Analysis of Boolean functions

Yuval Filmus

January 14, 2016

Contents

1 Basics (21 October 2015)

(Roughly [O'D14, Chapter 1].)

Analysis of Boolean functions studies Boolean functions by treating them as real-valued functions that happen to be Boolean. The main tool is Fourier analysis. In computer science and electrical engineering it is more usual to encounter Fourier analysis on $\mathbb{Z}_n$ or on the real line. In contrast, here the relevant group is $\mathbb{Z}_2^n$, which is a product space, and the flavor of the subject is quite different.

The most popular object of study in the area is a Boolean function on the Boolean cube, $f\colon \{0,1\}^n \to \{0,1\}$. Sometimes it will be nicer to change the domain to $\{1,-1\}^n$ and/or the range to $\{-1,1\}$. Where does the Boolean function come from? It can be a function computed by some circuit. Or it could be a family of sets, a voting scheme or some graph property. It could also be a probablistically checkable proof.

An excellent reference on the field is the recent monograph by Ryan O'Donnell [O'D14]. The library should have two copies of this book.

Fourier expansion It is somewhat nicer to consider functions on $\{1,-1\}^n$ rather than on $\{0,1\}^n$.

Claim 1. *Every function $f\colon \{1,-1\}^n \to \mathbb{R}$ has a unique representation as a multilinear polynomial.*

We will provide two proofs of this claim, which is the basis for everything that follows.

Proof. Let $t_1, \ldots, t_n \in \{1,-1\}^n$, and consider the function

$$\delta_{\mathbf{t}}(\mathbf{x}) = \prod_{i=1}^{n} \frac{1 + t_i x_i}{2}.$$

It's not too hard to check that $\delta_{\mathbf{t}}(\mathbf{t}) = 1$, and $\delta_{\mathbf{t}}(\mathbf{x}) = 0$ everywhere else. This easily implies that every function $f\colon \{1,-1\}^n \to \mathbb{R}$ can be represented as a multilinear polynomial.

To show that this representation is unique, let us show that the only multilinear polynomial that vanishes on $\{1,-1\}^n$ is the zero polynomial. The proof is by induction on n. The result is true for $n = 0$. Suppose that it's true for some n, and consider some representation P of zero. Write

$$P = x_{n+1}Q + R,$$

where Q, R are multilinear polynomials over $x_1, \ldots, x_n$. The induction hypothesis shows that $R + Q$ and $R - Q$ are both the zero polynomials. This implies that $R = Q = 0$, and so $P = 0$ is also the zero polynomial. □

The unique representation of f as a multilinear polynomial is known as the *Fourier expansion* of f:

$$f = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S.$$

Here $\chi_S(\mathbf{x}) = \prod_{i \in S} x_i$ is known as a *Fourier character* of *Fourier basis vector* (it is indeed a multiplicative character of the group $\mathbb{Z}_2^n$). When the domain is $\{0,1\}^n$ instead of $\{1,-1\}^n$, the Fourier character is given by $\chi_S = (-1)^{\sum_{i \in S} x_i}$.

Degree The *degree* of f is the degree of its multilinear representation. There is also a spatial interpretation of degree: it is the minimum d such that f can be written as a linear combination of d-juntas (functions depending on d coordinates). Indeed, on one hand, χ_S is an $|S|$-junta, and on the other, any d-junta has degree d.

Boolean functions of degree d. Suppose we have a Boolean function of degree d, say from $\{-1,1\}^n$ to $\{-1,1\}$. On how many variables can the function depend? (A function f *depends* on a variable x_i if there is a pair of inputs x, y differing only in the ith coordinate such that $f(x) \neq f(y)$.) Let $N(d)$ denote the maximal number of such variables for unbounded n, which turns out to be finite. It is easy to see that $N(0) = 0$.

A bit harder is seeing that $N(1) = 1$. First of all, the function $f(x_1) = x_1$ shows that $N(1) \geq 1$. In the other direction, let us consider a general degree 1 function f. I claim that $\hat{f}(\{i\}) \neq 0$ for at most one index i. For suppose that $\hat{f}(\{i\}), \hat{f}(\{j\}) \neq 0$ for $i \neq j$. By substituting values for all other variables, we can assume without loss of generality that $i = 1$, $j = 2$ and $n = 2$. The function is then

$$f = \hat{f}(\emptyset) + \hat{f}(\{1\})x_1 + \hat{f}(\{2\})x_2.$$

We now have to consider several cases, according to the signs of $\hat{f}(\{1\}), \hat{f}(\{2\})$. All cases are similar, so let us consider the case that both are positive. This implies that

$$f(-1,-1) < f(-1,+1) < f(+1,+1),$$

which contradicts the fact that f is Boolean.

What about $N(d)$ for general d? At first one might conjecture that $N(d) = d$, but in fact $N(d)$ is much larger. The following construction shows that $N(d) \geq 2^d - 1$. We consider a decision tree of depth d, where at each internal node we have a variable dictating which way to go, and each leaf contains the desired output variable. (A more common example with only $2^{d-1} + d - 1$ variables is the *addressing function*, in which all internal nodes at a given depth are assigned the same variable.) We can take all variables to be distinct, and the result is a Boolean function of degree d depending on $2^d - 1$ variables (try it out!).

This construction is not optimal. For example, for $d = 2$ we can show that $N(d) \geq 4$ by considering the function

$$\frac{a(x+y) + b(x-y)}{2}.$$

If $a = b$ then the value of this function is ax, and otherwise it is ay; this shows that the function is Boolean. This construction turns out to be optimal, so that $N(d) = 4$. Indeed, using influences one can show the upper bound $N(d) \leq d2^{d-1}$. The asymptotic value of $N(d)$ isn't known. Perhaps you can determine it?

Levels of the Fourier expansion We partition the set of Fourier coefficients according to the size of the set S. All coefficients $\hat{f}(S)$ for $|S| = k$ form level k of the Fourier expansion, and we collect them together in

$$f^{=k} = \sum_{|S|=k} \hat{f}(S)\chi_S.$$

Written out, all monomials in the multilinear representation of $f^{=k}$ have the same degree k, and so $f^{=k}$ is *homogeneous*. Clearly

$$f = \sum_{k=0}^{n} f^{=k},$$

and this coarse decomposition suffices in many cases, though not when we discuss specific coordinates. Similar self-explanatory notations are $f^{<k}, f^{\leq k}, f^{>k}, f^{\geq k}$.

Orthonormality We define an inner product on the space of functions from $\{1,-1\}^n \to \mathbb{R}$ by

$$\langle f, g \rangle = \mathop{\mathbb{E}}_{x}[f(x)g(x)] = \frac{1}{2^n} \sum_{x \in \{1,-1\}^n} f(x)g(x).$$

Under this inner product, it is easy to check that each χ_S has unit norm, simply since $\chi_S^2 \equiv 1$. What about the inner product between χ_S and χ_T when $S \neq T$? It is

$$\langle \chi_S, \chi_T \rangle = \mathop{\mathbb{E}}_{x}\left[\prod_{i \in S} x_i \prod_{i \in T} x_i\right] = \mathop{\mathbb{E}}_{x}\left[\prod_{i \in S \triangle T} x_i\right] = \mathop{\mathbb{E}}_{x}[\chi_{S \triangle T}(x)].$$

Since $S \neq T$, there is some $i \in S \triangle T$, and so

$$\langle \chi_S, \chi_T \rangle = \mathop{\mathbb{E}}_{x_i}\left[x_i \mathop{\mathbb{E}}_{x_{-i}}\left[\chi_{(S \triangle T) \setminus i}(x)\right]\right] = 0.$$

(Here x_{-i} consists of all variables except x_i.) So the different characters are orthogonal (this is actually true in every group). In particular, they are linearly independent. Since there are 2^n of them, and the dimension of the space of functions $\{1,-1\}^n \to \mathbb{R}$ is also 2^n, we deduce that the Fourier characters form an orthonormal basis for this space. This is another proof of Claim 1.

Consequences of the orthonormality of characters The orthonormality of characters immediately implies an important identity, known as Parseval's identity:

$$\langle f, g\rangle = \sum_{S,T} \hat{f}(S)\hat{g}(T)\,\mathbb{E}[\chi_S\chi_T] = \sum_S \hat{f}(S)\hat{g}(S).$$

In particular, we get an expression for the L2 norm of f:

$$\mathop{\mathbb{E}}_x[f(x)^2] = \|f\|^2 = \langle f, f\rangle = \sum_S \hat{f}(S)^2.$$

Orthonormality of characters also implies a formula for the individual Fourier coefficients:

$$\langle f, \chi_U\rangle = \sum_S \hat{f}(S)\langle \chi_S, \chi_U\rangle = \hat{f}(U).$$

In particular, since $\chi_\emptyset = 1$, we get that $\hat{f}(\emptyset) = \mathbb{E}[f]$. This implies that

$$\mathbb{V}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = \sum_{S\neq\emptyset} \hat{f}(S)^2.$$

Orthonormality also implies formulas of the style

$$\|f\|^2 = \sum_{k=0}^{n} \|f^{=k}\|^2.$$

Linearity A function $f\colon \{\pm1\}^n \to \{\pm1\}$ is *linear* if $f(xy) = f(x)f(y)$. (This is a piece of terminology we will only use in this lecture.) Here xy is given by $(xy)_i = x_iy_i$; in the $\{0,1\}$ world this corresponds to XOR. It is easy to check that every Fourier character is linear. It is not too difficult to see that the converse also holds. First, $f(1) = f(1)^2 = 1$. This implies that $f(x) = \prod_{x_i=-1} f(e_i)$, where e_i is the vector which is -1 only in coordinate i. We conclude that $f = \chi_S$, where $S = \{i : f(e_i) = -1\}$.

1.1 Property testing

Already at this stage, we can present an application of Fourier analysis, to property testing. The goal in property testing is to decide whether a given function f satisfies a given property using a few random samples of f. We would like the tester to say YES with high probability if f satisfies the property, and to say NO with high probability if f doesn't satisfy it. However, this is usually too much to ask. Consider the property of being linear. Given a linear function f, if we change f at a few places then it stops being linear, but any property tester reading just a few values of f has very small chance of noticing. In order to fix that, we change our requirements:

- If f satisfies the property, then the property tester always answers YES.
- If f is ϵ-far from every function satisfying the property, then the property tester answers NO with constant probability. (The probability itself can be enhanced by repetition.)

Blum–Luby–Rubinfeld A natural tester for linearity samples $x, y \in \{\pm1\}^n$ at random, and checks that $f(xy) = f(x)f(y)$. This tester always answers YES if f is linear. What happens if f is far from linear? To analyze this, we first find an expression for the success probability of the tester on an arbitrary function f:

$$\sigma = \mathop{\mathbb{E}}_{x,y}\left[\frac{1 + f(x)f(y)f(xy)}{2}\right].$$

What is this expression equal to?

$$\mathop{\mathbb{E}}_{x,y}[f(x)f(y)f(xy)] = \sum_{S,T,U} \hat{f}(S)\hat{f}(T)\hat{f}(U) \mathop{\mathbb{E}}_{x,y}[\chi_S(x)\chi_T(y)\chi_U(xy)].$$

Since χ_U is linear,

$$\mathop{\mathbb{E}}_{x,y}[\chi_S(x)\chi_T(y)\chi_U(xy)] = \mathop{\mathbb{E}}_{x}[\chi_S(x)\chi_U(x)] \mathop{\mathbb{E}}_{y}[\chi_T(y)\chi_U(y)].$$

This vanishes unless $S = U = T$, and so

$$\mathop{\mathbb{E}}_{x,y}[f(x)f(y)f(xy)] = \sum_S \hat{f}(S)^3.$$

Therefore

$$\sigma = \frac{1}{2} + \frac{1}{2}\sum_S \hat{f}(S)^3.$$

Suppose that the tester succeeds with probability $1/2+\epsilon$ for some $\epsilon > 0$. This implies that $\sum_S \hat{f}(S)^3 = 2\epsilon$. On the other hand,

$$\sum_S \hat{f}(S)^3 \leq \left(\max_S \hat{f}(S)\right) \sum_S \hat{f}(S)^2 = \max_S \hat{f}(S).$$

Therefore there exists some S such that $\hat{f}(S) \geq 2\epsilon$. In other words, $\langle f, \chi_S \rangle \geq 2\epsilon$. Now

$$\Pr[f = \chi_S] = \mathbb{E}\left[\frac{1 + f\chi_S}{2}\right] = \frac{1}{2} + \frac{1}{2}\langle f, \chi_S \rangle \geq \frac{1}{2} + \epsilon,$$

and so f is $(1/2 - \epsilon)$-close to a linear function.

Stated differently, if f is ϵ-far from a linear function, then the tester succeeds with probability at most $1/2 - \epsilon$, and so fails with probability at least $1/2 + \epsilon$. By repeating it $1/\epsilon$ times, we get a tester which samples the function at $O(1/\epsilon)$ points and makes an error with some constant probability.

This analysis is due to Bellare, Coppersmith, Håstad, Kiwi and Sudan [BCH+96]. The original analysis of Blum, Luby and Rubinfeld [BLR93] was combinatorial. Recently a new combinatorial argument was worked out by David, Dinur, Goldenberg, Kindler, and Shinkar [DDG+15].

2 More basic notions (28 October 2015)

(Roughly [O'D14, Chapter 2], additionaly covering Section 3.4 and perhaps Section 3.5.)

Influence Given a Boolean function f, how important is the input x_i? This notion is captured by a quantity called the *influence*:

$$\mathrm{Inf}_i[f] = \Pr_x[f(x) \neq f(x \oplus e_i)].$$

What does this have to do with Fourier analysis? There is a different way of writing the same formula, which works when the function f is $\{\pm 1\}$-valued (a very similar formula works for $\{0,1\}$-valued functions):

$$\mathrm{Inf}_i[f] = \frac{1}{4}\mathbb{E}_x[(f(x) - f(x \oplus e_i))^2].$$

As a bonus, this definition makes sense even for non-Boolean functions. Why did we choose $(f(x) - f(x \oplus e_i))^2$ over, say, $|f(x) - f(x \oplus e_i)|$? Since our definition facilitates the use of Fourier analysis. We can define the *derivative in direction i* of a function f by $(L_i f)(x) = f(x) - f(x \oplus e_i)$, and then $\mathrm{Inf}_i[f] = (1/4)\|L_i f\|^2$. We use the notation $L_i f$ since the *Laplacian* of f, an important quantity in spectral graph theory that (might) show up later, is given by $Lf = \sum_i L_i f$.

In order to compute the Fourier expansion of $L_i f$, let us start by computing the Fourier expansion of $f(x \oplus e_i)$. If $f(x) = \sum_S \hat{f}(S)\chi_S(x)$ then $f(x \oplus e_i) = \sum_S \hat{f}(S)\chi_S(x)\chi_S(e_i)$, and so the Fourier coefficient of S is $\hat{f}(S)\chi_S(e_i)$. Therefore the Fourier expansion of $L_i f$ is

$$L_i f = \sum_S \hat{f}(S)(1 - \chi_S(e_i))\chi_S = 2\sum_{S \ni i} \hat{f}(S)\chi_S.$$

Parseval's identity immediately shows that

$$\mathrm{Inf}_i[f] = \frac{1}{4}\|L_i f\|^2 = \sum_{S \ni i} \hat{f}(S)^2.$$

This very neat formula will be very useful in the future.

A related notion is the *total influence*, given by

$$\mathrm{Inf}[f] = \sum_i \mathrm{Inf}_i[f].$$

If f is a Boolean function that encodes some set $A \subseteq 2^{[n]}$, then the total influence measures the *edge-perimeter* of A. The edge-perimeter of A, written $\mathrm{ep}(A)$, is the number of edges of the Boolean hypercube whose one end is in A, and whose other end is in $\overline{A}$ (those are called *bichromatic edges*). The relation is $\mathrm{Inf}[f] = \mathrm{ep}(A)/2^n$. If f depends on d coordinates then clearly $\mathrm{Inf}[f] = \mathrm{ep}(A)/2^n \leq d$, since only edges in these directions can be bichromatic. A stronger bound is in fact true: $\mathrm{Inf}[f] \leq \deg f$. This is due to the spectral formula for $\mathrm{Inf}[f]$, which we obtain directly from the formulas for $\mathrm{Inf}_i[f]$:

$$\mathrm{Inf}[f] = \sum_S |S|\hat{f}(S)^2.$$

If $\deg f \leq d$ then we can bound $|S| \leq d$ and so Parseval's identity shows that $\mathrm{Inf}[f] \leq d$. Another easy inequality is $\mathbb{V}[f] \leq \mathrm{Inf}[f]$, since $\mathbb{V}[f]$ sums $\hat{f}(S)^2$ over all non-zero $|S|$. Altogether, we get the double-sided Poincaré inequality:

$$\mathbb{V}[f] \leq \mathrm{Inf}[f] \leq (\deg f)\,\mathbb{V}[f].$$

(In fact, only the lower bound $\mathbb{V}[f] \leq \mathrm{Inf}[f]$ should be called Poincaré's inequality.)

An alternative formula for $\mathrm{Inf}[f]$ is

$$\mathrm{Inf}[f] = \sum_{d=0}^{n} d\|f^{=d}\|^2.$$

This shows that $\mathrm{Inf}[f]$ only depends on the weight distribution of the Fourier expansion on the various levels.

L1 influences What about the alternative definition $\mathrm{Inf}_i^{(1)}[f] = \frac{1}{2}\mathbb{E}_x[|f(x) - f(x \oplus e_i)|]$? In this case we know that if $\|f\|_\infty \leq 1$ and $\deg f = d$ then $\mathrm{Inf}^{(1)}[f] \leq \min(d^2, n)$ (see [FHKL]), but it is conjectured that in fact $\mathrm{Inf}^{(1)}[f] \leq d$.

The upper bound $\mathrm{Inf}^{(1)}[f] \leq n$ is trivial. The other upper bound $\mathrm{Inf}^{(1)}[f] \leq d^2$ relies on a generalization of the classical Markov's inequality from approximation theory. Markov's inequality states that if f is a univariate polynomial of degree d satisfying $|f(x)| \leq 1$ for all $|x| \leq 1$ then $|f'(x)| \leq d^2$ for all $|x| \leq 1$; the maximum is attained by the Chebyshev polynomials at the point $x = 1$ (among else).

Using the classical inequality itself, one gets the slightly weaker upper bound of $2d^2$, which is what we show here. We actually show that at *each* point x, $\frac{1}{2}\sum_i |f(x) - f(x \oplus e_i)| \leq 2d^2$; for convenience we focus on the point $x = \mathbf{1}$. Let $S = \{i : f(x) \geq f(x \oplus e_i)\}$. Then

$$\frac{1}{2}\sum_i |f(x) - f(x \oplus e_i)| = \frac{1}{2}\sum_{i \in S}[f(x) - f(x \oplus e_i)] - \frac{1}{2}\sum_{i \notin S}[f(x) - f(x \oplus e_i)].$$

We will bound both terms by d^2. In fact, for *every* S we will bound $\frac{1}{2}\sum_{i \in S}[f(x) - f(x \oplus e_i)]$ by d^2 in absolute value. (The generalization of Markov's inequality allows us to bound both terms at once.) Calculation along the lines above gives

$$\frac{1}{2}\sum_{i \in S}[f(x) - f(x \oplus e_i)] = \sum_T |T \cap S| \hat{f}(T)\chi_T(x).$$

In particular, at the point $x = \mathbf{1}$ we get

$$\frac{1}{2}\sum_{i \in S}[f(\mathbf{1}) - f(\mathbf{1} \oplus e_i)] = \sum_T |T \cap S| \hat{f}(T).$$

Consider now the function $g(y) = f(\overbrace{y}^{S}, \overbrace{1}^{\overline{S}})$. Easy calculation shows that $g'(1)$ is exactly the quantity we wish to bound. In view of Markov's inequality, we can deduce that $|g'(1)| \leq d^2$ if we can show that $|g(y)| \leq 1$ for all $|y| \leq 1$. In fact, more is true: $|f(x)| \leq 1$ for all $x \in [-1, 1]^n$, since f is multilinear and so obtains its optima at endpoints of the cube. Indeed, consider any point $x \in [-1, 1]^n$. As a function of x_1 (fixing all other variables), f is of the form $f(x_1) = \alpha x_1 + \beta$, and so attains its maximum and minimum on one of the endpoints. Repeating the same argument $n-1$ more times, we see that f attains its optimal at endpoints, and so $f([-1, 1]^n) \subseteq [-1, 1]$.

Nisan–Szegedy The notion of influences allows us to prove a theorem of Nisan and Szegedy [NS94]: a Boolean function f of degree d depends on at most $d2^{d-1}$ variables. We do this by showing that if f depends on a variable x_i then $\mathrm{Inf}_i[f] \geq 1/2^{d-1}$. Since the total influence of f is at most d, it follows that f depends on at most $d2^{d-1}$ variables.

To prove this claim, we first prove a version of the Schwartz–Zippel lemma (called Schwartz's lemma by Nisan and Szegedy) for the Boolean cube: *Let $f\colon \{-1, 1\}^n \to \mathbb{R}$ be a non-zero function of degree d. Then f has at least 2^{n-d} non-zeroes.* The proof is by induction on n and d (technically, the induction is on $n + d$). When $n = d$ the result is clear. Consider now some $n > d$, and write $f = x_n g + h$. We now have to consider three cases:

1. Case 1: $g + h = 0$. In this case we can write $f = (x_n - 1)g$, where $\deg g \leq d - 1$ is non-zero. By induction we known that over the cube $\{-1, 1\}^{n-1}$, g has at least $2^{(n-1)-(d-1)} = 2^{n-d}$ non-zeroes. These correspond to 2^{n-d} non-zeroes of f.

2. Case 2: $-g + h = 0$. In this case we can write $f = (x_n + 1)g$, and reason as in case 1.

3. Case 3: Both $g + h$ and $-g + h$ are non-zero. In that case by induction we know that each of these have at least 2^{n-1-d} non-zeroes, and we obtain at least 2^{n-d} non-zeroes of f.

Recall that $\mathrm{Inf}_i[f] = \frac{1}{4}\|L_i f\|^2 = \Pr[L_i f \neq 0]$. While $L_i f$ itself has degree d, it is divisible by x_i, and in fact $\Pr[L_i f \neq 0] = \Pr[(L_i f)/x_i \neq 0]$. Since $(L_i)f/x_i$ has degree $d - 1$, Schwartz's lemma shows that either $(L_i f)/x_i = 0$ or $\Pr[(L_i f)/x_i \neq 0] \geq 1/2^{d-1}$, completing the proof.

Noise stability Another useful basic notion is that of *noise stability*, which measures how sensitive a given function is to noise applied to its inputs. Given a vector x, define a vector y by taking $y_i = x_i$ with probability $\frac{1+\rho}{2}$ and $y_i = -x_i$ with probability $\frac{1-\rho}{2}$ (where $\rho \in [-1,1]$). We denote this distribution $N_\rho(x)$, and define the *noise operator* T_ρ by

$$(T_\rho f)(x) = \mathop{\mathbb{E}}_{y \sim N_\rho(x)}[f(y)].$$

What happens if we apply the noise operator to a Fourier character? Let's start with a simple example, the function $f = \chi_i$:

$$T_\rho \chi_i = \frac{1+\rho}{2}(x_i) + \frac{1-\rho}{2}(-x_i) = \rho x_i.$$

More generally, $T_\rho \chi_S = \rho^{|S|}\chi_S$, and so

$$T_\rho f = \sum_S \rho^{|S|}\hat{f}(S)\chi_S = \sum_d \rho^d f^{=d}.$$

Applying noise reduces the higher-order parts of f more heavily than it does the lower-order parts.

A simple calculation reveals the following connection between T_ρ and total influence:

$$\mathrm{Inf}[f] = \left\| \frac{\partial T_\rho f}{\rho}(\rho = 1) \right\|^2.$$

In the future, we will be interested in expressions of the form

$$\langle T_\rho f, g \rangle = \mathop{\mathbb{E}}_x \mathop{\mathbb{E}}_{y \sim N_\rho(x)}[f(x)g(y)] = \mathop{\mathbb{E}}_{(x,y)\sim N_\rho}[f(x)g(y)].$$

Here N_ρ is a pair of *ρ-correlated* inputs: each of the marginals is uniform, and $\mathbb{E}[x_i y_i] = \rho$. Note that the relation between x and y is symmetric: if we take y as uniformly random and $x \sim N_\rho(y)$ then we obtain the same distribution. (This means that the operator T_ρ is self-adjoint.)

2.1 Learning

PAC learning is a central area in computational complexity. Initiated by Lesley Valiant, it attempts to model the same situations encountered in practice in the discipline of machine learning. Learning has very close ties to the subject of our course, but we will only touch it briefly.

Consider the following situation. There is a function f, which is given to us as a blackbox. We also know that f belongs to some *concept class* $\mathcal{F}$. Our task is to come up with a function g which is *close* to f. In order to do that, we are given access to *random samples* of the function f. Our goal is to design an algorithm that makes as few samples as possible, and efficiently generates a function g which is ϵ-close to f, where ϵ is a parameter. Our algorithm is allowed to fail, but we require that it succeed with high probability (or at least large constant probability).

Fourier learning Many concept classes satisfy a property known as *Fourier concentration*, in which there are only a few Fourier coefficients which are "significant". We say that the Fourier expansion of a function $f\colon \{-1,1\}^n \to \{-1,1\}$ is *ϵ-concentrated on the set of coefficients $\mathcal{E}$* if

$$\sum_{S \in \mathcal{E}} \hat{f}(S)^2 \geq 1 - \epsilon.$$

In words, all but an ϵ-fraction of the Fourier weight resides in the set of coefficients $\mathcal{E}$. A concept class is ϵ-concentrated on $\mathcal{E}$ if all functions in the concept class satisfy the condition.

Theorem 1. *Suppose that a concept class $\mathcal{F}$ is ϵ-concentrated on $\mathcal{E}$. Then $\mathcal{F}$ can be efficiently learned up to an error of ϵ using $\tilde{O}(|\mathcal{E}|/\epsilon)$ samples.*

To prove the theorem, we need to use the ubiquitous Chernoff bound, in the following form. Suppose that $X_1, \ldots, X_m$ are independent random variables such that $X_i \in [a,b]$, and let $\bar{X}$ be their mean. Then for all $t \geq 0$,

$$\Pr[|\bar{X} - \mathbb{E}[\bar{X}]| \geq t] \leq 2e^{-2mt^2/(b-a)^2}.$$

Proof. The idea is to estimate $\hat{f}(S)$ for all $S \in \mathcal{E}$. How do we do that? Given m samples $(x, f(x))$, we calculate an approximation $\tilde{f}(S)$ by taking the empirical average of $f(x)\chi_S(x)$. Chernoff's bound states that

$$\Pr[|\tilde{f}(S) - \hat{f}(S)| > \delta] \leq 2e^{-m\delta^2/2}.$$

Choose $\delta^2 = \epsilon/|\mathcal{E}|$ and $m = 2|\mathcal{E}|\log(10|\mathcal{E}|)/\epsilon$. Calculation shows that with constant probability, $|\tilde{f}(S) - \hat{f}(S)| \leq \delta$ for all $S \in \mathcal{E}$. For the rest of the proof, suppose that this event happens.

Define $h = \sum_{S\in\mathcal{E}} \tilde{f}(S)\chi_S$. We have

$$\|h - f\|^2 = \sum_{S\in\mathcal{E}} (\tilde{f}(S) - \hat{f}(S))^2 + \sum_{S\notin\mathcal{E}} \hat{f}(S)^2 \leq 2\epsilon.$$

This is great, but there is one problem: h is not a Boolean function! Fortunately, this is easy to fix. Let $g = \operatorname{sgn} h$. What can we say about $\|f - g\|^2$? Consider a specific point x, and suppose that $h(x) \geq 0$, so that $g(x) = 1$. If $f(x) = 1$ then $|g(x) - f(x)| \leq |h(x) - f(x)|$. If $f(x) = -1$ then $|g(x) - f(x)| = 2$ while $|h(x) - f(x)| \geq 1$. Either way, $|g(x) - f(x)| \leq 2|h(x) - f(x)|$, and so $\|f - g\|^2 \leq 8\epsilon$.

Finally, notice that $\|f - g\|^2 = \mathbb{E}[(f - g)^2] = 8\Pr[f \neq g]$, and so $\Pr[f \neq g] \leq 2\epsilon$. □

Learning monotone functions As a sample application, we'll show how to learn monotone functions. These are functions such that $x \leq y$ implies $f(x) \leq f(y)$, and are popular in computational complexity. If f is monotone, then

$$\mathrm{Inf}_i[f] = \Pr[f(x) \neq f(x \oplus e_i)] = \Pr_{x_{-i}}[f(x_{-i}, 1) = 1, f(x_{-i}, -1) = -1] = \frac{1}{2}\mathop{\mathbb{E}}_{x_{-i}}[f(x_{-i}, 1) - f(x_{-i}, -1)] = \hat{f}(\{i\}).$$

What is the maximal total influence that a monotone function can have?

$$\mathrm{Inf}[f] = \sum_{i=1}^{n} \hat{f}(\{i\}) = \mathop{\mathbb{E}}_{x}[f(x)(x_1 + \cdots + x_n)].$$

Clearly this expression is maximized when $f(x) = \operatorname{sgn}(x_1 + \cdots + x_n)$ (the majority function), in which case it is

$$\mathrm{Inf}[f] \leq \mathbb{E}[|x_1 + \cdots + x_n|] = O(\sqrt{n}).$$

Where did we get this last estimate? The central limit theorem shows that $x_1 + \cdots + x_n$ is close to a normal distribution with zero mean and variance n. Dividing by $\sqrt{n}$, we get that $\frac{x_1+\cdots+x_n}{\sqrt{n}}$ is close to a standard Gaussian, and so the expectation of $x_1 + \cdots + x_n$ should scale like $\sqrt{n}$. The constant can also be calculated if we really care. We will discuss more how to formalize this argument later in the course (we can also just use the central limit theorem directly).

