Boolean Function Analysis

Yuval Filmus

June 2, 2022

Contents

1 Introduction: linearity testing

Boolean function analysis [O'D14] studies functions $f\colon \{0,1\}^n \to \{0,1\}^n$ (known as Boolean functions) from a spectral perspective. (Often we will replace $0,1$ by ± 1.) These functions could come from a Boolean circuit, from a probabilistically checkable proof, from an error-correcting code, from an intersecting family, and so on. Much of the area is dedicated to understanding the structure of functions which satisfy given properties. By way of introduction, we will consider one of the simplest applications of Boolean function analysis: *linearity testing*.

A function $f\colon \{\pm 1\}^n \to \{\pm 1\}$ is a *homomorphism* or *linear* if for all $x, y \in \{\pm 1\}^n$ we have

$$f(xy) = f(x)f(y).$$

Which functions are linear? First, note that $f(1) = f(1)^2$, and so $f(1) = 1$. Next, let e^i be the vector given by $e^i_i = -1$ and $e^i_j = 1$ for all $j \neq i$. Then for all $x \in \{\pm 1\}^n$,

$$f(x) = \prod_{i\colon x_i = -1} f(e^i).$$

If S is the set of e^i such that $f(e^i) = -1$, this shows that

$$f(x) = \prod_{\substack{i \in S \\ x_i = -1}} (-1) = \prod_{i \in S} x_i.$$

In other words, $f(x)$ is linear if it is a *monomial*. Such functions are also known as *Fourier characters*, since they are characters of the group $\mathbb{Z}_2^n$.

What can we say about a function f which is *close* to linear, in the sense that $f(xy) = f(x)f(y)$ for most x, y? Concretely, suppose that

$$\Pr[f(xy) = f(x)f(y)] = 1 - \epsilon.$$

What can we say about f? Does it have to be close to a linear function? This is what we will show, using Fourier analysis.

The basic idea is to express the function f as a *mixture* of Fourier characters:

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

How do we know that such a representation exists? Is it unique?

First of all, notice that if every function can be represented in this way, then the representation must be unique: if we think of the space of functions on the Boolean cube $\{\pm 1\}^n$ as a vector space, then it has dimension 2^n, which exactly coincides with the number of Fourier characters.

There are many ways to show that every function can be represented in the form above, in other words, as a multilinear polynomial. For example, here is such a representation:

$$f(x) = \sum_{y \in \{\pm 1\}^n} f(y) \prod_{i=1}^n \frac{x_i y_i + 1}{2}.$$

The idea is that if $x_i = y_i$ then $x_i y_i = 1$, and otherwise $x_i y_i = -1$.

> The unique representation $f = \sum_{S \subseteq [n]} \hat{f}(S) x_S$ is known as the *Fourier expansion* of f, and the coefficients $\hat{f}(S)$ are known as the *Fourier coefficients* of f.

With this representation in hand, let us try to express the assumption in a different way. First, notice that $f(xy) = f(x)f(y)$ is the same as $f(x)f(y)f(xy) = 1$. Second, since $f(x)f(y)f(xy) \in \{\pm 1\}$, if we know

the probability that $f(x)f(y)f(xy) = 1$, then we also know the probability that $f(x)f(y)f(xy) = -1$, and so we can compute the expectation of $f(x)f(y)f(xy)$:

$$\mathbb{E}[f(x)f(y)f(xy)] = \Pr[f(x)f(y)f(xy) = 1] - \Pr[f(x)f(y)f(xy) = -1] = 1 - 2\epsilon.$$

At this point, we substitute the Fourier expansion of f and apply linearity of expectation to obtain

$$1 - 2\epsilon = \sum_{S,T,U} \hat{f}(S)\hat{f}(T)\hat{f}(U)\,\mathbb{E}[x_S y_T (xy)_U].$$

Notice that $(xy)_U = x_U y_U$. Moreover, $x_S x_U = x_{S \Delta U}$, where Δ is symmetric difference. This is because $x_i^2 = 1$. Altogether,

$$1 - 2\epsilon = \sum_{S,T,U} \hat{f}(S)\hat{f}(T)\hat{f}(U)\,\mathbb{E}[x_{S\Delta U}]\,\mathbb{E}[y_{T\Delta U}],$$

since x, y are independent.

What is the expectation of x_R? If $R = \emptyset$ then $x_R = 1$, and so $\mathbb{E}[x_R] = 1$. Otherwise,

$$\mathbb{E}[x_R] = \prod_{i \in R} \mathbb{E}[x_i] = 0,$$

since each x_i has zero mean. We conclude that

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

Before continuing, let us pause to notice something that came up during the calculation: $\mathbb{E}[x_S x_T] = 0$ if $S \neq T$, and $\mathbb{E}[x_S^2] = \mathbb{E}[1] = 1$. In other words, the Fourier characters x_S form an orthonormal basis for the space of real-valued functions on the Boolean cube. This gives us a way to compute the Fourier coefficients of a function h:

$$\mathbb{E}[h x_S] = \sum_T \hat{h}(T)\,\mathbb{E}[x_T x_S] = \hat{h}(S).$$

More generally,

$$\mathbb{E}[gh] = \sum_{S,T} \hat{g}(S)\hat{h}(T)\,\mathbb{E}[x_S x_T] = \sum_S \hat{g}(S)\hat{h}(S).$$

In particular,

$$\mathbb{E}[h^2] = \sum_S \hat{h}(S)^2.$$

This is known as *Parseval's identity*. The left-hand side is often written as $\|h\|^2$, since it is the square of the L_2 norm of h.

Back to our function f, which satisfies $f^2 = 1$, and so

$$\sum_S \hat{f}(S)^2 = 1.$$

Combining this with the preceding equation, we get

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

In other words, some Fourier coefficient $\hat{f}(S)$ must be close to 1 in value!

Our earlier formula for the Fourier coefficient shows that

$$1 - 2\epsilon \leq \hat{f}(S) = \mathbb{E}[f x_S].$$

Since both f and x_S are ± 1-valued,

$$\mathbb{E}[fx_S] = \Pr[fx_S = 1] - \Pr[fx_S = -1] = \Pr[f = x_S] - \Pr[f \neq x_S] = 2\Pr[f = x_S] - 1.$$

In other words,

$$\Pr[f = x_S] \geq 1 - \epsilon.$$

To conclude, we have shown that if $f(x)f(y) = f(xy)$ with probability $1 - \epsilon$, then there exists a set S such that $f = x_S$ with probability at least $1 - \epsilon$.

Property testing view Linearity testing is often viewed from the perspective of property testing. In this view, we are given a function $f\colon \{\pm 1\}^n \to \{\pm 1\}$ as a black box, and our goal is to find out whether f is a Fourier character or not, by sampling only few values of f.

What properties can we require of such a test? It is natural to require a Fourier character to always pass the test, and this is known as *perfect completeness*.

What if f is not a Fourier character? If f is very close to a Fourier character, say results from a Fourier character by changing only a few entries, no test that samples only a few values of f will be able to tell the difference. Hence the most we can say is that if f passes the test then it is probably close to a Fourier character. More accurately, the *soundness* guarantee is that if f passes the test with probability close to 1, then f is close to a Fourier character.

We can view the above as analyzing the following natural test: choose x, y at random, query f at locations x, y, xy, and check whether $f(x)f(y) = f(xy)$. If f is a Fourier character then the test always passes (perfect completeness). Conversely, if the test passes with probability $1 - \epsilon$, then f is ϵ-close to a Fourier character (soundness), meaning that $\Pr[f \neq x_S] \leq \epsilon$ for some Fourier character x_S.

List decoding regime A random function f satisfies $f(x)f(y) = f(xy)$ with probability very close to $1/2$. Intuitively, this is because for any given x, y, the probability that $f(x)f(y) = f(xy)$ is exactly $1/2$. (To formalize this argument, we need to show that $\Pr[f(x)f(y) = f(xy)]$ is concentrated around its mean $1/2$, which we can do using Chebyshev's inequality.) Therefore if a function satisfies $f(x)f(y) = f(xy)$ with probability noticeably larger than $1/2$, the function is not random. What can we say about such functions?

Our argument above actually shows that

$$\max_S \Pr[f = x_S] \geq \Pr[f(x)f(y) = f(xy)].$$

Therefore if $f(x)f(y) = f(xy)$ happens more often than in a random function, this implies that f has non-trivial correlation with some Fourier character.

Exercise Repeat the analysis above, replacing the test $f(x)f(y) = f(xy)$ with the test $f(x)f(y)f(z) = f(xyz)$.

2 Polymorphisms of majority

Here is another way of viewing Fourier characters: they are *polymorphisms* of the predicate $P(x, y, z)$ on $\{\pm 1\}^3$ which holds when $xyz = 1$. A function $f\colon D^n \to D$ is a polymorphism of a predicate $P \subseteq D^m$ if whenever vectors $x^1, \ldots, x^n$ satisfy the predicate P, then so does the vector $f(x^1_1, \ldots, x^n_1), \ldots, f(x^1_m, \ldots, x^n_m)$. Pictorially, we can think of an $n \times m$ table:

$$\begin{array}{ccc} x^1_1 & \cdots & x^1_m \\ x^2_1 & \cdots & x^2_m \\ \vdots & \ddots & \vdots \\ x^n_1 & \cdots & x^n_m \\ \hline f(x^1_1, \ldots, x^n_1) & \cdots & f(x^1_m, \ldots, x^n_m) \end{array}$$

Here the last row is generated from the rest of the table by applying f columnwise. The guarantee is that if all original rows satisfy P, then so does the final one.

A particular case of interest is when the predicate P is *truth-functional*, that is, arises from a function $\phi\colon D^{m-1} \to D$. In this case, $P(x_1, \ldots, x_m)$ holds if $x_m = \phi(x_1, \ldots, x_{m-1})$. The predicate P considered above arises from the function $\phi = xy$, which becomes the XOR function if we switch from ± 1 to $0, 1$.

Today we would like to consider the predicate arising from the majority function $\mathrm{MAJ}\colon \{\pm 1\}^3 \to \{\pm 1\}$. A function $f\colon \{\pm 1\}^n \to \{\pm 1\}$ is a polymorphism of MAJ if for all $x, y, z \in \{\pm 1\}^n$,

$$\mathrm{MAJ}(f(x), f(y), f(z)) = f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n)). \tag{1}$$

In contrast to the analysis in Section 1, in this case we won't get far if we just try to substitute the Fourier expansion in all places. Instead, we will fix x and average over y, z.

To see what we get on the left-hand side, we need to compute the Fourier expansion of MAJ. The easiest way to do this is using the formula from Section 1:

$$\widehat{\mathrm{MAJ}}(\emptyset) = \mathbb{E}[\mathrm{MAJ}] = 0,$$
$$\widehat{\mathrm{MAJ}}(\{1\}) = \mathbb{E}[\mathrm{MAJ}\, x_1] = \frac{1}{2}\,\mathbb{E}[\mathrm{MAJ}\,|x_1 = 1] - \frac{1}{2}\,\mathbb{E}[\mathrm{MAJ}\,|x_1 = -1] = \frac{1}{2}\left(\frac{3}{4} - \frac{1}{4}\right) - \frac{1}{2}\left(-\frac{3}{4} + \frac{1}{4}\right) = \frac{1}{2},$$
$$\widehat{\mathrm{MAJ}}(\{1,2,3\}) = \mathbb{E}[\mathrm{MAJ}\, x_1x_2x_3] = \frac{1 - 3 - 3 + 1}{8} = -\frac{1}{2}.$$

What about $\widehat{\mathrm{MAJ}}(\{1,2\}) = \mathbb{E}[\mathrm{MAJ}\, x_1x_2]$? Since MAJ is an odd function, that is $\mathrm{MAJ}(-x_1, -x_2, -x_3) = -\mathrm{MAJ}(x_1, x_2, x_3)$, we have

$$\mathbb{E}[\mathrm{MAJ}(x_1, x_2, x_3)x_1x_2] = \mathbb{E}[-\mathrm{MAJ}(-x_1, -x_2, -x_3)x_1x_2)] = -\mathbb{E}[\mathrm{MAJ}(y_1, y_2, y_3)(-y_1)(-y_2)],$$

where $y_i = -x_i$. Since $(-y_1)(-y_2) = y_1y_2$, we conclude that $\mathbb{E}[\mathrm{MAJ}(x_1, x_2, x_3)x_1x_2] = 0$. Altogether,

$$\mathrm{MAJ}(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3 - x_1x_2x_3}{2}.$$

Plugging this Fourier expansion and taking expectation over y, z, the left-hand side of (1) becomes

$$\frac{f(x) + 2\,\mathbb{E}[f] - f(x)\,\mathbb{E}[f]^2}{2} = \frac{1 - \mathbb{E}[f]^2}{2} f(x) + \mathbb{E}[f].$$

Now let us turn to the right-hand side of (1):

$$\mathop{\mathbb{E}}_{y,z}[f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n)] = \sum_S \hat{f}(S) \prod_{i \in S} \mathop{\mathbb{E}}_{y_i, z_i}[\mathrm{MAJ}(x_i, y_i, z_i)].$$

If $x_i = 1$, then $\mathrm{MAJ}(x_i, y_i, z_i) = 1$ with probability $3/4$, and so $\mathbb{E}[\mathrm{MAJ}(x_i, y_i, z_i)] = 1/2$. Similarly, if $x_i = -1$ then the expectation equals $-1/2$, and so we can say that it equals $x_i/2$. Therefore the right-hand side of (1) equals

$$\sum_S \hat{f}(S) \prod_{i \in S} \frac{x_i}{2} = \sum_S \left(\frac{1}{2}\right)^{|S|} \hat{f}(S) x_S.$$

This function is usually denoted $T_{1/2}f$, where T_ρ is the *noise operator*, which multiplies the S-th Fourier coefficient by $\rho^{|S|}$.

This sort of dependence on $|S|$ is very common in Fourier analysis, and it suggests decomposing the Fourier expansion into *levels* according to the size of S:

$$f = \sum_{d=0}^{n} \sum_{|S|=d} \hat{f}(S) x_S.$$

The d'th level of the Fourier expansion consists of coefficients $\hat{f}(S)$ with $|S| = d$. We often use the notation $f^{=d}$ for the sum above:

$$f^{=d} = \sum_{|S|=d} \hat{f}(S) x_S.$$

Using this notation, we can express the noise operator more succinctly:

$$T_\rho f = \sum_{d=0}^{n} \rho^d f^{=d}.$$

For future reference, let us give another interpretation of $T_\rho f$. On input x, $T_\rho f(x) = \mathbb{E}[f(w)]$, where $w_i = x_i$ with probability $\frac{1+\rho}{2}$ and $w_i = -x_i$ with probability $\frac{1-\rho}{2}$ (this interpretation makes sense as long as $|\rho| \leq 1$). Indeed,

$$\mathbb{E}[w_i] = \frac{1+\rho}{2} x_i - \frac{1-\rho}{2} x_i = \rho x_i,$$

and so this generalizes the case of $\mathrm{MAJ}(x_i, y_i, z_i)$.

Back to (1), which we have shown to be equivalent to

$$\frac{1 - \mathbb{E}[f]^2}{2} f(x) + \mathbb{E}[f] = T_{1/2} f(x).$$

This equation holds for any x, and so we can think of it as an identity of functions. Replacing each side with its Fourier expansion, we obtain

$$\frac{1 - \mathbb{E}[f]^2}{2} \hat{f}(\emptyset) + \mathbb{E}[f] + \sum_{S \neq \emptyset} \frac{1 - \mathbb{E}[f]^2}{2} \hat{f}(S) x_S = \sum_S \frac{1}{2^{|S|}} \hat{f}(S) x_S.$$

Since the Fourier expansion is unique, we can compare coefficients on both sides. In particular, the free coefficients must be equal:

$$\frac{1 - \mathbb{E}[f]^2}{2} \hat{f}(\emptyset) + \mathbb{E}[f] = \hat{f}(\emptyset).$$

This is a good point to mention that $\mathbb{E}[f] = \mathbb{E}[f x_\emptyset] = \hat{f}(\emptyset)$, which allows us to simplify the equation:

$$\frac{1 - \mathbb{E}[f]^2}{2} \mathbb{E}[f] + \mathbb{E}[f] = \mathbb{E}[f].$$

Therefore either $\mathbb{E}[f] = 0$, or else $\mathbb{E}[f]^2 = 1$. In the latter case, $\mathbb{E}[f] = \pm 1$, and so $f = \pm 1$ is constant.

The more interesting case is when $\mathbb{E}[f] = 0$. The Fourier expansions simplify to

$$\frac{1}{2} \sum_S \hat{f}(S) x_S = \sum_S \frac{1}{2^{|S|}} \hat{f}(S) x_S.$$

This shows that $\hat{f}(S) \neq 0$ only if S belongs to the first level, and so f has the form

$$f = \sum_{i=1}^{n} c_i x_i.$$

We say that f has *degree* 1, since the largest non-zero Fourier coefficient is on level 1, and moreover f is *homogeneous*, since all non-zero Fourier coefficients belong to the same level.

What does f look like? We claim that at most *one* coefficient c_i is non-zero. Indeed, suppose that $c_1, c_2 \neq 0$, and assume without loss of generality that $c_1, c_2 > 0$. Then

$$f(1, 1, 1, \ldots, 1) > f(1, -1, 1, \ldots, 1) > f(-1, -1, 1, \ldots, 1),$$

which is impossible since f is ± 1-valued. Thus $f = c_i x_i$. Since f is Boolean, in fact $f = \pm x_i$.

Concluding, we have shown that if f is a polymorphism of MAJ, then f is either a constant or of the form $\pm x_i$. We call functions of the form $\pm x_i$ *dictators*, since they are dictated by the i'th coordinate of the input. (Sometimes only x_i is called a dictator, and $-x_i$ is called an anti-dictator.)

Exercise Determine all Boolean functions (that is, functions $f\colon \{\pm 1\}^n \to \{\pm 1\}$) which have degree at most 1.

Exercise Find the polymorphisms of the predicate NAE on $\{\pm 1\}^3$, which holds when the inputs are not all equal.

2.1 Approximate polymorphisms

Exact polymorphisms of XOR are Fourier characters. Section 1 shows something stronger: if $f\colon \{\pm 1\}^n \to \{\pm 1\}$ is an *approximate polymorphism* of XOR, meaning that the polymorphism property holds for most tables, that f is close to some Fourier character.

What happens in the case of MAJ? Suppose that $f\colon \{\pm 1\}^n \to \{\pm 1\}$ satisfies

$$\Pr[\mathrm{MAJ}(f(x), f(y), f(z)) = f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n))] = 1 - \epsilon.$$

Is it true that f must be close to a constant or to a dictatorship?

We would like to repeat the analysis we had before. Let us try to see which parts of it we can salvage. If we denote the left-hand side by $g(x, y, z)$ and the right-hand side by $h(x, y, z)$, then we are given that $\Pr[g = h] = 1 - \epsilon$. Above we have calculated

$$\begin{aligned} G(x) &:= \mathop{\mathbb{E}}_{y,z}[g(x, y, z)] = \frac{1 - \mathbb{E}[f]^2}{2} f + \mathbb{E}[f], \\ H(x) &:= \mathop{\mathbb{E}}_{y,z}[h(x, y, z)] = T_{1/2} f, \end{aligned}$$

and then compared the Fourier expansions of G and H.

If $G = H$ then the Fourier expansions of G and H must be equal. In our case, $G \approx H$, a notion that we will have to formalize. What do we require of this notion? We need it to follow from the assumption $\Pr[g = h] = 1 - \epsilon$, and we want it to imply something about the Fourier expansions of G and H.

It turns out that the correct way to formalize $G \approx H$ is by considering $\|G - H\|^2 = \mathbb{E}[(G - H)^2]$. Indeed, Parseval's identity shows that

$$\|G - H\|^2 = \sum_S (\hat{G}(S) - \hat{H}(S))^2.$$

It remains to bound $\|G - H\|^2$ using the promise on g, h. The first step is to "undo" the expectation over y, z:

$$\mathbb{E}[(G - H)^2] = \mathop{\mathbb{E}}_x \left[\left(\mathop{\mathbb{E}}_{y,z}[g(x, y, z) - h(x, y, z)] \right)^2 \right] \leq \mathop{\mathbb{E}}_{x,y,z}[(g(x, y, z) - h(x, y, z))^2].$$

Indeed, for every x, we can think of $g(x, y, z) - h(x, y, z)$ as a random variable R_x (corresponding to the experiment of choosing y, z uniformly at random). The inequality then reads

$$\mathop{\mathbb{E}}_x[\mathbb{E}[R_x]^2] \leq \mathop{\mathbb{E}}_x[\mathbb{E}[R_x^2]],$$

which follows from $\mathbb{E}[R_x]^2 \leq \mathbb{E}[R_x^2]$ (this basic inequality states that $\mathbb{V}[R_x] \geq 0$, or follows from convexity of t^2 by Jensen's inequality).

Now $(g(x, y, z) - h(x, y, z))^2$ equals 4 if $g(x, y, z) \neq h(x, y, z)$ and 0 if $g(x, y, z) = h(x, y, z)$, and so

$$\mathbb{E}[(G - H)^2] \leq 4 \Pr[g \neq h] = 4\epsilon.$$

Substituting the Fourier expansions of G and H, we conclude that

$$\left(\frac{1 - \mathbb{E}[f]^2}{2} \mathbb{E}[f] \right)^2 + \sum_{S \neq \emptyset} \left(\frac{1 - \mathbb{E}[f]^2}{2} - \frac{1}{2^{|S|}} \right)^2 \hat{f}(S)^2 \leq 4\epsilon.$$

We will now try to follow our steps in the analysis of the case $\epsilon = 0$. The first step was to consider the expectation. Using the triangle inequality,

$$\begin{aligned}|\mathop{\mathbb{E}}_{x,y,z}[\mathrm{MAJ}(f(x), f(y), f(z))] - \mathop{\mathbb{E}}_{x,y,z}[f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n))]| \leq \\ \mathop{\mathbb{E}}_{x,y,z}[|\,\mathrm{MAJ}(f(x), f(y), f(z)) - f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n))|] = \\ 2\Pr[\mathrm{MAJ}(f(x), f(y), f(z)) \neq f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n))] = 2\epsilon.\end{aligned}$$

Substituting the expressions for the expectations on the left-hand side,

$$\left|\frac{1 - \mathbb{E}[f]^2}{2}\,\mathbb{E}[f] + \mathbb{E}[f] - \mathbb{E}[f]\right| \leq 2\epsilon \Longrightarrow |\,\mathbb{E}[f] - 1| \cdot |\,\mathbb{E}[f] + 1| \cdot |\,\mathbb{E}[f]| \leq 4\epsilon.$$

Now suppose that among $\{0, \pm 1\}$, $\mathbb{E}[f]$ is closest to a. Then for any other $b \in \{0, \pm 1\}$, the distance from $\mathbb{E}[f]$ to b is at least $1/2$. Therefore $\mathbb{E}[f]$ is 16ϵ-close to some $a \in \{0, \pm 1\}$.

If $\mathbb{E}[f]$ is 16ϵ-close to $a \in \{\pm 1\}$ then $\Pr[f = a] = 1 - 8\epsilon$. Otherwise, $|\,\mathbb{E}[f]| = 16\epsilon$, and we concentrate on the rest of the sum:

$$\sum_{S \neq \emptyset} \left(\frac{1 - \mathbb{E}[f]^2}{2} - \frac{1}{2^{|S|}}\right)^2 \hat{f}(S)^2 \leq 4\epsilon.$$

If $|S| = 1$ then the coefficient $(1 - \mathbb{E}[f]^2)/2 - 1/2$ could be close to zero. But for larger $|S|$, on the one hand $(1 - \mathbb{E}[f]^2)/2 \geq 1/2 - 128\epsilon^2$, which is at least $1/3$ when $\epsilon \leq 1/100$, and on the other hand, $1/2^{|S|} \leq 1/4$. Therefore the coefficient is at least $1/144$, assuming that $\epsilon \leq 1/100$ (if $\epsilon \geq 1/100$ then f is trivially 100ϵ-close to any Boolean function, so this case is not interesting). Concentrating only on that part of the sum yields

$$\sum_{|S|>1} \hat{f}(S)^2 \leq 576\epsilon.$$

The sum on the left is the squared norm of the function $f^{>1}$ which consists of all levels of f beyond level 1. We are therefore led to the following question:

What can we say about Boolean functions f satisfying $\|f^{>1}\|^2 \leq \epsilon$?

We will answer this question next week.

3 Friedgut–Kalai–Naor theorem

Suppose that $f\colon \{\pm 1\}^n \to \{\pm 1\}$ satisfies $\|f^{>1}\|^2 \leq \epsilon$. What can we say about f? Does it have to be close to a constant or a dictator?

Let us rephrase this question in terms of $g = f^{\leq 1}$, that is,

$$g = \hat{f}(\emptyset) + \sum_{i=1}^{n} \hat{f}(\{i\})x_i.$$

This is the *orthogonal projection* of f to the space of functions of degree 1 (by which we really mean the space of functions of degree *at most* 1), which means that it is the degree 1 function minimizing $\|f - g\|^2$. This is because the Fourier characters form an orthonormal basis.

What can we say about the function g? It is close to f, in the sense that $\|g - f\|^2 \leq \epsilon$. In particular, using the notation

$$\mathrm{dist}(y, S) = \min_{z \in S} |y - z|,$$

since f is ± 1-valued, we deduce that

$$\mathbb{E}[\mathrm{dist}(g, \{\pm 1\})^2] \leq \mathbb{E}[(g-f)^2] \leq \epsilon.$$

Our working hypothesis is that f should be close to some Boolean function r, which is a constant or a dictator, in the sense that $\Pr[f \neq r] \leq \delta$, where δ depends on ϵ. If this is the case, then we expect g to be close to r as well. Indeed,

$$\|g-r\|^2 = \mathbb{E}[(g-r)^2] \leq \mathbb{E}[2(g-f)^2 + 2(f-r)^2] \leq 2\|g-f\|^2 + 8\Pr[f \neq r] = O(\epsilon + \delta).$$

Here we used two useful facts: the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ (a special case of Cauchy–Schwarz), and $\|f-r\|^2 = 4\Pr[f \neq r]$ (since $(f-r)^2 \in \{0, 4\}$).

Conversely, suppose that we show that $\|g-r\|^2 \leq \eta$ for some Boolean function r which is either a constant or a dictator. Then an identical argument shows that $\|f-r\|^2 = O(\epsilon + \eta)$, and so $\Pr[f \neq r] = O(\epsilon + \eta)$.

This shows that the following questions are essentially equivalent:

1. Understanding the structure of Boolean functions f such that $\|f^{>1}\|^2$ is small.
2. Understanding the structure of degree 1 functions g such that $\mathbb{E}[\mathrm{dist}(g, \{\pm 1\})^2]$ is small.

We will focus on the second.

Getting rid of the constant term Our goal is to show that if $\deg g \leq 1$ and $\mathbb{E}[\mathrm{dist}(g, \{\pm 1\})^2] = \epsilon$, then g is close to a constant or a dictator. Since $\deg g \leq 1$, we can represent it as

$$g = c_0 + \sum_{i=1}^{n} c_i x_i.$$

It will be convenient to get rid of the constant coefficient, by considering the function

$$h = \sum_{i=0}^{n} c_i x_i.$$

This function satisfies

$$\begin{aligned} h(+1, x_1, \ldots, x_n) &= g(x_1, \ldots, x_n), \\ h(-1, x_1, \ldots, x_n) &= -g(-x_1, \ldots, -x_n). \end{aligned}$$

This shows that $\mathbb{E}[\mathrm{dist}(h, \{\pm 1\})^2] = \mathbb{E}[\mathrm{dist}(g, \{\pm 1\})^2]$, hence it suffices to understand functions like h. We will show that h must be close to a dictator, and it will follow that g must be close to a constant or a dictator.

Getting rid of large coefficients If h is close to a dictator r, say $r = \pm x_0$, then this would mean that $\|h-r\|^2$ is small, and so the following is small:

$$(\pm 1 - c_0)^2 + \sum_{i=1}^{n} c_i^2.$$

In other words, one of the coefficients is close to ± 1, and the other are close to zero. It turns out that it is easy to show this for *individual* coefficients; most of the effort would be to show that this holds *in aggregate*.

Consider the coefficient c_0. We know that

$$\mathop{\mathbb{E}}_{x_1, \ldots, x_n} \mathop{\mathbb{E}}_{x_0} [\mathrm{dist}(h, \{\pm 1\})^2] \leq \epsilon,$$

hence there is a choice of $x_1, \ldots, x_n$ such that

$$\mathrm{dist}(H + c_0, \{\pm 1\})^2 + \mathrm{dist}(H - c_0, \{\pm 1\})^2 \leq 2\epsilon, \quad \text{where } H = \sum_{i=1}^{n} c_i x_i.$$

Suppose without loss of generality that $H + c_0$ is closer to $+1$, say $H + c_0 = 1 + \gamma$, where $\gamma^2 \leq 2\epsilon$. There are now two cases to consider. First, suppose that $H - c_0$ is also closer to $+1$. Since $H - c_0 = 1 + \gamma - 2c_0$, we have $(2c_0 - \gamma)^2 \leq 2\epsilon$, and so

$$(2c_0)^2 \leq 2\gamma^2 + 2(2c_0 - \gamma)^2 = O(\epsilon),$$

and so $c_0^2 = O(\epsilon)$. If $H - c_0$ is close to -1, then $(2c_0 - \gamma - 2)^2 \leq \epsilon$, and so

$$(2c_0 - 2)^2 \leq 2\gamma^2 + 2(2c_0 - \gamma - 2)^2 = O(\epsilon),$$

hence $(c_0 - 1)^2 = O(\epsilon)$. If $H + c_0$ were close to -1, then we would also have the option $(c_0 + 1)^2 = O(\epsilon)$.

