1}\left(\frac{1}{2\pi}\int_{\mathbb{T}} f(x)\,e^{-irx}\,dx\right) e^{irt} \\
&= \frac{1}{2\pi}\int_{\mathbb{T}} f(x) \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1}\,e^{ir(t-x)}\,dx \\
&= \frac{1}{2\pi}\int_{\mathbb{T}} f(x)K_n(t-x)\,dx \quad \text{for } K_n(z) \stackrel{\text{def}}{=} \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1}\,e^{irz} \\
&= \frac{1}{2\pi}\int_{\mathbb{T}} f(t-y)K_n(y)\,dy \quad \text{for } y = t-x
\end{aligned}$$

which is the convolution of f with kernel K_n. Note that if K_n were the Dirac delta function, then $\int_{\mathbb{T}} f(t-y)K_n(y)\,dy$ would evaluate exactly to $f(t)$. Fejer's kernels K_n approximate this behavior.

Proposition 1.5. *K_n satisfies the following:*

$$K_n(s) = \begin{cases} \frac{1}{n+1}\left(\frac{\sin\frac{n+1}{2}s}{\sin\frac{n}{2}}\right)^2 & \text{if } s \neq 0 \\ n+1 & \text{if } s = 0 \end{cases}$$

The kernels K_n have three useful properties.

1. $\forall u : K_n(u) \geq 0$
2. $\forall \delta > 0 : K_n(s) \to 0$ uniformly outside the interval $[-\delta, \delta]$, i.e. $\forall \epsilon > 0, \exists n_0 : s \notin [-\delta,\delta] \Rightarrow |K_n(s)| < \epsilon$
3. $(1/2\pi)\int_{\mathbb{T}} K_n(s)\,ds = 1$

Given any $\epsilon > 0$, we seek a large enough n_0 such that for all $n > n_0$, $|\int_{\mathbb{T}} f(t-y)K_n(y)\,dy - f(t)| < \epsilon$. Divide this integral into two intervals:

$$\int_{\mathbb{T}} f(t-y)K_n(y)\,dy = \int_{-\delta}^{\delta} f(t-y)K_n(y)\,dy + \int_{\mathbb{T}\setminus[-\delta,\delta]} f(t-y)K_n(y)\,dy$$

The first integral on the RHS converges to $2\pi f(t)$ because $f(t-y)$ is almost constant and equals $f(t)$ in the range $y \in [-\delta,\delta]$ and $\int_{\mathbb{T}\setminus[-\delta,\delta]} K_n(s)\,ds$ converges to 2π because of property 3 of K_n. The second integral converges to 0 because f is bounded and because of property 2 of K_n. Hence the LHS converges to $2\pi f(t)$, finishing the proof. □

Corollary 1.6. *If $f, g : \mathbb{T} \to \mathbb{C}$ are continuous functions and $\forall r \in \mathbb{Z} : \hat{f}(r) = \hat{g}(r)$, then $f = g$.*

Proof. Let $h \stackrel{\text{def}}{=} f - g$. h is also continuous. $\forall r : \hat{h}(r) = \hat{f}(r) - \hat{g}(r) = 0$. By Fejer's Theorem, $h \equiv 0$. □

1.4.3 Connection with Weierstrass' Theorem

Because of the uniform convergence part of Fejer's Theorem, we have proved that for all $f : \mathbb{T} \to \mathbb{C}$ continuous and all $\epsilon > 0$, there exists a trigonometric polynomial P such that for all $t \in \mathbb{T}$, $|f(t) - P(t)| < \epsilon$. This implies **Weierstrass' Theorem** which states that "under $l_\infty[a, b]$ norm, polynomials are dense in $C[a, b]$," i.e., for all $f : [a, b] \to \mathbb{R}$ continuous and all $\epsilon > 0$, there exists a polynomial P such that for all $x \in [a, b]$, $|f(x) - P(x)| < \epsilon$.

Informally, Weierstrass' Theorem says that given any continuous function over a finite inverval and an arbitrarily small envelope around it, we can find a polynomial that fits inside that envelope in that interval. To see why this is implied by Fejer's Theorem, simply convert the given function $f : [a, b] \to \mathbb{C}$ into a symmetric function g over an interval of size $2(b - a)$, identify the end points of the new interval so that it is isomorphic to $\mathbb{T}$, and use Fejer's Theorem to conclude that $\sigma_n(g, .)$ is a trigonometric polynomial close to g (and hence f). To see why Weierstrass' Theorem implies Fejer's Theorem, recall that $\cos rt$ can be expressed as a degree r polynomial in $\cos t$. Use this to express the promised trigonometric polynomial $P(t)$ as a linear combination of $\cos rt$ and $\sin rt$ with $-n \leq r \leq n$.

Remark. Weierstrass' Theorem can alternatively be proved using Bernstein's polynomials even though normal interpolation polynomials do not work well for this purpose. Consider $f : [0, 1] \to \mathbb{R}$. The n^{th} Bernstein polynomial is $B_n(f, x) \stackrel{\text{def}}{=} \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k}x^k(1 - x)^{n-k}$. The key idea is that this involves the fact that the Binomial distribution $P(k) = \binom{n}{k}x^k(1 - x)^{n-k}$ is highly concentrated around $k = xn$ and thus approximates the behavior of the Dirac delta function.

Lecture 2

Introduction to Some Convergence theorems

Friday 14, 2005
Lecturer: Nati Linial
Notes: Mukund Narasimhan and Chris Ré

2.1 Recap

Recall that for $f : \mathbb{T} \to \mathbb{C}$, we had defined

$$\hat{f}(r) = \frac{1}{2\pi} \int_{\mathbb{T}} f(t) e^{-irt}\, dt$$

and we were trying to *reconstruct* f from $\hat{f}$. The classical theory tries to determine if/when the following is true (for an appropriate definition of equality).

$$f(t) \stackrel{??}{=} \sum_{r \in \mathbb{Z}} \hat{f}(r) e^{irt}$$

In the last lecture, we proved Fejér's theorem $f * k_n \to f$ where the $*$ denotes convolution and k_n (Fejér kernels) are trignometric polynomials that satisfy

1. $k_n \geq 0$
2. $\int_{\mathbb{T}} k_n = 1$
3. $k_n(s) \to 0$ uniformly as $n \to \infty$ outside $[-\delta, \delta]$ for any $\delta > 0$.

If X is a finite abelian group, then the space of all functions $f : X \to \mathbb{C}$ forms an algebra with the operations $(+, *)$ where $+$ is the usual pointwise sum and $*$ is convolution. If instead of a finite abelian group, we take X to be $\mathbb{T}$ then there is no unit in this algebra (i.e., no element h with the property that $h * f = f$ for all f). However the k_n behave as *approximate units* and play an important role in this theory. If we let

$$S_n(f, t) = \sum_{r=-n}^{n} \hat{f}(r) e^{irt}$$

Then $S_n(f, t) = f * D_n$, where D_n is the Dirichlet kernel that is given by

$$D_n(x) = \frac{\sin\left(n + \frac{1}{2}\right) s}{\sin \frac{s}{2}}$$

The Dirichlet kernel does not have all the nice properties of the the Fejér kernel. In particular,

1. D_n changes sign.

2. D_n does not converge uniformly to 0 outside arbitrarily small $[-\delta, \delta]$ intervals.

Remark. The choice of an appropriate kernel can simplify applications and proofs tremendously.

2.2 The Classical Theory

Let G be a locally compact abelian group.

Definition 2.1. A character on G is a *homomorphism* $\chi : G \to \mathbb{T}$. Namely a mapping satisfyin $\chi(g_1+g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$.

If χ_1, χ_2 are any two characters of G, then it is easily verified that $\chi_1\chi_2$ is also a character of G, and so the set of characters of G forms a commutative group under multiplication. An important role is played by $\hat{G}$, the group of all continuous characters. For example, $\hat{\mathbb{T}} = \mathbb{Z}$ and $\hat{\mathbb{R}} = \mathbb{R}$.

For any function $f : G \to \mathbb{C}$, associate with it a function $\hat{f} : \hat{G} \to \mathbb{C}$ where $\hat{f}(\chi) = \langle f, \chi \rangle$. For example, if $G = \mathbb{T}$ then $\chi_r(t) = e^{irt}$ for $r \in \mathbb{Z}$. Then we have $\hat{f}(\chi_r) = \hat{f}(r)$. We call $\hat{f} : \hat{G} \to \mathbb{C}$ the Fourier transform of f. Now $\hat{G}$ is also a locally compact abelian group and we can play the same game backwards to construct $\hat{\hat{f}}$. Pontryagin's theorem asserts that $\hat{\hat{G}} = G$ and so we can ask the question: Does $\hat{\hat{f}} = f$? While in theory Fejér answered the question of when $\hat{f}$ uniquely determines f, this question is still left unanswered.

For the general theory, we will also require a normalized nonnegative measure μ on G that is translation invariant: $\mu(S) = \mu(a + S) = \mu(\{a + s \,|\, s \in S\})$ for every $S \subseteq G$ and $a \in G$. There exists a unique such measure which is called the Haar measure.

2.3 L_p spaces

Definition 2.2. If (X, Ω, μ) is a measure space, then $L_p(X, \Omega, \mu)$ is the space of all measureable functions $f : X \to \mathbb{R}$ such that

$$\|f\|_p = \left[\int_X |f|^p \cdot d\mu\right]^{\frac{1}{p}} < \infty$$

For example, if $X = \mathbb{N}$, Ω is the set of all finite subsets of X, and μ is the counting measure, then $\|(x_1, x_2, \ldots, x_n, \ldots)\|_p = (\sum |x_i|^p)^{\frac{1}{p}}$. For $p = \infty$, we define

$$\|x\|_\infty = \sup_{i \in \mathbb{N}} |x_i|$$

Symmetrization is a technique that we will find useful. Loosely, the idea is that we are averaging over all the group elements.

Given a function $f : G \to \mathbb{C}$, we symmetrize it by defining $g : G \to \mathbb{C}$ as follows.

$$g(x) = \int_G f(x + a)\, d\mu(a)$$

We will use this concept in the proof of the following result.

Proposition 2.1. *If G is a locally compact abelian group, with a normalized Haar measure μ, and if $\chi_1, \chi_2 \in \hat{G}$ are two distinct characters then $\langle \chi_1, \chi_2 \rangle = 0$. i.e.,*

$$I = \int_X \chi_1(x)\overline{\chi_2(x)}\, d\mu(x) = \delta_{\chi_1,\chi_2} = \begin{cases} 0 & \chi_1 \neq \chi_2 \\ 1 & \chi_1 = \chi_2 \end{cases}$$

Proof. For any fixed $a \in G$, $I = \int_X \chi_1(x)\overline{\chi_2(x)}\, d\mu(x) = \int_X \chi_1(x+a)\overline{\chi_2(x+a)}\, d\mu(x)$. Therefore,

$$\begin{aligned} I &= \int_X \chi_1(x+a)\overline{\chi_2(x+a)}\, d\mu(x) \\ &= \int_X \chi_1(x)\chi_1(a)\overline{\chi_2(x)\chi_2(a)}\, d\mu(x) \\ &= \chi_1(a)\overline{\chi_2(a)} \int_X \chi_1(x)\overline{\chi_2(x)}\, d\mu(x) \\ &= \chi_1(a)\overline{\chi_2(a)} I \end{aligned}$$

This can only be true if either $I = 0$ or $\chi_1(a) = \chi_2(a)$. If $\chi_1 \neq \chi_2$, then there is at least one a such that $\chi_1(a) \neq \chi_2(a)$. It follows that either $\chi_1 = \chi_2$ or $I = 0$. □

By letting χ_2 be the character that is identically 1, we conclude that $\chi \in \hat{G}$ with $\chi \neq 1$ for any $\int_G \chi(x)\, d\mu(x) = 0$.

2.4 Approximation Theory

Weierstrass's theorem states that the polynomials are dense in $L_\infty[a,b] \cap C[a,b]$[1] Fejér's theorem is about approximating functions using trignometric polynomials.

Proposition 2.2. $\cos nx$ *can be expressed as a degree n polynomial in* $\cos x$.

Proof. Use the identity $\cos(u+v) + \cos(u-v) = 2\cos u \cos v$ and induction on n. □

The polynomial $T_n(x)$ where $T_n(\cos x) = \cos(nx)$ is called n^{th} Chebyshev's polynomial. It can be seen that $T_0(s) = 1$, $T_1(s) = s$, $T_2(s) = 2s^2 - 1$ and in general $T_n(s) = 2^{n-1}s^n$ plus some lower order terms.

Theorem 2.3 (Chebyshev). *The normalized degree n polynomial $p(x) = x^n + \ldots$ that approximates the function $f(x) = 0$ (on $[-1,1]$) as well as possible in the $L_\infty[-1,1]$ norm sense is given by $\frac{1}{2^{n-1}}T_n(x)$. i.e.,*

$$\min_{p \text{ a normalized polynomial}} \max_{-1 \leq x \leq 1} |p(x)| = \frac{1}{2^{n-1}}$$

This theorem can be proved using linear programming.

[1] This notation is intended to imply that the norm on this space is the sup-norm (clearly $C[a,b] \subseteq L_\infty[a,b]$)

2.4.1 Moment Problems

Suppose that X is a random variable. The simplest information about X are its moments. These are expressions of the form $\mu_r = \int f(x)x^r\,dx$, where f is the probability distribution function of X. A *moment problem* asks: Suppose I know all (or some of) the moments $\{\mu_r\}_{r\in\mathbb{N}}$. Do I know the distribution of X?

Theorem 2.4 (Hausdorff Moment Theorem). *If $f, g : [a, b] \to \mathbb{C}$ are two continuous functions and if for all $r = 0, 1, 2, \ldots,$ we have*

$$\int_a^b f(x)x^r\,dx = \int_a^b g(x)x^r\,dx$$

then $f = g$. Equivalently, if $h : [a, b] \to \mathbb{C}$ is a continuous function with $\int_a^b h(x)x^r\,dx = 0$ for all $r \in \mathbb{N}$, then $h \equiv 0$.

Proof. By Weierstrass's theorem, we know that for all $\epsilon > 0$, there is a polynomial P such that $\left\|\overline{h} - P\right\|_\infty < \epsilon$. If $\int_a^b h(x)x^r\,dx = 0$ for all $r \in \mathbb{N}$, then it follows that $\int_a^b h(x)Q(x)\,dx = 0$ for every polynomial $Q(x)$, and so in particular, $\int_a^b h(x)P(x)\,dx$. Therefore,

$$0 = \int_a^b h(x)P(x)\,dx = \int_a^b h(x)\overline{h(x)}\,dx + \int_a^b h(x)\left(P(x) - \overline{h(x)}\right)\,dx$$

Therefore,

$$\langle h, \overline{h}\rangle = -\int_a^b h(x)\left(P(x) - \overline{h(x)}\right)\,dx$$

Since h is continuous, it is bounded on $[a, b]$ by some constant c and so on $[a, b]$ we have $\left|h(x)\left(P(x) - \overline{h(x)}\right)\right| \le c \cdot \epsilon \cdot |b - a|$. Therefore, for any $\delta > 0$ we can pick $\epsilon > 0$ so that so that $\|h\|_2^2 \le \delta$. Hence $h \equiv 0$. □

2.4.2 A little Ergodic Theory

Theorem 2.5. *Let $f : \mathbb{T} \to \mathbb{C}$ be continuous and γ be irrational. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{r=1}^{n} f\left(e^{2\pi ir}\right) = \int_{\mathbb{T}} f(t)\,dt$$

Proof. We show that this result holds when $f(t) = e^{ist}$. Using Fejér's theorem, it will follow that the result holds for any continuous function. Now, clearly $\frac{1}{2\pi}\int_{\mathbb{T}} e^{ist}\,dt = 0$. Therefore,

$$\begin{aligned}\left|\frac{1}{n}\sum_{r=1}^{n} e^{2\pi irs\gamma} - \frac{1}{2\pi}\int_{\mathbb{T}} e^{ist}\,dt\right| &= \left|\frac{1}{n}\sum_{r=1}^{n} e^{2\pi irs\gamma}\right| \\ &= \left|\frac{1}{n}e^{2\pi is\gamma}\right|\left|\frac{1 - e^{2\pi ins\gamma}}{1 - e^{2\pi is\gamma}}\right| \\ &\le \frac{2}{n\cdot(1 - e^{2\pi is\gamma})}\end{aligned}$$

Since γ is irrational, $1 - e^{2\pi is\gamma}$ is bounded away from 0. Therefore, this quantity goes to zero, and hence the result follows. □

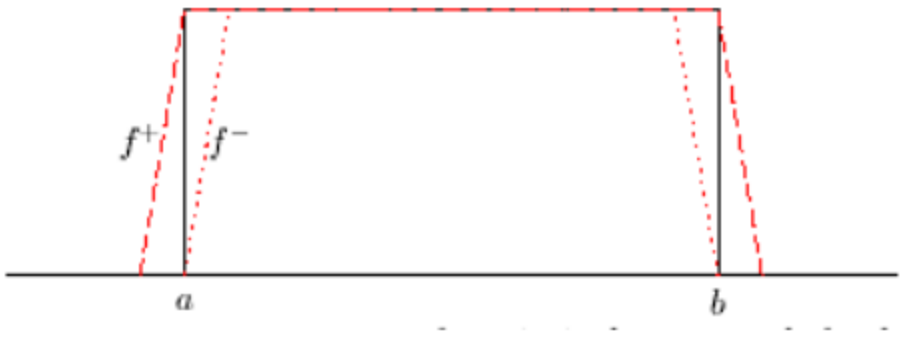


Figure 2.1: Probability of Property v. p

This result has applications in the evaluations of integrals, volume of convex bodies. Is is also used in the proof of the following result.

Theorem 2.6 (Weyl). *Let γ be an irrational number. For $x \in \mathbb{R}$, we denote by $\langle x \rangle = x - [x]$ the fractional part of x. For any $0 < a < b < 1$, we have*

$$\lim_{n\to\infty} \frac{|\{1 \le r \le n : a \le \langle r\gamma \rangle < b\}|}{n} = b - a$$

Proof. We would like to use Theorem 2.5 with the function $f = 1_{[a,b]}$. However, this function is not continuous. To get around this, we define functions $f^+ \ge 1_{[a,b]} \ge f^-$ as shown in the following diagram.

f^+ and f^- are continuous functions approximating f. We let let them approach f and pass to the limit. □

This is related to a more general ergodic theorem by Birkhoff.

Theorem 2.7 (Birkhoff, 1931). *Let $(\Omega, \mathcal{F}, p)$ be a probability measure and $T : \Omega \to \Omega$ be a measure preserving transformation. Let $X \in L_1(\Omega, \mathcal{F}, p)$ be a random variable. Then*

$$\frac{1}{n}\sum_{k=1}^{n} X \circ T^k \to E[X; \mathcal{I}]$$

Where $\mathcal{I}$ is the σ-field of T-invariant sets.

2.5 Some Convergence Theorems

We seek conditions under which $S_n(f,t) \to f(t)$ (preferably uniformly). Some history:

- DuBois Raymond gave an example of a continuous function such that $\limsup S_n(f,0) = \infty$.
- Kolmogorov [1] found a Lebesgue measureable function $f : \mathbb{T} \to \mathbb{R}$ such that for all t, $\limsup S_n(f,t) = \infty$.

- Carleson [2] showed that if $f : \mathbb{T} \to \mathbb{C}$ is a continuous function (even Riemann integrable), then $S_n(f,t) \to f(t)$ almost everywhere.
- Kahane and Katznelson [3] showed that for every $E \subseteq \mathbb{T}$ with $\mu(E) = 0$, there exists a continuous function $f : \mathbb{T} \to \mathbb{C}$ such that $S_n(f,t) \not\to f(t)$ if and only if $t \in E$.

Definition 2.3. $\ell_p = L_p(\mathbb{N}, \text{Finite sets}, \text{ counting measure}). = \{\boldsymbol{x} | (x_0, \dots)|^p < \infty\}$.

Theorem 2.8. *Let $f : \mathbb{T} \to \mathbb{C}$ be continuous and suppose that $\sum_{r\in\mathbb{Z}} |\hat{f}(r)| < \infty$ (so $\hat{f} \in \ell_1$). Then $S_n(f,t) \to f$ uniformly on $\mathbb{T}$.*

Proof. See lecture 3, theorem 3.1. □

2.6 The L_2 theory

The fact that $e(t) = e^{ist}$ is an orthonormal family of functions allows to develop a very satisfactory theory. Given a function f, the best coefficients $\lambda_1, \lambda_2, \dots, \lambda_n$ so that $\|f - \sum_{i=1}^n \lambda_j e_j\|_2$ is minimized is given by $\lambda_j = \langle f, e_j \rangle$. This answer applies just as well in any inner product normed space (Hilbert space) whenever $\{e_j\}$ forms an orthonormal system.