How does a bound on the influence imply spectral concentration? Let $\mathcal{S}$ be a random variable with $\Pr[\mathcal{S} = S] = \hat{f}(S)^2$ ($\mathcal{S}$ is known as the *spectral sample*). Then $\mathrm{Inf}[f] = \mathbb{E}[|\mathcal{S}|]$, and so $\Pr[|\mathcal{S}| > \mathrm{Inf}[f]/\epsilon] < \epsilon$. In other words,

$$\sum_{|S|>\mathrm{Inf}[f]/\epsilon} \hat{f}(S)^2 < \epsilon.$$

We deduce that monotone functions are ϵ-concentrated on Fourier coefficients of degree $O(\sqrt{n}/\epsilon)$. Theorem 1 thus implies that they can be ϵ-learned using $\tilde{O}(n^{O(\sqrt{n}/\epsilon)}/\epsilon)$ samples. This may not look like much, but it is better than the trivial $\tilde{O}(2^n)$ coupon-collectors bound.

Goldreich–Levin algorithm So far we have considered learning from *random* samples. Another popular model of learning is learning by *queries*. In this model, we can query particular values of the function f. This model comes up naturally in cryptography, and it was in this context that Goldreich and Levin came up with their algorithm. Previously we have shown how to *test* that a function is linear. Suppose that f is close to linear. We know that it must be close to some character. Can we find this character? The Goldreich–Levin algorithm acomplishes just that, and more: it can find *all* significant Fourier coefficients.

The idea is to do some sort of binary search. To that end, for $J \subseteq [n]$ and $x \in \{-1,1\}^J$, define $f|_{J=x}(y) = f(x,y)$ to be a function $\{-1,1\}^{\overline{J}} \to \{-1,1\}$. What is the Fourier expansion of $f_{J=x}$? For $T \subseteq \overline{J}$,

$$\mathop{\mathbb{E}}_{y}[f_{J=x}(y)\chi_T(y)] = \sum_S \hat{f}(S)\chi_{S\cap J}(x)\,\mathbb{E}[\chi_{S\cap\overline{J}}(y)\chi_T(y)] = \sum_{S\cap\overline{J}=T} \hat{f}(S)\chi_{S\cap J}(x).$$

Written differently,

$$\hat{f}_{J=x}(T) = \sum_{U\subseteq J} \hat{f}(T\cup U)\chi_U(x).$$

Squaring:

$$\hat{f}_{J=x}(T)^2 = \sum_{U,V\subseteq J} \hat{f}(T\cup U)\hat{f}(T\cup V)\chi_U(x)\chi_V(x).$$

Taking expectation over x:

$$\mathop{\mathbb{E}}_{x}[\hat{f}_{J=x}(T)^2] = \sum_{U,V\subseteq J} \hat{f}(T\cup U)\hat{f}(T\cup V)\,\mathbb{E}[\chi_U(x)\chi_V(x)] = \sum_{U\subseteq J} \hat{f}(T\cup U)^2.$$

As before, we can estimate this quantity for all T, J by sampling, say using the formula $\mathbb{E}_{x,y,z}[f(x,y)f(x,z)\chi_T(yz)]$.

Suppose one Fourier coefficient dominates the Fourier expansion of f, say $\hat{f}(S)^2 \geq 2/3$. At the first step, we find out whether $n \in S$ or $n \notin S$. How do we do that? Choosing $J = \{1,\ldots,n-1\}$ and $T = \emptyset, \{n\}$, we estimate

$$\sigma_0 = \sum_{n\notin S} \hat{f}(S)^2, \quad \sigma_1 = \sum_{n\in S} \hat{f}(S)^2.$$

If (say) $n \in S$, then by taking enough samples, it will be extremely likely that $\sigma_0 < 1/2 < \sigma_1$, while if $n \notin S$, it will be extremely likely that $\sigma_1 < 1/2 < \sigma_0$. We can therefore determine whether $n \in S$, and continuing this way we recover all of S.

A simple modification of this algorithm (left to the reader) finds all $1/\tau^2$ (or less) Fourier coefficients which are at least τ in magnitude.

3 Biased Fourier expansion (4 November 2015)

(Roughly [O'D14, Section 8.4] and parts of [Fri08].)

The idea The classical Fourier basis makes sense when we are interested in properties of function $f\colon \{-1,1\}^n \to \{-1,1\}$ under the uniform distribution. However, in many cases we are actually interested in a different distribution, namely the μ_p distribution. It is nicer to consider this distribution as a distribution over $\{0,1\}^n$, or over subsets of $[n]$. The distribution μ_p is given by

$$\mu_p(S) = p^{|S|}(1-p)^{n-|S|}.$$

When $p = 1/2$, we recover the uniform distribution. There is also a simple generative model: to generate a set according to μ_p, put each element inside with probability p. Stated differently, we generate a vector in $\{0,1\}^n$ by taking each coordinate independently to be a Bernoulli p random variable.

Where does μ_p come up? Random graphs. Percolation. Extremal combinatorics. And even computer science!

Biased characters One important property of the Fourier basis is that if a function f depends only on the coordinates in a set J, then its Fourier expansion is concentrated in coefficients $S \subseteq J$. Another is that the basis is orthonormal. It turns out that these properties determine the Fourier characters almost uniquely (up to sign). (It is a special case of a more general expansion, the Efron–Stein decomposition.)

What is the corresponding basis for μ_p? It's instructive to consider the case $n = 1$ first. We are looking for two elements, $\omega_\emptyset$ and $\omega_{\{1\}}$, that form an orthonormal basis, and furthermore $\omega_\emptyset$ is constant (due to the "junta" property). This forces $\omega_\emptyset \equiv 1$ (up to sign). What about $\omega_{\{1\}}$? Suppose $\omega_{\{1\}}(0) = \alpha$ and $\omega_{\{1\}}(1) = \beta$. Then $(1-p)\alpha^2 + p\beta^2 = 1$ (unit norm) and $(1-p)\alpha + p\beta = 0$ (orthogonality). The latter implies that $\alpha = \frac{-p}{1-p}\beta$, and so

$$(1-p)\frac{p^2}{(1-p)^2}\beta^2 + p\beta^2 = 1 \implies \frac{p^2 + p(1-p)}{1-p}\beta^2 = 1 \implies \beta^2 = \frac{1-p}{p}.$$

Therefore $\beta = \pm\sqrt{\frac{1-p}{p}}$ and $\alpha = \mp\sqrt{\frac{p}{1-p}}$. We choose the solution $\omega_{\{1\}}(0) = \sqrt{\frac{p}{1-p}}$ and $\omega_{\{1\}}(1) = -\sqrt{\frac{1-p}{p}}$ so that when $p = 1/2$, we get the usual Fourier basis (under the mapping $0 \mapsto 1$, $1 \mapsto -1$).

How do we extend this basis to many coordinates? Tensorisation! With some abuse of notation, we define

$$\omega_S(x_1, \ldots, x_n) = \prod_{i=1}^n \omega_{S\cap\{i\}}(x_i).$$

Why does this work? Let's calculate $\mathbb{E}[\omega_S\omega_T]$:

$$\mathbb{E}[\omega_S\omega_T] = \prod_{i=1}^n \mathop{\mathbb{E}}_{x_i}[\omega_{S\cap\{i\}}(x_i)\omega_{T\cap\{i\}}(x_i)].$$

Note that the individual coordinates x_i also have the measure μ_p (this is because μ_p is a *product measure*). Therefore if $S = T$ then we get $\mathbb{E}[\omega_S^2] = 1$, whereas if $i \in S \triangle T$ then $\mathbb{E}_{x_i}[\omega_{S\cap\{i\}}(x_i)\omega_{T\cap\{i\}}(x_i)] = 0$ and so $\mathbb{E}[\omega_S\omega_T] = 0$. So the ω_S are an orthonormal basis with respect to μ_p!

It is also pretty clear that ω_S satisfies the junta property. Therefore ω_S is the basis we were after. While the basis ω_S satisfies some properties of the Fourier basis (which we use below), it lacks others. For example, the "characters" ω_S are no longer multiplicative! So we have to be careful.

Influence and noise stability We can extend the definitions of influence and noise stability. Out of the several possible normalizations, the most convenient one for defining the influence is

$$\mathrm{Inf}_i[f] = \sum_{S\ni i} \hat{f}(S)^2.$$

This is a spectral definition, and the corresponding spatial definition is

$$\mathrm{Inf}_i[f] = p(1-p)\,\mathbb{E}[(f(x) - f(x \oplus e_i))^2].$$

Indeed, let us compute the Fourier expansion of $g = f|_{x_i=1} - f|_{x_i=0}$. If $i \in S$ then it is easy to check that $\hat{g}(S) = 0$ (indeed, g doesn't depend on i). If $i \notin S$ then

$$\hat{g}(S) = \mathop{\mathbb{E}}_{x_{-i}}[(f(x_{-i},1) - f(x_{-i},0))\omega_S(x_{-i})],$$

whereas

$$\hat{f}(S \cup \{i\}) = \mathop{\mathbb{E}}_{x_{-i}}\left[-\left(p\sqrt{\frac{1-p}{p}}f(x_{-i},1) + (1-p)\sqrt{\frac{p}{1-p}}f(x_{-i},0)\right)\omega_S(x_{-i})\right] = -\sqrt{p(1-p)}\hat{g}(S).$$

There is also a simpler formula for monotone Boolean f. Let $g = f|_{x_i=1} - f|_{x_i=0}$. When f is monotone, $g^2 = g$, and so

$$\mathrm{Inf}_i[f] = p(1-p)\,\mathbb{E}[g^2] = p(1-p)\,\mathbb{E}[g] = p(1-p)\hat{g}(\emptyset) = -\sqrt{p(1-p)}\hat{f}(\{i\}).$$

For noise stability, we similarly choose the normalization satisfying

$$T_\rho f = \sum_S \rho^{|S|}\hat{f}(S)\omega_S.$$

What is the corresponding spatial definition? Let $y_i = x_i$ with probability ρ, and $y_i = z_i \sim \mu_p$ with probability $1-\rho$. This extends the previous definition, since when $p = 1/2$, we get that $y_i = x_1$ with probability $\rho + (1-\rho)/2 = (1+\rho)/2$. We claim that $(T_\rho f)(x) = \mathbb{E}_y[f(y)]$. To check this, it is enough to verify that the formula works for the characters $\omega_\emptyset$ and $\omega_{\{1\}}$ (why?). Clearly $T_\rho\omega_\emptyset = \omega_\emptyset$, and $(T_\rho\omega_{\{1\}})(x) = \rho\omega_{\{1\}}(x) + (1-\rho)\,\mathbb{E}[\omega_{\{1\}}(z)] = \rho\omega_{\{1\}}(x)$, due to orthogonality of characters.

Russo–Margulis Let $f\colon \{0,1\}^n \to \{0,1\}$ be a Boolean function. How does the measure of f change with respect to p? To find the answer, express f as a multilinear polynomial in $x_1, \ldots, x_n$. Recall that one way of showing that such a representation exists is using the formula

$$f(x) = \sum_y f(y)\delta_y(x), \quad \delta_y(x) = \prod_{i=1}^n \begin{cases} 1-x_i & y_i = 0, \\ x_i & y_i = 1. \end{cases}$$

This shows that

$$f(p,\ldots,p) = \sum_y f(y)\delta_y(p,\ldots,p) = \sum_y f(y)\mu_p(y) = \mathop{\mathbb{E}}_{\mu_o}[f].$$

We are going to calculate $\frac{\partial f}{\partial x_i}$. Since f is multilinear, it is easy to check that

$$\frac{\partial f}{\partial x_i}(x_{-i}) = f(x_{-i},1) - f(x_{-i},0).$$

Indeed, a monomial not depending on x_i just vanishes, and a monomial of the form $x_i m$ turns to $m - 0 = m$. Now, we have already seen this formula: it is the function g from above! So

$$\mathop{\mathbb{E}}_{\mu_p}\left[\frac{\partial f}{\partial x_i}\right] = \mathop{\mathbb{E}}_{\mu_p}[g] = \hat{g}(\emptyset) = -\frac{1}{\sqrt{p(1-p)}}\hat{f}_p(\{i\}),$$

where $\hat{f}_p$ signifies which Fourier expansion we are talking about. Applying the chain rule,

$$\frac{\partial}{\partial p}\mathop{\mathbb{E}}_{\mu_p}[f] = \frac{\partial}{\partial p}f(p,\ldots,p) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(p,\ldots,p) = \sum_{i=1}^n \mathop{\mathbb{E}}_{\mu_p}\left[\frac{\partial f}{\partial x_i}\right] = -\frac{1}{\sqrt{p(1-p)}}\sum_{i=1}^n \hat{f}_p(\{i\}).$$

When f is furthermore monotone and Boolean, we obtain the simple formula

$$\frac{\partial}{\partial p}\mathop{\mathbb{E}}_{\mu_p}[f] = \frac{\mathrm{Inf}^{(p)}[f]}{p(1-p)} = \sum_{i=1}^n \Pr_{x\sim\mu_p}[f(x) \neq f(x \oplus e_i)].$$

We will see later on what the Russo–Margulis lemma is good for.

3.1 Erdős–Ko–Rado

What is the μ_p Fourier basis good for? It allows a simple proof of a version of the Erdős–Ko–Rado theorem [EKR61]. It's not quite the classical version, but it's definitely cleaner.

Let $\mathcal{F}$ be a family of subsets of $[n]$. We say that $\mathcal{F}$ is *intersecting* if whenever $S, T \in \mathcal{F}$ then $S \cap T \neq \emptyset$. How large can the measure of $\mathcal{F}$ be? If $p > 1/2$ then we can take all sets of size at least $n/2 + 1$. This is an intersecting family, and the central limit theorem implies that its measure tends to 1 as p tends to ∞. So the problem is only interesting for $p \leq 1/2$.

Can you think of a "large" intersecting family? One example is a *star*: all sets containing some point i. A star has measure p. The Erdős–Ko–Rado theorem states that no family has larger measure, and when $p < 1/2$, stars are the unique maximizers (when $p = 1/2$ there are other optimal families, such as those intersecting $[2t + 1]$ in at least $t + 1$ points).

Here is a magical spectral proof due to Friedgut. We will design a matrix A with the following characteristics:

1. If f is the characteristic vector of an intersecting family then $\langle f, Af \rangle_p = 0$ (here $\langle f, g \rangle_p = \mathbb{E}_{\mu_p}[f(x)g(x)]$). This is equivalent to $(\mu_p f)' Af = 0$, where $(\mu_p f)(x) = \mu_p(x) f(x)$.
2. The eigenvectors of A are the ω_S.

We'll see later why we want these properties; they had been identified previously by Hoffman and Lovász. How do we find the matrix A? We solve the one-dimensional problem, and tensorise. Let B be the matrix for $n = 1$:

$$B = \begin{pmatrix} \alpha & \beta \\ \gamma & 0 \end{pmatrix}.$$

Note that at the corner we must have 0, to accommodate the intersecting family $\{1\}$. We can assume without loss of generality that the eigenvalue corresponding to $\omega_\emptyset$ is 1, and so $\alpha + \beta = 1$ and $\gamma = 1$. What about the other eigenvector? A vector $\begin{pmatrix} x & y \end{pmatrix}$ is a multiple of $\omega_{\{1\}}$ if $(1-p)x + py = 0$. One of the solutions is $x = p$, $y = -(1-p)$, and so we must have

$$0 = (1-p)(p\alpha - (1-p)(1-\alpha)) + p^2 = (1-p)\alpha + p^2 - (1-p)^2 = (1-p)\alpha + 2p - 1.$$

We conclude that $\alpha = \frac{1-2p}{1-p}$ and so $\beta = 1 - \alpha = \frac{p}{1-p}$. Altogether,

$$B = \begin{pmatrix} \frac{1-2p}{1-p} & \frac{p}{1-p} \\ 1 & 0 \end{pmatrix}.$$

What are the eigenvalues corresponding to $\omega_\emptyset$ and $\omega_{\{1\}}$? We have already calculated $\lambda_\emptyset = 1$, and it is easy to see that $\lambda = \lambda_{\{1\}} = \frac{\beta}{\gamma} = -\frac{p}{1-p}$.

Now that we have the matrix B, how do we form the matrix A? We use the tensor product construction once again! This is sometimes known as the *Kronecker product* of matrices (illustrate on the board). It is not hard to check that the eigenvectors and eigenvalues also tensorise — for the latter, tensorization is just multiplication! So $A = B^{\otimes n}$ (the n-fold tensor product of B) has ω_S as eigenvector, with eigenvalue $\lambda^{|S|}$. This implies that $\widehat{Af}(S) = \lambda^{|S|} \hat{f}(S)$, and so if f is intersecting,

$$0 = \langle f, Af \rangle_p = \sum_S \lambda^{|S|} \hat{f}(S)^2.$$

Note that $\lambda^{|S|} \geq \lambda$, since $\lambda \geq -1$.

Suppose that f is the characteristic vector of a family $\mathcal{F}$. It's easy to check that $\hat{f}(\emptyset) = \mu_p(\mathcal{F})$. Since $f^2 = f$, we also have $\mu_p(\mathcal{F}) = \|f\|^2 = \sum_S \hat{f}(S)^2$. Therefore

$$0 = \mu_p(\mathcal{F})^2 + \sum_{S \neq \emptyset} \lambda^{|S|} \hat{f}(S)^2 \geq \mu_p(\mathcal{F})^2 + \lambda \sum_{S \neq \emptyset} \hat{f}(S)^2 = \mu_p(\mathcal{F})^2 + \lambda(\mu_p(\mathcal{F}) - \mu_p(\mathcal{F}))^2.$$

Therefore $\mu_p(\mathcal{F})^2 \leq -\lambda(\mu_p(\mathcal{F}) - \mu_p(\mathcal{F})^2)$, and so $(1 + \lambda)\mu_p(\mathcal{F}) \leq -\lambda$, implying

$$\mu_p(\mathcal{F}) \leq \frac{-\lambda}{1 - \lambda} = \frac{\frac{p}{1-p}}{\frac{1}{1-p}} = p.$$

We got the bound we wanted!

Uniqueness The proof is a bit long, but we can get more out of it. When $p < 1/2$, $\lambda > -1$, and so $\mu_p(\mathcal{F}) = p$ is possible only if all inequalities were strict, that is, if $\hat{f}(S) = 0$ for $|S| > 1$. Therefore

$$f(x_1, \ldots, x_n) = \mu_p(\mathcal{F}) + \sum_{i=1}^{n} c_i \omega_{\{i\}}(x_i),$$

for some coefficients c_i. Since f is Boolean, it's easy to check that at most one c_i can be non-zero. This implies that f is a dictatorship.

Stability The big advantage of this proof is that it implies even more. Suppose that $p < 1/2$ and that $\mu_p(\mathcal{F}) \geq p - \epsilon$. Then *most* of the Fourier weight should be concentrted on the Fourier coefficients $\hat{f}(S)$ for $|S| \leq 1$. In fact, one can calculate that all but an $O_p(\epsilon)$-fraction of the weight lies there. Therefore f is $O_p(\epsilon)$-close to a *dictatorship* g, which however need not be Boolean! The fundamental theorem of Friedgut, Kalai and Naor implies that f is $O_p(\epsilon)$-close to a *Boolean* dictatorship h, which must be a star. So the only way that $\mu_p(\mathcal{F})$ can be close to p is if it's close to an optimal family! This is known as *stability*.

Katona's proof There is a much simpler proof due to Katona [Kat72], using a method known as *Katona's circle method*. Take the unit circumference circle, and a window of length p sitting on it. Throw n points randomly on the circumference. The set of points lying inside the window has distribution μ_p, and so $\mu_p(\mathcal{F})$ is the probability that the set inside the window is in $\mathcal{F}$.

We can generate the same distribution by throwing n points and then placing the window randomly on the circumference. I claim that for any *deterministic* point locations, the measure of window locations corresponding to sets in $\mathcal{F}$ is at most p, which immediately implies the bound $\mu_p(\mathcal{F}) \leq p$. Indeed, consider some placement of the window for which the contents is in $\mathcal{F}$; we call such a location a $\mathcal{F}$-*location*. Rotate it clockwise all the way while keeping it a $\mathcal{F}$-location. Now rotate it counter-clockwise all the way. Since $\mathcal{F}$ is intersecting, any two $\mathcal{F}$-locations must intersect (as windows). This makes it clear then when rotating the window counter-clockwise, we could have rotated it at most by p, since beyond that the windows are disjoint. This completes the proof.

Uniform families The classical Erdős–Ko–Rado theorem is stated a bit differently. Let $\binom{[n]}{k}$ be the collection of all subsets of $[n] = \{1, \ldots, n\}$ of cardinality k. When $k > n/2$, all sets in $\binom{[n]}{k}$ intersect. The Erdős–Ko–Rado theorem states that when $k \leq n/2$, an intersecting subfamily of $\binom{[n]}{k}$ has cardinality at most $\binom{n-1}{k-1}$, and this is attained for stars (when $k < n/2$, *only* for stars).

Dinur and Safra [DS05] showed how to derive the μ_p-version of the theorem from its uniform version. Let $p < 1/2$, and let $\mathcal{F}$ be an intersecting family on n points. The idea is that we can view $\mathcal{F}$ also as an intersecting family on N points for $N > n$. Taking $N \to \infty$ will allow us to recover the theorem, using the following calculation:

$$\begin{aligned}
\mu_p(\mathcal{F}) &= \sum_{k=0}^{N} p^k (1-p)^{N-k} \left| \mathcal{F} \cap \binom{[N]}{k} \right| \\
&\leq \sum_{k=1}^{N/2} p^k (1-p)^{N-k} \binom{N-1}{k-1} + \sum_{k=N/2+1}^{N} p^k (1-p)^{N-k} \binom{N}{k} \\
&\leq \sum_{k=1}^{N-1} p^{(k-1)+1} (1-p)^{(N-1)-(k-1)} \binom{N-1}{k-1} + \Pr[\mathrm{Bin}(N,p) > N/2] \\
&= p + \Pr[\mathrm{Bin}(N,p) > N/2].
\end{aligned}$$

Since $p < 1/2$, as $N \to \infty$ the error terms tends to 0, and we deduce $\mu_p(\mathcal{F}) \leq p$.

We can also go the other way but the argument is more complicated and there is some loss (see Friedgut [Fri08]).

t-intersecting families There are many extensions of the basic Erdős–Ko–Rado theorem. For example, Ahlswede and Khachatrian [AK97, AK99] proved the optimal analog for *t-intersecting families*, in which any two sets must have at least t points in common. They show that if $\mathcal{F} \subseteq \binom{[n]}{k}$ is t-intersecting and for some integer $r \geq 0$,

$$(k-t+1)\left(2+\frac{t-1}{r+1}\right) \leq n \leq (k-t+1)\left(2+\frac{t-1}{r}\right),$$

then $|\mathcal{F}|$ is at most the size of the family

$$\mathcal{F}_{t,r} = \left\{A \in \binom{[n]}{k} : |A \cap [t+2r]| \geq [t+r]\right\}.$$

Furthermore, if n is strictly inside this interval then $\mathcal{F}_{t,r}$ is the unique optimum (up to coordinate renaming), and at the "breakpoints" there are two optimal families $\mathcal{F}_{t,r}$ and $\mathcal{F}_{t,r\pm1}$ (depending on the end of the interval).

The μ_p-analog of the Ahlswede–Khachatrian theorem is nicer to state: if $\mathcal{F}$ is t-intersecting and for some integer $r \geq 0$,

$$\frac{r}{t+2r-1} \leq p \leq \frac{r+1}{t+2r+1},$$

then $\mu_p(\mathcal{F})$ is at most the μ_p-measure of the family

$$\mathcal{F}_{t,r} = \{A \subseteq [t+2r] : |A| \geq t+r\}.$$

Again this family is unique (for the "breakpoints" this is first proved in [Fil13]).

Wilson [Wil84] and Friedgut [Fri08] gave spectral proofs of the case $r = 0$, Wilson in the uniform setting and Friedgut in the μ_p setting. We don't know how to prove the cases $r \geq 1$ using spectral methods.

4 Hypercontractivity (11 November 2015)

([O'D14, Sections 9.1,9.2,9.4,9.5,10.1].)

In this lecture we will concentrate on the classical Fourier expansion, though everything also works for the skewed Fourier expansion. For this reason, our functions will be on the domain $\{-1,1\}^n$. We will frequently use L_p norms, defined by $\|f\|_p = \mathbb{E}[|f|^p]^{1/p}$. For $p \geq 1$, these norms satisfy the triangle inequality.

We will require the classical Hölder inequality:

$$\langle f, g\rangle \leq \|f\|_p \|g\|_q, \text{ for } \frac{1}{p} + \frac{1}{q} = 1, \quad 1 \leq p, q \leq \infty.$$

This inequality also implies a dual definition of the L_p norm:

$$\|f\|_p = \sup_{g \neq 0} \frac{\langle f, g\rangle}{\|g\|_q}.$$

The $\geq$ direction follows immediately from the inequality. For the $\leq$ direction, take $g = f^{p-1}$, so that $\langle f, g\rangle = \|f\|_p^p$. On the other hand, since $q = p/(p-1)$, $\|g\|_q^q = \mathbb{E}[g^q] = \mathbb{E}[f^p] = \|f\|_p^p$, and so the right-hand side is $\|f\|_p^{p(1-1/q)} = \|f\|_p$.

More generally,

$$\langle f, g\rangle \leq \|f\|_p^\alpha \|g\|_q^\beta, \text{ for } \frac{\alpha}{p} + \frac{\beta}{q} = 1, \alpha + \beta = 2, \quad 1 \leq p, q \leq \infty, 0 \leq \alpha, \beta.$$

Contractivity One of the fundamental techniques in analysis of Boolean functions is *hypercontractivity*, which states that applying noise to a function smoothens it. Often, this is applied to low-degree polynomials, and the conclusion is that low-degree polynomials are *reasonable*, that is, they don't behave too crazily.

Before getting into *hyper*contractivity, let us mention a much simpler property, namely, *contractivity*. What happens when we apply the noise operator T_ρ to a function f (for $|\rho| \leq 1$)?

$$\|T_\rho f\|^2 = \left\| \sum_S \rho^{|S|} \hat{f}(S) \chi_S \right\|^2 = \sum_S \rho^{2|S|} \hat{f}(S)^2 \leq \sum_S \hat{f}(S)^2 = \|f\|^2.$$

So T_ρ contracts the L2 norm of f.

Another way of seeing this is using the formula

$$(T_\rho f)(x) = \mathop{\mathbb{E}}_{y \sim N_\rho(x)}[f(y)],$$

where y is obtained from x by flipping each bit with probability $\frac{1-\rho}{2}$. Another way of stating this is

$$T_\rho f = \mathop{\mathbb{E}}_{z \sim N_\rho(\mathbf{1})}[f^{\oplus z}], \quad \text{where } f^{\oplus z}(x) = f(x_1 z_1, \ldots, x_n z_n).$$

The triangle inequality implies that

$$\|T_\rho f\| = \| \mathop{\mathbb{E}}_{z \sim N_\rho(\mathbf{1})}[f^{\oplus z}]\| \leq \mathop{\mathbb{E}}_{z \sim N_\rho(\mathbf{1})}[\|f^{\oplus z}\|] = \|f\|,$$

since f and $f^{\oplus z}$ have the same norm (here we use the fact that the uniform measure on $\{-1,1\}^n$ is invariant under XOR). This inequality actually holds for *every* L_p norm, not only L_2.

Hypercontractivity I Surprisingly, we can say more:

$$\|T_{1/\sqrt{3}} f\|_4 \leq \|f\|_2.$$

In words, $T_{1/\sqrt{3}}$ smoothens f so much that its L4 norm becomes comparable to its original L2 norm! The proof is by induction. The case $n = 0$ is obvious, so consider some $n \geq 1$. Write

$$f(x_1, \ldots, x_n) = x_n g(x_1, \ldots, x_{n-1}) + h(x_1, \ldots, x_{n-1}).$$

Why is this decomposition a good idea? Since

$$\begin{aligned} T_{1/\sqrt{3}}f &= \frac{1+\frac{1}{\sqrt{3}}}{2}\left[x_n T_{1/\sqrt{3}}g + T_{1/\sqrt{3}}h\right] + \frac{1-\frac{1}{\sqrt{3}}}{2}\left[-x_n T_{1/\sqrt{3}}g + T_{1/\sqrt{3}}h\right] \\ &= \frac{1}{\sqrt{3}}T_{1/\sqrt{3}}x_n g + T_{1/\sqrt{3}}h. \end{aligned}$$

Another way of seeing that is through the spectral formula for T_ρ.