So far we have shown that each individual coefficient is somewhat close to $\{0, \pm 1\}$. Next, we will show that there cannot be two coefficients which are "large", using an argument similar to how we showed that Boolean degree 1 functions are constants or dictators.

For this part of the argument, we will need ϵ to be "small enough", that is, smaller than some absolute constant. This is an assumption which we commonly make, since usually structure results are trivial when ϵ is large. For example, suppose that we aim to conclude, eventually, that h is $O(\epsilon)$-close to a dictator. If $\epsilon \geq 1/100$ then this is automatically satisfied for *any* dictator, by choosing the big O constant appropriately. Indeed, if r is any dictator and $\mathrm{round}(h, \{\pm 1\})$ results from rounding h to the nearest value in $\{\pm 1\}$, then

$$\|h - r\|^2 = \mathbb{E}[(h - r)^2] \leq 2\,\mathbb{E}[(h - \mathrm{round}(h, \{\pm 1\}))^2] + 2\,\mathbb{E}[(\mathrm{round}(h, \{\pm 1\}) - r)^2] \leq 2\epsilon + 2 \leq 202\epsilon.$$

Suppose that c_0, c_1 are both large, say $(c_0 - 1)^2, (c_1 - 1)^2 = O(\epsilon)$. We know that

$$\mathop{\mathbb{E}}_{x_2,\ldots,x_n} \mathop{\mathbb{E}}_{x_0,x_1} [\mathrm{dist}(h, \{\pm 1\})^2] \leq \epsilon,$$

hence there is a choice of $x_2, \ldots, x_n$ such that

$$\mathrm{dist}(H + c_0 + c_1, \{\pm 1\})^2 + \mathrm{dist}(H - c_0 - c_1, \{\pm 1\})^2 \leq 4\epsilon.$$

(We removed two terms.) The idea now is that since $c_0 + c_1 \approx 2$, it is impossible for $H + (c_0 + c_1)$ and $H - (c_0 + c_1)$ to both be close to $\{\pm 1\}$, assuming ϵ is small enough.

Formally, suppose that $H + c_0 + c_1$ is closer to $a \in \{\pm 1\}$ and that $H - c_0 - c_1$ is closer to $b \in \{\pm 1\}$. Then

$$(c_0 + c_1 - a - (-c_0 - c_1 - b))^2 \leq 2(H + c_0 + c_1 - a)^2 + 2(-(H - c_0 - c_1 - b))^2 \leq 8\epsilon.$$

This implies that

$$(c_0 + c_1 - (a + b)/2)^2 \leq 2\epsilon.$$

On the other hand,

$$(c_0 + c_1 - 2)^2 \leq 2(c_0 - 1)^2 + 2(c_1 - 1)^2 \leq 4\epsilon.$$

Altogether, this shows that $((a + b)/2 - 2)^2 \leq 12\epsilon$. Since $(a + b)/2 \leq 1$, this is impossible as long as $\epsilon < 1/12$.

Summarizing, assuming that $\epsilon < 1/12$, at most one of the c_i can be "large". If such a coefficient exists, let us assume that it is c_0. We are now at the following situation: we know that c_0 is close to $C \in \{0, \pm 1\}$, and that $c_1, \ldots, c_n$ are each *individually* close to 0. This suggests that h is close to Cx_0. However,

$$\|H - Cx_0\|^2 = (c_0 - C)^2 + \sum_{i=1}^{n} c_i^2.$$

We would like to show that the right-hand side is small. So far, all we know is that each *particular* summand is $O(\epsilon)$, but it doesn't follow that the sum itself is small, since there are $n + 1$ many summands! The main part of the proof is to show that the coefficients are close to $\{0, \pm 1\}$ *in aggregate*.

Main argument The intuition here is that since the coefficients $c_1, \ldots, c_n$ are small, the distribution of $\sum_i c_i x_i$ is close to a normal distribution with zero mean and variance $\sum_i c_i^2$. On the other hand, we know that $\sum_i c_i x_i$ must be close to $\{\pm 1\} - c_0$. Since a normal distribution is "smooth", this can only happen if $\sum_i c_i x_i$ is concentrated around *one* of the values $\{\pm 1\} - c_0$, and in particular, $\sum_i c_i^2$ must be small.

Arguing this formally requires some work. We will do so by induction. Under the assumption that $c_1^2, \ldots, c_n^2 \leq K\epsilon$ (which is what we get from the preceding step), we will show, by induction on m, that in fact

$$\sum_{i=1}^{m} c_i^2 \leq K\epsilon.$$

The base case $m = 1$ is trivial, so let us assume that $\sum_{i=1}^{m} c_i^2 \leq K\epsilon$, and show that $\sum_{i=1}^{m+1} c_i^2 \leq K\epsilon$, assuming that $c_{i+1}^2 \leq K\epsilon$ and that ϵ is small enough.

First of all, we note that there is a setting of $x_0, x_{m+2}, \ldots, x_n$ for which

$$\mathbb{E}\left[\operatorname{dist}\left(H + \sum_{i=1}^{m+1} c_i x_i, \{\pm 1\}\right)^2\right] \leq \epsilon, \quad H = c_0 x_0 + \sum_{i=m+2}^{n} c_i x_i.$$

The remaining coefficients $c_1, \ldots, c_{m+1}$ satisfy

$$\sum_{i=1}^{m+1} c_i^2 \leq 2K\epsilon.$$

Let us now eliminate H. We can write

$$2\sum_{i=1}^{m+1} c_i x_i = \left(H + \sum_{i=1}^{m+1} c_i x_i\right) - \left(H + \sum_{i=1}^{m+1} c_i(-x_i)\right).$$

Denoting by $a, b \in \{\pm 1\}$ the values that the two expressions on the right are closest to, this gives

$$\left(2\sum_{i=1}^{m+1} c_i x_i - (a-b)\right)^2 \leq 2\left(H + \sum_{i=1}^{m+1} c_i x_i - a\right)^2 + 2\left(H + \sum_{i=1}^{m+1} c_i(-x_i) - b\right)^2.$$

Taking expectation and dividing by 4, this shows that

$$\mathbb{E}\left[\operatorname{dist}\left(\sum_{i=1}^{m+1} c_i x_i, \{0, \pm 1\}\right)^2\right] \leq \epsilon,$$

since $(a-b)/2 \in \{0, \pm 1\}$.

Since the variance of $\sum_{i=1}^{m+1} c_i x_i$ is small, this sum is concentrated around its mean, which is zero. Hence we expect that most of the time, $\sum_{i=1}^{m+1} c_i x_i$ would be closest to 0, and this would imply that

$$\sum_{i=1}^{n} c_i^2 = \mathbb{E}\left[\left(\sum_{i=1}^{n} c_i x_i\right)^2\right] \approx \mathbb{E}\left[\operatorname{dist}\left(\sum_{i=1}^{m+1} c_i x_i, \{0, \pm 1\}\right)^2\right] \leq \epsilon.$$

In order to argue this formally, let us notice that if $s := \sum_{i=1}^{m+1} c_i x_i$ is not closest to 0, then $|s| \geq 1/2$, and so $s^2 \leq 4s^4$ (we will see below why this is useful). This shows that

$$\left(\sum_{i=1}^{m+1} c_i x_i\right)^2 \leq \operatorname{dist}\left(\sum_{i=1}^{m+1} c_i x_i, \{0, \pm 1\}\right)^2 + 4\left(\sum_{i=1}^{m+1} c_i x_i\right)^4.$$

Indeed, if the sum is closest to 0 then the term on the left equals the first term on the right, and otherwise it is bounded by the second term on the right. We conclude that

$$\sum_{i=1}^{m+1} c_i^2 = \mathbb{E}\left[\left(\sum_{i=1}^{m+1} c_i x_i\right)^2\right] \leq \mathbb{E}\left[\operatorname{dist}\left(\sum_{i=1}^{m+1} c_i x_i, \{0, \pm 1\}\right)^2\right] + 4\,\mathbb{E}\left[\left(\sum_{i=1}^{m+1} c_i x_i\right)^4\right] \leq \epsilon + 4\,\mathbb{E}\left[\left(\sum_{i=1}^{m+1} c_i x_i\right)^4\right].$$

What do we do about the second term on the right? It is four times the expectation of

$$\sum_{i,j,k,\ell} c_i c_j c_k c_\ell x_i x_j x_k x_\ell.$$

Most of the terms here vanish: indeed, if $i \neq j, k, \ell$ then $\mathbb{E}[x_i x_j x_k x_\ell] = \mathbb{E}[x_i]\,\mathbb{E}[x_j x_k x_\ell] = 0$. There are two types of terms that survive: $i = j \neq k = \ell$ (and their two permutations), and $i = j = k = \ell$. Since $\mathbb{E}[x_i^2 x_k^2] = \mathbb{E}[x_i^4] = 1$, we can bound

$$\mathbb{E}\left[\left(\sum_{i=1}^{m+1} c_i x_i\right)^4\right] \leq 3\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} c_i^2 c_j^2 + \sum_{i=1}^{m+1} c_i^4 \leq 3\left(\sum_{i=1}^{m+1} c_i^2\right)^2 + K\epsilon \sum_{i=1}^{m+1} c_i^2,$$

since $c_1^2, \ldots, c_{m+1}^2 \leq K\epsilon$. By assumption, $\sum_i c_i^2 \leq 2K\epsilon$, and so putting everything together, we conclude that

$$\sum_{i=1}^{m+1} c_i^2 \leq \epsilon + 4 \cdot [3 \cdot (2K\epsilon)^2 + (K\epsilon) \cdot (2K\epsilon)] = \epsilon + 56K\epsilon^2.$$

We would like this to be at most $K\epsilon$, assuming that ϵ is small enough. Possibly increasing K so that it is at least 2, it suffices to assume that $\epsilon \leq 1/(56K)$ to make the inductive step go through.

Concluding the argument Our inductive proof shows that

$$\sum_{i=1}^{n} c_i^2 = O(\epsilon).$$

We also know that $(c_0 - C)^2 = O(\epsilon)$, where $C \in \{0, \pm 1\}$. This shows that

$$\|h - Cx_0\|^2 = (c_0 - C)^2 + \sum_{i=1}^{n} c_i^2 = O(\epsilon).$$

It remains to show that $C \neq 0$. Indeed,

$$\operatorname{dist}(C, \{\pm 1\})^2 = \mathbb{E}[\operatorname{dist}(Cx_0, \{\pm 1\})^2] \leq 2\,\mathbb{E}[(h - Cx_0)^2] + 2\,\mathbb{E}[\operatorname{dist}(h, \{\pm 1\})^2] = O(\epsilon),$$

which for small enough ϵ implies that $C \neq 0$.

Altogether, we have proved the following results, due to Friedgut, Kalai and Naor [FKN02]:

> If $g\colon \{\pm 1\}^n \to \mathbb{R}$ is a degree 1 function satisfying $\mathbb{E}[\operatorname{dist}(g, \{\pm 1\})^2] \leq \epsilon$, then there is a Boolean function $r\colon \{\pm 1\}^n \to \{\pm 1\}$, which depends on at most one input, such that $\|g - r\|^2 = O(\epsilon)$.

> If $f\colon \{\pm 1\}^n \to \{\pm 1\}$ satisfies $\|f^{>1}\|^2 \leq \epsilon$ then there is a Boolean function $r\colon \{\pm 1\}^n \to \{\pm 1\}$, which depends on at most one input, such that $\Pr[f \neq r] = O(\epsilon)$.

Together with the results of Section 2.1, we have completed the proof of the following result:

If $f\colon \{\pm 1\}^n \to \{\pm 1\}$ satisfies

$$\Pr_{x,y,z\in\{\pm1\}^n}[\mathrm{MAJ}(f(x), f(y), f(z)) = f(\mathrm{MAJ}(x_1, y_1, z_1), \ldots, \mathrm{MAJ}(x_n, y_n, z_n))] \geq 1 - \epsilon$$

then there exists a function $r\colon \{\pm 1\}^n \to \{\pm 1\}$, depending on at most one input, such that $\Pr[f \neq r] = O(\epsilon)$.

Exercise Find all functions $f\colon \{\pm 1\}^n \to \{0, \pm 1\}$ such that $\deg f \leq 1$. Then characterize all functions $f\colon \{\pm 1\}^n \to \{0, \pm 1\}$ such that $\|f^{>1}\|^2 \leq \epsilon$.

4 Voting and influences

Consider an election between two candidate, -1 and 1. There are n voters, and each one votes either -1 or 1. The outcome of the election is given by a function $f\colon \{\pm 1\}^n \to \{\pm 1\}$. In many places f is the majority function, but sometimes more sophisticated functions are used, for example in the United States. It is natural to assume that f is *monotone*, that is, if a voter changes their vote from -1 to 1, then the outcome cannot change from 1 to -1.

Suppose furthermore that each voter independently tosses a fair coin, a simplifying assumption which is not too far from reality. How many votes need to be "bought" in order to force the outcome, with probability $2/3$, say?

In the case of majority, it suffices to bribe $\Theta(\sqrt{n})$ voters. To see this, denote the votes by $x_1, \ldots, x_n$, and note that by the central limit theorem, $x_1 + \cdots + x_n$ has roughly Gaussian distribution with zero mean and standard deviation $\sqrt{n}$. If we bribe the last $C\sqrt{n}$ voters to vote 1, then $x_1 + \cdots + x_n$ has roughly Gaussian distribution with mean $C\sqrt{n}$ and standard deviation $\sqrt{n - C\sqrt{n}} \leq \sqrt{n}$. In particular, the probability that the sum is positive is roughly the probability that a standard Gaussian is at least $-C$, which tends to a positive constant; the constant is $1 - \Theta(e^{-C^2/2}/C)$, and in particular, it tends to 1 as $C \to \infty$.

Ajtai and Linial [AL93] constructed a function in which $\Omega(n/\log^2 n)$ voters need to be bribed. This is almost optimal, due to a fundamental result of Kahn, Kalai and Linial [KKL88], which shows that for any function f, there is a set of $O(n/\log n)$ voters whose bribing makes the outcome biased, in the sense that one of the candidates wins with probability $2/3$.

Kahn, Kalai and Linial choose which voters to bribe sequentially. The first voter to bribe is the one with the largest *influence* on the outcome of the election. For a Boolean function f, it is natural to define the i'th influence of f by

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

where $x^{(i)}$ results from negating the i'th coordinate.

The i'th influence is closely related to the Laplacian in direction i, which is given by

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

Indeed, if f is Boolean, then $|L_i f(x)|$ is the indicator of $f(x) \neq f(x^{(i)})$.

The effect of negating the i'th coordinate on a Fourier character x_S is easy to describe: if $i \notin S$ then the character stays the same, and otherwise it is negated. Therefore

$$L_i f = \frac{1}{2}\sum_S \hat{f}(S)x_S - \frac{1}{2}\sum_{i\notin S} \hat{f}(S)x_S + \frac{1}{2}\sum_{i\in S} \hat{f}(S)x_S = \sum_{i\in S} \hat{f}(S)x_S.$$

In particular, Parseval's identity shows that

$$\|L_i f\|^2 = \sum_{i\in S} \hat{f}(S)^2.$$

All of this makes sense even for non-Boolean f. When f is Boolean, $L_i f(x)^2$ is exactly the indicator of $f(x) \neq f(x^{(i)})$, since $L_i f \in \{0, \pm 1\}$. This shows that

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

We adopt this definition of influence even for non-Boolean f.

What happens if we bribe voter i to always vote 1? By how much does that increase the probability that the outcome is 1? We can choose a random x by first choosing all coordinates other than x_i, collectively known as x_{-i}, and then choosing i. If $f(x_{-i}, -1) = f(x_{-i}, 1)$, where the second argument is x_i, then bribing voter i makes no difference. Otherwise, since f is monotone, bribing voter i increases the probability of the outcome 1 from 1/2 to 1. Overall, this shows that

$$\Pr_{x_i = 1}[f(x) = 1] = \Pr[f(x) = 1] + \frac{1}{2}\mathrm{Inf}_i[f].$$

In the extreme case when $f = x_i$, we have $\mathrm{Inf}_i[f] = 1$, and indeed the probability of the outcome 1 increases from 1/2 to 1.

Before continuing with the Kahn–Kalai–Linial strategy, let us see what happens if we choose who to bribe *at random*. The effect depends on an important quantity known as the *total influence*:

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

Indeed, the calculation above shows that the output is skewed by $\mathrm{Inf}[f]/(2n)$. The total influence has a nice formula in terms of the Fourier coefficients:

$$\mathrm{Inf}[f] = \sum_{i=1}^{n} \mathrm{Inf}_i[f] = \sum_{i=1}^{n} \sum_{i \in S} \hat{f}(S)^2 = \sum_{S} |S| \hat{f}(S)^2 = \sum_{d=0}^{n} \|f^{=d}\|^2.$$

Total influence also has nice combinatorial interpretation as the *edge boundary*. Suppose that f is a Boolean function, which is the indicator function of some subset $A \subseteq \{\pm 1\}^n$ of the Boolean cube. The i'th influence $\mathrm{Inf}_i[f]$ measures the number of edges of the cube crossing from A to $\overline{A}$ in direction i, and so $\mathrm{Inf}[f]$ measures the total number of edges crossing from A to its complement.

A variant of this interpretation is *average sensitivity*. The sensitivity of f at a point x is the number of coordinates i such that $f(x) \neq f(x^{(i)})$. The average sensitivity of f is simply $\mathrm{Inf}[f]$.

Finally, total influence can also be defined in terms of the Laplacian of f, which is $Lf = \sum_i L_i f$: $\mathrm{Inf}[f] = \mathbb{E}[f \cdot Lf]$ (we prove this in Section 13.1).

An important inequality involving total influence is the Poincaré inequality: $\mathrm{Inf}[f] \geq \mathbb{V}[f]$. As a special case, if a Boolean function is balanced, then its average sensitivity is at least 1, which is tight for dictators. The Fourier formula for total influence immediately implies the Poincaré inequality, once we notice that

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

Indeed,

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

This bound is almost tight for low-degree functions: if f has degree d then

$$\mathrm{Inf}[f] = \sum_{S} |S| \hat{f}(S)^2 \leq d \hat{f}(S)^2 = d\,\mathbb{V}[f].$$

Poincaré's inequality implies that if we bribe a random voter to 1, then we bias the outcome by at least $\mathbb{V}[f]/(2n)$. This is tight for dictatorships, but far from tight in the case of majority. Indeed, a voter is

influential if the other votes split exactly evenly, which happens with probability $\Theta(1/\sqrt{n})$, which is much larger than $1/(2n)$.

Bribing a random voter is not a good strategy in general, as the case of a dictatorship demonstrates. In order to show that every election can be biased by bribing only $O(n/\log n)$ voters, we will show that every function f has a coordinate whose influence is $\Omega(\frac{\log n}{n}\mathbb{V}[f])$, a fundamental result known as the *KKL theorem*. This theorem is tight, as shown by the examples of the Tribes function:

$$\mathrm{Tribes}(x) = \bigvee_{i=1}^{n/m} \bigwedge_{j=1}^{m} x_{i,j}, \quad m = \log n - \log\log n,$$

where $\vee$ is the *max* operator and $\wedge$ is the *min* operator.

Let us check that Tribes is more or less balanced. Each of the n/m "tribes" evaluates to 1 with probability 2^{-m}, and so the function itself evaluates to -1 with probability

$$(1-2^{-m})^{n/m} \approx e^{-n/(2^m m)}.$$

Since $2^m m = (n/\log n)(\log n - \log\log n) \approx n$, this probability is roughly e^{-1}. The same calculation also allows us to estimate the influences of Tribes. For $x_{i,j}$ to be influential, we need all other tribes to evaluate to -1, which happens with probability roughly e^{-1}, and the other coordinates in the tribe to evaluate to 1, which happens with probability $2^{1-m} = 2\log n/n$. Overall, all influences are $O(\frac{\log n}{n})$.

Tribes is extremal from the point of view of the maximal influence, and dictators are extremal from the point of view of the average influence. Can we characterize functions which are extremal on either front? The answer in the case of maximal influence is not completely clear, but the answer in the case of average influence was worked out by Friedgut [Fri98]. Let us first make the question precise: What can we say about functions with average influence $O(1/n)$? Equivalently, what can we say about functions with total influence $O(1)$?

First, let us try to see which functions satisfy this. One example is constants and dictators. More generally, if a function depends on $O(1)$ coordinates, then its total influence is $O(1)$. Such a function is called a *junta*. Friedgut's junta theorem shows that every function with total influence $O(1)$ is close to a junta.

The proofs of the KKL theorem and of Friedgut's theorem are quite similar, but since the second one is more intuitive, we will start by proving Friedgut's theorem. Afterwards we will prove the KKL theorem, and finally, we will show how to use it to bias elections.

4.1 Bonus: L1 influences

We defined the influences of $f\colon \{\pm1\}^n \to \{\pm1\}$ as

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

and observed that

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

In fact, more is true:

$$\mathrm{Inf}_i[f] = \|L_i f\|_p^p = \mathbb{E}[|L_i f(x)|^p]$$

for all $p > 0$. We usually concentrate on the case $p = 2$ since it leads to a formula for $\mathrm{Inf}_i[f]$ in terms of Fourier coefficients.

When we consider non-Boolean functions, the choice of p does matter. Aaronson and Ambainis, in the conference version of their work [AA14], implicitly considered the case $p = 1$. They considered *bounded* functions $f\colon \{\pm1\}^n \to [-1,1]$, and implicitly assumed that

$$\sum_{i=1}^{n} \mathbb{E}[|L_i f|] \leq \deg(f),$$

an inequality which we saw holds for the usual influences, but isn't known to hold for these "L_1 influences". It turns out that the argument of Aaronson and Ambainis can be fixed to use the usual influences.

Bačkurs and Bavarian [BB14] showed that the "total L_1 influence" is $O(\deg(f)^3)$, and this was improved to $\deg(f)^2$ in [FHKL16], using approximation theory. Below we present the slightly weaker upper bound $2\deg(f)^2$. The best known lower bound is $\deg(f)$, achieved by Fourier characters, by other Boolean functions, and by some non-Boolean functions (see [FHKL16, Section 4]). Bačkurs and Bavarian conjecture that the true answer is $O(\deg(f))$; it might well be $\deg(f)$.

Suppose that $f\colon \{\pm1\}^n \to [-1,1]$. We will bound $\sum_{i=1}^n \mathbb{E}[|L_i f|]$ by showing that for *every* $x \in \{\pm1\}^n$,

$$\sum_{i=1}^n |L_i f(x)| \leq 2\deg(f)^2.$$

(This can be improved to $\deg(f)^2$, which is tight for Chebyshev polynomials.) Let's first get rid of the absolute values: it suffices to show that for all $x, y \in \{\pm1\}^n$,

$$\sum_{i=1}^n y_i L_i f(x) \leq 2\deg(f)^2.$$

If $S = \{i \in [n] : y_i = 1\}$, then the left-hand side is

$$\sum_{i\in S} L_i f(x) - \sum_{i\notin S} L_i f(x) = \sum_{i\in S} L_i f(x) + \sum_{i\notin S} L_i(-f)(x),$$

and so it suffices to show that for all $x \in \{\pm1\}^n$ and $S \subseteq [n]$,

$$\sum_{i\in S} L_i f(x) \leq \deg(f)^2.$$

We convert f into a *univariate* polynomial ϕ so that the left-hand side equals some derivative of ϕ. We choose

$$\phi(t) = f(tx|_S, x|_{\overline{S}}),$$

where the first part of the input corresponds to the coordinates in S, and the second part to the coordinates outside S. In other words,

$$\phi(t) = \sum_{T\subseteq[n]} t^{|S\cap T|}\hat{f}(T)x_T.$$

Therefore

$$\phi'(1) = \sum_{T\subseteq[n]} |S\cap T|\hat{f}(T)x_T = \sum_{i\in S} L_i f(x).$$

So $\phi'(1)$ is what we want to bound.

What do we know about ϕ? If $S = [n]$ then $\phi(t) = T_t f(x)$, and so for $t \in [-1,1]$, $\phi(t)$ is an average of values of f on $\{\pm1\}^n$. The same property holds for arbitrary S, and we conclude that $|\phi(t)| \leq 1$ when $|t| \leq 1$. Moreover, clearly $\deg(\phi) \leq \deg(f)$.

We are now left with the following problem: Given a degree d polynomial ϕ such that $|\phi(t)| \leq 1$ for all $|t| \leq 1$, how large can $\phi'(1)$ be? The answer, due to Bernstein and Markov, is d^2, which is achieved *uniquely* by the Chebyshev polynomial $T_d(x) = \cos(d\cos^{-1}(x))$.

5 Friedgut and KKL

5.1 Friedgut's junta theorem

Let f be a Boolean function with total influence I. Friedgut's junta theorem states that if I is small, then f is close to a junta, which is a function depending on a small number of coordinates. Which coordinates

belong to the junta? It is natural to conjecture that the junta J is composed of all influential coordinates, say J consists of all coordinates whose influence is at least τ (we will determine τ later on).

Once we have decided on the junta coordinates J, it is easy to construct the function itself: for every setting of the junta coordinates, we simply take the majority value, obtaining a Boolean function g. In order to understand how close f and g are, it will be useful to consider a third function h, which results from *averaging* f over all coordinates outside the junta, that is,

$$h(x_J, x_{-J}) = \mathop{\mathbb{E}}_{y_{-J}}[f(x_J, y_{-J})], \quad g(x_J, x_{-J}) = \mathrm{round}(h(x_J, x_{-J}), \{\pm 1\}).$$

It is easy to compute the Fourier expansion of h given that of f. Indeed, it is enough to see what happens to a Fourier character x_S. If all variables in S belong to the junta, then the character survives. Otherwise, the character is averaged out (since $\mathbb{E}[x_i] = 0$). Therefore

$$h = \sum_{S \subseteq J} \hat{f}(S) x_S,$$

which implies that

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

We will come back to this expression later, but first, let us see how to relate $\|f - h\|^2$ and $\Pr[f \neq g]$. The idea is very simple. We consider an arbitrary point x, and the three values $f(x), g(x), h(x)$. We know that on average $(f(x) - h(x))^2$ is small, and furthermore $f(x), g(x)$ are Boolean, and $g(x)$ is obtained from $h(x)$ by rounding. If $|h(x)| > 1$, then rounding actually brings $g(x)$ closer to $f(x)$. Otherwise, suppose that $0 \leq h(x) \leq 1$. If $f(x) = 1$ then $g(x)$ is again close to $f(x)$, and otherwise $|g(x) - f(x)| = 2$ whereas $|h(x) - f(x)| \geq 1$. This shows that $(g(x) - f(x))^2 \leq 4(h(x) - f(x))^2$, and so

$$\Pr[f \neq g] = \frac{1}{4}\|f - g\|^2 \leq \|f - h\|^2.$$

Therefore it suffices to bound $\|f - h\|^2$.

The formula for $\|f - h\|^2$ sums over all squared Fourier coefficients which intersect $\overline{J}$. Each such coefficient involves some coordinate $i \notin J$. By construction, $\sum_{i \in S} \hat{f}(S)^2 < \tau$. This shows that

$$\|f - h\|^2 \leq \sum_{i \notin J} \sum_{i \in S} \hat{f}(S)^2 < n\tau.$$

This bound is clearly not good enough. The problem is that we are counting each $\hat{f}(S)^2$ multiple times — in fact, $S \cap \overline{J}$ times. This suggests trying to get rid of Fourier coefficients corresponding to large sets S. Indeed,

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

so we can disregard large coefficients. It remains to bound

$$\sum_{\substack{S \not\subseteq J \\ |S| < M}} \hat{f}(S)^2.$$

At this point we invoke the magic wand of Boolean function analysis, *hypercontractivity*. An operator on functions is *contractive* if it reduces norm. For example, recall the noise operator T_ρ from Section 2. When $|\rho| \leq 1$, this operator is contractive:

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

It turns out that T_ρ is actually *hypercontractive*, which means that it satisfies an inequality of the form $\|T_\rho f\|_p \leq \|f\|_q$, for $p > q$.

Let us first recall what the L_p norms are:

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

This turns out to be a norm for $p \geq 1$ (including the limit $p = \infty$), that is, $\|cf\|_p = |c|\|f\|_p$ (which is easy to see), and the triangle inequality $\|f + g\|_p \leq \|f\|_p + \|g\|_p$ (which requires some argument). Another standard result states that $\|f\|_p$ is nondecreasing in p (when $p = \infty$, $\|f\|_\infty$ is just the maximum value of $|f|$).

The noise operator T_ρ is contractive for any norm (when $|\rho| \leq 1$). To see this, let us describe the noise operator in a slightly different way. One of the definitions we gave was: $T_\rho f(x) = \mathbb{E}[f(y)]$, where $y_i = x_i$ with probability $\frac{1+\rho}{2}$, and $y_i = -x_i$ otherwise. Instead, we can let $z_i = 1$ with probability $\frac{1+\rho}{2}$ and $z_i = -1$ otherwise, and then $T_\rho f(x) = \mathbb{E}[f(xz)]$. This shows that $T_\rho f$ is the average of functions f_z defined by $f_z(x) = f(xz)$. Since $\|f_z\|_p = \|f\|_p$ for all z, the triangle inequality immediately implies that $\|T_\rho f\|_p \leq \|f\|_p$.

