Introduction to Boolean Analysis (ECOPRO)

Yuval Filmus

August 2025

1 Introduction

Boolean function analysis studies functions on the Boolean cube $\{-1,1\}^n$ (or sometimes $\{0,1\}^n$) from a spectral perspective. Often the functions themselves are also Boolean. It has applications in many areas, such as combinatorics, social choice theory, probability theory, and theoretical computer science.

The starting point is the following observation:

Theorem 1. *Every function $f\colon \{-1,1\}^n \to \mathbb{R}$ can be written as a multilinear polynomial, in a unique way:*

$$f = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S, \text{ where } \chi_S(x) = \prod_{i \in S} x_i.$$

Stated differently, the functions χ_S (called *Fourier characters*) constitute a basis of the space of all functions from $\{-1,1\}^n$ to $\mathbb{R}$. In fact, more can be shown: the characters χ_S constitute an orthonormal basis with respect to the inner product

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

where the expectation is with respect to the uniform distribution.

Since there are 2^n different characters, matching the dimension of the space of all real-valued functions on $\{-1,1\}^n$, it suffices to show that different characters are orthogonal, and each has norm 1. The second property follows immediately from the observation $\chi_S(x)\chi_S(x) = 1$. As for the first property, since $\chi_S(x)\chi_T(x) = \chi_{S \triangle T}(x)$, it suffices to show that $\mathbb{E}[\chi_S] = 0$ whenever $S \neq \emptyset$. Indeed,

$$\mathop{\mathbb{E}}_x[\chi_S(x)] = \mathop{\mathbb{E}}_{x_1,\ldots,x_n \in \{-1,1\}}\left[\prod_{i \in S} x_i\right] = \prod_{i \in S} \mathop{\mathbb{E}}_{x_i \in \{-1,1\}}[x_i] = \prod_{i \in S} 0 = 0.$$

Some terminology:

- The functions χ_S are called (Fourier) basis functions or (Fourier) characters.
- The numbers $\hat{f}(S)$ are called Fourier coefficients.
- The entire formula is the Fourier expansion.
- The *degree* of f is the degree of its Fourier expansion as a polynomial.

Since the Fourier characters are orthonormal, we can extract them by taking an inner product:

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

In particular, since $\chi_\emptyset \equiv 1$,

$$\hat{f}(\emptyset) = \mathop{\mathbb{E}}_x[f(x)].$$

Another useful property is Parseval's identity:

$$\langle f, g\rangle = \sum_S \hat{f}(S)\hat{g}(S).$$

In particular, taking $f = g$,

$$\|f\|^2 = \sum_S \hat{f}(S)^2,$$

where $\|f\|$ is the L_2 norm, that is, $\|f\| = \sqrt{\langle f, f\rangle}$.

1.1 Linearity testing

A function $f\colon \{0,1\}^n \to \{0,1\}$ is *linear* if it satisfies

$$f(x \oplus y) = f(x) \oplus f(y) \text{ for all } x, y \in \{0,1\}^n.$$

If we think of $\{0,1\}$ as $\mathbb{Z}_2$, then this is the usual definition of linearity. It is easy to check that if f is linear then f has the form

$$f(x) = \bigoplus_{i \in S} x_i$$

for some $S \subseteq [n]$.

What can we say about f if it satisfies the following property (where x, y are drawn uniformly at random)?

$$\Pr_{x,y \sim \{0,1\}^n}[f(x \oplus y) = f(x) \oplus f(y)] \geq 1 - \epsilon$$

The natural guess is that f has to be close to a truly linear function. This has a very short proof using Fourier analysis.

First, let us switch from $\{0,1\}$ to $\{-1,1\}$, using the mapping $b \mapsto (-1)^b$. Linearity is now the property

$$f(xy) = f(x)f(y) \text{ for all } x, y \in \{-1,1\}^n.$$

The functions satisfying this property are the Fourier characters χ_S.

Suppose that a function $f\colon \{-1,1\}^n \to \{-1,1\}$ satisfies

$$\Pr_{x,y \sim \{-1,1\}^n}[f(xy) = f(x)f(y)] \geq 1 - \epsilon.$$

The first step is to convert the probability into an expectation:

$$\mathop{\mathbb{E}}_{x,y}[f(x)f(y)f(xy)] = \Pr_{x,y}[f(xy) = f(x)f(y)]\cdot 1 + \Pr_{x,y}[f(xy) \neq f(x)f(y)]\cdot(-1) = 2\Pr_{x,y}[f(xy) = f(x)f(y)] - 1 \geq 1 - 2\epsilon.$$

Next, we substitute the Fourier expansion of f:

$$1 - 2\epsilon \leq \mathop{\mathbb{E}}_{x,y}[f(x)f(y)f(xy)] = \mathop{\mathbb{E}}_{x,y}\left[\left(\sum_S \hat{f}(S)\chi_S(x)\right)\left(\sum_T \hat{f}(T)\chi_T(y)\right)\left(\sum_R \hat{f}(R)\chi_R(xy)\right)\right].$$

Using linearity of expectation:

$$1 - 2\epsilon \leq \sum_{S,T,R} \hat{f}(S)\hat{f}(T)\hat{f}(R) \mathop{\mathbb{E}}_{x,y}[\chi_S(x)\chi_T(y)\chi_R(xy)].$$

Since $\chi_R(xy) = \chi_R(x)\chi_R(y)$, we can rewrite this as

$$1 - 2\epsilon \leq \sum_{S,T,R} \hat{f}(S)\hat{f}(T)\hat{f}(R) \mathop{\mathbb{E}}_{x}[\chi_S(x)\chi_R(x)] \mathop{\mathbb{E}}_{y}[\chi_T(y)\chi_R(y)] = \sum_{S,T,R} \hat{f}(S)\hat{f}(T)\hat{f}(R)\langle \chi_S, \chi_R\rangle\langle \chi_T, \chi_R\rangle.$$

If $S \neq R$ or $T \neq R$ then one of the inner products vanishes. Therefore the only summands which remain correspond to $S = T = R$, leading to

$$1 - 2\epsilon \leq \sum_S \hat{f}(S)^3.$$

Recall that $\sum_S \hat{f}(S)^2 = \|f\|^2 = \mathbb{E}_x[f(x)^2] = 1$. Intuitively, the only way to reconcile this with $\sum_S \hat{f}(S)^3 \gtrsim 1$ is if one of the Fourier coefficients is large. Mathematically,

$$1 - 2\epsilon \leq \sum_S \hat{f}(S)^2 \cdot \hat{f}(S) \leq \sum_S \hat{f}(S)^2 \max_T \hat{f}(T) = \max_T \hat{f}(T) \sum_S \hat{f}(S)^2 = \max_T \hat{f}(T).$$

Therefore there exists T such that $\hat{f}(T) \geq 1 - 2\epsilon$. We would like to show that f agrees with χ_T on most inputs. Indeed,

$$1 - 2\epsilon \leq \hat{f}(T) = \langle f, \chi_T \rangle = \mathop{\mathbb{E}}_x[f(x)\chi_T(x)] = 2 \Pr_x[f(x) = \chi_T(x)] - 1,$$

and so

$$\Pr_x[f(x) = \chi_T(x)] \geq 1 - \epsilon.$$

Summarizing:

Theorem 2 (Linearity testing)**.** *Suppose that* $f\colon \{-1,1\}^n \to \{-1,1\}$ *satisfies*

$$\Pr_{x,y}[f(xy) = f(x)f(y)] \geq 1 - \epsilon.$$

Then there exists a character χ_S *such that*

$$\Pr_x[f(x) = \chi_S(x)] \geq 1 - \epsilon.$$

This result has a property testing interpretation. Suppose that you are given a function $f\colon \{-1,1\}^n \to \{-1,1\}$ as a black-box. Someone claims that f is linear, and you would like to test this by probing f at only a few (randomly chosen) places. Since you are only probing f at a few places, no testing procedure can guarantee that f is linear, but you could hope that f is close to a linear function, in the sense that f agrees with some linear function up to some small error set which is hard to hit using only a few applications of f.

Theorem 2 gives us exactly such a test: choose x, y at random, and verify that $f(xy) = f(x)f(y)$. This test has two properties:

Completeness If f is linear then the test always passes.

Soundness If the test passes w.p. $1 - \epsilon$, then f is ϵ-close to a linear function, meaning that there is a linear function g such that $\Pr_x[f(x) \neq g(x)] \leq \epsilon$.

A random function passes the test with probability 1/2. Hence we can think of the test as a pseudo-randomness measure. In particular, if the probability that the test passes is far away from 1/2, then the function is "far from random" — but in which sense?

