Boolean Function Analysis

Arpon Basu

Last updated November 20, 2024

Contents

Notation

Let $n \in \mathbb{N} = \{1, 2, \ldots\}$. Then we refer to the set $\{1, 2, \ldots, n\}$ as $[n]$.
Given two sets $S, T \subseteq [n]$, we use $S \triangle T$ or $S \oplus T$ to denote the symmetric difference of S and T, i.e. $(S \setminus T) \cup (T \setminus S)$.
We shall canonically identify $\mathbb{F}_2^n$ with $\{-1, 1\}^n$, with $(x_1, \ldots, x_n)$ mapping to $((-1)^{x_1}, \ldots, (-1)^{x_n})$. Note that $1 \in \mathbb{F}_2$ maps to $-1 \in \{-1, 1\}$, i.e. -1 is the "true" bit according to our scheme.

Acknowledgements

All the material in these notes is from Ryan O'Donnell's CMU Lecture series [O'D12], and his notes [O'D21]. The order, and the depth in which topics have been covered, though, is our own prerogative.

§1. Boolean Function Analysis: Introduction and Preliminaries

Theorem 1.1 (Fourier Analysis on the Boolean hypercube)**.** Let n be a natural number. Consider any function $f : \{-1, 1\}^n \mapsto \mathbb{R}$. Then there exists a unique function $\widehat{f} : 2^{[n]} \mapsto \mathbb{R}$ such that

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) x_S$$

for every $x = (x_1, \ldots, x_n) \in \{-1, 1\}^n$.
The function $\widehat{f}$ is also known as the *Fourier transform* of f.

Proof. We prove this statement by induction on n. For $n = 1$, note that any function $f : \{-1, 1\} \mapsto \mathbb{R}$ can be written as $f(x) = \left(\frac{f(1)+f(-1)}{2}\right) + \left(\frac{f(1)-f(-1)}{2}\right) \cdot x$, and further note that this representation is the unique representation of the form $\widehat{f}(\varnothing) + \widehat{f}(\{1\}) \cdot x$.
Thus the base case of our induction hypothesis is verified. Now, suppose the statement is true for some $n = k-1, k \geqslant 2$. Then note that any function $f : \{-1, 1\}^k \mapsto \mathbb{R}$ can be written as

$$f(x_1, x_2, \ldots, x_k) = \left(\frac{f(1, x_2, \ldots, x_k) + f(-1, x_2, \ldots, x_k)}{2}\right) + \left(\frac{f(1, x_2, \ldots, x_k) - f(-1, x_2, \ldots, x_k)}{2}\right) \cdot x_1$$

But $g(x_2, \ldots, x_k) := \frac{f(1,x_2,\ldots,x_k)+f(-1,x_2,\ldots,x_k)}{2}$ and $h(x_2, \ldots, x_k) := \frac{f(1,x_2,\ldots,x_k)-f(-1,x_2,\ldots,x_k)}{2}$ are functions on the $(k-1)$-dimensional Boolean hypercube and thus by the induction hypothesis possess a unique Fourier transform. Then combining the Fourier transforms for those two functions yields a Fourier transform for f.
We shall shortly see why this expansion is unique. ■

Definition 1.1 (Multilinear Polynomials)**.** A multivariate polynomial is called multilinear if it is linear (affine) in each of its variables. For example, $3x - 4xy + 5z - 2$ is a multilinear polynomial in x, y, z, but $x^2 + 4xy$ is not.

Corollary 1.2. *Any* function on the Boolean hypercube is equivalent to a multilinear polynomial of degree at most n.

Proposition 1. Every polynomial of degree d over the Boolean hypercube is equivalent to a multilinear polynomial of degree at most d. Furthermore, because of the uniqueness of the Fourier transform, this multilinear polynomial is also the Fourier transform of our polynomial.

Proof. Note that over the Boolean hypercube, every polynomial is equivalent to a multilinear polynomial of lower degree: One can see this even without invoking the Fourier expansion of the polynomial. Indeed, note that if $x_i \in \{-1, 1\}$, then $x_i^2 = 1$. Consequently, every term $\prod_{i=1}^n x_i^{k_i}$ in the polynomial can be replaced by the multilinear term $\prod_{i=1}^n x_i^{k_i \bmod 2}$, and thus we get an equivalent multilinear polynomial with a degree at most the original polynomial, as desired. ∎

Corollary 1.3. Multilinear polynomials are their own Fourier decompositions.

The material covered upto here can also be found, verbatim, in my notes on the SoS hierarchy.
We now define the *character functions*.

Definition 1.2. For every $S \subseteq [n]$, we define $\chi_S := x_S$, i.e. $\chi_S : \{-1, 1\}^n \mapsto \{-1, 1\} \subseteq \mathbb{R}$ is a function.

Remark. When we represent the boolean cube as $\mathbb{F}_2^n$, then the character function $\chi_S : \mathbb{F}_2^n \mapsto \{-1, 1\}$ is:

$$\chi_S(x) = (-1)^{\sum_{i \in S} x_i}$$

Note that $\chi_S(x + y) = \chi_S(x)\chi_S(y)$.

Then note that every function on the Boolean hypercube can be written as $f = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S$, i.e. $\{\chi_S\}_{S \subseteq [n]}$ *span* $\mathbb{R}^{\{-1,1\}^n}$. Now, we shall quickly define the notion of a dot product in our vector space $\mathbb{R}^{\{-1,1\}^n}$:

Definition 1.3 (Dot Product)**.** Given $f, g \in \mathbb{R}^{\{-1,1\}^n}$, we define $\langle f, g \rangle := 2^{-n} \sum_{x \in \{-1,1\}^n} f(x)g(x) = \mathbb{E}_{x \sim \{-1,1\}^n} [fg]$, where x is sampled uniformly from $\{-1, 1\}^n$.