Using $\mathbb{E}[x_n] = \mathbb{E}[x_n^3] = 0$ and $\mathbb{E}[x_n^2] = \mathbb{E}[x_n^4] = 1$, we calculate

$$\begin{aligned} \mathbb{E}[(T_{1/\sqrt{3}}f)^4] &= \mathbb{E}\left[\left(\frac{1}{\sqrt{3}}x_n T_{1/\sqrt{3}}g + T_{1/\sqrt{3}}h\right)^4\right] \\ &= \frac{1}{9}\,\mathbb{E}[(T_{1/\sqrt{3}}g)^4] + \frac{6}{3}\,\mathbb{E}[(T_{1/\sqrt{3}}g)^2(T_{1/\sqrt{3}}h)^2] + \mathbb{E}[(T_{1/\sqrt{3}}h)^4] \\ &\le \frac{1}{9}\,\mathbb{E}[(T_{1/\sqrt{3}}g)^4] + 2\sqrt{\mathbb{E}[(T_{1/\sqrt{3}}g)^4]\,\mathbb{E}[(T_{1/\sqrt{3}}h)^4]} + \mathbb{E}[(T_{1/\sqrt{3}}h)^4] \\ &\le \left(\sqrt{\mathbb{E}[(T_{1/\sqrt{3}}g)^4]} + \sqrt{\mathbb{E}[(T_{1/\sqrt{3}}h)^4]}\right)^2, \end{aligned}$$

using Cauchy–Schwartz. Applying the induction hypothesis,

$$\mathbb{E}[(T_{1/\sqrt{3}}f)^4] \le (\mathbb{E}[g^2] + \mathbb{E}[h^2])^2 = \mathbb{E}[f^2]^2.$$

Taking fourth roots, we obtain $\|T_{1/\sqrt{3}}f\|_4 \le \|f\|_2$.

Applying Hölder's inequality The hypercontractive inequality we have just proved has L2 norm on the right. We can get a similar hypercontractive inequality with an L2 norm on the left using Hölder's inequality. First, notice that for every ρ,

$$\|T_\rho f\|_2^2 = \sum_S \rho^{2|S|}\hat{f}(S)^2 = \langle f, T_\rho^2 f\rangle.$$

Applying this for $\rho = 1/\sqrt{3}$ together with Hölder's inequality (using the fact that $4/3$ and 4 are conjugate norms), we get

$$\|T_{1/\sqrt{3}}f\|_2^2 = \langle f, T_{1/\sqrt{3}}^2 f\rangle \le \|f\|_{4/3}\|T_{1/\sqrt{3}}^2 f\|_4.$$

Applying our earlier hypercontractivity result, we deduce

$$\|T_{1/\sqrt{3}}f\|_2^2 \le \|f\|_{4/3}\|T_{1/\sqrt{3}}f\|_2,$$

and so

$$\|T_{1/\sqrt{3}}f\|_2 \le \|f\|_{4/3}.$$

Low-degree functions are reasonable Suppose that f has degree d. Then for $\rho > 1$ (say $\rho = \sqrt{3}$),

$$\|T_\rho f\|_2^2 = \sum_S \rho^{2|S|}\hat{f}(S)^2 \le \rho^{2d}\|f\|_2^2.$$

Similarly, for $\rho \in (0,1)$ we have $\|T_\rho f\|_2^2 \ge \rho^{2d}\|f\|_2^2$. Combining this with our hypercontractivity results, we deduce

$$\begin{aligned} &\|f\|_4 = \|T_{1/\sqrt{3}}T_{\sqrt{3}}f\|_4 \le \|T_{\sqrt{3}}f\|_2 \le \sqrt{3}^d\|f\|_2, \\ &\frac{1}{\sqrt{3}^d}\|f\|_2 \le \|T_{1/\sqrt{3}}f\|_2 \le \|f\|_{4/3}. \end{aligned}$$

Hypercontractivity II With more work, we can deduce similar inequalities involving other pairs of norms: for all $1 \le p \le q \le \infty$ and $f\colon \{-1,1\}^n \to \mathbb{R}$,

$$\|T_\rho f\|_q \le \|f\|_p \text{ for all } 0 \le \rho \le \sqrt{\frac{p-1}{q-1}}.$$

The starting point is the following elementary inequality, which holds for all $1 \le p \le q \le \infty$ and all functions $f\colon \{-1,1\} \to \mathbb{R}$:

$$\|T_\rho f\|_q \le \|f\|_p \text{ for all } 0 \le \rho \le \sqrt{(p-1)/(q-1)}.$$

The proof is not very enlightening so we skip it. It is usually known as the "two-point inequality", since if we write $f(x) = a + bx$ then we can write it as

$$\sqrt[q]{\frac{(a+\rho b)^q + (a-\rho b)^q}{2}} \le \sqrt[p]{\frac{(a+b)^p + (a-b)^p}{2}},$$

which is a (parametric) inequality depending only on the two parameters a, b (and really only on one of them, since we can fix $a = 1$).

Applying Hölder's inequality, we deduce that for any $f, g \in \{-1,1\} \to \mathbb{R}$,

$$\langle T_\rho f, g\rangle \le \|T_\rho f\|_q \|g\|_{q/(q-1)} \le \|f\|_p \|g\|_{q/(q-1)}.$$

The crucial point is that this inequality tensorises. That is, if it works for functions $\{-1,1\} \to \mathbb{R}$, then it also works for functions $\{-1,1\}^n \to \mathbb{R}$ for every n. We prove this by induction on n.

We already have the case $n = 1$. For general n, write $x = (x', x_n)$, $y = (y', y_n)$, and let $f_b(x') = f(x', b)$, $g_b(y') = g(y', b)$. Recall that

$$\langle T_\rho f, g\rangle = \mathop{\mathbb{E}}_{x,y\sim N_\rho}[f(x)g(y)].$$

Since N_ρ is a product distribution, we can write

$$\langle T_\rho f, g\rangle = \mathop{\mathbb{E}}_{x_n,y_n\sim N_\rho} \mathop{\mathbb{E}}_{x',y'\sim N_\rho}[f_{x_n}(x')g_{y_n}(y')] \le \mathop{\mathbb{E}}_{x_n,y_n\sim N_\rho}[\|f_{x_n}\|_p \|g_{y_n}\|_{q/(q-1)}],$$

using the induction hypothesis. Now write $F(x) = \|f_x\|_p$ and $G(y) = \|g_y\|_{q/(q-1)}$. Applying the case $n = 1$ again, we have

$$\langle T_\rho f, g\rangle \le \mathop{\mathbb{E}}_{x,y\sim N_\rho}[F(x)G(y)] \le \|F\|_p \|G\|_{q/(q-1)} = \|f\|_p \|g\|_{q/(q-1)},$$

since for example

$$\|F\|_p^p = \mathop{\mathbb{E}}_x[F(x)^p] = \mathop{\mathbb{E}}_x \mathop{\mathbb{E}}_{x'}[f_x(x')^p] = \|f\|_p^p.$$

So we have proved that for all $f, g \in \{-1,1\}^n \to \mathbb{R}$,

$$\langle T_\rho f, g\rangle \le \|f\|_p \|g\|_{q/(q-1)}.$$

Finally, to deduce the actual hypercontractivity result, we need to use the dual definition of L_q norm:

$$\|T_\rho f\|_q = \sup_{g\ne 0} \frac{\langle T_\rho f, g\rangle}{\|g\|_{q/(q-1)}} \le \sup_{g\ne 0} \|f\|_p.$$

It is useful to consider the cases $p = 2$ and $q = 2$:

$$\|T_{1/\sqrt{q-1}} f\|_q \le \|f\|_2, \quad \|T_{\sqrt{p-1}} f\|_2 \le \|f\|_p.$$

As before, when $\deg f = d$ we deduce

$$\|f\|_q \le \sqrt{q-1}^d \|f\|_2, \quad \|f\|_2 \le \sqrt{p-1}^{-d} \|f\|_p.$$

4.1 Applications

Large deviation bounds Hypercontractivity implies that low-degree polynomials satisfy large deviation bounds: they are rarely very far from their mean (as measured in standard deviations). One such large deviation bound states that for all $t \geq \sqrt{2e}^d$,

$$\Pr[|f| \geq t\|f\|_2] \leq \exp(-\frac{d}{2e}t^{2/d}).$$

When $d = 1$, we get the expected e^{-t^2} dependency familiar from the central limit theorem. For higher-degree polynomials, the tails decay slower, but still very fast. The proof follows the usual method of moments. Assume for simplicity that $\|f\|_2 = 1$. For an appropriate $q \geq 2$, Markov's inequality gives

$$\Pr[|f| \geq t] = \Pr[|f|^q \geq t^q] \leq \frac{\mathbb{E}[|f|^q]}{t^q} = \frac{\|f\|_q^q}{t^q}.$$

Since $\|f\|_q \leq \sqrt{q-1}^d \|f\|_2 = \sqrt{q-1}^d$, we deduce that

$$\Pr[|f| \geq t] \leq ((q-1)^{d/2}/t)^q.$$

Choosing $q = t^{2/d}/e$, the bound is at most

$$(q^{d/2}/t)^q = (1/e^{d/2})^q = \exp(-\frac{d}{2}q) = \exp(-\frac{d}{2e}t^{2/d}).$$

The condition $t \geq \sqrt{2e}^d$ ensures that $q = t^{2/d}/e \geq 2$.

L2 norm versus L1 norm Hypercontractivity shows that for a low-degree polynomial, the L2 and L1 norms are comparable: if $\deg f = d$ then

$$\|f\|_2 \leq e^d \|f\|_1.$$

Choose $\epsilon > 0$. The generalized Hölder inequality implies that

$$\langle f, f \rangle \leq \|f\|_{2+\epsilon}^{(2+\epsilon)/(1+\epsilon)} \|f\|_1^{\epsilon/(1+\epsilon)}.$$

Hypercontractivity shows that $\|f\|_{2+\epsilon} \leq \sqrt{1+\epsilon}^d \|f\|_2$, and so

$$\|f\|_2^2 \leq \sqrt{1+\epsilon}^{d(2+\epsilon)/(1+\epsilon)} \|f\|_2^{(2+\epsilon)/(1+\epsilon)} \|f\|_1^{\epsilon/(1+\epsilon)}.$$

Rearranging,

$$\|f\|_2^{\epsilon/(1+\epsilon)} \leq \sqrt{1+\epsilon}^{d(2+\epsilon)/(1+\epsilon)} \|f\|_1^{\epsilon/(1+\epsilon)}.$$

Raising to the power $(1+\epsilon)/\epsilon$,

$$\|f\|_2 \leq \sqrt{1+\epsilon}^{d(2+\epsilon)/\epsilon} \|f\|_1.$$

As $\epsilon \to 0$, it is easy to check that $(1+\epsilon)^{(2+\epsilon)/\epsilon} \to e^2$, and so

$$\|f\|_2 \leq e^d \|f\|_1.$$

As an application, we can lower bound the probability that f exceeds its expectation. Consider for simplicity the case $\mathbb{E}[f] = 0$. Since $\mathbb{E}[f1_{f>0}] = -\mathbb{E}[f1_{f<0}]$, we see that $\mathbb{E}[f1_{f>0}] = \|f\|_1/2$. Cauchy–Schwartz implies that

$$\frac{1}{2}\|f\|_1 = \mathbb{E}[f1_{f>0}] \leq \|f\|_2 \sqrt{\Pr[f>0]} \leq e^d \|f\|_1 \sqrt{\Pr[f>0]},$$

and we deduce that unless f is constant,

$$\Pr[f > 0] \geq \frac{e^{-2d}}{4}.$$

Anticoncentration Can low-degree functions be very concentrated around any point? Hypercontractivity shows this is impossible. For the proof, we need the Paley–Zygmund inequality: if $Z \geq 0$ then for all $\theta \in (0,1)$,

$$\Pr[Z \geq \theta \mathbb{E}[Z]] \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

The idea is to write $Z = Z_{<\theta \mathbb{E}[Z]} + Z_{\geq \theta \mathbb{E}[Z]}$. Clearly

$$\mathbb{E}[Z] = \mathbb{E}[Z_{<\theta \mathbb{E}[Z]}] + \mathbb{E}[Z_{\geq \theta \mathbb{E}[Z]}] \leq \theta \mathbb{E}[Z] + \mathbb{E}[Z 1_{Z \geq \theta \mathbb{E}[Z]}].$$

Applying Cauchy–Schwartz, we can bound the second term by $\sqrt{\mathbb{E}[Z^2]}\sqrt{\Pr[Z \geq \theta \mathbb{E}[Z]]}$, and so

$$(1-\theta)^2 \mathbb{E}[Z]^2 \leq \mathbb{E}[Z^2] \Pr[Z \geq \theta \mathbb{E}[Z]],$$

implying the inequality.

Suppose $\deg f = d$. Hypercontractivity implies that $\|f\|_4 \leq \sqrt{3}^d \|f\|_2$, and so the Paley–Zygmund inequality (with $\theta = 1/4$) implies that

$$\Pr[|f| \geq \|f\|_2/2] = \Pr[f^2 \geq \|f\|_2^2/4] \geq \frac{9}{16} \frac{\|f\|_2^4}{\|f\|_4^4} \geq \frac{9^{1-d}}{16}.$$

In particular,

$$\Pr[|f - \mathbb{E}[f]| \geq \sqrt{\mathbb{V}[F]}/2] \geq \frac{9^{1-d}}{16}.$$

Small set expansion Suppose that $A \subseteq \{-1,1\}^n$ has measure α, that we pick some random point in A, and then apply some noise. What is the probability that we stay inside A? This probability turns out to be small, and this property is known as *small-set expansion*. To see this, let $f = 1_A$, so that $\mathbb{E}[f] = \alpha$. Hypercontractivity with $q = 2$ shows that

$$\|T_{\sqrt{p-1}} f\|_2 \leq \|f\|_p = \alpha^{1/p}.$$

On the other hand,

$$\|T_{\sqrt{p-1}} f\|_2^2 = \langle f, T_{p-1} f \rangle = \Pr_{x,y \in N_{p-1}}[x, y \in A] = \alpha \Pr_{\substack{x \in A, \\ y \in N_{p-1}(x)}}[y \in A].$$

If we put $p = 1 + \rho$, then this implies that

$$\alpha \Pr_{\substack{x \in A, \\ y \in N_\rho(x)}}[y \in A] \leq \alpha^{2/(1+\rho)},$$

or equivalently,

$$\Pr_{\substack{x \in A, \\ y \in N_\rho(x)}}[y \in A] \leq \alpha^{(1-\rho)/(1+\rho)}.$$

Biased Fourier expansion So far we have only discussed the uniform distribution on $\{-1,1\}^n$, but everything carries over to μ_p for general p. In particular, for every $q \geq 2 \geq r$ there exists some $\rho > 0$ such that

$$\|T_\rho f\|_q \leq \|f\|_r,$$

all norms with respect to μ_p. There is an explicit expression for ρ, but usually it is not needed; all we need is that ρ is continuous in q, r.

5 Consequences (18 November 2015)

([O'D14, Sections 9.1,9.6,10.3].)

Hypercontractivity is essential in the proof of many basic theorems in analysis of Boolean functions.

5.1 Friedgut–Kalai–Naor

Suppose that a function $f\colon \{-1,1\}^n \to \{-1,1\}$ is *affine*, that is, of the form $f(x_1,\ldots,x_n) = a_0 + \sum_i a_i x_i$. What can we say about the coefficients? It is not hard to see that at most one x_i can be non-zero, and so f must be constant or a *dictatorship*: a function depending on a single coordinate. The Friedgut–Kalai–Naor theorem [FKN02] is a relaxed version of this result, stating that if f is *almost* affine, then f is *close* to a dictatorship. While the theorem can be proved directly (as was done in the original paper as well as by Kindler and Safra [KS04]), it has a particularly simple proof using hypercontractivity.

The Friedgut–Kalai–Naor theorem states that if $f\colon \{-1,1\}^n \to \{-1,1\}$ satisfies $\|f^{>1}\|^2 = \epsilon$ then there exists a Boolean dictatorship D such that $\|f - D\|^2 = O(\epsilon)$, or equivalently, $\Pr[f \neq D] = O(\epsilon)$. For the proof, we will assume that ϵ is "small enough", since otherwise the theorem becomes trivial.

Reduction to odd case The first step is to define an auxiliary function which has zero mean:

$$g(x_0,x_1,\ldots,x_n) = x_0 f(x_0x_1,\ldots,x_0,x_n) = \sum_{\substack{S\subseteq[n]\\ |S| \text{ even}}} \hat{f}(S)\chi_{S\cup\{0\}} + \sum_{\substack{S\subseteq[n]\\ |S| \text{ odd}}} \hat{f}(S)\chi_S.$$

This function clearly has zero mean, but is it Boolean? Let's check:

$$\begin{aligned} g(1,x_1,\ldots,x_n) &= \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x_1,\ldots,x_n) = f(x_1,\ldots,x_n), \\ g(-1,x_1,\ldots,x_n) &= -\sum_{S\subseteq[n]} \hat{f}(S)\chi_S(-x_1,\ldots,-x_n) = -f(-x_1,\ldots,-x_n). \end{aligned}$$

Also, clearly $\|g^{>1}\|^2 = \|f^{>1}\|^2 = \epsilon$.

Proof for odd functions Let $\ell = g^{=1}$. We know that $\mathbb{E}[\ell^2] = \|g^{=1}\|^2 = 1-\epsilon$, and the idea of the proof is to show that ℓ^2 is concentrated around $1-\epsilon$. Recall that we have shown that for all functions ϕ of degree d,

$$\Pr[|\phi - \mathbb{E}[\phi]| \geq \sqrt{\mathbb{V}[\phi]}/2] \geq \frac{9^{1-d}}{16}.$$

We apply this for the function $\phi = \ell^2$ which has degree 2, and deduce that

$$\Pr[|\ell^2 - (1-\epsilon)| \geq \sqrt{\mathbb{V}[\ell^2]}/2] = \Omega(1).$$

On the other hand, we know that

$$\mathbb{E}[(|\ell|-1)^2] \leq \mathbb{E}[(\ell - g)^2] = \epsilon.$$

The idea now is that if ℓ^2 is far from $1-\epsilon$ then $|\ell|$ is far from 1. Indeed, suppose that $|\ell^2 - (1-\epsilon)| \geq (C+1)\sqrt{\epsilon}$. Then either $\ell^2 \geq 1 + C\sqrt{\epsilon}$ or $\ell^2 \leq 1 - C\sqrt{\epsilon}$. For small ϵ (small as a function of C!), this means that either $\ell \geq 1 + (C/3)\sqrt{\epsilon}$ or $\ell \leq 1 - (C/3)\sqrt{\epsilon}$, and in both cases $(|\ell|-1)^2 = \Omega(C\epsilon)$. Altogether, since this event happens with probability $\Omega(1)$, we deduce that $\mathbb{E}[(|\ell|-1)^2] = \Omega(C\epsilon)$, and so for an appropriate choice of C, we can conclude that

$$\sqrt{\mathbb{V}[\ell^2]}/2 \leq (C+1)\sqrt{\epsilon} \Longrightarrow \mathbb{V}[\ell^2] = O(\epsilon).$$

What is $\mathbb{V}[\ell^2]$? First, note that ℓ^2 has the Fourier expansion

$$\left(\sum_{i=0}^n \hat{g}(i)\chi_{\{i\}}\right)^2 = \sum_{i=0}^n \hat{g}(i)^2\chi_\emptyset + \sum_{i<j} 2\hat{g}(i)\hat{g}(j)\chi_{\{i,j\}},$$

Therefore

$$\mathbb{V}[\ell^2] = \sum_{i<j} 4\hat{g}(i)^2\hat{g}(j)^2 = 2\left(\sum_{i=0}^n \hat{g}(i)^2\right)^2 - 2\sum_{i=0}^n \hat{g}(i)^4 = 2(1-\epsilon)^2 - 2\sum_{i=0}^n \hat{g}(i)^4.$$

Since $\mathbb{V}[\ell^2] = O(\epsilon)$, we see that

$$\sum_{i=0}^n \hat{g}(i)^4 \geq (1-\epsilon)^2 - O(\epsilon) = 1 - O(\epsilon).$$

On the other hand,

$$\sum_{i=0}^n \hat{g}(i)^4 \leq \sum_{i=0}^n \hat{g}(i)^2 \max_{i=0}^n \hat{g}(i)^2 = (1-\epsilon)\max_{i=0}^n \hat{g}(i)^2.$$

We conclude that $\hat{g}(i)^2 \geq 1 - O(\epsilon)$ for some i, implying also that $|\hat{g}(i)| \geq 1 - O(\epsilon)$.

Let $s = \operatorname{sgn} \hat{g}(i)$. The dictatorship we're after is $D = s\chi_{\{i\}}$. It satisfies

$$\|g - D\|^2 = \sum_{j\neq i} \hat{g}(j)^2 + (\hat{g}(i) - s)^2 \leq O(\epsilon) + O(\epsilon^2) = O(\epsilon).$$

Only one thing is missing: we want to approximate f rather than g! If $i \neq 0$, then it is easy to check that in fact $\|f - D\|^2 = \|g - D\|^2$. Similarly, when $i = 0$, we need to replace D with the function s, obtaining $\|f - s\|^2 = \|g - D\|^2$.

More general statement It turns out that the Friedgut–Kalai–Naor theorem remains true under the slightly weaker assumption that f is an affine function satisfying $\mathbb{E}[(|f| - 1)^2] = \epsilon$. This is weaker since if f is a Boolean function satisfying $\|f^{>1}\|^2 = \epsilon$ then

$$\mathbb{E}[(|f^{\leq 1}| - 1)^2] \leq \mathbb{E}[(f^{\leq 1} - f)^2] = \epsilon.$$

Let us suppose, then, that f is an affine function satisfying $\mathbb{E}[(|f| - 1)^2] = \epsilon$, and define $F = \operatorname{sgn} f$. Note that

$$\|F^{>1}\|^2 = \mathbb{E}[(F^{\leq 1} - F)^2] \leq \mathbb{E}[(f - F)^2] = \mathbb{E}[(|f| - 1)^2] = \epsilon,$$

since $F^{\leq 1}$ is the projection of F to the span of $\chi_\emptyset, \chi_{\{1\}}, \ldots, \chi_{\{n\}}$. Applying the Friedgut–Kalai–Naor theorem to the function F, we obtain a Boolean function D such that $\|F - D\|^2 = O(\epsilon)$. The L2 triangle inequality then shows that

$$\|f - D\|^2 \leq 2\|f - F\|^2 + 2\|F - D\|^2 = O(\epsilon).$$

5.2 Kahn–Kalai–Linial

Suppose that $f\colon \{-1,1\}^n \to \{-1,1\}$ is balanced. Can it be that all the individual influences of f are small? A priori, all we know is that some coordinate satisfies $\operatorname{Inf}_i[f] \geq \operatorname{Inf}[f]/n$, which could be as small as $O(1/n)$. Surprisingly, Kahn, Kalai and Linial [KKL88] showed that there is always a coordinate i satisfying $\operatorname{Inf}_i[f] = \Omega(\frac{\log n}{n})$. More generally,

$$\max_i \operatorname{Inf}_i[f] = \Omega\left(\mathbb{V}[f]\frac{\log n}{n}\right).$$

The proof is a straightforward but mysterious application of hypercontractivity. Recall that $\operatorname{Inf}_i[f] = \|f_i\|^2$, where

$$f_i = \sum_{S\ni i} \hat{f}(S)\chi_S.$$

We can also define $f_i(x) = (f(x) - f(x \oplus e_i))/2$, and this shows that $f_i(x) \in \{0, 1, -1\}$. The basic idea is to apply hypercontractivity to the functions f_i and use the fact that $|f_i| \in \{0, 1\}$. Hypercontractivity shows that

$$\sum_{S\ni i} \hat{f}(S)^2 3^{-|S|} = \|T_{1/\sqrt{3}} f_i\|_2^2 \leq \|f_i\|_{4/3}^2 = \mathbb{E}[|f_i|^{4/3}]^{3/2} = \mathbb{E}[|f_i|^2]^{3/2} = \operatorname{Inf}_i[f]^{3/2}.$$

Summing over all i, we obtain

$$\sum_S \frac{|S|}{3^{|S|}} \hat{f}(S)^2 \leq \sum_i \mathrm{Inf}_i[f]^{3/2} \leq \sqrt{\max_i \mathrm{Inf}_i[f]}\,\mathrm{Inf}[f].$$

What can we say about the left-hand side?

$$\sum_S \frac{|S|}{3^{|S|}} \hat{f}(S)^2 \geq \sum_{|S|\geq 1} 3^{-|S|} \hat{f}(S)^2 = \mathbb{V}[f]\,\mathbb{E}[3^{-|\mathcal{S}|}],$$

where $\mathcal{S}$ is the spectral sample restricted to non-zero subsets (that is, $\Pr[\mathcal{S} = S] = \hat{f}(S)^2/\mathbb{V}[f]$ for $S \neq \emptyset$). Since 3^{-x} is convex, $\mathbb{E}[3^{-|\mathcal{S}|}] \geq 3^{-\mathbb{E}[|\mathcal{S}|]}$. What is $\mathbb{E}[|\mathcal{S}|]$? It is

$$\mathbb{E}[|\mathcal{S}|] = \frac{1}{\mathbb{V}[f]} \sum_{S\neq\emptyset} \hat{f}(S)^2 |S| = \frac{\mathrm{Inf}[f]}{\mathbb{V}[f]}.$$

Therefore

$$\frac{\mathbb{V}[f]}{\mathrm{Inf}[f]} 3^{-\mathrm{Inf}[f]/\mathbb{V}[f]} \leq \sqrt{\max_i \mathrm{Inf}_i[f]}.$$

Let $R = \mathrm{Inf}[f]/\mathbb{V}[f]$, and note that the left-hand side is a decreasing function of R. Fix a threshold α. If $R \geq \alpha$ then $\max_i \mathrm{Inf}_i[f] \geq \mathbb{V}[f](\alpha/n)$, and otherwise $\max_i \mathrm{Inf}_i[f] \geq (1/R3^R)^2 \geq 1/\alpha^2 9^\alpha$. Choose α so that both bounds are identical: $\mathbb{V}[f]/n = 1/\alpha^3 9^\alpha$; roughly speaking, $\alpha = \Theta(\log(n/\mathbb{V}[f]))$. Thus

$$\max_i \mathrm{Inf}_i[f] = \Omega\left(\frac{\mathbb{V}[f]}{n} \log \frac{n}{\mathbb{V}[f]}\right).$$

Tight example Can this bound be achieved? Let's concentrate on the most interesting case, $\mathbb{E}[f] \approx 0$. In this case, the bound $\frac{\log n}{n}$ is achieved by the so-called *tribes function*. There are n/m tribes $x_1, \ldots, x_{n/m}$ of size m, and the function is given by

$$f = \bigvee_{i=1}^{n/m} \bigwedge_{j=1}^{m} x_{i,j}.$$

(For convenience, we think of f as a function on $\{0,1\}^n$.) The probability that a tribe is all 1 is 2^{-m}, and so $\Pr[f = 0] = 1 - (1-2^{-m})^{n/m}$. If we choose m so that $n/m \approx 2^m$ then we get that $\Pr[f = 0] \approx 1 - 1/e$ (we can tweak that to $\Pr[f = 0] \approx 1/2$). Solving for m, we find that $m \approx \log(n/\log n)$.

What are the influences of f? For a variable $x_{i,j}$ to be influential at a given point, the ith tribe must be the only tribe which is all 1. This happens with probability $2^{-m}(1-2^{-m})^{n/m-1} \approx 2^{-m} \approx \frac{\log n}{n}$.

Sharp thresholds Although we have proved KKL only for the probability measure $\mu_{1/2}$, the same proof works for any *fixed* p. In particular, if $p_L \leq p \leq p_H$ for some $0 < p_L < p_H < 1$, then at every p there exists an index i such that $\mathrm{Inf}_i^{(p)}[f] = \Omega(\mathbb{V}[f]\frac{\log n}{n})$. Now suppose that f is monotone and *transitive-symmetric*, that is, symmetric with respect to some transitive permutation group (for example, f could be a monotone graph property). Then all its influences are the same, and so $\mathrm{Inf}^{(p)}[f] = \Omega(\mathbb{V}^{(p)}[f] \log n)$. Recall now the Russo–Margulis lemma:

$$\frac{d}{dp} \mathop{\mathbb{E}}_{\mu_p}[f] = \frac{\mathrm{Inf}^{(p)}[f]}{p(1-p)}.$$

This means that for $p \in [p_L, p_H]$, the derivative of $F(p) = \mathbb{E}_{\mu_p}[f]$ is $\Omega(\mathbb{V}^{(p)}[f] \log n)$. Since f is Boolean, in fact $\mathbb{V}^{(p)}[f] = F(p)(1-F(p))$, and so

$$F'(p) = \Omega(F(p)(1-F(p)) \log n).$$

For all non-trivial f, we have $F(0) = 0$ and $F(1) = 1$. Define the *critical probability* of F by $F(p_c) = 1/2$. How fast does F approach 0 or 1 as we get away from the critical probability? Pretty fast! Indeed,

suppose that $F(p_c + C/\log n) = 1 - \delta$. Then for $p \in [p_c, p_c + C/\log n]$, we have $F'(p) = \Omega(\delta \log n)$, and so $1/2 - \delta = \Omega(\delta C)$, implying that $\delta = O(1/C)$. In other words,

$$F\left(p_c + \frac{C}{\log n}\right) \geq 1 - O\left(\frac{1}{C}\right), \quad F\left(p_c - \frac{C}{\log n}\right) \leq O\left(\frac{1}{C}\right).$$

Therefore F goes from 0 to 1 at an interval of width $1/\log n$, assuming p_c is bounded away from 0 and 1. We call this kind of behavior *sharp threshold*. More careful analysis shows that the width of the interval is proportional to $p_c \log(1/p_c)/\log n$ (assuming $p_c \leq 1/2$); the only change is that we take into account the asymptotics of ρ in the hypercontractive inequality. This result is due to Friedgut and Kalai [FK96].