Theorem 2 turns out to solve this question as well. Suppose first that

$$\Pr_{x,y}[f(xy) = f(x)f(y)] \geq \frac{1}{2} + \delta.$$

Choosing $\epsilon = 1 - (1/2 + \delta) = 1/2 - \delta$, the theorem implies that there exists a character χ_S such that

$$\Pr_x[f(x) = \chi_S] \geq \frac{1}{2} + \delta.$$

In other words, f *correlates* non-trivially with some character.

When the probability is too small, that is

$$\Pr_{x,y}[f(xy) = f(x)f(y)] \leq \frac{1}{2} - \delta,$$

we apply the same analysis to $g(x) = -f(x)$. The function g satisfies

$$\Pr_{x,y}[g(xy) = g(x)g(y)] = \Pr_{x,y}[f(xy) \neq f(x)f(y)] \geq \frac{1}{2} + \delta,$$

and so g correlates with some character χ_S, implying that

$$\Pr_x[f(x) = -\chi_S(x)] = \Pr_x[g(x) = \chi_S(x)] \geq \frac{1}{2} + \delta.$$

The only known proof of Theorem 2 in the regime $\epsilon \approx 1/2$ is using Fourier analysis (when ϵ is small, self-correction also works).

2 Influences

There are n different people trying to toss a single coin together. Each of them chooses $x_i \in \{-1, 1\}$ uniformly at random, and the result is aggregated using a function $f\colon \{-1,1\}^n \to \{-1,1\}$, which should be balanced: $\Pr_x[f(x) = 1] = 1/2$. How many people do I need to bribe in order to increase the probability of 1 to 2/3? We would like to find a function f which maximizes this number.

A natural choice for f is the Majority function (assuming n is odd). The central limit theorem shows that with probability 2/3, the number of 1 inputs is at least $n/2 - C\sqrt{n}$ for some constant $C > 0$. This means that if we know the input, it suffices to bribe $O(\sqrt{n})$ people. Now, this assumption is quite unrealistic. However, we can still bias the voting by bribing $O(\sqrt{n})$ people using a similar argument: the central limit theorem shows that with probability 2/3, the number of 1 inputs along the first $n - 2C\sqrt{n}$ people is at least $(n - 2C\sqrt{n})/2 - C\sqrt{n - 2C\sqrt{n}} \geq n/2 - 2C\sqrt{n}$. Bribing the remaining $2C\sqrt{n}$ people brings the total to at least $n/2$.

Are there better functions? Ajtai and Linial constructed a function which requires bribing $\Omega(\frac{n}{\log^2 n})$ people. We do not know whether this is optimal. However, Kahn, Kalai and Linial shows that it always suffices to bribe $O(\frac{n}{\log n})$ people, using their celebrated KKL theorem.

The KKL theorem is usually stated in terms of *influence*. Given a function $f\colon \{-1,1\}^n \to \{-1,1\}$, the i'th influence is

$$\mathrm{Inf}_i[f] = \Pr_x[f(x) \neq f(x^{\oplus i})],$$

where $x^{\oplus i}$ is obtained from x by flipping the i'th coordinate. In other words, the i'th influence is the probability that the i'th coordinate influences the output.

Theorem 3 (KKL)**.** *Every function $f\colon \{-1,1\}^n \to \{-1,1\}$ has a variable whose influence is*

$$\Omega\left(\frac{\log n}{n}\,\mathbb{V}[f]\right).$$

The variance has to make an appearance here: for example, if f is constant then all variables have zero influence.

The KKL theorem has two implications:

- For any balanced f, a briber who sees x can change $O(\frac{n}{\log n})$ entries so that with probability 2/3 (over x), the output is 1.
- For monotone balanced f, a briber can set $O(\frac{n}{\log n})$ *fixed* inputs to 1 in such a way that with probability 2/3, the output is 1.

The proof is a repeated application of the KKL theorem, where each additional coordinate increases the probability of 1 by roughly $\frac{\log n}{n}$. Since we stop once $\Pr[f = 1]$ exceeds $2/3$, the variance is always $\Omega(1)$.

While we don't know whether $\frac{n}{\log n}$ is tight for the original problem, the KKL theorem itself is tight, for the Tribes function:

$$\mathrm{Tribes}(x) = \bigvee_{i \in [n/m]} \bigwedge_{j \in [m]} x_{i,j},$$

for an appropriate choice of m. Here we momentarily switch back to $0, 1$.

We think of each i as a tribe consisting of m people. The function equals 1 if there is a tribe which "votes" 1 unanimously. This happens with probability

$$1 - (1 - 2^{-m})^{n/m} \approx 1 - \exp\left(-\frac{n}{2^m m}\right).$$

If $\frac{n}{2^m m} \approx \ln 2$ then the function will be roughly balanced. This happens for $m \approx \log_2 n - \log_2 \log_2 n$.

All variables have the same influence. The variable $x_{i,j}$ is influential if (i) all other tribes do not vote 1 unanimously, and (ii) the rest of tribe i votes 1. Event (i) happens with probability $1-(1-2^{-m})^{n/m-1} \approx 1/2$, and event (ii) with probability $2^{1-m} = 2/2^m = \Theta(m/n) = \Theta(\frac{\log n}{n})$, since $\frac{n}{2^m m} \approx \ln 2$.

2.1 Fourier formula

The influences have nice formulas in terms of the Fourier expansion. Define

$$L_i f(x) = \frac{f(x) - f(x^{\oplus i})}{2}.$$

This is a function from $\{-1, 1\}^n$ to $\{-1, 0, 1\}$, and

$$\mathrm{Inf}_i[f] = \Pr[L_i f \neq 0] = \|L_i f\|^2.$$

Substituting the Fourier expansion of f,

$$L_i f(x) = \sum_S \hat{f}(S) \frac{\chi_S(x) - \chi_S(x^{\oplus i})}{2}.$$

If $i \notin S$ then $\chi_S(x) = \chi_S(x^{\oplus i})$. If $i \in S$ then $\chi_S(x^{\oplus i}) = -\chi_S(x)$, and so

$$L_i f(x) = \sum_{i \in S} \hat{f}(S) \chi_S(x).$$

Consequently,

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

2.2 Total influence

The sum of all influences is called the *total influence*:

$$\mathrm{Inf}[f] = \sum_{i=1}^{n} \mathrm{Inf}_i[f].$$

This is a very natural quantity: $\frac{\mathrm{Inf}[f]}{n}$ is the probability that $f(x) \neq f(y)$, where (x, y) is a random edge in the Boolean cube $\{-1, 1\}^n$.

In terms of the Fourier expansion, we have

$$\mathrm{Inf}[f] = \sum_{i=1}^{n} \sum_{i \in S} \hat{f}(S)^2 = \sum_S |S| \hat{f}(S)^2.$$

Since

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

we can deduce the *Poincaré inequality*

$$\mathbb{V}[f] \leq \mathrm{Inf}[f].$$

This implies that the random walk on the Boolean cube has a spectral gap of $1/n$.

2.3 Proving the KKL theorem

We can now describe the proof of the KKL theorem, up to a certain lemma which will require the introduction of one more tool.

Recall that the KKL theorem states that given a function $f\colon \{-1,1\}^n \to \{-1,1\}$, there is always a variable whose influence is $\Omega(\frac{\log n}{n} \mathbb{V}[f])$. To simplify notation, from now on we assume that $\mathbb{V}[f] = 1$ (equivalently, f is balanced), but the general case is almost the same.

The KKL theorem is obvious if the *total* influence is large enough: if the total influence is at least $c \log n$, then there surely exists a variable whose influence is at least $c\frac{\log n}{n}$. Therefore we can assume that the total influence is small: at most $c \log n$.

The total influence measures how much the Fourier expansion is concentrated on the low levels. Indeed, recall that $\sum_S \hat{f}(S)^2 = 1$. This means that we can construct a probability distribution on subsets of $[n]$ in which the probability to sample S is $\hat{f}(S)^2$. The total influence is just $\mathbb{E}[|S|]$. Markov's inequality shows that

$$\Pr[|S| \geq 2\,\mathrm{Inf}[S]] \leq \frac{1}{2},$$

and so

$$\sum_{|S| \geq 2\,\mathrm{Inf}[S]} \hat{f}(S)^2 \leq \frac{1}{2} \Longrightarrow \sum_{|S| \leq 2\,\mathrm{Inf}[S]} \hat{f}(S)^2 \geq \frac{1}{2}.$$

How is this helpful? We can write

$$\frac{1}{2} \leq \sum_{|S| \leq 2\,\mathrm{Inf}[S]} \hat{f}(S)^2 \leq \sum_{i=1}^{n} \sum_{\substack{|S| \leq 2\,\mathrm{Inf}[S] \\ i \in S}} \hat{f}(S)^2 = \sum_{i=1}^{n} \|(L_i f)^{\leq 2\,\mathrm{Inf}[f]}\|^2,$$

where $g^{\leq d} = \sum_{|S| \leq d} \hat{f}(S)\chi_S$ is the low-degree part of g. Here we used the fact that $\hat{f}(\emptyset) = 0$, since f is balanced.