Theorem 2.9 (Bessel's Inequality). *For every $\lambda_1, \lambda_2, \dots, \lambda_n$,*

$$\left\| f - \sum_{i=1}^n \lambda_i e_i \right\|^2 \geq \|f\|^2 - \sum_{i=1}^n \langle f, e_i \rangle^2$$

with equality when $\lambda_i = \langle f, e_i \rangle$

Proof. We offer a proof here for the real case, in the next lecture the complex case will be done as well.

$$\begin{aligned}\left\| f - \sum_{i=1}^n \lambda_i e_i \right\|^2 &= \left\| (f - \sum_{i=1}^n \langle f, e_i \rangle e_i) + (\sum_{i=1}^n \langle f, e_i \rangle e_i - \sum_{i=1}^n \lambda_i e_i) \right\|^2 \\ &= \left\| (f - \sum_{i=1}^n \langle f, e_i \rangle e_i) \right\|^2 + \left\| (\sum_{i=1}^n \langle f, e_i \rangle e_i - \sum_{i=1}^n \lambda_i e_i) \right\|^2 + \text{ cross terms}\end{aligned}$$

$$\text{cross terms } = 2\langle f - \sum_{i=1}^n \langle f, e_i \rangle e_i, \sum_{i=1}^n \langle f, e_i \rangle e_i - \sum_{i=1}^n \lambda_i e_i \rangle$$

Observe that the terms in the cross terms are orthogonal to one another since $\forall i \langle f - \langle f, e_i \rangle e_i, e_i \rangle = 0$. We write

$$2 \sum \langle f, e_i \rangle \langle f - \sum_{j=1}^n \langle f, e_j \rangle e_j, e_i \rangle - \sum_{i}^n \lambda_i \langle f - \sum_{j=1}^n \langle f, e_j \rangle e_i, e_i \rangle$$

Observe that each innter product term is 0. Since if $i = j$, then we apply $\forall i \langle f - \langle f, e_i \rangle e_i, e_i \rangle = 0$. If $i \neq j$, then they are orthogonal basis vectors.

We want to make this as small as possible and have only control over the λ_is. Since this term is squared and therefore non-negative, the sum is minimized when we set $\forall i\ \lambda_i = \langle f, e_i \rangle$. With this choice,

$$\begin{aligned}\left\| f - \sum_{i=1}^{n} \lambda_i e_i \right\|^2 &= \langle f - \sum_{i=1}^{n} \lambda_i e_i, f - \sum_{i=1}^{n} \lambda_i e_i \rangle \\ &= \langle f, f \rangle - 2 \sum_{i=1}^{n} \lambda_i \langle f, e_i \rangle + \sum_{i=1}^{n} \lambda_i^2 \\ &= \|f\|^2 - \sum_{i=1}^{n} \langle f, e_i \rangle^2\end{aligned}$$

where the last inequality is obtained by setting $\lambda_i = \langle f, e_i \rangle$. □

References

[1] A. N. Kolmogorov, *Une série de Fourier-Lebesgue divergente partout*, CRAS Paris, 183, pp. 1327-1328, 1926.

[2] L. Carleson, *Convergence and growth of partial sums of Fourier series*, Acta Math. 116, pp. 135-157, 1964.

[3] J-P Kahane and Y. Katznelson, *Sur les ensembles de divergence des séries trignométriques*, Studia Mathematica, 26 pp. 305-306, 1966

Lecture 3

Harmonic Analysis on the Cube and Parseval's Identity

Jan 28, 2005
Lecturer: Nati Linial
Notes: Pete Couperus and Neva Cherniavsky

3.1 Where we can use this

During the past weeks, we developed the general machinery which we will apply to problems in discrete math and computer science in the following weeks. In the general setting, we can ask how much information can we determine about a function f given its Fourier coefficients $\hat{f}$. Or, given f what can we say about $\hat{f}$? There is some distinction between properties which will hold in the general setting, and those that make sense for the specific spaces we have dealt with. So far, we have looked at

1. $\mathbb{T}$ (the unit circle/Fourier Series).
2. $\mathbb{Z}/n\mathbb{Z}$ (Discrete Fourier Transform).
3. $\mathbb{R}$ (Real Fourier Transform).
4. $\{0,1\}^n = \mathbf{GF}(2)^n = (\mathbb{Z}/2\mathbb{Z})^n$ (the n-cube).

For the n-cube (or for any space we wish to do Harmonic Analysis on), we need to determine the characters. We can view elements of $\{0,1\}^n$ as subsets of $[n] = \{1, ..., n\}$, and then to each subset $S \subseteq [n]$, let $\chi_S(T) = (-1)^{|S \cap T|}$. Then:

$$\langle \chi_{S_1}, \chi_{S_2} \rangle = \frac{1}{2^n} \sum_{T \subseteq [n]} (-1)^{|S_1 \cap T| + |S_2 \cap T|}$$

To see that the χ_S form an orthonormal basis, suppose that $x \in S_1 - S_2$. Then, the function

$$\phi(A) = \begin{cases} A - \{x\} & x \in A \\ A \cup \{x\} & x \notin A \end{cases}$$

gives a bijection between $\{A : |S_1 \cap A| \equiv |S_2 \cap A| \pmod 2\}$ and $\{A : |S_1 \cap A| \not\equiv |S_2 \cap A| \pmod 2\}$. So, $\langle \chi_{S_1}, \chi_{S_2} \rangle = 0$ for $S_1 \neq S_2$. If $S_1 = S_2$, then $|T \cap S_1| + |T \cap S_2|$ is always even, so $\langle \chi_S, \chi_S \rangle = 1$.

Hence, the χ_S form an orthonormal basis for functions from $\{0,1\}^n \to \mathbb{R}$. (This is, of course, true in general, but it's useful to see this explicitly for this special case). Then for any $f : \{0,1\}^n \to \mathbb{R}$, we can write $f = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S$, where

$$\hat{f}(S) = \langle f, \chi_S \rangle = \frac{1}{2^n} \sum_{T \subseteq [n]} f(T) \cdot (-1)^{|S \cap T|}.$$

There is an equivalent and often useful way of viewing this. We can also view the n-cube as $\{-1,1\}^n$ with coordinate-wise multiplication. In this case, any function $f : \{-1,1\}^n \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial:

$$f = \sum_{S \subset \{0,1\}^n} a_S \prod_{i \in S} x_i$$

where $\prod_{i \in S} x_i$ corresponds to χ_S.

There is an advantage to the fact that we now deal with a finite group. Note that $f = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S$ is always the case for functions over the n-cube, unlike working over $\mathbb{T}$. Working over $\mathbb{T}$, we made some assumptions on f to be able have a similar formula to recover f from its fourier coefficients.

Now we can ask, what can be said about $\hat{f}$ when f is boolean (when the range of f is $\{0,1\}$)? More specifically, how do the properties of f get reflected in $\hat{f}$? In general, this is too hard a question to tackle. But what sorts of relationships between properties are we looking for? In the case of $\mathbb{T}$, the smoothness of f roughly corresponds to its fourier coefficients $\hat{f}(r)$ decaying rapidly as $r \to \infty$. E.g.

$$\begin{array}{rcl} f : \mathbb{T} \to \mathbb{C} & \leftrightarrow & \{f(r) | r \in \mathbb{Z}\} \\ \text{smooth} & \leftrightarrow & \hat{f}(r) \text{ decays rapidly} \end{array}$$

An instance of this relationship can be seen from the following theorems.

Theorem 3.1. *Let $f : \mathbb{T} \to \mathbb{C}$ be continuous, and suppose that $\sum_{r=-\infty}^{\infty} |\hat{f}(r)|$ converges. Then $S_n(f) \to f$ uniformly.*

We can derive this theorem from another.

Theorem 3.2. *Suppose that the sequence $\sum_{r=-n}^{n} |a_r|$ converges (as $n \to \infty$). Then $g_n(t) = \sum_{r=-n}^{n} a_r e^{irt}$ converges uniformly as $n \to \infty$ on $\mathbb{T}$ to $g : \mathbb{T} \to \mathbb{C}$, where g is continuous and $\hat{g}(r) = a_r$ for all r.*

This (roughly) says that if we have a sequence that is decreasing rapidly enough (its series converges absolutely), then we can choose these to be the Fourier coefficients for some continuous function.

To see that Theorem 3.2 implies Theorem 3.1, if $\hat{f}(r) = \hat{g}(r) = a_r$ for all r, and both f and g are continuous, then $f = g$. This is based on Fejer's Theorem (or Weierstrauss).

So to prove Theorem 3.1, all that remains is to prove Theorem 3.2.

Proof. The underlying idea for the proof of Theorem 3.2 is that $C(\mathbb{T})$ with the ∞-norm is a **complete metric space**, meaning that all Cauchy sequences converge. Recall, a sequence (a_n) is Cauchy if for $\epsilon > 0$, there is some N so for $n, m \geq N$, we have $d(a_n, a_m) < \epsilon$ (where d is whatever metric we are using). So, to prove the theorem, we only need to check that $\{f_n\} = \{\sum_{-n}^{n} a_r e^{irt}\}$ is a Cauchy sequence with the ∞-norm. Since $s_n := \sum_{r=-n}^{n} |a_r|$ converges, for $\epsilon > 0$, there is some N so that $|s_m - s_n| < \epsilon$ for $n, m \geq N$ (basically, the tail end is small), hence

$$\left|\left| \sum_{m \geq |r| > n} a_r e^{irt} \right|\right| \leq \left| \sum_{m \geq |r| > n} |a_r| \right| < \epsilon.$$

So, the $\{g_n\}$ forms a Cauchy sequence.

Hence,

$$\begin{gathered} \sum_{r=-n}^{n} a_r e^{-irt} \to g \text{ uniformly, so} \\ e^{-ikt} \sum_{r=-n}^{n} a_r e^{-irt} \to e^{-ikt} g(t) \text{ uniformly.} \\ \int_{\mathbb{T}} e^{-ikt} \sum_{r=-n}^{n} a_r e^{-irt} dt \to \int_{\mathbb{T}} e^{-ikt} g(t). \end{gathered} \tag{3.1}$$

□

Recall, du Bois Raymond gives an example of $f : \mathbb{T} \to \mathbb{C}$ such that $\overline{\lim} |S_n(f, 0)| = +\infty$. However, if the first derivative is somewhat controlled, we can say more.

Theorem 3.3. *Let $f : \mathbb{T} \to \mathbb{C}$ be continuous and suppose that f' is defined for all but a finite subset of $\mathbb{T}$. Then $S_n(f) \to f$ uniformly.*

f smooth $\leftrightarrow$ $\hat{f}$ decays rapidly $\Rightarrow$ "$S_n f \to f$".

Recall from basic analysis, if f_n are continuously differentiable and if $f_n \to f$ uniformly and $f_n' \to g$ uniformly then $f' = g$ and g is continuous. This will allow us to show that the Fourier Series of f' is attained by termwise derivatives of the Fourier Series of f.

Theorem 3.4. *Let $f : \mathbb{T} \to \mathbb{C}$ be continuous and suppose that $\sum_{r=-\infty}^{\infty} r|\hat{f}(r)|$ converges. Then f is continuously differentiable and $\sum_{r=-n}^{n} ir\hat{f}(r)e^{irt} \to f'$ uniformly.*

Proof. We would like to show that we can apply this when $f_n = S_n f$. But if $\sum_{r=-\infty}^{\infty} r|\hat{f}(r)|$ converges, then $\sum_{r=-\infty}^{\infty} |\hat{f}(r)|$ converges (since the each term is smaller). So we have

$$|\hat{f}(r)| \leq |r\hat{f}(r)| \Rightarrow \sum_{-n}^{n} |\hat{f}(r)| \text{ converges}$$

So by Theorem 3.1, $f_n = S_n f \to f$ uniformly. By the same theorem, $f_n' = \sum_{r=-n}^{n} ir\hat{f}(r)e^{irt} \to g$ is continuous. By the statement above due to basic analysis, we know that this implies that f is continuously differentiable. □

A similar argument will provide a stronger connection between the idea that $\hat{f}(r)$ are rapidly decreasing implies that f is "smoother".

Proposition 3.5. *Let $f : \mathbb{T} \to \mathbb{C}$ satisfy $f^{(n-1)}$ is continuously differentiable except possibly finitely many points X, and $|f^{(n)}(x)| \leq M$ for $x \notin X$. Then $\forall r \neq 0 |\hat{f}(r)| \leq Mr^{-n}$.*

Proof. (Integration by parts).

$\hat{f}(r) = \frac{1}{2\pi}\int_{\mathbb{T}} f(t)e^{-irt}dt$. Let $u = f(t), dv = e^{-irt}dt$. Then $du = f'(t)dt, v = \frac{e^{-irt}}{-ir}$.

$$\begin{aligned} \hat{f}(r) = \frac{1}{2\pi}\int_{\mathbb{T}} f(t)e^{-irt}dt = \\ \frac{1}{2\pi}\left[f(t)\frac{e^{-irt}}{-ir}|_{-\pi}^{\pi} - \int_{\mathbb{T}} f'(t)\frac{e^{-irt}}{-ir}dt\right] = \\ \frac{1}{2\pi}\left[0 - \int_{\mathbb{T}} f'(t)\frac{e^{-irt}}{-ir}dt\right] = \cdots = \text{ (first term is 0 since } f \text{ is periodic)} \\ \frac{1}{2\pi(-ir)^n}\int_{\mathbb{T}} f^{(n)}(t)e^{-irt}dt. \end{aligned} \tag{3.2}$$

So

$$\left|\hat{f}(r)\right| \leq \left|\frac{(-ir)^{-n}}{2\pi}\right| \int_{-\pi}^{\pi} |f^{(n)}(t)e^{-irt}|dt = O(\frac{1}{r^n})$$

□

Corollary 3.6. *If $f : \mathbb{T} \to \mathbb{C}$ is in C^2 (twice continuously differentiable), then $S_n f \to f$ uniformly.*

Proof.

$$\hat{f}(r) = O\left(\frac{1}{r^2}\right) \Rightarrow \sum_{r=-\infty}^{\infty} \left|\hat{f}(r)\right| \text{ converges.}$$

So, $S_n f \to f$ uniformly. □

3.2 Rate of Convergence

Until now, we haven't really addressed the rate of convergence, meaning when $S_n(f)$ does converge to f, how fast does it converge to f? Examine $g(x) = \pi - |x|$ for $x \in [-\pi, \pi]$, and extend g periodically to $h(x)$. Direction calculation gives $|S_n(h, 0) - \pi| > \frac{1}{n+2}$. By using L_2 theory, it can be further shown that every trigonometric polynomial P of degree n has the property $\|P - h\|_\infty > \Omega(n^{-3/2})$. Kolmogorov showed the following.

Theorem 3.7. *(Kolmogorov) For all $A > 0$, there is a trigonometric polynomial f such that:*

1. $f \geq 0$.
2. $\frac{1}{2\pi} \int_{\mathbb{T}} f(t)dt \leq 1$
3. *For every $x \in \mathbb{T}$,* $\sup_n |S_n(f, x)| \geq A$.

Hence, there is a Lebesgue integrable function f such that for all $x \in \mathbb{T}$, $\overline{\lim}|S_n(f, x)| = +\infty$.

3.2.1 Convergence Results

In 1964, Carleson proved the following.

Theorem 3.8. *(Carleson) If f is continuous (or only Riemann integrable), then $S_n f \to f$ almost everywhere.*

Later, Kahane and Katznelson proved that this result is tight.

Theorem 3.9. *For all $E \subseteq \mathbb{T}$ with $\mu(E) = 0$, there is a continuous f such that $S_n f \to f$ exactly on $\mathbb{T} - E$.*

Notice that these results make somewhat weak assumptions on f. We will now work on seeing how things improve in the situation where f is an L_2 function.

3.3 L_2 theory for Fourier Series

Recall part of original question was "how are f and $\hat{f}$ related"? Our immediate goal will be to show that in the L_2 case, their norms are identical, which is the Parseval identity. Recall, $\|f\|_2 = \sqrt{\int_{\mathbb{T}} |f(t)|^2 dt}$. Then the Parseval identity states $\|f\|_2 = \|\hat{f}\|_2$. For the Discrete Fourier Transform, this essentially means that the transform matrix is an orthonormal matrix.

We will procede by focusing on Hilbert Spaces. A Hilbert Space $\mathcal{H}$ is a normed ($\mathbb{C}$-linear) space with an inner product $\langle \cdot, \cdot \rangle$ satisfying the following axioms.

1. $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.
2. $\langle x, y \rangle = \overline{\langle y, x \rangle}$.
3. $\langle x, x \rangle = \|x\|^2 \geq 0$ with equality $\iff x = 0$.

There are a number of facts that we know about familiar Hilbert spaces (like $\mathbb{R}^n$) that hold for general Hilbert spaces as well.

Theorem 3.10. *If $\mathcal{H}$ is a Hilbert space, then the Cauchy-Schwarz Inequality holds, namely if $f, g \in \mathcal{H}$, then $\|f\| \cdot \|g\| \geq \langle f, g \rangle$.*

Proof. We will show the proof for real Hilbert spaces.

$$0 \leq \langle f - \lambda g, f - \lambda g \rangle = \|f\|^2 - 2\lambda \langle f, g \rangle + \lambda^2 \|g\|^2. \tag{3.3}$$

Viewing this as a degree 2 polynomial in λ, it is non-negative, so has at most one real root. Hence, the discriminant $(-2\langle f, g \rangle)^2 - 4\|f\|^2\|g\|^2 \leq 0$. Hence, $\langle f, g \rangle^2 \leq \|f\|^2\|g\|^2$. □

One may ask, if we have an element $f \in \mathcal{H}$, how can we best approximate f with respect to some basis? Specifically, let $e_{-n}, ..., e_0, e_1, ..., e_n$ be an orthonormal system in $\mathcal{H}$ (meaning, $\langle e_i, e_j \rangle = \delta_{i,j}$). Given $f \in \mathcal{H}$, the question is to find $\lambda_i \in \mathbb{C}$ such that $\|f - \sum_i \lambda_i e_i\|$ is minimized.

Theorem 3.11. *Let $\mathcal{H}$, $\{e_i\}$, f be as above. Set $g = \sum_{j=-n}^{n} \lambda_j e_j$, and $g_0 = \sum_{j=-n}^{n} \langle f, e_j \rangle e_j$. Then*

$$\|f\|_2^2 \geq \sum_{j=-n}^{n} \langle f, e_j \rangle^2, \|f - g\|_2 \geq \|f - g_0\|_2 = \sqrt{\|f\|_2^2 - \sum_{j=-n}^{n} \langle f, e_j \rangle^2} \tag{3.4}$$

with equality iff $\lambda_j = \langle f, e_j \rangle$ for all j.

Proof.

$$\begin{aligned} \|f - g\|_2^2 = \langle f - g, f - g \rangle = \langle f - \sum_j \lambda_j e_j, f - \sum_j \lambda_j e_j \rangle = \\ \|f\|_2^2 - (\sum_j \lambda_j \overline{\langle f, e_j \rangle} + \overline{\lambda_j} \langle f, e_j \rangle) + \sum_j |\lambda_j|^2 = \\ \langle f, f \rangle + \sum_j |\lambda_j - \langle f, e_j \rangle|^2 - \sum_j |\langle f, e_j \rangle|^2 \geq \\ \langle f, f \rangle - \sum_j |\langle f, e_j \rangle|^2 = \|f - g_0\|^2. \end{aligned} \tag{3.5}$$

Note that equality in the last step occurs exactly when $\lambda_j = \langle f, e_j \rangle$ for all j. □

Corollary 3.12. *(Approximation and Bessel's Inequality).*

1. *$S_n f$ is the closest (in the L_2 sense) degree n trigonometric polynomial approximation to f.*
2. *(Bessel's Inequality). If $f \in L_2(\mathbb{T})$, then*

$$\|f\|^2 = \frac{1}{2\pi} \int_{\mathbb{T}} |f(t)|^2 dt \geq \sum_{r=-n}^{n} |\hat{f}(r)|^2,$$

and $\|f\|^2 \geq \sum_{r=-\infty}^{\infty} |\hat{f}(r)|^2$.

This shows one side of the Parseval Identity, namely $\|f\|^2 \geq \|\hat{f}\|^2$.
Recall by Theorem 3.1, if f continuous and $\hat{f} \in l_1$ (meaning $\sum_r |\hat{f}(r)|$ converges), then $S_n f \to f$ uniformly. We will show that f having continuous first derivative in fact implies that $\hat{f}$ is in l_1.

Corollary 3.13. *If $f \in C^1$, then $S_n f \to f$ uniformly.*

Proof.

$$\begin{aligned}
\sum_{r=-n}^{n} |\hat{f}(r)| &= |\hat{f}(0)| + \sum_{1\leq|r|\leq n} |r\hat{f}(r)| \cdot \frac{1}{|r|} \\
\text{(by Cauchy-Schwartz)} \quad &\leq |\hat{f}(0)| + \sqrt{2\sum_{1\leq r\leq n} \frac{1}{r^2}} \cdot \sqrt{\sum_{1\leq|r|\leq n} |\hat{f}'(r)|^2} \\
\text{(by identity } \sum_{1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}) \quad &\leq |\hat{f}(0)| + \sqrt{\frac{\pi^2}{3} \cdot \frac{1}{2\pi} \int_{\mathbb{T}} |f'(t)|^2 dt}
\end{aligned}$$

which is bounded since the first derivative is bounded (continuous on $\mathbb{T}$). □

3.3.1 Parseval's Identity

We are now ready to complete the proof of the Parseval Identity.

Theorem 3.14. *If $f : \mathbb{T} \to \mathbb{C}$ is continuous, then $\|f - S_n f\| \to 0$.*

Proof. By Weierstrass (or Fejer) approximation, for any $\epsilon > 0$, there is some trigonometric polynomial P such that $\|f - P\|_\infty < \epsilon$. So,

$$\|f - S_n f\|_2 \leq \|f - P\|_2 + \|S_n P - S_n f\|_2 \leq \|f - P\|_\infty + \|S_n(P - f)\|_2$$

We use the fact that $S_n P = f$ for every trigonometric polynomial of degree $\leq n$. Then Bessel's inequality tells us $\|S_n(P - f)\|_2 \leq \|P - f\|_2$. Since $\|P - f\|_2 \leq \|P - f\|_\infty < \epsilon$, we see that $\|f - S_n f\|_2 < 2\epsilon$. This completes the proof. □

Hence, it is easy to see $\|f\|_2 = \|\hat{f}\|_2$, and we have the Parseval Identity.