As an example, consider again the tribes function, with parameters chosen so that $F(1/2) \approx 1/2$. Since the tribes function is monotone and transitive-symmetric, this result implies that $F(1/2 - C/\log n)$ is very close to 0, while $F(1/2 + C/\log n)$ is very close to 1.

Other sharp threshold theorems The main problem with this result is that in many situations, we are interested in functions for which the critical probability is subconstant. For example, the threshold for k-colorability is $\Theta(1/n)$. We cannot expect a similar result in general: for example, the critical probability for the OR function is $\Theta(1/n)$, but the function $F(p) = 1 - (1-p)^n$ doesn't exhibit a sharp threshold since for $p = \Theta(1/n)$ we have $F(p) \approx 1 - e^{-pn}$. How can we ensure that a function has a sharp threshold, then?

Further symmetries One option is to assume that the function has more symmetries. Bourgain and Kalai [BK97] proved a sharp threshold result of that sort. A special case of their result states that if f is affine-invariant ($f(x) = f(Ax+b)$ for all invertible A) then $\log n$ can be strengthened to $\log n \log\log n$. While this might seem as a small difference, it has been very important in a recent result about Reed–Muller codes, which shows that they achieve capacity on erasure channels (Kumar and Pfister [KP15]). What is the connection? In very general terms, there is a function (the EXIT function) which measures the probability that a certain bit is decoded correctly, as a function of the error probability. The sharp threshold phenomenon ensures that if the error probability is slightly lower than the threshold, then the probability that each bit is decoded correctly is very close to 1.

Friedgut's sharp threshold theorem Another option is to show that if a function does not exhibit a sharp threshold, then there must be an "explanation". The most celebrated theorem of this kind is due to Friedgut [Fri99], who showed that if a function corresponding to a graph property exhibits a coarse threshold then near the threshold it can be approximated by a narrow DNF; other results in that direction are due to Bourgain (in an appendix to Friedgut's paper) and Hatami [Hat12]. Friedgut's theorem has been very influential in random graph theory.

Elementary argument Every non-trivial monotone property P_n has a threshold function $\theta(n)$ in the following sense: if $p(n) = o(\theta(n))$ then $\mu_p(P_n) \to 0$, whereas if $p(n) = \omega(\theta(n))$ then $\mu_n(P_n) \to 1$. Indeed, one can choose $\theta(n)$ so that $\mu_{\theta(n)}(P_n) = 1/2$ (such a point exists since $\mu_p(P_n)$ is polynomial in p and so continuous).

Here is why it works. For every integer C, generate a vector $x \in \{0,1\}^n$ by taking $x_1, \ldots, x_C \sim \mu_{\theta(n)}$ and letting $x = \max(x_1, \ldots, x_C)$ (coordinate-wise). Note that $x \sim \mu_{q(n)}$ for $q(n) = 1 - (1-\theta(n))^C \leq C\theta(n)$. Since P_n is monotone, the probability that x satisfies P_n is at least the probability that one of $x_1, \ldots, x_C$ satisfies it, and so

$$\mu_{C\theta(n)}(P_n) \geq \mu_{q(n)}(P(n)) \geq 1 - 2^{-C}.$$

This implies that if $p(n) = \omega(\theta(n))$ then $\mu_{p(n)}(P_n) \to 1$. The other property is proved analogously.

5.3 Friedgut's junta theorem

Suppose that $f\colon \{-1,1\}^n \to \{-1,1\}$ depends on k variables. In particular, it has degree at most k, and so total influence at most k. Conversely, what can we say if f has influence at most k? Friedgut's junta theorem [Fri98] implies such an approximate converse, though with some loss of parameters: f must be

close to a Boolean K-junta, where K is *exponential* in k. The exponential loss is unavoidable: the tribes function depends on all inputs but has total influence $O(\log n)$.

The first step in proving Friedgut's junta theorem is to identify the coordinates in the junta. This is easy: we just take all the influential coordinates. For some parameter τ to be determined, we define

$$J = \{i \in [n] : \mathrm{Inf}_i[f] \geq \tau\}.$$

Since the total influence of f is k, we immediately see that $|J| \leq k/\tau$, so as long as τ depends only on k (but not on n), the size of the junta will be bounded. It is then natural to define the junta g by

$$g(x_J, x_{\bar{J}}) = \mathop{\mathbb{E}}_{x_{\bar{J}}}[f(x_J)] = \sum_{S \subseteq J} \hat{f}(S)\chi_S.$$

The function g is not quite Boolean, and we will fix this later. But first, we show that f is close to g. Parseval's identity shows that

$$\|f - g\|^2 = \sum_{S \not\subseteq J} \hat{f}(S)^2.$$

In order to bound this, notice that Markov's inequality immediately shows that

$$\sum_{|S| \geq k/\epsilon} \hat{f}(S)^2 = \Pr[|\mathcal{S}| \geq k/\epsilon] \leq \frac{\mathbb{E}[|\mathcal{S}|]}{k/\epsilon} = \epsilon.$$

Here $\mathcal{S}$ is the spectral sample. This shows that

$$\|f - g\|^2 \leq \epsilon + \sum_{\substack{S \not\subseteq J \\ |S| < k/\epsilon}} \hat{f}(S)^2.$$

How do we bound the right hand side? We need to somehow use the fact that coordinates not in J have small influence:

$$\sum_{\substack{S \not\subseteq J \\ |S| < k/\epsilon}} \hat{f}(S)^2 \leq \sum_{i \notin J} \sum_{\substack{S \ni i \\ |S| < k/\epsilon}} \hat{f}(S)^2.$$

Each summand on the right looks pretty similar to $\mathrm{Inf}_i[f]$, but we are only summing over small sets S. This suggests the following line of thought:

$$\mathrm{Inf}_i[f]^{3/2} \geq \sum_{S \ni i} \hat{f}(S)^2 3^{-|S|} \geq 3^{-k/\epsilon} \sum_{\substack{S \ni i \\ |S| < k/\epsilon}} \hat{f}(S)^2,$$

using an inequality we have seen before. Combining the estimates, we get

$$\|f - g\|^2 \leq \epsilon + 3^{k/\epsilon} \sum_{i \notin J} \mathrm{Inf}_i[f]^{3/2} \leq \epsilon + 3^{k/\epsilon} k\sqrt{\tau}.$$

Choosing τ small enough, say $10^{-k/\epsilon}/(k\epsilon)^2$, we deduce $\|f - g\|^2 = O(\epsilon)$, with $|J| \leq (k^3/\epsilon^2)10^{k/\epsilon}$.

It remains to sort out the fact that g is not Boolean. The idea is to consider $h = \operatorname{sgn} g$, which is also a J-junta. Consider any point x, and suppose that $f(x) = 1$. If $h(x) = 1$ then certainly $(f(x) - h(x))^2 \leq (f(x) - g(x))^2$. If $h(x) = -1$ then $(f(x) - h(x))^2 = 4$ while $g(x) \leq 0$ implies that $(f(x) - g(x))^2 \geq 1$. Therefore $(f(x) - h(x))^2 \leq 4(f(x) - g(x))^2$. We conclude that $\|f - h\|^2 = O(\|f - g\|^2) = O(\epsilon)$.

5.4 Kindler's proof of the Friedgut–Kalai–Naor theorem.

As a bonus, we give another proof of the Friedgut–Kalai–Naor theorem, taken from Guy Kindler's thesis [Kin02] (see also his paper with Muli Safra [KS04]). We prove the following strong version of the theorem. Suppose that $f(x_1, \ldots, x_n) = \sum_{i=1}^n a_i x_i$ satisfies $\mathbb{E}[(|f| - 1)^2] = \epsilon$. Then for some i, $\sum_{j \neq i} a_j^2 < (1 + o(1))\epsilon$ (here $o(1)$ is a function tending to 0 as $\epsilon \to 0$). Alternatively, $\|f - (a_0 + a_i x_i)\|^2 < (1 + o(1))\epsilon$.

In order to deduce the classical version of the Friedgut–Kalai–Naor theorem from this result, let F be a balanced Boolean function satisfying $\|F^{>1}\|^2 = \epsilon$, and define $f = F^{\leq 1}$. Kindler's version of the theorem shows that for some function g depending on at most one coordinate, $\|f - g\|^2 < (1 + o(1))\epsilon$. Since $\deg g \leq 1$, $\|F - g\|^2 = \|f - g\|^2 + \|f^{>1}\|^2 < (2 + o(1))\epsilon$. Let $G = \operatorname{sgn} g$, another function depending on at most one coordinate. As we have seen previously, $\|F - G\|^2 < (8 + o(1))\epsilon$, and so $\Pr[F \neq G] < (2 + o(1))\epsilon$. Using ranodmized rounding [O'D14, Exercise 3.34], this can be improved to $(1 + o(1))\epsilon/2$.

For the proof, we arrange the indices so that $|a_1| \geq |a_i|$ for all $2 \leq i \leq n$. Our goal is thus to show that $\sum_{i=2}^n a_i^2 < (1 + o(1))\epsilon$. We start with an easy observation: $|a_i| = O(\sqrt{\epsilon})$ for $i \geq 2$. Indeed, otherwise $|a_1|, |a_i| = \Omega(\sqrt{\epsilon})$, and an easy case analysis shows that with constant probability $||f| - 1| = \Omega(\sqrt{\epsilon})$ and so $\mathbb{E}[(|f| - 1)^2] = \Omega(\epsilon)$, which for a proper choice of hidden constants contradicts the assumption on f.

The idea now is to prove by (backward) induction on m that $\sum_{i=m}^n a_i^2 < (1 + o(1))\epsilon$. This clearly holds for $m = n + 1$. Now suppose it holds for $m > 2$. Since $a_m^2 = O(\epsilon)$ and by induction $\sum_{i=m+1}^n a_i^2 < (1 + o(1))\epsilon$, we conclude that $\sum_{i=m}^n a_i^2 = O(\epsilon)$. The next step is to get rid of the variables $x_1, \ldots, x_{m-1}$. Indeed, we can always find some setting of these variables for which the restricted function ϕ satisfies $\mathbb{E}[(|\phi| - 1)^2] \leq \epsilon$. The restricted function is of the form $\phi(x_m, \ldots, x_n) = C + \sum_{i=m}^n a_i x_i$. Furthermore, the triangle inequality shows that C must be close to ± 1:

$$||C| - 1| = \||C| - 1\| \leq \||\phi| - 1\| + \||\phi| - |C|\| \leq \sqrt{\epsilon} + \|\phi - C\| = O(\sqrt{\epsilon}).$$

In other words, $||C| - 1| = O(\sqrt{\epsilon})$. We assume for simplicity that C is positive, and so $C \geq 1 - O(\sqrt{\epsilon})$. What we want to say now is that since $\sum_{i=m}^n a_i^2$ is small, it is unlikely for ϕ to dip below zero, and so $\mathbb{E}[(\phi - 1)^2] \approx \mathbb{E}[(|\phi| - 1)^2] = \epsilon$. Indeed, an application of Chernoff's bound shows that

$$\Pr[\phi < -t] = \Pr\left[\sum_{i=m}^n a_i x_i < C + t\right] \leq \exp \frac{-(C + t)^2}{2 \sum_{i=m}^n a_i^2}.$$

We can assume that $\sum_{i=m}^n a_i^2 \geq \epsilon$ (otherwise we have already proved what we wanted). Therefore

$$\mathbb{E}[|\phi| 1_{\phi<0}] = \int_{t=0}^\infty \Pr[\phi < -t]\, dt \leq \int_{t=C}^\infty e^{-t^2/2\epsilon}\, dt.$$

The integral on the right equals

$$\sqrt{2\pi\epsilon} \Pr[N(0, 1) > C/\epsilon] \leq e^{-(C/\epsilon)^2/2} = o(\epsilon),$$

using $C \geq 1 - O(\sqrt{\epsilon})$. Notice now that

$$\mathbb{E}[(|\phi| - 1)^2] = \mathbb{E}[(\phi - 1)^2] + 2\,\mathbb{E}[\phi - |\phi|] = \mathbb{E}[(\phi - 1)^2] - 4\,\mathbb{E}[|\phi| 1_{\phi<0}].$$

Therefore $\mathbb{E}[(\phi - 1)^2] \leq (1 + o(1))\epsilon$. Finally,

$$\sum_{i=m}^n a_i^2 = \mathbb{V}[\phi] \leq \mathbb{E}[(\phi - 1)^2] = (1 + o(1))\epsilon.$$

This completes the proof.

6 Hardness of approximation of Vertex Cover (25 November 2015)

(Hardness of approximation: standard material. Vertex cover: [KR08])

6.1 Hardness of approximation

The classical theory of NP-completeness describes tasks which are (presumably) hard to accomplish efficiently. In many cases, the problems we are interested in are optimization problems, and while they may be hard to solve exactly, it might be that they are easy to approximate. For example, while vertex cover is hard to solve exactly, it is easy to come up with 2-approximation algorithms. Hardness of approximation attempts to find the *inapproximability threshold*, which is the number K such that for every $\epsilon > 0$ there is a $(K + \epsilon)$-approximation algorithm, but it is NP-hard to $(K - \epsilon)$-approximate the problem.

PCP Surprisingly, hardness of approximation is connected to the problem of *proof verification*. A proof verification procedure probes a proof at a small number of locations, and decides whether the proof is valid or not. Its decision must be correct with high probability. More concretely, one common definition is the classes $\mathsf{PCP}(r(n), q(n))$ of probablistically checkable proofs. A language L is in $\mathsf{PCP}(r(n), q(n))$ if there exists a randomized polytime algorithm V which accepts an input x of size n and a proof w, uses $r(n)$ random bits to decide (non-adaptively) on $q(n)$ locations, reads these locations from w and answers YES or NO, and satisfying the following two properties, for some arbitrary constant $\delta > 0$:

Completeness If $x \in L$ then there exists a proof w which *always* convinces V.

Soundness If $x \notin L$ then for all w, the probability that V is convinced is at most $1 - \delta$.

By trying all possible proofs (they can be assumed to be of size at most $q(n)2^{r(n)}$), a non-deterministic machine can decide each $L \in \mathsf{PCP}(r(n), q(n))$ in time $poly(n)q(n)2^{r(n)}$. In particular, $\mathsf{PCP}(O(\log n), O(1)) \subseteq \mathsf{NP}$. The PCP theorem [ALM+98, Din07] states a surprising converse: $\mathsf{NP} = \mathsf{PCP}(O(\log n), O(1))$. This means that every language L in NP has a proof system and an associated verifier which can make an educated guess on whether $x \in L$ by reading only a constant number of bits from the proof! A similar (but much easier) theorem [BFL90] states that $\mathsf{NEXP} = \mathsf{PCP}(poly(n), O(1)) = \mathsf{PCP}(poly(n), poly(n))$.

Håstad showed that the constant $O(1)$ can be chosen to be 3. We can represent the verifier as a set of polynomially many tests $\phi_1, \ldots, \phi_N$ (depending on the input x), each involving three bits of the witness:

$$\phi_1(w_{i_1}, w_{j_1}, w_{k_1}), \ldots, \phi_N(w_{i_N}, w_{j_N}, w_{k,N}).$$

If $x \in L$ then some truth assignment satisfies all these constraints, otherwise any truth assignment satisfies at most a $1 - \delta$ fraction of the constraints. Each constraint ϕ_t is equivalent to a conjunction of at most 8 clauses, and by taking their conjuction, we obtain an instance of MAX-3SAT. If $x \in L$ then this instance is satisfiable, otherwise every assignment satisfies at most a $1 - \delta/8$-fraction of the constraints. More explicitly, what this means is that for every $L \in \mathsf{NP}$ there is a polytime reduction f (which is polytime since V is) such that $f(x)$ is an instance of MAX-3SAT, and

Completeness If $x \in L$ then $f(x)$ is satisfiable.

Soundness If $x \notin L$ then at most a $1 - \delta/8$-fraction of the clauses of $f(x)$ can be satisfied.

This implies that we cannot approximate MAX-3SAT better than $1-\delta/8$ (unless P=NP), since otherwise we could use such an approximation algorithm to solve every problem in NP!

Using analysis of Boolean functions, Håstad [Hås01] improved this and showed a threshold of 7/8 for approximating MAX-3SAT. This approximation ratio is achieved by the random assignment algorithm (which can be derandomized using the method of conditional expectations), so in this case, the trivial algorithm cannot be improved upon!

Two-prover games Another way to look at this result, which is useful for constructing reductions, is via two-player games. Consider the verifier that we mentioned earlier. We can assume without loss of generality that each test is actually of the form $\phi(x, y, z) = x \vee y \vee z$, where x, y, z are *literals* rather than *variables*. Earlier we discussed a one-player game, in which the prover chooses an assignment for all variables, the verifier chooses a constraint ϕ, queries the values of the associated variables, and checks whether the constraint is satisfied.

An alternative formulation has two different provers. One prover keeps an assignment for all variables. The other prover keeps, for each clause ϕ, a satisfying assignment. The verifier chooses a clause ϕ and a variable $v \in \phi$ at random, queries the first prover for v, the second prover for a satisfying assignment of ϕ, and checks that they both match on v. If this is a YES instance, then the provers can decide on a satisfying assignment, and then the verifier will always be convinced. Otherwise, the verifier can be convinced with probability at most $1 - \delta/3$.

Another way of presenting this formulation is as a LABEL COVER problem. We construct a bipartite graph in which one bipartition corresponds to vertices, and the other to clauses. Whenever a variable participates in a clause we draw an edge, and annotate it with the function mapping the clause label to the variable label. We call these edges *constraints*, and a labeling of the vertices satisfies a constraint if the clause label gets mapped to the variable label. If this is a YES instance, there is a labeling of the vertices which satisfies all constraints. Otherwise, every labeling satisfies at most $1 - \delta/3$ of the constraints.

Parallel repetition In many situations we would like to boost the soundness from $1 - \delta/3$ to an arbitrarily small probability. A natural way to do this is to start with an instance of LABEL COVER and compose it with itself. Thinking of this as a two-prover game, we now choose m clauses $\phi_1, \ldots, \phi_m$, identify one variable v_i in each, and then ask the two provers for the truth assignments of all marked variables and all clauses, respectively. The verifier then accepts only if the assignments to all variables match. Now it would seem that if the original game had success probability at most α, then the new game has success probability at most α^m, but this is surprisingly not true! Nevertheless, Ran Raz [Raz98] showed that as $m \to \infty$, the success probability does tend to 0, exponentially fast. This is known as the *parallel repetition theorem*.

Unique label cover Many hardness of approximation results can be proved using LABEL COVER, but at some point research got stuck. Khot [Kho02] suggested assuming as a hypothesis that LABEL COVER is still NP-hard even if the constraints are *bijections*. One must be careful here: if the constraints are bijections and the instance is fully satisfiable, then it is easy to verify it by guessing one label per connected component. We therefore have to forego perfect completness, and in the YES case only ask that at least a $1 - \epsilon$ fraction of the constraints can be satisfied.

There is also a dependency on the *alphabet size*, which is the size of the set of labels. For classical LABEL COVER, parallel repetition magnified the soundness at the expense of expanding the alphabet. This led Khot to the following version of his conjecture. For each $\epsilon > 0$ there exists an alphabet size for which it is NP-hard to distinguish whether a UNIQUE LABEL COVER instance is at least $(1 - \epsilon)$-satisfiable or at most ϵ-satisfiable. This is known as the *Unique Games Conjecture*.

Arora, Barak and Steurer constructed a sub-exponential time algorithm for solving this problem, and this has led to some speculation whether the UGC is actually true or not. The jury is still out. While most results concerning UGC are conditional on the conjecture, it has also yielded some unconditional results such as the celebrated Khot–Vishnoi [KV15] refutation of the Goemans–Linial conjecture.

Sparse set expansion A related problem is *Small Set Expansion*. For a d-regular graph and a subset of vertices, we define its *expansion* by $\phi(S) = |E(S, \overline{S})|/d|S|$, a quantity which is always in $[0, 1]$. Small Set Expansion is the following problem. Given a regular graph $G = (V, E)$, distinguish the following two cases:

YES case There is a set S of size $\gamma|V|$ such that $\phi(S) \leq \epsilon$.

NO case All sets S of size $\gamma|V|$ satisfy $\phi(S) \geq 1 - \epsilon$.

Raghavendra and Steurer [RS10] put forward the conjecture that for all $\epsilon > 0$ there exists $\gamma > 0$ such that Small Set Expansion with these parameters is NP-hard. They showed that this conjecture implies the unique games conjecture. The reverse implications is not known.

6.2 Vertex cover

Given a graph $G = (V, E)$, a subset S of the vertices is a *vertex cover* if every edge intersects S. VERTEX COVER is the problem of finding a vertex cover of minimal size. The decision version is, given a graph G and an integer k, to decide whether G has a vertex cover of size at most k. This formulation shows that VERTEX COVER is in NP, and in fact it is NP-complete.

VERTEX COVER is related to another well-known problem, INDEPENDENT SET: a set S is a vertex cover iff its complement is an independent set. In terms of approximability, however, the problems are quite different: VERTEX COVER can be 2-approximated, while it is NP-hard to approximate INDEPENDENT SET within $n^{1-\epsilon}$ for any $\epsilon > 0$!

It is also natural to consider the weighted version of VERTEX COVER, in which vertices have non-negative weights, and the goal is to choose a vertex cover of minimal weight. By duplicating vertices, we can reduce the weighted version to the original version with an arbitrarily small loss in accuracy.

There is a simple 2-approximation algorithm for VERTEX COVER, which goes as follows. Find a maximal matching M in the graph (a *maximal* matching is one to which no edge can be added; even though a *maximum* matching can also be found efficiently, a maximal matching suffices here). Every vertex cover of G must contain at least one vertex out of each edge in M, and so has size at least $|M|$. Conversely, since M is maximal, every edge touches M, and so the $2|M|$ vertices in M are a vertex cover.

Another simple 2-approximation algorithm is based on linear programming. For each vertex i we have a variable $x_i \in [0, 1]$, and for each edge (i, j) we add the constraint $x_i + x_j \geq 1$. The objective function we want to minimize is $\sum_i x_i$. Clearly, any vertex cover is a solution to this linear program. Conversely, given a solution $(x_1, \ldots, x_n)$, let $S = \{i : x_i \geq 1/2\}$. For each edge (i, j), it cannot be that $i, j \notin S$ since then $x_i + x_j < 1$. Hence S is a vertex cover. Also, $|S| \leq 2 \sum_i x_i$, and so $|S|$ is at most twice as large as the optimal vertex cover.

A third simple 2-approximation algorithm, due to Bar-Yehuda and Even, goes as follows. Go over all edges, and for each edge (x, y), reduce both weights by $\min(w(x), w(y))$. After going over all edges, pick the vertices whose weight is zero. This is clearly a vertex cover, since the operation of reducing weights always reduces at least one of the weights to zero. Consider now each step of the algorithm. Reducing the weights by $\omega = \omega(x, y) = \min(w(x), w(y))$ reduces the weight of the optimal solution by at least ω, and the weight of the solution chosen by the algorithm by at most 2ω. The weight of the final solution is zero in terms of the new weights, and at most $2 \sum \omega(x, y)$ in terms of the original weights, whereas the weight of the optimal solution is at least $\sum \omega(x, y)$, showing that this is a 2-approximation.

A fourth simple 2-approximation algorithm, due to Savage, proceeds as follows. Run DFS from one of the nodes, and take all non-leaf vertices. This is a vertex cover since all edges from leaves of the DFS tree point to non-leaves. To see that this is a 2-approximation, it is enough to show that if a tree has t non-leaves then it admits a $t/2$-matching. We prove this by induction on the size of the tree. The claim is trivial for a single vertex. Given a tree T in which the root's children are $T_1, \ldots, T_n$ and T_1's children are $S_1, \ldots, S_m$, we construct a matching by taking the edge (T, T_1) and inductively constructed matchings of $S_1, \ldots, S_m, T_2, \ldots, T_n$. In total, this gives a matching having at least $1 + \sum_i t(S_i)/2 + \sum_{j>1} t(T_j)/2 = t(T)/2$ edges.

The best hardness of approximation result based on NP-hardness is due to Dinur and Safra [DS05], who showed that it is NP-hard to approximate VERTEX COVER better than $10\sqrt{5} - 21 \approx 1.36$. Assuming the unique games conjecture, one can prove more: Khot and Regev [KR08] showed that it is UGC-hard to approximate VERTEX COVER better than 2. So the trivial algorithms are optimal!

6.3 The reduction

In order to show UGC-hardness, we use the following version of unique label cover. For every $\delta > 0$ and t there exists an alphabet R such that the following problem is UGC-hard. Given a multigraph $G = (X, E)$ along with a constraint $\psi_e \in S_R$ (the symmetric group on R) for every edge $e \in E$, distinguish the following two options:

YES case Some labeling satisfies all constraints for a $(1-\delta)$-fraction of the vertices.

NO case No t-labeling satisfies all constraints for any δ-fraction of the vertices.

(In both cases we are only interested in constraints involving two satisfied vertices.)

Here a t-labeling gives t possible labels for each vertex, and a constraint $\psi_{(x,y)}$ is satisfied by a labeling L if $\psi_{(x,y)}(a) = b$ for some $a \in L(x)$ and $b \in L(y)$.

For any $\epsilon > 0$, we will show that it is UGC-hard to distinguish the case in which there is an independent set of weight $1/2 - \epsilon_0$ from the case in which there is no independent set of weight ϵ_1, where $\epsilon_0, \epsilon_1 \to 0$. In terms of vertex covers, in the first case there is a vertex cover of weight $1/2 + \epsilon_0$ and in the second every vertex cover has weight at least $1 - \epsilon_1$, leading to an inapproximability threshold of 2.

Let $p = 1/2 - \epsilon$. We choose the parameters δ and t to ensure that $\epsilon_0, \epsilon_1 \to 0$. Given an instance $G = (R, X, \Psi)$ of unique label cover, we construct a weighted graph $G = (V, E, w)$ as follows:

Vertices For each $x \in X$ and $F \subseteq R$ there is a vertex (x, F) of weight $\mu_p(F)/|X|$; note that $w(V) = 1$.

Edges For every constraint $\psi_{xy} \in \Psi$ we add an edge $(x, F), (y, G)$ whenever no $a \in F, b \in G$ satisfies $\psi_{xy}(a) = b$.

To understand this construction better, let us see what happens in the YES case. Suppose L is a labeling that satisfies all constraints in a set X_0 of size $(1-\delta)|X|$. Define

$$I = \{(x, F) : x \in X_0, L(x) \in F\}.$$

Clearly $w(I) = 1/2 - \delta/2$. Why is I an independent set? Consider any pair of vertices $(x, F), (y, G) \in I$. Since $L(x) \in F$ and $L(y) \in G$ satisfy $\psi_{xy}(L(x)) = L(y)$, we see that $(x, F), (y, G)$ cannot be an edge.

The independent set I encodes the labeling L using what is known as the (biased) *long code*. In the set I, the assignment $L(x) = a$ is encoded as a dictatorship: the function $F \mapsto (x, F) \in I$ depends only on the coordinate $F(a)$. The difficult part of the proof shows how to decode a meaningful assignment from *any* independent set I which is not too small.

6.4 Soundness

Let I be an independent set of weight γ, where γ is some constant which will only depend on ϵ. For $x \in X$, let $I_x = \{F : (x, F) \in I\}$, so that $\mathbb{E}_x[w(I_x)] = \gamma$. This shows that a $\gamma/2$-fraction of xs satisfies $w(I_x) \geq \gamma/2$, since otherwise

$$\mathop{\mathbb{E}}_x[w(I_x)] < (\gamma/2) \cdot 1 + (1 - \gamma/2) \cdot (\gamma/2) < \gamma.$$

Let $X_0 = \{x : w(I_x) \geq \gamma/2\}$, so that $|X_0| \geq (\gamma/2)|X|$. We will show how to associate with each $x \in X_0$ a set of vertices $L(x)$ of constant size (depending only on ϵ; this will be our t) such that L satisfies all constraints in X_0. Before doing this, we mention that we can assume that I_x is monotone: indeed, if we replace each I_x with its upset, then it remains an independent set, while its weight can only increase. To see why, note that I is independent if for every constraint ψ_{xy} and every $F \in I_x, G \in I_y$ there are $a \in F, b \in G$ such that $\psi_{xy}(a) = b$.

We define $L(x)$ in two stages. First, recalling that $p = 1/2 - \epsilon$, note that $\mu_{1/2-\epsilon}(I_x) \geq 0$ while $\mu_{1/2-\epsilon/2}(I_x) \leq 1$. Hence there must exist $p_x \in [p, 1/2 - \epsilon/2]$ at which the derivative of $\mu_{p_x}(I_x)$ is at most $2/\epsilon$. According to the Russo–Margulis lemma, this means that $\mathrm{Inf}^{(p_x)}[I_x] \leq 2/\epsilon$, and so Friedgut's theorem shows that I_x is η-close to an $O_\eta(1)$-junta (we will choose η later on). Note that the size of the junta depends on p_x, but we can bound it in terms of p alone. We define $C(x)$ to consist of this junta.