At this point we invoke a *level-d inequality*, stating that for every function $g\colon \{-1,1\}^n \to \{-1,0,1\}$, we have

$$\|g^{\leq d}\|^2 \leq 3^d \|g\|^3.$$

What this roughly means is that if $\|g\|$ is small then most of the Fourier mass of g is concentrated on high degrees (since the total mass is $\|g\|^2$, and $\|g\|^3 = \|g\|^2 \cdot \|g\|$). This is the case for *subcubes*. For example, the indicator function of $x_1 = \cdots = x_m = 1$ can be written as

$$\prod_{i=1}^{m} \frac{1 + x_i}{2},$$

and so its Fourier expansion has $\hat{g}(S) = 2^{-m}$ for all $S \subseteq [m]$. Here $\|g\|^2 = \mathbb{E}[g^2] = \mathbb{E}[g] = 2^{-m}$, and $\|g^{\leq d}\|^2 = \binom{m}{\leq d} 2^m$, which is small unless $d \gtrsim m/2$.

The level d inequality can be quantitatively improved, but the version we use has a simpler proof and will suffice for us. Taking $g = L_i f$ and $d = 2\operatorname{Inf}[S] \le 2c\log n$, we obtain

$$\frac{1}{2} \le 3^d \sum_{i=1}^{n} \|L_i f\|^3 = n^{c\log 9} \sum_{i=1}^{n} \operatorname{Inf}_i[f]^{3/2} \le n^{c\log 9} \cdot \operatorname{Inf}[f] \cdot \sqrt{\max_i \operatorname{Inf}_i[f]} \le n^{c\log 9} \cdot c\log n \cdot \sqrt{\max_i \operatorname{Inf}_i[f]}.$$

Therefore one of the influences is at least $n^{-2c\log 9}/4c^2\log^2 n$, which for an appropriate choice of c is $\Omega(\frac{1}{\sqrt{n}})$, say, completing the proof of the KKL theorem (assuming the level d inequality).

Some intuition If the total influence is $\Omega(\log n)$, then obviously one of the influences has to be $\Omega(\frac{\log n}{n})$. Otherwise, the total influence is $O(\log n)$, which means that the Fourier mass of the function is concentrated on degree up to $O(\log n)$. This implies that for the average i, at least $\Omega(\frac{1}{\log n})$ of the Fourier mass is concentrated up to degree $O(\log n)$. This means that for the average i, the i'th influence cannot be too small, since otherwise almost all of the mass will be beyond degree $O(\log n)$. Making this argument quantitative gives a bound of $\Omega(\frac{1}{\sqrt{n}})$ (say) on the maximal influence.

The proof of the KKL theorem might seem quite lossy. For example, the upper bound

$$\sum_{|S|\le 2\operatorname{Inf}[S]} \hat{f}(S)^2 \le \sum_{|S|\le 2\operatorname{Inf}[S]} |S|\hat{f}(S)^2 = \sum_{i=1}^{n} \|(L_i f)^{\le 2\operatorname{Inf}[f]}\|^2$$

is potentially loose. However, since $\mathbb{E}[|S|] = O(\log n)$, the loss is 'only' $O(\log n)$. Also, whereas we get a lower bound of $\Omega(\frac{\log n}{n})$ on the maximal influence when the total influence is large, we got a much better lower bound $\Omega(\frac{1}{\sqrt{n}})$ on the maximal influence when the total influence is small. In addition, the level-d inequality that we use can be substantially improved.

In contrast, we know that the KKL theorem is tight up to constant factors. This means that optimizing the proof by tightening the loose parts can only improve the result by a constant factor. Typically we do not care about such factors in Boolean function analysis.

2.4 Level-d inequality

In order to complete the proof of the KKL theorem, it remains to prove the following statement.

Lemma 4 (Weak level-d inequality)**.** *For all $g\colon \{-1,1\}^n \to \{-1,0,1\}$,*

$$\|g^{\le d}\|^2 \le 3^d \|g\|^3.$$

The proof will follow from the following concentration bound.

Lemma 5 (Concentration of degree d functions)**.** *Let $h\colon \{-1,1\}^n \to \mathbb{R}$ have degree at most d (this means that the degree of its Fourier expansion, considered as a polynomial, is at most d). Then*

$$\|h\|_4 \le \sqrt{3}^d \|h\|_2, \text{ where } \|h\|_p = \mathbb{E}_x[|h(x)|^p]^{1/p} \text{ is the } L_p \text{ norm of } h.$$

Generally speaking, the L_4-norm of a function can be much larger than its L_2-norm. As an extreme example, if h is the δ function of a point (that is, $h(x) = 1$ for some $x \in \{-1,1\}^n$, and $h(y) = 0$ otherwise) then

$$\|h\|_4 = 2^{-n/4} \text{ whereas } \|h\|_2 = 2^{-n/2},$$

and so $\|h\|_4$ is much larger than $\|h\|_2$; in particular, we cannot bound $\|h\|_4 \le C\|h\|_2$ for any C which is independent of n. Lemma 5 states such a bound in which C depends only on the degree of h.

Before proving Lemma 5, let us see how it implies Lemma 4:

$$\|g^{\le d}\|^2 \overset{(1)}{=} \langle g^{\le d}, g\rangle \overset{\text{Hölder}}{\le} \|g^{\le d}\|_4 \|g\|_{4/3} \overset{\text{Lem 5}}{\le} \sqrt{3}^d \|g^{\le d}\| \|g\|_{4/3} \overset{(2)}{=} \sqrt{3}^d \|g^{\le d}\| \|g\|^{3/2}.$$

Canceling out $\|g^{\leq d}\|$ and squaring the inequality, we deduce Lemma 4.

The calculation used two simple identities. Equality (1) follows from Parseval's identity:

$$\langle g^{\leq d}, g\rangle = \sum_S \widehat{g^{\leq d}}(S)\hat{g}(S) = \sum_{|S|\leq d} \hat{g}(S)^2 = \|g^{\leq d}\|^2,$$

since $\widehat{g^{\leq d}}(S) = 0$ unless $|S| \leq d$.

Equality (2) holds since $|g(x)|^{4/3} = |g(x)|^2$ for all x, itself a consequence of $g(x) \in \{-1, 0, 1\}$:

$$\|g\|_{4/3} = \mathop{\mathbb{E}}_x[|g(x)|^{4/3}]^{3/4} = \mathop{\mathbb{E}}_x[|g(x)|^2]^{(1/2)\cdot(3/2)} = \|g\|_2^{3/2}.$$

Strong level-d inequality Chin Ho Lee's level-d inequality for $[0, 1]$-valued function gives the bound

$$\|g^{=d}\|^2 \leq 4\alpha^2(2e\ln(e/\alpha^{1/d}))^d, \text{ where } \alpha = \mathbb{E}[g] = \|g\|^2.$$

In other words, $\|g\|^3$ can be improved to roughly $\|g\|^4$, up to a poly-logarithmic factor.

Interpretation as a concentration inequality Suppose that $f\colon \{\pm 1\}^n \to \mathbb{R}$ is centered: $\mathbb{E}[f] = 0$. Chebyshev's inequality shows that

$$\Pr_x[|f(x)| \geq C\|f\|] \leq \frac{1}{C^2}.$$

Lemma 5 implies a stronger concentration inequality in terms of the degree $d = \deg f$:

$$\Pr_x[|f(x)| \geq C\|f\|] = \Pr_x[f(x)^4 \geq C^4\|f\|^4] \overset{\text{Markov}}{\leq} \frac{\mathbb{E}[f^4]}{C^4\|f\|^4} = \frac{\|f\|_4^4}{C^4\|f\|^4} \leq \frac{9^d}{\|f\|^4}.$$

Using higher norms (and a corresponding generalization of Lemma 5), one can get an even better concentration inequality, showing that f has exponentially small tails:

$$\Pr_x[|f(x)| \geq C\|f\|] \leq \exp\left(-\frac{d}{2e}C^{2/d}\right) \text{ for all } C \geq (2e)^{d/2}.$$

3 Hypercontractivity

We move on to the proof of Lemma 5. This proof employs *hypercontractivity*, a technical tool which is behind many results in the area.

The statement of hypercontractivity involves the *noise operator* T_ρ. This is an operator, parametrized by a real number ρ, which takes a function $f\colon \{-1, 1\}^n \to \mathbb{R}$ and outputs another function $T_\rho f\colon \{-1, 1\}^n \to \mathbb{R}$ which is a smoothed version of f.