Hypercontractivity is the stronger property that $\|T_\rho f\|_p \leq \|f\|_q$ for $p > q$ (depending on ρ). Using an inductive argument similar to an argument which we encountered in Section 3, we will show that

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

This actually holds for *every* function f, not just Boolean functions. From this, we will deduce that

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

We will apply this not to f itself, but rather to $L_i f$:

$$\sum_{i \in S} 3^{-|S|} \hat{f}(S)^2 = \|T_{1/\sqrt{3}} L_i f\|_2^2 \leq \|L_i f\|_{4/3}^2 = \mathbb{E}[|L_i f(x)|^{4/3}]^{3/2}.$$

Since $L_i f(x) \in \{0, \pm 1\}$, the right-hand side is in fact $\mathrm{Inf}_i[f]^{3/2}$. This shows that

$$\sum_{S \not\subseteq J} 3^{-|S|} \hat{f}(S)^2 \leq \sum_{i \notin J} \sum_{i \in S} 3^{-|S|} \hat{f}(S)^2 \leq \sum_{i \notin J} \mathrm{Inf}_i[f]^{3/2} \leq \sqrt{\tau}\, \mathrm{Inf}[f].$$

At this point it becomes apparent why it is useful to separate the coefficients into small S and large S: the above inequality is only useful if $3^{-|S|}$ is not too small, that is, when $|S|$ is not too large. Altogether, we obtain

$$\begin{aligned}
\|f - h\|^2 &\leq \sum_{|S| \geq M} \hat{f}(S)^2 + \sum_{\substack{S \not\subseteq J \\ |S| < M}} \hat{f}(S)^2 \\
&\leq \frac{\mathrm{Inf}[f]}{M} + 3^M \sum_{S \not\subseteq J} 3^{-|S|} \hat{f}(S)^2 \\
&\leq \frac{\mathrm{Inf}[f]}{M} + 3^M\, \mathrm{Inf}[f] \sqrt{\tau}.
\end{aligned}$$

Suppose we are aiming at $\|f - h\|^2 \leq \epsilon$. The easiest way to satisfy this is to ask for both summands to be at most $\epsilon/2$. Looking at the first summand, we should choose $M = 2\,\mathrm{Inf}[f]/\epsilon$, and so the second summand is $2^{O(\mathrm{Inf}[f]/\epsilon)}\, \mathrm{Inf}[f] \sqrt{\tau}$, which means that we need to choose $\tau = 2^{-\Theta(\mathrm{Inf}[f]/\epsilon)}$ (this requires some calculation).

How many coordinates does J contain? Each coordinate contributes $\mathrm{Inf}_i[f] \geq \tau$ to the total influence, and so the number of coordinates is at most $\mathrm{Inf}[f]/\tau = \mathrm{Inf}[f] 2^{O(\mathrm{Inf}[f]/\epsilon)} = 2^{O(\mathrm{Inf}[f]/\epsilon)}$. This concludes the proof of Friedgut's junta theorem:

> Let $f \colon \{\pm 1\}^n \to \{\pm 1\}$. For any ϵ, there is a Boolean junta h, depending on $2^{O(\mathrm{Inf}[f]/\epsilon)}$ coordinates, such that $\Pr[f \neq h] \leq \epsilon$.

5.2 Kahn–Kalai–Linial theorem

The Kahn–Kalai–Linial theorem states that every Boolean function has a somewhat influential coordinate. Our starting point is the inequality

$$\sum_{S \not\subseteq J} \hat{f}(S)^2 \leq \frac{\mathrm{Inf}[f]}{M} + 3^{M/2}\,\mathrm{Inf}[f]\sqrt{\tau},$$

where M is arbitrary and J is the collection of all coordinates whose influence is at least τ.

Suppose we are aiming at an influence of at least $\frac{\kappa}{n}\,\mathbb{V}[f]$, where κ is a function of n. If $\mathrm{Inf}[f] \geq \kappa\,\mathbb{V}[f]$, then the maximal influence is obviously at least $\frac{\kappa}{n}\,\mathbb{V}[f]$, so we can assume that $\mathrm{Inf}[f] \leq \kappa\,\mathbb{V}[f]$.

A natural place to find an influential variable is in the set J. Indeed,

$$\sum_{i \in J} \mathrm{Inf}_i[f] = \sum_S |S \cap J| \hat{f}(S)^2 \geq \sum_{S \neq \emptyset} \hat{f}(S)^2 - \sum_{S \not\subseteq J} \hat{f}(S)^2,$$

since if $S \neq \emptyset$ is a subset of J then $|S \cap J| \geq 1$. Now, the first term is $\mathbb{V}[f]$, and we bounded the other one above, so averaging over all variables in J, there must be one whose influence is at least

$$\frac{\mathbb{V}[f] - \mathrm{Inf}[f]/M - 3^{M/2}\,\mathrm{Inf}[f]\sqrt{\tau}}{|J|} \geq \frac{\mathbb{V}[f] - \mathrm{Inf}[f]/M - 3^{M/2}\,\mathrm{Inf}[f]\sqrt{\tau}}{\mathrm{Inf}[f]/\tau},$$

since $|J| \leq \mathrm{Inf}[f]/\tau$. This suggests choosing M, τ so that the two subtrahends are at most $\mathbb{V}[f]/10$, say. Accordingly, we choose $M = \mathrm{Inf}[f]/(10\,\mathbb{V}[f])$, and so $\tau = 2^{-\Theta(\mathrm{Inf}[f]/\mathbb{V}[f])}$. This gives us a variable whose influence is at least

$$\Theta\left(\frac{\mathbb{V}[f]}{\mathrm{Inf}[f]} 2^{-\Theta(\mathrm{Inf}[f]/\mathbb{V}[f])}\right).$$

Since $\mathrm{Inf}[f] \leq \kappa\,\mathbb{V}[f]$, this is at least

$$\Theta(2^{-\Theta(\kappa)}/\kappa) = 2^{-\Theta(\kappa)}.$$

The best choice of κ is the one which balances the two terms $2^{-\Theta(\kappa)}$ and $\frac{\kappa}{n}\,\mathbb{V}[f]$. We want $2^{\Theta(\kappa)}\kappa = n/\mathbb{V}[f]$, and so $\kappa = \Theta(\log(n/\mathbb{V}[f]))$. Altogether, we obtain the KKL theorem:

> Let $f\colon \{\pm 1\}^n \to \{\pm 1\}$. There exists a variable i whose influence is at least
>
> $$\Omega\left(\frac{\log(n/\mathbb{V}[f])}{n}\,\mathbb{V}[f]\right).$$

How do we use the KKL theorem to influence elections? As we mentioned in Section 4, the idea is to iteratively bribe the most influential voter. Suppose that we want to bribe voters until the probability that one of the candidates wins is at least $2/3$. The variance of f is

$$\mathbb{V}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = 1 - (\Pr[f = 1] - \Pr[f \neq -1])^2,$$

and so one of the candidates wins with probability at least $2/3$ when the variance drops below $1-(2/3-1/3)^2 = 8/9$.

If the original variance is below $8/9$, then there is nothing to do. Otherwise, we repeatedly bribe the most influential voter to vote for candidate 1. As long as the variance is above $8/9$ and there are m voters left, we can find a voter whose influence is at least $\Omega(\frac{\log m}{m}) = \Omega(\frac{\log n}{n})$. Bribing this voter increases the probability that candidate 1 wins by $\Omega(\frac{\log n}{n})$. Hence this process necessarily stops after $O(\frac{n}{\log n})$ steps.

The formulation of the KKL theorem above is not dimension-free, that is, it involves n. The general philosophy in Boolean function analysis is to prove statements where n does not appear. We can obtain such a statement directly from our proof. What the proof shows is that for a parameter κ of our choice, either $\mathrm{Inf}[f] \geq \kappa\,\mathbb{V}[f]$, or $\max_i \mathrm{Inf}_i[f] \geq 2^{-\Theta(\kappa)}$. Stated in terms of $\delta = 2^{-\Theta(\kappa)}$, this gives the following dimension-free version of the KKL theorem:

Let $f\colon \{\pm 1\}^n \to \{\pm 1\}$. For every $\delta > 0$, one of the following must hold:

$$\max_i \mathrm{Inf}_i[f] \geq \delta \text{ or } \mathrm{Inf}[f] = \Omega(\log(1/\delta)\,\mathbb{V}[f]).$$

That is, for balanced functions, if all influences are small, then the total influence is large. We can deduce the previous formulation of the KKL theorem as before, by balancing both terms.

6 Hypercontractivity

Hypercontractivity is the secret spice behind much of Boolean function analysis. One way to think about it is that it encapsulates a conceptually useful proof by induction. Another is via convergence of Markov chains, a point of view which we will not discuss here.

Let us try to prove an inequality of the form $\|T_\rho f\|_4 \leq \|f\|_2$ by induction. It would be simpler to raise everything to the fourth power, proving instead $\mathbb{E}[(T_\rho f)^4] \leq \mathbb{E}[f^2]^2$. The induction is on the number of inputs n. When $n = 0$, the inequality trivially holds for any ρ. Now suppose that we can prove this inequality for functions on n inputs, and try to prove it for functions on $n+1$ inputs. In order to reduce the number of variables, we will separate the variable x_{n+1}, writing

$$f = x_{n+1} \sum_{S \subseteq [n]} \hat{f}(S \cup \{n+1\}) x_S + \sum_{S \subseteq [n]} \hat{f}(S) x_S.$$

For the sake of succinctness, we will write this in the following way:

$$f = x_{n+1} g + h,$$

where g, h are functions on n variables. We then have

$$T_\rho f = \rho x_{n+1} T_\rho g + T_\rho h.$$

Now we can attempt the proof by induction:

$$\mathbb{E}[(T_\rho f)^4] = \sum_{i=0}^{4} \binom{4}{i} \rho^i \,\mathbb{E}[x_{n+1}^i]\,\mathbb{E}[(T_\rho g)^i (T_\rho h)^{4-i}].$$

It is easy to compute $\mathbb{E}[x_{n+1}^i] = 1$ for $i = 0, 2, 4$ and $\mathbb{E}[x_{n+1}^i] = 0$ for $i = 1, 3$, and so

$$\mathbb{E}[(T_\rho f)^4] = \rho^4\,\mathbb{E}[(T_\rho g)^4] + 6\rho^2\,\mathbb{E}[(T_\rho g)^2 (T_\rho h)^2] + \mathbb{E}[(T_\rho h)^4].$$

We can bound $\mathbb{E}[(T_\rho g)^4] \leq \mathbb{E}[g^2]^2$ and $\mathbb{E}[(T_\rho h)^4] \leq \mathbb{E}[h^2]^2$ by induction. As for the mixed term, the Cauchy–Schwarz inequality shows that

$$\mathbb{E}[(T_\rho g)^2 (T_\rho h)^2] \leq \sqrt{\mathbb{E}[(T_\rho g)^4]\,\mathbb{E}[(T_\rho h)^4]} \leq \mathbb{E}[g^2]\,\mathbb{E}[h^2].$$

Altogether, this gives

$$\mathbb{E}[(T_\rho f)^4] \leq \rho^4\,\mathbb{E}[g^2]^2 + 6\rho^2\,\mathbb{E}[g^2]\,\mathbb{E}[h^2] + \mathbb{E}[h^2]^2.$$

Our target is $\mathbb{E}[f^2]^2$. By Parseval's identity,

$$\mathbb{E}[f^2] = \|f\|^2 = \sum_{S \subseteq [n+1]} \hat{f}(S)^2 = \sum_{S \subseteq [n]} \hat{g}(S)^2 + \sum_{S \subseteq [n]} \hat{h}(S)^2 = \mathbb{E}[g^2] + \mathbb{E}[h^2],$$

and so we are looking for a value of ρ for which the following always holds:

$$\rho^4\,\mathbb{E}[g^2]^2 + 6\rho^2\,\mathbb{E}[g^2]\,\mathbb{E}[h^2] + \mathbb{E}[h^2]^2 \leq \mathbb{E}[g^2]^2 + 2\,\mathbb{E}[g^2]\,\mathbb{E}[h^2] + \mathbb{E}[h^2]^2.$$

Comparing coefficients, we need $\rho^4 \leq 1$ and $6\rho^2 \leq 2$, and so $|\rho| \leq 1/\sqrt{3}$. We have proved hypercontractivity in the following form:

For all functions $f\colon \{\pm 1\}^n \to \mathbb{R}$ and all $|\rho| \leq 1/\sqrt{3}$:

$$\|T_\rho f\|_4 \leq \|f\|_2.$$

Crucially, this inequality doesn't involve n. We say that it is *dimension-independent*. Boolean function analysis concerns itself mostly with such dimension-independent results. We have seen several examples above: the Friedgut–Kalai–Naor theorem, Friedgut's junta theorem and one version of the Kahn–Kalai–Linial theorem.

When f has constant degree d, we can get rid of the noise operator, by writing $f = T_\rho T_\rho^{-1} f$. The spectral formula for the noise operator makes it clear that T_ρ is indeed invertible, and $T_\rho^{-1} = T_{\rho^{-1}}$, and so

$$\|f\|_4 = \|T_{1/\sqrt{3}} T_{\sqrt{3}} f\|_4 \leq \|T_{\sqrt{3}} f\|_2 = \sqrt{\sum_S 3^{|S|} \hat{f}(S)^2} \leq \sqrt{3}^d \|f\|_2.$$

The proofs in Section 5 used a different form of hypercontractivity, in which the L_2 norm was on the left-hand side rather than on the right-hand side. This version is easily deducible from the current version, via Hölder's inequality, which states that $\langle f, g \rangle \leq \|f\|_p \|g\|_q$, where $1/p + 1/q = 1$. If $p = q = 1/2$ then we just get the Cauchy–Schwarz inequality. Here we will be interested in $p = 4$ and $q = 4/3$. Using this, if $|\rho| \leq 1/\sqrt{3}$ then

$$\|T_\rho f\|_2^2 = \langle f, T_\rho^2 f \rangle \leq \|f\|_{4/3} \|T_\rho^2 f\|_4 \leq \|f\|_{4/3} \|T_\rho f\|_2,$$

and so $\|T_\rho f\|_2 \leq \|f\|_{4/3}$. The first step uses the symmetry of the operator T_ρ:

$$\langle T_\rho g, h \rangle = \sum_S \rho^{|S|} \hat{g}(S) \hat{h}(S) = \sum_S \hat{g}(S) \rho^{|S|} \hat{h}(S) = \langle g, T_\rho h \rangle.$$

Altogether, we get

For all functions $f\colon \{\pm 1\}^n \to \mathbb{R}$ and all $|\rho| \leq 1/\sqrt{3}$:

$$\|T_\rho f\|_2 \leq \|f\|_{4/3}.$$

6.1 Another proof of FKN

As another illustration of hypercontractivity, let us give an alternative proof of the Friedgut–Kalai–Naor theorem which we proved in Section 3.

The Friedgut–Kalai–Naor theorem states, in one formulation, that if $F\colon \{\pm 1\}^n \to \{\pm 1\}$ is close to degree 1, in the sense that $\|F^{>1}\|^2 = \epsilon$, then F is close to a Boolean function G which is either constant or a dictator, in the sense that $\Pr[F \neq G] = O(\epsilon)$.

As we have shown in Section 3, we can assume, without loss of generality, that $\mathbb{E}[F] = 0$. Therefore, $f = F^{\leq 1}$ has the form

$$f = \sum_{i=1}^n c_i x_i,$$

where $c_i = \hat{F}(\{i\}) = \mathbb{E}[F x_i]$. In this case our goal is to show that f is close to a Boolean dictator $\pm x_i$.

In order to show that f is close to a dictator, it suffices to show that some c_i is close to ± 1. Indeed,

$$\mathbb{E}[F x_i] = \Pr[F = x_i] - \Pr[F = -x_i] = 2\Pr[F = x_i] - 1 = 1 - 2\Pr[F = -x_i],$$

and so if $c_i = 1 - \delta$ then $\Pr[F = x_i] = 1 - \delta/2$, and if $c_i = -1 + \delta$ then $\Pr[F = -x_i] = 1 - \delta/2$.

Since $\|F^{>1}\|^2 = \epsilon$ while $\|F\|^2 = 1$, we can conclude that $\|f\|^2 = 1 - \epsilon$, that is,

$$\sum_{i=1}^n c_i^2 = 1 - \epsilon.$$

Also, we know that each c_i^2 is $O(\epsilon)$-close to $\{0, \pm 1\}$, as we have shown in Section 3. It could therefore conceivably be the case that all c_i are small. In order to rule this case out, we will consider

$$\sum_{i=1}^n c_i^4.$$

If all c_i were small, then this sum would be at most $O(\epsilon)$ (since by assumption $c_i^4 \leq O(\epsilon) c_i^2$), so to rule this case out, all we need to do is give a lower bound on $\sum_i c_i^4$, which we expect to be close to 1.

In order to get a handle on $\sum_i c_i^4$, we consider $\mathbb{E}[f^4]$:

$$\mathbb{E}[f^4] = \sum_{i,j,k,\ell=1}^n c_i c_j c_k c_\ell \, \mathbb{E}[x_i x_j x_k x_\ell] = \sum_{i=1}^n c_i^4 + 3 \sum_{i=1}^n \sum_{j \neq i} c_i^2 c_j^2 = 3 \left(\sum_{i=1}^n c_i^2 \right)^2 - 2 \sum_{i=1}^n c_i^4.$$

We expect the left-hand side to be close to 1. Since the right-hand side is $3\,\mathbb{E}[f^2]^2 - 2\sum_i c_i^4 \approx 3 - 2\sum_i c_i^4$, this will show that $\sum_i c_i^4 \approx 1$.

Instead of estimating $\mathbb{E}[f^4]$ directly, we will consider the related quantity $\mathbb{E}[(f^2-1)^2]$, prompted by the known property $\mathbb{E}[f^2] = 1 - \epsilon$.

We know that $\mathbb{E}[\mathrm{dist}(f, \{\pm 1\})^2] \leq \mathbb{E}[(f-F)^2] \leq \epsilon$, and so with probability $1 - 1/C$, it holds that $\mathrm{dist}(f, \{\pm 1\})^2 \leq C\epsilon$ (we will choose C later on). This implies that $f = \pm 1 + \tau$, where $|\tau| \leq \sqrt{C\epsilon}$, and so $f^2 = 1 + \Theta(\tau)$ (since $\epsilon \leq 1$), implying that $(f^2-1)^2 = O(\tau^2) = O(C\epsilon)$. This shows that

$$\mathbb{E}[(f^2-1)^2] \leq O(C\epsilon) + \mathbb{E}[(f^2-1)^2 1_{\mathrm{dist}(f,\{\pm 1\})^2 > C\epsilon}].$$

The hard part is to bound the behavior of f on the bad inputs, which cause it to be abnormally large. This is where hypercontractivity comes in. But first, we need a trick, the most standard one — Cauchy–Schwarz:

$$\mathbb{E}[(f^2-1)^2 1_{\mathrm{dist}(f,\{\pm 1\})^2 > C\epsilon}] \leq \sqrt{\mathbb{E}[(f^2-1)^4]} \sqrt{\mathbb{E}[1_{\mathrm{dist}(f,\{\pm 1\})^2 > C\epsilon}]} \leq \frac{1}{\sqrt{C}} \|f^2 - 1\|_4^2.$$

Since $f^2 - 1$ has degree 2, we know that $\|f^2-1\|_4 \leq 3\|f^2-1\|_2$, and so altogether,

$$\mathbb{E}[(f^2-1)^2] \leq O(C\epsilon) + \frac{9}{\sqrt{C}} \mathbb{E}[(f^2-1)^2].$$

Choosing any $C > 81$, we conclude that

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

Since $(f^2-1)^2 = f^4 - 2f^2 + 1$ and $\mathbb{E}[f^2] = 1 - \epsilon$, this shows that

$$\mathbb{E}[f^4] = O(\epsilon) + 2(1-\epsilon) - 1 = 1 + O(\epsilon),$$

as predicted above. Therefore

$$2\sum_{i=1}^n c_i^4 = 3\,\mathbb{E}[f^2]^2 - \mathbb{E}[f^4] = 3(1-\epsilon)^2 - (1 + O(\epsilon)) \geq 2 - O(\epsilon),$$

implying that $\sum_i c_i^4 \geq 1 - O(\epsilon)$. It is now easy to show that some c_i must be large. This follows from

$$\sum_{i=1}^n c_i^4 \leq \max_i c_i^2 \cdot \sum_{i=1}^n c_i^2 = (1-\epsilon) \max_i c_i^2.$$

Altogether, we conclude that some c_i satisfies $c_i^2 \geq 1 - O(\epsilon)$, and so $|c_i| \geq 1 - O(\epsilon)$. As we have seen above, this implies that F is $O(\epsilon)$-close to either x_i or $-x_i$.

6.2 General norms

Most applications of hypercontractivity in Boolean function analysis use one of the two forms stated above. Sometimes, however, we need to consider larger norms. Here is the most general form of hypercontractivity:

> For all functions $f\colon \{\pm 1\}^n \to \{\pm 1\}$, all $1 \leq p \leq q \leq \infty$, and all $|\rho| \leq \sqrt{\frac{p-1}{q-1}}$:
>
> $$\|T_\rho f\|_q \leq \|f\|_p.$$

This theorem is proved by induction. In the base case $n = 1$, we have a function $f = a + bx$. We need to prove that

$$\sqrt[q]{\frac{|a+\rho b|^q + |a-\rho b|^q}{2}} \leq \sqrt[p]{\frac{|a+b|^p + |a-b|^p}{2}}.$$

If $a = 0$, then the inequality reads $|\rho b| \leq |b|$, which holds as long as $|\rho| \leq 1$. Hence we can assume that $a \neq 0$. Since the inequality is homogeneous, we can consider $f/a = 1 + (b/a)x$ instead of f. Writing $t = a/b$, we need to prove that

$$\sqrt[q]{\frac{|1+\rho t|^q + |1-\rho t|^q}{2}} \leq \sqrt[p]{\frac{|1+t|^p + |1-t|^p}{2}}.$$

When t is small, a Taylor expansion shows that

$$|1+\rho t|^q = (1+\rho t)^q \approx 1 + q\rho t + \binom{q}{2}\rho^2 t^2.$$

We get a similar expression for $|1-\rho t|^q$, and so the left-hand side is roughly

$$\sqrt[q]{1 + \binom{q}{2}\rho^2 t^2} \approx 1 + \frac{q-1}{2}\rho^2 t^2.$$

Similarly, the left-hand side is roughly $1 + \frac{p-1}{2}t^2$. Comparing coefficients, we see that we need $(q-1)\rho^2 \leq p-1$, and so $|\rho| \leq \sqrt{\frac{p-1}{q-1}}$. It turns out that the base case does hold for such ρ, but we will not prove so here.

The more interesting part of the proof is *tensorization*, which is the way in which we deduce the general case from the base case. While a direct inductive proof is possible, it is a bit tricky. Instead, we will consider an equivalent form of hypercontractivity, due to Ryan O'Donnell, for which the inductive proof is straightforward.

Suppose that $\|T_\rho f\|_q \leq \|f\|_p$. Using Hölder's inequality, we can obtain a two-function version:

$$\langle T_\rho f, g\rangle \leq \|T_\rho f\|_q \|g\|_{q'} \leq \|f\|_p \|g\|_{q'}, \text{ where } \frac{1}{q} + \frac{1}{q'} = 1.$$

Coincidentally, the left-hand side is a very natural bilinear form:

$$\langle T_\rho f, g\rangle = \mathop{\mathbb{E}}_{x}[T_\rho f(x) g(x)] = \mathop{\mathbb{E}}_{\substack{x \sim \{\pm 1\}^n \\ y \sim N_\rho(x)}}[f(y)g(x)],$$

where $N_\rho(x)$ is obtained by flipping each coordinate with probability $\frac{1-\rho}{2}$. The connection between x and y is symmetric: we could also have sampled $y \sim \{\pm 1\}^n$ and then $x \sim N_\rho(y)$ to obtain the same distribution. We write this symmetric distribution as $(x, y) \sim N_\rho$.

The two-function version will be easier to tensorize. Before seeing that, let us show how to deduce hypercontractivity in its original form. The idea is very simple: assuming for simplicity that $f \geq 0$, we take $g = (T_\rho f)^{q-1}$ to obtain

$$\|T_\rho f\|_q^q = \mathbb{E}[(T_\rho f)^q] = \langle T_\rho f, (T_\rho f)^{q-1}\rangle \leq \|f\|_p \|(T_\rho f)^{q-1}\|_{q/(q-1)} = \|f\|_p \|T_\rho f\|_q^{q-1},$$

and so $\|T_\rho f\|_q \leq \|f\|_p$. The proof for arbitrary f is similar, using $g = |T_\rho f|^{q-1}\operatorname{sgn}(T_\rho f)$.

Now let us prove that $\langle T_\rho f, g\rangle$ holds by induction. We have already seen the base case $n = 1$, so suppose that f, g are functions on $n+1$ variables. The basic idea is to write

$$\langle f, T_\rho g\rangle = \mathop{\mathbb{E}}_{(x_{n+1},y_{n+1})\sim N_\rho} \mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_n,y_n)\sim N_\rho} [f(x_1,\ldots,x_n,x_{n+1})g(y_1,\ldots,y_n,y_{n+1})].$$

If we fix the value of x_{n+1}, y_{n+1}, then the functions f, g now depend only on n variables, and so we can apply the induction hypothesis:

$$\langle f, T_\rho g\rangle \leq \mathop{\mathbb{E}}_{(x_{n+1},y_{n+1})\sim N_\rho} \left[\mathop{\mathbb{E}}_{x_1,\ldots,x_n} [|f(x_1,\ldots,x_n,x_{n+1})|^p]^{1/p} \mathop{\mathbb{E}}_{x_1,\ldots,x_n} [|g(x_1,\ldots,x_n,y_{n+1})|^{q'}]^{1/q'} \right].$$

The outer expectation is also of the form $\langle F, T_\rho G\rangle$, on a single coordinate. Applying the base case gives

$$\begin{aligned}\langle f, T_\rho g\rangle \leq{}& \\ &\mathop{\mathbb{E}}_{x_{n+1}} \left[\left| \mathop{\mathbb{E}}_{x_1,\ldots,x_n} [|f(x_1,\ldots,x_n,x_{n+1})|^p]^{1/p} \right|^p \right]^{1/p} \cdot \mathop{\mathbb{E}}_{x_{n+1}} \left[\left| \mathop{\mathbb{E}}_{x_1,\ldots,x_n} [|g(x_1,\ldots,x_n,x_{n+1})|^{q'}]^{1/q'} \right|^{q'} \right]^{1/q'} \\ &= \|f\|_p \|g\|_{q'}.\end{aligned}$$

7 Constant degree functions: Kindler–Safra theorem

In Section 2 we showed that every Boolean degree 1 function is a dictator. Similarly, every Boolean degree d function is a junta, although the argument is less trivial.

Let $f\colon \{\pm1\}^n \to \{\pm1\}$ have degree d. If f is a junta, then f must depend on all variables with positive influence. We call such variables *influential*. Since there are only finitely many juntas, the influence of any influential variable in a junta is $\Omega(1)$. This suggests showing that every influential variable in f has influence $\Omega(1)$. To show this, we consider the function $L_i f$, define in Section 4. The proof is a simple application of hypercontractivity:

$$\operatorname{Inf}_i[f] = \|L_i f\|_2^2 \leq 3^d \|T_{1/\sqrt{3}} L_i f\|_2^2 \leq 3^d \|L_i f\|_{4/3}^2 = 3^d \|L_i f\|_2^3 = 3^d \operatorname{Inf}_i[f]^{3/2}.$$

If $\operatorname{Inf}_i[f] \neq 0$ then $\operatorname{Inf}_i[f] \geq 9^{-d}$. Conversely, as we have shown in Section 4, $\operatorname{Inf}[f] \leq d\,\mathbb{V}[f] \leq d$. Therefore at most $9^d d$ variables are influential in f. In other words, f depends on at most $9^d d$ variables. In fact, this can be improved to $O(2^d)$ [CHS20, Wel20].

What about Boolean functions which are merely *close* to degree d? In Section 3, we showed that such functions are close to a constant or a dictator, that is, to a Boolean degree 1 function. Using a similar strategy, we will show that the same holds for degree d functions, a result first proved by Kindler and Safra [KS04, Kin02].

Let $f\colon \{\pm1\}^n \to \{\pm1\}$ satisfy $\|f^{>d}\|^2 = \epsilon$. Following the lead of Section 3, we will consider the function $g = f^{\leq d}$, which is a degree d function satisfying $\mathbb{E}[\operatorname{dist}(g, \{\pm1\})^2] \leq \epsilon$.

Just as in Section 3, we can assume that ϵ is small enough, say $\epsilon \leq 1$. This implies that $\|g\|^2 \leq 2\|f\|^2 + 2\|f^{>d}\|^2 = O(1)$, and so $\operatorname{Inf}[g] \leq d\|g\|^2 = O(d)$.

Identifying the junta variables The first step is identifying the junta variables, which we accomplish by modifying the argument for Boolean degree d functions. For any variable i,

$$\sqrt{\operatorname{Inf}_i[g]} = \|L_i g\|_2 \leq \sqrt{3}^d \|L_i g\|_{4/3}.$$

This time we cannot directly relate $\|L_i g\|_{3/2}$ and $\operatorname{Inf}_i[g]$. However, rounding g to $G = \operatorname{round}(g, \{\pm1\})$,

$$\|L_i g\|_{4/3} \leq \|L_i G\|_{4/3} + \|L_i g - L_i G\|_{4/3} \leq \|L_i G\|_{4/3} + \|L_i g - L_i G\|_2 \leq \|L_i G\|_{4/3} + \|g - G\|_2 \leq \|L_i G\|_{4/3} + \sqrt{\epsilon}.$$

Since G is Boolean,

$$\|L_i G\|_{4/3} = \|L_i G\|_2^{3/2} \leq (\|L_i g\|_2 + \|L_i G - L_i g\|_2)^{3/2} = (\sqrt{\mathrm{Inf}_i[g]} + \sqrt{\epsilon})^{3/2}.$$

Therefore

$$\sqrt{\mathrm{Inf}_i[g]} \leq \sqrt{3}^d \big((\sqrt{\mathrm{Inf}_i[g]} + \sqrt{\epsilon})^{3/2} + \sqrt{\epsilon}\big).$$

If $\sqrt{\mathrm{Inf}_i[g]} \geq 2\sqrt{3}^d \sqrt{\epsilon}$ then

$$\frac{1}{4}\,\mathrm{Inf}_i[g] \leq 3^d (\sqrt{\mathrm{Inf}_i[g]} + \sqrt{\epsilon})^3 = O(\mathrm{Inf}_i[g]^{3/2}),$$

and so $\mathrm{Inf}_i[g] = \Omega(1)$.