If I_x actually depended only on the variables in $C(x)$, then we could take $L(x) = C(x)$. Why is that? Consider any constraint ψ_{xy} such that $x, y \in X_0$. Since I is an independent set, any $F \in I_x$ and $G \in I_y$ satisfy $\psi_{xy}(a) = b$ for some $a \in F, b \in G$. In particular, we could take $F = C(x) \in I_x$ and $G = C(y) \in I_y$ (here we are using monotonicity), and the proof will be complete. Unfortunately, this argument doesn't quite work since we only know that I_x is *close* to a $C(x)$-junta. This necessitates including in $L(x)$ additionally all variables with influence at least η' (we will choose η' later on) with respect to μ_{p_x}:

$$L(x) = C(x) \cup \{i : \mathrm{Inf}_i^{(p_x)}[I_x] \geq \eta'\}.$$

Since the total influence of I_x at μ_{p_x} is at most $2/\epsilon$, the size of $L(x)$ is bounded as a function of ϵ, η, η' alone.

In the remainder, we show that for an appropriate choice of η, η', the resulting multi-assignment $L(x)$ satisfies all constraints involving X_0. In order to do that, we will consider some specific constraint ψ_{xy}, which for simplicity we assume is the identity permutation. Thus we have to show that $L(x)$ intersects $L(y)$.

Suppose that $L(x)$ and $L(y)$ are disjoint. Our plan is to come up with sets $A \in I_x, B \in I_y$ which are disjoint, which contradicts the fact that I is an independent set, since there is an edge connecting (x, A) and (y, B). How are we going to accomplish that? Our plan is to find $A_0 \subseteq C(x)$ and $B_0 \subseteq C(y)$ such that I_x contains most of $A_0 \times 2^{\overline{C(x) \cup C(y)}}$ and I_y contains most of $B_0 \times 2^{\overline{C(x) \cup C(y)}}$. Quantitatively, we will show that

$$\mu_{p_x}(\{F \subseteq \overline{C(x) \cup C(y)} : A_0 \cup F \in I_x\}) \geq 1 - O(\eta/\gamma),$$
$$\mu_{p_y}(\{F \subseteq \overline{C(x) \cup C(y)} : B_0 \cup F \in I_y\}) \geq 1 - O(\eta/\gamma).$$

Since $p_x, p_y \leq 1/2 - \epsilon/2 =: q$ and I_x, I_y are monotone, the same estimates hold also for μ_q. If I_x, I_y contained *all* of these fibers, then $A_0 \in I_x$ and $B_0 \in I_y$ would be disjoint. Here, by taking η/γ small enough, we can force

$$\mu_q(\{F \subseteq \overline{C(x) \cup C(y)} : A_0 \cup F \in I_x, B_0 \cup F \in I_y\}) > q,$$

and since $q < 1/2$, the Erdős–Ko–Rado theorem implies that $\{F \subseteq \overline{C(x) \cup C(y)} : A_0 \cup F \in I_x, B_0 \cup F \in I_y\}$ is too big to be intersecting. This family therefore contains two disjoint sets A_1, B_1, and then $A_0 \cup A_1 \in I_x$ and $B_0 \cup B_1 \in I_y$ are disjoint, completing the proof[1].

It remains to show that for some $A_0 \subseteq C(x)$,

$$\mu_{p_x}(\{F \subseteq \overline{C(x) \cup C(y)} : A_0 \cup F \in I_x\}) \geq 1 - O(\eta/\gamma).$$

(The same statement for y follows in the same way.) Here we will use the fact that all variables in $C(y)$ have low influence on I_x (since $C(y)$ is disjoint from $L(x)$ by assumption).

Suppose first that I_x were completely independent of $C(y)$. Recall that I_x is η-close to a junta J_x depending only on the coordinates $C(x)$. Assuming $\eta \leq \gamma/4$, since I_x is monotone we have $\mu_{p_x}(I_x) \geq \mu_p(I_x) \geq \gamma/2$, and so $\mu_{p_x}(J_x) \geq \gamma/2 - \eta \geq \gamma/4$. For every $A \subseteq C(x)$, define

$$m(A) = \mu_{p_x}^{[\overline{C(x) \cup C(y)}]}(\{F \subseteq \overline{C(x) \cup C(y)} : F \cup A \in I_x\}).$$

(The square brackets signify that the measure is taken with respect to the given universe.) We know that

$$\sum_{A \in J_x \cap C(x)} \mu_{p_x}^{[\overline{C(x) \cup C(y)}]}(A)(1 - m(A)) = \mu_{p_x}(J_x \setminus I_x) \leq \eta$$

while

$$\sum_{A \in J_x \cap C(x)} \mu_{p_x}^{[\overline{C(x) \cup C(y)}]}(A) = \mu_{p_x}(J_x) \geq \gamma/4.$$

This implies that some $A_0 \in J_x \cap C(x)$ must satisfy $1 - m(A_0) \leq \eta/(\gamma/4)$, or $m(A_0) \geq 1 - 4\eta/\gamma$.

While in general I_x does depend on $C(y)$, the dependence is very slight, since all variables in $C(y)$ have low influence. It is natural to define another set K_x which doesn't depend on $C(y)$ by:

$$K_x = \{Z \subseteq \overline{C(y)} : Z \times 2^{C(y)} \subseteq I_x\} \times 2^{C(y)}.$$

In words, for every $Z \subseteq \overline{C(y)}$, we look at the fiber $Z \times 2^{C(y)}$. If I_x contains *all* of this fiber, then we include it in K_x. Otherwise we don't. This makes it clear that $K_x \subseteq I_x$. On the other hand, we can show that I_x is not much larger than K_x.

[1] This argument looks a bit odd, and indeed the "correct" argument uses the cross-intersecting Erdős–Ko–Rado theorem: the two sets $\mathcal{F}_A = \{F \subseteq \overline{C(x) \cup C(y)} : A_0 \cup F \in I_x\}$ and $\mathcal{F}_B = \{F \subseteq \overline{C(x) \cup C(y)} : B_0 \cup F \in I_y\}$ satisfy $\sqrt{\mu_q(\mathcal{F}_A)\mu_q(\mathcal{F}_B)} > q$, and so they cannot be *cross-intersecting*, that is, there must be a set in $\mathcal{F}_A$ which is disjoint from some set in $\mathcal{F}_B$.

Indeed, suppose that $Z \times 2^{C(y)}$ is some fiber which intersects I_x but is absent from K_x. Since I_x is monotone, there must be some maximal set $W \subseteq C(y)$ such that $Z \cup W \notin I_x$ but $Z \cup W \cup \{i\} \in I_x$ for all $i \in C(y) \setminus W$. We assign each such fiber to some $i \in C(y)$, and call it an i-fiber. Each i-fiber $Z \times 2^{C(y)}$ contributes at most $\mu_{p_x}^{[\overline{C(y)}]}(Z)$ to $\mu_{p_x}(I_x \setminus K_x)$. It also contributes at least $\mu_{p_x}^{[\overline{C(y)}]}(Z)\mu_{p_x}^{[C(y)]}(W)$ to $[p_x(1-p_x)]^{-1}\operatorname{Inf}_i[I_x]$. Since $\mu_{p_x}^{[C(y)]}(W) \geq p_x^{|C(y)|} \geq p_x^{2/\epsilon}$, we see that the contribution of all i-fibers to $\mu_{p_x}(I_x \setminus K_x)$ is at most

$$\frac{\operatorname{Inf}_i[I_x]}{p_x(1-p_x)p_x^{2/\epsilon}} \leq \frac{\eta'}{p_x(1-p_x)p_x^{2/\epsilon}}.$$

The size of $C(y)$ itself is also bounded (in terms of η alone), and so by choosing η' small enough, we can ensure that

$$\mu_{p_x}(I_x \setminus K_x) \leq \eta.$$

Now we are in good shape. The triangle inequality shows that $\mu_{p_x}(J_x \setminus K_x) \leq 2\eta$, and so there exists some $A_0 \in J_x \cap C(x)$ satisfying $m_K(A_0) \geq 1 - 8\eta/\gamma$, where m_K is the same as m but with K_x instead of I_x. Since $K_x \subseteq I_x$, this implies that $m(A_0) \geq 1 - 8\eta/\gamma$, completing the proof.

6.5 Review of the whole argument

Let us recap the entire argument. Given an instance (R, X, E, Ψ) of unique label cover, we construct a graph whose vertex set is $X \times 2^R$, and there is an edge connecting $(x, A), (y, B)$ if $(x, y) \in E$ and $\psi_{(x,y)}(a) \neq b$ for all $a \in A, b \in B$. The weight of a vertex (x, A) is $\mu_p(A)/|X|$, where $p = 1/2 - \epsilon$. An independent set in this graph is a set $I = \bigcup_{x \in X} \{x\} \times I_x$ such that for all edges $(x, y) \in E$ and all $A \in I_x, B \in I_y$ there exist $a \in A, b \in B$ such that $\psi_{(x,y)}(a) = b$.

If the instance is a YES instance, with the assignment L satisfying all constraints for a $(1 - \delta)$-fraction of the vertices, then we can take $I_x = \{A \subseteq 2^R : L(x) \in A\}$ for all satisfied vertices x. This is an independent set since for all edges (x, y) such that both x, y are satisfied and all $A \in I_x, B \in I_y$ we have $L(x) \in A, L(y) \in B$ and $\psi_{(x,y)}(a) = b$. This independent set has measure $(1 - \delta)p \approx 1/2$.

Now suppose that there is an independent set of measure γ. Our goal is to construct a t-labeling satisfying a δ-fraction of the vertices, for some constants t, δ of our choice, showing that the instance cannot be a NO instance. This will finish the proof since in the YES case we get a vertex cover of measure roughly $1/2$, whereas in the NO case we get that the minimum vertex cover has measure at least $1 - \gamma \approx 1$.

Good vertices A calculation shows that at least a $\gamma/2$ fraction of the vertices satisfy $\mu_p(I_x) \geq \gamma/2$. We focus only on these vertices. For simplicity, we assume that all constraints linking them are identity constraints. We are thus given that for any two good vertices x, y, the families I_x, I_y are cross-intersecting. We can also assume that I_x, I_y are monotone.

Applying the junta theorem The Russo–Margulis lemma shows that for every good x there exists $p_x \in [p, q]$ (where $q = 1/2 - \epsilon/2$) such that the total influence of I_x at p_x is at most $2/\epsilon$. Friedgut's theorem then shows that with respect to μ_{p_x}, the family I_x is η-close to a junta J_x depending on the set constant-size set $C(x)$. We will choose η later.

Decoding If $\eta = 0$ then we could use the labeling $L(x) = C(x)$ since for any good x, y, $C(x) \in J_x, C(y) \in J_y$ and $I_x = J_x, I_y = J_y$ are cross-intersecting. Since $\eta > 0$, we need a more delicate argument. We let $L(x) = C(x) \cup \{i : \operatorname{Inf}_i^{(p_x)}[I_x] \geq \eta'\}$ for some η'. We want to show that $L(x), L(y)$ intersect, so we will assume that they are disjoint and will reach a contradiction by showing that I_x, I_y are not cross-intersecting.

We will show that for some assignments $A_0 \subseteq C(x)$ and $B_0 \subseteq C(y)$,

$$\mu_{p_x}(I_x|_{(C(x),C(y))=(A_0,\emptyset)}) \geq 1 - O(\gamma/\eta), \quad \mu_{p_y}(I_y|_{(C(x),C(y))=(\emptyset,B_0)}) \geq 1 - O(\gamma/\eta).$$

The same is true for the μ_q measures since I_x, I_y are monotone. The cross-intersecting Erdős–Ko–Rado theorem shows that the restrictions cannot be cross-intersecting if $1 - O(\gamma/\eta) > q$ (which we can arrange), and it follows that I_x, I_y are not cross-intersecting.

Proving the technical lemma Since we assumed $L(x), L(y)$ are disjoint, in particular all variables in $C(y)$ have low influence in F_x. Suppose that I_x were actually independent of $C(y)$. In that case, a simple averaging argument proves the existence of such a set A_0: on the one hand the measure of J_x is at least $\gamma/2 - \eta \geq \gamma/4$ (say), and on the other hand the measure of $J_x \setminus I_x$ is at most η.

To handle the general case, define a family $K_x \subseteq I_x$ which is independent of $C(y)$ in the most economical way: $K_x = (I_x \cap \overline{C(y)}) \times C(y)$. The fact that the influences of variables in $C(y)$ on I_x are small allows us to bound the measure of $I_x \setminus K_x$, say by η, and then the measure of $J_x \setminus K_x$ is at most 2η, and we can run the same argument as before.

7 Gaussian space (2 December 2015)

([O'D14, Sections 11.1–11.3], [DMN13])

Suppose we want to understand a binomial random variable $\mathrm{Bin}(n,p)$ for constant p and large n. Many aspects of this random variable are captured by the central limit theorem, which states that its distribution approaches that of a normal random variable $N(np, np(1-p))$. The invariance principle states something similar for much more general situations. This prompts us to study *Gaussian space*, which is $\mathbb{R}^n$ subject to the standard Gaussian measure.

7.1 Basic definitions

The n-dimensional Gaussian space is the measure space consisting of $\mathbb{R}^n$ and the product Gaussian measure $N(0,1)^{\otimes n}$. We say that a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is in L^p if $\mathbb{E}[|f|^p] < \infty$. We will mostly be interested in functions which are L^2.

Every function on the Boolean cube has a unique expansion as a multilinear polynomial. This is clearly not the case in Gaussian space: the function x_1^2, for example, has no such expansion. We need to expand the Fourier basis to a more expressive basis known as the *Hermite basis*. Consider first the case $n = 1$. The Fourier basis in this case consists of the two functions $1, x$. The most important property of the Fourier basis is that it is orthonormal. What quadratic polynomial can we add to this list so that it continues to be orthonormal? If we take $ax^2 + bx + c$ then we have

$$\langle ax^2 + bx + c, 1\rangle = \mathbb{E}[ax^2 + bx + c] = a + c,$$

and so $a = -c$. Similarly,

$$\langle ax^2 + bx + c, x\rangle = \mathbb{E}[ax^3 + bx^2 + cx] = b,$$

and so $b = 0$. We conclude that the polynomial must be a multiple of $x^2 - 1$. The norm of $x^2 - 1$ is

$$\|x^2 - 1\|^2 = \mathbb{E}[x^4 - 2x^2 + 1] = 3 - 2 + 1 = 2,$$

and so $(x^2 - 1)/\sqrt{2}$ is the polynomial we are after. We can continue this way and construct an infinite orthonormal sequence $h_i(x)$ of univariate polynomials, where $\deg h_i = i$. They are given explicitly by the formula

$$h_i(x) = \frac{(-1)^i}{\sqrt{i!}e^{-x^2/2}} \frac{d^i}{dx^i} e^{-x^2/2}.$$

It turns out that the Hermite polynomials form a *complete* basis for L^2, that is, every function in $L^2(\mathbb{R})$ can be expanded as an infinite linear combination of Hermite polynomials. Indeed, for any function f, let $p(f)$ be its projection to the span of the Hermite polynomials. Since the Hermite polynomials are orthogonal, $\phi = f - p(f)$ is orthogonal to all Hermite polynomials, and so to all polynomials. In particular, $\langle \phi, e^{-itx}\rangle = 0$. This expression is just the Fourier transform of $\phi(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Since the Fourier transform is invertible, $\phi = 0$ and so $f = p(f)$.

How did we get this formula, and how do we know that these polynomials are orthogonal? The starting point is the following Taylor expansion:

$$\begin{aligned}
e^{tx-t^2/2} &= e^{x^2/2-(t-x)^2/2} \\
&= e^{x^2/2} \sum_{i=0}^{\infty} \frac{d^i}{dt^i} e^{-(t-x)^2/2}\Big|_{t=0} \frac{t^i}{i!} \\
&\overset{t=x-s}{=} e^{x^2/2} \sum_{i=0}^{\infty} (-1)^i \frac{d^i}{ds^i} e^{-s^2/2}\Big|_{s=x} \frac{t^i}{i!} \\
&= \sum_{i=0}^{\infty} h_i(x) \frac{t^i}{\sqrt{i!}}.
\end{aligned}$$

Multiplying two copies and taking expectations over $x \sim N(0,1)$, we obtain

$$\mathbb{E}[e^{tx-t^2/2} e^{sx-s^2/2}] = \sum_{i,j=0}^{\infty} \mathbb{E}[h_i(x)h_j(x)] \frac{t^i s^j}{\sqrt{i!j!}}.$$

On the other hand, a classical calculation shows that

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

Therefore

$$\mathbb{E}[e^{tx-t^2/2} e^{sx-s^2/2}] = e^{(t+s)^2/2 - t^2/2 - s^2/2} = e^{ts} = \sum_{i=0}^{\infty} \frac{t^i s^i}{i!}.$$

Comparing coefficients, we see that the Hermite polynomials indeed form an orthonormal basis. (The signs are there only to keep the leading coefficient positive.)

What about n-dimensional Gaussian space? To obtain a complete basis, all we need to do is to take the nth tensor power of the Hermite basis for $\mathbb{R}$. Each basis function is of the form $h_{i_1}(x_1) \cdots h_{i_n}(x_n)$, and this collection forms a complete basis for $L^2(\mathbb{R}^n)$. It is natural to grade this basis by putting $h_{i_1}(x_1) \cdots h_{i_n}(x_n)$ at level $i_1 + \cdots + i_n$. Every function $f \in L^2$ can then be decomposed into its homogeneous components: $f = \sum_{i \geq 0} f^{=i}$. This decomposition behaves nicely with respect to the Gaussian noise operator, which we turn to next.

7.2 Noise operator

Analogous to the noise operator on the cube, we have the Ornstein–Uhlenbeck noise operator U_ρ in Gaussian space. There are several equivalent ways of defining it. One way is:

$$U_\rho f(x) = \mathop{\mathbb{E}}_{y \sim N(0,1)}[f(\rho x + \sqrt{1-\rho^2} y)].$$

The strange coefficient ensures that if $x \sim N(0,1)$ that so is the input to f, and this shows that

$$\langle f, U_\rho g \rangle = \mathop{\mathbb{E}}_{(x,y) \sim N(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right))}[f(x) g(y)].$$

It can also be defined via Brownian motion.

What is the effect of the noise operator on the Hermite basis? Consider first the case $n = 1$. The following calculations are all easy:

$$U_\rho h_0(x) = \mathop{\mathbb{E}}_y[1] = 1 = h_0(x),$$

$$U_\rho h_1(x) = \mathop{\mathbb{E}}_y[\rho x + \sqrt{1-\rho^2} y] = \rho x = \rho h_1(x),$$

$$U_\rho h_2(x) = \frac{1}{\sqrt{2}} \mathop{\mathbb{E}}_y[(\rho x + \sqrt{1-\rho^2} y)^2 - 1] = \frac{1}{\sqrt{2}}(\rho^2 x^2 + 1 - \rho^2 - 1) = \frac{\rho^2(x^2-1)}{\sqrt{2}} = \rho^2 h_2(x).$$

The pattern persists, and $U_\rho h_d(x) = \rho^d h_d(x)$. This quickly implies that in the general case,

$$U_\rho f = \sum_{d=0}^{\infty} \rho^d f^{=d}.$$

(This is why we defined the grading in this particular way.)

In more detail, consider the following way of generating ρ-correlated normal random variables: choose two unit vectors a, b in Euclidean space whose inner product is ρ, and let z be a random multivariate Gaussian in the same space. It is not hard to check that $x = \langle a, z \rangle$ and $y = \langle b, z \rangle$ are standard Gaussians, and furthermore $\mathbb{E}[\langle a, z \rangle \langle b, z \rangle] = \langle a, b \rangle = \rho$. So x, y are ρ-correlated Gaussians. Now

$$\mathbb{E}[e^{sx+ty}] = \mathbb{E}[e^{\sum_i (sa_i + tb_i) z_i}] = e^{\sum_i (sa_i + tb_i)^2/2} = e^{s^2/2 + t^2/2 + \rho st}.$$

Therefore

$$\mathbb{E}[e^{sx-s^2/2} e^{ty-t^2/2}] = e^{\rho st} = \sum_{i=0}^{\infty} \rho^i \frac{s^i t^i}{i!}.$$

On the other hand,

$$\mathbb{E}[e^{sx-s^2/2} e^{ty-t^2/2}] = \sum_{i,j=0}^{\infty} \mathbb{E}[h_i(x) h_j(y)] \frac{s^i t^j}{\sqrt{i!j!}}.$$

Comparing coefficients, we see that $\langle U_\rho h_i, h_i \rangle = \rho^i$.

7.3 Laplacian and noise stability

Recall the Laplacian operator $Lf = \frac{1}{2}\sum_i [f - f^{\oplus i}] = \sum_S |S| \hat{f}(S) \chi_S$ which we defined on the cube (actually without the normalization!) and satisfies $\mathrm{Inf}[f] = \langle Lf, f \rangle$. Another way to derive the same operator is by taking the derivative of the noise operator, in the following sense:

$$\frac{d}{d\rho} \langle T_\rho f, g \rangle = \frac{d}{d\rho} \sum_S \rho^{|S|} \hat{f}(S) \hat{g}(S) = \sum_S |S| \rho^{|S|-1} \hat{f}(S) \hat{g}(S) = \rho^{-1} \langle L T_\rho f, g \rangle.$$

Another parameterization is even more illuminating:

$$\frac{d}{dt} \langle T_{e^{-t}} f, g \rangle = \frac{d}{dt} \sum_S e^{-t|S|} \hat{f}(S) \hat{g}(S) = -\sum_S |S| e^{-t|S|} \hat{f}(S) \hat{g}(S) = -\langle L T_{e^{-t}} f, g \rangle.$$

We can thus write $T_\rho = \rho^L$ and $T_{e^{-t}} = e^{-tL}$.

When doing the same in Gaussian space, we obtain the following expression for the Laplacian:

$$Lf(x) = \langle x, \nabla f(x) \rangle - \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i^2}(x).$$

The proof is via a simple Taylor expansion: for a univariate f and $z \sim N(0,1)$,

$$\begin{aligned} U_\rho f(x) - f(x) &= \mathbb{E}[f(\rho x + \sqrt{1-\rho^2} z)] - f(x) \\ &= f(\rho x) - f(x) + \sqrt{1-\rho^2}\, \mathbb{E}[z] f'(\rho x) + \frac{1}{2}(1-\rho^2)\, \mathbb{E}[z^2] f''(\rho x) + O((1-\rho^2)^{3/2}\, \mathbb{E}[z^3]) \\ &= f(\rho x) - f(x) + \frac{1}{2}(1-\rho^2) f''(\rho x) + o(1-\rho^2). \end{aligned}$$

Dividing by $\rho - 1$ and taking the limit $\rho \to 1$, we obtain

$$\begin{aligned} Lf(x) = \lim_{\rho \to 1} \frac{f(\rho x) - f(x)}{\rho - 1} - f''(x) = \\ \frac{d}{d\rho} f(\rho x)\Big|_{\rho=1} - f''(x) = x f'(\rho x)|_{\rho=1} - f''(x) = x f'(x) - f''(x). \end{aligned}$$

As in the case of the cube, we also have a spectral expression:

$$Lf = \sum_{d=0}^\infty d f^{=d}.$$

Furthermore, we have the nice formula

$$\langle Lf, g \rangle = \langle \nabla f, \nabla g \rangle.$$

This formula follows from integration by parts, and the fact that the normal density φ satisfies the differential equation $\varphi'(x) = -x\varphi(x)$: in the one-dimensional case,

$$\begin{aligned} \langle Lf, g \rangle = \int_{-\infty}^\infty (x f'(x) - f''(x)) g(x) \varphi(x)\, dx = \int_{-\infty}^\infty x f'(x) g(x) \varphi(x)\, dx + \int_{-\infty}^\infty f'(x) (g\varphi)'(x)\, dx = \\ \int_{-\infty}^\infty x f'(x) g(x) \varphi(x)\, dx + \int_{-\infty}^\infty f'(x)(g'(x)\varphi(x) - x g(x) \varphi(x))\, dx = \int_{-\infty}^\infty f'(x) g'(x) \varphi(x)\, dx. \end{aligned}$$

7.4 Hypercontractivity

The Gaussian noise operator is contractive in L^2 due to the explicit formula we gave above. But more generally, for every $p \geq 1$ we have

$$\|U_\rho f\|_p^p = \mathop{\mathbb{E}}_x[|U_\rho f(x)|^p] = \mathop{\mathbb{E}}_x[|\mathop{\mathbb{E}}_y f(\rho x + \sqrt{1-\rho^2} y)|^p].$$

Applying Jensen's inequality (using the convexity of $|t|^p$), we see that

$$\|U_\rho f\|_p^p \leq \mathop{\mathbb{E}}_{x,y}[|f(\rho x + \sqrt{1-\rho^2}y)|^p] = \|f\|_p^p.$$

Hypercontractivity remains true even in Gaussian space, and we can prove this using the central limit theorem. Here is the basic idea. Consider for simplicity the case $n = 1$. The central limit theorem tells us that if $x_1, \ldots, x_m$ are Bernoulli random variables then the distribution of $\sigma_m = \frac{x_1+\cdots+x_m}{\sqrt{m}}$ approaches that of a standard Gaussian. Moreover, if $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_m)$ are ρ-correlated ($\mathbb{E}[x_i y_i] = \rho$) and we define $\tau_m = \frac{y_1+\cdots+y_m}{\sqrt{m}}$ then $\mathbb{E}[\sigma_m \tau_m] = \rho$, and so the joint distribution of (σ_m, τ_m) tends to $N(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right))$.

Recall that our hypercontractive estimates ultimately follow from inequalities of the type

$$\langle T_\rho f, g \rangle \leq \|f\|_q \|g\|_{q/(q-1)}.$$

Suppose we know such an estimate for some ρ, q. We lift it to Gaussian space by considering, for each $f, g \in L^2$ and m, the functions f_m, g_m on the m-dimensional cube (we are still assuming that $n = 1$!) given by $f_m(x_1, \ldots, x_m) = f(\sigma_m)$, $g_m(y_1, \ldots, y_m) = g(\tau_m)$. The central limit theorem shows that $\|f_m\|_q \to \|f\|_q$ (the first norm on the cube, the second in Gaussian space), and similarly $\|g_m\|_{q/(q-1)} \to \|g\|_q$. Moreover, since

$$\langle T_\rho f_m, g_m \rangle = \mathop{\mathbb{E}}_{x,y \ \rho\text{-correlated}}[f_m(x) g_m(y)],$$

the multidimensional central limit theorem also shows that $\langle T_\rho f_m, g_m \rangle \to \langle U_\rho f, g \rangle$. Taking the limit $m \to \infty$, we recover the same inequality for Gaussian space.

Differentiating the optimal hypercontractivity estimate yields the Gaussian log Sobolev inequality:

$$\mathbb{E}[f^2 \log(f^2)] - \mathbb{E}[f^2] \log \mathbb{E}[f^2] \leq 2\,\mathbb{E}[\|\nabla f\|^2].$$

7.5 Isoperimetry

Our future application for Majority is Stablest requires an isoperimetric theorem due to Borell. It states that among all functions $f \colon \mathbb{R}^n \to [0,1]$ with given mean, the one minimizing $\langle f, U_\rho f \rangle$, for *any* $\rho \in [0,1]$, is the indicator of a hyperplane (say $x_1 \leq t$ for an appropriate t). We can rotate the hyperplane so that this function is the indicator of an event of the form $x_1 + \cdots + x_n \leq t$, which is the Gaussian analog of majority. Using the invariance principle, we will deduce a corresponding statement for the Boolean cube. (The statement can also be proved directly, but for didactic purposes we don't do that.)

There are several known proofs of Borell's theorem [Bor75]. Apart from Borell's proof which used Erhard symmetrization, there is an elementary proof due to De, Mossel and Neeman using convexity [DMN13], another proof due to Mossel and Neeman [MN14] using the semigroup method, and a proof of Eldan [Eld14] using Brownian motion. We present some ideas from the proof of De, Mossel and Neeman. They actually prove a two-function version of Borell's theorem. For all $\rho \in (0,1)$ and $\alpha, \beta \in [0,1]$, define

$$\Lambda_\rho(\alpha, \beta) = \Pr_{(x,y) \sim N_\rho}[\Phi(x) \leq \alpha, \Phi(y) \leq \beta].$$

Here Φ is the CDF of the standard Gaussian, and $N_\rho = N(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right))$. We will also use ϕ for its density. What Mossel and Neeman show is that every two functions $f, g \in L^2(\mathbb{R}^n)$ bounded by $[0,1]$ satisfy

$$\mathop{\mathbb{E}}_{(x,y) \sim N_\rho^n}[\Lambda_\rho(f(x), g(y))] \leq \Lambda_\rho(\mathbb{E}[f], \mathbb{E}[g]).$$

In order to deduce Borell's theorem, take $g = f$, and suppose first that f is Boolean. In this case it is not hard to check that $\Lambda_\rho(f(x), f(y)) = f(x)f(y)$, and so the left-hand side is just $\langle f, U_\rho f \rangle$. On the right-hand side we have $\Lambda_\rho(\mathbb{E}[f], \mathbb{E}[f])$, which is the noise stability of a hyperplane.