We define T_ρ via a distribution $N_\rho(x)$ on $\{-1, 1\}^n$, which is defined for every $x \in \{-1, 1\}^n$ and $\rho \in [0, 1]$. We sample $y \sim N_\rho(x)$ by sampling each coordinate of y independently according to the following law:

$$y_i = \begin{cases} x_i & \text{w.p. } \frac{1+\rho}{2}, \\ -x_i & \text{w.p. } \frac{1-\rho}{2}. \end{cases}$$

In words, $N_\rho(x)$ is obtained by flipping each coordinate with probability $\frac{1-\rho}{2}$. Taking some extreme examples, $N_1(x) = x$ always, and $N_0(x)$ is a uniform sample from $\{-1, 1\}^n$.

Here is an equivalent definition, which will be useful for later generalization:

$$y_i = \begin{cases} x_i & \text{w.p. } \rho, \\ \text{random} & \text{w.p. } 1-\rho. \end{cases}$$

To see that these definitions are equivalent, observe that $\Pr[y_i = -x_i] = (1-\rho)\cdot\frac{1}{2} = \frac{1-\rho}{2}$, since $y_i = -x_i$ can only happen if we obtain y_i by "resampling" rather than by copying x_i. Similarly, $\Pr[y_i = x_i] = \rho + (1-\rho)\cdot\frac{1}{2} = \frac{1+\rho}{2}$.

The operator T_ρ is defined by smoothing f according to N_ρ:

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

In words, $T_\rho f(x)$ is obtained by averaging f over a "soft neighborhood" of x. Extreme examples include $T_1 f = f$ (the identity operator) and $T_0 f = \mathbb{E}[f]$ (the operator replacing f by the constant function $\mathbb{E}[f]$).

Since $T_\rho f$ is an averaging operator, for every $p \geq 1$ we have

$$\|T_\rho f\|_p \leq \|f\|_p.$$

To see this, notice that we can sample $y \sim N_\rho(x)$ by sampling $z \sim N_\rho(\mathbf{1})$ and taking $y = zx$. This shows that

$$T_\rho f = \mathop{\mathbb{E}}_{z\sim N_\rho(\mathbf{1})}[f^z], \text{ where } f^z(x) = f(zx).$$

The functions f^z all have the same norms (since if x is a uniform sample from $\{-1,1\}^n$ then so is zx), and so $\|T_\rho f\|_p \leq \|f\|_p$ follows from the triangle inequality, which holds for all $p \geq 1$.

Hypercontractivity (in the special case of the L_4 and L_2 norms) states a stronger inequality:

Lemma 6 (L_4-L_2 hypercontractivity)**.** *For every function $f\colon \{-1,1\}^n \to \mathbb{R}$,*

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

Intuitively, after averaging f, the resulting function is somewhat concentrated, and so its L_4-norm can be bounded. Similar inequalities hold for any two L_p-norms (with appropriate noise rates ρ), but this version has a particularly simple proof.

Before proving Lemma 6, let us see how it implies Lemma 5. For this, we need to compute the Fourier expression for $T_\rho f$. Since T_ρ is linear, it suffices to compute $T_\rho \chi_S$:

$$(T_\rho \chi_S)(x) = \mathop{\mathbb{E}}_{y\sim N_\rho(x)}\left[\prod_{i\in S} y_i\right] = \prod_{i\in S}\mathop{\mathbb{E}}_{y\sim N_\rho(x)}[y_i] = \prod_{i\in S}\left(\frac{1+\rho}{2}x_i + \frac{1-\rho}{2}(-x_i)\right) = \prod_{i\in S}(\rho x_i) = \rho^{|S|}\chi_S(x).$$

Therefore

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

An alternative argument uses the other definition of N_ρ:

$$(T_\rho \chi_S)(x) = \prod_{i\in S}\mathop{\mathbb{E}}_{y\sim N_\rho(x)}[y_i] = \prod_{i\in S}(\rho\cdot x_i + (1-\rho)\cdot \mathop{\mathbb{E}}_{z_i\sim\text{random}}[z_i]) = \prod_{i\in S}(\rho x_i) = \rho^{|S|}\chi_S(x),$$

since the expected value of a random sample out of $\{\pm 1\}$ is 0.

When $\rho \in (0,1)$, the operator T_ρ has the effect of discounting the high-degree parts of f, acting as a kind of "low-pass filter".

The Fourier expression for $T_\rho f$ makes sense for any value of ρ. In particular, the inverse operator of T_ρ is $T_{\rho^{-1}}$ (since $\rho^{-|S|}\rho^{|S|} = 1$). If you prefer, you can think of T_ρ for $\rho > 1$ as the inverse operator of $T_{\rho^{-1}}$, which has a probabilistic interpretation.

We can now present the proof of Lemma 5. Recall that we are given a function h of degree at most d. Writing $h = T_{1/\sqrt{3}} T_{\sqrt{3}} h$:

$$\|h\|_4^2 = \|T_{1/\sqrt{3}} T_{\sqrt{3}} h\|_4^2 \overset{\text{Lem 6}}{\leq} \|T_{\sqrt{3}} h\|_2^2 = \sum_S (\sqrt{3}^{|S|}\hat{h}(S))^2 \overset{\deg h \leq d}{\leq} 3^d \sum_S \hat{h}(S)^2 = 3^d \|h\|_2^2.$$

3.1 Bonami's lemma

It remains to prove Lemma 6, also known as Bonami's lemma. The proof is a simple induction on the dimension.

When $n = 0$, f is a constant, and $T_{1/\sqrt{3}}f = f$, so there is nothing to proof.

Now suppose that Lemma 6 holds for some n. We prove it for $n+1$. Given a function $f\colon \{-1,1\}^{n+1} \to \mathbb{R}$, we construct from it two functions on the n-dimensional cube by writing

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

Explicitly,

$$\begin{aligned} g &= \sum_{S\subseteq[n]} \hat{f}(S)\chi_S, \\ h &= \sum_{S\subseteq[n]} \hat{f}(S\cup\{n+1\})\chi_S. \end{aligned}$$

The Fourier formula for $T_\rho f$ shows that

$$T_{1/\sqrt{3}}f(x_1,\ldots,x_{n+1}) = T_{1/\sqrt{3}}g(x_1,\ldots,x_n) + \tfrac{1}{\sqrt{3}}x_{n+1}T_{1/\sqrt{3}}h(x_1,\ldots,x_n).$$

Indeed, T_ρ has the effect of multiplying every x_i in the Fourier expansion by ρ. It follows that

$$\begin{aligned} \|T_{1/\sqrt{3}}f\|_4^4 &= \mathop{\mathbb{E}}_{x_1,\ldots,x_{n+1}}\left[\left(T_{1/\sqrt{3}}g(x_1,\ldots,x_n) + \tfrac{1}{\sqrt{3}}x_{n+1}T_{1/\sqrt{3}}h(x_1,\ldots,x_n)\right)^4\right] = \\ &\underbrace{\mathbb{E}[(T_{1/\sqrt{3}}g)^4]}_{(0)} + \underbrace{\tfrac{4}{\sqrt{3}}\,\mathbb{E}[(T_{1/\sqrt{3}}g)^3(T_{1/\sqrt{3}}h)]\,\mathbb{E}[x_{n+1}]}_{(1)} + \underbrace{\tfrac{6}{\sqrt{3}^2}\,\mathbb{E}[(T_{1/\sqrt{3}}g)^2(T_{1/\sqrt{3}}h)^2]\,\mathbb{E}[x_{n+1}^2]}_{(2)} + \\ &\qquad \underbrace{\tfrac{4}{\sqrt{3}^3}\,\mathbb{E}[(T_{1/\sqrt{3}}g)(T_{1/\sqrt{3}}h)^3]\,\mathbb{E}[x_{n+1}^3]}_{(3)} + \underbrace{\tfrac{1}{\sqrt{3}^4}\,\mathbb{E}[(T_{1/\sqrt{3}}h)^4]\,\mathbb{E}[x_{n+1}^4]}_{(4)}. \end{aligned}$$

This is not as bad as it looks. Since $\mathbb{E}[x_{n+1}] = \mathbb{E}[x_{n+1}^3] = 0$, terms (1) and (3) drop. Using $x_{n+1}^2 = 1$, the situation simplifies to

$$\|T_{1/\sqrt{3}}f\|_4^4 \le \underbrace{\mathbb{E}[(T_{1/\sqrt{3}}g)^4]}_{(0)} + \underbrace{2\,\mathbb{E}[(T_{1/\sqrt{3}}g)^2(T_{1/\sqrt{3}}h)^2]}_{(2)} + \underbrace{\tfrac{1}{9}\,\mathbb{E}[(T_{1/\sqrt{3}}h)^4]}_{(4)}.$$

We can bound terms (0) and (4) using the induction hypothesis, by $\|g\|_2^4$ and $\|h\|_2^4$, respectively. In order to bound (2), we apply the Cauchy–Schwarz inequality:

$$(2) \le 2\sqrt{\mathbb{E}[(T_{1/\sqrt{3}}g)^4]}\sqrt{\mathbb{E}[(T_{1/\sqrt{3}}h)^4]} \le 2\|g\|_2^2\|h\|_2^2.$$

Combining the three bounds, we have

$$\|T_{1/\sqrt{3}}f\|_4^4 \le \|g\|_2^4 + 2\|g\|_2^2\|h\|_2^2 + \tfrac{1}{9}\|h\|_2^4 \le \|g\|_2^4 + 2\|g\|_2^2\|h\|_2^2 + \|h\|_2^4 = (\|g\|_2^2 + \|h\|_2^2)^2.$$

In order to complete the proof, observe that

$$\|g\|_2^2 + \|h\|_2^2 = \sum_{S\subseteq[n]} \hat{f}(S)^2 + \sum_{S\subseteq[n]} \hat{f}(S\cup\{n+1\})^2 = \sum_{T\subseteq[n+1]} \hat{f}(T)^2 = \|f\|_2^2.$$

Therefore

$$\|T_{1/\sqrt{3}}f\|_4^4 \le (\|g\|_2^2 + \|h\|_2^2) = \|f\|_2^4,$$

completing the proof of the inductive step.