We have shown that the influence of every variable is either $O(\epsilon)$ or $\Omega(1)$. Since $\mathrm{Inf}[g] = O(d)$, there can be at most $O(1)$ of the latter variables, forming a set J. We reorder the variables so that $J = \{1, \ldots, k\}$. Variables outside of J have influence at most $K\epsilon$, for some $K > 0$. We can assume that $K > 1$.

Inductive argument The heart of the proof is an inductive argument, mimicking the proof in Section 3. We will prove by induction that for all $m > k$,

$$\sum_{\max(S) \geq m} \hat{g}(S)^2 \leq K\epsilon.$$

The base case, $m = n + 1$, is trivial. Now suppose that this inequality holds for some $m + 1$. Since $\mathrm{Inf}_m[g] \leq K\epsilon$,

$$\sum_{\max(S) \geq m} \hat{g}(S)^2 \leq \sum_{\max(S) > m} \hat{g}(S)^2 + \sum_{m \in S} \hat{g}(S)^2 \leq 2K\epsilon.$$

For any subset $T \subseteq \{m, \ldots, n\}$, let $g^{\oplus T}$ be the function formed by negating the coordinates in T. Thus

$$\Delta_T := \frac{g - g^{\oplus T}}{2} = \sum_{|S \cap T| \text{ odd}} \hat{g}(S) x_S.$$

Since $\mathbb{E}[\mathrm{dist}(g/2, \{\pm 1/2\})^2] = \mathbb{E}[\mathrm{dist}(-g^{\oplus T}/2, \{\pm 1/2\})^2] \leq \epsilon/4$,

$$\mathbb{E}[\mathrm{dist}(\Delta_T, \{0, \pm 1\})^2] \leq 2\,\mathbb{E}[\mathrm{dist}(g/2, \{\pm 1/2\})^2] + 2\,\mathbb{E}[\mathrm{dist}(-g^{\oplus T}/2, \{\pm 1/2\}^2)] \leq \epsilon.$$

(This is because the sum of two values in $\{\pm 1/2\}$ lies in $\{0, \pm 1\}$.) On the other hand, we know that

$$\|\Delta_T\|^2 = \sum_{|S \cap T| \text{ odd}} \hat{g}(S)^2 \leq \sum_{\max(S) \geq m} \hat{g}(S)^2 \leq 2K\epsilon.$$

Following the steps of Section 3, we want to say that the only way that Δ_T can be simultaneously close to $\{0, \pm 1\}$ and have small norm is if it is in fact close to 0. As in Section 3, we observe that

$$\Delta_T^2 \leq \mathrm{dist}(\Delta_T, \{0, \pm 1\})^2 + 4\Delta_T^4,$$

since either $|\Delta_T| \leq 1/2$, in which case Δ_T^2 equals the first term, or $|\Delta_T| > 1/2$, in which case $\Delta_T^2 \leq 4\Delta_T^4$. Taking expectations and using hypercontractivity, we can bound $\mathbb{E}[\Delta_T^4]$ in terms of $\mathbb{E}[\Delta_T^2]^2$ (in Section 3 we resorted to a direct calculation instead):

$$\|\Delta_T\|_2^2 \leq \epsilon + 4\|\Delta_T\|_4^4 \leq \epsilon + 4 \cdot 9^d \|\Delta_T\|_2^4 \leq \epsilon + O(K^2\epsilon^2).$$

So far so good, but how does Δ_T relate to the quantity we are actually interested in? The idea is to pick T at random. If $\max(S) < m$, then $|S \cap T|$ is never odd. Otherwise, the probability that $|S \cap T|$ is odd is at

least $2^{-|S\cap\{m,\dots,n\}|} \geq 2^{-d}$ (this is the probability that one element in $S \cap \{m,\dots,n\}$ is inside T, and the rest are outside T). Therefore

$$\sum_{\max(S)\geq m} \hat{g}(S)^2 \leq 2^d \mathop{\mathbb{E}}_T \|\Delta_T\|_2^2 \leq 2^d\epsilon + O(K^2\epsilon^2).$$

If $K > 2^d$ and ϵ is small enough (a valid assumption, as in Section 3), the right-hand side is at most $K\epsilon$, completing the proof by induction.

Finishing the proof Where do we stand? Assuming that $J = \{1,\dots,k\}$, we have shown that

$$\sum_{S\colon \max(S)>k} \hat{g}(S)^2 = O(\epsilon).$$

For general J, this shows that

$$\sum_{S\not\subseteq J} \hat{g}(S)^2 = O(\epsilon).$$

As in Section 5, this shows that the function h obtained by averaging h over the coordinates outside of J satisfies $\|g-h\|^2 = O(\epsilon)$. Furthermore, as in Section 5, the function $H = \mathrm{round}(h, \{\pm1\})$ is a Boolean junta which also satisfies $\|g-H\|^2 = O(\epsilon)$.

We would like to argue that H is not just a junta, but actually has degree d. Indeed, up to the choice of the junta coordinates, there are only finitely many options for H. Therefore either $\deg H \leq d$, or $\|H^{>d}\|^2 = \Omega(1)$. In the latter case, $\|g-H\|^2 \geq \|H^{>d}\|^2 = \Omega(1)$, since $\deg g \leq d$. This possibility can be ruled out if ϵ is small enough. Summarizing, we have proved the following results, known as the Kindler–Safra theorem:

> If $g\colon \{\pm1\}^n \to \mathbb{R}$ is a degree d function satisfying $\mathbb{E}[\mathrm{dist}(g,\{\pm1\})^2] < \epsilon$, then there is a Boolean degree d function $r\colon \{\pm1\}^n \to \{\pm1\}$, depending on a constant number of inputs, such that $\|g-r\|^2 = O(\epsilon)$.

> If $f\colon \{\pm1\}^n \to \{\pm1\}$ satisfies $\|f^{>d}\|^2 \leq \epsilon$ then there is a Boolean degree d function $r\colon \{\pm1\}^n \to \{\pm1\}$, depending on a constant number of inputs, such that $\Pr[g \neq r] = O(\epsilon)$.

How does the big O constant scale with d? Carefully keeping track of all big O constants, we see that $\|g-H\|^2 \leq 2^{O(d)}\epsilon$. Similarly, H depends on $2^{O(d)}$ coordinates.

In order to guarantee that H has degree d, we appealed to there being only finitely many possible H. We can make this argument constructive. Suppose that $\deg H > d$, say $\hat{H}(S) \neq \emptyset$, where $|S| > d$. Since H depends on $M = 2^{O(d)}$ coordinates, this means that $|\hat{H}(S)| = |\mathbb{E}[Hx_S]| \geq 2^{-M}$, since $\mathbb{E}[Hx_S]$ is the average of 2^M many integers. Therefore, in order to guarantee that H has degree d, we need $\epsilon \leq 2^{-2^{O(d)}}$.

Keller and Klein [KK20] used a different proof (carried out in the more difficult setting of the slice) that directly constructs a Boolean function H of degree d such that $\|g-H\|^2 \leq 2^{O(d)}\epsilon$. They also showed that $\Pr[f \neq H] \leq 4\epsilon + 2^{O(d)}\epsilon^2$, using a simple application of hypercontractivity.

8 Biased Fourier analysis: Erdős–Ko–Rado

So far we have been considering functions with respect to the uniform distribution. This distribution is hidden in our expectations, which are always taken with respect to the input distribution. However, in many cases we are interested in other distributions. Perhaps the best example is $G(n,p)$ random graphs, in which each edge appears in the graph with probability p. When $p = 1/2$, this is just the uniform distribution, but we are often interested in other values of p.

When $p = 1/2$, it is convenient to think of the Boolean cube as $\{\pm1\}^n$, and of Boolean functions as ±1-valued functions. For general p, this convention makes less sense, and instead, typically we replace $\{\pm1\}$

with $\{0,1\}$. The distribution on $\{0,1\}^n$ in which each coordinate equals 1 with probability p independently is known as μ_p.

How do we generalize the Fourier expansion to the p-biased case? There are two basic properties of the Fourier basis: it is an orthonormal basis, and it is "coordinate-based", in the sense that there is a natural correspondence between subsets of $[n]$ and Fourier characters. We can express "coordinate-based" in a more formal way: if a function depends only on the coordinates in J, then its Fourier expansion is supported on subsets of J; and conversely, χ_S depends only on the coordinates in S.

These "axioms" suffice to derive the p-biased Fourier basis, up to negation. Fixing p, we will denote this basis by $(\omega_S)_{S\subseteq[n]}$. First of all, $\omega_\emptyset$ is constant, and by orthonormality, the constant is ± 1. We choose $\omega_\emptyset = 1$. Continuing, $\omega_{\{i\}}$ depends only on x_i, say $\omega_{\{i\}} = \alpha x + \beta$. By orthogonality,

$$0 = \mathop{\mathbb{E}}_{\mu_p}[\omega_{\{i\}}\omega_\emptyset] = \mathop{\mathbb{E}}_{\mu_p}[\alpha x_i + \beta] = \alpha p + \beta.$$

By orthonormality,

$$1 = \mathop{\mathbb{E}}_{\mu_p}[\omega_{\{i\}}^2] = \mathop{\mathbb{E}}_{\mu_p}[\alpha^2 x_i^2 + 2\alpha\beta x_i + \beta^2] = (\alpha^2 + 2\alpha\beta)p + \beta^2.$$

The first equation shows that $\beta = -\alpha p$, and so the second equation gives

$$\alpha^2 p - 2\alpha^2 p^2 + \alpha^2 p^2 = 1 \Longrightarrow \alpha^2 = \frac{1}{p(1-p)}.$$

We choose

$$\omega_{\{i\}} = \frac{x_i - p}{\sqrt{p(1-p)}}.$$

In this formula, $p = \mathbb{E}[x_i]$ and $p(1-p) = \mathbb{V}[x_i]$.

At this point one can already guess the general formula:

$$\omega_S = \prod_{i\in S} \omega_{\{i\}} = \prod_{i\in S} \frac{x_i - p}{\sqrt{p(1-p)}}.$$

Indeed, $\mathbb{E}[\omega_S^2] = \prod_{i\in S} \mathbb{E}[\omega_{\{i\}}^2] = 1$, and if $S \neq T$, then without loss of generality there is some $i \in S \setminus T$, and then $\mathbb{E}[\omega_S\omega_T] \propto \mathbb{E}[\omega_{\{i\}}] = 0$. In both cases, we crucially used the independence of the coordinates — equivalently, the fact that μ_p is a *product measure*.

A basis of this form is known as *tensorial* — by defining the tensor product accordingly, we can think of the p-biased Fourier basis as the tensor product

$$\{\omega_\emptyset, \omega_{\{1\}}\} \otimes \cdots \otimes \{\omega_\emptyset, \omega_{\{n\}}\}.$$

This basis is unique up to (individual) negation. To see this, suppose that we have constructed ω_T for $T \subsetneq S$. The space of all functions on S has dimension $2^{|S|}$. The function ω_S is orthogonal to the $2^{|S|} - 1$ functions ω_T for $T \subsetneq S$. These functions are orthogonal, and so linearly independent, hence ω_T belongs to some one-dimensional subspace, which contains exactly two points of unit norm.

8.1 Intersecting families

We now give a simple application of the p-biased Fourier basis, to the study of intersecting families. A subset $\mathcal{F} \subseteq \{0,1\}^n$ is called an *intersecting family* if every $A, B \in \mathcal{F}$ intersect (that is, they are not disjoint). What is the largest μ_p-measure of an intersecting family? The answer depends qualitatively on p.

When $p > 1/2$, one can take the family of all sets of size $\lceil \frac{n+1}{2} \rceil$. This is an intersecting family whose μ_p-measure tends to 1 as $n \to \infty$, and so this regime is not so interesting.

When $p = 1/2$, every intersecting family has measure at most $1/2$, since it can contain at most one set out of each pair $A, \overline{A}$. There are many different intersecting families of measure $1/2$, for example all families of the form

$$\{S : |S \cap [2r+1]| \geq r+1\},$$

which can also be viewed as the majority function on the first $2r+1$ coordinates.

The most interesting regime is when $p < 1/2$. One obvious construction is a *star*, which consists of all sets containing the point i, for some arbitrary $i \in [n]$. This family has μ_p-measure p. Is there any better construction? If not, are there any other constructions achieving the same measure? If not, are there any other constructions *approaching* the same measure?

We will approach the subject using a technique originating in the work of Hoffman [Hof70] and Lovász [Lov79], although for the purpose of exposition, the relation to their work will not be immediately apparent.

The idea is to define a "noise operator" T with appropriate properties. Just as the noise operator T_ρ, we will define $Tf(x) = \mathbb{E}[f(y)]$, where y is some noisy version of x, say $y \sim N(x)$. It is natural to require that if $x \sim \mu_p$ then also $y \sim \mu_p$. Moreover, noise should be applied to different coordinates independently.

We will design our noise operator in such a way that if f is the characteristic function of an intersecting family, then $\langle f, Tf\rangle = 0$ (this mimics a construction of Hoffman, in which T is the normalized adjacency matrix of some graph). This property will be satisfied if $N(1_A)$ is guaranteed to be a set which is disjoint from A. Considering an individual coordinate, this means that N must change 1 to 0, and can act arbitrarily on 0, under the constraint that if the input to N is distributed μ_p, then so is the output. If we denote by t the probability that N changes 0 to 1, then the probability that N outputs 1 is $(1-p)t$, and so $t = \frac{p}{1-p}$.

If $1_A \sim \mu_p$ and $1_B \sim N(1_A)$ then $A \cap B = \emptyset$ and both 1_A and 1_B have distribution μ_p. As seen above, this determines the distribution of (A, B) uniquely. Since the constraints are symmetric, we conclude that we can also sample (A, B) by first sampling $1_B \sim \mu_p$ and then sampling $1_A \sim N(1_B)$.

This symmetry implies that ω_S is an eigenfunction of T. The proof is by induction. Suppose that we have shown that $T\omega_R = \lambda_R \omega_R$ for all $R \subsetneq S$. Then

$$\langle \omega_R, T\omega_S\rangle = \langle T\omega_R, \omega_S\rangle = \lambda_R \langle \omega_R, \omega_S\rangle = 0.$$

Hence $T\omega_S$ is a function, depending only on the coordinates in S, which is orthogonal to ω_R for all $R \subsetneq S$. This means that $T\omega_S = \lambda_S \omega_S$ for some λ_S.

If $1_R \sim N(1_S)$ then $R \cap S = \emptyset$, and so $T\omega_S(1_S) = \mathbb{E}[\omega_S(R)] = \omega_S(\emptyset)$. This allows us to calculate λ_S by substituting 1_S:

$$\lambda_S \left(\frac{1-p}{\sqrt{p(1-p)}}\right)^{|S|} = \lambda_S \omega_S(1_S) = T\omega_S(1_S) = \omega_S(\emptyset) = \left(\frac{-p}{\sqrt{p(1-p)}}\right)^{|S|},$$

and so

$$\lambda_S = \left(-\frac{p}{1-p}\right)^{|S|}.$$

This means that

$$Tf = \sum_S \left(\frac{-p}{1-p}\right)^{|S|} \hat{f}(S)\omega_S.$$

If f is indeed the characteristic function of an intersecting family, then $\langle f, Tf\rangle = 0$, which by Parseval's identity implies that

$$0 = \langle f, Tf\rangle = \sum_S \left(\frac{-p}{1-p}\right)^{|S|} \hat{f}(S)^2 \geq \hat{f}(\emptyset)^2 - \frac{p}{1-p}\sum_{S\neq\emptyset} \hat{f}(S)^2.$$

We know that $\hat{f}(\emptyset) = \mathbb{E}[f] = \mu_p(\mathcal{F})$. Parseval's identity shows that

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

Substituting this in the inequality, we obtain

$$0 \geq \mu_p(\mathcal{F})^2 - \frac{p}{1-p}\mu_p(\mathcal{F})(1-\mu_p(\mathcal{F})) = \mu_p(\mathcal{F})(1-\mu_p(\mathcal{F}))\left(\frac{\mu_p(\mathcal{F})}{1-\mu_p(\mathcal{F})} - \frac{p}{1-p}\right),$$

and so $\mu_p(\mathcal{F}) \leq p$, since the function $\frac{x}{1-x} = \sum_{n\geq 1} x^n$ is increasing for $x \geq 0$.

This calculation allows us to say much more. First of all, we can understand which intersecting families have μ_p-measure exactly p. If f is the characteristic function of such an intersecting family, all inequalities above must be tight. Since the only inequality we used was $(\frac{-p}{1-p})^{|S|} \geq \frac{-p}{1-p}$ and $\frac{p}{1-p} < 1$ (since $p < \frac{1}{2}$), this means that the Fourier expansion of f is supported on the first two levels, that is, $\deg f \leq 1$. Just as in the unbiased case, this implies that f is a dictator, and so a star (since stars are the only intersecting dictators). In other words, stars are the *unique* extremal families.

We can say even more. If f is the characteristic function of an intersecting family whose measure is close to p, then the inequalities above must be nearly tight. This translates to $\|f^{>1}\|$ being small (see exercise below). In the unbiased case ($p = 1/2$), this implies that f is close to a dictator via the Friedgut–Kalai–Naor theorem, and so (since f is the characteristic function of an intersecting family) to a star.

Our proof of the Friedgut–Kalai–Naor theorem in Section 3 goes through with little changes for any *constant* p, once we replace x_i with $\omega_i = \frac{x_i - p}{\sqrt{p(1-p)}}$; perhaps the biggest difference is that $\mathbb{E}[\omega_i^4] \neq 1$ in general. Therefore, for any *constant* p, an intersecting family of almost extremal μ_p-measure is close to a star (this is known as *stability*).

Most results in Boolean function analysis generalize from $p = 1/2$ to any constant p; the most glaring exception is that $\omega_S\omega_T \neq \omega_{S\Delta T}$ in general. Moreover, these results usually hold uniformly for all $p \in [\eta, 1-\eta]$, where $\eta > 0$ is arbitrary. The case of small p is more interesting, and we will consider it carefully later on.

Exercise Fix $p < 1/2$, and suppose that $\mathcal{F}$ is an intersecting family.

(a) Show that if $\mu_p(\mathcal{F}) = p$ then $\mathcal{F}$ is a star.

(b) Assuming the Friedgut–Kalai–Naor theorem for μ_p, show that if $\mu_p(\mathcal{F}) \geq p - \epsilon$ then there exists a star $\mathcal{G}$ such that $\mu_p(\mathcal{F}\Delta\mathcal{G}) = O(\epsilon)$.

(c) Prove the Friedgut–Kalai–Naor theorem for μ_p, and show that it holds uniformly for all $p \in [\eta, 1-\eta]$, where $\eta > 0$ is arbitrary.

Exercise Two families $\mathcal{F}, \mathcal{G}$ are *cross-intersecting* if any $A \in \mathcal{F}$ and $B \in \mathcal{G}$ intersect. Show that if $p < 1/2$ and $\mathcal{F}, \mathcal{G}$ are cross-intersecting then $\mu_p(\mathcal{F})\mu_p(\mathcal{G}) \leq p^2$, with equality achieved only if $\mathcal{F} = \mathcal{G}$ is a star.

8.2 Hypercontractivity

Does hypercontractivity extend to the p-biased setting? Before answering this question, we need to define an analog of the noise operator T_ρ. A natural choice is to define it in terms of the Fourier expansion:

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

In the unbiased case, there was an equivalent definition of T_ρ, namely $T_\rho f(x) = \mathbb{E}[f(y)]$, where y is obtained from x by flipping each entry with probability $\frac{1-\rho}{2}$. A similar interpretation exists in the p-biased case, which also serves to clarify the unbiased case, as we now explain.

We are looking for a noise distribution $N_\rho(x)$ satisfying two constraints. First, if $x \sim \mu_p$ then $N_\rho(x) \sim \mu_p$. Second, the corresponding noise operator coincides with the spectral definition above. Let $x \sim \mu_p$ and $y \sim N_\rho(x)$. If the two constraints hold, then

$$\mathop{\mathbb{E}}_{x,y}\left[\frac{x-p}{\sqrt{p(1-p)}} \cdot \frac{y-p}{\sqrt{p(1-p)}}\right] = \mathop{\mathbb{E}}_{x,y}\left[\rho\left(\frac{x-p}{\sqrt{p(1-p)}}\right)^2\right] = \rho.$$

Let us consider two extreme cases. In order to get $\rho = 1$, we can simply choose $y = x$. In order to get $\rho = 0$, we can choose $y \sim \mu_p$ independent of x. By linearity, we can get any value of ρ by a mixture of these two extremes: $N_\rho(x)$ is obtained by retaining x with probability ρ, and resampling it (according to μ_p) with probability $1 - \rho$.

Armed with a definition of T_ρ, we can try to mimic the proof of hypercontractivity in Section 6 (only the simple case). We are looking for a value of ρ that satisfies $\|T_\rho f\|_4 \leq \|f\|_2$, and aiming at an inductive proof. The base case $n = 0$ holds for any value of ρ, so it suffices to consider the inductive step.

An arbitrary function f on $n+1$ variables can be written as

$$f = \omega_{n+1} g + h, \text{ where } g = \sum_{S \subseteq [n]} \hat{f}(S \cup \{n+1\}) \omega_S, h = \sum_{S \subseteq [n]} \hat{f}(S) \omega_S.$$

Retracing our steps in Section 6,

$$\mathbb{E}[(T_\rho f)^4] = \sum_{i=0}^{4} \binom{4}{i} \rho^i \, \mathbb{E}[\omega_{n+1}^i] \, \mathbb{E}[(T_\rho g)^i (T_\rho h)^{4-i}].$$

As in the unbiased case, we have $\mathbb{E}[\omega_{n+1}] = 0$ and $\mathbb{E}[\omega_{n+1}^2] = 1$. In contrast to the unbiased case, the following two moments are not $0, 1$:

$$\kappa_3 := \mathbb{E}[\omega_{n+1}^3] = \frac{(1-p)(-p)^3 + p(1-p)^3}{(p(1-p))^{3/2}} = \frac{1-2p}{\sqrt{p(1-p)}},$$
$$\kappa_4 := \mathbb{E}[\omega_{n+1}^4] = \frac{(1-p)(-p)^4 + p(1-p)^4}{(p(1-p))^2} = \frac{1-3p(1-p)}{p(1-p)}.$$

The exact values do not matter as long as p is constant. The only significant difference from the unbiased case is that κ_3 is non-zero. Nevertheless, let us try to bound $\mathbb{E}[(T_\rho f)^4]$ in terms of $\mathbb{E}[f^2]^2 = (\mathbb{E}[g^2] + \mathbb{E}[h^2])^2$:

$$\mathbb{E}[(T_\rho f)^4] = \rho^4 \kappa_4 \, \mathbb{E}[(T_\rho g)^4] + 4\rho^3 \kappa_3 \, \mathbb{E}[(T_\rho g)^3 (T_\rho h)] + 6\rho^2 \, \mathbb{E}[(T_\rho g)^2 (T_\rho h)^2] + \mathbb{E}[(T_\rho h)^4].$$

We can bound $\mathbb{E}[(T_\rho g)^4] \leq \mathbb{E}[g^2]^2$ and $\mathbb{E}[(T_\rho h)^4] \leq \mathbb{E}[h^2]^2$ by induction. Applying Cauchy–Schwarz, we can similarly bound $\mathbb{E}[(T_\rho g)^2 (T_\rho h)^2] \leq \sqrt{\mathbb{E}[(T_\rho g)^4] \, \mathbb{E}[(T_\rho h)^4]} \leq \mathbb{E}[g^2] \, \mathbb{E}[h^2]$. As for the remaining term,

$$\mathbb{E}[(T_\rho g)^3 (T_\rho h)] \leq \sqrt{\mathbb{E}[(T_\rho g)^4] \, \mathbb{E}[(T_\rho g)^2 (T_\rho h)^2]} \leq \mathbb{E}[g^2] \sqrt{\mathbb{E}[g^2] \, \mathbb{E}[h^2]} \leq \frac{\mathbb{E}[g^2]^2 + \mathbb{E}[g^2] \, \mathbb{E}[h^2]}{2},$$

using the AM-GM inequality. Altogether, we obtain

$$\mathbb{E}[(T_\rho f)^4] \leq (\rho^4 \kappa_4 + 2\rho^3 \kappa_3) \, \mathbb{E}[g^2]^2 + (2\rho^3 \kappa_3 + 6\rho^2) \, \mathbb{E}[g^2] \, \mathbb{E}[h^2] + \mathbb{E}[h^2]^2.$$

In order for this to be bounded by $(\mathbb{E}[g^2] + \mathbb{E}[h^2])^2$, we need ρ to satisfy

$$\rho^4 \kappa_4 + 2\rho^3 \kappa_3 \leq 1, \quad \rho^3 \kappa_3 + 3\rho^2 \leq 1.$$

For any fixed p, there will be a fixed ρ satisfying this. Moreover, there is a single ρ which works for all $p \in [\epsilon, 1-\epsilon]$, for any fixed $\epsilon > 0$.

How does ρ vary with p? Let us consider the setting $p \leq 1/2 - \epsilon$, in which $\kappa_3 = \Theta(\frac{1}{\sqrt{p}})$ and $\kappa_4 = \Theta(\frac{1}{p})$. The optimal choice of ρ satisfies

$$\rho = \Theta(\min((1/\kappa_4)^{1/4}, (1/\kappa_3)^{1/3}, 1)) = \Theta(\min(p^{1/4}, \sqrt{p}^{1/3}, 1)) = \Theta\left(\sqrt[4]{p}\right).$$

9 Russo–Margulis

One of the most famous applications of p-biased Boolean function analysis is to understand threshold phenomena in random graphs. The starting point is a simple connection between the μ_p-measure of a monotone property and total influence, known as the *Russo–Margulis formula* [Rus82].

A *monotone property* of subsets of $[n]$ (in the case of random graphs, subsets of $[\binom{n}{2}]$) is a collection $\mathcal{F}$ of subsets of $[n]$ which is closed upwards: if $S \in \mathcal{F}$ and $T \supseteq S$ then $T \in \mathcal{F}$. For example, the following properties are monotone: containing a certain point; containing at least $n/2$ points; the Tribes function; (for a graph) containing a triangle; being connected. If f is the characteristic function of $\mathcal{F}$, then f is monotone.

Let $\phi(p) = \mu_p(\mathcal{F})$. What does the derivative of ϕ look like? To answer this, we will compare $\phi(p)$ and $\phi(p+\epsilon)$ using a coupling of μ_p and $\mu_{p+\epsilon}$. Let $t_1, \ldots, t_n \sim U([0,1])$ and define $x_i = [t_i < p]$ and $y_i = [t_i < p+\epsilon]$. Clearly $x \sim \mu_p$ and $y \sim \mu_{p+\epsilon}$. Moreover, $x \leq y$, and so $\phi(p+\epsilon) - \phi(p)$ is the probability that $x \notin \mathcal{F}$ but $y \in \mathcal{F}$ (we identify sets with Boolean vectors).

The probability that y differs from x in more than one coordinate is $O(n^2\epsilon^2)$. The probability that y differs from x in a specific coordinate i is $\epsilon(1-\epsilon)^{n-1} = \epsilon + O(n\epsilon^2)$. It follows that

$$\phi(p+\epsilon) - \phi(p) = \epsilon \sum_{i=1}^{n} \Pr_{\mu_p}[x|_{i=0} \notin \mathcal{F} \text{ and } x|_{i=1} \in \mathcal{F}] + O_n(\epsilon^2).$$

(The notation $O_n(\epsilon^2)$ denotes a quantity bounded by $C_n\epsilon^2$, where C_n depends on n.) Taking the limit $\epsilon \to 0$, we see that

$$\phi'(p) = \sum_{i=1}^{n} \Pr_{\mu_p}[f(x_{-i}, 0) \neq f(x_{-i}, 1)].$$

This looks very similar to the formula for the total influence in the unbiased case: while we defined the i'th influence by comparing $f(x)$ and $f(x^{\oplus i})$, we would get the same result if we compared instead $f(x_{-i}, 0)$ and $f(x_{-i}, 1)$. How does this look in terms of the Fourier expansion? It suffices to consider the Fourier basis vectors. If $i \notin S$ then $\omega_S(x_{-i}, 0) = \omega_S(x_{-i}, 1)$, and otherwise

$$\omega_S(x_{-i}, 1) - \omega_S(x_{-i}, 0) = \omega_{S-i}(x)\left(\frac{1-p}{\sqrt{p(1-p)}} - \frac{-p}{\sqrt{p(1-p)}}\right) = \frac{1}{\sqrt{p(1-p)}}\omega_{S-i}(x).$$

(This is also the derivative of $\omega_S(x)$ with respect to x_i.) Therefore

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

This suggests two different ways of defining influence, differing by a factor of $p(1-p)$. We could either define it as the expectation above, or as the sum of squared Fourier coefficients. We choose the latter, although some sources prefer the former. That is, we define

$$\mathrm{Inf}_i[f] = p(1-p) \mathop{\mathbb{E}}_{\mu_p}[(f(x_{-i}, 1) - f(x_{-i}, 0))^2] = \sum_{i \in S} \hat{f}(S)^2.$$

Equivalently, if y is obtained from x by *resampling* the i'th coordinate, then $x_i \neq y_i$ with probability $2p(1-p)$, and so

$$\mathrm{Inf}_i[f] = \frac{1}{2} \mathop{\mathbb{E}}_{\mu_p}[(f(x) - f(y))^2].$$

The total influence is defined by summing over the individual influences.

Having defined total influence, we see that

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

where the superscript denotes that total influence is taken with respect to μ_p.