When f is not Boolean, we can express it as an infinite convex combination $\sum_i \alpha_i f_i$ of Boolean functions *with the same mean*. Now

$$\sqrt{\langle f, U_\rho f \rangle} = \|U_{\sqrt{\rho}} f\| \leq \sum_i \alpha_i \|U_{\sqrt{\rho}} f_i\| \leq \sum_i \alpha_i \sqrt{\Lambda_\rho(\mathbb{E}[f], \mathbb{E}[f])} = \sqrt{\Lambda_\rho(\mathbb{E}[f], \mathbb{E}[f])}.$$

We can also go the other way, deducing the functional version in n dimensions from the set version in $n+1$ dimensions. Define sets F and G by $(x,z) \in F$ if $\Phi(z) \leq f(x)$, and $(y,z) \in G$ if $\Phi(z) \leq g(y)$. Notice that

$$\mathbb{E}[F] = \mathop{\mathbb{E}}_x[\Pr_z[\Phi(z) \leq f(x)]] = \mathop{\mathbb{E}}_x[f(x)] = \mathbb{E}[f],$$

and similarly $\mathbb{E}[G] = \mathbb{E}[g]$. For all x, y we have

$$\mathop{\mathbb{E}}_{(z,w)\sim N_\rho}[\Lambda_\rho(F(x,z),G(y,w))] = \mathop{\mathbb{E}}_{(z,w)\sim N_\rho}[F(x,z)G(y,w)] = \mathop{\mathbb{E}}_{(z,w)\sim N_\rho}[\Pr[\Phi(z) \leq f(x), \Phi(w) \leq g(y)]] = \Lambda_\rho(f(x),g(y)).$$

Therefore

$$\mathop{\mathbb{E}}_{(x,y)\sim N_\rho^n} \mathop{\mathbb{E}}_{(z,w)\sim N_\rho}[\Lambda_\rho(F(x,z),G(y,w))] = \mathop{\mathbb{E}}_{(x,y)\sim N_\rho^n}[\Lambda_\rho(f(x),g(y))].$$

We can thus deduce the inequality for f, g from the inequality for F, G.

We will actually prove this inequality only for $n = 1$. The result for general n follows by a simple induction. Suppose that we know the result for some n, and that $f, g \in L^2(\mathbb{R}^{n+1})$. For each a, b, by restricting the last coordinate we obtain functions $f_a, g_b \in L^2(\mathbb{R}^n)$. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim N_\rho^{n+1}}[\Lambda_\rho(f(x),g(y))] &= \mathop{\mathbb{E}}_{(a,b)\sim N_\rho} \mathop{\mathbb{E}}_{(x,y)\sim N_\rho^n}[\Lambda_\rho(f_a(x),g_b(y))] \\
&\leq \mathop{\mathbb{E}}_{(a,b)\sim N_\rho}[\Lambda_\rho(\mathbb{E}[f_a],\mathbb{E}[g_b])] \\
&\leq \Lambda_\rho(\mathop{\mathbb{E}}_a \mathbb{E}[f_a], \mathop{\mathbb{E}}_b \mathbb{E}[g_b]) = \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g]).
\end{aligned}$$

Calculation shows that the following modified Hessian matrix is negative semidefinite at any point x, y, for any $\sigma \in [0, \rho]$:

$$M_\sigma(x,y) = \begin{pmatrix} \frac{\partial^2 \Lambda_\rho(x,y)}{\partial x^2} & \sigma \frac{\partial^2 \Lambda_\rho(x,y)}{\partial x \partial y} \\ \sigma \frac{\partial^2 \Lambda_\rho(x,y)}{\partial x \partial y} & \frac{\partial^2 \Lambda_\rho(x,y)}{\partial y^2} \end{pmatrix}.$$

This can be summarized by saying that Λ_ρ is *ρ-concave*. A Taylor expansion shows that

$$\Lambda_\rho(\mathbb{E}[f]+a,\mathbb{E}[g]+b) = \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g]) + a\frac{\partial \Lambda_\rho}{\partial x}(\mathbb{E}[f]) + b\frac{\partial \Lambda_\rho}{\partial y}(\mathbb{E}[g]) + \frac{1}{2}\begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x^2} & \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x \partial y} \\ \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x \partial y} & \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial y^2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \epsilon(a,b),$$

where ϵ consists of third-order terms. Therefore

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim N_\rho}[\Lambda_\rho(f(x),g(y))] = \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g]) + \mathbb{E}[f-\mathbb{E}[f]]\frac{\partial \Lambda_\rho}{\partial x}(\mathbb{E}[f]) + \mathbb{E}[g-\mathbb{E}[g]]\frac{\partial \Lambda_\rho}{\partial y}(\mathbb{E}[g]) + \\
&\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim N_\rho}\begin{pmatrix} f(x)-\mathbb{E}[f] & g(y)-\mathbb{E}[g] \end{pmatrix} \begin{pmatrix} \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x^2} & \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x \partial y} \\ \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x \partial y} & \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial y^2} \end{pmatrix} \begin{pmatrix} f(x)-\mathbb{E}[f] \\ g(y)-\mathbb{E}[g] \end{pmatrix} + \mathbb{E}[\epsilon(f,g)].
\end{aligned}$$

Since $\mathbb{E}[f-\mathbb{E}[f]] = \mathbb{E}[g-\mathbb{E}[g]] = 0$, the linear terms cancel. If we expand $f-\mathbb{E}[f], g-\mathbb{E}[g]$ in the Hermite expansion, then we find that

$$\mathop{\mathbb{E}}_{(x,y)\sim N_\rho}[(f(x)-\mathbb{E}[f])(g(y)-\mathbb{E}[g])] = \sum_{i=1}^\infty \rho^i \hat{f}(i)\hat{g}(i) \leq \rho\sqrt{\sum_{i=1}^\infty \hat{f}(i)^2}\sqrt{\sum_{i=1}^\infty \hat{g}(i)^2} = \rho\sqrt{\mathbb{V}[f]\,\mathbb{V}[g]},$$

using Cauchy–Schwartz. If we put $\sigma_f = \sqrt{\mathbb{V}[f]}$, $\sigma_g = \sqrt{\mathbb{V}[g]}$, and $\sigma = \mathbb{E}[(f-\mathbb{E}[f])(g-\mathbb{E}[g])]/\sigma_x\sigma_y \leq \rho$, then the quadratic term is

$$\frac{1}{2}\begin{pmatrix} \sigma_f & \sigma_g \end{pmatrix} \begin{pmatrix} \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x^2} & \sigma\frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x \partial y} \\ \sigma\frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial x \partial y} & \frac{\partial^2 \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g])}{\partial y^2} \end{pmatrix} \begin{pmatrix} \sigma_f \\ \sigma_g \end{pmatrix} \leq 0,$$

and so

$$\mathop{\mathbb{E}}_{(x,y)\sim N_\rho}[\Lambda_\rho(f(x),g(y))] \leq \Lambda_\rho(\mathbb{E}[f],\mathbb{E}[g]) + \mathbb{E}[\epsilon(f,g)].$$

This is almost what we want, the only problem being the third-order error term! There are two ways of getting rid of it. The first, used by Mossel and Neeman, is to use the semigroup method instead, interpolating smoothly between the left-hand side and the right-hand side; the same ρ-concavity property has to be used. De, Mossel and Neeman instead estimate the error term, showing that it tends to zero if f, g have low influences (we will see similar phenomena when we discuss Majority is Stablest). An application of the central limit theorem then completes the proof.

Balanced functions We now derive an explict formula for the bound in Borell's theorem when $\mathbb{E}[f] = 1/2$. It is more natural to consider ± 1-valued functions rather than $\{0,1\}$-valued functions, since now the condition is $\mathbb{E}[f] = 0$. The optimal function is still a hyperplane; for definiteness, we can choose f to simply be the sign function. Its noise stability is

$$\Pr_{(x,y)\sim N_\rho}[x, y \leq 0 \text{ or } x, y \geq 0] - \Pr_{(x,y)\sim N_\rho}[x \leq 0 \leq y \text{ or } y \leq 0 \leq x] = 1 - 2\Pr_{(x,y)\sim N_\rho}[x \leq 0 \leq y \text{ or } y \leq 0 \leq x].$$

We will calculate the probability that x, y have different signs. Let $\rho = \cos\theta$. One way to generate such correlated x, y is to choose two unit norm 2-dimensional vectors u, v at an angle of θ, and taking $x = \langle u, g\rangle$, $y = \langle v, g\rangle$, where g is a 2-dimensional standard Gaussian. Indeed, clearly $\mathbb{E}[x] = \mathbb{E}[y] = 0$, $\mathbb{E}[x^2] = \mathbb{E}[y^2] = 1$, and $\mathbb{E}[xy] = \langle u, v\rangle = \cos\theta = \rho$. We can assume that u is the vector at angle 0, and v is the vector at angle θ. If the angle of g is γ then $\langle u, g\rangle \geq 0$ if $\gamma \in [-\pi/2, \pi/2]$, and $\langle v, g\rangle \geq 0$ if $\gamma \in [\theta - \pi/2, \theta + \pi/2]$. Therefore x, y have different signs with probability θ/π. We conclude that the noise stability is

$$1 - \frac{2\cos^{-1}\rho}{\pi}.$$

8 Invariance principle (16 December 2015)

([O'D14, Sections 11.5–11.7])

8.1 Berry–Esseen

The central limit theorem states that if $X_1, X_2, \ldots$ is an infinite collection of independent "reasonable" random variables then the random variable

$$\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{\sum_{i=1}^n \mathbb{V}[X_i]}}$$

converges in distribution to a Gaussian $N(0,1)$. The Berry–Esseen theorem is a quantitative version of the central limit theorem, bounding the speed of convergence.

Let $X_1, \ldots, X_n$ be an infinite collection of independent random variables with zero mean and unit norm. Our object of study will be the random variable

$$X = \sum_i c_i X_i.$$

We would like to say that the distribution of X is similar to the distribution of the corresponding normal random variable $N(0, \sum_i c_i^2)$, but this is not always the case. Suppose for example that $c_1 = 1$ while $c_2 = \cdots = c_n = 0$. In this case we cannot really say much, since $X = X_1$. The Berry–Esseen theorem states, qualitatively, then if none of the c_i is too large compared to the others, then X is indeed close to the corresponding normal.

The idea of the proof is to replace the random variables X_i one by one by independent standard Gaussians $G_1, \ldots, G_n$. This will show that

$$X \approx \sum_i c_i G_i,$$

and the latter random variable has the required distribution. Although the measure of approximation we're eventually after is CDF distance, the proof becomes much easier by considering *test functions*, which are functions satisfying certain properties, in this case having a bounded third derivative. We will show that for each test function ψ satisfying $\|\psi'''\|_\infty \leq B$,

$$\mathbb{E}[\psi(c_1 X_1 + \cdots + c_n X_n)] \approx \mathbb{E}[\psi(c_1 G_1 + \cdots + c_n G_n)].$$

In order to understand the dependence of X on X_1, we use a Taylor expansion. Let $Y \sim c_2 X_2 + \cdots + c_n X_n$. Then

$$\mathbb{E}[\psi(X)] = \mathop{\mathbb{E}}_{Y,X_1}[\psi(Y + c_1 X_1)] = \mathop{\mathbb{E}}_{Y}[\psi(Y) + c_1 X_1 \psi'(Y) + \tfrac{1}{2} c_1^2 X_1^2 \psi''(Y) + \tfrac{1}{6} c_1^3 X_1^3 \psi'''(Z)],$$

where Z is some point in the interval connecting Y and $Y + c_1 X_1$. Using the bound on the third derivative, we obtain

$$\mathbb{E}[\psi(X)] = \mathop{\mathbb{E}}_{Y}[\psi(Y)] + \frac{1}{2} c_1^2 \mathop{\mathbb{E}}_{Y}[\psi''(Y)] \pm \frac{1}{6} c_1^3 \,\mathbb{E}[|X_1|^3] B.$$

We get a very similar estimate if we replace X_1 by G_1:

$$\mathbb{E}[\psi(c_1 G_1 + Y)] = \mathop{\mathbb{E}}_{Y}[\psi(Y)] + \frac{1}{2} c_1^2 \mathop{\mathbb{E}}_{Y}[\psi''(Y)] \pm \frac{1}{6} c_1^3 \,\mathbb{E}[|G_1|^3] B.$$

Therefore, using $\mathbb{E}[|G_1|^3] = \gamma_3$,

$$|\,\mathbb{E}[\psi(c_1 X_1 + Y)] - \mathbb{E}[\psi(c_1 G_1 + Y)]| \leq \frac{B}{6} c_1^3 (\mathbb{E}[|X_1|^3] + \gamma_3).$$

Continuing this way, replacing all of $X_1, \ldots, X_n$, we deduce

$$|\,\mathbb{E}[\psi(c_1 X_1 + \cdots + c_n X_n)] - \mathbb{E}[\psi(c_1 G_1 + \cdots + c_n G_n)]| \leq \frac{B}{6} \sum_{i=1}^n c_i^3 (\mathbb{E}[|X_1|^3] + \gamma_3).$$

Note that since $\mathbb{E}[|X_i|^3] = \|X_i\|_3^3 \geq \|X_i\|_2^3 = 1$, $\mathbb{E}[|X_i|^3] + \gamma_3 \leq (1 + \gamma_3)\,\mathbb{E}[|X_i|^3]$. When is the right-hand side small? It is natural to normalize by requiring $\sum_i c_i^2 = 1$, and then we would like $\sum_{i=1}^n c_i^3$ to be small. Since $\sum_{i=1}^n c_i^3 \leq \max_i |c_i|$, it suffices to assume that all of the individual coefficients are small.

CDF distance What we are really interested in is CDF distance, and to that end we would like the function ψ to be $\psi(x) = 1_{x<t}$. However, this function isn't smooth. Instead, we use a function $\psi = \psi_{t,\eta}$ with the following properties:

- $\psi(x) = 0$ for $x \leq t$.
- $0 \leq \psi(x) \leq 1$ for $t \leq x \leq t + \eta$.
- $\psi(x) = 1$ for $x \geq t + \eta$.
- $\|\psi'''\|_\infty = O(1/\eta^3)$.

Here η is a parameter that we can choose. Constructing ψ is a technical exercise which we happily skip.

What does this give us? Let $K = \sum_{i=1}^n c_i^3 \mathbb{E}[|X_i|^3]$ and $G = \sum_i c_i G_i = N(0, \sum_i c_i^2)$. For simplicity, assume that in fact $\sum_i c_i^2 = 1$. For every $\eta > 0$,

$$|\mathbb{E}[\psi_{t,\eta}(X)] - \mathbb{E}[\psi_{t,\eta}(G)]| = O(K/\eta^3).$$

For any random variable R,

$$\Pr[R \leq t] \leq \mathbb{E}[\psi_{t,\eta}(R)] \leq \Pr[R \leq t + \eta].$$

This allows us to get the estimate

$$\Pr[X \leq t] \leq \mathbb{E}[\psi_{t,\eta}(X)] \leq \mathbb{E}[\psi_{t,\eta}(G)] + O(K/\eta^3) \leq \Pr[G \leq t + \eta] + O(K/\eta^3).$$

Similarly, by considering $\psi_{t-\eta,\eta}$ instead, we can show that

$$\Pr[X \leq t] \geq \Pr[G \leq t - \eta] - O(K/\eta^3).$$

Choosing $\eta = K^{1/4}$, we see that the *Lévy distance* between X and G is at most $O(K^{1/4})$.

This is almost what we wanted. To get what we actually wanted, we need to use the anticoncentration of Gaussians:

$$\Pr[t \leq G \leq t + \eta] \leq \frac{\eta}{\sqrt{2\pi}} = O(\eta),$$

a property which follows from the fact that the density of G is pointwise at most $1/\sqrt{2\pi}$. How does this help? As we have seen,

$$\Pr[X \leq t] \leq \Pr[G \leq t + \eta] + O(K/\eta^3) \leq \Pr[G \leq t] + O(\eta + K/\eta^3).$$

Choosing $\eta = K^{1/4}$, we get that $\Pr[X \leq t] \leq \Pr[G \leq t] + O(K^{1/4})$. The other inequality is proved in the same way, and so

$$|\Pr[X \leq t] - \Pr[G \leq t]| = O(K^{1/4}).$$

The actual Berry–Esseen theorem is even stronger: it gives an upper bound of the form $O(K)$. The advantage of our particular proof is that it generalizes to low-degree polynomials.

8.2 Invariance principle

The celebrated non-linear invariance principle of Mossel, O'Donnell and Oleszkiewicz [MOO10] is a generalization of the Berry–Esseen theorem to low-degree polynomials. The condition that all c_i be small is replaced by the condition that all variables have low influence; otherwise the proof is very similar. For concreteness, we will treat the case in which X_i are balanced Bernoulli random variables. In this case we can improve on the argument above by taking a fourth-order Fourier expansion instead of a third-order one. Therefore our test functions will need to have a bounded fourth derivative: $B = \|\psi''''\|_\infty$.

Consider some multilinear polynomial $f(X_1, \ldots, X_n)$ of degree d. Let us try to mimic the same proof as above. Write $f(x_1, \ldots, x_n) = g(x_2, \ldots, x_n) + x_1 h(x_2, \ldots, x_n)$. Then

$$\mathbb{E}[\psi(f(X_1, \ldots, X_n))] = \mathop{\mathbb{E}}_{X_2,\ldots,X_n} \mathop{\mathbb{E}}_{X_1} [\psi(X_1 g(X_2, \ldots, X_n) + h(X_2, \ldots, X_n))].$$

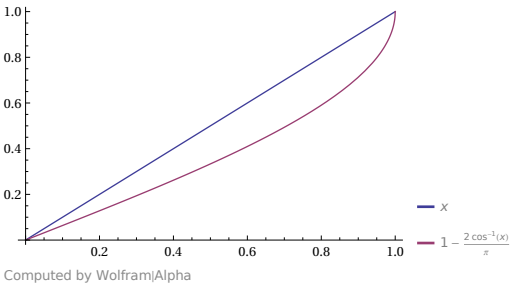

Figure 1: Noise stability of dictator (blue) and majority (red)

We now take a Taylor expansion, and working as before conclude that

$$\mathbb{E}[\psi(f(X_1,\ldots,X_n))] = \mathop{\mathbb{E}}_{X_2,\ldots,X_n}[h + \tfrac{1}{2}g^2] \pm O(B \mathop{\mathbb{E}}_{X_2,\ldots,X_n}[g^4]).$$

A similar expression holds for Gaussians, and so

$$|\,\mathbb{E}[\psi(f(X_1,X_2,\ldots,X_n)) - \psi(f(G_1,X_2,\ldots,X_n))]| = O(B \mathop{\mathbb{E}}_{X_2,\ldots,X_n}[g^4]).$$

What can we say about the error term? Note that g is nothing else then the discrete derivative of f, and so $\mathbb{E}[g^2] = \mathrm{Inf}_1[f]$. Hypercontractivity thus allows us to bound

$$\mathbb{E}[g^4] \leq O(1)^d\,\mathbb{E}[g^2]^2 = C^d\,\mathrm{Inf}_1[f]^2,$$

for some constant C. Continuing this way (using the fact that both Bernoullis and Gaussians are hypercontractive), we see that

$$|\,\mathbb{E}[\psi(f(X_1,\ldots,X_n))] - \mathbb{E}[\psi(f(G_1,\ldots,G_n))]| = O(BC^d \sum_i \mathrm{Inf}_i[f]^2) \leq O(BC^d d\|f\|^2\tau), \text{ where } \tau = \max_i \mathrm{Inf}_i[f].$$

Here we used $\sum_i \mathrm{Inf}_i[f] \leq d\,\mathbb{V}[f] \leq d\|f\|^2$. We see that what controls the quality of the approximation is the parameter τ. Also, this time $f(G_1,\ldots,G_n)$ need not be a Gaussian, but rather it is generally some arbitrary function in Gaussian space.

Corollaries As before, we can deduce as corollaries bounds on the Lévy and CDF distances. The bound on the Lévy distance is proved in much the same way, the resulting bound being of the form $O(C_1^d\tau^{1/4})$. In order to bound the CDF distance, we need to use an anticoncentration result for Gaussian polynomials due to Carbery and Wright [CW01]. The result states that a degree d Gaussian polynomial of unit norm can lie in an interval of length ϵ with probability at most $O(d\epsilon^{1/d})$. This leads to a bound on the CDF distance of the form $O(C_2^d\tau^{1/O(d)})$.

8.3 Majority is Stablest

Borell's theorem states that among all Boolean functions in Gaussian space with mean $1/2$, the most noise stable is a hyperplane. What happens for functions on the Boolean cube? The analog of a hyperplane is the majority function, which the central limit theorem shows has roughly the same noise stability as a hyperplane, and so (for a 0/1-valued function) roughly $1/2 - \cos^{-1}\rho/2\pi$. Are there any better functions? Certainly: the function $f(x) = x_1$ is balanced, and has noise stability $(1+\rho)/4$, which is much larger than that of majority (see Figure 1, which plots the noise stabilities of the corresponding ± 1-valued functions). However, it turns out that if we bound the maximal influence of f, then we get a bound approaching $1/2 - \cos^{-1}\rho/2\pi$. This is the Majority is Stablest conjecture (now theorem).

What prompted the authors to prove the invariance principle was the Majority is Stablest conjecture. Unfortunately, deducing it from Borell's theorem via the invariance principle is slightly messy. One major problem is that an arbitrary Boolean function doesn't have low degree! Fortunately, this can be fixed

by applying a little noise. Let f be a Boolean function with mean α and maximal influence τ. For some small δ, let $g = T_{1-\delta}f$. Also, let ψ be any C-Lipschitz function (this condition is somewhat more useful than the condition on bounded fourth derivative). For any d we can write

$$|\mathbb{E}[\psi(g(X))] - \mathbb{E}[\psi(g(G))]| \leq |\mathbb{E}[\psi(g^{\leq d}(X))] - \mathbb{E}[\psi(g^{\leq d}(G))]| + O(C\|g^{>d}(X)\|_1) + O(C\|g^{>d}(G)\|_1).$$

We can bound $\|g^{>d}\|_1 \leq \|g^{>d}\|_2 \leq (1-\delta)^d\|f\|_2 \leq (1-\delta)^d \leq e^{-d\delta}$. In order to bound the first term, we would like to use the invariance principle. To that end, we construct some smooth approimation to ψ. It turns out that such smooth approximations ψ_η exist, with the following properties:

- $\|\psi_\eta''''\|_\infty = O(1/\eta^3)$.
- $\|\psi - \psi_\eta\|_\infty \leq \eta$.

Applying the invariance principle to ψ_η allows us to bound the first term above by $O((C^d/\eta^3)\tau + \eta)$. Choosing $\eta = C^{d/4}\tau^{1/4}$, we conclude that

$$|\mathbb{E}[\psi(g(X))] - \mathbb{E}[\psi(g(G))]| = O(C^d\tau^{1/4} + e^{-d\delta}).$$

A good choice for d is $\log_C(1/\tau^{1/8})$, and this gives an overall bound of $O(\tau^{1/8} + \tau^{-O(\delta)})$.

Why is it kosher to apply $T_{1-\delta}$? Since $\langle f, T_\rho f\rangle = \langle g, T_{\rho'}g\rangle$, where $\rho = \rho'(1-\delta)^2$. Consider the function S given by

$$S(x) = \begin{cases} 0 & x \leq 0, \\ x^2 & 0 \leq x \leq 1, \\ 1 & 1 \leq x. \end{cases}$$

Since g is bounded by $[0,1]$, over the cube $\langle g, T_{\rho'}g\rangle = \mathbb{E}[(T_{\sqrt{\rho'}}g)^2] = \mathbb{E}[S(T_{\sqrt{\rho'}}g)]$. Applying the invariance principle, we see that

$$\mathbb{E}[S(U_{\sqrt{\rho'}}g)] = \langle g, T_{\rho'}g\rangle \pm O(\tau^{1/8} + \tau^{-O(\delta)}).$$

The idea now is to apply Borell's theorem to the function g. Unfortunately, while on the Boolean cube g is certainly bounded by $[0,1]$, this is not necessarily the case in Gaussian space. However, the invariance principle shows that it is *almost* the case. To that end, define the truncation of g by $h = \min(1, \max(0, g))$. This is a function to which Borell's theorem does apply, and using the invariance principle, we can show that it is quite close to g. Indeed, the function $\mathrm{dist}_{[0,1]}$, the distance to the interval $[0,1]$, is 1-Lipschitz, and on the Boolean cube $\mathbb{E}[\mathrm{dist}_{[0,1]}(g)] = 0$. Therefore

$$\mathbb{E}[\mathrm{dist}_{[0,1]}(g(G))] = O(\tau^{1/8} + \tau^{-O(\delta)}).$$

This quantity bounds $\|g - h\|_1$ and so $|\mathbb{E}[g] - \mathbb{E}[h]|$. Since $\Lambda_\rho(\alpha, \alpha)$ is itself Lipschitz, Borell's theorem shows that

$$\mathbb{E}[S(U_{\sqrt{\rho'}}h)] = \mathbb{E}[(U_{\sqrt{\rho'}}h)^2] = \langle h, U_{\rho'}h\rangle \leq \frac{1}{2} - \frac{\cos^{-1}\rho'}{2\pi} + O(\tau^{1/8} + \tau^{-O(\delta)}).$$

Note that we find here ρ' instead of ρ, but luckily $\rho' = \rho/(1-\delta)^2$ is close to ρ, and so assuming δ is small enough, $|\cos^{-1}\rho' - \cos^{-1}\rho| = O(|\rho - \rho'|) = O(\delta)$. Since S is Lipschitz,

$$\mathbb{E}[S(U_{\sqrt{\rho'}}g)] \leq \mathbb{E}[S(U_{\sqrt{\rho'}}h)] + O(\tau^{1/8} + \tau^{-O(\delta)}) \leq \frac{1}{2} - \frac{\cos^{-1}\rho}{2\pi} + O(\tau^{1/8} + \tau^{-O(\delta)} + \delta).$$

Putting everything together, we deduce that

$$\langle f, T_\rho f\rangle = \langle g, T_{\rho'}g\rangle \leq \frac{1}{2} - \frac{\cos^{-1}\rho}{2\pi} + O(\tau^{1/8} + \tau^{-O(\delta)} + \delta).$$

For any fixed δ, as $\tau \to 0$, the error term tends to 0. By choosing δ appropriately, this shows that as $\tau \to 0$, the error term tends to 0, completing the proof of Majority is Stablest.

If f is $\{-1,1\}$-valued with $\mathbb{E}[f] = 0$ then $F = (f+1)/2$ is $\{0,1\}$-valued with $\mathbb{E}[F] = 1/2$, and

$$\langle F, T_\rho F\rangle = \frac{1}{4}\langle f+1, T_\rho f + 1\rangle = \frac{1}{4}(\langle f, T_\rho f\rangle + \mathbb{E}[f] + \mathbb{E}[T_\rho f] + 1) = \frac{1}{4}(\langle f, T_\rho f\rangle + 1).$$

Therefore

$$\langle f, T_\rho f\rangle = 4\langle F, T_\rho F\rangle - 1 \leq 1 - \frac{2}{\pi}\cos^{-1}\rho + O(\tau^{1/8} + \tau^{-O(\delta)} + \delta).$$

9 Max Cut (28 December 2015)

([O'D14, Section 11.7], [KKMO07], [GO08])

MAX CUT is the following optimization problem. Given an edge-weighted graph $G = (V, E)$, find a partition $V = S \cup T$ that maximizes the total weight of edges crossing the partition. This problem is NP-hard. How well can we approximate it? In a paper describing semidefinite programming, Goemans and Williamson [GW95] suggested the following semidefinite relaxation. Each vertex i is associated with a unit vector v_i. If we could force the v_i to be one-dimensional then $v_i \in \{-1, 1\}$, and then the indicator of $v_i \neq v_j$ (i.e., an edge is cut) is simply $\frac{1}{2} - \frac{1}{2} v_i v_j$. This suggests maximizing the following objective function:

$$\sum_{(i,j)\in E} \frac{w(e)}{2} \left[1 - \langle v_i, v_j \rangle\right],$$

under the constraints $\|v_i\| = 1$. Alternatively, we can write this as follows:

$$\begin{aligned} & \max \sum_{(i,j)\in E} \frac{w(e)}{2} \left[1 - A_{ij}\right] \\ s.t. \quad & A_{ii} = 1 \text{ for all } i \\ & A \succeq 0 \end{aligned}$$

Here $A \succeq 0$ means that A is a symmetric positive semidefinite matrix. Given a solution A, we can recover vectors v_i using Cholesky decomposition.

Suppose that we solve this semidefinite program, obtaining unit vectors $v_1, \ldots, v_n$. How do we come up with a partition? Goemans and Williamson suggest picking a random direction g, and putting a vertex x in S if $\langle v_x, g \rangle \geq 0$. How good is this solution? Suppose that $\langle v_i, v_j \rangle = \cos\theta$. As we have seen before (when calculating the noise stability of hyperplanes), the probability that this edge is cut is θ/π. This shows that the approximation ratio of the algorithm is

$$\min_\theta \frac{\theta/\pi}{(1-\cos\theta)/2} = \min_\theta \frac{2}{\pi} \frac{\theta}{1-\cos\theta}.$$

It turns out that this minimum is roughly 0.8786, achieved at some angle $\theta \approx 0.74\pi$.

Amazingly, assuming the Unique Games Conjecture, this is optimal! This was shown by Khot, Kindler, Mossel and O'Donnell [KKMO07] conditional on the Majority is Stablest conjecture, and it was this application that provided the impetus for the machinery of the invariance principle.

9.1 Integrality gap

There are two natural questions that one could ask. First, is the analysis of the Goemans–Williamson algorithm tight? Second, is there a better rounding scheme? Even though the first question is more relevant for us, we start with the second one, which was answered by Feige and Schechtman [FS02]. They constructed an infinite graph G such that the gap between the maximum cut and the optimal solution of the SDP is exactly 0.8786.