3.2 General hypercontractivity

If $p \geq q \geq 1$ then

$$\|T_\rho f\|_p \leq \|f\|_q \text{ for } \rho = \sqrt{\frac{q-1}{p-1}}.$$

This is proved in two steps:

1. The base case $n = 1$, known as a *two-point inequality*, since it involves a function whose domain consists of two points.

2. A short inductive argument.

By optimizing over q, one obtains the sharp level-d inequality mentioned above.

3.3 Small set expansion

Hypercontractivity has an interpretation in terms of the edge-expansion of sets. Taking $p = 2$ and $q = 1 + \rho$ in the general statement of hypercontractivity gives

$$\|T_{\sqrt{\rho}} f\|_2 \leq \|f\|_{1+\rho}.$$

Observe now that

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

If f is $\{0,1\}$-valued then $\|f\|_{1+\rho}^{1+\rho} = \|f\|_2^2$, and so

$$\langle T_\rho f, f\rangle \leq \|f\|_{1+\rho}^2 = (\|f\|_2^2)^{2/(1+\rho)} = \mathbb{E}[f]^{2/(1+\rho)}.$$

Suppose that $f = 1_S$. The inner product $\langle T_\rho f, f\rangle$ can be expanded as

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

Therefore

$$\Pr_{x,y \sim N_\rho(x)}[x, y \in S \mid x \in S] = \frac{\langle T_\rho f, f\rangle}{\mu(S)} \leq \mu(S)^{(1-\rho)/(1+\rho)},$$

where $\mu(S) = |S|/2^n$ is the measure of S. In words, if we choose a random point in S and perform a short "walk" corresponding to the noise operator T_ρ,[1] then the probability that we stay in S is only $\mu(S)^{(1-\rho)/(1+\rho)}$, which is small when ρ is constant and $\mu(S)$ is small. In other words, small sets expand.

3.4 Log-Sobolev inequality

From a broader perspective, hypercontractivity is equivalent to fast mixing of the random walk on the hypercube $\{-1, 1\}^n$.

There are many ways to measure how fast a Markov chain mixes. Often one is interested in mixing in total variation distance, but this can be hard to control. Another popular notion is L_2 mixing: we measure how fast the random walk mixes by computing the L_2 difference between the stationary distribution and the distribution after a given number of steps (or, in the continuous time setting, after a given amount of time). As is well-known, the rate of convergence in the L_2 metric is controlled by the *spectral gap* of the walk.

Diaconis and Saloff-Coste showed that hypercontractivity is equivalent to mixing in the Kullback–Leibler metric, which is controlled by the so-called *log-Sobolev constant*, which is the best constant in the eponymous

[1] The distribution $N_\rho(x)$ corresponds a rate 1 continuous time random walk on the hypercube stopped at time $\frac{1}{2}\ln(1/\rho)$.

log-Sobolev inequality. The log-Sobolev constant is bounded by the spectral gap, meaning that a log-Sobolev inequality is stronger than a spectral gap (formally, having a log-Sobolev inequality implies having a spectral gap, but the opposite doesn't hold).

Mossel, Oleszkiewicz and Sen showed that *reverse hypercontractivity* (hypercontractivity for L_p-"norms" for $p < 1$; the direction of the inequality is reversed!) is equivalent to satisfying a *modified log Sobolev inequality*, an inequality which is weaker than log-Sobolev (but equivalent to it in the continuous setting) but stronger than just having a spectral gap. This means that in some cases reverse hypercontractivity holds while hypercontractivity doesn't. Reverse hypercontractivity implies a useful hitting property: for any two sets A, B, the probability that $x \in A$ and $N_\rho(x) \in B$ is lower-bounded in terms of the measures of A and B.

3.5 Other noise operators

The operator T_ρ is the standard noise operator used in Boolean function analysis. It has several important properties:

- If x has the uniform distribution over $\{\pm 1\}^n$, then so does $N_\rho(x)$. This makes T_ρ self-adjoint with respect to the inner product $\langle \cdot, \cdot \rangle$, which is defined using the uniform measure.
- T_ρ is hypercontractive.
- The Fourier characters are eigenfunctions of T_ρ, with eigenvalues $T_\rho \chi_S = \rho^{|S|} \chi_S$.

Sometimes other noise operators are studied. There are several reasons:

- We are trying to understand a particular random walk. One example is the noise operator that flips exactly $\frac{1-\rho}{2} n$ coordinates (rather than each coordinate with probability $\frac{1-\rho}{2}$), which appears in coding theory.
- We are interested in a different measure over $\{\pm 1\}^n$, or in a different domain altogether. We will see an example below, the biased measure over $\{\pm 1\}^n$.

In the latter case, we don't necessarily care about which noise operator we use, but rather about what we can prove with it. In many applications, we aim at proving a level-d inequality. Reflecting at the proof, we used two properties: hypercontractivity (this was used to obtain $\|f\|^3$ rather than just $\|f\|^2$ on the right-hand side; every power larger than 2 would be useful) and a lower bound on the eigenvalues corresponding to low-degree functions (the factor 3^d is just $(1/\sqrt{3}^d)^{-2}$, where $1/\sqrt{3}^d$ is the minimal eigenvalue of an eigenfunction of degree at most d). One way to get such operators is via random walks, but this is not the only way. In recent years, a different approach has been developed, by Noam Lifshitz and others. The idea is to use the properties of one domain X to reason about another domain Y via a coupling between these two domains. The noise operator we use for Y is derived from that of X via the coupling, and so it doesn't necessarily correspond to a random walk.

4 Sharp thresholds

One of the celebrated applications of Boolean function analysis is sharp threshold theorems. Our goal is to outline a new proof of one of them, Bourgain's booster theorem, using *global hypercontractivity*. In this section we switch gears, moving faster while providing fewer details, and completely omitting some of the proofs.

In the $G(n, p)$ random graph model, there are n fixed vertices, and each of the $\binom{n}{2}$ edges between them is added with probability p independently. Here are two classical results about this model:

$$\Pr\left[G\big(n, \frac{c}{n}\big) \text{ contains a triangle}\right] \longrightarrow 1 - e^{-c^3/6},$$
$$\Pr\left[G\big(n, \frac{\log n + c}{n}\big) \text{ is connected}\right] \to e^{-e^{-c}}.$$

Both limits are as $n \to \infty$.

Although this is not the subject of these notes, let us briefly explain where these two expressions come from. The expected number of triangles in $G(n, \frac{c}{n})$ is $\binom{n}{3}(c/n)^3 \approx c^3/6$, and the distribution of the number of triangles converges to a Poisson distribution, since the different triangles are mostly edge-disjoint. The expression $1 - e^{-c^3/6}$ is just the probability that a Poisson random variable with expectation $c^3/6$ is non-zero.

Similarly, it turns out that the main obstruction for connectivity is isolated vertices. The probability that a vertex in $G(n, \frac{\log n + c}{n})$ is isolated is $(1 - \frac{\log n + c}{n})^{n-1} \approx e^{-c}/n$, and so the expected number of isolated vertices is e^{-c}. Once again, the distribution is roughly Poisson, and $e^{-e^{-c}}$ is just the probability that the Poisson random variable equals zero.

Both properties considered above are *monotone*: if they hold for a given graph, then they continue to hold even if we add edges. This implies that there is a critical probability p_c for which the probability of the event is exactly 1/2. The critical probability for containing a triangle is $\Theta(\frac{1}{n})$, and for being connected is $\frac{\log n + \Theta(1)}{n}$.