This simple statement already has interesting implications. If $\mathcal{F}$ is non-trivial (isn't empty and doesn't contain everything), then ϕ is strictly increasing, and so we can define $\tau(q) = \phi^{-1}(q)$ for any $q \in [0,1]$. How large can $\tau(3/4) - \tau(1/4)$ be? Nothing interesting can be said in general, but in many cases, including that of random graphs, the property in question has some symmetries, and in particular, is invariant under some transitive permutation group (a permutation group is *transitive* is for any i, j there is a permutation mapping i to j). This implies that all influences are the same. Using a p-biased version of the KKL theorem, this allows us to show that $\tau(3/4) - \tau(1/4)$ is small.

We will consider the case in which $0 \ll \tau(1/2) \ll 1$ (that is, $\epsilon \leq \tau(1/2) \leq 1 - \epsilon$ for some fixed $\epsilon > 0$), although a similar argument works in general, once we take care of the dependence on p, as shown by Friedgut and Kalai [FK96].

Let $p_0 = \tau(1/2)$. If we choose $x, y \sim \mu_{p_0}$ independently then their coordinate-wise minimum $z = \min(x, y)$ is distributed $\mu_{p_0^2}$. Also,

$$\Pr[z \in \mathcal{F}] \leq \Pr[x, y \in \mathcal{F}] = \Pr[x \in \mathcal{F}]^2 = 1/4.$$

Therefore $\tau(1/4) \geq p_0^2$. Similarly, $\tau(3/4) \leq 1 - (1-p_0)^2$. By assumption, $0 \ll \tau(1/2) \ll 1$, and so the same holds for $\tau(1/4)$ and $\tau(3/4)$.

As we stated above, a theorem such as the KKL theorem holds uniformly for all $p \in [\eta, 1-\eta]$. In particular, for all $p \in [\tau(1/4), \tau(3/4)]$ the maximal influence is $\Omega(\frac{\log n}{n} \mathbb{V}[f]) = \Omega(\frac{\log n}{n})$ (since the variance is at least $\frac{3}{16}$). Since all influences are the same, the total influence is $\Omega(\log n)$. In other words, $\phi'(p) = \Omega(\log n)$, and so

$$\tau(3/4) - \tau(1/4) = O\left(\frac{1}{\log n}\right).$$

In contrast, if f is a junta such as x_i, then $\tau(3/4) - \tau(1/4)$ is constant. This shows that coarse threshold behavior is associated with f depending on a small number of coordinates. The sharp threshold theorems of Friedgut [Fri99], Bourgain and Hatami [Hat12] prove this in a formal sense, crucially also for $p = o(1)$, which is the interesting regime for random graphs and many other settings.

Exercise

(a) Extend the KKL theorem to the p-biased setting, with an explicit dependence on p.

(b) Obtain a bound on $\tau(3/4) - \tau(1/4)$ for monotone functions invariant under a transitive permutation group, as a function of $p_0 = \tau(1/2)$ and n.

9.1 Russo–Margulis + Friedgut

One winning combination is a that of Russo–Margulis and Friedgut's junta theorem, which is often used when analyzing PCPs as part of hardness of approximation proofs. As a toy example, we will prove the following result. Fix $p < 1/2$ and $\epsilon > 0$. We will show how to associate with each monotone family $\mathcal{F}$ of μ_p-measure at least ϵ a set $L(\mathcal{F})$ of $O_{p,\epsilon}(1)$ labels such that if $\mathcal{F}$ and $\mathcal{G}$ are cross-intersecting (that is, any set in $\mathcal{F}$ intersects any set in $\mathcal{G}$) then $L(\mathcal{F})$ and $L(\mathcal{G})$ intersect. This is the technical heart of the proof that the trivial 2-approximation algorithms for vertex cover cannot be improved (assuming the unique games conjecture), due to Khot and Regev [KR08]. In this proof, the approximation ratio is roughly $1/(1-p)$, and so we want p to be close to $1/2$.

The idea is very simple. Since $\mu_p(\mathcal{F}) \geq \epsilon \geq 0$, and clearly $\mu_{1/2}(\mathcal{F}) \leq 1$, there must be a point $p_{\mathcal{F}} \in [p, 1/2]$ where the derivative of $\phi(q) = \mu_q(\mathcal{F})$ is at most $1/(1/2 - p)$. According to the Russo–Margulis formula, the total influence of the characteristic function f at $\mu_{p_{\mathcal{F}}}$ is $O_p(1)$. It is natural to choose as $L(\mathcal{F})$ all coordinates whose influence is at least some $\eta = \eta(p) > 0$. Since the total influence is $O_p(1)$, the number of labels only depends on p.

Now suppose we are given two monotone families $\mathcal{F}, \mathcal{G}$ whose label sets $L(\mathcal{F}), L(\mathcal{G})$ are disjoint. We will show that $\mathcal{F}, \mathcal{G}$ are not cross-intersecting.

The proof of Friedgut's junta theorem (which works for $\mu_{p_\mathcal{F}}$ as well) shows that $\mathcal{F}$ is δ-close with respect to $\mu_{p_\mathcal{F}}$ to a junta depending on some subset $C(\mathcal{F}) \subseteq L(\mathcal{F})$ of at most $M = M(p,\delta)$ coordinates (that is, the probability that $\mathcal{F}$ differs from the junta is at most δ). Similarly, g is δ-close with respect to $\mu_{p_\mathcal{G}}$ to a junta depending on some subset $C(\mathcal{G}) \subseteq L(\mathcal{G})$ of coordinates.

Our first step is to remove a few sets from $\mathcal{F}$ so that it doesn't depend on the coordinates in $C(\mathcal{G})$ (we will see later why this is necessary). To accomplish this, we think of $\mathcal{F}$ as a $2^{\overline{C(\mathcal{G})}} \times 2^{C(\mathcal{G})}$ Boolean table. We form a new family $\mathcal{F}'$ by removing every row which is non-constant; the resulting family is monotone since $\mathcal{F}$ is monotone. The restriction f_S of f to any such row $S \subseteq \overline{C(\mathcal{G})}$ has variance at least $p_\mathcal{F}^M(1-p_\mathcal{F}^M) \geq p^M(1-p^M)$ with respect to $\mu_{p_\mathcal{F}}$. Therefore

$$\begin{aligned}
\mu_{p_\mathcal{F}}(\mathcal{F} \setminus \mathcal{F}') &\leq \Pr_{S \subseteq \overline{C(\mathcal{G})}}[f_S \text{ is not constant}] \\
&\leq [p^M(1-p^M)]^{-1} \mathop{\mathbb{E}}_{S \subseteq \overline{C(\mathcal{G})}}[\mathbb{V}[f_S]] \\
&\overset{(1)}{\leq} [p^M(1-p^M)]^{-1} \mathop{\mathbb{E}}_{S \subseteq \overline{C(\mathcal{G})}}[\mathrm{Inf}[f_S]] \\
&= [p^M(1-p^M)]^{-1} \sum_{i \in C(\mathcal{G})} \mathop{\mathbb{E}}_{S \subseteq \overline{C(\mathcal{G})}}[\mathrm{Inf}_i(f_S)] \\
&= [p^M(1-p^M)]^{-1} \sum_{i \in C(\mathcal{G})} \mathrm{Inf}_i(f) \\
&\overset{(2)}{\leq} [p^M(1-p^M)]^{-1} M\eta,
\end{aligned}$$

where (1) is Poincaré's inequality, and (2) follows since $C(\mathcal{G}) \subseteq L(\mathcal{G})$ is disjoint from $L(\mathcal{F})$.

We will choose η so that the right-hand side is at most $\min(\delta, \epsilon/2)$, implying that

$$\mu_{p_\mathcal{F}}(\mathcal{F}') \geq \mu_{p_\mathcal{F}}(\mathcal{F}) - \frac{\epsilon}{2} \overset{(*)}{\geq} \mu_p(\mathcal{F}) - \frac{\epsilon}{2} \geq \frac{\epsilon}{2},$$

where $(*)$ holds since $\mathcal{F}$ is monotone. Note that η depends on M, δ, and therefore to prevent circular choice, δ will need to depend only on p, ϵ.

Recall that $\mathcal{F}$ is δ-close with respect to $\mu_{p_\mathcal{F}}$ to a junta depending on $C(\mathcal{F})$. Since $\mathcal{F}$ and $\mathcal{F}'$ are δ-close, this implies that $\mathcal{F}'$ is 2δ-close to such a junta. For each $A \subseteq C(\mathcal{F})$, consider the restriction $\mathcal{F}'_A$ of $\mathcal{F}'$ to those sets whose intersection with $C(\mathcal{F})$ is A. Thus

$$\mathop{\mathbb{E}}_A[\min(\mu_{p_\mathcal{F}}(\mathcal{F}'_A), 1 - \mu_{p_\mathcal{F}}(\mathcal{F}'_A))] \leq \delta,$$

where the expectation on A is taken with respect to $\mu_{p_\mathcal{F}}$ restricted to $C(\mathcal{F})$. In particular, $\min(\mu_{p_\mathcal{F}}(\mathcal{F}'_A), 1 - \mu_{p_\mathcal{F}}(\mathcal{F}'_A)) \leq \sqrt{\delta}$ for all but a $\sqrt{\delta}$-fraction of A's (with respect to $\mu_{p_\mathcal{F}}$!). Such A's contribute at most $\sqrt{\delta}$ to the total measure of $\mathcal{F}'$. Sets A such that $\mu_{p_\mathcal{F}}(\mathcal{F}'_A) \leq \sqrt{\delta}$ contribute another $\sqrt{\delta}$, and so if $\epsilon/2 > 2\sqrt{\delta}$, there must be a set A such that $\mu_{p_\mathcal{F}}(\mathcal{F}'|_A) \geq 1 - \sqrt{\delta}$. Accordingly, we choose $\delta = (\epsilon/5)^2$.

Since $\mathcal{F}'$ doesn't depend on the coordinates in $C(\mathcal{G})$, the same holds for $\mathcal{F}'_A$, and we conclude that the following set has $\mu_{p_\mathcal{F}}$-measure at least $1 - \epsilon/5$, with respect to the coordinates outside of $C(\mathcal{F}) \cup C(\mathcal{G})$:

$$\mathcal{F}^* = \{S \in \mathcal{F}' : S \cap (C(\mathcal{F}) \cup C(\mathcal{G})) = A\}.$$

Since $\mathcal{F}'$ is monotone, $\mu_{1/2}(\mathcal{F}^*) \geq 1 - \epsilon/5$ as well.

In a completely analogous way, we can find $B \subseteq C(\mathcal{G})$ such that the following family has $\mu_{1/2}$-measure at least $1 - \epsilon/5$, again with respect to the coordinates outside of $C(\mathcal{F}) \cup C(\mathcal{G})$:

$$\mathcal{G}^* = \{S \in \mathcal{F}' : S \cap (C(\mathcal{F}) \cup C(\mathcal{G})) = B\}.$$

We are finally ready to show that $\mathcal{F}, \mathcal{G}$ are not cross-intersecting. Choose a set $T \subseteq \overline{C(\mathcal{F}) \cup C(\mathcal{G})}$ uniformly at random. The probability that $A \cup T \in \mathcal{F}^*$ and $B \cup (C(\mathcal{F}) \cup C(\mathcal{G}) \setminus T) \in \mathcal{G}^*$ is at least $1 - 2\epsilon/5 > 0$, and so such a set T exists. Since $A \subseteq C(\mathcal{F})$ and $B \subseteq C(\mathcal{G})$, the two sets $A \cup T$ and $B \cup (C(\mathcal{F}) \cup C(\mathcal{G}) \setminus T)$ are disjoint, completing the proof.

10 Very biased Fourier analysis: Biased FKN theorem

How does p-biased Fourier analysis differ from standard Fourier analysis (the case $p = 1/2$)? When $0 \ll p \ll 1$, the behavior is very similar, although one crucial property of the Fourier characters is lost, namely, they are no longer characters of the group $\mathbb{Z}_2^n$ (no longer satisfy $x_S x_T = x_{S \Delta T}$), which makes it harder to analyze linearity testing, for example.

The situation greatly differs when p is very small (or very close to 1), which is of interest in random graph theory, among else: $G(n,p)$ random graphs often exhibit their most interesting behavior for sub-constant p. For example, the threshold for connectivity is $p = \frac{\log n}{n}$.

The Russo–Margulis formula allows understanding the speed of threshold phenomena in terms of total influence: using the terminology of Section 9, if $\tau(2/3) - \tau(1/3)$ is large compared to $\tau(1/2)$, then this means that the total influence around $\tau(1/2)$ is relatively small. If the threshold is bounded away from $0, 1$, then Friedgut's junta theorem implies that the graph property is essentially a junta, which is impossible for graph properties. For sub-constant p, the characterization (due to Friedgut, Bourgain and Hatami) is much more complex, and only states that the graph property is "global", that is, affected only slightly by local events.

As an example of very biased Fourier analysis, we generalize the FKN theorem 3 to this setting. Recall that the Friedgut–Kalai–Naor (FKN) theorem states that if f is a degree 1 function which is close to Boolean, in the sense that $\mathbb{E}[\mathrm{dist}(f, \{\pm 1\})^2] = \epsilon$, then f is $O(\epsilon)$-close to a Boolean dictatorship. What happens for other values of p?

When dealing with several different values of p, it is often advantageous to switch to $\{0,1\}$-valued functions and variables, which to avoid confusion we will name $y_1, \ldots, y_n$. Thus the function f can be written in the form

$$f = c_0 + \sum_{i=1}^{n} c_i y_i,$$

and the FKN theorem states that if f is close to Boolean then it is close to one of the following functions: $0, 1, y_i, 1 - y_i$. The same result holds for any constant value of p. Note that if we replaced y_i with $\omega_i = \frac{y_i - p}{\sqrt{p(1-p)}}$, the coefficients in front of ω_i would have to depend on p.

What happens when p is small? Other examples arise. For example, the function $y_1 + y_2$ is quite close to Boolean, since the probability that both y_1 and y_2 equal 1 is only p^2. More generally, we can consider the function $f_m = \sum_{i=1}^{m} y_i$ or its negation. If $m = \lambda/p$ then the distribution of f_m is roughly Poisson with expectation λ, and so

$$\Pr[f_m \notin \{0,1\}] \approx 1 - e^{-\lambda}(1 + \lambda) = \Theta(\lambda^2).$$

Moreover,

$$\mathbb{E}[f_m^2] = m\,\mathbb{E}[y_i^2] + m(m-1)\,\mathbb{E}[y_i y_j] \approx \lambda + \lambda^2.$$

This shows that as long as $\lambda = O(\sqrt{\epsilon})$, the function f_m will be $O(\epsilon)$-close to Boolean.

The FKN theorem for small p [Fil16] states that this is the only example: if $p \leq 1/2$ and f is a degree 1 function such that $\mathbb{E}[\mathrm{dist}(f, \{0,1\})^2] \leq \epsilon$, then either f or $1 - f$ is close to a sum of $O(\sqrt{\epsilon}/p)$ coordinates. (In fact, this is slightly wrong, can you see why?)

Recall that the FKN theorem can also be stated analogously for Boolean functions F which are close to degree 1, in the sense that $\|F^{>1}\|^2 \leq \epsilon$. A simple corollary of the formulation above is that either F or $1 - F$ is close to the maximum of $O(\sqrt{\epsilon}/p)$ coordinates.

In the rest of this subsection, we prove the FKN theorem for small p in full, starting with its first formulation. That is, we are given a linear function

$$f = c_0 + \sum_{i=1}^{n} c_i y_i,$$

where $y_i \in \{0,1\}$, and we are promised that for some $p \leq 1/2$,

$$\mathop{\mathbb{E}}_{\mu_p}[\mathrm{dist}(f,\{0,1\})^2] \leq \epsilon.$$

As in the proof of the unbiased FKN theorem, we can assume that ϵ is "small enough", that is, smaller than some constant. Indeed, if $\epsilon \geq \epsilon_0$ for some $\epsilon_0 > 0$ then since $x^2 \leq 2(x-1)^2 + 2$,

$$\mathbb{E}[f^2] \leq 2 + 2\,\mathbb{E}[\mathrm{dist}(f,\{0,1\})^2] \leq 2 + 2\epsilon = O(\epsilon),$$

since $2 \leq (2/\epsilon_0)\epsilon$.

The main idea behind the proof is that we can sample from μ_p in two steps: first, sample a set S according to μ_{2p}, and substitute 0 for every variable outside of S. Then, sample a point according to $\mu_{1/2}$ restricted to S. In other words, if we sample $S \sim \mu_{2p}$ and a subset $T \sim \mu_{1/2}(S)$ (that is, each element of S is included in T with probability $1/2$), then $T \sim \mu_p$.

Accordingly, for every $S \subseteq [n]$ we define a function $f_S \colon \{0,1\}^S \to \mathbb{R}$ by $f_S(y_S) = f(y_S, 0)$, where the second argument on the right-hand side corresponds to the coordinates outside S. Let $\epsilon_S = \mathbb{E}_{\mu_{1/2}}[\mathrm{dist}(f_S,\{0,1\})^2]$. Then

$$\mathop{\mathbb{E}}_{S\sim\mu_{2p}}[\epsilon_S] = \epsilon.$$

Let $\tilde{f}_S = 2f_S - 1$, a function which is $O(\epsilon_S)$-close to $\{\pm 1\}$. How does the function $\tilde{f}_S$ look like in the unbiased Fourier basis? Using the conversion formula $x_i = 2y_i - 1$,

$$\tilde{f}_S = 2c_0 - 1 + 2\sum_{i=1}^{n} c_i \frac{x_i + 1}{2} = 2c_0 + \sum_{i=1}^{n} c_i - 1 + \sum_{i=1}^{n} c_i x_i.$$

The unbiased FKN theorem states that there exists a Boolean dictatorship g_S such that $\|\tilde{f}_S - g_S\|^2 = O(\epsilon_S)$. Since g_S is a Boolean dictatorship, $\widehat{g_S}(\{i\}) \in \{0, \pm 1\}$, and so

$$O(\epsilon_S) \geq \|\tilde{f}_S - g_S\|^2 \geq \sum_{i\in S} (\hat{f}(\{i\}) - \widehat{g_S}(\{i\}))^2 \geq \sum_{i\in S} \mathrm{dist}(c_i, \{0,\pm 1\})^2.$$

The idea now is to take expectation with respect to $S \sim \mu_{2p}$. On the left-hand side we get $O(\epsilon)$. Since $\Pr[i \in S] = 2p$ for every $i \in [n]$, on the right-hand side we get $2p\sum_i \mathrm{dist}(c_i, \{0,\pm 1\})^2$. Over all, this shows that

$$\sum_{i=1}^{n} \mathrm{dist}(c_i, \{0,\pm 1\})^2 = O\left(\frac{\epsilon}{p}\right).$$

Let $d_i = \mathrm{round}(c_i, \{0,\pm 1\})$. It is natural to "round" the coefficients c_i to d_i. In order to do this properly, we switch to the orthogonal variables $\omega_i = \frac{y_i - p}{\sqrt{p(1-p)}}$. Writing f in these terms, we get

$$f = c_0 - p\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} c_i(y_i - p) = c_0 - p\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} \sqrt{p(1-p)}\, c_i \omega_i.$$

We form a new function g by replacing c_i with d_i:

$$g = c_0 - p\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} c_i(y_i - p) = c_0 - p\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} \sqrt{p(1-p)}\, d_i \omega_i.$$

We only changed the level one Fourier coefficients, since this ensures that we can estimate $\|f-g\|^2$:

$$\|f-g\|^2 = \sum_{i=1}^{n} (\sqrt{p(1-p)}c_i - \sqrt{p(1-p)}d_i)^2 = p(1-p)\sum_{i=1}^{n}(c_i-d_i)^2 = O(\epsilon).$$

In particular, using $(a+b)^2 \leq 2a^2+2b^2$, we see that g is also close to Boolean:

$$\mathbb{E}[\mathrm{dist}(g,\{0,1\})^2] \leq 2\,\mathbb{E}[(g-f)^2] + 2\,\mathbb{E}[\mathrm{dist}(f,\{0,1\})^2] = O(\epsilon).$$

A similar calculation shows that not too many d_i's can be non-zero. On the one hand, $\|g\|^2 \leq 2 + 2\,\mathbb{E}[\mathrm{dist}(g,\{0,1\})^2] = O(1)$. On the other hand,

$$\|g\|^2 \geq \sum_{i=1}^{n} \hat{g}(\{i\})^2 = p(1-p)\sum_{i=1}^{n} d_i^2.$$

This shows that $\sum_{i=1}^{n} d_i^2 = O(1/p)$, and so at most $O(1/p)$ of the d_i can be non-zero.

How does the function g look like in terms of the original variables y_i? Reversing the calculations above, for some $b \in \mathbb{R}$ we have

$$g = b + \sum_{i=1}^{n} d_i y_i.$$

We can write this even more simply: let I be the set of indices such that $d_i = 1$, and let J be the set of indices such that $d_j = -1$. Then

$$g = b + \sum_{i\in I} y_i - \sum_{j\in J} y_j.$$

Since at most $O(1/p)$ are non-zero, $|I|, |J| = O(1/p)$.

The probability that all y_i in $I \cup J$ are zero is $(1-p)^{|I|\cup|J|} = (1-p)^{O(1/p)} = \Omega(1)$. This shows that

$$\mathbb{E}[\mathrm{dist}(b,\{0,1\})^2] = \mathbb{E}[\mathrm{dist}(g,\{0,1\})^2 \mid y_i = 0 \text{ for all } i \in I\cup J] \leq \frac{\mathbb{E}[\mathrm{dist}(g,\{0,1\})^2]}{\Pr[y_i = 0 \text{ for all } i \in I\cup J]} = O(\epsilon).$$

Taking $d = \mathrm{round}(b,\{0,1\})$, this shows that the function

$$h = d + \sum_{i\in I} y_i - \sum_{j\in J} y_j$$

satisfies $\|g-h\|^2 = O(\epsilon)$ and so, as before, $\|f-h\|^2 = O(\epsilon)$ and $\mathbb{E}[\mathrm{dist}(h,\{0,1\})^2] = O(\epsilon)$.

The cases $d=0$ and $d=1$ are quite similar, so from now on we assume that $d=0$, and approximate the structure of f; if $d=1$, the same structure will hold for $1-f$. Thus from now on, we assume that

$$h = \sum_{i\in I} y_i - \sum_{j\in J} := h_I - h_J.$$

Let us start by getting rid of J. We know that $\Pr[h_I = 0] = \Omega(1)$, and so $\mathbb{E}[\mathrm{dist}(-h_J,\{0,1\})^2] = O(\epsilon)$. Since $h_J \geq 0$, this implies that $\mathbb{E}[h_J^2] = O(\epsilon)$. Therefore $\|h_I - f\|^2 = O(\epsilon)$ and $\mathbb{E}[\mathrm{dist}(h_I,\{0,1\})^2] = O(\epsilon)$.

Now we bound the size of I, by looking at the probability that $h_I = 2$:

$$\Pr[h_I = 2] = \binom{|I|}{2} p^2 (1-p)^{|I|-2} \overset{(*)}{=} \Omega(|I|^2 p^2).$$

The starred lower bound holds as long as $|I| \neq 1$. This shows that either $|I| = 1$, or $|I| = O(\sqrt{\epsilon}/p)$.

Notice that

$$\|h_I\|^2 \leq |I|p + |I|^2 p^2 = O(p + \sqrt{\epsilon}),$$

and so the same holds for $\|f\|^2$. In other words, f is actually somewhat close to a *constant* function. Summarizing:

Let $f\colon \{0,1\}^n \to \mathbb{R}$ be a degree 1 polynomial satisfying $\mathbb{E}_{\mu_p}[\mathrm{dist}(f,\{0,1\})^2] \leq \epsilon$, where $p \leq 1/2$. Then there exists a set I, of size at most $\max(1, O(\sqrt{\epsilon}/p))$, such that

$$\min\left(\left\|\sum_{i\in I} y_i - f\right\|^2, \left\|1 - \sum_{i\in I} y_i - f\right\|^2\right) = O(\epsilon).$$

Moreover, there exists $d \in \{0,1\}$ such that

$$\|f - d\|^2 = O(p + \sqrt{\epsilon}).$$

Finally, let us see what all this implies for Boolean functions $F\colon \{0,1\}^n \to \{0,1\}$ which are close to degree one, in the sense that $\|F^{>1}\|^2 = \epsilon$. Taking $f = F^{\leq 1}$, we get $\|f - F\|^2 = \epsilon$. Since F is Boolean, in particular $\mathbb{E}[\mathrm{dist}(f,\{0,1\})^2] \leq \epsilon$. Therefore we can approximate either f or $1-f$ by a function of the form h_I, where $|I| \leq \max(1, O(\sqrt{\epsilon}/p))$.

Without loss of generality, let us assume that it is f which is close to h_I, that is, $\|f - h_I\|^2 = O(\epsilon)$. Since we are going to replace f with the Boolean function F, it is natural to replace h_I with the Boolean function H_I given by

$$H_I = \bigvee_{i\in I} y_i,$$

that is, the *maximum* of the coordinates I. If $|I| = 1$ then $H_I = h_I$. Otherwise,

$$\|h_I - H_I\|^2 = \mathbb{E}[(h_I - 1)^2 1_{h_i \geq 2}] = \mathbb{E}[(h_I - 1)^2] - \Pr[h_I = 0].$$

Since $\mathbb{E}[h_I^2] \leq |I|p + |I|^2p^2$, $\mathbb{E}[h_I] = |I|p$, and $\Pr[h_I = 0] \geq 1 - |I|p$, this shows that

$$\|h_I - H_I\|^2 = \mathbb{E}[h_I^2] - 2\,\mathbb{E}[h_I] + 1 - \Pr[h_I = 0] \leq (|I|p + |I|^2p^2) - 2|I|p + 1 - (1 - |I|p) = |I|^2p^2 = O(\epsilon).$$

Since $\|f - h_I\|^2 = O(\epsilon)$, $\|f - F\|^2 = \epsilon$, and $\|h_I - H_I\|^2 = O(\epsilon)$, we deduce that $\|F - H_I\|^2 = O(\epsilon)$. Since F and H_I are both Boolean, in fact $\Pr[F \neq H_I] = O(\epsilon)$. In summary:

Let $F\colon \{0,1\}^n \to \{0,1\}$ be a Boolean function satisfying $\|F^{>1}\|^2 \leq \epsilon$ with respect to μ_p, where $p \leq 1/2$. Then there exists a set I, of size at most $\max(1, O(\sqrt{\epsilon}/p))$, such that

$$\min\left(\Pr\Big[F \neq \bigvee_{i\in I} y_i\Big], \Pr\Big[1 - F \neq \bigvee_{i\in I} y_i\Big]\right) = O(\epsilon).$$

Moreover, there exists $d \in \{0,1\}$ such that

$$\Pr[F \neq d] = O(p + \sqrt{\epsilon}).$$

In this theorem, it is important to note that the definition of $F^{>1}$ itself also depends on p.

Exercise Generalize the results of this subsection to the case of an arbitrary product measure on $\{0,1\}^n$.

11 Invariance principle

One of the most celebrated results in probability theory is the *central limit theorem*, which states (in one version) that the sum of many i.i.d. "reasonable" random variables behaves like a Gaussian random variable.

The assumption that the variables are not only independent but also identically distributed is crucial. Indeed, a sum of the form

$$x_1 + \frac{x_2 + \cdots + x_n}{n},$$

where $(x_1, \ldots, x_n)$ is a random point in $\{\pm 1\}^n$, does *not* look like a Gaussian. The problem is that x_1 has too much influence on the sum, compared to the other variables.

The Berry–Esseen[1] theorem is one way to quantify this phenomenon. Suppose that $X_1, \ldots, X_n$ are independent zero-mean variables with second moments $\mathbb{E}[X_i^2] = \sigma_i^2$ and third moments $\mathbb{E}[|X_i|^3] = \rho_i$. The sum $X_1 + \cdots + X_n$ has zero mean and variance $\sigma^2 = \sum_i \sigma_i^2$. The Berry–Esseen theorem states that for every $t \in \mathbb{R}$,

$$|\Pr[X_1 + \cdots + X_n < t] - \Pr[N(0, \sigma^2) < t]| = O\left(\frac{\sum_{i=1}^n \rho_i}{\sigma^3}\right),$$

where $N(\mu, \sigma^2)$ is a Gaussian random variable with mean μ and variance σ^2.

Let us see what this implies in the special case $X_i = c_i x_i$, where $(x_1, \ldots, x_n)$ is a random point in $\{\pm 1\}^n$. The second and third moments are easily calculated to be

$$\sigma_i^2 = c_i^2, \quad \rho_i = |c_i|^3.$$

Therefore the sum is close to a Gaussian if

$$\sum_{i=1}^n |c_i|^3 \ll \left(\sum_{i=1}^n c_i^2\right)^{3/2}.$$

This condition is a bit hard to work with. To better understand the picture, let us notice that

$$\sum_{i=1}^n |c_i|^3 \leq \max_i |c_i| \cdot \sum_{i=1}^n c_i^2,$$

and so the condition holds whenever

$$\max_i c_i^2 \ll \sum_{i=1}^n c_i^2,$$

that is, whenever none of the squared weights is "prominent" compared to the sum of squared weights.

The invariance principle is a similar statement for polynomials. In contrast to the linear case, when dealing with polynomials, we cannot say that a "smooth" polynomial behaves like a Gaussian. For example, suppose that $(x_1, \ldots, x_n)$ is a random point in $\{\pm 1\}^n$, and consider the polynomial

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

This behaves not like a Gaussian, but rather like the *square* of a Gaussian. Similarly,

$$\left(\sum_{i=1}^{n/2} x_i\right)^2 + \sum_{i=n/2+1}^n x_i$$

behaves like the sum of a squared Gaussian and an independent Gaussian. Therefore we need to change the statement of the result.