Corollary 3.15. *(Parseval) If $f : \mathbb{T} \to \mathbb{C}$ is continuous, then $\frac{1}{2\pi}\int_{\mathbb{T}} |f(t)|^2 dt = \|f\|_2^2 = \sum_{r=-\infty}^{\infty} |\hat{f}(r)|^2$.*

Proof. Since $\|f - S_n f\|_2^2 = \|f\|_2^2 - \sum_{r=-n}^{n} |\hat{f}(r)|^2$ goes to 0 as $n \to \infty$, we conclude that $\sum_{r=-n}^{n} |\hat{f}(r)|^2 \to \|f\|_2^2$ as $n \to \infty$. □

In other words, $f \to \hat{f}$ is an isometry in L_2.

3.4 Geometric Proof of the Isoperimetric Inequality

We will complete this next time. Here, we will present Steiner's idea to resolve the following question. What is the largest area of a planar region with fixed circumference L? We will present Steiner's idea here. Suppose that C is a curve such that the area enclosed is optimal.

C encloses a convex region.

If not, then there are points A, B such that the line segment joining A and B lies outside the region. By

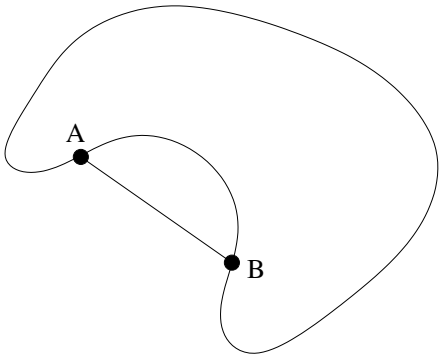


Figure 3.1: Joining A and B yields more area.

replacing the arc from A to B with the line segment from A to B, we increase the area, and decrease the circumference. See Figure 3.1.

C encloses a centrally symmetric region.

If not, pick points A, B such that the arc length from A to B is the same for both directions travelled.

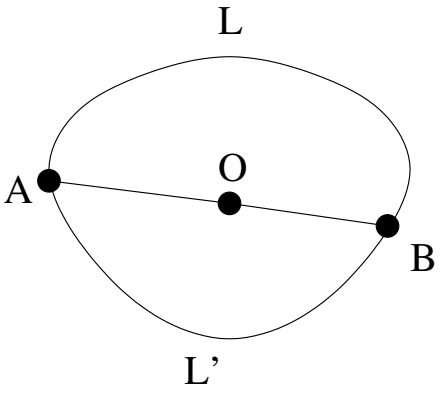


Figure 3.2: Reflecting larger area yields more area.

Suppose that the region enclosed by $\overline{AB} \cup L$ has area at least that of the region enclosed by $\overline{AB} \cup L'$. We can then replace the latter by a mirror copy of the first. This can only increase the total area and yields a region that is centrally symmetric with respect to the middle of the segment $[A, B]$. **C is a circle.**

Recall the following fact from Euclidean geometry: A circle is the locus of all points x such that xA is perpendicular to xB where AB is some segment (which is the diameter of the circle). Therefore, if this is not so, then there is some parallelogram a, b, c, d, with $\overline{ac}$ passing the the center, inscribed in C (since C is centrally symmetric), and with the angle at b not equal to $\frac{\pi}{2}$. Now, "move" sides a, b and c, d to sides a', b'

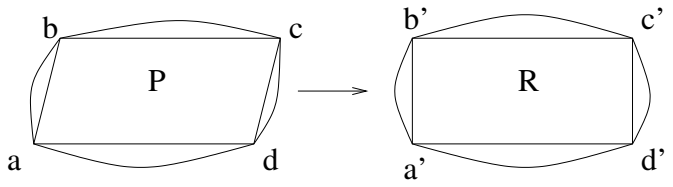


Figure 3.3: Changing Parallelogram to Rectangle Yields more area.

and c', d' such that a', b', c', d' forms a rectangle. See Figure 3.3. We obtain a new curve C' such that the area outside of rectangle $R = [a', b', c', d']$ is the same as the area outside of the parallelogram $P = [a, b, c, d]$. Since the side lengths of R and P are the same, the area enclosed by R must exceed the area enclosed by P, so the area enclosed by C' must exceed the area enclosed by C. Hence, C was not optimal. Hence, our parallelogram P must have angles equal to $\frac{\pi}{2}$.

Although these ideas are pretty and useful, this is still not a proof of the isoperimetric inequality. We do not know that an optimal C exists, only that if it does, it must be a circle.

Lecture 4

Applications of Harmonic Analysis

February 4, 2005
Lecturer: Nati Linial
Notes: Matthew Cary

4.1 Useful Facts

Most of our applications of harmonic analysis to computer science will involve only Parseval's identity.

Theorem 4.1 (Parseval's Identity).

$$\|f\|_2 = \|\hat{f}\|_2$$

Corollary 4.2.

$$\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle.$$

Proof. Note that $\langle f+g, f+g\rangle = \|f+g\|_2 = \|\widehat{f+g}\|_2 = \|\hat{f}+\hat{g}\|_2$. Now as $\langle f+g, f+g\rangle = \|f\|_2^2 + \|g\|_2^2 + 2\langle f, g\rangle$, and similarly $\|\hat{f}+\hat{g}\|_2^2 = \|\hat{f}\|_2^2 + \|\hat{g}\|_2^2 + 2\langle \hat{f}, \hat{g}\rangle$, applying Parseval to $\|f\|_2$ and $\|g\|_2$ and equating finishes the proof. □

The other basic identity is the following.

Lemma 4.3.

$$\widehat{f * g} = \hat{f} \cdot \hat{g}$$

Proof. We will show this for the unit circle $\mathbb{T}$, but one should note that it is true more generally. Recall that by definition $h = f * g$ means that

$$h(t) = \frac{1}{2\pi}\int_{\mathbb{T}} f(s)g(t-s)ds.$$

Now to calculate $\widehat{f * g}$ we manipulate $\hat{h}$.

$$\begin{aligned}
\hat{h}(r) &= \frac{1}{2\pi}\int_{\mathbb{T}} h(x)e^{-irx}dx \\
&= \frac{1}{4\pi^2}\iint_{\mathbb{T}^2} f(s)g(x-s)e^{-irx}ds\,dx \\
&= \frac{1}{4\pi^2}\iint_{\mathbb{T}^2} f(s)g(x-s)e^{-irs}e^{-ir(x-s)}dx\,ds
\end{aligned}$$

using $e^{-irx} = e^{-irs}e^{-ri(x-s)}$ and interchanging the order of integration. Then by taking $u = x - s$ we have

$$\begin{aligned}
&= \frac{1}{4\pi^2}\iint_{\mathbb{T}^2} f(s)g(u)e^{-irs}e^{-iru}du\,ds \\
&= \frac{1}{4\pi^2}\int_{\mathbb{T}} f(s)e^{-irs}ds \int_{\mathbb{T}} g(u)e^{-iru}du \\
&= \left(\frac{1}{2\pi}\int_{\mathbb{T}} f(s)e^{-irs}ds\right)\left(\frac{1}{2\pi}\int_{\mathbb{T}} g(u)ds\right) \\
&= \hat{f}\cdot\hat{g}.
\end{aligned}$$

□

4.2 Hurwitz's Proof of the Isoperimetric Inequality

Recall from last lecture that the isoperimetric problem is to show that a circle encloses the largest area for all curves of a fixed length. Formally, if L is the length of a curve and A the area enclosed, then we want to show that $L^2 - 4\pi A \geq 0$ with equality if and only if the curve is a circle. We will prove the following stronger theorem.

Theorem 4.4. *Let $(x, y) : \mathbb{T} \to \mathbb{R}^2$ be an anticlockwise arc length parametrization of a non self-intersecting curve Γ of length L enclosing an area A. If $x, y \in C^1$, then*

$$L^2 - 4\pi A = 2\pi^2\Big(\sum_{n\neq 0} |n\hat{x}(n) - i\hat{y}(n)|^2 + |n\hat{y}(n) + i\hat{x}(n)|^2 + (n^2-1)\big(|\hat{x}(n)|^2 + |\hat{y}(n)|^2\big)\Big).$$

In particular, $L^2 \geq 4\pi A$ with equality if and only if Γ is a circle.

We will not define "arc length parameterization" formally, only remark that intuitively it means that if one views the parameterization as describing the motion of a particle in the plane, then an arc length parameterization is one so that the speed of the particle is constant. In our context, where we view time as the unit circle $\mathbb{T}$ of circumference 2π, we have that $(\dot{x})^2 + (\dot{y})^2$ is a constant so that the total distance covered is $\left(\frac{L}{2\pi}\right)^2$.

Proof. First we use our identity about the parameterization to relate the length to the transform of the parameterization.

$$\begin{aligned}
\left(\frac{L}{2\pi}\right)^2 &= \frac{1}{2\pi}\int_{\mathbb{T}} \left(\big(\dot{x}(s)\big)^2 + \big(\dot{y}(s)\big)^2\right) ds \\
&= \|\hat{\dot{x}}\|_2^2 + \|\hat{\dot{y}}\|_2^2 && \text{by Parseval's} \\
&= \sum_{-\infty}^{\infty} |in\hat{x}(n)|^2 + |in\hat{y}(n)|^2 && \text{by Fourier differentiation identities} \\
&= \sum_{-\infty}^{\infty} -n^2\big(|\hat{x}(n)|^2 + |\hat{y}(n)|^2\big) && (4.1)
\end{aligned}$$

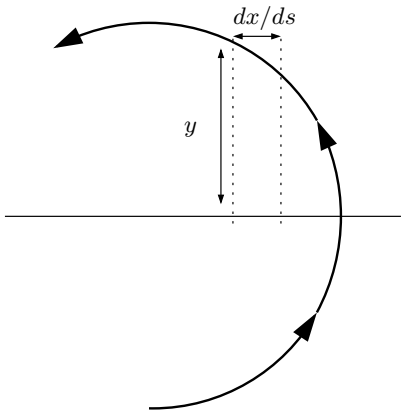


Figure 4.1: Computing the area enclosed by a curve

Now we compute the area. As the curve is anticlockwise,

$$A = -\int y \frac{dx}{ds} ds,$$

where the negative sign comes from the fact that the curve is anticlockwise. See Figure 4.1. This area integral looks like an inner product, so we write

$$\frac{A}{2\pi} = -\langle y, \dot{x} \rangle = -\langle \hat{y}, \hat{\dot{x}} \rangle.$$

By symmetry, considering the area integral from the other direction, we also have that

$$\frac{A}{2\pi} = \langle \hat{x}, \hat{\dot{y}} \rangle,$$

note there is no negative sign in this expression. Hence by adding we have that

$$\begin{aligned} \frac{A}{\pi} &= \langle \hat{x}, \hat{\dot{y}} \rangle - \langle \hat{y}, \hat{\dot{x}} \rangle \\ &= \sum_{-\infty}^{\infty} in\big(\hat{x}(n)^* \hat{y}(n) - \hat{x}(n)\hat{y}(n)^*\big), \end{aligned} \tag{4.2}$$

using the Fourier differentiation identities and using the notation a^* for the complex conjugate of a. Now subtract (4.1) and (4.2) to prove the theorem.

To see why the right hand side is zero if and only if Γ is a circle, consider when the right hand side vanishes. As it is a sum of many squares, $\hat{x}$ and $\hat{y}$ must vanish for all $n \neq 0$ or ± 1. Looking carefully at what those terms mean shows that they vanish if and only if Γ is a circle. □

4.3 Harmonic Analysis on the Cube for Coding Theory

The theory of error correcting codes is broad and has numerous practical applications. We will look at the asymptotic theory of block coding, which like many problems in coding theory is well-known, has a long

history and is still not well understood. The Boolean or Hamming cube $\{0,1\}^n$ is the set of all n-bit strings over $\{0,1\}$. The usual distance on $\{0,1\}^n$ is the *Hamming distance* $d_H(x,y)$, defined over $x, y \in \{0,1\}^n$ by the number of positions where x and y are not equal: $d_H(x,y) = |\{i : x_i \neq y_i\}|$. A code $\mathcal{C}$ is a subset of $\{0,1\}^n$. The *minimum distance* of $\mathcal{C}$ is the minimum distance between any two elements of $\mathcal{C}$:

$$\mathrm{dist}(\mathcal{C}) = \min\{d_H(x,y) : x, y \in \mathcal{C}\}.$$

The asymptotic question is to estimate the size of the largest code for any given minimum distance,

$$A(d,n) = \max\{|\mathcal{C}|, \mathcal{C} \subset \{0,1\}^n, \mathrm{dist}(\mathcal{C}) \geq d\},$$

as $n \to \infty$. The problem is easier if we restrict the parameter space by fixing d to be a constant fraction of the bit-length n, that is, consider $A(\delta n, n)$. Simple constructions show for $1/2 > \delta > 0$ that $A(\delta n, n)$ is exponential in n, so the interesting quantity will be the bit-rate of the code. Accordingly, we define the *rate* of a code as $\frac{1}{n}\log|\mathcal{C}|$, and then define the asymptotic rate limit as

$$R(\delta) = \limsup_{n\to\infty} \left\{ \frac{1}{n}\log|\mathcal{C}| : \mathcal{C} \subset \{0,1\}^n, \mathrm{dist}(\mathcal{C}) \geq \delta n \right\}.$$

It is a sign of our poor knowledge of the area that we do not even know if in the above we can replace the $\limsup$ by lim, i.e., if the limit exists. If $|\mathcal{C}| = 2^k$, we may think of the code as mapping k-bit strings into n-bit strings which are then communicated over a channel. The rate is then the ratio k/n, and measures the efficiency with which we utilize the channel.

A code is *linear* if $\mathcal{C}$ is a linear subspace of $\{0,1\}^n$, viewed as a vector space over $\mathbf{GF}(2)$. In a linear code, if the minimum distance is realized by two codewords x and y, then $x - y$ is a codeword whose (Hamming) length equals the minimum distance. Hence for linear codes we have that

$$\mathrm{dist}(\mathcal{C}) = \min\big\{|w| : w \in \mathcal{C} \setminus \{0\}\big\}.$$

Here we use $|\cdot|$ to indicate the *Hamming weight* of a codeword, the number of nonzero positions. Note that this is equal to several other, common norms on $\mathbf{GF}(2)$.

A useful entity is the *orthogonal code* to a given code. If $\mathcal{C}$ a linear code, we define

$$\mathcal{C}^\perp = \{y : \forall x \in \mathcal{C}, \langle x, y\rangle = 0\},$$

where we compute the inner product $\langle\cdot,\cdot\rangle$ over $\mathbf{GF}(2)$, that is, $\langle x, y\rangle = \sum_{i=1}^n x_i y_i \pmod 2$.

4.3.1 Distance Distributions and the MacWilliams Identities

Our first concrete study of codes will be into the *distance distribution*, which are the probabilities

$$\Pr[|x - y| = k : x, y \text{ chosen randomly from } \mathcal{C}]$$

for $0 \leq k \leq n$. If $\mathcal{C}$ is linear, our discussion above shows that the question of distance distribution is identical to the weight distribution of a code, the probabilities that a randomly selected codeword has a specified weight.

The MacWilliams Identities are important identities about this distribution that are easily derived using Parseval's Identity. Let $f = 1_{\mathcal{C}}$, the indicator function for the code. We first need the following lemma.

Lemma 4.5.

$$\hat{f} = \frac{|\mathcal{C}|}{2^n} 1_{\mathcal{C}^\perp}$$

Proof.

$$\begin{aligned}\hat{f}(u) &= \frac{1}{2^n} \sum_v f(v)\chi_v(u) \\ &= \frac{1}{2^n} \sum_v f(v)(-1)^{\langle u,v\rangle} \\ &= \frac{1}{2^n} \sum_{v\in\mathcal{C}} (-1)^{\langle u,v\rangle}\end{aligned}$$

If $u \in \mathcal{C}^\perp$, then $\langle u, v\rangle = 0$ for all $v \in \mathcal{C}$, so that $\hat{f}(u) = |\mathcal{C}|/2^n$. Suppose otherwise, so that $\sum_{\mathcal{C}}(-1)^{\langle u,v\rangle} = |\mathcal{C}_0| - |\mathcal{C}_1|$, where $\mathcal{C}_0$ are the codewords of $\mathcal{C}$ that are perpendicular to u, and $\mathcal{C}_1 = \mathcal{C} \setminus \mathcal{C}_1$. As $u \notin \mathcal{C}^\perp$, $\mathcal{C}_1$ is nonempty. Pick an arbitrary w in $\mathcal{C}_1$. Then, any $y \in \mathcal{C}_1 \setminus \{w\}$ corresponds to a unique $x \in \mathcal{C}_0$, namely $w + y$. Similarly, any $x \in \mathcal{C}_0 \setminus \{0\}$ corresponds to $w + x \in \mathcal{C}_1 \setminus w$. As $w \in \mathcal{C}_1$ corresponds to $0 \in \mathcal{C}_0$, we have that $|\mathcal{C}_0| = |\mathcal{C}_1|$. Hence $\sum_{\mathcal{C}}(-1)^{\langle e_1,v\rangle} = 0$, so that

$$\hat{f}(u) = \begin{cases} |\mathcal{C}|/2^n & \text{if } u \in \mathcal{C}^\perp \\ 0 & \text{otherwise} \end{cases}$$

which proves the lemma. □

We now define the *weight enumerator* of a code to be

$$P_{\mathcal{C}}(x, y) = \sum_{w\in\mathcal{C}} x^{|w|} y^{n-|w|}.$$

The MacWilliams Identity connects the weight enumerators of $\mathcal{C}$ and $\mathcal{C}^\perp$ for linear codes.

Theorem 4.6 (The MacWilliams Identity).

$$P_{\mathcal{C}}(x, y) = |\mathcal{C}| P_{\mathcal{C}^\perp}(y - x, y + x)$$

Proof. Harmonic analysis provides a nice proof of the identity by viewing it as an inner product. Define $f = 1_{\mathcal{C}}$ and $g(w) = x^{|w|} y^{n-|w|}$. Then, using Parseval's,

$$P_{\mathcal{C}}(x, y) = 2^n \langle f, g\rangle = 2^n \langle \hat{f}, \hat{g}\rangle.$$

$\hat{f}$ has already been computed in Lemma 4.5, so we turn our attention to $\hat{g}$.

$$\begin{aligned}\hat{g}(u) &= \frac{1}{2^n} \sum_v g(v)(-1)^{\langle u,v\rangle} \\ &= \frac{1}{2^n} \sum_v x^{|v|} y^{n-|v|} (-1)^{\langle u,v\rangle}\end{aligned}$$

Let u have k ones and $n-k$ zeros. For a given v, let s be the number of ones of v that coincide with those of u, and let t be the number ones of v coinciding with the zeros of u. Then we rewrite the sum as

$$\begin{aligned}
&= \frac{1}{2^n} \sum_{s,t,k} \binom{k}{s} \binom{n-k}{t} x^{s+t} y^{n-s-t} (-1)^s \\
&= \frac{y^n}{2^n} \sum_{s} \binom{k}{s} \left(-\frac{x}{y}\right)^s \sum_{t} \binom{n-k}{t} \left(\frac{x}{y}\right)^t \\
&= \frac{y^n}{2^n} \left(1 - \frac{x}{y}\right)^k \left(1 + \frac{x}{y}\right)^{n-k} \\
&= \frac{1}{2^n} (y-x)^k (y+x)^{n-k} \\
&= \frac{1}{2^n} (y-x)^{|u|} (y+x)^{n-|u|}
\end{aligned}$$

Now, as $\langle f, g \rangle = \langle \hat{f}, \hat{g} \rangle = 2^{-n} P_{\mathcal{C}}(x, y)$, we plug in our expressions for $\hat{f}$ and $\hat{g}$ to get

$$\begin{aligned}
2^{-n} P_{\mathcal{C}}(x, y) &= \frac{|\mathcal{C}|}{2^n} \sum_{w \in \mathcal{C}^\perp} \frac{1}{2^n} (y-x)^{|w|} (y+x)^{n-|w|} \\
&= \frac{|\mathcal{C}|}{2^n} P_{\mathcal{C}^\perp}(y-x, y+x),
\end{aligned}$$

which implies

$$P_{\mathcal{C}} = |\mathcal{C}| P_{\mathcal{C}^\perp}(y-x, y+x).$$

□

4.3.2 Upper and Lower Bounds on the Rate of Codes

We now turn our attention to upper and lower bounds for codes. We remind any complexity theorists reading these notes that the senses of "upper bound" and "lower bound" are reversed from their usage in complexity theory. Namely, a lower bound on $R(\delta)$ shows that good codes exist, and an upper bound shows that superb codes don't.

In the remainder of this lecture we show several simple upper and lower bounds, an then set the stage for the essentially strongest known upper bound on the rate of codes, the MacEleiece, Rumsey, Rodemich and Welsh (MRRW) upper bound. This is also referred to as the JPL bound, after the lab the authors worked in, or the linear programming (LP) bound, after its proof method.

Our first bound is a lower bound. Recall the binary entropy function

$$H(x) = -x \log x - (1-x) \log(1-x).$$

Theorem 4.7 (Gilbert-Varshamov Bound).

$$R(\delta) \geq 1 - H(\delta),$$

and there exists a linear code satisfying the bound.

Proof. We will sequentially pick codewords where each new point avoids all δn-spheres around previously selected points. The resulting code $\mathcal{C}$ will satisfy

$$|\mathcal{C}| \geq \frac{2^n}{\text{vol(sphere of radius } \delta n)} = \frac{2^n}{\sum_{j=0}^{\delta n} \binom{n}{j}}.$$

Now, note that $\log \binom{n}{\alpha n}/n \to H(\alpha)$ as $n \to \infty$, so that $2^n / \sum^{\delta n} \binom{n}{j} \sim 2^{n(1-H(\delta))}$, and take logs to prove the first part of the theorem.