Here is their infinite graph; corresponding finite examples can be obtained through discretization. The vertices are all vectors of unit norm in $\mathbb{R}^n$. The edges are given by the following random experiment: two random unit vectors x, y are chosen subject to the restriction $\langle x, y \rangle \leq \rho := \cos\theta$; most of them satisfy $\langle x, y \rangle \approx \cos\theta$. It is easy to see that the SDP has a feasible solution of value

$$\mathop{\mathbb{E}}_{x,y} \frac{1 - \langle x, y \rangle}{2} \approx \frac{1 - \cos\theta}{2}.$$

On the other hand, Feige and Schechtman use the method of symmetrization to show that the optimal cut is a hemisphere consisting, say, of all vectors x such that $x_1 \geq 0$. How many edges does a hemisphere cut? Given two random unit vectors x, y whose angle is roughly θ, the probability that $x_1 \geq 0 \geq y_1$ is roughly θ/π (consider their projection into two dimensions). The resulting gap matches the promise of the algorithm as $n \to \infty$.

While this shows that the gap 0.8786 is optimal, the integrality gap instance itself is actually solved optimally by the algorithm. This prompts the following question: does the algorithm ever produce a *bad* approximation?

9.2 Algorithmic gap

More relevant to us is an example due to Karloff [Kar00] in which the Goemans–Williamson algorithm is off by a factor of 0.8786. This time we present the discretized version of the construction. Our vertices are all points in $\{\pm 1/\sqrt{n}\}^n$ (all unit vectors). Edge weights are again determined by a random experiment, this time as follows: choose a vertex x at random, then choose $y \sim N_\rho(x)$, and add the edge (x, y).

As before, the graph comes with an embedding whose objective value is

$$\mathop{\mathbb{E}}_{x,y} \frac{1 - \langle x, y\rangle}{2} = \frac{1-\rho}{2}.$$

How did we get $\mathbb{E}[\langle x, y\rangle] = \rho$? If we identify vertices with ± 1 vectors, then the inner product of two vertices x, y (without normalization) corresponds to the usual inner product of the corresponding ± 1 vectors (with normalization), and we know that

$$\mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y \sim N_\rho(x)} [\langle x, y\rangle] = \mathop{\mathbb{E}}_{x} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_i y_i] = \rho.$$

This is in fact an optimal embedding, as we show below. Moreover, the central limit theorem shows that for most edges (x, y), $\langle x, y\rangle \approx \rho$. Therefore random hyperplane rounding results in a solution whose expected value is only roughly θ/π (per our earlier analysis, since most angles are very close to θ). In contrast, the cut given by $x_1 = 1/\sqrt{n}$ separates two vectors with probability $(1-\rho)/2$ (the probability that $x_i \neq y_i$), and so the resulting approximation ratio is only 0.8786.

It remains to show that the embedding we gave is optimal. Any embedding is given by n functions $F_1, \ldots, F_n \colon \{\pm 1/\sqrt{n}\}^n \to \mathbb{R}$ such that $\sum_{i=1}^n F_i(x)^2 = 1$ for all vertices x. For each coordinate i,

$$\mathbb{E}[F_i(x) F_i(y)] = \sum_S \rho^{|S|} \hat{F}_i(S)^2 \geq \rho \sum_S \hat{F}_i(S)^2 = \rho\, \mathbb{E}[F_i^2],$$

since $\rho < 0$. Therefore denoting the complete embedding by F, we have

$$\mathbb{E}[\langle F(x), F(y)\rangle] = \sum_{i=1}^{n} \mathbb{E}[F_i(x) F_i(y)] \geq \rho \sum_{i=1}^{n} \mathbb{E}[F_i^2] = \rho.$$

This shows that the embedding is indeed optimal.

Where does the suboptimality stem from? Another optimal solution embeds x at the point $(\operatorname{sgn} x_1, 0, \ldots, 0)$. This embedding has objective value

$$\mathop{\mathbb{E}}_{x,y} \frac{1 - x_1 y_1}{2} = \frac{1-\rho}{2}.$$

If we use random hyperplane rounding on this embedding, then we obtain the optimal solution: if the random hyperplane is w, then x gets into part $\operatorname{sgn}(w_1 \operatorname{sgn} x_1) = \operatorname{sgn} w_1 \operatorname{sgn} x_1$. The semidefinite program cannot distinguish between these two embeddings, and so it could choose the one that loses a factor of 0.8786 during rounding. Moreover, even if the semidefinite program finds the correct semidefinite matrix, when the algorithm uses the Cholesky decomposition to find an embedding, it encounters a degree of freedom which corresponds to a random rotation (since $FF^T = FQQ^TF^T$ for all orthogonal Q). This makes it even harder to find a good embedding.

9.3 Hardness of approximation result

We will use a different version of Unique Label Cover this time. For every $\delta > 0$, there is an alphabet R such that the following problem is UGC-hard. Given a left-regular bipartite graph (V, W, E) and permutations ψ_e labeling the edges, distinguish the following two cases:

YES instance There is a labeling satisfying a $1 - \delta$ fraction of the edges.

NO instance Every labeling satisfies at most a δ fraction of the edges.

We construct an instance of Max Cut as follows, where ρ is some parameter:

Vertices For each $w \in W$ and $F \subseteq R$ there is a vertex (w, F). We identify F with a function $R \to \{-1, 1\}$.

Edges Instead of describing a set of edges and weights, we show how to sample an edge. We pick a vertex $v \in V$ and two of its neighbors $w, w' \in W$ at random. We then pick $x \in \{-1,1\}^R$ at random, $y \sim N_\rho(x)$, and generate the edge $(w, \psi_{vw}^{-1}(x)), (w', \psi_{vw'}^{-1}(y))$.

As usual, it is easier to understand the construction by considering what happens on a YES instance. Given a good labeling L, we generate a cut by taking the function $f(w, F) = F(L(w))$. A random edge is satisfied by L with probability at least $1 - \delta$, and so by the union bound, both edges $(v, w), (v, w')$ are satisfied with probability at least $1 - 2\delta$ (each one is a random edge since the Label Cover graph is left-regular). In that case,

$$\begin{aligned} f(w, \psi_{vw}^{-1}(x)) &= \psi_{vw}^{-1}(x)(L(w)) = x(L(v)), \\ f(w', \psi_{vw'}^{-1}(y)) &= \psi_{vw'}^{-1}(y)(L(w')) = y(L(v)). \end{aligned}$$

Therefore the edge is cut with probability $\frac{1-\rho}{2}$. Overall, the value of the maximum cut is at least $(1 - 2\delta)\frac{1-\rho}{2}$. This suggests choosing $\rho = \cos(0.74\pi) < 0$. It might be alarming that ρ is negative, but later we will see that this is not so bad.

If we want to prove that the Goemans–Williamson algorithm is optimal, then we need to consider instances having a maximum cut of value at least $\cos^{-1}\rho/\pi + \epsilon$, for some small ϵ. Our aim is to show that these cannot be NO instances. Our starting point is finding a formula for the weight of edges cut by some function f:

$$\begin{aligned} w(f) &= \mathbb{E}\left[\frac{1}{2} - \frac{1}{2} f(w, \psi_{vw}^{-1}(x)) f(w', \psi_{vw'}^{-1}(y))\right] \\ &= \frac{1}{2} - \frac{1}{2} \mathop{\mathbb{E}}_{v,x,y} \left(\mathbb{E}[f(w, \psi_{vw}^{-1}(x))]\, \mathbb{E}[f(w', \psi_{vw'}^{-1}(y))]\right), \end{aligned}$$

since w, w' are chosen independently. Define $g_v(x) = \mathbb{E}_w[f(w, \psi_{vw}^{-1}(x))]$, where the expectation is over neighbors of v. Using this notation,

$$w(f) = \frac{1}{2} - \frac{1}{2} \mathop{\mathbb{E}}_{v,x,y}[g_v(x) g_v(y)] = \frac{1}{2} - \frac{1}{2} \mathop{\mathbb{E}}_{v}[\langle g_v, T_\rho g_v \rangle].$$

Since $w(f) \geq \cos^{-1}\rho/\pi + \epsilon$, we deduce that

$$\mathop{\mathbb{E}}_{v}[\langle g_v, T_\rho g_v \rangle] = 1 - 2w(f) \leq 1 - \frac{2}{\pi}\cos^{-1}\rho - 2\epsilon.$$

It follows that an ϵ-fraction of the vertices satisfies $\langle g_v, T_\rho g_v \rangle \leq 1 - (2/\pi)\cos^{-1}\rho - \epsilon$, since otherwise

$$\mathop{\mathbb{E}}_{v}[\langle g_v, T_\rho g_v \rangle] > (1-\epsilon)(1 - (2/\pi)\cos^{-1}\rho - \epsilon) \geq 1 - (2/\pi)\cos^{-1}\rho - 2\epsilon.$$

Call such a vertex *good*. We would like to apply the Majority is Stablest theorem (for $\{-1,1\}$-valued functions), but there is a slight problem, since $\rho < 0$; and moreover, we don't know anything about $\mathbb{E}[g_v]$.

To fix this, let h be the odd part of g_v, that is $h(x) = (g_v(x) - g_v(-x))/2$. Note that $\mathbb{E}[h] = 0$ and $\mathrm{Inf}_i[h] \leq \mathrm{Inf}_i[g_v]$ (by considering, for example, the spectral formula). Moreover,

$$\langle g_v, T_\rho g_v \rangle = \sum_S \rho^{|S|} \hat{g}_v(S)^2 \geq \sum_{|S| \text{ odd}} \rho^{|S|} \hat{g}_v(S)^2 = \langle h, T_\rho h \rangle = -\langle h, T_{-\rho} h \rangle.$$

Majority is Stablest states that if all influences of h are small, then $\langle h, T_{-\rho} h \rangle \lesssim 1 - \frac{2}{\pi}\cos^{-1}(-\rho)$ (this is the version for $\{-1,1\}$-valued functions). Since $\cos^{-1}(-\rho) = \pi - \cos^{-1}(\rho)$, altogether we get that if all influences of h are small then

$$\langle g_v, T_\rho g_v \rangle \gtrsim -(1 - \tfrac{2}{\pi}(\pi - \cos^{-1}\rho)) = 1 - \frac{2}{\pi}\cos^{-1}\rho.$$

We conclude that if v is good then g_v must have an influential variable i. In fact we need a bit more: that $\mathrm{Inf}_i[g_v^{\leq k}] \geq \tau$ for some constant k, τ (depending on ϵ, ρ). If we carefully go over our proof of Majority is Stablest, we see that because we applied a bit of noise in the beginning, we actually get this slightly stronger promise (this is essentially because the operator $T_{1-\delta}$ cuts $g_v^{>k}$ very sharply). Let us label v with this $i = L(v)$. Since $\mathrm{Inf}_i[g_v^{\leq k}] \geq \tau$, we have

$$\tau \leq \sum_{\substack{S \ni i \\ |S| \leq k}} \hat{g}_v(S)^2 = \sum_{\substack{S \ni i \\ |S| \leq k}} \mathop{\mathbb{E}}_w[\hat{f}_w(\psi_{vw}^{-1}(S))]^2,$$

where $f_w(x) = f(w, x)$. Convexity of t^2 implies that

$$\tau \leq \sum_{\substack{S \ni i \\ |S| \leq k}} \mathop{\mathbb{E}}_w[\hat{f}_w(\psi_{vw}^{-1}(S))^2] = \mathop{\mathbb{E}}_w[\mathrm{Inf}_{\psi_{vw}^{-1}(i)}[f_w^{\leq k}]].$$

Therefore at least a $\tau/2$-fraction of w satisfy $\mathrm{Inf}_{\psi_{vw}^{-1}(i)}[f_w^{\leq k}] \geq \tau/2$; call such a w good with repsect to v.

The reason we had to insist that $\mathrm{Inf}_i[g_v^{\leq k}] \geq \tau$ rather than just $\mathrm{Inf}_i[g_v] \geq \tau$ is that for every w, there are at most $2k/\tau$ elements satisfying $\mathrm{Inf}_j[f_w^{\leq k}] \geq \tau/2$. Let us put all of them in a set $C(w)$. We have shown above that a $\tau/2$ fraction of w satisfy $\psi_{vw}^{-1}(i) \in C(w)$. Therefore the multi-labeling $L(v), C(w)$ satisfies an $\epsilon(\tau/2)$-fraction of the constraints: if v is good and w is good (which happens with probability $\epsilon(\tau/2)^2$) then $L(v) \in \psi_{vw}(C(w))$. In order to construct an actual labeling, choose a label for w at random from $C(w)$. Such a labeling satisfies a fraction $\epsilon(\tau/2)(\tau/2k)$ of the constraints in expectation. By choosing the parameters carefully, we can make this quantity larger than δ, completing the proof.

9.4 Extensions

Raghavendra's theorem Raghavendra [Rag09] proved a far-reaching generalization of this result. Max Cut is a particular *constraint satisfaction problem* (CSP). A general CSP is given by a list of allowable constraints. In the case of Max Cut, there is only one constraint, $x \neq y$. In the case of MAX-3SAT, there are eight constraints, corresponding to the functions $x \vee y \vee z$, $x \vee y \vee \bar{z}$, and so on. How well can we approximate a CSP? Raghavendra constructed a canonical semidefinite relaxation and showed how to round it, and proved that the resulting approximation ratio is optimal assuming the unique games conjecture. His main idea was to convert any *integrality gap* (an instance in which the SDP produces a fractional solution much better than the integral optimum) to a PCP reduction.

Approximation resistance A particularly intriguing question here is for which predicates we can beat the random assignment algorithm. Håstad [Hås01] showed that it is NP-hard to approximate 3SAT better than 7/8 and 3LIN (in which disjunctions are replaced by XORs of triplets of variables) better than 1/2; and these approximations can be produced by choosing a random assignment (if you're averse to randomized algorithms, the method of conditional expectations can be used to derandomize these algorithms). A predicate for which this is true is called *approximation resistant*. Khot, Tulsiani and Worah [KTW14] characterize approximation resistant binary predicates assuming the unique games conjecture, though the characterization is not simple, and perhaps not even decidable.

10 Analysis on $\mathbb{Z}_r^n$

(Parts of [O'D14, Chapter 8] and [ADFS04].)

So far the functions we have been considering were on either the Boolean cube or on Gaussian space. Another common domain is a generalization of the Boolean cube, $\mathbb{Z}_r^n$. It will be convenient to think of $\mathbb{Z}_r$ as the set Ω_r of rth root of unity. We start by describing the Fourier expansion of functions $\Omega_r^n \to \mathbb{C}$, in a way completely analogous to our description of the case $r = 2$.

We start by showing, in two ways, that every function $\Omega_r^n \to \mathbb{C}$ has a unique representation as a polynomial in which the degree of each variable is at most $r - 1$; we say that such a polynomial has *individual degree* at most $r - 1$. As before, we will give two proofs: spatial and spectral. We start with the spatial proof. We first show that every function has some representation as a polynomial of individual degree at most $r - 1$, and then that this representation is unique.

The proof of the *existence* part is by induction on n. The case $n = 0$ is obvious. Suppose now that the result holds for some n, and consider a function $f\colon \Omega_r^{n+1} \to \mathbb{C}$. For $\omega \in \Omega_r$ let $f_\omega(x_1, \ldots, x_n) = f(x_1, \ldots, x_n, \omega)$. By induction, we can write each f_i as a polynomial P_i of individual degree at most $r - 1$. Define

$$P = \frac{1}{r} \sum_{i=0}^{r-1} x_{n+1}^i \sum_{\omega \in \Omega} \omega^{-i} P_\omega(x_1, \ldots, x_n).$$

For $\tau \in \Omega$ we have

$$P(x_1, \ldots, x_n, \tau) = \frac{1}{r} \sum_{i=0}^{r-1} \sum_{\omega \in \Omega} (\tau/\omega)^i f(x_1, \ldots, x_n, \omega).$$

Since $(\tau/\omega)^r = \tau^r/\omega^r = 1$, τ/ω is also an rth root of unity. Therefore if $\tau \neq \omega$ we have

$$\sum_{i=0}^{r-1} (\tau/\omega)^i = \frac{(\tau/\omega)^r - 1}{\tau/\omega - 1} = 0.$$

Conversely, when $\tau = \omega$, the same sum equals r. We conclude that P indeed represents f.

For *uniqueness*, as before it suffices to prove that the zero function has a unique representation has a polynomial of individual degree at most $r - 1$. The proof is by induction on n. The base case $n = 0$ is obvious. Suppose that the result holds for some n, and consider some polynomial P over $x_1, \ldots, x_{n+1}$ having individual degree at most $r - 1$ which represents zero. We can write $P = \sum_{i=0}^{r-1} x_{n+1}^i P_i$, where each P_i depends only on $x_1, \ldots, x_n$. For fixed $0 \leq j \leq r - 1$ we have

$$\frac{1}{r} \sum_{\omega \in \Omega} \omega^{-j} P(x_1, \ldots, x_n, \omega) = \frac{1}{r} \sum_{i=0}^{r-1} \sum_{\omega \in \Omega} \omega^{i-j} P_i(x_1, \ldots, x_n).$$

If $i \neq j$ then $\sum_{\omega \in \Omega} \omega^{i-j} = 0$ (why? one way to see that is by considering that Ω consists of all powers of some primitive root of unity), and otherwise the sum equals r. We conclude that

$$\frac{1}{r} \sum_{\omega \in \Omega} \omega^{-j} P(\cdot, \omega) = P_i.$$

Since P represents the zero function, the left-hand side is always zero, and so by induction, P_i is the zero polynomial. We conclude that P is the zero polynomial.

Fourier basis The more standard proof of this expansion is spectral. We define an inner product on Ω_r^n by $\langle f, g \rangle = \mathbb{E}[\bar{f} g]$, where $\bar{f}$ is complex conjugation. With each function $\sigma\colon [n] \to \{0, \ldots, r - 1\}$ we associate a monomial $\chi_\sigma = \prod_{i=1}^n x_i^{\sigma(i)}$. We claim that these monomials are orthogonal and have unit norm. Indeed, since $\bar{\omega} = \omega^{-1}$ for $\omega \in \Omega$, this follows from the calculation

$$\langle \chi_\sigma, \chi_\tau \rangle = \mathbb{E}\left[\prod_{i=1}^n x_i^{\tau(i)-\sigma(i)}\right] = \prod_{i=1}^n \frac{1}{r} \sum_{\omega \in \Omega} \omega^{\tau(i)-\sigma(i)}.$$

There are r^n monomials and the dimension of $\mathbb{C}[\Omega_r^n]$ is r^n, so the Fourier characters χ_σ must form an orthonormal basis. The corresponding Fourier expansion is

$$f = \sum_\sigma \hat{f}(\sigma)\chi_\sigma.$$

Orthonormality implies that $\langle f, \chi_\sigma \rangle = \hat{f}(\sigma)$. Parseval's identity now reads

$$\mathbb{E}[f^2] = \sum_\sigma |\hat{f}(\sigma)|^2.$$

Notice the absolute value.

Greenwell–Lovász As an application of this Fourier expansion, we prove a result of Greenwell and Lovász [GL74], follows the method of [ADFS04]. Suppose that we have a traffic light with r possible lights, which is controlled by n many r-way switches. We are given that whenever *all* switches change position, the light always changes. Greenwell and Lovász proved that the traffic light is actually controlled by a single switch! This is in fact a form of the Erdős–Ko–Rado theorem.

Let f be the indicator function for the set of switch settings giving rise to some specific color. We know that f is the indicator of an independent set in the graph K_r^n in which two vertices are connected if they agree on some coordinate. Let A be the adjacency matrix of the graph. Then $\langle f, Af \rangle = f^* Af = 0$. What is the effect of the operator A on the Fourier expansion? Consider some basis vector χ_σ. We have

$$\begin{aligned}
(A\chi_\sigma)(x_1, \ldots, x_n) &= \sum_{y_1 \neq x_1} \cdots \sum_{y_n \neq x_n} \chi_\sigma(y_1, \ldots, y_n) \\
&= \sum_{y_1 \neq x_1} y_1^{\sigma(1)} \cdots \sum_{y_n \neq x_n} y_n^{\sigma(n)} \\
&= \prod_{\sigma(i)=0} (r-1) \times \prod_{\sigma(i)\neq 0} (-x_i^{\sigma_i}) \\
&= (r-1)^{n-|\sigma|}(-1)^{|\sigma|}\chi_\sigma,
\end{aligned}$$

where $|\sigma|$ is the Hamming weight of σ, that is, the number of entries different from zero. Substituting this in the identity $\langle f, Af \rangle = 0$, we obtain

$$\sum_\sigma |\hat{f}(\sigma)|^2 (r-1)^{n-|\sigma|}(-1)^{|\sigma|} = 0.$$

We also know that $\sum_\sigma |\hat{f}(\sigma)|^2 = \|f\|^2 = \mathbb{E}[f]$ and that $\hat{f}(0) = \langle f, 1 \rangle = \mathbb{E}[f]$. Therefore

$$\begin{aligned}
0 &= (r-1)^n \,\mathbb{E}[f]^2 + \sum_{\sigma \neq 0} |\hat{f}(\sigma)|^2 (r-1)^{n-|\sigma|}(-1)^{|\sigma|} \\
&\geq (r-1)^n \,\mathbb{E}[f]^2 - (r-1)^{n-1} \sum_{\sigma \neq 0} |\hat{f}(\sigma)|^2 \\
&\geq (r-1)^n \,\mathbb{E}[f]^2 - (r-1)^{n-1}(\mathbb{E}[f] - \mathbb{E}[f]^2).
\end{aligned}$$

We conclude that $1 - \mathbb{E}[f] \geq (r-1)\,\mathbb{E}[f]$ or $\mathbb{E}[f] \leq 1/r$.

Considering the indicator functions f_ω for all possible colors $\omega \in \Omega$, the upper bound $\mathbb{E}[f_\omega] \leq 1/r$ implies that $\mathbb{E}[f_\omega] = 1/r$ for all $\omega \in \Omega$. Equality is only possible if the only non-zero Fourier coefficients are those whose support has size at most 1. We thus deduce that

$$f_\omega(x_1, \ldots, x_n) = \sum_{i=1}^n f_{\omega,i}(x_i).$$

(This representation isn't unique.) The fact that f_ω is Boolean forces at most one $f_{\omega,i}$ to be non-constant. In other words, f_ω depends on a single coordinate i_ω. Since $\mathbb{E}[f_\omega] = 1/r$, there is some τ_ω such that $f_\omega(x) = [x_{i_\omega} = \tau_\omega]$. It is not hard to check that since the traffic light can have at most one light at a time, the indices i_ω must be the same for all $\omega \in \Omega$. This completes the proof.

Influence We can define influence and noise in this case as well (though we don't present any applications). We start with influence.

Let $L_i f(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) - \frac{1}{r}\sum_{y_i} f(\ldots, y_i, \ldots)$. In order to understand the effect of L_i on the Fourier expansion, we compute $L_i x_i^d$ when $d \neq 0$ (clearly $L_i 1 = 0$):

$$L_i x_i^d = x_i^d - \frac{1}{r}\sum_{y_i} y_i^d = x_i^d.$$

This shows that

$$L_i f = \sum_{\sigma(i)\neq 0} \hat{f}(\sigma)\chi_\sigma.$$

This prompts the definition

$$\mathrm{Inf}_i[f] = \|L_i f\|^2 = \sum_{\sigma(i)\neq 0} |\hat{f}(\sigma)|^2.$$

The corresponding spatial definition is

$$\mathrm{Inf}_i[f] = \mathbb{E}[(f(x) - f(y))^2],$$

where x is a random point and y is a random neighbor, where x is considered a neighbor of itself. Alternatively,

$$\mathrm{Inf}_i[f] = \mathop{\mathbb{E}}_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} \mathop{\mathbb{E}}_{x_i}[(f(\ldots, x_i, \ldots) - \mathop{\mathbb{E}}_{y_i} f(\ldots, y_i, \ldots))^2] = \mathop{\mathbb{E}}_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} \mathop{\mathbb{V}}_{x_i}[f].$$

When f is 0/1-valued, $\mathrm{Inf}_i[f]$ is the probability that f changes when we randomize its ith coordinate.

We can define the total influence as before $\mathrm{Inf}[f] = \sum_i \mathrm{Inf}_i[f]$, and we get the nice formula

$$\mathrm{Inf}[f] = \sum_\sigma |\sigma||\hat{f}(\sigma)|^2.$$

This prompts us define the degree-d part of f by

$$f^{=d} = \sum_{|\sigma|=d} \hat{f}(\sigma)\chi_\sigma.$$

We then have $\mathrm{Inf}[f] = \sum_d d\|f^{=d}\|^2$. This is a special case of the Efron–Stein decomposition (which also has many other names), see [O'D14, Section 8.3].

Noise We proceed to define the noise operator, for $\rho \in [0, 1]$. Given a point $x \in \Omega_r^n$, let $y_i = x_i$ with probability ρ, and y_i is chosen randomly from Ω_r otherwise. We define $(T_\rho f)(x) = \mathbb{E}[f(y)]$. Clearly $T_\rho 1 = 1$. The effect of T_ρ on x_i^d for $d \neq 0$ is

$$T_\rho x_i^d = \rho x_i^d + (1-\rho)\mathop{\mathbb{E}}_{\omega\in\Omega_r}[\omega^d] = \rho x_i^d.$$

Therefore

$$T_\rho f = \sum_\sigma \rho^{|\sigma|}\hat{f}(\sigma)\chi_\sigma = \sum_d \rho^d f^{=d}.$$

Our proof of Friedgut–Kalai–Naor used the hypercontractive inequality. Alon et al. [ADFS04] generalize this proof to this setting, and as a result obtain a stability version of the Erdős–Ko–Rado theorem in this setting. So there is at least one application of these generalizations!

Efron–Stein decomposition There is a decomposition which is in between the Fourier expansion itself and the very rough decomposition into degree d parts, which is a special case of the Efron–Stein decomposition. For each subset $S \subseteq [n]$ we define $f^{\subseteq J}$ as f averaged over coordinates not in J, and $f^{=S}$ by

$$f^{=S} = \sum_{J\subseteq S} (-1)^{|S|-|J|} f^{\subseteq J}.$$

(This formula is a result of *Möbius inversion* on the formula $f^{\subseteq S} = \sum_{J \subseteq S} f^{=J}$.) What does this give in our case? There is an obvious guess, and to verify it we first need to compute $f^{\subseteq J}$. At this point, a routine calculation gives

$$f^{\subseteq J} = \sum_{\sigma\colon \operatorname{supp}\sigma \subseteq J} \hat{f}(\sigma)\chi_\sigma,$$

where $\operatorname{supp}(\sigma) = \{i : \sigma(i) \neq 0\}$. Therefore

$$f^{=S} = \sum_{J \subseteq S} (-1)^{|S|-|J|} \sum_{\sigma\colon \operatorname{supp}\sigma \subseteq J} \hat{f}(\sigma)\chi_\sigma = \sum_{\sigma\colon \operatorname{supp}\sigma = S} \hat{f}(\sigma)\chi_\sigma,$$

using Möbius inversion again. In particular, this shows that

$$f = \sum_{S \subseteq [n]} f^{=S}$$

is an orthogonal decomposition, another generalization of the usual Fourier transform.

The formula used to define $f^{=S}$ makes sense in more general situations, and one can prove that the resulting decomposition is always orthogonal (see [O'D14, Theorem 8.35]). This decomposition is variously named Hoeffding decomposition, Efron–Stein decompositions, ANOVA decomposition, or simply orthogonal decomposition.

11 Roth's theorem for $\mathbb{Z}_3^n$

(Ryan O'Donnell's lecture notes [O'D07, Lecture 27].)

The classical theorem of van der Waerden in Ramsey theory states that if the integers are colored using finitely many colors, then there are arbitrarily long monochromatic arithmetic progressions. In fact, for every number of colors c and any length k there exists a number n such that any c-coloring of $[n]$ contains a k-term arithmetic progression. The corresponding "density" version of this theorem would state that for any c, k there exists a number n such that any subset of $[n]$ of density $1/c$ contains a k-term arithmetic progression. This clearly implies van der Waerden's theorem.

The density version of van der Waerden's theorem was proven by Roth using Fourier analysis on $\mathbb{Z}_n$. Let $r_k(n)$ be the minimum density such that any subset of $[n]$ of this density must contain a k-term arithmetic progression. (A simple averaging argument due to Varnavides shows that any set of density $(1+\epsilon)r_k(n)$ must contain many k-term arithmetic progressions.) The density version of van der Waerden's theorem implies that $r_3(n) \to 0$. How fast does $r_3(n)$ tend to 0? Behrend gave a construction showing that $r_3(n) = \Omega(n/e^{-\sqrt{\log n}})$, and the best upper bound, due to Sanders, is $r_3(n) = O(n(\log\log n)^5/\log n)$. Erdős and Turán conjectured that any sequence (a_i) satisfying $\sum_i 1/a_1 = \infty$ contains arbitrarily large arithmetic progressions.

If we could improve Sanders' bound to $O(n/\log n)$, then it would follow that the primes, whose density is $n/\log n$, contain infinitely many 3-term arithmetic progressions. (In fact, $r_3(n) = O(n\log\log n/\log n)$ would suffice, as Gowers [Gow13] observes.) Van der Corput proved unconditionally that the primes contain infinitely many 3-term arithmetic progressions.

Roth proved his result using Fourier analysis (Croot and Sisask later showed how to modify the argument to avoid Fourier analysis), but his proof doesn't immediately generalize to longer arithmetic progressions. Szemerédi extended Roth's theorem to k-term arithmetic progressions for any k, using his regularity lemma. Furstenberg and Katznelson reproved the result (without explicit density bounds) using ergodic theory. Finally, Gowers was able to generalize Roth's original proof using higher-order Fourier analysis. Green and Tao were able to transfer these results to the primes, thus showing that the primes (and in fact any subset of primes with positive density) contain arbitrarily long arithmetic progressions.