The "action" in the case of connectivity lies in a narrow window of width $\Theta(\frac{1}{n})$, which is much smaller than the critical probability, which is roughly $\frac{\log n}{n}$. We say that connectivity has a *sharp threshold*. In contrast, in the case of containing a triangle, both are of order $\frac{1}{n}$, corresponding to a *coarse threshold*.

Formally, given a monotone graph property P, let p_- be the probability such that $G(n, p_-)$ satisfies P w.p. 1/3, and let p_+ be the probability such that $G(n, p_+)$ satisfies P w.p. 2/3 (the constants 1/3 and 2/3 are arbitrary). The property P has a sharp threshold if $p_+ - p_- = o(p_c)$.

Knowing that a threshold is sharp helps locating it, since $p_- \approx p_c \approx p_+$. Bourgain's booster theorem gives a criterion: *If P is a monotone graph property with a coarse threshold and $p_c = o(1)$ then there is a subgraph H consisting of $O(1)$ edges such that adding H to $G(n,p)$ boosts the probability of satisfying P by $\Omega(1)$, for some $p \in [p_-, p_+]$.* For example, we can boost the property of containing a triangle to 1 by adding a triangle (for a less trivial example, consider the property of containing at least 10 triangles). In contrast, no such boosting happens for connectivity, since it is unlikely that the fixed graph H touches one of the few isolated vertices, which are randomly distributed.

Bourgain's original proof was short but tricky. The recently developed technique of global hypercontractivity, developed by Noam Lifshitz and others, enabled them to give a simple conceptual proof. The argument works for arbitrary monotone $f\colon \{0,1\}^n \to \{0,1\}$.

4.1 Russo–Margulis formula

For any non-constant monotone function $f\colon \{0,1\}^n \to \{0,1\}$ there is a critical probability p_c such that $\Pr_{x \sim \mu_{p_c}}[f(x) = 1] = 1/2$, where μ_p is the distribution over $\{0,1\}^n$ in which each coordinate equals 1 with probability p independently. The critical probability exists since the function

$$\phi_f(p) = \Pr_{x \sim \mu_p}[f(x) = 1]$$

is continuous (since it can be expressed as a polynomial), increasing, and satisfies $\phi_f(0) = 0$ and $\phi_f(1) = 1$.

The property corresponding to f has a coarse threshold if $p_+ - p_- = \Omega(p_c)$, where $p_- = \phi_f^{-1}(1/3)$ and $p_+ = \phi_f^{-1}(2/3)$. This suggests that in order to understand coarse thresholds, we should take a look at the derivative of ϕ_f: the condition above implies that $\phi_f'(p) \le \frac{2/3 - 1/3}{p_+ - p_-} = O(1/p_c)$ for some $p \in [p_-, p_+]$.

We can compute $\phi_f'(p) = \lim_{\epsilon \to 0} \frac{\phi_f(p+\epsilon) - \phi_f(p)}{\epsilon}$ using a coupling between $\mu_{p+\epsilon}$ and μ_p. For each i, let $u_i \sim U([0,1])$ be a uniform random variable. Let $x \sim \mu_p$ be obtained by setting $x_i = 1$ if $u_i < p$, and similarly obtain $y \sim \mu_{p+\epsilon}$ by setting $y_i = 1$ if $u_i < p + \epsilon$; thus $x_i \le y_i$ always, and $x_i < y_i$ w.p. ϵ.

For a small $\epsilon > 0$ we have

$$\phi_f(p+\epsilon) - \phi_f(p) = \mathop{\mathbb{E}}_{x,y}[f(y) - f(x)] = \Pr_{x,y}[f(x) = 0 \text{ and } f(y) = 1].$$

If $y = x$ then certainly $f(x) = f(y)$. Otherwise, it is likely that x, y differ only in a single coordinate: any other outcome has probability $O(\epsilon^2)$. For any given coordinate i, the two inputs differ only on i with probability $\epsilon - O(\epsilon^2)$. Assuming this happens, the remaining coordinates are still sampled according to μ_p, and so

$$\phi_f(p+\epsilon) - \phi_f(p) = \epsilon \sum_{i=1}^n \Pr_{z \sim \mu_p}[f(z|_{i \leftarrow 0}) = 0 \text{ and } f(z|_{i \leftarrow 1}) = 1] + O(\epsilon^2),$$

where $z|_{i \leftarrow b}$ sets z_i to b.

We immediately deduce the Russo–Margulis formula:

$$\phi'_f(p) = \sum_{i=1}^n \Pr_{z \sim \mu_p}[f(z|_{i \leftarrow 0}) = 0 \text{ and } f(z|_{i \leftarrow 1}) = 1].$$

The expression on the right should ring a bell: it is very similar to the definition of *total influence*.

4.2 Biased Fourier analysis

In order to proceed, we need to introduce biased Fourier analysis, which is Fourier analysis on $\{0,1\}^n$ with respect to the biased measure μ_p. Up to now we have considered the case $p = 1/2$ (though using ± 1 rather than $0, 1$, only a minor difference).

One salient property of the Fourier basis χ_S was its orthonormality. In our new setting, we also seek a basis ω_S which is orthonormal with respect to the inner product

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

Moreover, we would like ω_S to depend only on the coordinates in S. Up to sign, this completely determines ω_S:

$$\omega_S(x_1, \ldots, x_n) = \prod_{i \in S} \frac{x_i - p}{\sqrt{p(1-p)}}.$$

Where does this formula come from? We simply subtracted from x_i its expectation (this implies orthogonality), and divided by the standard deviation (this implies orthonormality).

As in the classical case of $\{-1,1\}^n$, here as well we define the Fourier expansion to be

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

The parameter p does not appear anywhere in this expression! It will have to be understood from context.

Next, we would like to obtain a Fourier expression for ϕ'_f. The first step is to replace the existing expression by one which is more amenable to "L_2 manipulations":

$$\phi'_f(p) = \sum_{i=1}^n \Pr_{z \sim \mu_p}[f(z|_{i \leftarrow 0}) = 0 \text{ and } f(z|_{i \leftarrow 1}) = 1] = \sum_{i=1}^n \mathop{\mathbb{E}}_{z \sim \mu_p}[\big(f(z|_{i \leftarrow 0}) - f(z|_{i \leftarrow 1})\big)^2].$$

Substituting the Fourier expansion,

$$f(z|_{i \leftarrow 0}) - f(z|_{i \leftarrow 1}) = \sum_S \hat{f}(S)(\omega_S(z|_{i \leftarrow 0}) - \omega_S(z|_{i \leftarrow 1})) = \sum_{i \in S} \hat{f}(S) \omega_{S \setminus \{i\}}(z) \frac{(0-p)-(1-p)}{\sqrt{p(1-p)}} =$$

$$\frac{-1}{\sqrt{p(1-p)}} \sum_{i \in S} \hat{f}(S) \omega_{S \setminus \{i\}}(z).$$

This implies that

$$\mathop{\mathbb{E}}_{z \sim \mu_p}[\big(f(z|_{i \leftarrow 0}) - f(z|_{i \leftarrow 1})\big)^2] = \frac{1}{p(1-p)} \sum_{i \in S} \hat{f}(S)^2.$$

Suggestively, we define the i'th influence using the same Fourier expression as before:

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

In these terms, the Russo–Margulis formula becomes

$$\phi'_f(p) = \frac{1}{p(1-p)} \mathrm{Inf}[f].$$

Recall that if the threshold is coarse then $\phi'_f(p) = O(1/p)$ for some $p \in [p_-, p_+]$. The assumption $p_c = o(1)$ implies that also $p = o(1)$,[2] and so $\phi'_f(p) = O(1/p(1-p))$. In total, we deduce that

$$\mathrm{Inf}[f] = O(1)$$

for some $p \in [p_-, p_+]$.

If p were constant, then Friedgut's celebrated junta theorem (whose proof is very similar to the proof of the KKL theorem) states that f is close to a *junta*, which is a function depending on $O(1)$ coordinates.

Theorem 7 (Friedgut's junta theorem). *Let $f\colon \{\pm 1\}^n \to \{\pm 1\}$, where $\{\pm 1\}^n$ is considered with respect to the uniform distribution. For every $\epsilon > 0$, there exists a function $g\colon \{\pm 1\}^n \to \{\pm 1\}$ depending on $2^{O(\mathrm{Inf}[f]/\epsilon)}$ coordinates such that $\Pr[f \neq g] = O(\epsilon)$ (we say that f is $O(\epsilon)$-close to g).*

This cannot be the case for graph properties such as containing a triangle, since they are isomorphism-invariant, implying that all coordinates "look the same". Something in the Fourier argument thus has to break down, and this thing is hypercontractivity.

For completeness, we include a sketch of the proof of Friedgut's junta theorem.