We now show that there's a linear code satisfying this rate bound. This proof is different than the one given in class, as I couldn't get that to work out. The presentation is taken from Trevison's survey of coding theory for computational complexity. We can describe a linear k-dimensional code $\mathcal{C}_A$ by a $k \times n$ 0-1 matrix A by $\mathcal{C}_A = \{Ax : x \in \{0,1\}^k\}$. We'll show that if $k/n \leq 1 - H(\delta)$, with positive probability $\text{dist}(\mathcal{C}_A) \geq \delta n$. As the code is linear, it suffices to show that the weight of all nonzero codewords is at least δn. As for a given $x \in \{0,1\}^k$, Ax is uniformly distributed over $\{0,1\}^n$, we have

$$\Pr[|Ax| < \delta n] = 2^{-n} \sum_{i=0}^{\delta n - 1} \binom{n}{i} \leq 2^{-n} 2^{nH(\delta)+o(n)},$$

using our approximation to the binomial sum. Now we take a union bound over all 2^k choices for x to get

$$\Pr[\exists x \neq 0 : Ax < d] \leq 2^k \cdot 2^{-n} \cdot 2^{nH(\delta)+o(n)} = 2^{k+n(H(\delta)-1)+o(1)} < 1$$

by our choice of $k \leq n(1 - H(\delta))$. □

We now turn to upper bounds on $R(\delta)$.

Theorem 4.8 (Sphere-Packing Bound).

$$R(\delta) \leq 1 - H(\delta/2)$$

Proof. The theorem follows from noting that balls of radius $\delta n/2$ around codewords must be disjoint, and applying the approximations used above for the volume of spheres in the cube. □

We note in Figure 4.2 that the sphere-packing bound is far from the GV bound. In particular, that GV bound reaches zero at $\delta = 1/2$, while the sphere-packing bound is positive until $\delta = 1$. However, we have the following simple claim.

Claim 4.1. $R(\delta) = 0$ for $\delta > 1/2$.

Proof. We will show the stronger statement that if $|\mathcal{C}|$ is substantial then not only is it impossible for $d_H(x,y) > \delta n$ for all $x, y \in \mathcal{C}$, but even the average of all $x, y \in \mathcal{C}$ will be at most $n/2$. This average distance is

$$\frac{1}{\binom{|\mathcal{C}|}{2}} \sum_{\mathcal{C} \times \mathcal{C}} d(x,y),$$

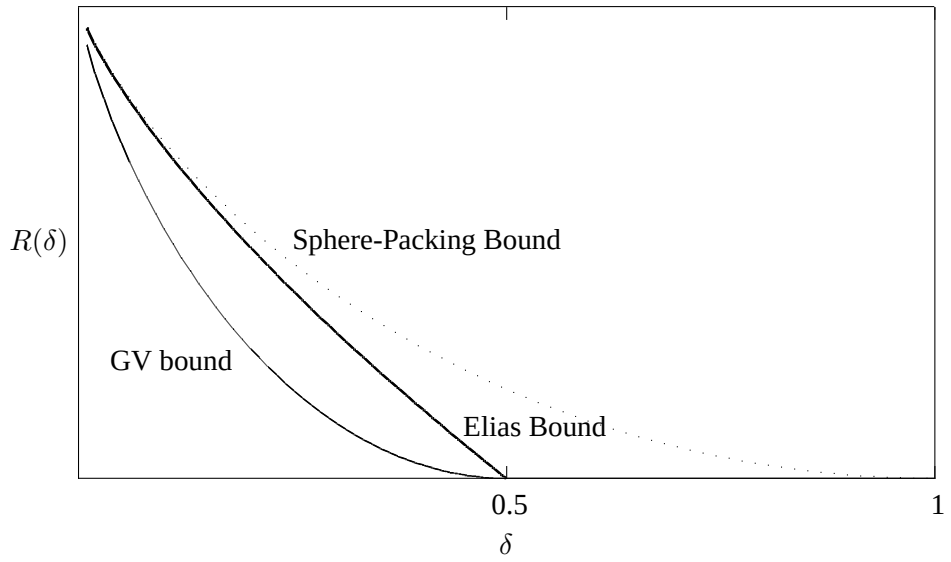


Figure 4.2: The GV bound contrasted with the Sphere-Packing Bound

and we will expand the distance $d(x, y) = |\{i : x_i \neq y_i\}|$. Reversing the order of summation,

$$\begin{aligned}\text{Average distance} &= \frac{1}{\binom{|\mathcal{C}|}{2}} \sum_i \sum_{x,y} 1_{x_i \neq y_i} \\ &= \frac{1}{\binom{|\mathcal{C}|}{2}} \sum_i z_i(|\mathcal{C}| - z_i),\end{aligned}$$

where z_i is the number of zeros in the i^{th} position of all the codewords of $\mathcal{C}$.

$$\begin{aligned}&\leq \frac{1}{\binom{|\mathcal{C}|}{2}} \sum \frac{|\mathcal{C}|^2}{4} \\ &\leq \frac{1}{2} n \cdot \frac{|\mathcal{C}|}{|\mathcal{C}| - 1}.\end{aligned}$$

So unless $\mathcal{C}$ is very small, the average distance is essentially $n/2$. □

Our next upper bound improves on the sphere packing bounds, at least achieving $R(\delta) = 0$ for $\delta > 1/2$. It still leaves a substantial gap with the GV bound.

Theorem 4.9 (Elias Bound).

$$R(\delta) \leq 1 - H\left(\frac{1 - \sqrt{1 - 2\delta}}{2}\right)$$

Proof. The proof begins by considering the calculation of average distance from the previous theorem. It follows from Jensen's inequality that if the average weight of the vectors in $\mathcal{C}$ is αn, then the maximum of $\sum z_i(|\mathcal{C}| - z_i)$ is obtained if for all i, $z_i = (1 - \alpha)\mathcal{C}$ for some α. We sketch the argument for those not

familiar with Jensen's inequality. The inequality states that if f is convex, then for $x_1, \ldots, x_n$, $\frac{1}{n}\sum f(x_i) \leq f(\sum x_i/n)$, with equality if and only if $x_1 = \cdots = x_n$. For our case, the function $f(x) = x(|\mathcal{C}| - x)$ is easily verified convex and so its maximum over a set of z_i is achieved when the z_i are all equal. This makes the average distance in $\mathcal{C}$ at most $2\alpha(1-\alpha)n$.

With this calculation in mind, chose a spherical shell S in $\{0,1\}^n$ centered at some x_0 with radius r such that

$$|S \cap \mathcal{C}| \geq |C| \cdot \frac{|S|}{2^n}.$$

Such a shell exists as the right hand side of the inequality is the expected intersection size if the sphere is chosen randomly. Set $r = pn$ so that $|S| \approx 2^{nH(p)}$, which means

$$|S \cap \mathcal{C}| \geq \frac{|\mathcal{C}|}{2^{n\left(1-H(p)\right)}}.$$

Now apply the argument above on $x_0 + \mathcal{C} \cap S$. It follows from our discussion that we actually have a p fraction of ones in each row, so if $\delta > 2p(1-p)$, the $|S \cap \mathcal{C}|$ is subexponential, but this is equal to $|\mathcal{C}|2^{-n\left(1-H(p)\right)}$, implying

$$\frac{1}{n}\log|\mathcal{C}| < 1 - H(p).$$

Let us rewrite our condition $\delta > 2p(1-p)$ as follows:

$$1 - 2p \geq \sqrt{1-2\delta} \Rightarrow p = \frac{1-\sqrt{1-2\delta}}{2}.$$

This is the critical value of p—when p is below this the code is insubstantially small. □

Figure 4.2 shows how the Elias bound improves the sphere-packing bound to something reasonable. The gap between it and the GV bound is still large, however.

4.4 Aside: Erdoös-Ko-Rado Theorem

The proof of the Elias bound that we just saw is based on the following clever idea: we investigate and unknown object (the code $\mathcal{C}$) by itersecting it with random elements of a cleverly chosen set (the sphere). This method of "a randomly chosen fish-net" is also the basis for the following beautiful proof, due to Katona, of the Erdös-Ko-Rado theorem.

Definition 4.1. An *intersecting family* is a family $\mathcal{F}$ of k-sets in $1 \ldots n$ (compactly, $\mathcal{F} \subseteq \binom{[n]}{k}$), with $2k \leq n$, such that for any $A, B \in \mathcal{F}$, $A \cap B \neq \emptyset$.

Informally, an intersecting family is a collection of sets which are pairwise intersecting. One way to construct such a set is to pick a common point of intersection, and then choose all possible $(k-1)$-sets to fill out the sets. The Erdoös-Ko-Rado Theorem says that this easy construction is the best possible.

Theorem 4.10 (Erdös-Ko-Rado). *If $\mathcal{F} \subseteq \binom{[n]}{k}$ is an intersecting family with $2k \leq n$, then*

$$\mathcal{F} \leq \binom{n-1}{k-1}.$$

Proof (Katona). Given an intersecting family $\mathcal{F}$, arrange $1 \dots n$ in a random permutation π along a circle, and count the number of sets $A \in \mathcal{F}$ such that A appears as an arc in π. This will be our random fish-net.

There are $(n-1)!$ cyclic permutations—that is, $n!$ permutations, divided by n as rotations of the circle are identical. There are $k!$ ways for a given k-set to be arranged, and $(n-k)!$ ways of the other elements not interfering with that arc, so that the set appears consecutively on the circle. Hence the probability that a given k-set appears as an arc is

$$\frac{k!(n-k)!}{(n-1)!} = \frac{n}{\binom{n}{k}},$$

which by linearity of expectation implies

$$E\begin{bmatrix}\text{\# arcs belonging} \\ \text{to } \mathcal{F}\end{bmatrix} = \frac{n|\mathcal{F}|}{\binom{n}{k}}.$$

Now, as $2k \leq n$, at most k member of an intersecting family can appear as arcs on the circle, otherwise two of the arcs wouldn't intersect. Hence

$$\frac{n|\mathcal{F}|}{\binom{n}{k}} \leq k$$

implying

$$|\mathcal{F}| \leq \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}$$

□

Lecture 5

Isoperimetric Problems

Feb 11, 2005
Lecturer: Nati Linial
Notes: Yuhan Cai & Ioannis Giotis

Codes: densest sphere packing in $\{0,1\}^n$.

$A(n,d) = max\{|\varphi|, \varphi \subseteq \{0,1\}^n, dist(\varphi) \geq d\}$

$R(\delta) = limsup\{\frac{1}{n}\log_2(\varphi) | \varphi \subseteq \{0,1\}^n, dist(\varphi) \geq \delta_n\}$

'Majority is the stablest' -

- Gaussian: $\frac{1}{(2\pi)^{(n/2)}} e^{-\|x\|^2/2}$
- Borell: isoperimetric problem is solved by a half-space

Isoperimetric Questions on the cube (Harper): Vertex and Edge isoperimetric questions.

The edge problem is defined as follows: Given that $S \subseteq \{0,1\}^n, |S| = R$, how small $e(S, \bar{S})$ be?

Answer: $\forall S \subseteq \{0,1\}^n, e(S) \leq 1/2|S|\log_2|S|, |S| = 2^k, S = \{(* \ldots * 0 \ldots 0\}$ with k *s.

Proof (induction on dim):

$e(S) \leq e(S_0) + e(S_1) + |S_0|, |S| = x, |S_0| \geq \alpha x, \alpha < 1/2.$

$1/2x\log_2 x \geq 1/2(\alpha x)\log_2 \alpha x) + 1/2(1-\alpha)x\log[(1-\alpha)x] + \alpha x$

$0 \geq \alpha\log\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha$

$H(\alpha) \geq 2\alpha$ at $\alpha = 0, 1/2$.

The vertex isoperimetric problem is defined as $min\sharp\{y \mid y \notin S, \exists x \in S\}$ such that $xy \in E(\langle 0,1\rangle^n)$, $S \subseteq \{0,1\}^n$, $|S| \leq k$. The answer is an optimal S-ball. Specifically, if $k = |S| = \sum_{j=0}^{t}\binom{n}{j}$, then $|S| \geq \binom{n}{t+1}$.

We will use the Kraskal-Katona theorem. If $f \subseteq \binom{[n]}{k}$, then the *shadow of* f is

$$\sigma(f) = \left\{ y \in \binom{[n]}{k} \mid \exists x \in f, x \supseteq y \right\}$$

We wish to minimize $|\sigma(f)|$.

To do this, take f as an initial segment in the reverse lexicographic order. The lexicographic order is defined as

$$A < B, \text{ if } min(A \backslash B) < min(B \backslash A)$$

while the reverse lexicographic order is

$$A <_{RL} B, \text{ if } max(A \backslash B) <_{RL} \max(B \backslash A)$$

For example:

$$\begin{aligned} Lex &: \langle 1,2\rangle\langle 1,3\rangle\langle 1,4\rangle, \ldots \\ RLex &: \langle 1,2\rangle\langle 1,3\rangle\langle 2,3\rangle, \ldots \end{aligned}$$

Margulis and Talagrand gave the following definition for $S \subseteq \langle 0,1\rangle^n$

$$h(x) = \{y \notin S \mid xy \in E\}, x \in S$$

We now have the 2 problems

- Vertex Isoperimetric, $min_{|S|=k} \sum_{x \in S} (h(x))^{0 \to \rho = 0}$
- Edge Isoperimetric, $min_{|S|=k} \sum_{x \in S} (h(x))^{\to \rho = 1}$

We have $|S| \geq 2^{n-1} \Rightarrow \Sigma \sqrt{h(x)} \geq \Omega(2^n)$, for $p = 1/2$.

Kleitman: $|S| = \Sigma_{j=0}^{t} \binom{n}{j}$, $S \subseteq \{0,1\}^n$, $t < n/2 \Rightarrow diam(S) \geq 2t$. Can you show that S necessarily contains a large code?

Question: (answered by Friedgut) suppose that $|S| \simeq 2^{n-1}$ and $\varphi(S, S^C) \sim 2^{n-1}$, then is S roughly a dictatorship?

Answer: yes. subcube $x_1 = 0 \Leftrightarrow f(x_1, \ldots, x_n) = x_1$. $R(\delta) = limsup_{n \to \infty} \{\frac{1}{n} \log(\varphi) | \varphi \subseteq \{0,1\}^n, dist(\varphi) \geq \delta n\}$.

5.1 Delsarte's LP

Having $g = 1_C$, $f = 2^n g * g / |C|$, Delsarte's LP is

$$\begin{aligned} A(n,d) \leq max \Sigma_{x \in \{0,1\}^n} f(x) \\ f \geq 0 \\ f(\mathbf{0}) = 1 \\ \hat{f} \geq 0 \\ f|_{1,\ldots,d-1} = 0 \end{aligned}$$

Some useful equations

$$\begin{aligned} g * g(0) &= \frac{1}{2^n} \sum g(y) g(y) = \frac{|C|}{2^n} \\ g * g(S) &= \frac{1}{2^n} \sharp \{x, y \in C \mid x \oplus y = S\} \end{aligned}$$

We start with an observation. Without loss of generality, f is symmetric or in other words $f(x)$ depends only on $|x| = \alpha_{|x|}$. We look for $\alpha_0 = 1, \alpha_1 = \ldots = \alpha_{d-1} = 0$, $\alpha_d, \ldots, \alpha_n \geq 0$.

We've expressed $f \geq 0, f(\hat{0}) = 1$ and we are trying to maximize $\sum \binom{n}{j} \alpha_j$.

$$\begin{aligned} L_j &= \{x \in \{0,1\}^n, |x| = j\} \\ f &= \sum_{j=0}^{n} \alpha_j 1_{L_j} \\ \hat{f} &= \sum_{j} \alpha_j \hat{1}_{L_j} \end{aligned}$$

Note that L_j is symmetric. It also follows that $\hat{1}_{L_j}$ is symmetric. We need to know $\hat{1}_{L_j}$ if $|y| = t$.

$$\begin{aligned} \hat{\phi}(T) &= \sum \phi(S)(-1)^{|S \cap T|} \\ \hat{1}_{L_j}(T) &= \sum_{|S|=j} (-1)^{|S \cap T|} \\ K_j^{(n)}(x) &= \sum_i (-1)^i \binom{t}{i} \binom{n-t}{j-i} \end{aligned}$$

This is the *Krawtchouk* polynomial presented in the next section.

5.2 Orthogonal Polynomials on $\mathbb{R}$

Interesting books for this section are "Interpolation and Approximation" by Davis and "Orthogonal polynomials" by Szegö.

The weights of orthogonal polynomials on $\mathbb{R}$ are defined by

$$w : \mathbb{R} \to \mathbb{R}^+ \, , \, \int_{\mathbb{R}} w(x) < \infty$$

The inner product on $f : \mathbb{R} \to \mathbb{R}$ is

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x) g(x) w(x) \, dx$$

and with weights $w_1, w_2, \ldots$, and points $x_1, x_2, \ldots$

$$\langle f, g \rangle = \sum w_i f(x_i) g(x_i)$$

Let's now talk about orthogonality. Start from the functions $1, x, x^2, \ldots$ and carry out a Gram-Schmidt orthogonalization process. You'll end up with a sequence of polynomials $P_0(x), P_1(x), \ldots$ s.t. P_i has degree i and $\langle P_i, P_j \rangle = \delta_{ij}$.

One case of orthogonal polynomials are the *Krawtchouk* polynomials, on discrete points $x_0 = 0, x_1 = 1, \ldots, x_n = n$ with $w_j = \binom{n}{j}/2^n$. The j-th Krawtchouk polynomial $K_j(x)$ is a degree j polynomial in x. It is also the value of $\hat{1}_{L_j}(T)$ whenever $|T| = x$.

$$K_j^{(n)}(x) = \sum_{i=0}^{n} (-1)^i \binom{x}{i}\binom{n-x}{j-i}$$

Let's see why are they orthogonal or in other words

$$\frac{1}{2^n}\sum_{i=0} K_p(i)K_q(i)\binom{n}{i} = \delta_{pq}\binom{n}{p}$$

Starting from

$$\langle 1_p, 1_q\rangle = \frac{1}{2^n}\binom{n}{p}\delta_{pq}$$

and using Parseval's identity we get

$$\langle \hat{1}_{L_p}, \hat{1}_{L_q}\rangle = \frac{1}{2^n}\sum K_p(|S|)K_q(|S|) = \frac{1}{2^n}\sum_{i=0} K_p(i)K_q(i)\binom{n}{i}$$

The first K_j's are

$$K_0(x) = 1, K_1(x) = n - 2x, K_2(x) = \binom{x}{2} - (n-x) + \binom{n-x}{2} = \frac{(n-2x)^2 - n}{2}$$

We also have the following identity

$$K_j(n-x) = (-1)^j K_j(x)$$

Lemma 5.1. *Every system of orthogonal polynomials satisfies a 3-term recurrence*

$$xP_j = \alpha_j P_{j+1} + \beta_j P_j + \gamma_j P_{j-1}$$

Proof.

$$\begin{aligned} 1_{L_i} * 1_{L_j}(S) &= \frac{1}{2^n}\sum_i 1_{L_j}(S \oplus i) = \\ &= \frac{1}{2^n}((j+1)1_{L_{j+1}} + (n-j+1)1_{L_{j-1}}) = \\ &= \frac{1}{2^n}((j+1)1_{L_{j+1}} + (n-j+1)1_{L_{j-1}}) \end{aligned}$$

For the Krawtchouk polynomials

$$\begin{aligned} K_iK_j &= (j+1)K_{j+1} + (n-j+1)K_{j-1} \\ (n-2x)K_j &= (j+1)K_{j+1} + (n-j+1)K_{j-1} \end{aligned}$$

□

Theorem 5.2. *For every family of orthogonal polynomials there is*

1. *a 3-term recurrence relation*

$$x \cdot P_j = \alpha_j P_{j+1} + \beta_j P_j + \gamma_j P_{j-1}$$

2. *P_j has j real roots all in $conv[supp\ w]$.*

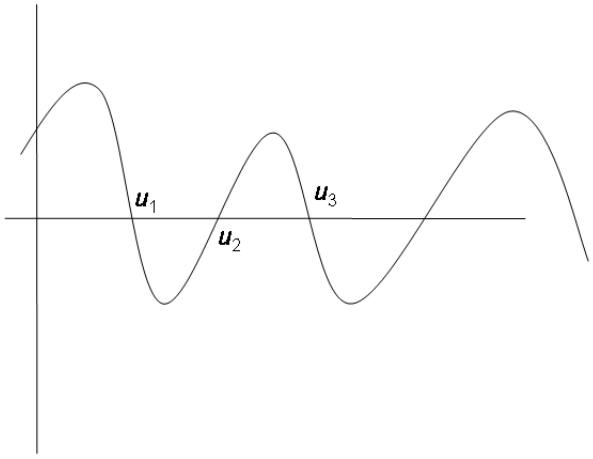
Proof. Observe that $P_0, P_1, \ldots, P_t$ form a basis for the space of all polynomials of degree $\leq t$, which means that $\langle P, Q\rangle = 0$, $\forall Q$ polynomials of degree j

$$x \cdot P_j = \sum_{i=0}^{j+1} \lambda_i P_i \tag{5.1}$$

We now claim that $\lambda_0 = \lambda_1 = \cdots = \lambda_{j-2} = 0$. Let's take in (5.1) an inner product with P_l,$l < j-1$.

$$\begin{aligned} \langle xP_j, P_l\rangle &= \sum_{i=0}^{j+1} \lambda_i \langle P_i, P_j\rangle = \lambda_l \|P_l\|^2 \\ \langle P_j, xP_l\rangle &= \lambda_l \|P_l\|^2 \end{aligned}$$

which is 0 for P_l of degree $\leq j-1$. □



If u_i's are the zeros of P_j of odd multiplicity then

$$0 = \langle P_j, \prod(x - u_i)\rangle = P_j \prod(x - u_j) > 0$$

Lecture 6

MRRW Bound and Isoperimetric Problems

Feb 18, 2005
Lecturer: Nati Linial
Notes: Ethan Phelps-Goodman and Ashish Sabharwal

6.1 Preliminaries

First we recall the main ideas from the last lecture. Let

$$g = 1_C, \qquad f = \frac{g * g}{|C|}.$$

Then we can bound the code size $A(n, d)$ using Delsarte's linear program:

$$A(n,d) \leq \max_f \sum_{x \in \{0,1\}^n} f(x)$$

subject to

$$f \geq 0 \qquad f(\mathbf{0}) = 1$$
$$\widehat{f} \geq 0 \qquad f_{|1,\ldots,d-1} = 0$$

By averaging over a solution f, we can get an equivalent solution that is symmetric about permutations of the input bits. That is, we can assume w.l.o.g. that f that depends only on the hamming weight of the input. f is then determined by $n+1$ coordinate weights A_j by

$$A_j = \sum_{x \mid |x|=j} f(x)$$

Or equivalently,

$$f = \sum_{j=0}^{n} \frac{A_j}{\binom{n}{j}} 1_{L_j}$$

Central to our proof will be the Krawtchouk polynomials, which are related to our linear program by

$$\widehat{1}_{L_r} = K_r(x) = \sum_{j=0}^{r} (-1)^j \binom{x}{j}\binom{n-x}{r-j}$$

$$\widehat{f} = \sum_{j=0}^{n} \frac{A_j}{\binom{n}{j}} K_j$$

6.2 Primal and Dual Programs

Making the substitutions above we can now write Delsarte's program in terms of Krawtchouk polynomials and symmeterized f.

$$A(n,d) \leq \max_{A_0,\ldots,A_n} \sum_{i=0}^{n} A_i$$

subject to

$$\begin{aligned} & A_0 = 1 \\ & A_1, \ldots, A_{d-1} = 0 \\ \forall k \in \{0,\ldots,n\} \quad & \sum_{i=0}^{n} \frac{A_i}{\binom{n}{i}} K_i(k) \geq 0. \end{aligned}$$

This can be further simplified with the following identity for Krawtchouk polynomials.