Roth's theorem follows from a similar result for the group $\mathbb{Z}_n$ (a subset of $[n]$ can be embedded in $\mathbb{Z}_{3n}$ so that a 3-term arithmetic progression in the latter corresponds to one in the former). More generally, arithmetic progressions make sense for any Abelian group G, where an arithmetic progression is a sequence of the form $x, x+d, x+2d, \ldots, x+(k-1)d$. For k-term arithmetic progressions it is prudent to demand that no non-zero element satisfies $(k-1)d = 0$. In particular, for $k = 3$ we need G to have odd order. Meshulam [Mes95] proved an analog of Roth's theorem for finite Abelian groups of odd order and high *rank* (number of factors in a maximal decomposition); his proof is much simpler and less technical than Roth's.

We will concentrate on the case $G = \mathbb{Z}_3^n$. Our goal would be to show that for some constant C, if $A \subset \mathbb{Z}_3^n$ contains at least $C3^n/n$ elements then A contains a 3-term arithmetic progression. We follow the exposition in Ryan O'Donnell's lecture notes.

Let f be the characteristic function of A, and let $\mu = \mathbb{E}[f] \geq C/n$. We can represent an arithmetic progression $x, x+d, x+2d$ in a simpler form: letting $y = x+d$, the progression becomes $x, y, 2x-y = -x-y$. If we were to choose A randomly by putting in each element with probability μ, then for random x, y, the probability that $x, y, -x-y \in A$ is μ^3. Hence if A is "random-looking", we expect A to have many 3-term arithmetic progressions. How do we quantify the "randomness" of A? It turns out that from the point of view of 3-term arithmetic progressions, A is random unless it has a prominent Fourier coefficient. We get this by explicitly counting the number of 3-term arithmetic progressions, repeating calculations that we had encountered earlier while analyzing the BLR linearity test.

The number of 3-term arithmetic progressions is the number of pairs $x \neq y$ satisfying $x, y, -x-y \in A$.

In order to get an expression for this, we first ignore the constraint $x \neq y$, and calculate:

$$\begin{aligned}\Pr[x, y, -x-y \in A] &= \mathbb{E}[f(x)f(y)f(-x-y)] \\ &= \mathop{\mathbb{E}}_{x,y} \sum_{\sigma,\tau,\upsilon} \hat{f}(\sigma)\hat{f}(\tau)\hat{f}(\upsilon)\chi_\sigma(x)\chi_\tau(y)\chi_\upsilon(-x-y) \\ &= \mathop{\mathbb{E}}_{x,y} \sum_{\sigma,\tau,\upsilon} \hat{f}(\sigma)\hat{f}(\tau)\hat{f}(\upsilon)\chi_\sigma(x)\overline{\chi_\upsilon(x)}\chi_\tau(y)\overline{\chi_\upsilon(y)} \\ &= \sum_\sigma \hat{f}(\sigma)^3 \\ &= \mu^3 + \sum_{\sigma \neq 0} \hat{f}(\sigma)^3.\end{aligned}$$

(In the calculation we used orthogonality of characters, together with the fact that $\chi_\sigma(-x) = \overline{\chi_\sigma(x)}$, which follows from $\omega^{-1} = \overline{\omega}$ for any root of unity ω.) We call such a formula a *counting formula*.

The actual number of 3-term arithmetic progressions is thus

$$9^n \Pr[x, y, -x-y \in A] - |A| = 9^n \left(\mu^3 - 3^{-n}\mu + \sum_{\sigma \neq 0} \hat{f}(\sigma)^3 \right).$$

We want to find a condition for the right-hand side to be positive, and to this end we proceed to estimate the error term:

$$\begin{aligned}\left| \sum_{\sigma \neq 0} \hat{f}(\sigma)^3 \right| &\leq \sum_{\sigma \neq 0} |\hat{f}(\sigma)|^3 \\ &\leq \sum_{\sigma \neq 0} |\hat{f}(\sigma)|^2 \max_{\sigma \neq 0} |\hat{f}(\sigma)| \\ &= (\mu - \mu^2) \max_{\sigma \neq 0} |\hat{f}(\sigma)|.\end{aligned}$$

Therefore if all non-constant Fourier coefficients are small compared to μ^2, then A contains 3-term arithmetic progressions, just by counting. This suggests measuring the randomness of A in terms of the largest magnitude of a non-constant Fourier coefficient. We say that A is *ϵ-pseudorandom* if $|\hat{f}(\sigma)| \leq \epsilon$ for all $\sigma \neq 0$ (recall that f is the characteristic function of A). If A is μ^2-pseudorandom then the number of 3-term arithmetic progressions is at least

$$9^n \left(\mu^3 - 3^{-n}\mu - (\mu - \mu^2)\mu^2 \right) = 9^n \mu \left(\mu^4 - 3^{-n}\mu \right),$$

which is positive as long as $\mu^3 > 3^{-n}$.

What if A is not μ^2-pseudorandom? In that case f must correlate noticeably with some Fourier character χ_σ, and this gives A some *structure*; this is known as the *structure versus pseudorandomness* paradigm, first appearing in Szemerédi's work. The idea now is to find a substructure H of $\mathbb{Z}_3^n$ inside which A has larger density. If $A \cap H$ is pseudorandom, then $A \cap H$ contains a 3-term arithmetic progression. Otherwise, we find a substructure of H inside which A has an even larger density, and so on. This process cannot go on forever, since the density of A keeps increasing. This kind of argument is known as a *density increase* argument, and it first appears in the same work of Szemerédi.

What kind of structure is implied by $|\hat{f}(\sigma)|$ being large? It is natural to partition $\mathbb{Z}_3^n$ into three subsets, according to the value of χ_σ. Notice that

$$\hat{f}(\sigma) = 3^{-n} \sum_x f(x)\chi_\sigma(x) = 3^{-n} \sum_{\omega \in \Omega} \omega \sum_{\chi_\sigma(x) = \omega} f(x) = \mathop{\mathbb{E}}_{\omega \in \Omega} \mathbb{E}[f(x) | \chi_\sigma(x) = \omega].$$

(Here $\Omega = \Omega_3$ consists of all third roots of unity.) In particular, if $|\hat{f}(\sigma)| \geq \mu^2$ then $|\mathbb{E}[f(x)|\chi_\sigma(x) = \omega]| \geq \mu^2$ for some $\omega \in \Omega$, which implies that the density of f is at least μ^2 on $\{x : \chi_\sigma(x) = \omega\}$. Since $\mu^2 < \mu$, this is not really what we wanted.

To fix this, we look at $g = f - \mu$ instead; since $\sigma \neq 0$, $|\hat{g}(\sigma)| = |\hat{f}(\sigma)| \geq \mu^2$. Let $\delta_\omega = \mathbb{E}[g(x)|\chi_\sigma(x) = \omega]$ and $\mu_\omega = \mathbb{E}[f(x)|\chi_\sigma(x) = \omega]$, and note that $\mu_\omega = \mu + \delta_\omega$. We thus want to show that one of the δ_ω is positive and large. The preceding argument shows that $|\delta_\omega| \geq \mu^2$ for some ω, but we get no guarantee that $\delta_\omega > 0$. To fix this, note that $\mathbb{E}[\delta_\omega] = \mathbb{E}[g] = 0$, and so

$$\mu^2 \leq |\hat{g}(\sigma)| \leq \mathop{\mathbb{E}}_{\omega\in\Omega}[|\delta_\omega|] = \mathop{\mathbb{E}}_{\omega\in\Omega}[|\delta_\omega| + \delta_\omega] = 2 \mathop{\mathbb{E}}_{\omega\in\Omega}[\delta_\omega 1_{\delta_\omega > 0}].$$

It follows that $\delta_\omega \geq \mu^2/2$ for some $\omega \in \Omega$.

Concluding, for some $\omega \in \Omega$ it holds that the density of A inside $H = \{x \in \Omega^n : \chi_\sigma(x) = \omega\}$ is at least $\mu + \mu^2/2$. We can think of H as a subset of $\mathbb{Z}_3^n$ obtained as the set of solutions to the single equation of the form $\sum_i a_i x_i = b$, and then we can write H in the form $H = T\mathbb{Z}_3^{n-1} + z$, where T is a full rank linear transformation. It is then natural to consider $B = T^{-1}(A \cap H - z)$, which is a subset of $\mathbb{Z}_3^{n-1}$ of density at least $\mu + \mu^2/2$. An arithmetic progression $x, y, -x - y$ in B translates to the sequence $Tx + z, Ty + z, T(-x - y) + z$ in A, which is an arithmetic progression since the sum of the outer terms is double the inner term: $Tx + z + T(-x - y) + z = -T(y) + 2z = 2(T(y) + z)$. Thus if B contains a 3-term arithmetic progression, so does A.

Summarizing our work so far, we have shown that if A is a subset of $\mathbb{Z}_3^n$ of measure $\mu > 3^{-n/3}$ without a 3-term arithmetic progression, then there is a subset $B \subseteq \mathbb{Z}_3^{n-1}$ of measure $\mu + \mu^2/2$ without a 3-term arithmetic progression. Continuing this way $2/\mu$ times, we have doubled the density; and so continuing this way $4/\mu$ times, we have made the density infinite, which is absurd. This implies that after some $i < 4/\mu$ steps, we have reached a set of measure $\mu_i \leq 3^{-(n-i)/3}$. This definitly cannot happen if $\mu > 3^{-(n-4/\mu)/3}$. If $\mu > C/n$ then $3^{-(n-4/\mu)/3} < 3^{-(1-4/C)(n/3)}$, and so for any $C > 4$ and large enough n, $\mu > C/n$ implies that $\mu > 3^{-(n-4/\mu)/3}$ and the density increment argument works. We have shown that any subset of $\mathbb{Z}_3^n$ of density $\Omega(3^n/n)$ contains a 3-term arithmetic progression.

We can try to replicate the same proof for longer arithmetic progressions. We will find out that the pseudorandomness condition isn't good enough for the counting lemma to imply that the set has 4-term arithmetic progressions. Gowers found stronger pseudorandomness conditions given by his *(Gowers) uniformity norms*, and this led to higher-order Fourier analysis.

12 Reed–Muller codes

(Kumar and Pfister [KP15].)

One of the central problems in information theory is that of channel capacity. We want to transfer data through a channel, and the question is how to overcome noise in the channel. Two classical channels are the *binary symmetric channel* and the *binary erasure channel*. In the binary symmetric channel $BSC(p)$, data is sent in as bits. Independently, each bit is flipped with probability p. The capacity of this channel (we explain later what this means) is $1 - h(p)$, where $h(p) = -p \log_2 p - (1-p)\log_2(1-p)$ is the entropy function. In the binary erasure channel $BEC(p)$, data is also sent in bits. Independently, each bit is *erased* (changed to $*$) with probability p. The capacity of this channel is $1 - p$. (The binary *deletion* channel is the superficially similar channel in which each bit is deleted with probability p; we don't see the positions of the deleted bits, rather they are just absent from the output stream. The capacity of this channel is still unknown.)

Suppose that we want to send n bits of information through the binary erasure channel. The number of bits coming out is (usually) roughly $(1-p)n$, and so we don't expect to be able to reliably send more than that many bits of data. Moreover, with some small probability p^n *all* bits are erased, so we cannot guarantee being able to send *any* amount of information. The definition of capacity is thus a bit subtle.

We start by defining codes. A *code* is a collection of 2^k binary vectors of some length n. The *rate* of the code is k/n. We use the code in the following way. We translate a message m of length k bits into a codeword X. We send X through the channel to get Y. Now we run some decoding procedure to get a codeword Z. If $X = Z$ then we say that the decoding was successful. The error probability of a decoding procedure is the probability that decoding was successful given a uniformly random message.

For a given channel, a rate r is *achievable* if there is a sequence of codes C_n whose rate tends to r which can be decoded with error probability tending to zero. The *capacity* of a channel is the maximum rate achievable. Shannon gave a formula for the rate of a channel, and this formula implies the claimed capacities of BSC and BEC. The codes that Shannon uses to prove his theorem are random, and so cannot be used in practice. One of the main aims of coding theory is finding codes with good parameters for which encoding and decoding is efficient. We say that a sequence of codes C_n is *capacity-achieving* for a channel if their rates tend to the capacity of the channel and they can be decoded with error probability tending to zero. We are thus looking for capacity-achieving codes which have low encoding and decoding complexity.

While capacity-achieving families of codes are known for both BEC and BSC, some of the common families of codes in standard use are not known to achieve capacity. One of these standard codefamilies is Reed–Muller codes, which we describe later on. Recently, a simple connection between this question and sharp threshold theorems has been discovered by two groups, Kudekar et al. [KMEU15] and Kumar and Pfister [KP15]. Both groups analyzed Reed–Muller codes with respect to BEC, showing that in some sense they are capacity-achieving. The result obtained by the second group is stronger, and to obtain their stronger result they have to use a deep sharp threshold theorem of Bourgain and Kalai [BK97]. It is their work which we follow here.

Let $C \subseteq \{0,1\}^n$ be a *linear code* of dimension k. This just means that C is a vector subspace of $\{0,1\}^n$ (under addition modulo 2). Since C has dimension k, it has 2^k codewords and so its rate is $r = k/n$. One particularly simple decoding procedure for this code under the BEC is as follows. Denoting by X the input to the channel and Y the output of the channel, if $Y_i \neq *$ then we know that $X_i = Y_i$. Even if $Y_i = *$ then it might be the case that all codewords consistent with Y have the same value for the ith coordinate, and again we recover X_i. Otherwise we declare failure. We will analyze the success probability of this decoding procedure with respect to the binary erasure channel in which the probability of erasing the ith bit is p_i.

Let X be a uniformly random codeword of C, and let Y be the corresponding output of the binary erasure channel with erasure probabilities $\mathbf{p} = (p_1, \ldots, p_n)$. Denote by $\alpha_i(\mathbf{p})$ the probability that the decoding procedure outlined above fails in decoding the ith bit, and by $\beta_i(\mathbf{p})$ the same probability under the assumption that $Y_i = *$. Clearly $\alpha_i(\mathbf{p}) = p_i \beta_i(\mathbf{p})$, and $\beta_i(\mathbf{p})$ is independent of p_i.

Given Y, the set of codewords which could map to Y form an affine subspace $V(Y)$. We define $D(\mathbf{p}) = \mathbb{E}[\dim V(Y)]$. (The original paper defines $D(\mathbf{p}) = H(X|Y)$, where H is entropy, but in our case we can avoid the concept of entropy.) How does $D(\mathbf{p})$ depend on p_i? Let $Y^{[i]}$ result from setting the ith bit to $*$. If $Y_i = *$ then $V(Y) = V(Y^{[i]})$. Otherwise, either $V(Y) = V(Y^{[i]})$ or $\dim V(Y) = \dim V(Y^{[i]}) - 1$,

the latter case happening with probability $\beta_i(\mathbf{p})$. Therefore $D(\mathbf{p}) = \mathbb{E}[\dim V(Y^{[i]})] - (1-p_i)\beta_i(\mathbf{p})$. Since the first term doesn't depend on p_i, we conclude that

$$\frac{\partial D(\mathbf{p})}{\partial p_i} = \beta_i(\mathbf{p}).$$

Why is this useful? Let $D(p)$ be the value of D at the constant p vector. The formula implies that

$$\frac{dD(p)}{dp} = \sum_i \frac{\partial D(p)}{\partial p_i} = \sum_{i=1}^n \beta_i(p).$$

Since clearly $D(1) = k$ whereas $D(0) = 0$, integration gives the so-called *area formula*

$$\int_0^1 \frac{1}{n} \sum_{i=1}^n \beta_i(p)\, dp = \frac{k}{n}.$$

Let $\beta(p) = (1/n)\sum_i \beta_i(p)$ (in fact, all our codes will be symmetric and so $\beta(p) = \beta_i(p)$ for all i.) A simple coupling argument shows that β is increasing. If β had a sharp threshold then the area formula implies that this threshold must occur close to k/n, and so as long as $p < k/n$, $\beta(p) \approx 0$, that is, the *average* bit is decoded correctly with high probability. If $\beta(p) = o(1/n)$ then the entire codeword is decoded correctly with high probability.

In order to analyze $\beta_i(p)$, let us find an alternative expression for it. Let $S = \{i : Y_i = *\} \cup \{i\}$. It is not hard to check that the ith bit cannot be decoded iff S contains the support of some codeword x which furthermore satisfies $x_i = 1$. If we let $C_i = \{x \in C : x_i = 1\}$ and $U_i = \{A \setminus i : A \supset \operatorname{supp} x \text{ for some } x \in C_i\}$, then $\beta_i(p) = \mu_p^{[n]\setminus i}(U_i)$. Since the set U_i is monotone, the Russo–Margulis lemma shows that

$$\frac{d\beta_i(p)}{dp} = \frac{\mathrm{Inf}^{(p)}[f]}{p(1-p)}.$$

If each U_i is invariant under some transitive permutation group (this happens, for example, when C is invariant under some 2-transitive permutation group), then all influences are equal, and so the KKL theorem shows that

$$\frac{d\beta_i(p)}{dp} = \Omega(\beta_i(p)(1-\beta_i(p))\log(n-1)) = \Omega(\beta_i(p)(1-\beta_i(p))\log n).$$

If $\beta_i(p_0) = \epsilon$ and $\beta_i(p_1) = 1-\epsilon$, then this implies that $1 - 2\epsilon = \Omega((p_1 - p_0)\epsilon(1-\epsilon)\log n)$, and so

$$\beta_i^{-1}(1-\epsilon) - \beta^{-1}(\epsilon) = O\left(\frac{1}{\epsilon \log n}\right).$$

When the code is invariant under some transitive permutation group, we can replace β_i by β. The area formula implies that

$$\frac{k}{n} \geq (1-\epsilon)(1-\beta^{-1}(1-\epsilon)) \geq (1-\epsilon)(1-\beta^{-1}(\epsilon) - O(1/\epsilon \log n)).$$

Rearranging gives

$$1 - \beta^{-1}(\epsilon) \leq (1-\epsilon)^{-1}\frac{k}{n} + O\left(\frac{1}{\epsilon \log n}\right) = \frac{k}{n} + O\left(\epsilon + \frac{1}{\epsilon \log n}\right).$$

Stated differently,

$$\beta\left(1 - \frac{k}{n} - \Omega\left(\epsilon + \frac{1}{\epsilon \log n}\right)\right) \leq \epsilon.$$

In other words, if a code satisfies the requisite conditions (say it is invariant under some 2-transitive permutation group), then the probability that decoding fails for a fixed bit at erasure probability p is at most ϵ as long as the rate of the code is $1 - p - \Omega(\epsilon + 1/\epsilon \log n)$. In other words, such a code is capacity-achieving with respect to decoding individual bits.

In order to show that a code is capacity-achieving, we need to show that $n\beta(1-k/n-\delta) \to 0$ rather than just $\beta(1-k/n) \to 0$. To this end, let us start by analyzing the implications of KKL more carefully. For $p \leq \beta^{-1}(1/2)$, we can rewrite the guarantee of KKL as $\beta'(p) = \Omega(\beta(p)\log n)$. The solution to $B'(p) = CB(p)$ is $B(p) = Ae^{Cp}$, and so for $p_0 \leq p_1 \leq \beta^{-1}(1/2)$ we actually have $\beta(p_1)/\beta(p_0) \geq \exp\Omega((p_1-p_0)\log n)$. In particular, $(1/2)/\epsilon \geq \exp\Omega((\beta^{-1}(1/2)-\beta^{-1}(\epsilon))\log n)$, or

$$\beta^{-1}(1/2) - \beta^{-1}(\epsilon) = O\left(\frac{\log(1/2\epsilon)}{\log n}\right).$$

A similar estimate holds for $\beta^{-1}(1-\epsilon) - \beta^{-1}(1/2)$, and we conclude that

$$\beta^{-1}(1-\epsilon) - \beta^{-1}(\epsilon) = O\left(\frac{\log(1-\epsilon)/\epsilon}{\log n}\right).$$

Repeating the preceding calculations, we get

$$\beta\left(1 - \frac{k}{n} - \Omega\left(\epsilon + \frac{\log(1/\epsilon)}{\log n}\right)\right) \leq \epsilon,$$

and so the probability that decoding fails for a particular bit at erasure probability p is at most ϵ if the rate of the code is

$$1 - p - \Omega\left(\epsilon + \frac{\log(1/\epsilon)}{\log n}\right).$$

Our goal is to achieve $\epsilon = o(1/n)$ (so that all bits are decoded correctly with high probability) while keeping the rate error term $o(1)$. This is impossible using our current bounds.

Suppose, however, that we could strengthen KKL and show that

$$\frac{d\beta(p)}{dp} = \Omega(\beta(p)(1-\beta(p))\log n \cdot \gamma(n))$$

for some function $\gamma(n) = \omega(1)$. The same calculations as before show that we only need the rate to be

$$1 - p - \Omega\left(\epsilon + \frac{\log(1/\epsilon)}{\log n \cdot \gamma(n)}\right).$$

In particular, choosing $\epsilon = e^{-\log(n)\sqrt{\gamma(n)}} = o(1/n)$, the rate is $1 - p - o(1)$.

Reed–Muller codes At this point we can describe Reed–Muller codes. Let $\mathbb{F} = GF(q)$ (we will be interested in the case $q = 2$). For parameters m, d, the Reed–Muller code $RM(m,d)$ is the graphs of all polynomials of degree at most d on $\mathbb{F}^m$ (when $d = 1$ we get Reed–Solomon codes). Thus $n = q^m$ and $q^k = \sum_{D \leq d} \binom{m}{D}(q-1)^D$. This is a linear code, and as m tends to infinity, we can find values of d that will ensure that the rate is roughly r for any r of our choosing. If f is a polynomial of degree at most d then so is $x \mapsto f(Ax+b)$, where A is any invertible linear transformation; this can be seen through the Fourier expansion of f. This shows that the code is invariant under the action of the affine group, which is 2-transitive. It is thus eligible for our calculations. A deep result of Bourgain and Kalai [BK97] strengthens KKL to $\gamma(n) = \log\log n$ (this result applies to any linear-invariant code, that is, a code invariant under the group of linear transformations), and allows us to conclude that these codes achieve capacity.

What about the binary symmetric channel? It is conjectured that Reed–Muller codes achieve capacity in this case as well, with a different simple decoder, the MAP or ML decoder (both are the same in this case), which decodes each bit separately according to which value of that bit was more likely to result in Y assuming that all other bits were chosen at random. Perhaps the proof above can be generalized to cover this case as well.

References

[ADFS04] Noga Alon, Irit Dinur, Ehud Friedgut, and Benny Sudakov. Graph products, Fourier analysis and spectral techniques. *Geometric and functional analysis*, 14(5):913–940, 2004.

[AK97] Rudolf Ahlswede and Levon H. Khachatrian. The complete intersection theorem for systems of finite sets. *European Journal of Combinatorics*, 18(2):125–136, 1997.

[AK99] Rudolf Ahlswede and Levon H. Khachatrian. A pushing-pulling method: New proofs of intersection theorems. *Combinatorica*, 19(1):1–15, 1999.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998.

[BCH+96] Mihir Bellare, Don Coppersmith, Johan Håstad, Marcos A. Kiwi, and Madhu Sudan. Linearity testing in characteristic two. *IEEE Transactions on Information Theory*, 42:1781–1795, 1996.

[BFL90] László Babai, Lance Fortnow, and Carsten Lund. Nondeterministic exponential time has two-prover interactive protocols. In *SFCS '90: Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, pages 16–25, 1990.

[BK97] Jean Bourgain and Gil Kalai. Influences of variables and threshold intervals under group symmetries. *Geometric and Functional Analysis*, 7(3):438–361, 1997.

[BLR93] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences*, 47(3):549–595, 1993.

[Bor75] Christer Borell. The Brunn–Minkowski inequality in Gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975.

[CW01] Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

[DDG+15] Roee David, Irit Dinur, Elazar Goldenberg, Guy Kindler, and Igor Shinkar. Direct sum testing. In *The 6th Innovations in Theoretical Computer Science (ITCS) conference*, 2015.

[Din07] Irit Dinur. The PCP theorem by gap amplification. 54(3):12, 2007.

[DMN13] Anindya De, Elchanan Mossel, and Joe Neeman. Majority is stablest: discrete and SoS. In *STOC '13*, pages 477–486, 2013.

[DS05] Irit Dinur and Samuel Safra. On the hardness of approximating minimum vertex cover. *Annals of Mathematics*, 162(1):439–485, 2005.

[EKR61] Paul Erdős, Chao Ko, and Richard Rado. Intersection theorems for systems of finite sets. *Quarterly Journal of Mathematics*, 12(1):313–320, 1961.

[Eld14] Ronen Eldan. A two-sided estimate for the Gaussian noise stability deficit. *Inventiones Mathematicae*, 2014.

[FHKL] Yuval Filmus, Hamed Hatami, Nathan Keller, and Noam Lifshitz. Bounds on the sum of L1 influences. *Israel Journal of Matematics*.

[Fil13] Yuval Filmus. *Spectral methods in extremal combinatorics*. PhD thesis, University of Toronto, 2013.

[FK96] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the American Mathematical Society*, 124:2993–2002, 1996.

[FKN02] Ehud Friedgut, Gil Kalai, and Assaf Naor. Boolean functions whose Fourier transform is concentrated on the first two levels. *Advances in Applied Mathematics*, 29(3):427–437, 2002.

[Fri98] Ehud Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–36, 1998.

[Fri99] Ehud Friedgut. Sharp threshold of graph properties, and the k-SAT problem. *Journal of the American Mathematical Society*, 12(4):63–70, 1999.

[Fri08] Ehud Friedgut. On the measure of intersecting families, uniqueness and stability. *Combinatorica*, 28(5):503–528, 2008.

[FS02] Uriel Feige and Gideon Schechtman. On the optimality of the random hyperplane rounding technique for Max Cut. *Random Structures and Algorithms*, 20(3):403–440, 2002.

[GL74] Donald L. Greenwell and László Lovász. Applications of product colouring. *Acta Mathematica Academiae Scientiarum Hungarica*, 25(3):335–340, 1974.

[GO08] Anupam Gupta and Ryan O'Donnell. Advanced approximation algorithms. Online lecture notes, Spring 2008.

[Gow13] W. Timothy Gowers. Erdős and arithmetic progressions. *Bolyai society mathematical studies*, 25:265–287, 2013.

[GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001.

[Hat12] Hamed Hatami. A structure theorem for Boolean functions with small total influences. *Annals of Mathematics*, 176(1):509–533, 2012.

[Kar00] Howard Karloff. How good is the Goemans–Williamson Max Cut algorithm? *SIAM Journal on Computing*, 29(1):336–350, 2000.

[Kat72] Gyula O. H. Katona. A simple proof of the Erdős–Chao Ko–Rado theorem. *Journal of Combinatorial Theory, Series B*, 13(2):183–184, 1972.

[Kho02] Subhash Khot. On the power of unique 2-prover 1-round games. In *STOC '02*, pages 767–775, 2002.

[Kin02] Guy Kindler. *Property testing, PCP, and juntas*. PhD thesis, Tel-Aviv University, 2002.

[KKL88] Jeff Kahn, Gil Kalai, and Nati Linial. The influence of variables on Boolean functions. In *29th Annual Symposium on Foundations of Computer Science (FOCS 1988)*, pages 68–80, 1988.

[KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for MAXCUT and other 2variable CSPs? *SIAM Journal on Computing*, 37(1):319–257, 2007.

[KMEU15] Shrinivas Kudekar, Marco Mondelli, Soğlu Eren Sa and Rüdiger Urbanke. Reed–Muller codes achieve capacity on the binary erasure channel under MAP decoding. *Preprint*, 2015.

[KP15] Santhosh Kumar and Henry D. Pfister. Reed-Muller codes achieve capacity on erasure channels. *Preprint*, 2015.

[KR08] Subhash Khot and Oded Regev. Vertex cover might be hard to approximate to within $2-\epsilon$. *Journal of Computer and System Sciences*, 74(3):335–349, 2008.

[KS04] Guy Kindler and Shmuel Safra. Noise-resistant Boolean functions are juntas, 2004. Unpublished manuscript.

[KTW14] Subhash Khot, Madhur Tulsiani, and Pratik Worah. A characterization of strong approximation resistance. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC '14)*, pages 634–643, 2014.

[KV15] Subhash Khot and Nisheeth K. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative-type metrics into ℓ_1. *Journal of the ACM*, 62(1):8, 2015.

[Mes95] Roy Meshulam. *Journal of Combinatorial Theory, Series A*, 71(1):168–172, 1995.

[MN14] Elchanan Mossel and Joe Neeman. Robust optimality of Gaussian noise stability. *Journal of the European Mathematical Society*, 2014.

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. *Annals of Mathematics*, 171:295–341, 2010.

[NS94] Noam Nisan and Mario Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994.

[O'D07] Ryan O'Donnell. Lecture notes for analysis of boolean functions, 2007.

[O'D14] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

[Rag09] Prasad Raghavendra. *Approximating NP-hard Problems: Efficient Algorithms and their Limits*. PhD thesis, University of California, Berkeley, 2009.

[Raz98] Ran Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):768–803, 1998.

[RS10] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *STOC '10*, pages 755–764, 2010.

[Wil84] Richard M. Wilson. The exact bound in the Erdős–Ko–Rado theorem. *Combinatorica*, 4(2–3):247–257, 1984.