The basic idea is that in the context of the Berry–Esseen theorem, we can "convert" $X_1 + \cdots + X_n$ to $N(0, \sigma^2)$ (recall that $\sigma^2 = \sum_i \sigma_i^2$, where $\sigma_i^2 = \mathbb{E}[X_i^2]$) by replacing each X_i by a Gaussian random variable $G_i = N(0, \sigma_i^2)$ with the same mean and variance. This kind of replacement is something that makes sense for arbitrary polynomials.

We will be interested in functions on the Boolean cube $\{\pm 1\}^n$. For such functions, we cannot consider arbitrary polynomials. Indeed, consider the quadratic

$$\frac{\sum_{i=1}^n x_i^2}{\sqrt{n}}.$$

[1]Sometimes spelled Esséen. The name is apparently pronounced as in English, with the stress on the first syllable.

On the Boolean cube, this is just the constant $\sqrt{n}$, but if we replace each x_i by a standard Gaussian, then we obtain $N(0,3)$ (since $\mathbb{E}[N(0,1)^4] = 3$)! This kind of problem goes away if we concentrate on multilinear polynomials.

We have arrived at the following problem. Let P be a multilinear polynomial, let $X_1, \ldots, X_n$ be random i.i.d. variables uniformly distributed on $\{\pm 1\}$, and let $G_1, \ldots, G_n$ be random i.i.d. standard Gaussians. When are the distributions of $P(X_1, \ldots, X_n)$ and $P(G_1, \ldots, G_n)$ close?

There are many notions of closeness that one can consider. For example, the Berry–Esseen theorem considers closeness in CDF, also known as *Kolmogorov–Smirnov distance*. It turns out that it is much easier to consider closeness with respect to test functions. That is, given a "nice" function $\Phi\colon \mathbb{R} \to \mathbb{R}$, when can we bound

$$|\,\mathbb{E}[\Phi(P(X_1, \ldots, X_n))] - \mathbb{E}[\Phi(P(G_1, \ldots, G_n))]|?$$

It is natural to try a hybrid argument, in this context known as the replacement technique. In this argument, we don't compare $(X_1, \ldots, X_n)$ and $(G_1, \ldots, G_n)$ directly. Instead, we consider a sequence of distributions, each differing only in a single random variable:

$$\begin{aligned}
&(X_1, X_2, X_3, \ldots, X_n)\\
&(G_1, X_2, X_3, \ldots, X_n)\\
&(G_1, G_2, X_3, \ldots, X_n)\\
&\ldots\\
&(G_1, G_2, G_3, \ldots, G_n)
\end{aligned}$$

We bound the difference in expectation of two neighboring distributions, and then take the sum of differences. This is a standard technique in cryptography. Let's try to apply it here and see what happens.

We wish to bound

$$\begin{aligned}
&|\,\mathbb{E}[\Phi(P(G_1, \ldots, G_{i-1}, X_i, X_{i+1}, \ldots, X_n))] - \mathbb{E}[\Phi(P(G_1, \ldots, G_{i-1}, G_i, X_{i+1}, \ldots, X_n))]| \leq\\
&\qquad \mathop{\mathbb{E}}_{\substack{G_1,\ldots,G_{i-1}\\ X_{i+1},\ldots,X_n}} [|\mathop{\mathbb{E}}_{X_i}[\Phi(P(G_1, \ldots, G_{i-1}, X_i, X_{i+1}, \ldots, X_n))] - \mathop{\mathbb{E}}_{G_i}[\Phi(P(G_1, \ldots, G_{i-1}, G_i, X_{i+1}, \ldots, X_n)]|].
\end{aligned}$$

Once we fix $G_1, \ldots, G_{i-1}, X_{i+1}, \ldots, X_n$, the polynomial becomes a function on a single variable, say

$$P(G_1, \ldots, G_{i-1}, z, X_{i+1}, \ldots, X_n) = Az + B.$$

We wish to argue that $\mathbb{E}[\Phi(AX_i + B)] \approx \mathbb{E}[\Phi(AG_i + B)]$. We can expect this to be the case if A is small on average. We chose G_i in such a way that X_i and G_i have the same first and second moment, and they also happen to have an identical third moment. This suggests aiming at an expression involving powers of X_i. We can obtain such an expression using a Taylor series for Φ:

$$\Phi(Az + B) = \Phi(B) + \Phi'(B)Az + \Phi''(B)\frac{A^2z^2}{2} + \Phi'''(B)\frac{A^3z^3}{6} + \Phi''''(A\theta z + B)\frac{A^4z^4}{24},$$

for some $\theta \in [0,1]$. Suppose that $|\Phi''''|$ is always bounded by some K. Since X_i and G_i have identical first, second, and third moments,

$$|\,\mathbb{E}[\Phi(AX_i + B)] - \mathbb{E}[\Phi(AG_i + B)]| \leq \frac{KA^4}{24}|\,\mathbb{E}[X_i^4] + \mathbb{E}[G_i^4]| = O(KA^4).$$

In other words, the difference in expectation under the two hybrid distributions is small if $\mathbb{E}[A^4]$ is small. To see whether this is the case, we need a formula for A. Recalling the operator L_i from Section 4,

$$A = L_i P(G_1, \ldots, G_{i-1}, \cdot, X_{i+1}, \ldots, X_n),$$

where the value of the placeholder makes no difference. We have thus shown that

$$|\mathbb{E}[\Phi(P(G_1,\ldots,G_{i-1},X_i,X_{i+1},\ldots,X_n))] - \mathbb{E}[\Phi(P(G_1,\ldots,G_{i-1},G_i,X_{i+1},\ldots,X_n))]| \leq O(K)\cdot \mathop{\mathbb{E}}_{\substack{G_1,\ldots,G_{i-1}\\X_{i+1},\ldots,X_n}}[(L_iP)^4].$$

What can we say about $\mathbb{E}[(L_iP)^4]$? Before answering that, let us note that

$$\mathop{\mathbb{E}}_{\substack{G_1,\ldots,G_i\\X_{i+1},\ldots,X_n}}[(L_iP)^2] = \sum_{\substack{S\subseteq[n]\\i\in S}} \hat{P}(S)^2 = \mathrm{Inf}_i[P].$$

This is clear if we replaced $G_1,\ldots,G_{i-1}$ by $X_1,\ldots,X_{i-1}$, and it still holds with the Gaussian variables, since the calculation only uses the first two moments, which are identical for X_j and G_j.

Again replacing $G_1,\ldots,G_{i-1}$ with $X_1,\ldots,X_{i-1}$, we can bound $\mathbb{E}[(L_iP)^4]$ by $9^{\deg P}\,\mathbb{E}[(L_iP)^2]^2$, using hypercontractivity (see Section 6). The proof of hypercontractivity in Section 6 easily extends to the Gaussian case, and so

$$|\mathbb{E}[\Phi(P(G_1,\ldots,G_{i-1},X_i,X_{i+1},\ldots,X_n))] - \mathbb{E}[\Phi(P(G_1,\ldots,G_{i-1},G_i,X_{i+1},\ldots,X_n))]| \leq O(K9^{\deg P}\,\mathrm{Inf}_i[P]^2).$$

Summing over all i, this shows that

$$|\mathbb{E}[\Phi(P(X_1,\ldots,X_n))] - \mathbb{E}[\Phi(P(G_1,\ldots,G_n))]| = O\left(K9^{\deg P}\sum_{i=1}^n \mathrm{Inf}_i[P]^2\right).$$

We can simplify this expression using the bound $\mathrm{Inf}[p] \leq \deg P\cdot\mathbb{V}[P]$, proved in Section 4:

$$|\mathbb{E}[\Phi(P(X_1,\ldots,X_n))] - \mathbb{E}[\Phi(P(G_1,\ldots,G_n))]| = O\left(K2^{O(\deg P)}\,\mathbb{V}[P]\max_i \mathrm{Inf}_i[P]\right).$$

Apart from the normalizing factor $\mathbb{V}[P]$, this states that if P has low degree and low influences, then for any test function Φ with bounded fourth derivative, $\Phi(P)$ behaves the same under both the uniform distribution over $\{\pm1\}^n$ and the standard n-dimensional Gaussian distribution. This is one form of the celebrated *invariance principle* [MOO10]:

> Let $\Phi\colon\mathbb{R}\to\mathbb{R}$ be a function satisfying $|\Phi''''(z)|\leq K$ for all $z\in\mathbb{R}$. If P is a degree d multilinear polynomial with maximal influence τ then
>
> $$|\mathbb{E}[\Phi(P(X_1,\ldots,X_n))] - \mathbb{E}[\Phi(P(G_1,\ldots,G_n))]| = O(K2^{O(d)}\tau\,\mathbb{V}[P]),$$
>
> where $(X_1,\ldots,X_n)$ is the uniform distribution on $\{\pm1\}^n$, and $(G_1,\ldots,G_n)$ are i.i.d. standard Gaussians.

It is possible to deduce closeness in CDF (as in the Berry–Esseen theorem) from this result by using suitable test functions. The idea is to approximate the function $\Phi(z) = [z<t]$ by two other functions Φ_ℓ,Φ_u, with bounded fourth derivatives, that approximate Φ in the following senses:

1. $\Phi_\ell\leq\Phi\leq\Phi_u$.
2. Φ_ℓ and Φ_u agree with Φ outside a small neighborhood of t.

The invariance principle shows that

$$\mathbb{E}[\Phi_\ell(P(X_1,\ldots,X_n))] \approx \mathbb{E}[\Phi_\ell(P(G_1,\ldots,G_n))].$$

If we show that

$$\mathbb{E}[\Phi_\ell(P(G_1,\ldots,G_n))] \approx \mathbb{E}[\Phi(P(G_1,\ldots,G_n))],$$

then we can conclude that

$$\mathbb{E}[\Phi(P(X_1,\ldots,X_n))] \geq \mathbb{E}[\Phi_\ell(P(X_1,\ldots,X_n))] \approx \mathbb{E}[\Phi(P(G_1,\ldots,G_n))],$$

and repeating the same argument with Φ_u would show a similar upper bound.

Why does $\mathbb{E}[\Phi_\ell(P(\mathbf{G}))] \approx \mathbb{E}[\Phi(P(\mathbf{G}))]$? The two functions Φ_ℓ, Φ differ only in a neighborhood of t, so this amounts to showing that $P(G_1,\ldots,G_n)$ cannot be too concentrated in any small neighborhood, a phenomenon known as Gaussian anti-concentration and proved by Carbery and Wright [CW01].

Exercise Show that $\|T_{1/\sqrt{3}} f\|_4 \leq \|f\|_2$ when the input consists of m standard Gaussians and $n-m$ Boolean variables uniformly distributed on $\{\pm 1\}$, and conclude that $\mathbb{E}[f^4] \leq 9^{\deg f}\,\mathbb{E}[f^2]^2$ for such functions.

11.1 Application: Majority is Stablest

Noise sensitivity Suppose that $f\colon \{\pm 1\}^n \to \{\pm 1\}$ represents the outcome of an election, just like in Section 4. How sensitive is the election to random errors in reading individual votes? Suppose that each vote is flipped with probability $p < 1/2$. Denoting the original votes by $x = x_1,\ldots,x_n$, the new votes by $y = y_1,\ldots,y_n$, and assuming as usual that the original votes are random, the probability that the outcome changes is

$$\Pr[f(x) \neq f(y)] = \Pr[f(x)f(y) = -1] = \mathbb{E}\left[\frac{1 - f(x)f(y)}{2}\right].$$

Notice that $\mathbf{y} \sim N_\rho(\mathbf{x})$ for $\rho = 1-2p$, where N_ρ is the distribution used to define the noise operator T_ρ. Thus

$$\Pr[f(x) \neq f(y)] = \frac{1}{2} - \frac{1}{2}\,\mathbb{E}[f(x)T_\rho f(x)] = \frac{1}{2} - \frac{1}{2}\langle f, T_\rho f\rangle.$$

Thus the stability of f depends only $\langle f, T_\rho f\rangle$: the larger $\langle f, T_\rho f\rangle$ is, the more stable the elections are. The quantity $\langle f, T_\rho f\rangle$ is known as the *noise sensitivity* of f.

Which balanced function f is the most stable? Since f is balanced,

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

and this bound is achieved by functions of degree 1. As we have seen in Section 2, such functions are dictators.

Dictators are not a good choice for a voting system. What happens if we decree that no voter have too much of an influence on the outcome of the election? That is, what if we ask that the maximal influence be small? According to the invariance principle, in this case the function f behaves as if it lived on Gaussian space. What is the analog of noise sensitivity in Gaussian space?

Rotation sensitivity The noise operator T_ρ has a counterpart in Gaussian space. Recall that in Section 6.2, we defined a distribution N_ρ on *pairs* of points in $\{\pm 1\}^n$. This distribution had the following features: if $(x, y) \sim N_\rho$ then:

1. The marginals x, y are distributed uniformly over $\{\pm 1\}^n$.
2. For every i, the correlation between x_i and y_i is $\mathbb{E}[x_i y_i] = \rho$.
3. For any two functions f, g, $\langle f, T_\rho g\rangle = \mathbb{E}_{(x,y)\sim N_\rho}[f(x)g(y)]$.

We can construct a similar distribution on Gaussian space using bivariate Gaussians:

$$(x_i, y_i) \sim N\left(\begin{pmatrix}0 & 0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

In words, (x_i, y_i) is a bivariate Gaussian with zero mean and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Concretely, given x_i, we can sample y_i using the formula

$$y_i = \rho x_i + \sqrt{1-\rho^2} \cdot z_i, \text{ where } z_i \sim N(0,1).$$

Indeed, it is not hard to check that $\mathbb{E}[x_i y_i] = \rho$ and $\mathbb{E}[y_i^2] = \rho^2 + (\sqrt{1-\rho^2})^2 = 1$.

Now suppose that $w_i \sim N(0,1)$, and define

$$y_i = \rho_1 x_i + \sqrt{1-\rho_1^2} \cdot w_i,$$
$$z_i = \rho_2 x_i - \sqrt{1-\rho_2^2} \cdot w_i.$$

Clearly $(x_i, y_i) \sim N_{\rho_1}$ and $(x_i, z_i) \sim N_{\rho_2}$. What about (y_i, z_i)? Since

$$\mathbb{E}[y_i z_i] = \rho_1 \rho_2 - \sqrt{1-\rho_1^2}\sqrt{1-\rho_2^2},$$

also $(y_i, z_i) \sim N_\rho$ for a suitable ρ. We get a much simpler formula if we take $\theta_1 = \cos^{-1}\rho_1$ and $\theta_2 = \cos^{-1}\rho_2$: the addition formula for cosine shows that $\cos^{-1}\rho = \theta_1 + \theta_2$, if we choose $\theta_1, \theta_2 \in [0, \pi]$.[2]

If f is a Boolean function on Gaussian space then

$$\Pr[f(y) \neq f(z)] \leq \Pr[f(y) \neq f(x)] + \Pr[f(z) \neq f(x)],$$

and so, using the notation R_θ for $N_{\cos\theta}$, if $\theta_1, \theta_2 \in [0, \pi]$ then

$$\Pr_{(x,y)\sim R_{\theta_1+\theta_2}}[f(x) \neq f(y)] \leq \Pr_{(x,y)\sim R_{\theta_1}}[f(x) \neq f(y)] + \Pr_{(x,y)\sim R_{\theta_2}}[f(x) \neq f(y)].$$

The same formula holds for any number of summands, and in particular, for any integer $\ell \geq 1$,

$$\Pr_{(x,y)\sim R_{\pi/2}}[f(x) \neq f(y)] \leq \ell \Pr_{(x,y)\sim R_{\pi/2\ell}}[f(x) \neq f(y)].$$

Since $\cos(\pi/2) = 0$, the distribution $R_{\pi/2} = N_0$ consists of two independent Gaussians, and so the left-hand side is just

$$\Pr_{x,y}[f(x) \neq f(y)] = \mathop{\mathbb{E}}_{x,y}\left[\frac{1 - f(x)f(y)}{2}\right] = \frac{\mathbb{E}[f^2] - \mathbb{E}[f]^2}{2} = \frac{1}{2}\mathbb{V}[f].$$

Altogether, this shows that

$$\Pr_{(x,y)\sim R_{\pi/2\ell}}[f(x) \neq f(y)] \geq \frac{\mathbb{V}[f]}{2\ell}. \tag{2}$$

This inequality is tight for the sign function $f = \mathrm{sgn}(x_1)$, as we now show. The sign function satisfies $\mathbb{V}[f] = 1$. How do we compute its *rotation sensitivity*, that is the probability that $\mathrm{sgn}(x_1) \neq \mathrm{sgn}(y_1)$ when $(x_1, y_1) \sim R_{\pi/2\ell}$?

We can sample x_1, y_1 by sampling two independent Gaussians x_1, z_1 and letting $y_1 = \cos(\pi/2\ell)x_1 + \sin(\pi/2\ell)z_1$. Thus $\mathrm{sgn}(x_1)$ is the sign of the inner product of (x_1, z_1) with $(1, 0)$, and $\mathrm{sgn}(y_1)$ is the sign of the inner product of (x_1, z_1) with $(\cos(\pi/2\ell), \sin(\pi/2\ell))$.

Write $(x_1, z_1) = (r\cos\theta, r\sin\theta)$, which is what we get if we think of this pair as a point on the complex plane. Crucially, θ is uniformly distributed. The inner product of (x_1, z_1) with a vector $(\cos(\alpha), \sin(\alpha))$ equals $r\cos(\theta - \alpha)$, and in particular, the sign of the inner product only depends on α. In particular:

- x_1 is positive if θ lies in a sector of width π around 0.

[2] This is not a coincidence, and can be explained by viewing the noise operator in terms of rotations on the plane, similar to Figure 1.

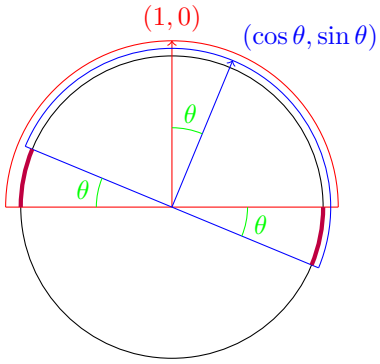

Figure 1: Red region is where x_1 is positive, blue region is where y_1 is positive, purple region is where they have different signs

- y_1 is positive if θ lies in a sector of width π around $\pi/2\ell$.

This shows that x_1 and y_1 have different signs with probability $2(\pi/2\ell)/(2\pi) = 1/2\ell$, see Figure 1. We have shown that the sign function satisfies

$$\Pr_{(x,y)\sim R_{\pi/2\ell}}[\mathrm{sgn}(x_1) \neq \mathrm{sgn}(y_1)] = \frac{1}{2\ell}.$$

Other functions satisfying this include the sign of any other linear form, for example $f(x) = \mathrm{sgn}(x_1 + \cdots + x_n)$; this is because the Gaussian distribution is rotation-invariant, and sign functions are scale-invariant. Such functions f are known as *(balanced) halfspaces*. More generally, we can consider functions of the form $\mathrm{sgn}(a_1x_1 + \cdots + a_nx_n + \theta)$, which are not balanced. It turns out that they are also tight for Equation (2).

Borell [Bor75], using completely different techniques (symmetrization), showed that Equation (2) holds for all angles θ, and that halfspaces are the unique optimizers (up to measure zero). Stability versions of Borell's theorem are also known [MN15, Eld15]. Such results show that if Equation (2) is almost tight then f is close to a halfspace.

Majority is Stablest The analog of halfspaces on the Boolean cube is threshold functions, that is, functions of the form $\mathrm{sgn}(x_1 + \cdots + x_n + \theta)$. Let us focus on the case of majority, $\mathrm{MAJ}(x) = \mathrm{sgn}(x_1 + \cdots + x_n)$. If x is a random point on the Boolean cube, then $X := (x_1 + \cdots + x_n)/\sqrt{n}$ has distribution quite close to Gaussian. If $(x, y) \sim N_\rho$ and $Y := (y_1 + \cdots + y_n)/\sqrt{n}$ then $\mathbb{E}[XY] = \rho$, and so the bivariate central limit theorem shows that (X, Y) roughly has the *Gaussian* distribution N_ρ. Therefore MAJ behaves much like the sign function considered above, and in particular,

$$\Pr_{(x,y)\sim N_\rho}[\mathrm{MAJ}(x) \neq \mathrm{MAJ}(y)] \approx \frac{\cos^{-1}\rho}{2\ell}.$$

In contrast, if f is a dictator then

$$\Pr_{(x,y)\sim N_\rho}[f(x) \neq f(y)] = \Pr_{(x,y)\sim N_\rho}[x_1 \neq y_1] = \frac{1-\rho}{2}.$$

The difference is illustrated in Figure 2.

The invariance principle implies (after several technical manipulations that we skip) that Borell's theorem approximately holds for functions on the Boolean cube $\{\pm 1\}^n$ as long as all influences of f are small. More accurately, for every $\delta > 0$ there exists $\tau > 0$ such that if $\max_i \mathrm{Inf}_i[f] \leq \tau$ then

$$\Pr_{(x,y)\sim N_\rho}[f(x) \neq f(y)] \geq \frac{\cos^{-1}\rho}{\pi} - \delta.$$

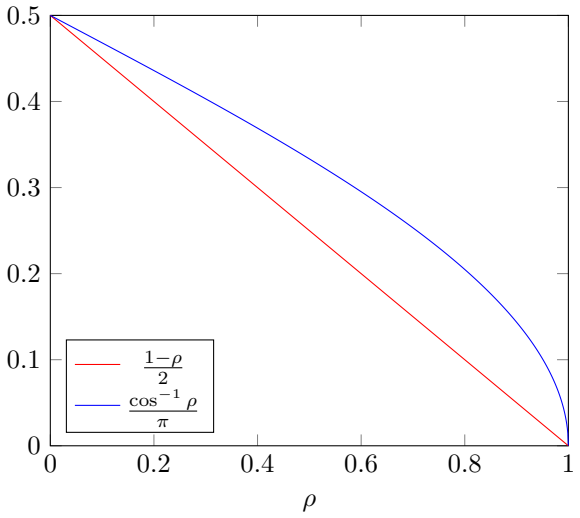

Figure 2: $\Pr[f(x) \neq f(y)]$ when $(x, y) \sim N_\rho$ for dictator (red) and majority (blue)

Stated differently, if a balanced Boolean function f is significantly more stable than $\frac{\cos^{-1}\rho}{\pi}$, then the only explanation is that it has an influential variable, which is preventing it from behaving as if it lived in Gaussian space.

What does all this have to do with theoretical computer science? It turns out that this result, known as Majority is Stablest, implies that the Goemans–Williamson algorithm [GW95] gives the optimal worst-case approximation ratio for MAX-CUT, as shown in [KKMO07]. This was the original impetus for proving the invariance principle.

While the original proof of Majority is Stablest used the invariance principle, since then direct inductive proofs were found [DMN16].

11.2 Application: Bourgain's tail bound

The Kindler–Safra theorem, proved in Section 7, states that a Boolean function which is concentrated up to constant degree is close to a junta. If we take the degree into account, the Kindler–Safra theorem states that if $f\colon \{\pm 1\}^n \to \{\pm 1\}$ satisfies $\|f^{>k}\|^2 \leq \epsilon$, then f is $2^{O(k)}\epsilon$-close to a Boolean junta depending on $2^{O(k)}$ variables. This result is meaningful only as long as $k \ll \log n$. What happens for larger k?

Let us see what happens for the majority function. Since all influences are very small, we instead consider the sign function, which is the Gaussian analog of majority. We have seen that

$$\Pr_{(x,y)\sim N_\rho}[\operatorname{sgn}(x) \neq \operatorname{sgn}(y)] = \frac{\cos^{-1}\rho}{\pi}.$$

Since sgn is Boolean,

$$\langle \operatorname{sgn}, T_\rho \operatorname{sgn}\rangle = \mathop{\mathbb{E}}_{(x,y)\sim N_\rho}[\operatorname{sgn}(x)\operatorname{sgn}(y)] = 1 - 2\Pr_{(x,y)\sim N_\rho}[\operatorname{sgn}(x) \neq \operatorname{sgn}(y)] = 1 - \frac{2\cos^{-1}\rho}{\pi}.$$

If sgn was a function on the Boolean cube, then the left-hand side would be

$$\sum_{d=0}^{\infty} \rho^d \| \operatorname{sgn}^{=d} \|^2.$$

(We can make this meaningful on Gaussian space using the Hermite expansion.) The right-hand side is

$$\frac{2}{\pi}\sum_{d=0}^{\infty}\frac{\binom{2d}{d}}{4^d}\frac{\rho^{2d+1}}{2d+1}.$$

This is just the Taylor series of arccosine. Using the well-known asymptotics of central binomial coefficients, we see that for odd d,

$$\|\operatorname{sgn}^{=d}\|^2 = \Theta\left(\frac{1}{d^{3/2}}\right).$$

Summing this over all $d \geq k$,

$$\|\operatorname{sgn}^{>k}\|^2 = \Theta\left(\frac{1}{\sqrt{k}}\right).$$

Assuming that majority behaves like the sign function, this shows that there exist functions which are far from being a junta and have mass $\Theta(\frac{1}{\sqrt{k}})$ beyond level k.

Bourgain's theorem [Bou02] states that this bound is optimal: any Boolean function which has less mass beyond level k is essentially a junta. Bourgain's original proof was complicated. Here we will follow the technique of Kindler–Kirshner–O'Donnell [KKO18], who also introduced rotation sensitivity, analyzed in the preceding section.

Gaussian tail bound The starting point is showing that *every* Boolean function f in Gaussian space has mass $\Omega(\frac{\mathbb{V}[f]}{\sqrt{k}})$ beyond the k'th level; the connection to functions on the Boolean cube is via the invariance principle. This will follow as an application of the inequality

$$\Pr_{(x,y)\sim R_{2\theta}}[f(x) \neq f(y)] \leq 2 \Pr_{(x,y)\sim R_\theta}[f(x) \neq f(y)],$$

applied to a suitable angle θ.

Let us first express both sides in terms of the Hermite expansion. Above we have shown that

$$\langle f, T_\rho f\rangle = 1 - 2\Pr_{(x,y)\sim N_\rho}[f(x) \neq f(y)],$$

and so, taking $\rho = \cos\theta$,

$$\Pr_{(x,y)\sim R_\theta}[f(x) \neq f(y)] = \frac{1-\langle f, T_\rho f\rangle}{2} = \frac{\langle f,f\rangle - \langle f, T_\rho f\rangle}{2} = \sum_{d=0}^{\infty}\frac{1-\rho^d}{2}\|f^{=d}\|^2.$$

(The Hermite expansion continues to infinity.) Therefore we can express the inequality above as follows:

$$\sum_{d=0}^{\infty}\frac{1-\cos^d(2\theta)}{2}\|f^{=d}\|^2 \leq \sum_{d=0}^{\infty}(1-\cos^d\theta)\|f^{=d}\|^2.$$

How do the two sides compare? Taylor expansion shows that

$$1-\cos^d\theta = \frac{d\theta^2}{2} - O(d^2\theta^4).$$

Similarly,

$$\frac{1-\cos^d(2\theta)}{2} = \frac{d(2\theta)^2}{4} - O(d^2\theta^4) = d\theta^2 - O(d^2\theta^4).$$

Thus when $d\theta^2 \ll 1$, the coefficient on the left is in fact *larger* than the coefficient on the right! The only explanation is that the function has significant component in higher levels.

In order to quantify this, let us pick a threshold $D = \Theta(1/\theta^2)$. When $d \leq D$, the "eigenvalue" on the left is roughly double that on the right, and so

$$\sum_{d=D+1}^{\infty} (1 - \cos^d \theta)\|f^{=d}\|^2 \geq \Omega(1) \cdot \sum_{d=0}^{\infty} \frac{1 - \cos^d(2\theta)}{2}\|f^{=d}\|^2.$$

The expression on the left is bounded by $\|f^{>D}\|^2$. The expression on the right is just

$$\Omega(1) \cdot \Pr_{(x,y)\sim R_\theta}[f(x) \neq f(y)] \overset{(2)}{\geq} \frac{\mathbb{V}[f]}{\pi/\theta},$$

assuming that π/θ is an even integer. Given k, choose θ to be an even integer such that $k = \Theta(1/\theta^2)$. Combining everything, we have shown that

$$\|f^{>k}\|^2 = \Omega\left(\frac{\mathbb{V}[f]}{\sqrt{k}}\right).$$

Junta conclusion Applying the invariance principle, we obtain a similar bound for functions on the Boolean cube when all influences are small. But we can show more. Let $f\colon \{\pm 1\}^n \to \{\pm 1\}$ be an arbitrary Boolean function such that $\|f^{>k}\|^2 \leq \frac{\epsilon}{\sqrt{k}}$. We will show that f is approximately a junta.

The junta variables are easy to identify, namely the variables having high influence. However, there is a subtle issue: f could have large total influence, and so there could be many variables with high influence. Therefore we actually choose as the junta variables those variables of high influence in $f^{\leq k}$, whose number we can bound. We ignore this subtlety (which is also required in order to apply the invariance principle in the first place!) in the sequel; roughly speaking, this is not an issue since f is close to $f^{\leq k}$, by assumption.

Denote by J the variables with high influence. We would like to move f to Gaussian space using the invariance principle. However, the invariance principle only applies to functions in which all variables have low influence. Therefore we only replace the variables *outside* J with Gaussians, obtaining a function on $\{\pm 1\}^J \times \mathbb{R}^{\overline{J}}$. Consequently, we can only apply the tail bound on the Gaussian part.