Fact 6.1.

$$\frac{K_i(k)}{\binom{n}{i}} = \frac{K_k(i)}{\binom{n}{k}}$$

Proof.

$$\begin{aligned} \frac{1}{\binom{n}{i}} \sum_{j=0}^{i} (-1)^j \binom{k}{j}\binom{n-k}{i-j} &= \sum_{j}^{i} (-1)^j \frac{i!(n-i)!k!(n-k)!}{n!j!(k-j)!(i-j)!(n-k-i+j)!} \\ &= \frac{1}{\binom{n}{k}} \sum_{j}^{i???} (-1)^j \binom{i}{j}\binom{n-i}{k-j} \end{aligned}$$

□

Using this in the last constraint, and removing the $1/\binom{n}{k}$ term, which pulls out of the sum and doesn't affect the sign, we get the constraints

$$\forall k \in \{0,\ldots,n\} \quad \sum_{i=0}^{n} A_i K_k(i) \geq 0.$$

Our approach will be to use LP duality to give a bound on the maximum of this program. Recall that duality tells us that the maximum value of the primal is at most the minimum value of the dual. Strong duality states that the optima are exactly equal, but we will not use this.

Start by multiplying each of the $\sum_{i=0}^{n} A_i K_k(i) \geq 0$ constraints by β_k, and summing all of the constraints. This gives

$$\sum_{k=1}^{n} \beta_k \sum_{i=0}^{n} A_i K_k(i) = \sum_{i=0}^{n} A_i \sum_{k=1}^{n} \beta_k K_k(i) \geq 0$$

Let $\gamma(x) = \sum_{k=1}^{n} \beta_k K_k(i)$. If we add the constraint that $\forall x, \gamma(x) \leq -1$, then using $A_0 = 1, A_1, \ldots, A_{d-1} = 0$, we get

$$\sum_{i=0}^{n} A_i \gamma(i) = \gamma(x) + \sum_{i=d}^{n} A_i \gamma(i) \geq 0$$

$$\gamma(0) \geq -\sum_{i=d}^{n} A_i \gamma(i)$$

$$\geq \sum_{i=d}^{n} A_i$$

$$\gamma(0) + 1 \geq \sum_{i=1}^{n} A_i \geq A(n, d)$$

What we have done here is just an explicit construction of the dual. The reader can check that this dual can be arrived at by any standard method for computing the dual.

Let $\beta(x) = 1 + \sum_{k=1}^{n} \beta_k K_k(x)$. Then our final program is given by

$$A(n, d) \leq \min_{\beta_k} \beta(0)$$

subject to:

$$\forall k = 1, \ldots, n, \quad \beta_k \geq 0$$
$$\forall j = d, \ldots, n, \quad \beta(j) \leq 0$$

6.3 The Upper Bound

To show an upper bound on $A(n, d)$ we need to demonstrate a feasible solution β and bound $\beta(0)$. First we need two additional facts about Krawtchouk polynomials.

Fact 6.2 (Christoffel-Darboux). Let $P_1, P_2, \ldots$ be a family of orthonormal polynomials, and let a_i be the leading coefficient of P_i. Then

$$\frac{P_k(x)P_{k+1}(y) - P_{k+1}(x)P_k(y)}{y - x} = \frac{a_{k+1}}{a_k} \sum_{i=0}^{k} P_i(x)P_i(y)$$

For the case of Krawtchouk polynomials, the leading term of $K_r(x)$ is $\frac{-2^r}{r!}$. Also, to normalize we need to divide K_r by $\sqrt{\binom{n}{r}}$. Putting these together, we get

$$\frac{K_{r+1}(x)K_r(y) - K_r(x)K_{r+1}(y)}{y - x} = \frac{2}{r+1}\binom{n}{r} \sum_{i=0}^{r} \frac{K_i(x)K_i(y)}{\binom{n}{i}}.$$

The second fact we need is that the product of two Krawtchouk polynomials can be expressed as a non-negative combination of Krawtchouk polynomials.

Fact 6.3. For any p, q, there exist $\alpha_0, \ldots, \alpha_{p+q} \geq 0$ such that

$$K_p \cdot K_q = \sum_{j=0}^{p+q} \alpha_j K_j$$

This can be seen easily from the harmonic analysis perspective since $K_p \cdot K_q = \widehat{1}_{L_p} \cdot \widehat{1}_{L_q} = \widehat{1_{L_p} * 1_{L_q}}$, and the convolution is a positive combination.

We can now present the feasible solution for the dual. Let

$$\alpha(x) = \frac{\big(K_t(a)K_{t+1}(x) - K_{t+1}(a)K_t(x)\big)^2}{a - x}.$$

Then set $\beta(x) = \frac{\alpha(x)}{\alpha_0}$, where α_0 is chosen to make the constant term equal 1. Now we need to set values for a and t. Denote by $x_r^{(l)}$ the leftmost root of K_r. We know from last lecture that the roots of the Krawtchouk polynomials are real, lie in $[0, n]$, and interleave with one another. Therefore we can pick a t such that $0 < x_{t+1}^{(l)} < x_t^{(l)} < d$. In the region $(x_{t+1}^{(l)}, x_t^{(l)})$, K_{t+1} is negative and K_t is positive, so we can pick an a such that $K_t(a) = -K_{t+1}(a)$.

Now we need to show that $\alpha(x)$ satisfies the two constraints from the dual. First, note that at all $x > d$, $\alpha(x) < 0$. Then we just need to show that $\alpha(x)$ is non-negative combination of Krawtchouk polynomials. Using the above settings, and Christoffel-Darboux, we can factor $\alpha(x)$ as

$$\begin{aligned}
\alpha(x) &= \big(K_t(a)K_{t+1}(x) - K_{t+1}(a)K_t(x)\big)\left[\frac{K_t(a)K_{t+1}(x) - K_{t+1}(a)K_t(x)}{a - x}\right] \\
&= K_t(a)(K_{t+1}(x) + K_t(x))\left[\frac{K_t(a)K_{t+1}(x) - K_{t+1}(a)K_t(x)}{a - x}\right] \\
&= K_t(a)(K_{t+1}(x) + K_t(x))\left[\frac{2}{r+1}\binom{n}{r}\sum_{i=0}^{r}\frac{K_i(x)K_i(y)}{\binom{n}{i}}\right]
\end{aligned}$$

Since all terms are positive, this can be expanded as a positive combination of Krawtchouk polynomials.

Now that we have a feasible solution to the dual, we just need to find the value of $\beta(0)$. We can use the fact that for $t \approx \tau n$, the leftmost root is at $x_t^{(l)} = (1 + o(1))(\frac{1}{2} - \sqrt{\tau(1-\tau)})n$. Given this we can conclude that $R(\delta) \leq H(\frac{1}{2} - \sqrt{\delta(1-\delta)})$. Both the lecture and van Lint [1] seem to imply that this step is obvious, but your scribe has been unable to see any connection.

6.4 More on Isoperimetric Problems on the Cube

We now turn our attention to isoperimetric problems. In a previous lecture, we studied isoperimetric questions on the n-dimensional cube, namely the vertex isoperimetric problem and the edge isoperimetric problem. Why is the study of such problems important? The reason is that Computer Science deals with Boolean functions which are simply partitions of the n-dimensional cube into two parts. Understanding the geometry of the cube is therefore critical to understand Boolean functions. Here is one more isoperimetric problem that is open.

Open Problem 6.1 (Chung-Füredi-Grahan-Seymour, 1988 J.C.T.A.). What is the largest $d = d(n)$ such that for all $S \subseteq \{0,1\}^n, |S| > 2^{n-1}$, there exists $x \in S$ with $d_S(x) \geq d$?

Here $d_S(x)$ denotes the number of neighbors of x in S. Note that for $|S| \leq 2^{n-1}$, S can be an independent set, i.e., $\forall x \in S. d_S(x) = 0$. Further, for $|S| > 2^{n-1}$, S may not be independent. In general, all we know is that $d(n)$ is both $O(\sqrt{n})$ and $\Omega(\log n)$. This leaves a huge gap open.

Consider any Boolean function $f : \{0,1\}^n \to \{0,1\}$ represented as a $0,1$-labeling of the n-dimensional cube seen as a layered lattice. This lattice has four types of edges as depicted in Figure 6.1. Let $S = f^{-1}(0)$. The two edges from 0 to 1 and from 1 to 0 belong to the cut $E(S, S^c)$ and thus contribute to the cut size $e(S, S^c)$.

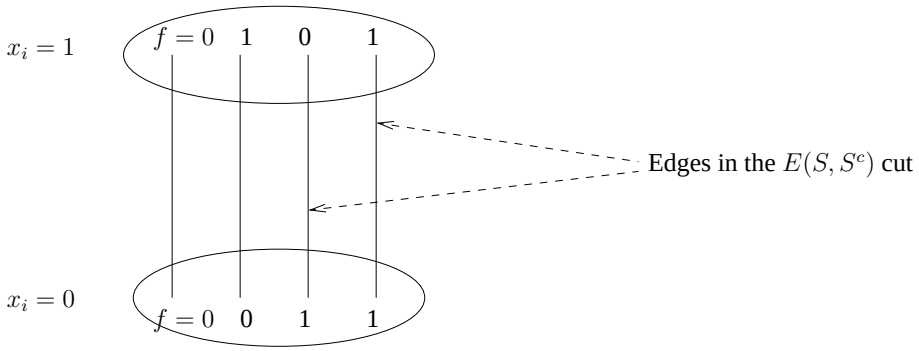


Figure 6.1: The cut defined in terms of the four types of edges in the lattice

If $|S| = 2^{n-1}$, then $e(S, S^c) \geq 2^{n-1}$. This is sharp for $S = \{x \mid x_1 = 0\}$. In the edge isoperimetry problem, given $|S|$, we want to minimize the cut size $e(S, S^c)$. What about trying to *maximize* the cut size instead? The maximum cut size can really be anything. Indeed, when f is the parity function, $e(S, S^c) = n2^{n-1}$.

6.4.1 Maximizing Edge Cut Size for Monotone Functions

Consider the setting of the previous section. How can we maximize the edge cut when f is *monotone*, i.e., $x \succ y \Rightarrow f(x) \geq f(y)$, where $x \succ y$ means $\forall i. x_i \geq y_i$? In the following, we use Parseval's identity to answer this question.

Theorem 6.1. *Let $S \subseteq \{0,1\}^n$ correspond to a monotone Boolean function $f : \{0,1\}^n \to \{0,1\}$. $f =$* majority *maximizes the edge cut size $e(S, S^c)$.*

Proof. It is clear from the lattice corresponding to $f =$ majority (see Figure 6.2) that the size of the cut corresponding to it is $\binom{n}{\lfloor n/2 \rfloor} = \Theta(\sqrt{n}\, 2^n)$. We will use Parseval's identity to prove that this is the optimal.

Let f be any monotone Boolean function in n dimensions. Recall that for characters $\chi_T(Z) = (-1)^{|Z \cap T|}$, the function f can be represented as $f = \sum_T \hat{f}(T)\chi_T$ where $\hat{f}(T) = \langle f, \chi_T \rangle$. What is $\hat{f}(\{i\})$?

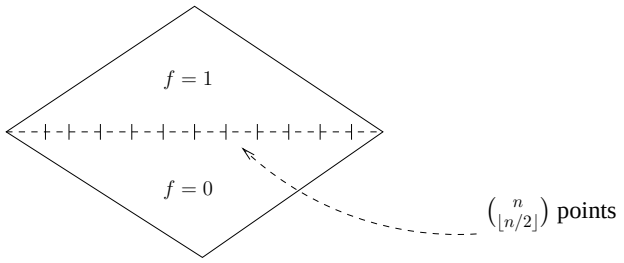


Figure 6.2: The lattice corresponding to the majority function

$\chi_{\{i\}}(Z) = (-1)^{|Z \cap \{i\}|}$ which is $+1$ if $i \notin Z$ and -1 if $i \in Z$. Therefore

$$\begin{aligned}
\hat{f}(\{i\}) &= \langle f, \chi_{\{i\}} \rangle \\
&= \frac{1}{2^n} \sum_Z f(Z) \chi_{\{i\}}(Z) \\
&= \frac{1}{2^n} \left(\sum_{Z \not\ni i} f(Z) - \sum_{Z \ni i} f(Z) \right) \\
&= -\frac{1}{2^n} \cdot \text{(number of cut edges in the } i\text{-direction)}
\end{aligned}$$

For ease of computation, convert everything from the $\{0,1\}$ basis to the $\{-1,+1\}$ basis. This quantity is then $(2/2^n)$ times the number of cut edges in the i-direction. Using Parseval's identity and Cauchy-Schwartz inequality, $1 = \|f\|_2^2 = \sum_S \left(\hat{f}(S)\right)^2 \geq \sum_i \left(\hat{f}(\{i\})\right)^2 \geq (1/n) \left(\sum_i \hat{f}(\{i\})\right)^2$. Hence $\sqrt{n} \geq \sum_i \hat{f}(\{i\}) = (2/2^n)\, e(S, S^c)$, which finishes the proof. □

We give an alternative *combinatorial* proof of the fact that $e(S, S^c) = 2^{n-1} \sum_i \hat{f}(\{i\})$ based on the following claim.

Claim 6.1. Let f be a monotone Boolean function. If the expectation of f is given and fixed, then to maximize $e(f^{-1}(0), f^{-1}(1))$, it is best to take f symmetric.

Proof of claim. Consider $\sum_{x:f(x)=0}(n - 2|x|)$. This is the sum of the first Krawtchouk polynomials and is equal to the cut size $e(f^{-1}(0), f^{-1}(1))$ because $(n - |x|)$ edges in the lattice corresponding to f that go upwards from x contributing $+1$ each while $|x|$ edges go downward from x contributing -1 each (see Figure 6.3. Maximizing this quantity means minimizing $\sum_{x:f(x)=0} |x|$ which happens exactly when f is "pushed down" as much as possible.

Formally, let us change the basis from $\{0,1\}$ to $\{-1,+1\}$ and reinterpret the summation. It is equal to $\sum_{x:f(x)=1}(n - 2|x|) - \sum_{x:f(x)=-1}(n - 2|x|) = 2^n \langle f, K_1 \rangle$. Observe however that $\sum_x (n - 2|x|) = \langle K_1, K_0 \rangle = 0$. Therefore $\sum_{x:f(x)=1}(n - 2|x|) = 2^{n-1} \langle f, K_1 \rangle$, which is the same as $\sum_i \hat{f}(\{i\})$ by the properties of Krawtchouk polynomials. □

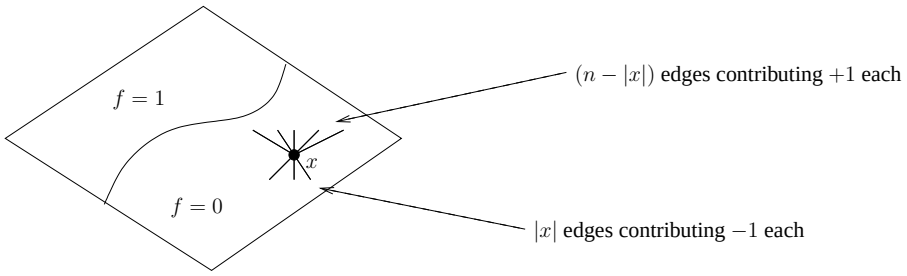


Figure 6.3: Contribution of f to the cut

6.4.2 The Brunn-Minkowski Inequality

Let v be a volume measure on subsets of $\mathbb{R}^n$.

Theorem 6.2 (Brunn-Minkowski [2]). *For A, B measurable subsets of $\mathbb{R}^n$,*

$$(v(A+B))^{1/n} \geq (v(A))^{1/n} + (v(B))^{1/n}.$$

Moreover, equality holds if and only if A and B are homothetic, i.e. $B = \lambda A + C$ for $\lambda \in \mathbb{R}$.

Here $A+B$ is the Minkowski sum defined as $\{a+b \mid a \in A, b \in B\}$, where $a+b$ is the standard vector sum over $\mathbb{R}^n$. For $\lambda \in \mathbb{R}$, λA is similarly defined as $\{\lambda a \mid a \in A\}$. We will not be using the second part of the theorem.

Let us try to understand what this inequality says. Take a convex body K in $\mathbb{R}^n$ and slide a hyperplane $A_t, t \in \mathbb{R}$, through it (see Figure 6.4). What can we say about the function $f(t) = \mu_{n-1}(A_t \cap K)$ which is the volume of the intersection of the body with the hyperplane? Brunn-Minkowski inequality says that $(f(t))^{1/(n-1)}$ is convex.

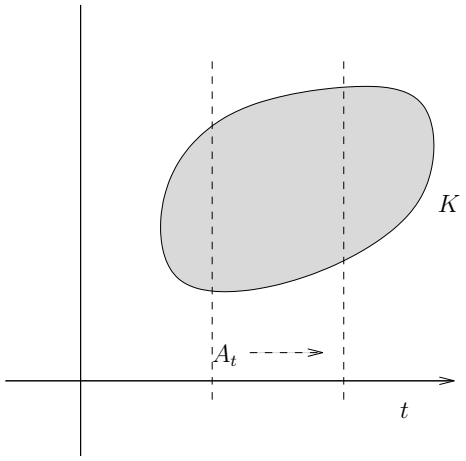


Figure 6.4: Sliding a hyperplane A_k through a convex body K

Theorem 6.3. *Brunn-Minkowski inequality implies the classical n-dimensional isoperimetric inequality.*

Proof. We want to show that if $K \subseteq \mathbb{R}^n$ and B is the unit ball in $\mathbb{R}^n$, then

$$\left(\frac{v(K)}{v(B)}\right)^{\frac{1}{n}} \leq \left(\frac{S(K)}{S(B)}\right)^{\frac{1}{n-1}}$$

where S denotes the surface area. For a 2-dimensional plane, the LHS equals $\sqrt{A/\pi}$ while the RHS equals $L/(2\pi)$. To prove LHS $\geq$ RHS, we need $L^2 \geq 4\pi A$, which we know to be true. Let's try to generalize this to higher dimensions.

The surface area of K is, by definition,

$$S(K) = \lim_{\varepsilon \to 0} \frac{v(K + \varepsilon B) - v(K)}{\varepsilon}.$$

By Brunn-Minkowski inequality,

$$\begin{aligned} S(K) &\geq \lim_{\varepsilon \to 0} \frac{\left((v(K))^{\frac{1}{n}} + \varepsilon \, (v(B))^{\frac{1}{n}} \right)^n - v(K)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{n\varepsilon \, (v(K))^{\frac{n-1}{n}} \, (v(B))^{\frac{1}{n}} + O(\varepsilon^2)}{\varepsilon} \\ &= n \ (v(K))^{\frac{n-1}{n}} \, (v(B))^{\frac{1}{n}} \\ &= S(B) \ \left(\frac{v(K)}{v(B)}\right)^{\frac{n-1}{n}} \ \frac{n \, v(B)}{S(B)} \end{aligned}$$

The last term $n \, v(B)/S(B)$ is, however, always 1 in any number of dimensions. We have therefore proved the isoperimetric inequality. □

References

[1] J.H. van Lint. *Introduction to Coding Theory.* Springer, 1999.

[2] J. Matousek. *Lectures on Discrete Geometry.* Springer-Verlag, 2002.