Proof. The function g will depend on all variables whose influence is at least $\tau > 0$, for some τ which arises from the proof. Let J be the set of these variables, and define a function G by averaging over all variables not in J. The function G is not necessarily Boolean, and g itself will be defined by rounding it to a Boolean function.

If χ_S is a character and we average it over all coordinates outside of J, then there are two cases. If $S \subseteq J$ then averaging over the coordinates outside of J has no effect. Otherwise, averaging "kills" the character (reduces it to the zero function):

$$\mathop{\mathbb{E}}_{x_{\overline{J}}}[\chi_S] = \prod_{i \in S \cap J} x_i \times \prod_{i \in S \setminus J} \mathbb{E}[x_i] = 0.$$

Therefore

$$G = \sum_{S \subseteq J} \hat{f}(S)\chi_S.$$

This formula allows us to compute the distance between f and G:

$$\mathbb{E}[(f-G)^2] = \mathbb{E}\left[\left(\sum_{S \not\subseteq J} \hat{f}(S)\chi_S\right)^2\right] = \sum_{S \not\subseteq J} \hat{f}(S)^2.$$

We now proceed as in the proof of the KKL theorem, using $\Pr[|S| \geq \mathrm{Inf}[f]/\epsilon] \leq \epsilon$:

$$\mathbb{E}[(f-G)^2] \leq \epsilon + \sum_{|S| \leq \mathrm{Inf}[f]/\epsilon} |S \setminus J| \hat{f}(S)^2 = \epsilon + \sum_{i \notin J} \|L_i f^{\leq \mathrm{Inf}[f]/\epsilon}\|^2.$$

[2] It suffices to observe that $p_+ \leq 2p_c = o(1)$. Indeed, sample $x, y \sim \mu_{p_c}$ and take their maximum $z = x \vee y$, whose distribution is $\mu_{2p_c - p_c^2}$. Since $f(z) = 1$ if either $f(x) = 1$ or $f(y) = 1$, we have $\phi_f(2p_c - p_c^2) \geq 3/4$. Thus $p_+ \leq 2p_c - p_c^2$.

Applying the level-d inequality,

$$\mathbb{E}[(f-G)^2] \le \epsilon + 3^{\mathrm{Inf}[f]/\epsilon} \sum_{i \notin J} \|L_i f\|^3 \le \epsilon + 3^{\mathrm{Inf}[f]/\epsilon} \mathrm{Inf}[f] \max_{i \notin J} \sqrt{\mathrm{Inf}_i[f]} \le \epsilon + 3^{\mathrm{Inf}[f]/\epsilon} \mathrm{Inf}[f] \sqrt{\tau}.$$

We can write this as

$$\mathbb{E}[(f-G)^2] \le \epsilon \left(1 + 3^{\mathrm{Inf}[f]/\epsilon} (\mathrm{Inf}[f]/\epsilon) \sqrt{\tau}\right).$$

Choosing $\tau = (\frac{1}{3^{\mathrm{Inf}[f]/\epsilon}(\mathrm{Inf}[f]/\epsilon)})^2$, we obtain $\mathbb{E}[(f-G)^2] \le 2\epsilon$.

There are at most $\mathrm{Inf}[f]/\tau = 2^{O(\mathrm{Inf}[f]/\epsilon)}$ many coordinates in J.[3] Thus f is close to a function G depending on the correct number of coordinates. The only problem is that G isn't Boolean. Fortunately, this is easy to fix. We define g by rounding G to $\{-1, 1\}$ pointwise. The function g clearly depends on the same coordinates as G. On the other hand, if $a \in \{-1, 1\}$, b is arbitrary, and B results from rounding b to $\{-1, 1\}$, then

$$(a - B)^2 \le 4(a - b)^2.$$

(The worst case is $a = 1$, $b = -\epsilon$, $B = -1$.) This shows that $\mathbb{E}[(f-g)^2] \le 8\epsilon$. Since both f and g are Boolean, this implies that $\Pr[f \ne g] = 2\epsilon$. □

5 Global hypercontractivity

Bonami's lemma states that when $p = 1/2$,

$$\|T_\rho f\|_4 \le \|f\|_2, \text{ where } \rho = \frac{1}{\sqrt{3}}.$$

Does this hold for other values of p, perhaps with a different ρ? In order to even state this question, we need to define $T_\rho f$ in general. We simply take the Fourier definition:

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

This is a reasonable definition since it will allow us to mimic arguments such as the proof of the KKL lemma, which involved only the Fourier interpretation of T_ρ.

That said, there is also a natural interpretation of T_ρ as an averaging operator with respect to the distributions $N_\rho(x)$, defined as follows:

$$y_i = \begin{cases} x_i & \text{w.p. } \rho, \\ \mu_p & \text{w.p. } 1 - \rho, \end{cases}$$

where in the second case we resample the coordinate according to μ_p. To see that this works, observe that (using T_ρ for the definition using N_ρ)

$$T_\rho \frac{x_i - p}{\sqrt{p(1-p)}} = \mathbb{E}\left[\frac{y_i - p}{\sqrt{p(1-p)}}\right] = \rho \frac{x_i - p}{\sqrt{p(1-p)}} + (1-\rho) \mathop{\mathbb{E}}_{z_i \sim \mu_p}\left[\frac{z_i - p}{\sqrt{p(1-p)}}\right] = \rho \frac{x_i - p}{\sqrt{p(1-p)}}.$$

Therefore $T_\rho \omega_i = \rho \omega_i$. Similarly, $T_\rho \omega_S = \rho^{|S|} \omega_S$.

Instead of considering hypercontractivity directly, it will be slightly easier to consider its immediate consequence, Lemma 5:

$$\|f\|_4 \le \rho^{-\deg f} \|f\|_2.$$

[3] For this calculation to work out we need the benign assumption $\epsilon = O(1)$ (if it doesn't hold, we can use any Boolean function for g).

Taking as f the function $f(x) = x_1$, this inequality reads

$$p^{1/4} \leq \rho^{-1} p^{1/2} \implies \rho \leq p^{1/4}.$$

It turns out that Bonami's lemma indeed holds for some $\rho = O(p^{1/4})$, with a very similar proof (the only complication is that the term we denoted by (3) no longer vanishes). This allows most results in the area to carry through for arbitrary p (though we do lose the property $\chi_S \chi_T = \chi_{S \triangle T}$ which was useful for analyzing linearity testing).

When $p = o(1)$ (as in the case of the critical probability of the two graph properties considered above), the parameter ρ has to go to zero with n, and hypercontractivity becomes useless, due to the appearance of ρ^{-1} in the proof (cf. the proof of KKL). This is not just an issue with hypercontractivity itself: results such as Friedgut's junta theorem, which hold for constant p, stop working, and a more nuanced picture appears.

It turns out that we can recover hypercontractivity if we assume that f is *global*. In order to define this concept, we first define the operator D_i, familiar from the Russo–Margulis formula:

$$D_i f(x) = f(x|_{i \leftarrow 1}) - f(x|_{i \leftarrow 0}).$$

This is just the derivative of the Fourier expansion of f with respect to x_i. Also, up to scaling, this is (almost) the same as the operator L_i considered in Section 2. Indeed, computing the derivative of the Fourier expansion gives

$$D_i f = \frac{1}{\sqrt{p(1-p)}} \sum_{i \in S} \hat{f}(S) \omega_{S \setminus \{i\}}.$$

We extend the definition of D_i to sets of coordinates:

$$D_{\{i_1, \ldots, i_\ell\}} f = D_{i_1} D_{i_2} \cdots D_{i_\ell} f.$$

The order of application doesn't matter since

$$D_S f(x) = \sum_{y \in \{0,1\}^S} (-1)^{|y| + |S|} f(x|_{S \leftarrow y}) = \frac{1}{(p(1-p))^{|S|/2}} \sum_{S \subseteq T} \hat{f}(T) \omega_{T \setminus S}.$$

A function f is β-*global* if

$$\|D_S f\|_2 \leq \beta \text{ for all } S.$$

When $p = 1/2$, $\|D_i f\|_2^2 = 4 \operatorname{Inf}_i[f]$, and so globalness means that all influences are small, as well as some higher-order extensions of them.

A function such as $f(x) = x_1$ is not global (that is, not β-global for any small β) since $D_1 f = 1$. Globalness also considers higher-order derivatives. This is necessary since functions such as $f(x) = x_1 x_2$ also falsify Lemma 5.

We can now state one version of global hypercontractivity: if f is β-global then

$$\|T_\rho f\|_4 \leq \sqrt{\beta \|f\|_2} \text{ for some \textit{universal} constant } \rho \text{ independent of } p.$$

Global hypercontractivity can be proved in several ways. While none of them is particularly difficult, all of them are somewhat long, and so we skip the proof.