For each assignment α to the variables in J, the function f_α obtained by substituting α to J lives in Gaussian space, and so satisfies

$$\|f_\alpha^{>k}\|^2 \geq \Omega\left(\frac{1}{\sqrt{k}}\right) \cdot \mathbb{V}[f_\alpha]. \tag{3}$$

How do these quantities relate to the Fourier spectrum of f? The Fourier expansion of f_α is

$$\sum_{S\subseteq\overline{J}} \left(\sum_{T\subseteq J} \alpha_T \hat{f}(S \cup T)\right) x_S.$$

Therefore

$$\|f_\alpha^{>k}\|^2 = \sum_{\substack{S\subseteq\overline{J} \\ |S|>k}} \left(\sum_{T\subseteq J} \alpha_T \hat{f}(S \cup T)\right)^2 = \sum_{\substack{S\subseteq\overline{J} \\ |S|>k}} \sum_{T_1,T_2\subseteq J} \alpha_{T_1}\alpha_{T_2}\hat{f}(S \cup T_1)\hat{f}(S \cup T_2).$$

Taking expectation over α gives

$$\mathop{\mathbb{E}}_{\alpha}[\|f_\alpha^{>k}\|^2] = \sum_{\substack{S\subseteq\overline{J} \\ |S|>k}} \sum_{T\subseteq J} \hat{f}(S \cup T)^2 = \sum_{|R\setminus J|>k} \hat{f}(R)^2,$$

due to orthogonality of characters.

Since variance is just the case $k = 0$, taking expectation over the tail bound (3) gives

$$\sum_{|R \setminus J| > k} \hat{f}(R)^2 \geq \Omega\left(\frac{1}{\sqrt{k}}\right) \sum_{R \not\subseteq J} \hat{f}(R)^2.$$

The left-hand side is at most $\|f^{>k}\|^2 \leq \frac{\epsilon}{\sqrt{k}}$. The right-hand side, as we have seen in Section 5, is the distance between f and the function g obtained by averaging over the coordinates outside J. Thus

$$\frac{\epsilon}{\sqrt{k}} \geq \|f^{>k}\|^2 \geq \Omega\left(\frac{1}{\sqrt{k}}\right) \|f - g\|^2,$$

and so $\|f - g\|^2 = O(\epsilon)$. As we have seen in Section 5, if we round g to a Boolean function G, then $\Pr[f \neq G] = O(\epsilon)$ as well.

How large is the junta J? This depends on the threshold required for the invariance principle to go through, an issue we have been vague about. It turns out that $|J| \leq 2^{O(k)}\mathsf{poly}(1/\epsilon)$.

12 Global hypercontractivity

Hypercontractivity is the crucial ingredient in many proofs in Boolean function analysis. Hypercontractivity holds in the p-biased cube, as we have shown in Section 8.2. However, the parameters deteriorate as p gets closer to 0. This is inevitable, as the following simple calculation shows.

Suppose that $\|T_\rho f\|_4 \leq \|f\|_2$ for some value of ρ, with respect to μ_p. As we have seen in Section 6, this implies that

$$\|f\|_4 = \|T_\rho T_{\rho^{-1}} f\|_4 \leq \|T_{\rho^{-1}} f\|_2 \leq \rho^{-\deg f} \|f\|_2.$$

Let us apply this to the function $f(y_1, \ldots, y_n) = y_1$, where $y_1, \ldots, y_n \sim \mu_p$. For any $q \geq 1$, $\|f\|_q = \mathbb{E}[|f|^q]^{1/q} = p^{1/q}$. Therefore the inequality above reads

$$\sqrt[4]{p} \leq \rho^{-1}\sqrt{p} \Longrightarrow \rho \leq \sqrt[4]{p}.$$

Conversely, in Section 8.2 we have shown that hypercontractivity does hold for $\rho = \sqrt[4]{p}$, with the usual inductive proof that we first explained in Section 6. For p close to zero, this result is next to useless, since $T_{\sqrt[4]{p}}$ reduces a function to essentially its average.

This examples shows that we cannot hope for a hypercontractivity estimate with ρ independent of p. However, it might be that such an estimate holds for *some* functions. This should remind us of the situation in Section 11, in which we showed that low degree functions with low influences behave in similar ways on the unbiased Boolean cube and on Gaussian space. Could it be that functions with low influences do obey hypercontractivity with ρ independent of p? Such a result was proved in [KLLM19]. Here we give an exposition based on unpublished notes of Noam Lifshitz.

The basic idea is to compare a function f on the p-biased cube with its analog F on the unbiased cube. By that we mean that f and F have the same Fourier coefficients:

$$f = \sum_{S \subseteq [n]} \hat{f}(S)\omega_S, \quad F = \sum_{S \subseteq [n]} \hat{f}(S)x_S,$$

where $\omega_S = \prod_{i \in S} \frac{y_i - p}{\sqrt{p(1-p)}}$, a p-biased Fourier character, depends on $(y_1, \ldots, y_n) \sim \mu_p$, and $x_S = \prod_{i \in S} x_i$, an unbiased Fourier character, depends on $(x_1, \ldots, x_n) \sim \mu_{1/2}$.

Just as hypercontractivity states that by applying noise we can convert an L_4 norm into an L_2 norm, we will show that by applying noise, we can convert the L_4 norm of f to the L_4 norm of F:

$$\mathbb{E}[(T_\rho f)^4] \leq \mathbb{E}[F^4] + \text{penalty terms},$$

where the penalty terms are the analogs of the error terms in the proof of the invariance principle. Applying even more noise, we would be able to convert the L_4 norm on the right into an L_2 norm $\mathbb{E}[F^2]^2$, and so obtain an expression involving only the original function, since f and F have the same L_2 norm.

The proof will proceed by induction, as in Section 6. Therefore we start with the base case. Suppose that $f = a\omega + b$, where $\omega = \frac{y-p}{\sqrt{p(1-p)}}$ for $y \sim \mu_p$, and let $F = ax + b$ for $x \sim \mu_{1/2}$. We would like to compare $\mathbb{E}[(T_\rho f)^4]$ and $\mathbb{E}[F^4]$, which equal

$$\mathbb{E}[(T_\rho f)^4] = \mathbb{E}[\omega^4]\rho^4 a^4 + 4\,\mathbb{E}[\omega^3]\rho^3 a^3 b + 6\rho^2 a^2 b^2 + b^4,$$
$$\mathbb{E}[F^4] = a^4 + 6a^2 b^2 + b^4.$$

We calculated $\mathbb{E}[\omega^3]$ and $\mathbb{E}[\omega^4]$ in Section 8.2. Looking at the expressions, we see that if p is bounded away from $1/2$ then

$$\kappa_3 := \mathbb{E}[\omega^3] = \Theta\left(\frac{1}{\sqrt{p}}\right), \quad \kappa_4 := \mathbb{E}[\omega^4] = \Theta\left(\frac{1}{p}\right).$$

Comparing the expressions for $\mathbb{E}[T_\rho f^4]$ and $\mathbb{E}[F^4]$, there are two main issues. First, κ_4 is not constant. We will take care of this using a penalty term. Second, we have an additional term $\kappa_3 a^3 b$. Using the AM-GM inequality, we can convert it to terms which we are able to handle:

$$\kappa_3 a^3 b = \sqrt{\kappa_3^2 a^6 b^2} = \sqrt{\kappa_3^2 a^4 \cdot a^2 b^2} \leq \frac{\kappa_3^2 a^4 + a^2 b^2}{2}.$$

Therefore

$$\mathbb{E}[(T_\rho f)^4] \leq (\kappa_4 \rho^4 + 2\kappa_3^2 \rho^3) a^4 + (2\rho^3 + 6\rho^2) a^2 b^2 + b^4.$$

For small enough constant ρ, this is at most

$$\mathbb{E}[F^4] + O\left(\frac{1}{p}\right) a^4.$$

We can extract a by taking a derivative of F. Renaming x to x_1, We have $L_1 F = a x_1$ (see Section 4), and so

$$\mathbb{E}[(T_\rho f)^4] \leq \mathbb{E}[F^4] + \alpha\,\mathbb{E}[(L_1 F)^4], \text{ where } \alpha = O\left(\frac{1}{p}\right).$$

This concludes the one-dimensional case.

It is not too difficult to guess what we get in the general case:

$$\mathbb{E}[(T_\rho f)^4] \leq \sum_{S \subseteq [n]} \alpha^{|S|}\,\mathbb{E}[(L_S F)^4],$$

where $L_S F$ is obtained by applying the operators L_i for all $i \in S$, in an arbitrary order (they commute).

We prove this by induction. We have already seen the base case $n = 1$. Now suppose $f = \omega_{n+1} g + h$, and let $F = x_{n+1} G + H$, where G, H are obtained from g, h by replacing ω_T by x_T.

For every value of $y_1, \ldots, y_n$, the function $T_\rho f$ becomes the one-dimensional function $T_\rho[\omega_{n+1} T_\rho g(y_1, \ldots, y_n) + T_\rho h(y_1, \ldots, y_n)]$ (where the first T_ρ is with respect to the last coordinate and the other T_ρ's are with respect to the first n coordinates), and so the base case shows that

$$\mathop{\mathbb{E}}_{y_{n+1}}[(T_\rho f)^4(y_1, \ldots, y_n, y_{n+1})] \leq \mathop{\mathbb{E}}_{x_{n+1}}[(x_{n+1} T_\rho g(y_1, \ldots, y_n) + T_\rho h(y_1, \ldots, y_n))^4] + \alpha (T_\rho g(y_1, \ldots, y_n))^4.$$

Taking expectation over $y_1, \ldots, y_n$, the second term is at most

$$\alpha \sum_{S \subseteq [n]} \alpha^{|S|}\,\mathbb{E}[(L_S G)^4] = \sum_{\substack{S \subseteq [n+1] \\ n+1 \in S}} \alpha^{|S+1|}\,\mathbb{E}[(L_S F)^4].$$

As for the first term, since $x_{n+1}G + H = F$, we can similarly bound it by

$$\mathop{\mathbb{E}}_{x_{n+1}} \sum_{S \subseteq [n]} \alpha^{|S|} \mathop{\mathbb{E}}_{x_1,\ldots,x_n} [(L_S F)^4] = \sum_{S \subseteq [n]} \alpha^{|S|} \mathbb{E}[(L_S F)^4].$$

Altogether, this shows that

$$\mathbb{E}[(T_\rho f)^4] \leq \sum_{S \subseteq [n+1]} \alpha^{|S|} \mathbb{E}[(L_S F)^4],$$

as needed. This completes the inductive proof.

Finally, in order to obtain an expression in terms of f on the right-hand side, we apply more noise. Let $\sigma = \rho/\sqrt{3}$. Then

$$\mathbb{E}[(T_\sigma f)^4] \leq \sum_{S \subseteq [n]} \alpha^{|S|} \mathbb{E}[(L_S T_{1/\sqrt{3}} F)^4] \leq \sum_{S \subseteq [n]} \alpha^{|S|} \mathbb{E}[(L_S F)^2]^2 = \sum_{S \subseteq [n]} \alpha^{|S|} \mathbb{E}[(L_S f)^2]^2,$$

since $L_S F$ and $L_S f$ have the same Fourier coefficients. We record this form of hypercontractivity:

> For any $p \leq 1/2$, the following holds with respect to μ_p and some constant $\sigma > 0$ independent of p:
>
> $$\|T_\sigma f\|_4^4 \leq \sum_{S \subseteq [n]} \alpha^{|S|} \|L_S f\|_2^4, \quad \alpha = O\left(\frac{1}{p}\right).$$

When is the right-hand side small? For starters, let us notice that

$$f(x_{-n},1) - f(x_{-n},0) = \sum_{\substack{S \subseteq [n] \\ n \in S}} \hat{f}(S) \omega_{S \setminus \{n\}} \frac{(1-p) - (-p)}{\sqrt{p(1-p)}} = \sum_{\substack{S \subseteq [n] \\ n \in S}} \hat{f}(S) \omega_{S \setminus \{n\}} \frac{1}{\sqrt{p(1-p)}}.$$

This shows that

$$\mathop{\mathbb{E}}_{x_{-n}} [(f(x_{-n},1) - f(x_{-n},0))^2] = \frac{1}{p(1-p)} \|L_n f\|_2^2.$$

(We have actually already seen this calculation in Section 9.) The expression $f(x_{-n},1) - f(x_{-n},0)$ can be cast as an operator $D_n f = (L_n f)/(\sqrt{p(1-p)}\omega_n)$. Defining $D_{\{i_1,\ldots,i_s\}} f = D_{i_1} \ldots D_{i_s} f$, this shows that

$$\|T_\sigma f\|_4^4 \leq \sum_{S \subseteq [n]} (p\alpha)^{|S|} \|D_S f\|_2^2 \|L_S f\|_2^2,$$

where $p\alpha = O(1)$. Now suppose that $\|D_S f\|_2 \leq \beta$ for all S. Then

$$\|T_\sigma f\|_4^4 \leq \sum_{S \subseteq [n]} (p\alpha)^{|S|} \beta^2 \|L_S f\|_2^2 = \beta^2 \sum_{S \subseteq [n]} \sum_{T \supseteq S} (p\alpha)^{|S|} \hat{f}(T)^2 = \beta^2 \sum_{T \subseteq [n]} \hat{f}(T)^2 \sum_{S \subseteq T} (p\alpha)^{|S|} = \beta^2 \sum_{T \subseteq [n]} (1+p\alpha)^{|T|} \hat{f}(T)^2.$$

Suppose now that we replace σ by $\tau = \sigma/(1+p\alpha)$. Then

$$\|T_\tau f\|_4^4 \leq \beta^2 \sum_{T \subseteq [n]} \hat{f}(T)^2 = \beta^2 \|f\|_2^2.$$

This is another form of hypercontractivity worth recording:

> For any $p \leq 1/2$, the following holds with respect to μ_p and some constant $\tau > 0$ independent of p:
>
> $$\|T_\tau f\|_4 \leq \sqrt{\beta \|f\|_2}, \text{where } \beta = \max_{S \subseteq [n]} \|D_S f\|_2.$$

12.1 Application: Bourgain's booster theorem

One of the classical topics in random graph theory is the threshold behavior of monotone graph properties. Here are two examples, connectivity and containing a triangle:

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

In both cases, the limit is as $n \to \infty$.

There is a crucial difference between these two properties. In the case of connectivity, the threshold is around $\frac{\log n}{n}$, and the probability jumps from 0 to 1 in an interval ("window") of width $\frac{1}{n} = o(\frac{\log n}{n})$. In contrast, in the case of containing a triangle, both the threshold and the window width are around $\frac{1}{n}$.

Let us recall a notation from Section 9: for a monotone property $f\colon \{0,1\}^n \to \{0,1\}$ and $q \in [0,1]$, we define $\tau(q)$ as the value of p such that $\mu_p(f) = q$, where $\mu_p(f) = \mathbb{E}_{\mu_p}[f]$ is the probability that $G(n,p)$ satisfies the property.

A simple argument (taking the union or intersection of several $G(n, \tau(1/2))$ graphs) shows that the window width is always at most the threshold itself. When the window width has the same order of magnitude as the threshold, we say that the threshold is *coarse*, and otherwise we say that it is *sharp*. Sharp thresholds are easier to locate (since we only need to show that we are inside the window in order to locate the threshold with high accuracy), so we would like a criterion that guarantees that a property manifests a sharp threshold.

The Russo–Margulis formula, proved in Section 9, shows that if f is a monotone property then $\phi(p) = \mu_p(f)$ satisfies $\phi'(p) = \mathrm{Inf}^{(p)}[f]/p(1-p)$, where $\mathrm{Inf}^{(p)}[f]$ is the total influence of f with respect to μ_p. If the critical probability is $p_c = \tau(1/2)$ and the window width is $w = \tau(3/4) - \tau(1/4)$, then some point $p \in [\tau(1/4), \tau(3/4)]$ has derivative at most $1/(2w)$. If the threshold is coarse then $p, w = \Theta(p_c)$ and so $\mathrm{Inf}^{(p)}[f] = O(p/w) = O(1)$.

What can we say about such functions? If p is constant, then Friedgut's junta theorem states that f is close to a junta, as we showed in Section 5. However, this is no longer the case for small p. For example, let us consider the case of containing a triangle. Since $\phi(p) \approx 1 - e^{-(pn)^3/6}$, we have $\phi'(p) \approx \frac{1}{2} n^3 p^2 e^{-(pn)^3/6}$, and so $\mathrm{Inf}^{(p)}[f] \approx p\phi'(p) \approx \frac{1}{2}(pn)^3 e^{-(pn)^3/6}$. When $p = c/n$, this is constant, yet f is far from being a junta.

Sharp threshold theorems describe the structure of functions f satisfying $\mathrm{Inf}^{(p)}[f] = O(1)$, under the additional assumption that $0 \ll \mu_p(f) \ll 1$; we do not know the answer when $\mu_p(f)$ is small. The most useful such theorem is due to Bourgain (appearing in an appendix to [Fri98]), which we will prove here using global hypercontractivity, following [KLLM19].

Let f be a monotone function such that

$$\mathrm{Inf}^{(p)}[f] \le K\,\mathbb{V}[f].$$

When $0 \ll \mu_p(f) \ll 1$, this is the same as requiring that $\mathrm{Inf}^{(p)}[f] = O(1)$.

An argument in the style of the proofs in Section 5 shows that

$$\mathbb{V}[f] = \sum_{1 \le |S| \le 2K} \hat{f}(S)^2 + \sum_{|S| > 2K} \hat{f}(S)^2.$$

The first term is $\mu_p(f)^2$. The second term is at most $\mathrm{Inf}[f^{\le 2K}]$. The third term is at most

$$\frac{1}{2K} \sum_S |S| \hat{f}(S)^2 = \frac{1}{2K} \mathrm{Inf}[f] \le \frac{1}{2}\,\mathbb{V}[f].$$

Altogether,

$$\mathbb{V}[f] \le \mathrm{Inf}[f^{\le 2K}] + \frac{1}{2}\,\mathbb{V}[f],$$

which implies that

$$\frac{1}{2}\,\mathbb{V}[f] \le \mathrm{Inf}[f^{\le 2K}]. \tag{4}$$

We now bound $\mathrm{Inf}[f^{\leq 2K}]$ using global hypercontractivity. For each individual i,

$$\mathrm{Inf}_i[f^{\leq 2K}] = \|L_i(f^{\leq 2K})\|_2^2 = \|(L_i f)^{\leq 2K}\|_2^2 = p(1-p)\|(D_i f)^{<2K}\|_2^2,$$

since $D_i f = L_i f/(\omega_n \sqrt{p(1-p)})$. Crucially, $D_i f \in \{0, \pm 1\}$. Applying Hölder's inequality $\langle g, h\rangle \leq \|g\|_4 \|h\|_{4/3}$, this shows that

$$\mathrm{Inf}_i[f^{\leq 2K}] = p(1-p)\langle (D_i f)^{<2K}, D_i f\rangle \leq p(1-p)\|(D_i f)^{<2K}\|_4 \|D_i f\|_{4/3} = p(1-p)\|(D_i f)^{<2K}\|_4 \|D_i f\|_2^{3/2}.$$

Global hypercontractivity shows that

$$\|(D_i f)^{<2K}\|_4 = \|T_\tau T_{\tau^{-1}} (D_i f)^{<2K}\|_4 \leq \sqrt{\beta_i}\|T_{\tau^{-1}}(D_i f)^{<2K}\|_2^{1/2} \leq \sqrt{\beta_i} e^{O(K)} \|D_i f\|_2^{1/2},$$

where

$$\beta_i = \max_{S \subseteq [n]} \|D_S T_{\tau^{-1}} (D_i f)^{<2K}\|_2.$$

Altogether,

$$\mathrm{Inf}_i[f^{\leq 2K}] \leq p(1-p)\sqrt{\beta_i}\|D_i f\|_2^2 = \sqrt{\beta_i}\,\mathrm{Inf}_i[f].$$

Substituting this in (4),

$$\frac{1}{2}\mathbb{V}[f] \leq \sqrt{\max_i \beta_i} e^{O(K)}\,\mathrm{Inf}[f] \Longrightarrow \max_i \beta_i \geq e^{-O(K)}.$$

Now let us explore β_i further. First, we can restrict S in the definition of β_i to sets of size less than $2K$. Second, $T_{\tau^{-1}}$ increases the Fourier coefficients of $(D_i f)^{<2K}$ by a factor of at most $e^{O(K)}$, and so we can remove it at the cost of increasing the entire expression by that factor. Altogether, we can bound

$$\max_i \beta_i \leq e^{O(k)} \max_{|S| \leq 2K} \|D_S f\|_2^2.$$

We conclude that

$$\max_{|S| \leq 2K} \|D_S f\|_2 \geq e^{-O(K)}.$$

Now take a set S such that $\|D_S f\|_2 \geq e^{-O(K)}$. We can give an explicit formula for $D_S f$:

$$D_S f(y) = \sum_{T \subseteq S} (-1)^{|S \setminus T|} f(y_{-S}, T),$$

where y_{-S} consists of the coordinates of y outside S, and we think of T as as assignment to the coordinates of y in S; this generalizes the explicit formula for $D_n F$ appearing above. Since f is Boolean, this expression is bounded by $2^{|S|}$, and so $\|D_S f\|_2^2 \leq 4^{|S|}\Pr[D_S f \neq 0]$, implying that $\Pr[D_S f \neq 0] \geq e^{-O(K)}$.

Consider a particular assignment y_{-S} to the coordinates outside of S. If $D_S f \neq 0$ then since f is monotone, necessarily $f(y_{-S}, \emptyset) = 0$ and $f(y_{-S}, S) = 1$. Assuming that $p \leq 1/2$, this shows that if we fix the coordinates in S to 1, then this increases the expected value (for this partial assignment y_{-S}) by at least $(1-p)^{|S|} \geq e^{-O(K)}$. Therefore,

$$\mathbb{E}[f \mid y_i = 1 \text{ for all } i \in S] \geq \mathbb{E}[f] + e^{-O(K)}.$$

We have deduced Bourgain's booster theorem:

> Let f be a monotone function and $p \leq 1/2$. If $\mathrm{Inf}^{(p)}[f] \leq K\mu_p(f)(1-\mu_p(f))$ then there exists a set S of size at most $2K$ such that
>
> $$\mathbb{E}[f \mid y_i = 1 \text{ for all } i \in S] \geq \mathbb{E}[f] + e^{-O(K)}.$$

13 Analysis on the slice: Erdős–Ko–Rado

Up to now we have considered Fourier analysis on the Boolean cube $\{\pm 1\}^n$ with respect to the uniform measure, on the Boolean cube $\{0,1\}^n$ with respect to the biased measure μ_p, and briefly on Gaussian space. All of these settings have the useful feature that the coordinates are independent. What happens when this property is absent?

The simplest such situation is the *slice* $\binom{[n]}{k}$, which is the uniform distribution ν_k over all vectors in $\{0,1\}^n$ of Hamming weight k. Intuitively, the slice behaves very similarly to the Boolean cube with respect to the measure $\mu_{k/n}$, and we will show a formal version of this later. But first, let us explore the following basic question: what is the right notion of Fourier expansion for functions on the slice?

Every function on $\{\pm 1\}^n$ can be written in a unique way as a multilinear polynomial in $x_1, \ldots, x_n$. This is no longer the case for the slice, where we consider the inputs to be the coordinates $x_1, \ldots, x_n \in \{0,1\}$, which are promised to sum to k. Indeed,

$$\sum_{i=1}^n x_i - k = 0,$$

and more generally, if we multiply the left-hand side by any polynomial P and multilinearize the result, then we also get a polynomial that vanishes on the slice. Therefore, in order to obtain a canonical representation, we need to add more constraints. Such a canonical representation was found by Dunkl [Dun76], who asked that the representing polynomial P be "orthogonal to the defining system", that is,

$$\sum_{i=1}^n \frac{\partial P}{\partial x_i} = 0.$$

This constraint, known as *harmonicity*, is not enough. We need the polynomial to be multilinear, to take into account the identity $x_i^2 = x_i$. Similarly, since $x_S = 0$ for any $|S| > k$, we need $\deg P \leq k$; dually, we also need $\deg P \leq n - k$. Putting all these constraints together, we do obtain a canonical representation.

Existence Let us first show that if f is any function on the slice $\binom{[n]}{k}$, where for simplicity we assume that $k \leq n/2$, then it can be represented by a harmonic multilinear polynomial of degree at most k.

Our starting point is the following trivial representation of f as a homogeneous multilinear polynomial:

$$P_k = \sum_{|S|=k} c_S x_S.$$

The idea is to find a homogeneous multilinear polynomial Q_k of the same degree such that $\Delta Q_k = 0$ and $P_k - Q_k$ can be represented by a homogeneous multilinear polynomial of degree $k-1$, and then repeat the process $k-1$ more times.

How do we find such a polynomial Q_k? If the coefficients of Q_k are d_S, then

$$\Delta Q_k = \sum_{|T|=k-1} \sum_{i \in [n] \setminus T} d_{T \cup \{i\}} x_T.$$

Therefore $\Delta Q_k = 0$ if $\sum_{i \notin T} d_{T \cup \{i\}} = 0$ for all T of size $k-1$.

It is natural to think of Q_k as a *formal sum* of the "indeterminates" x_S. Similarly, ΔQ_k is a formal sum of the indeterminates x_T. The operator Δ is a linear operator mapping $V_k := \operatorname{span}(\{x_S : |S| = k\})$ to $V_{k-1} := \operatorname{span}(\{x_T : |T| = k-1\})$.

Linear algebra tells us that V_k can be written as the direct sum of $\ker \Delta$ and $\operatorname{im} \Delta^T$. This is very promising, since applying this to P_k, this means that we can find a representation $P_k = Q_k + \Delta^T P_{k-1}$ where $\Delta Q_k = 0$ and $P_{k-1} \in V_{k-1}$. We have thus found a candidate for Q_k.

What can we say about $\Delta^T P_{k-1}$? The operator Δ^T maps x_T to $\sum_{i \in [n] \setminus T} x_{T \cup \{i\}}$. On the slice,

$$\sum_{i \in [n] \setminus T} x_{T \cup \{i\}} = x_T \sum_{i \in [n] \setminus T} x_i = x_T,$$

since if $x_T = 1$ then exactly one of the x_i's will evaluate to 1. Thus *on the slice*, we have $P_k = Q_k + P_{k-1}$.

We now have to repeat the same process for P_{k-1}. We can view Δ as an operator from V_{k-1} to $V_{k-2} := \mathrm{span}(\{x_U : |U| = k-2\})$, and so find a representation $P_{k-1} = Q_{k-1} + \Delta^T P_{k-2}$, where $\Delta Q_{k-1} = 0$. This time we have

$$\sum_{i \in [n] \setminus U} x_{U \cup \{i\}} = x_U \sum_{i \in [n] \setminus U} x_i = 2x_U,$$

on the slice. Thus on the slice, $P_{k-1} = Q_{k-1} + 2P_{k-2}$. Continuing in this vein, we obtain

$$P_k = \sum_{d=0}^{k} (k-d)! Q_d,$$

where Q_d is a degree d polynomial satisfying $\Delta Q_d = 0$. This is the representation we were looking for.

This argument also shows that if f can be represented by *some* polynomial of degree d, then it can be represented as a harmonic multilinear polynomial of degree at most d.

Uniqueness Now, let us show that the representation is unique. Equivalently, we need to show that if f is identically zero on the slice, then the only harmonic multilinear polynomial of degree at most k representing it is the zero polynomial. Let us therefore consider a harmonic multilinear polynomial

$$P = \sum_{|S| \leq k} c_S x_S$$

which evaluates to zero on the entire slice.

Since P evaluates to zero on the slice, in particular $\mathbb{E}[P] = 0$. Therefore

$$0 = c_\emptyset + \sum_{d=1}^{k} \sum_{|S|=d} \mathbb{E}[x_S] c_S = c_\emptyset + \sum_{d=1}^{k} \mathbb{E}[x_1 \cdots x_d] \sum_{|S|=d} c_S.$$

Since P is harmonic, we know that for every T,

$$\sum_{i \in [n] \setminus T} c_{T \cup \{i\}} = 0.$$

Summing this over all T of size $d-1$, we see that $(d+1) \sum_{|S|=d} c_S = 0$, and so $c_\emptyset = 0$. We have actually shown something stronger: $c_\emptyset = 0$ iff $\mathbb{E}[P] = 0$.

Next, we also know that $\mathbb{E}[Px_1] = 0$. Therefore

$$0 = \mathbb{E}[x_1] c_{\{1\}} + \sum_{d=1}^{k} \sum_{\substack{|S|=d \\ 1 \notin S}} (\mathbb{E}[x_1 \cdots x_d] c_S + \mathbb{E}[x_1 \cdots x_{d+1}] c_{S \cup \{1\}}).$$

If we sum the constraint $\sum_{i \in [n] \setminus T} c_{T \cup \{i\}} = 0$ over all T containing 1 of size d, then we see that

$$d \sum_{\substack{|S|=d+1 \\ 1 \in S}} c_S = 0 \Longrightarrow \sum_{\substack{|S|=d \\ 1 \notin S}} c_{S \cup \{1\}} = 0.$$

Similarly, if we sum the constraint over all T not containing 1 of size $d-1$, then we see that

$$0 = \sum_{\substack{|T|=d-1 \\ 1 \notin T}} \sum_{i \notin T} c_{T \cup \{i\}} = \sum_{\substack{|T|=d-1 \\ 1 \notin T}} \sum_{i \notin T \cup \{1\}} c_{T \cup \{i\}} + c_{T \cup \{1\}} = d \sum_{\substack{|S|=d \\ 1 \notin S}} c_S + \sum_{\substack{|S|=d-1 \\ 1 \notin S}} c_{S \cup \{1\}}.$$

We already know that the second summand vanishes, hence the first one also does. In total, we deduce that $c_{\{1\}} = 0$. Again, we have actually shown something stronger: assuming that $\mathbb{E}[P] = 0$, then $c_{\{1\}} = 0$ iff $\mathbb{E}[Px_1] = 0$. In particular, if $\mathbb{E}[P] = \mathbb{E}[Px_1] = 0$ then $c_\emptyset = c_{\{1\}} = 0$; and the converse also holds.