Lecture 7

The Brunn-Minkowski Theorem and Influences of Boolean Variables

Friday 25, 2005
Lecturer: Nati Linial
Notes: Mukund Narasimhan

Theorem 7.1 (Brunn-Minkowski). *If $A, B \subseteq \mathbb{R}^n$ satisfy some mild assumptions (in particular, convexity suffices), then*

$$[\mathrm{vol}\,(A+B)]^{\frac{1}{n}} \geq [\mathrm{vol}\,(A)]^{\frac{1}{n}} + [\mathrm{vol}\,(B)]^{\frac{1}{n}}$$

where $A + B = \{\, a + b : a \in A$ and $b \in B\}$.

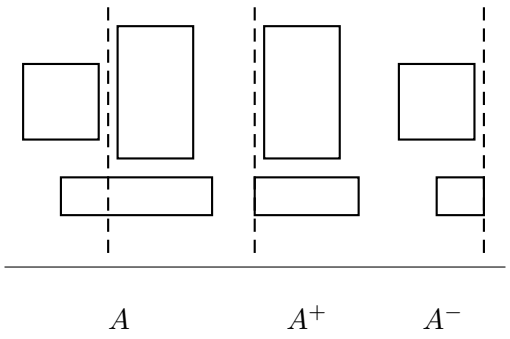
Proof. First, suppose that A and B are axis aligned boxes, say $A = \prod_{j=1}^{n} I_j$ and $B = \prod_{i=1}^{n} J_i$, where each I_j and J_i is an interval with $|I_j| = x_j$ and $|J_i| = y_i$. We may assume WLOG that $I_j = [0, x_j]$ and $J_i = [0, y_i]$ and hence $A + B = \prod_{i=1}^{n}[0, x_i + y_i]$. For this case, the BM inequality asserts that

$$\prod_{i=1}^{n}(x_i + y_i)^{\frac{1}{n}} \geq \prod_{i=1}^{n} x_i^{\frac{1}{n}} \cdot \prod_{i=1}^{n} y_i^{\frac{1}{n}}$$

$$\Leftrightarrow \qquad 1 \geq \left[\prod\left(\frac{x_i}{x_i + y_i}\right)\right]^{\frac{1}{n}} \cdot \left[\prod\left(\frac{y_i}{x_i + y_i}\right)\right]^{\frac{1}{n}}$$

Now, since the arithmetic mean of n numbers is bounded above by their harmonic mean, we have $(\prod \alpha_i)^{\frac{1}{n}} \leq \frac{\sum \alpha_i}{n}$ and $(\prod(1-\alpha_i))^{\frac{1}{n}} \leq \frac{\sum(1-\alpha_i)}{n}$. Taking $\alpha_i = \frac{x_i}{x_i+y_i}$ and hence $1 - \alpha_i = \frac{y_i}{x_i+y_i}$, we see that the above inequality always holds. Hence the BM inequality holds whenever A and B are axis aligned boxes.

Now, suppose that A and B are the disjoint union of axis aligned boxes. Suppose that $A = \bigcup_{\alpha \in \mathcal{A}} A_\alpha$ and $B = \bigcup_{\beta \in \mathcal{B}} B_\beta$. We proceed by induction on $|\mathcal{A}| + |\mathcal{B}|$. We may assume WLOG that $|\mathcal{A}| > 1$. Since the boxes are disjoint, there is a hyperplane separating two boxes in $\mathcal{A}$. We may assume WLOG that this hyperplane is $x_1 = 0$.



Let $A^+ = \{x \in A : x_1 \geq 0\}$ and $A^- = \{x \in A : x_1 \leq 0\}$ as shown in the figure above. It is clear that both A^+ and A^- are the disjoint union of axis aligned boxes. In fact, we may let $A^+ = \bigcup_{\alpha \in \mathcal{A}^+} A_\alpha$ and $A^- = \bigcup_{\alpha \in \mathcal{A}^-} A_\alpha$ where $|\mathcal{A}^+| < |\mathcal{A}|$ and $|\mathcal{A}^-| < |\mathcal{A}|$. Suppose that $\frac{\text{vol}\left(A^+\right)}{\text{vol}(A)} = \alpha$. Pick a λ so that

$$\frac{\text{vol}\left(\{x \in B : x_1 \geq \lambda\}\right)}{\text{vol}\left(B\right)} = \alpha$$

We can always do this by the mean value theorem because the function $f(\lambda) = \frac{\text{vol}(\{x \in B : x_1 \geq \lambda\})}{\text{vol}(B)}$ is continuous, and $f(\lambda) \to 0$ as $\lambda \to \infty$ and and $f(\lambda) \to 1$ as $\lambda \to -\infty$.

Let $B^+ = \{x \in B : x_1 \geq \lambda\}$ and $B^- = \{x \in B : x_1 \leq \lambda\}$. By induction, we may apply BM to both (A^+, B^+) and (A^-, B^-), obtaining

$$\left[\text{vol}\left(A^+ + B^+\right)\right]^{\frac{1}{n}} \geq \left[\text{vol}\left(A^+\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B^+\right)\right]^{\frac{1}{n}}$$
$$\left[\text{vol}\left(A^- + B^-\right)\right]^{\frac{1}{n}} \geq \left[\text{vol}\left(A^-\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B^-\right)\right]^{\frac{1}{n}}$$

Now,

$$\left[\text{vol}\left(A^+\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B^+\right)\right]^{\frac{1}{n}} = \alpha^{\frac{1}{n}}\left[\left[\text{vol}\left(A\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B\right)\right]^{\frac{1}{n}}\right]$$
$$\left[\text{vol}\left(A^-\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B^-\right)\right]^{\frac{1}{n}} = (1-\alpha)^{\frac{1}{n}}\left[\left[\text{vol}\left(A\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B\right)\right]^{\frac{1}{n}}\right]$$

Hence

$$\left[\text{vol}\left(A^+ + B^+\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(A^- + B^-\right)\right]^{\frac{1}{n}} \geq \left[\left[\text{vol}\left(A\right)\right]^{\frac{1}{n}} + \left[\text{vol}\left(B\right)\right]^{\frac{1}{n}}\right]$$

The general case follows by a limiting argument (without the analysis for the case where equality holds). $\square$

Suppose that $f : \mathbb{S}^1 \to \mathbb{R}$ is a mapping having a Lipshitz constant 1. Hence

$$\|f(x) - f(y)\| \leq \|x - y\|_2$$

Let μ be the median of f, so

$$\mu = \text{prob}\left[\{x \in \mathbb{S}^n : f(x) < \mu\}\right] = \frac{1}{2}$$

We assume that the probability distribution always admits such a μ (at least approximately). The following inequality holds for every $\epsilon > 0$ as a simple consequence of the isoperimetric inequality on the sphere.

$$\{\mathrm{x} \in \mathbb{S}^{\mathrm{n}} : |\mathrm{f} - \mu| > \epsilon\} < 2e^{-\epsilon n/2}$$

For $A \subseteq \mathbb{S}^n$ and for $\epsilon > 0$, let

$$A_\epsilon = \{x \in \mathbb{S}^n : \operatorname{dist} x, A < \epsilon\}$$

Question 7.1. Find a set $A \subseteq \mathbb{S}^n$ with $\mathrm{A} = a$ for which A_ϵ is the smallest.

The probability used here is the (normalized) Haar measure. The answer is always a spherical cap, and in particular if $a = \frac{1}{2}$, then the best A is the hemisphere (and so $A_\epsilon = \{x \in \mathbb{S}^n : x_1 < \epsilon\}$). We will show that for $A \subseteq \mathbb{S}^n$ with $\mathrm{A} = \frac{1}{2}$, $\mathrm{A}_\epsilon \geq 1 - 2e^{-\epsilon^2 n/4}$. If A is the hemisphere, then $\mathrm{A}_\epsilon = 1 - \Theta(e^{-\epsilon^2 n/2})$, and so the hemisphere is the best possible set.

But first, a small variation on BM :

$$\operatorname{vol}\left(\frac{A+B}{2}\right) \geq \sqrt{\operatorname{vol}(A) \cdot \operatorname{vol}(B)}$$

This follows from BM because

$$\begin{aligned}
\operatorname{vol}\left(\frac{A+B}{2}\right)^{\frac{1}{n}} &\geq \operatorname{vol}\left(\frac{A}{2}\right)^{\frac{1}{n}} + \operatorname{vol}\left(\frac{B}{2}\right)^{\frac{1}{n}} \\
&= \frac{1}{2}\left[\operatorname{vol}(A)^{\frac{1}{n}} + \operatorname{vol}(B)^{\frac{1}{n}}\right] \\
&\geq \sqrt{\operatorname{vol}(A)^{\frac{1}{n}} + \operatorname{vol}(B)^{\frac{1}{n}}}
\end{aligned}$$

For $A \subseteq \mathbb{S}^n$, let $\tilde{A} = \{\lambda a : a \in A, 1 \geq \lambda \geq 0\}$. Then $\mathrm{A} = \mu_{n+1}(\tilde{A})$. Let $B = \mathbb{S}^n \setminus A_\epsilon$.

Lemma 7.2. *If $\tilde{x} \in \tilde{A}$ and $\tilde{y} \in \tilde{B}$, then*

$$\left|\frac{\tilde{x} + \tilde{y}}{2}\right| \leq 1 - \frac{\epsilon^2}{8}$$

It follows that $\frac{\tilde{A}+\tilde{B}}{2}$ is contained in a ball of radius at most $1 - \frac{\epsilon^2}{8}$. Hence

$$\begin{aligned}
\left(1 - \frac{\epsilon^2}{8}\right)^{n+1} &\geq \operatorname{vol}\left(\frac{\tilde{A} + \tilde{B}}{2}\right) \\
&\geq \sqrt{\operatorname{vol}\left(\tilde{A}\right) \cdot \operatorname{vol}\left(\tilde{B}\right)} \\
&\geq \sqrt{\frac{\operatorname{vol}\left(\tilde{B}\right)}{2}}
\end{aligned}$$

Therefore, $2e^{-\epsilon^2 n/4} \geq \operatorname{vol}\left(\tilde{B}\right)$.

7.1 Boolean Influences

Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function. For a set $S \subseteq [n]$, the influence of S on f, $I_f(S)$ is defined as follows. When we pick $\{x_i\}_{i \notin S}$ uniformly at random, three things can happen.

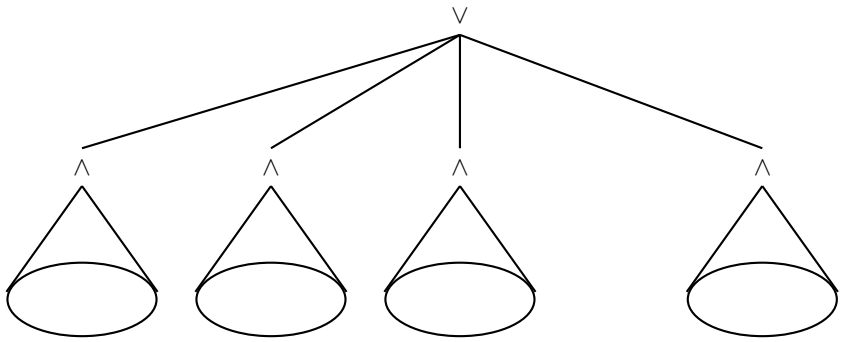
1. $f = 0$ regardless of $\{x_i\}_{i \in S}$ (suppose that this happens with probability q_0).
2. $f = 1$ regardless of $\{x_i\}_{i \in S}$ (suppose that this happens with probability q_1).
3. With probability $\mathrm{Inf}_f(S) := 1 - q_0 - q_1$, f is still undetermined.

Some examples:

- (Dictatorship) $f(x_1, x_2, \ldots, x_n) = x_1$. In this case
$$\mathrm{Inf}_{\mathsf{dictatorship}}(S) = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{if } i \notin S \end{cases}$$
- (Majority) For $n = 2k+1$, $f(x_1, x_2, \ldots, x_n)$ is 1 if and only if a majority of the x_i are 1. For example, if $S = \{1\}$,
$$\begin{aligned} \mathrm{Inf}_{\mathsf{majority}}(\{1\}) &= \mathrm{prob}\,(x_1 \text{ is the tie breaker }) \\ &= \frac{\binom{2k}{k}}{2^{2k}} = \Theta\left(\frac{1}{\sqrt{k}}\right) \end{aligned}$$
For fairly small sets S,
$$\mathrm{Inf}_{\mathsf{majority}}(S) = \Theta\left(\frac{|S|}{\sqrt{n}}\right)$$
- (Parity) $f(x_1, x_2, \ldots, x_n) = 1$ if and only if an even number of the x_i's are 1. In this case
$$\mathrm{Inf}_{\mathsf{parity}}(\{x_i\}) = 1$$
for every $1 \leq i \leq n$.

Question 7.2. What is the smallest $\delta = \delta(n)$ such that there exists a function $f : \{0,1\}^n \to \{0,1\}$ which is balanced (i.e., $Ef = \frac{1}{2}$) for which $\mathrm{Inf}_f(\{x_i\}) < \delta$ for all x_i ?

Consider the following example, called tribes. The set of inputs $\{x_1, x_2, \ldots, x_n\}$ is partitioned into tribes of size b each. Here, $f(x_1, x_2, \ldots, x_n) = 1$ if and only if there is a tribe that unanimously 1.



Since we want $Ef = \frac{1}{2}$, we must have $\text{prob}(f = 0) = \left(1 - \frac{1}{2^b}\right)^{\frac{n}{b}} = \frac{1}{2}$. Therefore, $\frac{n}{b} \ln\left(1 - \frac{1}{2^b}\right) = -\ln 2$. We use the Taylor series expansion for $\ln(1 - \epsilon) = -\epsilon - \epsilon^2/2 - \cdots = -\epsilon - O(\epsilon^2)$ to get $\frac{n}{b}\left(\frac{1}{2^b} + O\left(\frac{1}{4^b}\right)\right) = -\ln 2$. This yields $n = b \cdot 2^b \ln 2\,(1 + O(1))$. Hence $b = \log_2 n - \log_2 \ln n + \Theta(1)$.

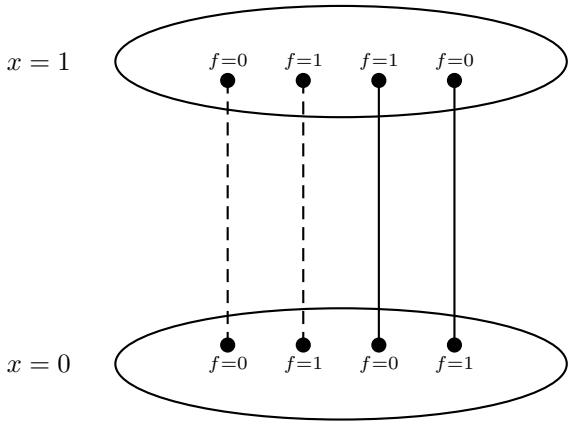
Hence,

$$\begin{aligned}
\text{Inf}_{\text{tribes}}(x) &= \left(1 - \frac{1}{2^b}\right)^{\frac{n/b}{-}1} \cdot \left(\frac{1}{2}\right)^{b-1} \\
&= \frac{\left(1 - \frac{1}{2^b}\right)^{\frac{n}{b}}}{1 - \frac{1}{2^b}} \cdot \frac{1}{2^{b-1}} \\
&= \frac{1}{1 - \frac{1}{2^b}} \cdot \frac{1}{2^b} \\
&= \frac{1}{2^{b-1}} = \Theta\left(\frac{\log b}{n}\right)
\end{aligned}$$

In this example, each individual variable has influence $\Theta(\log n/n)$. It was later shown that this is lowest possible influence.

Proposition 7.3. *If $Ef = \frac{1}{2}$, then $\sum_x \text{Inf}_f(x) \geq 1$.*

This is a special case of the edge isoperimetric inequality for the cube, and the inequality is tight if f is dictatorship.



The variable x is influential in the cases indicated by the solid lines, and hence

$$\text{Inf}_f(x) = \frac{\#\text{ of mixed edges}}{2^{n-1}}$$

Let $S = f^{-1}(0)$. Then $\sum \text{Inf}_f(x) = \frac{1}{2^{n-1}} e(S, S^c)$.

One can use $\hat{f}$ to compute influences. For example, if f is monotone (so $x \prec y \Rightarrow f(x) \leq f(y)$), then

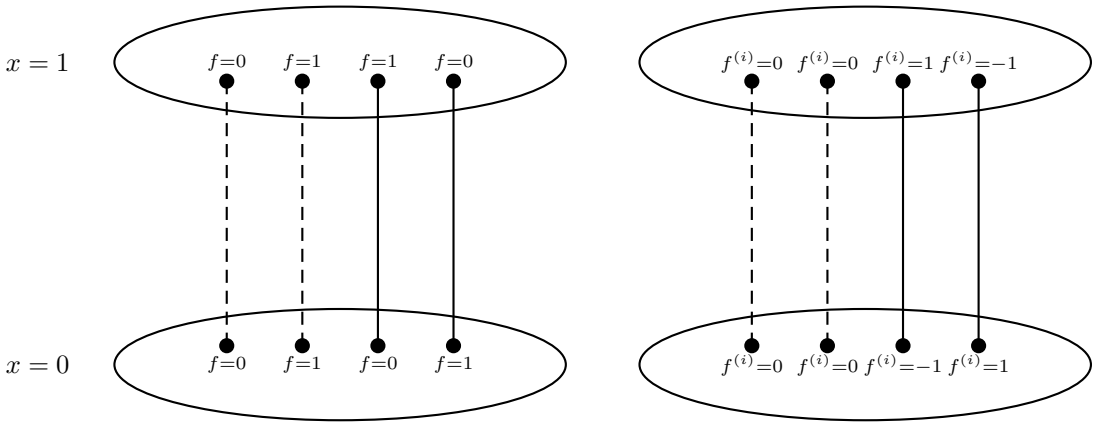
$$\hat{f}(S) = \sum_T \frac{(-1)^{|S \cap T|}}{2^n}$$

Therefore,

$$\begin{aligned}
\hat{f}(\{i\}) &= \frac{1}{2^n} \sum_{i \notin T} f(T) - \frac{1}{2^n} \sum_{i \in T} f(T) \\
&= \frac{1}{2^n} \sum_{i \notin T} (f(T) - f(T \cup \{i\})) \\
&= \frac{-1}{2^n} \cdot \# \text{ mixed edges in the direction of } i \\
&= -\frac{1}{2} \mathrm{Inf}_f(x_i)
\end{aligned}$$

Hence $\mathrm{Inf}_f(x_i) = -2\hat{f}(\{i\})$. What can be done to express $\mathrm{Inf}_f(x)$ for a general f? Define

$$f^{(i)}(z) = f(z) - f(z \oplus e_i)$$



Then

$$\mathrm{Inf}_f(x_i) = \left|\text{support} f^{(i)}\right| = \sum_w \left(f^{(i)}(w)\right)^2$$

The last term will be evaluated using Parseval. For this, we need to compute the Fourier expression of $f^{(i)}$ (expressed in terms of $\hat{f}$).

$$\begin{aligned}
\widehat{f^{(i)}}(S) &= \frac{1}{2^n}\sum_T f^{(i)}(T)(-1)^{|S\cap T|} \\
&= \frac{1}{2^n}\sum_T \Big[f(T) - f(T\oplus\{i\})\Big](-1)^{|S\cap T|} \\
&= \frac{1}{2^n}\sum_{i\notin T} \Big(\Big[f(T) - f(T\cup\{i\}\Big](-1)^{|S\cap T|} + \Big[f(T\cup\{i\}) - f(T)\Big](-1)^{|S\cap(T\cup\{i\})|}\Big) \\
&= \frac{1}{2^n}\sum_{i\notin T} \Big[f(T) - f(T\cup\{i\}\Big]\left((-1)^{|S\cap T|} - (-1)^{|S\cap(T\cup\{i\})|}\right) \\
&= \begin{cases} 0 & \text{if } i\notin S \\ 2\hat{f}(S) & \text{if } i\in S \end{cases}
\end{aligned}$$

Using Parseval on $\widehat{f^{(i)}}$ along with the fact that $\widehat{f^{(i)}}$ takes on only values $\{0,\pm 1\}$, we conclude that

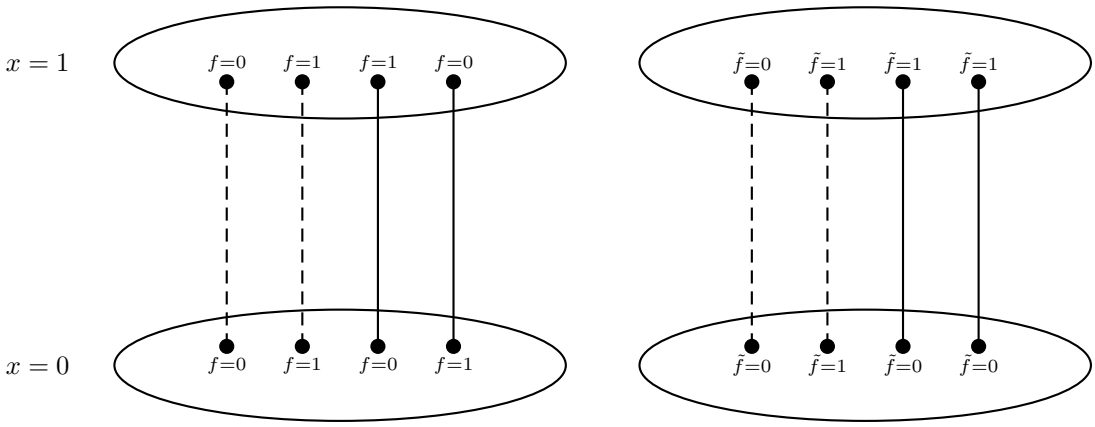
$$\mathrm{Inf}_f(x_i) = 4\sum_{i\in S} |hat f(S)|^2$$

Next time, we will show that if $Ef = \frac{1}{2}$, then there exists a i such that $\sum_{i\in S}\left(\hat{f}(S)\right)^2 > \Omega(\ln n/n)$.