5.1 Bourgain's booster theorem

Finally we have all the tools to prove Bourgain's booster theorem. Given a non-constant monotone function $f\colon \{0,1\}^n \to \{0,1\}$ such that $\operatorname{Inf}[f] \leq K$ with respect to some p such that $p \in [p_-, p_+]$, we would like to show that there is a set S of $O_K(1)$ coordinates such that $\Pr_{x \sim \mu_p}[f(x|_{S \leftarrow 1})] \geq \phi_f(p) + \Omega_K(1)$.

The proof is very similar to the proof of the KKL theorem, where the assumption $\operatorname{Inf}[f] \leq K$ replaces the assumption $\operatorname{Inf}[f] \leq c \log n$. On the one hand,

$$\|f\|^2 = \mathbb{E}[f^2] = \mathbb{E}[f] \geq \frac{1}{9},$$

and on the other hand,

$$\|f^{\geq 10K}\|^2 \leq \frac{1}{10}$$

using Markov's inequality. Thus

$$\|f^{\leq 10K}\|^2 \geq \frac{1}{90}.$$

Next, we mimic the proof of the weak level-d inequality, Lemma 4:

$$\frac{1}{90} \leq \|f^{\leq 10K}\|^2 = \langle f^{\leq 10K}, f\rangle \leq \|f^{\leq 10K}\|_4 \|f\|_{4/3}.$$

When proving the level-d inequality, we related $\|f\|_{4/3}$ and $\|f\|_2$. Here it is enough to use the simple bound

$$\|f\|_{4/3} \leq \|f\|_2 \leq 1,$$

which follows from the monotonicity of L_q-norms in q. Thus

$$\|f^{\leq 10K}\|_4 \geq \frac{1}{90}.$$

At this point we invoke global hypercontractivity. We first write $f^{\leq 10K} = T_\rho T_{\rho^{-1}} f^{\leq 10K}$, where ρ is the constant from global hypercontractivity. Let β be the globalness of $T_{\rho^{-1}} f^{\leq 10K}$, that is, the minimum β such that $T_{\rho^{-1}} f^{\leq 10K}$ is β-global. Then

$$\frac{1}{90^2} \leq \beta \|T_{\rho^{-1}} f^{\leq 10K}\|_2 \leq \beta \rho^{-10K} \|f\|_2 \leq \beta.$$

We deduce that the function $f^{\leq 10K}$ isn't so global! There must exist a set S such that

$$\|D_S T_{\rho^{-1}} f^{\leq 10K}\|_2 \geq \frac{\rho^{10K}}{90^2}.$$

Since D_S is the derivative according to the coordinates in S, we have $D_S T_{\rho^{-1}} f^{\leq 10K} = 0$ unless $|S| \leq 10K$. Therefore the set S above has size at most $10K$. On the other hand, the Fourier expression for D_S implies that $\|D_S T_{\rho^{-1}} f^{\leq 10K}\|_2 \leq \rho^{-10K} \|D_S f\|_2$. Summarizing, there exists a set S of size at most $10K$ such that

$$\frac{\rho^{40K}}{90^4} \leq \|D_S f\|_2^2.$$

We are almost done. Recall the explicit expression

$$D_S f(x) = \sum_{y \in \{0,1\}^S} (-1)^{|y|+|S|} f(x|_{S \leftarrow y}).$$

We always have $|D_S f(x)| \leq 2^{|S|-1}$, and furthermore $D_S f(x) = 0$ unless $f(x|_{S\leftarrow \mathbf{1}}) = 1$ and $f(x|_{S \leftarrow \mathbf{0}}) = 0$. Therefore

$$\frac{\rho^{40K}}{90^4} \leq \mathop{\mathbb{E}}_{x \sim \mu_p}[(D_S f(x))^2] \leq 4^{10K} \Pr_{x\sim\mu_p}[f(x|_{S\leftarrow \mathbf{1}}) = 1 \text{ and } f(x|_{S\leftarrow \mathbf{0}}) = 0] \Longrightarrow$$
$$\Pr_{x\sim\mu_p}[f(x|_{S\leftarrow \mathbf{1}}) = 1 \text{ and } f(x|_{S\leftarrow \mathbf{0}}) = 0] \geq e^{-\Omega(K)}.$$

The set S is the required booster, since $x|_S = \mathbf{0}$ happens with probability $O_K(p_c)$. Formally,

$$\mathop{\mathbb{E}}_{x\sim\mu_p}[f(x|_{S\leftarrow\mathbf{1}})] \geq \mathop{\mathbb{E}}_{x\sim\mu_p}[f(x|_{S\leftarrow\mathbf{0}})] + e^{-\Omega(K)} \geq \mathop{\mathbb{E}}_{x\sim\mu_p}[f(x)] - \Pr_{x\sim\mu_p}[x|_S \neq \mathbf{0}] + e^{-\Omega(K)} = \mathop{\mathbb{E}}_{x\sim\mu_p}[f(x)] + e^{-\Omega(K)} - 10Kp.$$

If $p = o(1)$ (which follows from $p_c = o(1)$), then the final error term is insignificant, completing the proof.

6 Bonus: Erdős–Ko–Rado

As a bonus, here is a short proof of the p-biased version of the Erdős–Ko–Rado theorem.

Theorem 8 (Erdős–Ko–Rado)**.** *If $p \le 1/2$ and $\mathcal{F}$ is an intersecting family then $\mu_p(\mathcal{F}) \le p$. If furthermore $p < 1/2$ and $\mu_p(\mathcal{F}) = p$ then $\mathcal{F} = \{S : i \in S\}$ for some i.*

Consider the following distribution:

- $x \sim \mu_p$.
- $y \sim N_\rho(x)$, where $\rho = -\frac{p}{1-p}$.

We claim that x, y are always disjoint (where we identify elements of $\{0,1\}^n$ with subsets of $[n]$). Indeed, considering the probabilistic interpretation of N_ρ (pretending that $\rho \in [0,1]$), the probability that $x_i = y_i = 1$ is

$$p \cdot (\rho + (1-\rho)p = p \cdot \left(-\frac{p}{1-p} + \frac{1}{1-p} \cdot p\right) = 0.$$

If $\mathcal{F}$ is intersecting and $f = 1_{\mathcal{F}}$ then

$$\langle f, T_\rho f \rangle = \mathop{\mathbb{E}}_{x,y}[f(x)f(y)] = 0.$$

On the other hand,

$$\langle f, T_\rho f \rangle = \sum_S \left(-\frac{p}{1-p}\right)^{|S|} \hat{f}(S)^2 \ge \hat{f}(\emptyset)^2 - \frac{p}{1-p} \sum_{S \ne \emptyset} \hat{f}(S)^2,$$

using $\frac{p}{1-p} \le 1$, which holds when $p \le 1/2$.

Recall that $\hat{f}(\emptyset) = \mu_p(\mathcal{F})$. Also,

$$\sum_{S \ne \emptyset} \hat{f}(S)^2 = \sum_S \hat{f}(S)^2 - \hat{f}(\emptyset)^2 = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = \mathrm{Var}[f] = \mu_p(\mathcal{F})(1 - \mu_p(\mathcal{F})).$$

It follows that

$$0 \ge \mu_p(\mathcal{F})^2 - \frac{p}{1-p}\mu_p(\mathcal{F})(1-\mu_p(\mathcal{F})) \Longrightarrow 0 \ge (1-p)\mu_p(\mathcal{F}) - p(1-\mu_p(\mathcal{F})) = \mu_p(\mathcal{F}) - p.$$

In other words, $\mu_p(\mathcal{F}) \le p$.

If $\mu_p(\mathcal{F}) = p$ then the inequality

$$\sum_S \left(-\frac{p}{1-p}\right)^{|S|} \hat{f}(S)^2 \ge \hat{f}(\emptyset)^2 - \frac{p}{1-p} \sum_{S \ne \emptyset} \hat{f}(S)^2$$

has to be tight. When $p < 1/2$ we have $\frac{p}{1-p} < 1$, and so this inequality can only be tight if all the Fourier mass of f lies on levels 0 and 1, that is, $\deg f \le 1$. A short argument (exercise!) implies that $f \in \{0, 1, x_i, 1 - x_i\}$ for some i. Since $f = 1_{\mathcal{F}}$, necessarily $f = x_i$, and so $\mathcal{F} = \{S : i \in S\}$.

This argument in fact shows a little bit more: if $\mu_p(\mathcal{F})$ is close to p then almost all the Fourier mass of f lies on levels 0 and 1. The Friedgut–Kalai–Naor theorem (which was proved in order to complete the proof of Kalai's quantitative Arrow's theorem) then implies that f is close to a function of one of the forms $0, 1, x_i, 1 - x_i$, and so $\mathcal{F}$ is close to a family of the form $\{S : i \in S\}$.