In the same way, we prove that all other coefficients vanish (we leave this as an exercise to the reader). Importantly, the proof actually shows something quite a bit stronger: f is orthogonal to all functions of degree at most d iff its harmonic representation (which is how we shall call the unique representation described above) has no terms of degree at most d. This means that if we write

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

where $f^{=d}$ is the d'th homogeneous part of the harmonic representation of f, then the functions $f^{=d}$ are *orthogonal*.

Generating set How do harmonic representations look like? Here is an example of a harmonic function which is homogeneous of degree d:

$$(x_1 - x_2) \cdots (x_{2d-1} - x_{2d}).$$

To check that this function is indeed harmonic, all we need to do is observe is that the product of harmonic polynomials is harmonic, by the product rule of the derivative; the claim then follows by considering the easy case of $x_1 - x_2$.

It turns out that every homogeneous harmonic multilinear polynomial of degree d can be written as a linear combination of functions of this form. To see this, let us first determine the dimension D_d of the space of harmonic multilinear polynomials which are homogeneous of degree d. Without the harmonicity constraint, the dimension is $\binom{n}{d}$. There are $\binom{n}{d-1}$ different harmonicity constraints, so $D_d \geq \binom{[n]}{d} - \binom{[n]}{d-1}$. On the other hand, by unique representation we know that $\sum_{d=0}^{k} D_d = \binom{n}{k}$. Thus

$$\binom{n}{k} = \sum_{d=0}^{k} D_d \geq \sum_{d=0}^{k} \binom{n}{d} - \binom{n}{d-1} = \binom{n}{k}.$$

This shows that all inequalities are tight, and so $D_d = \binom{n}{d} - \binom{n}{d-1}$.

Let us now say that a sequence $1 \leq j_1 < \cdots < j_d \leq n$ is *admissible* if there exist indices $i_1, \ldots, i_d$ such that all of $i_1, \ldots, i_d, j_1, \ldots, j_d$ are different, and additionally $i_t < j_t$ for all $t \in [d]$. It is a classical combinatorial result that the number of admissible sequences is $\binom{n}{d} - \binom{n}{d-1}$ (this is the famous Bertrand ballot problem, when we think of indices j_t as being votes to one candidate, and indices not in $j_1, \ldots, j_d$ as being votes to the other candidate).

For each admissible sequence $j_1 < \cdots < j_d$, choose some witness $i_1, \ldots, i_d$, and consider the polynomial

$$\chi_{j_1,\ldots,j_d} = \prod_{t=1}^{d} (x_{i_t} - x_{j_t}),$$

which is harmonic multilinear. We claim that these polynomials are linearly independent (as vectors of coefficients), and so must form a basis for all harmonic multilinear polynomials which are homogeneous of degree d. To see this, we construct a partial order on monomials of degree d. Say that $(i_1, \ldots, i_d) \prec (j_1, \ldots, j_d)$ if $i_t < j_t$ for all $t \in [d]$, and extend this partial order arbitrarily to a linear order. The matrix of coefficients of the polynomials $\chi_{j_1,\ldots,j_d}$ is triangular with 1s on the diagonal with respect to this order, and so these polynomials are linearly independent.

Exercise Complete the proof of uniqueness, and see where the condition $k \leq n/2$ comes into play.

13.1 Influence

Two concepts that were important in Boolean function analysis on the Boolean cube were influence and noise. How do we extend them to the slice?

We defined the influence in direction i via the operation of flipping the i'th coordinate. However, this operation doesn't preserve the Hamming weight. Instead, the minimal change is swapping two coordinates. Accordingly, one can define an influence in direction $(i\ j)$, but it is less useful than the analog of *total influence*. In the case of the Boolean cube, this is the sum of all influences, and also has an alternative characterization in terms of the Hamming graph, which is the graph corresponding to the Boolean cube (two vertices are connected if they are at Hamming distance 1): if f is a Boolean function, then $\mathrm{Inf}[f]$ is the average, over x, of the number of neighbors y of x such that $f(x) \neq f(y)$. For arbitrary functions,

$$\mathrm{Inf}[f] = \frac{1}{4}\mathop{\mathbb{E}}_{x}\left[\sum_{y\sim x}(f(x)-f(y))^2\right].$$

We also had another formula for total influence: using $y \sim x$ to go over all neighbors y of x in the Hamming graph,

$$\langle f, Lf\rangle, \text{ where } Lf(x) = \sum_{y\sim x}\frac{f(x)-f(y)}{2}.$$

Indeed,

$$\mathrm{Inf}[f] = \frac{1}{4}\mathop{\mathbb{E}}_{x}\left[\sum_{y\sim x}(f(x)-f(y))^2\right] = \frac{n}{4}\mathop{\mathbb{E}}_{x,y\sim x}[f(x)^2 - 2f(x)f(y) + f(y)^2] =$$

$$\frac{n}{4}\mathop{\mathbb{E}}_{x,y\sim x}[2f(x)^2 - 2f(x)f(y)] = \frac{1}{2}\mathop{\mathbb{E}}_{x}\left[\sum_{y\sim x} f(x)(f(x)-f(y))\right] = \langle f, Lf\rangle,$$

since if we choose a random x and a random neighbor y, then y is also random.

The analog of the Hamming graph in the case of the slice is the *Johnson graph*, in which two points x, y are connected if they differ in two coordinates, which is the minimal number. Accordingly, for functions on the slice we will define

$$\mathrm{Inf}[f] \propto \mathop{\mathbb{E}}_{x}\left[\sum_{y\sim x}(f(x)-f(y))^2\right].$$

The constant of proportionality is somewhat arbitrary — we will determine it so that the Fourier formula for $\mathrm{Inf}[f]$ will resemble the one for the Boolean cube.

Just as in the case of the Boolean cube, it will be easier to use the formula $\mathrm{Inf}[f] \propto \langle f, Lf\rangle$, where

$$Lf(x) = \sum_{y\sim x}(f(x)-f(y)).$$

(It is more natural to forego the factor of 2 in this case, since the x_i are $\{0,1\}$-valued rather than $\{\pm 1\}$-valued.) This is easier since we can compute Lf directly on the functions $\chi_d = (x_1 - x_2)\cdots(x_{2d-1} - x_{2d})$, and deduce its value for arbitrary f.

Suppose that x is a point such that $\chi_d(x) = 0$. This means that there is a pair of equal coordinates $x_{2i-1} = x_{2i}$. Suppose for definiteness that $x_{2i-1} = x_{2i} = 0$. The only neighbors of x on which χ_d possibly does not vanish are in directions $(2i-1\ j)$ and $(2i\ j)$, where $j \neq 2i-1, 2i$. Suppose that y is a neighbor in direction $(2i-1\ j)$ such that $\chi_d(y) \neq 0$. Necessarily $x_j = 1$, and so $y_{2i-1} - y_{2i} = 1$. If instead we look at the neighbor z in direction $(2i\ j)$, then the only difference between y and z is in coordinates $2i-1, 2i$, where $z_{2i-1} - z_{2i} = -1$. Therefore $\chi_d(z) = -\chi_d(y)$. Matching neighbors in this way, we see that $L\chi_d(x) = 0$.

The more interesting case is when $\chi_d(x) \neq 0$. Without loss of generality, let us say that $x_1 = x_3 = \cdots = x_{2d-1} = 1$ and $x_2 = x_4 = \cdots = x_{2d} = 0$. For a swap to change the value of the function, one of the coordinates being swapped must belong to $[2d]$. This can happen in the following ways:

1. Coordinates $2i-1, 2i \in [2d]$ are swapped. This changes the value of the function to -1. There are d such swaps.

2. A coordinate $2i-1 \in [2d]$ is swapped with a coordinate $2j \in [2d]$, where $i \neq j$. This changes the value of the function to 0. There are $d(d-1)$ such swaps.

3. A coordinate $2i-1 \in [2d]$ is swapped with a 0-coordinate outside $[2d]$. This also zeroes the function. There are $d(n-k-d)$ such swaps.

4. A coordinate $2i \in [2d]$ is swapped with a 1-coordinate outside $[2d]$. This also zeroes the function. There are $d(k-d)$ such swaps.

In total,

$$L\chi_d(x) = 2d + d(d-1) + d(n-k-d) + d(k-d) = d(n-d+1).$$

When $\chi_d(x) = -11$, we similarly get the negative of this value. Therefore

$$L\chi_d = d(n-d+1)\chi_d.$$

Since functions of the form χ_d span the d'th level (consisting of harmonic multilinear polynomials which are homogeneous of degree d), this shows that

$$Lf = \sum_{d=0}^{k} d(n-d+1) f^{=d},$$

and so, since the different levels are orthogonal to each other,

$$\langle f, Lf \rangle = \sum_{d=0}^{k} d(n-d+1) \|f^{=d}\|^2.$$

In comparison, the coefficients in the case of the Boolean cube were d. We therefore define the total influence by dividing the formula above by a factor of n:

$$\mathrm{Inf}[f] = \frac{1}{n}\langle f, Lf \rangle = \sum_{d=0}^{k} \left(1 - \frac{d-1}{n}\right) d \|f^{=d}\|^2.$$

One useful property of Lf is that the eigenvalues $d(n-d+1)$ are all distinct (assuming $k \leq n/2$). To see this, it suffices to notice that

$$(d+1)(n-d) - d(n-d+1) = n - 2d,$$

which is strictly positive when $d+1 \leq n/2$.

The operator L is closely related to the adjacency operator A of the Johnson graph. The relation is $Lf = k(n-k)f - Af$, since every vertex has degree $k(n-k)$. The eigenspaces of A are thus also the $k+1$ levels of the Fourier expansion. The same holds for all powers of A, and so for all polynomials in A, and so for all operators B such that $B(x,y)$ depends only on the distance between x and y in the Johnson graph.

Such operators are encountered very frequently, since if B is any linear operator which "doesn't care about names of elements" — formally, is invariant under the operation of permuting the indices $[n]$ — then $B(x,y)$ only depends on the Hamming distance between x and y, which is twice their distance in the Johnson graph. Any such operator will have the same eigenspaces as A, and so to compute its eigenvalues, it suffices to compute them on the functions χ_d.

13.2 Noise

In the case of the Boolean cube, we defined the noise operator T_ρ in two equivalent ways. First, using an operation which flips every coordinate with probability $\frac{1-\rho}{2}$. Second, using the Fourier expansion: T_ρ multiplies the d'th level by ρ^d. It is clear how to extend the second definition to the case of the slice. What about the first definition?

Here is another way at looking at the Fourier definition. Recall that

$$Lf = \sum_{d=0}^{n} df^{=d}.$$

In other words, the eigenspaces of L are the Fourier levels, with eigenvalues d. Therefore

$$T_\rho f = \sum_{d=0}^{n} \rho^d f = e^{-L\ln(1/\rho)} f.$$

As in the case of the slice, we can write L in terms of the adjacency operator A of the Hamming graph. It will be slightly nicer to consider instead the corresponding random walk operator $M = A/n$, which corresponds to taking a random neighbor. Then $L = (nI - A)/2 = (n/2)(I - M)$, and so

$$T_\rho f = e^{-(n/2)\ln(1/\rho)(I-M)}.$$

Defining $t = (n/2)\ln(1/\rho)$ and using the Taylor series for e^x, this gives

$$T_\rho f = e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} M^k f = \mathop{\mathbb{E}}_{k\sim \mathrm{P}(t)}[M^k f],$$

where $\mathrm{P}(t)$ is a Poisson random variable with expectation t. That is, $T_\rho f(x) = \mathbb{E}[f(y)]$, where y is obtained by taking $\mathrm{P}(t)$ random steps. (Equivalently, we can view this in terms of continuous-time Markov chains.)

We can generalize all of this to the slice by considering the adjacency and random walk operators of the Johnson graph. The upshot is that the d in ρ^d will be replaced by $(1 - \frac{d-1}{n})d$, leading to the following definition:

$$T_\rho f = \sum_{d=0}^{k} \rho^{(1-\frac{d-1}{n})d} f^{=d}.$$

How does this compare with the noise operator T^*_ρ whose eigenvalues are exactly ρ^d? We have

$$\|T_\rho f - T^*_\rho f\|^2 = \sum_{d=0}^{k} (\rho^{(1-\frac{d-1}{n})d} - \rho^d)^2 \|f^{=d}\|^2 \leq \max_d (\rho^{(1-\frac{d-1}{n})d} - \rho^d)^2 \|f\|^2.$$

When $d \leq \sqrt{n}$, using $e^x = 1 + O(x)$ (for $x = O(1)$) we can upper bound

$$\rho^{(1-\frac{d-1}{n})d} - \rho^d = \rho^d \left((1/\rho)^{d(d-1)/n} - 1 \right) = O_\rho\left(\frac{d^2\rho^d}{n} \right),$$

and otherwise

$$\rho^{(1-\frac{d-1}{n})d} - \rho^d \leq \rho^{d/2} \leq \rho^{\sqrt{n}/2}.$$

This shows that

$$\|T_\rho f - T^*_\rho f\|^2 = O_\rho\left(\frac{1}{n} \right) \|f\|^2.$$

Hypercontractivity still holds for constant p. We show this below for low-degree functions by relating the slice to the corresponding p-biased cube.

13.3 Application: Erdős–Ko–Rado

In Section 8.1, we considered the p-biased version of the Erdős–Ko–Rado theorem. The original version took place in the slice, and this is one motivation to study it. (Another motivation is $G(n,m)$ random graphs, which were the model of random graphs originally studied by Erdős and Rényi [ER60].)

A family $\mathcal{F} \subseteq \binom{[n]}{k}$ is *intersecting* if any two sets $A, B \in \mathcal{F}$ intersect. If $k > n/2$ then any two sets intersect, and so the concept is not interesting. If $k = n/2$ then an intersecting family contains at most one out of each pair $S, \overline{S}$, and so the measure of any such family (which is its size divided by $\binom{n}{k}$) is at most $1/2$, and this is achieved by *stars*, that is, all families containing some fixed element i.

The interesting case is when $k < n/2$. In this case, stars are intersecting families of measure k/n, but is this optimal? Let us try to mimic the proof in Section 8.1, constructing a noise operator T such that $\langle f, Tf \rangle = 0$ if f is the characteristic function of an intersecting family. We will have $Tf(x) = \mathbb{E}_{y \sim N(x)}[f(y)]$, where $N(x)$ is supported on sets disjoint from x. The only reasonable choice for $N(x)$ is a random set disjoint from x. This makes T the random walk operator of the Kronecker graph, in which two sets are connected if they are disjoint.

The operator T "doesn't care about names of elements", and so we know that its eigenspaces are the Fourier levels. To compute the eigenvalues, it suffices to compute $T\chi_d$ on some point x such that $\chi_d(x) = 1$, say one in which $x_1 = x_3 = \cdots = x_{2d-1} = 1$ and $x_2 = x_4 = \cdots = x_{2d} = 0$. Any y in the support of $N(x)$ has $y_1 = y_3 = \cdots = y_{2d-1} = 0$, so $T\chi_d(x)$ is $(-1)^d$ times the probability that $y_2 = y_4 = \cdots = y_{2d} = 1$. Out of the $\binom{n-k}{k}$ neighbors of x in the Kronecker graph, $\binom{n-k-d}{k-d}$ are of this form, and so

$$T\chi_d = (-1)^d \frac{\binom{n-k-d}{k-d}}{\binom{n-k}{k}} \chi_d.$$

Therefore if f is the characteristic function of an intersecting family $\mathcal{F}$, then

$$0 = \langle f, Tf \rangle = \sum_{d=0}^{k} (-1)^d \frac{\binom{n-k-d}{k-d}}{\binom{n-k}{k}} \|f^{=d}\|^2.$$

Following our footsteps in Section 8.1, we will single out $d = 0$, and lower bound all other eigenvalue by the minimal one, that is, the one maximizing $\binom{n-k-d}{k-d}$ over odd d. The interpretation of this quantity as the number of neighbors such that $y_2 = y_4 = \cdots = y_{2d} = 1$ makes it clear that the maximizer is $d = 1$ (the least number of constraints), and so

$$0 \geq \|f^{=0}\|^2 - \frac{\binom{n-k-1}{k-1}}{\binom{n-k}{k}} \|f^{>0}\|^2.$$

We can simplify the ratio of the binomials to $\frac{k}{n-k}$.

What is $f^{=0}$? It is the constant coefficient of the harmonic expansion of f. Since $\mathbb{E}[\chi_d] = 0$ for $d > 0$, we see that $f^{=0} = \mathbb{E}[f]$. Therefore, as in the p-biased case,

$$\frac{k}{n-k}(\mathbb{E}[f] - \mathbb{E}[f]^2) \geq \mathbb{E}[f]^2 \Longrightarrow \mathbb{E}[f] \leq \frac{k}{n-k}(1 - \mathbb{E}[f]).$$

This shows that $\frac{n}{n-k}\mathbb{E}[f] \leq \frac{k}{n-k}$, and so $\mathbb{E}[f] \leq k/n$. Furthermore, as in the p-biased case, equality can only happen if $\deg f \leq 1$, which implies that f is a dictator and so a star (exercise). We can even deduce that near-maximizers are close to stars, at least when k/n is bounded away from 0, using a slice version of the Friedgut–Kalai–Naor theorem.

Exercise Show that if $\deg f \leq 1$ then f is a dictator (unless $k \in \{1, n-1\}$), and conclude that if $2 \leq k < n/2$ and $\mathcal{F}$ is an intersecting family of measure k/n, then $\mathcal{F}$ is a star.

13.4 Coupling the slice and the cube

When k/n is constant (an assumption we make from now on), the slice behaves similarly to the p-biased cube. For example if we put $p = k/n$, then $\mathbb{E}[x_1] = p$, and $\mathbb{E}[x_1 x_2] = \frac{k(k-1)}{n(n-1)} \approx p^2$; the same holds for all monomials of degree $o(\sqrt{n})$.

One particularly transparent way to connect the slice $\binom{[n]}{k}$ and the corresponding p-biased cube (with $p = k/n$) was found by Noam Lifshitz (as yet unpublished). The idea is to consider a coupling of the slice and the p-biased cube. The coupling is very simple: we choose $x \sim \mu_p$, and then take a uniformly random subset (if $|x| \geq k$) or superset (if $|x| \leq k$) of x of size exactly k.

Let $T_{\mu\to\nu}$ be an operator mapping functions on the p-biased cube to functions on the slice as follows: $T_{\mu\to\nu} f(y) = \mathbb{E}[f(x)]$, where x is distributed as the first element of the coupling above, subject to the second element being y. Define $T_{\nu\to\mu}$ similarly in the other direction. We define a noise operator on the slice by

$$T^\nu_\rho f = T_{\mu\to\nu} T^\mu_\rho T_{\nu\to\mu},$$

where T^μ_ρ is the usual noise operator on the p-biased cube. Since $T_{\mu\to\nu}, T_{\nu\to\mu}$ are averaging operators they are contractive (by an application of the triangle inequality), and so (aiming, arbitrarily, at $(2, 4/3)$-hypercontractivity)

$$\|T^\nu_\rho f\|_2 = \|T_{\mu\to\nu} T^\mu_\rho T_{\nu\to\mu} f\|_2 \leq \|T^\mu_\rho T_{\nu\to\mu} f\|_2 \leq \|T_{\nu\to\mu} f\|_{4/3} \leq \|f\|_{4/3}.$$

It remains to find out how T^ν_ρ operates on the harmonic expansion of f. Since T^ν_ρ "doesn't care about names of coordinates", it suffices to consider $f = \chi_d$.

What can we say about $T_{\nu\to\mu}\chi_d$? Since the operators $T_{\nu\to\mu}$ and $T_{\mu\to\nu}$ are adjoint, we have $\langle T_{\nu\to\mu}\chi_d, g\rangle = \langle \chi_d, T_{\mu\to\nu} g\rangle$. If $\deg g < d$ then g can be written as a linear combination of monomials of degree $e < d$. Each such monomial x_S only depends on $e < d$ variables, and so $T_{\mu\to\nu} x_S$ can be written as a polynomial in these variables, hence its harmonic expansion has degree at most e. This shows that $\deg T_{\mu\to\nu} g < d$, and so the inner product vanishes. In other words, $T_{\nu\to\mu}\chi_d$ has no Fourier coefficients below level d. Conversely, each monomial in χ_d depends on only d variables, and so $T_{\nu\to\mu}\chi_d$ has degree at most d. We conclude that $T_{\nu\to\mu}\chi_d$ lies in the d'th level of the Fourier expansion. This is useful since it implies that

$$T^\nu_\rho \chi_d = \rho^d T_{\mu\to\nu} T_{\nu\to\mu} \chi_d.$$

It remains to understand the effect of $T_{\mu\to\nu} T_{\nu\to\mu}$ on χ_d. We know that χ_d is an eigenvector of $T_{\mu\to\nu} T_{\nu\to\mu}$, with some eigenvalue λ_d which can be calculated as

$$\lambda_d = \frac{\langle T_{\mu\to\nu} T_{\nu\to\mu}\chi_d, \chi_d\rangle}{\|\chi_d\|^2} = \frac{\|T_{\nu\to\mu}\chi_d\|^2}{\|\chi_d\|^2} = \frac{\Pr[T_{\nu\to\mu}\chi_d \neq 0]}{\Pr[\chi_d \neq 0]}.$$

The denominator is exactly

$$2^d \frac{k^{\underline{d}}(n-k)^{\underline{d}}}{n^{\underline{2d}}} \approx (2p(1-p))^d.$$

In order to estimate the numerator, let us consider the distribution of the Hamming weight of x. If $\ell \geq k$, then there are $\binom{n-k}{\ell-k}$ vectors $x \geq y$ of Hamming weight ℓ. Since y is obtained from such an x with probability $1/\binom{k}{\ell}$, the probability that $|x| = \ell$ is proportional to

$$p^\ell (1-p)^{n-\ell} \frac{\binom{n-k}{\ell-k}}{\binom{\ell}{k}}.$$

When ℓ increases by 1, this gets multiplied by a factor of

$$\frac{p}{1-p}\frac{n-\ell}{\ell+1} = \frac{k}{\ell+1}\frac{n-\ell}{n-k}.$$

Similarly, if $\ell \leq k$ then the probability that $|x| = \ell$ is proportional to

$$p^\ell (1-p)^{n-\ell} \frac{\binom{k}{\ell}}{\binom{n-\ell}{k-\ell}}.$$

When ℓ decreases by 1, this gets multiplied by a factor of

$$\frac{1-p}{p} \frac{\ell}{n-\ell+1} = \frac{n-k}{n-\ell+1} \frac{\ell}{k}.$$

In both cases, once ℓ deviates by $\sqrt{n}$ from k, the ratio becomes $1 - O(1/\sqrt{n})$, and so it drops exponentially every $\sqrt{n}$ steps. This means that $|x|$ is concentrated in a distance of roughly $O(\sqrt{n})$ from k.

If we denote by δ the deviation of $|x|$ from k, then fixing $|x|$, the probability that $x \setminus y$ (if $|x| \geq k$) or $y \setminus x$ (if $|x| \leq k$) contains one of the first $2d$ coordinates is $O(d\delta/n)$. Since δ is of order roughly $O(\sqrt{n})$, we see that $\Pr[T_{\nu\to\mu}\chi_d \neq 0]$ differs from $\Pr[\chi_d \neq 0]$ by at most roughly $O(d/\sqrt{n})$. The error term is much smaller than the main term as long as $d \ll \log n$, in which case

$$\lambda_d \approx 1 \pm \frac{e^{O(d)}}{\sqrt{n}}.$$

Just as in the case of T_ρ and T^*_ρ, this implies that $\|T^\nu_\rho f\|_2$ is close to $\|T_\rho f\|_2$ and $\|T^*_\rho f\|_2$, at least for low degree functions. This suffices to show that $\|f\|_4 = O(\|f\|_2)$ for functions of bounded degree, for example, and so can be used to derive many of the theorems proved in these lecture notes.

Exercise Obtain a more precise estimate for the eigenvalues of T^ν_ρ, and deduce a concrete statement of hypercontractivity for low-degree functions on the slice.

References

[AA14] Scott Aaronson and Andris Ambainis. The need for structure in quantum speedups. *Theory Comput.*, 10:133–166, 2014. doi:10.4086/toc.2014.v010a006.

[AL93] Miklós Ajtai and Nathal Linial. The influence of large coalitions. *Combinatorica*, 13(2):129–145, 1993. doi:10.1007/BF01303199.

[BB14] Artūrs Bačkurs and Mohammad Bavarian. On the sum of L_1 influences. In *IEEE 29th Conference on Computational Complexity—CCC 2014*, pages 132–143. IEEE Computer Soc., Los Alamitos, CA, 2014. doi:10.1109/CCC.2014.21.

[Bor75] Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30(2):207–216, 1975. doi:10.1007/BF01425510.

[Bou02] Jean Bourgain. On the distributions of the Fourier spectrum of Boolean functions. *Israel J. Math.*, 131:269–276, 2002. doi:10.1007/BF02785861.

[CHS20] John Chiarelli, Pooya Hatami, and Michael Saks. An asymptotically tight bound on the number of relevant variables in a bounded degree Boolean function. *Combinatorica*, 40(2):237–244, 2020. doi:10.1007/s00493-019-4136-7.

[CW01] Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Math. Res. Lett.*, 8(3):233–248, 2001. doi:10.4310/MRL.2001.v8.n3.a1.

[DMN16] Anindya De, Elchanan Mossel, and Joe Neeman. Majority is stablest: discrete and SoS. *Theory Comput.*, 12:Paper No. 4, 50, 2016. doi:10.4086/toc.2016.v012a004.

[Dun76] Charles F. Dunkl. A Krawtchouk polynomial addition theorem and wreath products of symmetric groups. *Indiana Univ. Math. J.*, 25(4):335–358, 1976. doi:10.1512/iumj.1976.25.25030.

[Eld15] Ronen Eldan. A two-sided estimate for the Gaussian noise stability deficit. *Invent. Math.*, 201(2):561–624, 2015. doi:10.1007/s00222-014-0556-6.

[ER60] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[FHKL16] Yuval Filmus, Hamed Hatami, Nathan Keller, and Noam Lifshitz. On the sum of the L_1 influences of bounded functions. *Israel J. Math.*, 214(1):167–192, 2016. doi:10.1007/s11856-016-1355-0.

[Fil16] Yuval Filmus. Friedgut-Kalai-Naor theorem for slices of the Boolean cube. *Chic. J. Theoret. Comput. Sci.*, pages Art. 14, 17, 2016. doi:10.4086/cjtcs.2016.014.

[FK96] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proc. Amer. Math. Soc.*, 124(10):2993–3002, 1996. doi:10.1090/S0002-9939-96-03732-X.

[FKN02] Ehud Friedgut, Gil Kalai, and Assaf Naor. Boolean functions whose Fourier transform is concentrated on the first two levels. *Adv. in Appl. Math.*, 29(3):427–437, 2002. doi:10.1016/S0196-8858(02)00024-6.

[Fri98] Ehud Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–35, 1998. doi:10.1007/PL00009809.

[Fri99] Ehud Friedgut. Sharp thresholds of graph properties, and the k-sat problem. *J. Amer. Math. Soc.*, 12(4):1017–1054, 1999. With an appendix by Jean Bourgain. doi:10.1090/S0894-0347-99-00305-7.

[GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42(6):1115–1145, 1995. doi:10.1145/227683.227684.

[Hat12] Hamed Hatami. A structure theorem for Boolean functions with small total influences. *Ann. of Math. (2)*, 176(1):509–533, 2012. doi:10.4007/annals.2012.176.1.9.

[Hof70] Alan J. Hoffman. On eigenvalues and colorings of graphs. In *Graph Theory and its Applications (Proc. Advanced Sem., Math. Research Center, Univ. of Wisconsin, Madison, Wis., 1969)*, pages 79–91. Academic Press, New York, 1970.

[Kin02] Guy Kindler. *Property testing, PCP and Juntas*. PhD thesis, Tel-Aviv University, 2002.

[KK20] Nathan Keller and Ohad Klein. A structure theorem for almost low-degree functions on the slice. *Israel J. Math.*, 240(1):179–221, 2020. doi:10.1007/s11856-020-2062-4.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on Boolean functions. In *Proc. 29th Ann. Symp. on Foundations of Comp. Sci.*, pages 68–80. Computer Society Press, 1988.

[KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM J. Comput.*, 37(1):319–357, 2007. doi:10.1137/S0097539705447372.

[KKO18] Guy Kindler, Naomi Kirshner, and Ryan O'Donnell. Gaussian noise sensitivity and Fourier tails. *Israel J. Math.*, 225(1):71–109, 2018. doi:10.1007/s11856-018-1646-8.

[KLLM19] Peter Keevash, Noam Lifshitz, Eoin Long, and Dor Minzer. Hypercontractivity for global functions and sharp thresholds, 2019. URL: https://arxiv.org/abs/1906.05568, arXiv:1906.05568.

[KR08] Subhash Khot and Oded Regev. Vertex cover might be hard to approximate to within $2-\epsilon$. *J. Comput. System Sci.*, 74(3):335–349, 2008. doi:10.1016/j.jcss.2007.06.019.

[KS04] Guy Kindler and Shmuel Safra. Noise-resistant Boolean functions are juntas, 2004. Unpublished manuscript.

[Lov79] László Lovász. On the Shannon capacity of a graph. *IEEE Trans. Inform. Theory*, 25(1):1–7, 1979. doi:10.1109/TIT.1979.1055985.

[MN15] Elchanan Mossel and Joe Neeman. Robust optimality of Gaussian noise stability. *J. Eur. Math. Soc. (JEMS)*, 17(2):433–482, 2015. doi:10.4171/JEMS/507.

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Ann. of Math. (2)*, 171(1):295–341, 2010. doi:10.4007/annals.2010.171.295.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi:10.1017/CBO9781139814782.

[Rus82] Lucio Russo. An approximate zero-one law. *Z. Wahrsch. Verw. Gebiete*, 61(1):129–139, 1982. doi:10.1007/BF00537230.

[Wel20] Jake Wellens. Relationships between the number of inputs and other complexity measures of boolean functions, 2020. arXiv:2005.00566.