Lemma 7.4. *For every $f:\{0,1\}^n\to\{0,1\}$, there is a monotone $g:\{0,1\}^n\to\{0,1\}$ such that*

- $Eg = Ef$.
- *For every $s\subseteq[n]$, $\mathrm{Inf}_g(S) \le \mathrm{Inf}_f(S)$.*

Proof. We use a shifting argument.



Clearly $E\tilde{f} = Ef$. We will show that for all S, $\mathrm{Inf}_{\tilde{f}}(S) \le \mathrm{Inf}_f(S)$. We may keep repeating the shifting step until we obtain a monotone function g. It is clear that the process will terminate by considering the progress measure $\sum f(x)\,|x|$ which is strictly increasing. Therefore, we only need show that $\mathrm{Inf}_{\tilde{f}}(()\,S) \le \mathrm{Inf}_f(S)$.

□

Lecture 8

More on the influence of variables on boolean functions

March 4, 2005
Lecturer: Nati Linial
Notes: Neva Cherniavsky & Atri Rudra

In this lecture we will look at the following natural question– do there exist balanced boolean functions f on n variables such that for every variable x the influence of x on f is "small" and how small can this bound be made? ("Balanced" means that $Pr(f = 0) = Pr(f = 1) = \frac{1}{2}$ but $\mathbb{E}f = \alpha$ for some α bounded away from 0 and 1 is just as interesting.) In the last lecture we showed that for the "tribes" function (which was defined by Ben-Or and Linial in [1]), every variable has influence $\Theta(\frac{\log n}{n})$. Today, we will prove the result of Kahn, Kalai and Linial [2] which shows that this quantity is indeed the best one can hope for. In the process we will look into the Bonami Beckner Inequality and will also look at threshold phenomena in random graphs.

8.1 The Kahn Kalai Linial Result

Recall the definition of influence. Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function and let $S \subset [n]$. The influence of the set of variables S on the function f, denoted by $Inf_f(S)$, is the probability that f is still undetermined when all variables in $[n] - S$ have been assigned values at random.

We also talk about influences in the case when the function is defined on a solid cube– $f : [0,1]^n \to \{0,1\}$. This formulation has connections to game theory– variables are controlled by the players. Note that in this case we can talk about things like "influence of a subset of variables towards 0".

The situation for the case when $|S| = 1$ is relatively better understood. As we saw in the last class, the situation for calculating $Inf_f(x)$ looks like Figure 8.1. In particular, calculating the influence is same as counting the number of non-solid edges in Figure 8.1.

The situation is much less understood for the more general case, for example when $|S| = 2$. In a nutshell, we are interested in situations other than those in Figure 8.2. This scenario is not well understood and is still a mysterious object. Unlike the case of a single variable, the number of zeroes and ones in a "mixed" 2-dimensional subcube can vary and the whole situation is consequently more complex. As an interesting special case consider the "tribes" example and the case when S is an entire tribe. It is easy to see that the influence of S towards 1 is exactly 1 (as any tribe can force the result to be 1) while it is not hard to see that the influence of S towards 0 is only $O(\frac{\log n}{n})$. As another example consider the case when S consists of one element from each tribe (one can consider each element of S as a "spy" in a tribe). Here the influence of S

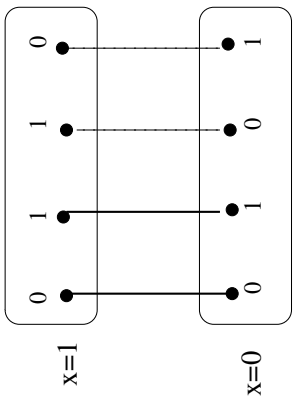


Figure 8.1: Influence of a variable

towards 0 is exactly 1 (as each spy can force its tribe to be 0). Further, its influence towards 1 can be shown to be $\frac{\sqrt{2}-1}{4} + o(1)$.

Let us now turn back to the motivating question for this lecture–

Question 8.1. Find boolean functions f with $\mathbb{E}f = \frac{1}{2}$ for which $Inf_f(x)$ is small for each variable x.

For any variable x_i define $\beta_i = Inf_f(x_i)$ and let $\boldsymbol{\beta} = \langle \beta_1, \cdots, \beta_n \rangle$. Thus, the quantity we are interested in is $||\boldsymbol{\beta}||_\infty$. Note that the edge isoperimetric inequality on the cube implies that $\sum_{i=1}^n \beta_i \geq 1$ which by an averaging argument gives $||\boldsymbol{\beta}||_\infty \geq \frac{1}{n}$. Also note that for the "tribes" example, $||\boldsymbol{\beta}||_\infty = \Theta(\frac{\log n}{n})$. The following result due to Kahn, Kalai and Linial shows that this is the best possible–

Theorem 8.1. *For any $f : \{-1,1\} \to \{-1,1\}$ with $\mathbb{E}f = 0$, there exists a variable x such that $Inf_f(x) \geq \Omega(\frac{\log n}{n})$.*

Before we start to prove the theorem, let us collect some facts that we know or follow trivially from what we have covered in previous lectures.

$$\sum_{S \subseteq [n]} (\hat{f}(S))^2 = 1 \tag{8.1}$$

$$\hat{f}(\emptyset) = 0 \tag{8.2}$$

$$\beta_i = 4 \cdot \sum_{S \ni i} (\hat{f}(S))^2 \tag{8.3}$$

$$\sum_{i=1}^n \beta_i = 4 \cdot \sum_{S \subseteq [n]} |S| (\hat{f}(S))^2 \tag{8.4}$$

Equation (8.1) follows from Parseval's identity and the fact that f takes values in $\{-1,1\}$. Equation (8.2) follows from the fact that $\chi_\emptyset = 1$ which implies $\hat{f}(\emptyset) = \langle f, \chi_\emptyset \rangle = 2^n \mathbb{E}f = 0$. Equation (8.3)was proved in the last lecture. Equation (8.4) follows from summing Equation (8.3) for all i.

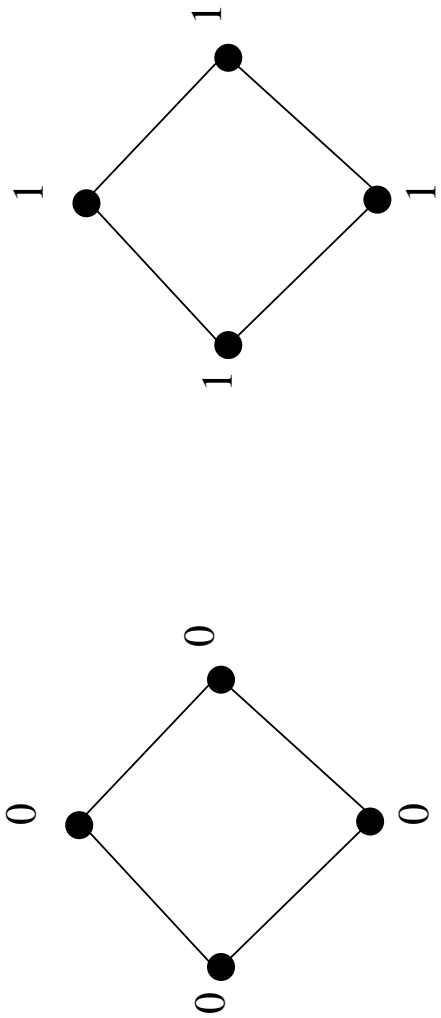


Figure 8.2: Influence of a general subset of variables

We will first show that if "most" of the $\sum_{S\subseteq[n]}(\hat{f}(S))^2$ comes from 'large' sets S then the conclusion of Theorem 8.1 holds. In fact, in this case even the *average* influence is $\Omega(\frac{\log n}{n})$. To be more precise let $T = \frac{\log n}{10}$ and $H = \sum_{|S|\geq T}(\hat{f}(S))^2$. Further assume that $H \geq \frac{1}{10}$. Then by (8.4), $\sum_{i=1}^{n}\beta_i \geq 4\cdot\sum_{|S|\geq T}|S|(\hat{f}(S))^2 \geq 4HT \geq \frac{\log n}{25}$.

It follows that it suffices to prove the theorem under the complementary assumption that $H < \frac{1}{10}$. In view of (8.1) this is the same as showing $\sum_{|S|<T}(\hat{f}(S))^2 > 0.9$. In the proof we will need to estimate the following quantity–

$$\sum_{|S|<T}(\hat{\phi}(S))^2 \equiv \sum_{S} W_T(S)(\hat{\phi}(S))^2$$

for $\phi = f^{(i)}$ (recall from the last lecture that $f^{(i)}(z) = f(z) - f(z\oplus e_i)$). Here $W_T(\cdot)$ is the step function which takes value 1 for any set $S\subseteq[n]$ such that $|S|\leq T$ and 0 otherwise. We use two ideas to solve the problem in hand–

- Try and approximate $W(\cdot)$ by functions which are sufficiently close to the step function $W_T(\cdot)$ and are easy to analyze.
- As the bound $||\boldsymbol{\beta}||_1 \geq 1$ is tight (which is insufficient for our purposes) and since it is difficult to work directly with $||\boldsymbol{\beta}||_\infty$, we could perhaps estimate $||\boldsymbol{\beta}||_p$ for some $\infty > p > 1$. In the proof we will use $p = \frac{4}{3}$ but this is quite arbitrary.

We focus on the second alternative and use the Bonami Beckner inequality which we consider in the next section.

8.2 Bonami Beckner Inequality

Let T_ϵ be a linear operator which maps real functions on $\{-1,1\}^n$ to real functions on $\{-1,1\}^n$. By linear, we mean that the following holds for functions f and g and scalars a and b: $T_\epsilon(a \cdot f + b \cdot g) = a \cdot T_\epsilon(f) + b \cdot T_\epsilon(g)$.

As $T_\epsilon(\cdot)$ is a linear operator, one can fully determine it by just specifying it at the basis of the functions $\{\chi_S\}$. We define the operator as follows

$$T_\epsilon(\chi_S) = \epsilon^{|S|}\chi_S$$

By linearity, $T_\epsilon(f) = \sum_{S\subseteq[n]} \epsilon^{|S|}\hat{f}(S)\chi_S(\cdot)$. Note that $T_1(f) = f$. In other words, T_1 is the identity operator.

We will now state the main result of this section–

Theorem 8.2. *Let $0 < \epsilon < 1$ and consider T_ϵ as an operator from L_p to L_2 where $p = 1 + \epsilon^2$. Then its operator norm is* 1.[1]

Let us first explain the terms used in the statement above. Let $T : (X, ||\cdot||_X) \to (Y, ||\cdot||_Y)$ be a linear operator– here X and Y are normed spaces and $||\cdot||_X$ and $||\cdot||_Y$ are their respective norms. The operator norm of T is defined as

$$||T||_{op} = sup_{x\in X}\frac{||Tx||_Y}{||x||_X}$$

This quantity measures how much the "length" (norm) of an element $x \in X$ can grow by an application of the operator T. We now turn to the proof.

What is the size of f? How expanding is the operator? These are very narrow passages; we have no wiggle room. We can only use Parseval's in L_2, so the norm on the right hand side needs to be L_2. On the left hand side, our norm is L_p, which is usually very difficult to calculate. But because our functions (the $f^{(i)}$) only take on the values $\{-1, 0, 1\}$, we can calculate the necessary L_p norms.

That the operator norm of T_ϵ is at least 1 is obvious. Let f be identically 1 everywhere. Then $\hat{f}(T) = \sum f(S)(-1)^{|S\cap T|} = 0$ for $T \neq \emptyset$ and $\hat{f}(\emptyset) = 1$. So $||T_\epsilon f||_2 = 1 = ||f||_p$. What the Bonami-Beckner inequality says is that for every $f : \{-1,1\}^n \to \mathbb{R}$, $||T_\epsilon f||_2 \leq ||f||_p$.

We'll do a part of the proof only. The proof is via induction on n, the dimension of the cube. The base case is the main part of the proof; the method for inducting is standard in analysis and we'll skip it.

[1] This is called a *hypercontractive inequality*.

Again, the surprising thing here is that $p < 2$ and still $||T||_{op} \leq 1$.

For $n = 1$, every function f has the form $f = a + bx$ and then $T_\epsilon f = a + \epsilon bx$. (There are only two characters, $\chi_\emptyset$ and $\chi_{\{x\}}$). Then at $x = -1$, $f = a - b$ and $T_\epsilon f = a - \epsilon b$. At $x = 1$, $f = a + b$ and $T_\epsilon f = a + \epsilon b$. So

$$\begin{aligned} ||f||_p &= \left[\frac{|a+b|^p + |a-b|^p}{2}\right]^{\frac{1}{p}} \\ ||T_\epsilon f||_2 &= \sqrt{\frac{(a+\epsilon b)^2 + (a-\epsilon b)^2}{2}} \\ &= \sqrt{a^2 + \epsilon^2 b^2} \end{aligned}$$

We want to prove $||T_\epsilon f||_2 \leq ||f||_p$, i.e., we want to prove

$$\left[\frac{|a+b|^p + |a-b|^p}{2}\right]^{\frac{1}{p}} \geq \sqrt{a^2 + \epsilon^2 b^2}. \tag{8.5}$$

Suppose $|a| \geq |b|$. Let $b = ta$ and divide by $|a|$:

$$\begin{aligned} \left(\frac{|a+b|^p + |a-b|^p}{2}\right)^{\frac{1}{p}} &= \left(\frac{|a+ta|^p + |a-ta|^p}{2}\right)^{\frac{1}{p}} \\ &= |a|\left(\frac{|1+t|^p + |1-t|^p}{2}\right)^{\frac{1}{p}} \\ \sqrt{a^2 + \epsilon^2 b^2} &= \sqrt{a^2 + \epsilon^2 (at)^2} \\ &= |a|\sqrt{1 + \epsilon^2 t^2} \end{aligned}$$

So we will prove

$$\frac{|1+t|^p + |1-t|^p}{2} \geq (1 + \epsilon^2 t^2)^{\frac{p}{2}} \text{ when } |t| \leq 1 \tag{8.6}$$

and (8.5) will follow. Note that if $|a| < |b|$, we'd substitute $a = bt$ and divide by $|b|$, and would want to prove

$$\frac{|t+1|^p + |t-1|^p}{2} \geq (\epsilon^2 + t^2)^{\frac{p}{2}} \text{ when } |t| \leq 1 \tag{8.7}$$

But since

$$(1 + \epsilon^2 t^2) \geq \epsilon^2 + t^2,$$

once we prove equation (8.6), (8.7) will follow immediately.

Proof of (8.6) is via the Taylor expansion. For the left hand side, terms in odd places will cancel out, and terms in even places will double. Recall $p = 1 + \epsilon^2$ and $|t| \leq 1$. The left hand side becomes

$$\sum_{j=0}^{\infty} t^{2j} \begin{pmatrix} p \\ 2j \end{pmatrix}$$

and the right hand side becomes

$$\sum_{j=0}^{\infty} \epsilon^{2j} t^{2j} \begin{pmatrix} p/2 \\ j \end{pmatrix}.$$

Let's examine the first two terms of each side. On both sides, the first term is 1. On the left hand side, the second term is $t^2(p(p-1))/2$ and on the right hand side the second term is $\epsilon^2 t^2 p/2$; since $\epsilon^2 = p-1$, this means the second terms are also equal on both sides. Therefore it is sufficient to compare the terms from $j \geq 2$.

What we discover is that on the left hand side, all terms are positive, whereas on the right hand side, the $j = 2k$ and $j = 2k+1$ terms have a negative sum for all $k \geq 1$. First we show the left hand side is positive. The $(2j)$th coefficient equals $p(p-1)(p-2)\ldots(p-2j+1)$ divided by some positive constant. Note that $p(p-1)$ is positive and all the terms $(p-2)\ldots(p-2j+1)$ are negative. But since there are an even number of these negative terms, the product as a whole is positive. Therefore, on the left hand side, all terms are positive.

Now consider the right hand side. We will show that the $j = 2k$ and $j = 2k+1$ terms have a negative sum for all $k \geq 1$. Consider the sum

$$\epsilon^{4k} t^{4k} \begin{pmatrix} p/2 \\ 2k \end{pmatrix} + \epsilon^{4k+2} t^{4k+2} \begin{pmatrix} p/2 \\ 2k+1 \end{pmatrix}.$$

We can divide out $\epsilon^{4k} t^{4k}$ without affecting the sign. Since the second term is the positive one, and $|t| \leq 1$, we can replace t^2 by 1 without loss of generality. So now we have

$$\begin{aligned}
\begin{pmatrix} p/2 \\ 2k \end{pmatrix} + \epsilon^2 \begin{pmatrix} p/2 \\ 2k+1 \end{pmatrix} &= \begin{pmatrix} p/2 \\ 2k \end{pmatrix} + (p-1) \begin{pmatrix} p/2 \\ 2k+1 \end{pmatrix} \\
&= \frac{p/2(p/2-1)\ldots(p/2-2k+1)}{2k!} + (p-1)\frac{p/2(p/2-1)\ldots(p/2-2k)}{(2k+1)!} \\
&= \left[(2k+1)\frac{p}{2}(\frac{p}{2}-1)\ldots(\frac{p}{2}-2k+1) + (p-1)\frac{p}{2}(\frac{p}{2}-1)\ldots(\frac{p}{2}-2k)\right]/(2k+1)! \\
&= \left[\frac{p}{2}(\frac{p}{2}-1)\ldots(\frac{p}{2}-2k+1)\right]\left[2k+1+(p-1)(\frac{p}{2}-2k)\right]/(2k+1)!
\end{aligned}$$

Notice that the first term in brackets is negative and the second term in brackets is positive. Thus the sum of the kth even and odd term is negative for all k, and we've proved equation (8.6). Equation (8.6) implies equation (8.7); equations (8.6) and (8.7) together imply equation (8.5); and equation (8.5) implies $||f||_p \geq ||T_\epsilon f||_2$ for $p = 1+\epsilon^2$. Thus we've proved the base case for the Bonami-Beckner inequality.

8.3 Back to Kahn Kalai Linial

In general it is not obvious how to utilize this inequality, since it's hard to compute the p-norm. But we're looking at an easy case. Specifically, if $g : \{0,1\}^n \to \{-1,0,1\}$, the inequality says that

$$\Pr(g \neq 0) = t \Rightarrow t^{\frac{2}{1+\delta}} \geq \sum_S \delta^{|S|}(\hat{g}(S))^2 \tag{8.8}$$

To see this, let $\delta = \epsilon^2$. We know $||g||_p \geq ||T_\epsilon g||_2$.

$$||g||_p = \left(\frac{1}{2^n}\sum |g(x)|^p\right)^{\frac{1}{p}} = \left(\frac{1}{2^n}\sum_{g(x)\neq 0} 1\right)^{\frac{1}{p}} = t^{\frac{1}{p}} = t^{\frac{1}{1+\epsilon^2}}$$

So the Bonami-Beckner inequality tells us

$$t^{\frac{1}{1+\epsilon^2}} \geq \sqrt{\sum_S \epsilon^{2|S|}(\hat{g}(S))^2}$$

and squaring both sides and substituting δ gives us equation (8.8).

We applied this inequality for $g = f^{(i)}$. Then $t = \beta_i$, the influence of the ith variable, which is exactly what we're looking for. Recall that

$$\widehat{f^{(i)}(S)} = \begin{cases} 0 & i \notin S \\ \hat{f}(S) & i \in S \end{cases}$$

We want to prove that $\max \beta_i \geq \Omega(\log n/n)$. By substituting the new values into (8.8), we get

$$\beta_i^{\frac{2}{1+\delta}} \geq \sum_{S\ni i} \delta^{|S|}(\hat{f}(S))^2$$

The $\delta^{|S|}$ terms are the W_S's that we want to behave sufficiently close to the step function. Recall, we also know that $\sum_{0\leq|S|<T}(\hat{f}(S))^2 > 0.9$ where $T = \log n/10$. Since we assume f to be balanced (i.e. $\Pr(f = 1) = 1/2$), we can ignore the 0 term because $\hat{f}(S) = 0$ when $S = \emptyset$, but we cannot ignore it for imbalanced functions. Still, it won't matter, and we'll come back to this point later.

So we know $\sum_{0<|S|<T}(\hat{f}(S))^2 > 0.9$. Let $\delta = 1/2$ (the choice is arbitrary).

$$\begin{aligned}
n \cdot \max \beta_i^{\frac{4}{3}} &\geq \sum_i \beta_i^{\frac{4}{3}} \\
&\geq \sum_i \sum_{S\ni i} (\frac{1}{2})^{|S|}(\hat{f}(S))^2 \\
&= \sum_S \left(\frac{1}{2}\right)^{|S|} |S|(\hat{f}(S))^2 \\
&\geq \sum_{|S|<T} \left(\frac{1}{2}\right)^{|S|} |S|(\hat{f}(S))^2 \\
&\geq \left(\frac{1}{2}\right)^T \sum_{|S|<T} (\hat{f}(S))^2 \\
&\geq (n^{-\frac{1}{10}})(0.9)
\end{aligned}$$

Therefore, $n \cdot \max \beta_i^{4/3} \geq 0.9/n^{1/10}$ and so

$$\max \beta_i \geq \left(\frac{c}{n^{11/10}}\right)^{3/4} = \Omega(n^{-33/40}) >> \frac{\log n}{n}$$

There is some progress in understanding what the vector of influences is in general, but we have a long way to go.

We return to an issue we had left open. What happens when we deal with functions that are imbalanced? What if $f : \{0,1\}^n \to \{0,1\}, \mathbb{E}f = p$ (not necessarily $p = 1/2$). Now we cannot ignore $\hat{f}(\emptyset)$ in the previous argument.

Indeed, $f(\emptyset) = p$. $\sum(\hat{f}(S))^2 = p$. We'd have to subtract $(\hat{f}(\emptyset))^2 = p^2$ off of p in general. But this is fine as long as $0 < p < 1$. We mention this because this technique is used often.

8.4 Sharp thresholds in graphs

Theorem 8.3. *Every monotone graph property has a sharp threshold.*

This theorem is also the title of the paper, by Friedgut and Kalai. The background is that in the late 1950s, Erdős and Renyi defined random graphs, which became an important ingredient for many investigations in modern discrete mathematics and theoretical computer science. A random graph $G(n,p)$ is a graph on n vertices in which the probability that (x,y) is an edge in G is p. That is, for each pair of vertices, flip a p-weighted coin, and put an edge in the graph if the coin comes up heads. We do this independently for each pair of vertices. So

$$Pr(G) = p^{e(G)}(1-p)^{\binom{n}{2}-e(G)}$$

Already Erdős and Renyi had noticed that some properties of graphs have special behavior. For example, take the property that G is connected. Write $Pr(G \text{ is connected})$ as $f(p)$. When p is small, the graph is unlikely to be connected; when p is big, it is almost certain to be connected. The shape of $f(p) = Pr(G \text{ is connected})$ is shown in Figure 8.3.

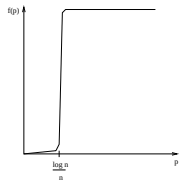


Figure 8.3: Sharp threshold

The transition from disconnected to connected is very rapid. This is related to a class of physical phenomena called phase transition; there is some physical system that is dependent on a single parameter, such as temperature or pressure, and the system goes from one phase to the other very sharply (for instance, the system goes from not magnetized to magnetized).

There are other graph properties that also exhibit this behavior (e.g Hamiltonian cycle, planarity, etc). But there was no satisfactory general theorem until Freidgut-Kalai.

To have a precise form of the theorem, we define the terms:

Definition 8.1. A *graph property* is a class of labeled graphs that is closed under permutations of the vertex labels.

Intuitively, a graph property holds or does not hold regardless of the labeling of the vertices, such as connectedness, "the graph contains a triangle", the graph is 17-colorable, etc.

Definition 8.2. A graph property is called *monotone* if it continues to hold after adding more edges to the graph.

Again, connectedness is monotone; non-planarity is also. An example of a graph property which is non-monotone is "the number of edges in the graph is even".

Let A be a monotone graph property.

$$\mu_p(A) = Pr(A|G \in_R G(n,p))$$

where G is sampled randomly. Clearly $\mu_p(A)$ is an increasing function of p. The theorem says that p_0, where $\mu_{p_0}(A) = \epsilon$, and p_1, where $\mu_{p_1}(A) = 1 - \epsilon$ are very close. Namely,

$$p_1 - p_0 \leq O\left(\frac{\log(1/\epsilon)}{\log n}\right)$$

In the connectivity example, this is not very interesting. The transition from almost surely disconnected to almost surely connected takes place around $p = \log n/n$, and $1/\log n$ is much bigger. In other words, the gap is much bigger than the critical value where the threshold occurs.

Later, Bourgain and Kalai showed that for all monotone graph properties, the same bound holds with $p_1 - p_0 \leq O\left(\frac{1}{\log^{2-\gamma} n}\right)$, for every $\gamma > 0$ and n large enough. This theorem is nearly tight, since there exist examples of monotone graph properties where the width of the gap is $\Theta\left(\frac{1}{\log^2 n}\right)$. For instance, "$G$ contains a k-clique" for specific $k = \Theta(\log n)$ where the critical $p = 1/2$.

So what can we do about the problem in the connectivity example, where the threshold comes at a point much smaller than the gap? We can ask a tougher question: is it true that the transition from $\mu_{p_0} = \epsilon$ to $\mu_{p_1} = 1 - \epsilon$ occurs between $(1 \pm o(1))q$, where q is the critical probability (i.e. $\mu_q(A) = 1/2$)?

If the answer were yes, we'd have a more satisfactory gap in relation to our critical value in the connectivity case. However, the answer is negative for certain graph properties: this is asking too much. For example, suppose the property is that G contains a triangle. Here the critical $p = c/n$. The expected number of triangles in G is

$$\binom{n}{3} p^3 = \frac{c^3}{6}(1 - \frac{1}{n})(1 - \frac{2}{n})$$

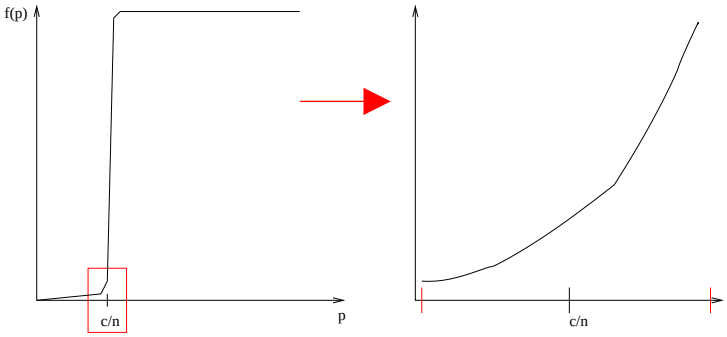


Figure 8.4: G contains a triangle

The picture looks like Figure 8.4. At the smaller scale, the threshold isn't as sharp. As we vary c in $p = c/n$, the probability that G contains no triangle changes from one constant to another constant, both bounded away from zero and from one. The reason behind this picture is the number of triangles is a random variable with a Poisson distribution. Therefore, the probability there is no triangle is $Pr(X = 0) = e^{-\mu}$, where μ is the expectation of X.

One of Freidgut's major contributions was to characterize which graph properties have a sharp threshold and which don't in this stronger sense. It's a little complicated to even state his result precisely, but the spirit of the theorem is this: properties like "G contains a triangle" are considered "local". Friedgut's theorem says that "a graph property has a sharp threshold (in the strong sense)" is roughly equivalent to the property being "non-local".

References

[1] M. Ben-Naor and N. Linial. Collective Coin Flipping. In *Randomness and Computation* S. Micali ed. Academic press, new York, 1990.

[2] J. Kahn, G. Kalai and N. Linial The influence of variables on boolean function. In *Proceedings of the 29th Symposium on the Foundations of Computer Science*, White Planes, 1988, Pages 68–80.

Lecture 9

Threshold Phenomena

March 7, 2005
Lecturer: Nati Linial
Notes: Chris Ré

9.1 Asides

There is a good survey of this area by Gil Kalai, Muli Safra called *Threshold Phenomena and Influence* due out very soon.

9.1.1 Percolation

Though our main technical result concerns random graphs in the G(n,p) model, let us mention other contexts in which threshold phenomena occur. One classical example is *Percolation*, an area started in physics. A typical question here is this: given a planar grid and $0 < p < 1$. Create a graph by keeping each edge of the planar grid with probability p and removing each edge with probability 1-p. The inclusion of edges is done independently. Our question is then: In the resulting graph is the origin in an infinite connected component?

It turns out that there is a critical probability, p_c, such that

$p < p_c$	with probability 1, the origin is not in an infinite component
$p > p_c$	with probability > 0, the origin is in an infinite component

You can imagine considering other similiar questions on higher dimensional grids. For the planar grids it turns out that $p_c = \frac{1}{2}$.

This problem comes up in mining in the following idealized model. Somewhere underground is a deposit of oil. It is surrounded by rocks whose structure is that of a 'random sponge', a solid with randomly placed cavities. The question is how far the oil is likely to flow away from its original location. Percolation in a 3-dimensional setting is a good abstraction of the above physical situation.

Now imagine graphing the probability of the property holding versus the p value from above. As an example see figure 9.1. The interesting questions are how does it behave around or slightly to the right of p_c. For example is this a smooth function? Is it differentiable? How large is its derivative? Figure 9.1 illustrates some curves that could happen. In this example, the property could be discontinuous at p_c or is continuous but not smooth at p_c.

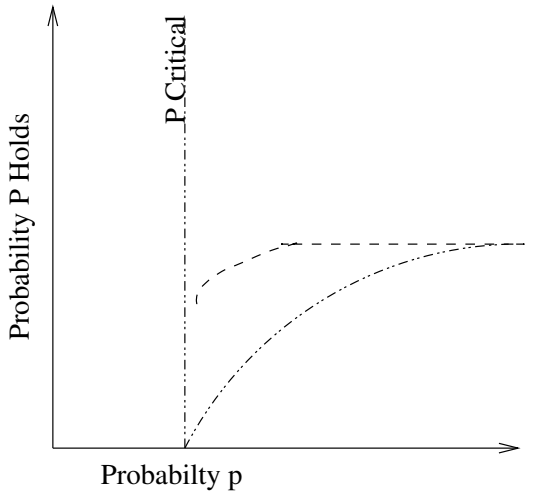


Figure 9.1: Probability of Property vs. p

9.2 Monotone Graph Properties

The main theorem we want to prove is:

Theorem 9.1 (Friedgut and Kalai). *Every monotone graph property has a sharp threshold*

To make this precise, we need some definitions. Let P be a graph property, that is a property invariant under vertex relabeling. A property P is *monotone* if $P(G_0)$ implies that $P(G)$ for all G such that G_0 is a subgraph of G. A property has a sharp threshold, if $Pr[A|G(n,p_1)] = \epsilon$, $Pr[A|G(n,p_2)] = 1-\epsilon$ and $p_2 - p_1 = o(1)$

Theorem 9.2 (Erdös and Renyi). *The threshold for graph connectivity is at* $p = \frac{\log n}{n}$

$p < (1-\epsilon)\frac{\log n}{n}$	G almost surely disconnected
$p > (1+\epsilon)\frac{\log n}{n}$	G almost surely connected

There is a 'counter-point' model to our deleting model, where we throw in edges. There are some surprising facts in this model. For example, when you throw in the edge that reaches the last isolated vertex, with almost certainty, you also connect the graph - at the exact same stage. At the same instant, you also make the graph hamiltonian.

It may be illustrative to see the form of these arguments.

Proof. sketch Let $p < (1-\epsilon)\frac{\log n}{n}$. Let X be a random variable representing the number of isolated vertices. Then $E[X] \to \infty$ since $E[x] = n(1-p)^{n-1}$. We also need a second moment argument like Chebyshev to deduce $X > 0$ almost surely. In particular, when $X > 0$, the graph is disconnected. □

Proof. Let Y_k be a random variable that counts the number of sets $S \subset V$ with $|S| = k$ that have no edges between S and its complement. Then the $E[Y_k] = \binom{n}{k}(1-p)^{k(n-k)}$. It can be checked that if $p > (1+\epsilon)\frac{\log n}{n}$,

then $\sum_{k \leq \frac{n}{2}} E[Y_k] = o(1)$. It follows that with probability $1 - o(1)$ no such sets exist. Clearly, when no such sets exist, the graph is connected. □

9.2.1 Relation to KKL

Why should we expect KKL to work like these examples?

If $f : \{0,1\}^n \to \{0,1\}$ with $E[f] = \frac{1}{2}$[1]. By KKL, $\exists x \in [n]\ Inf_f(x) > \Omega(\frac{\log n}{n})$. Let $N = \binom{[n]}{2}$ then each $z \in \{0,1\}^N$ is a description of an n vertex graph and the variables correspond to edges.

We can now view graph property as an N-variable boolean function. Notice also by symmetry if one edge (variable) is influential, then all edges (variables) are influential. As we will see later large influence entails a sharp threshold.

To generalize, we need to understand the role of p in $G(n,p)$. We have to work with $\{0,1\}^N$ not under the uniform distribution but under the following product distribution: $Pr[U] = p^{|U|}(1-p)^{N-|U|} = p^{E(G)}(1-p)^{\binom{n}{2}-E(G)}$. We are denoting the Hamming weight of U as $|U|$. and $E(G)$ is the edge set of the graph G.

9.3 BK^3L

9.3.1 A relation between influence and the derivative of $\mu_p(A)$

The new B and K in our theory are Bourgain and Katznelson. By $\mu_p(A)$ we denote the probability that the property A holds under the $p, 1-p$ product measure.

Lemma 9.3 (Margulis & Russo). *Let $A \subseteq \{0,1\}^n$ be a monotone subset and let $\mu_p(A)$ be the p-measure of A. For $x \in A$ let $h(x) = |\{y \notin A | x, y \in E(cube)\}|$ (number of neigbors of x outside of A).*

Let $\Psi_p(A) = \sum h(x)\mu_p(x)$, the weighted sum of these hs.

Additionally let $\Phi_p(A)$ be the sum of influences of individual variables. Then

$$\Phi_p(A) =_1 \frac{\Psi_p(A)}{p} =_2 \frac{d}{dp}\mu_p(A)$$

The subscripts on the equality are only for convenience in the proof.

Definition 9.1. We will say $x \succ y$, if x and y differ in exactly one coordinate, say the i^{th}, and $x_i = 1$ and $y_i = 0$.

Influences, more generally In general, if X is a probability space and if $f : X^n \to \{0,1\}$ (i.e. f can be viewed as an indicator function for a subset of X^n). For $1 \leq k \leq n$, we can say $Inf_f(k) = Pr_{X^{n-1}}[$ Obtain a non-constant fiber $]$. Here we are randomly choosing $n-1$ coordinates from X with the k^{th} coordinate missing, and checking if the resulting fiber is constant for f. Namely, if the value of f is fixed regardless of the choice of for the k^{th} variable.

[1] choosing $E[f] = \frac{1}{2}$ is not critical. Anything bounded away from 0,1 will do

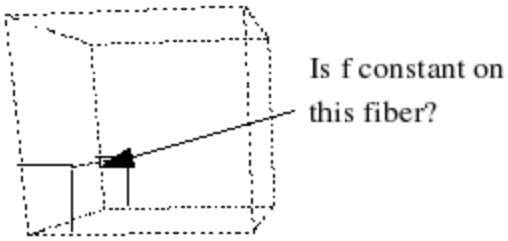


Figure 9.2: Cube with a Fiber

Proof. We prove equality 1. $\Phi_p(A)$ is the sum of all influences. The influence of the i^{th} variable is the weighted sum of all such edges such that $x \succ y$ where $x \in A$, $y \notin A$ and $x_i = 1, y_i = 0$. The probability of the relevent event is this: We have selected all coordinates except the i^{th} and the outcome should coincide with x. There are $|x| - 1$ coordinates which are 1 among those and $n - |x|$ coordinates for which are 0. So we can rewrite the formula as follows

$$\begin{aligned}
\Phi_p(A) &= \sum_{x\in A, y\notin A, x\succ y} p^{(|x|-1)}(1-p)^{n-|x|} \\
&= \frac{1}{p} \sum_{x\in A, y\notin A, x\succ y} p^{(|x|)}(1-p)^{n-|x|} \\
&= \frac{1}{p} \sum_{x\in A} p^{(|x|)}(1-p)^{n-|x|} \; |\{y | y \notin A, x \succ y\}| \\
&= \frac{1}{p} \sum_{x\in A} p^{(|x|)}(1-p)^{n-|x|} \; h(x) \\
&= \frac{1}{p} \sum_{x\in A} \mu_p(x) h(x) = \frac{1}{p} \Psi_p(A)
\end{aligned}$$

□

Proof. Equality 2.

$$\frac{d}{dp}\mu_p(A) = \sum_{x\in A} |x| p^{|x|-1}(1-p)^{n-|x|} - \sum_{x\in A} (n-|x|) p^{|x|}(1-p)^{n-|x|-1}$$

$$p\frac{d}{dp}\mu_p(A) = \sum_{x\in A} |x| p^{|x|}(1-p)^{n-|x|} - \frac{p}{1-p} \sum_{x\in A} (n-|x|) p^{|x|}(1-p)^{n-|x|}$$

For a fixed vertex of the cube, x, and e an edge incident with x define

$$w_{x,e} = \begin{cases} 1 & \text{e goes down from x} \\ -\frac{p}{1-p} & \text{e goes up from x} \end{cases}$$

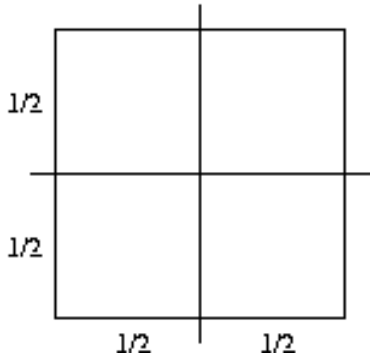


Figure 9.3: Partitioning the cube to derive KKL from BK^3L

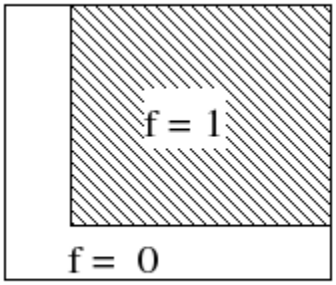


Figure 9.4: f on n=2 case

So we can rewrite (summing over x and e's incident).

$$p\frac{d}{dp}\mu_p(A) = \sum_{x\in A, e\sim x} w_{x,e}\mu_p(x)$$

This is because there are $|x|$ edges going down from x and $|n-x|$ edges going up from it. Notice that if $x \succ y$ are both in A and $e = (x, y)$, then $w_{x,e}\mu_p(x) + w_{y,e}\mu_p(y) = 0$. It follows that we can restrict to the sum to the edges in $E(A, A^c)$. In other words,

$$p\frac{d}{dp}\mu_p(A) = \sum_{x\in A, y\notin A, e=(x,y)} w_{x,e}\,\mu_p(x) = \sum_{x\in A, y\notin A, e=(x,y)} \mu_p(x) = \sum_{x\in A} h(x)\mu_p(x) = \Psi_p(A)$$

□

Returning to the proof that every monotone graph property has a sharp threshold. Let A be a monotone graph property and let us operatre in the probability space $G(n, p)$. We will show here that the p value where the property holds with less than ϵ is very close to where the property holds with $\frac{1}{2}$. A symmetric argument for $1-\epsilon$ will give us the full desired result.

9.3.2 Words about BKKKL

Theorem 9.4 (BKKKL). *Let $f : [0,1]^n \to \{0,1\}$ with $E[f] = t$, let $t' = min(t, 1-t)$. Then there exists $n \geq k \geq 1$ such that $Inf_f(k) \geq \Omega(t' \frac{\log n}{n})$*

Set version of KKL $\forall f : \{0,1\}^n \to \{0,1\}$ $E[f] \sim \frac{1}{2}$ and for every $\omega(n) \to \infty$ as $n \to \infty$, $\exists S \subseteq [n]$. $|S| \leq \frac{n(\omega(n))}{\log n}$ with $Inf_f(S) = 1 - o(1)$. This result follows from repeated application of KKL.

Remark. It is interesting to note that the analgous statement for $f : [0,1]^n \to [0,1]$ does not hold.

Consider the following f, represented in figure 9.4. Let $f(x_1, \ldots, x_n) = 0$ iff $\exists i$ $0 \leq x_i \leq \frac{c}{n}$ where $c = \log_e(2)$. In other words, $f^{-1}(1) = \prod_i^n [\frac{c}{n}, 1]$. Let $|S| = \alpha$. In this example, $Inf(S) = Pr$[f still undetermined when all variables outside of S are set at random]. The function is still undetermined iff all others outside the set are 1. This happens with probability $(1 - \frac{c}{n})^{n(1-\alpha)} \approx e^{-c(1-\alpha)}$, which is bounded away from 1.

This is a 'close-cousin' of the tribes example. Recall in the tribes example we broke the variables into 'tribes' of size $\sim \log n - \log\log n$. Each tribe contributed if all variables take on the value 1, that is there is one assignment out of the $2^{\log n - \log\log n} = \frac{n}{\log n}$ such that the tribe had value 1. In our setting, we can identify tribes with single variables. The 0 region of the continuous case corresponds to the assigment where all variables in the discrete case are set to 1, since this determines the function.

Proof. By BK^3L there exist influential variables. By symmetry all variables are influential. Sum of all individual influences are at least as large as $\Phi_p(A) \geq \Omega(\epsilon \log N) = \Omega(\epsilon \log n)$

$$\Phi_p(A) \geq \Omega(\mu_p(A) \log n)$$

By Margulis-Russo Lemma we know $\Phi_p(A) = \frac{d}{dp} \mu_p(A)$.

$$\frac{d}{dp}\mu_p(A) \geq \Omega(\mu_p(A) \log n)$$

$$(\frac{d}{dp}\mu_p(A))/\mu_p(A) \geq \Omega(\log n)$$

$$\frac{d}{dp}(\log(\mu_p(A)) \geq \Omega(\log n)$$

let p_1, p_2 be defined by $Pr_{G(n,p_1)}[A] = \epsilon$ and $Pr_{G(n,p_2)}[A] = \frac{1}{2}$. From above we know that $d(\log \mu_p(A)) > \Omega(\log n)$ so $p_1 - p_2 < O(\frac{\log \frac{1}{\epsilon}}{\log n})$. □

Remark. We will not give a proof here, but note that Freidgut showed using standard measure theory how to derive BK^3L from KKL. Namely, how we can reach the same conclusion for any $f : X^n \to \{0,1\}$, where X is any probability space. To derive KKL from BK^3L, is easy: Given $f : \{0,1\}^n \to \{0,1\}$ define $F : [0,1]^n \to \{0,1\}$ by breaking the cube to 2^n subcubes and letting F be constant on each subcube that is equal to f at the correpsoning vertex of the cube. For a simple illustration of the case $n = 2$, see figure 9.3.