This dot product is just the rescaled version of the usual dot product on $\mathbb{R}$-vector spaces. In the case that f, g are themselves Boolean-valued functions, the dot product measures "similarity" between them, i.e. if $f, g : \{-1, 1\}^n \mapsto \{-1, 1\}$ are two functions, then

$$\mathbb{E}_x [fg] = \Pr_x(f(x) = g(x)) - \Pr_x(f(x) \neq g(x)) = 1 - 2 \Pr_x(f(x) \neq g(x)) = 1 - 2\,\mathrm{dist}(f, g)$$

where $\mathrm{dist}(f, g)$ is the fractional Hamming distance between f and g.
We now lay the basis (no pun intended) for Boolean function analysis.

Theorem 1.4. $\{\chi_S\}_{S \subseteq [n]}$ form an orthonormal basis of $\mathbb{R}^{\{-1,1\}^n}$ under the aforementioned dot product, i.e. $\langle \chi_S, \chi_T \rangle = \mathbb{1}_{S=T}$. In particular, the Fourier expansion of a function $f \in \mathbb{R}^{\{-1,1\}^n}$ is unique, since the Fourier expansion of f is the expression of f in terms of a basis of $\mathbb{R}^{\{-1,1\}^n}$.

Proof. Note that $\langle \chi_S, \chi_T \rangle = \mathbb{E}_x [x_S \cdot x_T] = \mathbb{E}_x [x_{S \oplus T}]$. If $S \neq T$, then $x_{S\oplus T}$ is not identically equal to 1 and thus will vanish when we take expectations. ∎

Corollary 1.5. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, we have $\widehat{f}(S) = \langle f, \chi_S \rangle$.

Corollary 1.6 (Plancherel's theorem)**.** For any $f, g \in \mathbb{R}^{\{-1,1\}^n}$, we have $\langle f, g \rangle = \sum_{S \subseteq [n]} \widehat{f}(S)\widehat{g}(S)$.

Corollary 1.7 (Parseval's Theorem)**.** For any $f \in \mathbb{R}^{\{-1,1\}^n}$, we have $\|f\|_2^2 := \langle f, f \rangle = \sum_{S \subseteq [n]} \widehat{f}(S)^2$. In particular, if $f : \{-1,1\}^n \mapsto \{-1,1\}$ is a Boolean valued function, then $\sum_{S \subseteq [n]} \widehat{f}(S)^2 = 1$.

Remark. For any $p \geqslant 1$, we can similarly define $\|f\|_p := \mathbb{E}\left[|f|^p\right]^{1/p}$, and $\|\cdot\|_p$ is a norm on $\mathbb{R}^{\{-1,1\}^n}$. Recall from analysis, that if $a > b \geqslant 1$, then $\|f\|_a \geqslant \|f\|_b$ with equality holding iff f is a constant function. Also recall Hölder's inequality, which says that $\|f\|_p \cdot \|g\|_q \geqslant \|fg\|_1$ for any $p, q \geqslant 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

Definition 1.4. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, we define the *weight* of f at S to be $\widehat{f}(S)^2$. Given any $0 \leqslant k \leqslant n$, we also define $W^k[f] := \sum_{|S|=k} \widehat{f}(S)^2$, $W^{\leqslant k}[f] := \sum_{|S| \leqslant k} \widehat{f}(S)^2$, $W^{>k}[f] := \sum_{|S|>k} \widehat{f}(S)^2$.

Remark. As our intuition about Boolean functions develops further, we shall see that "complicated" Boolean functions have a non-negligible fraction of their total weight in their high-frequency components. Conversely, functions for which $W^{\leqslant k}/W^{\leqslant n}$ ratio is close to 1 for some "small" k, are easy to deal with, and more easily understood and characterized.

Proposition 2. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, $\mathbb{E}[f] = \widehat{f}(\varnothing)$.

Proof. Note that $\mathbb{E}[f] = \mathbb{E}[f \cdot \chi_\varnothing]$, since $\chi_\varnothing$ is identically equal to 1. The proposition follows. ∎

Proposition 3. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, $\mathrm{Var}(f) = \mathbb{E}\left[f^2\right] - \mathbb{E}[f]^2 = \sum_{S \subseteq [n], S \neq \varnothing} \widehat{f}(S)^2 = W^{>0}[f]$.

Proof. Note that $\mathbb{E}\left[f^2\right] = \|f\|_2^2$, and thus, by Corollary 1.7, $\mathbb{E}\left[f^2\right] = \sum_{S \subseteq [n]} \widehat{f}(S)^2$. We are then done by Proposition 2. ∎

We end by proving a short identity about the variance of a Boolean function.

Lemma 1.8. Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a Boolean-valued function. Then $\mathrm{Var}(f) = 4\Pr(f(x) = 1) \cdot \Pr(f(x) \neq 1)$.

Proof. Write $\alpha = \Pr(f(x) = 1)$. Note that $\mathbb{E}[f] = \alpha - (1-\alpha) = 2\alpha - 1$. Thus $\mathrm{Var}(f) = \mathbb{E}\left[f^2\right] - \mathbb{E}[f]^2 = 1 - (2\alpha-1)^2 = 4\alpha(1-\alpha)$, as desired. ■

We now recast the basic notions of probability in our language. This will be useful later on, when we have to deal with the Fourier transforms of the convolutions of PDFs.

Definition 1.5 (Probability Density Functions)**.** A function $\varphi : \mathbb{F}_2^n \mapsto \mathbb{R}_{\geqslant 0}$ is a PDF if $\mathbb{E}_{x\sim\mathbb{F}_2^n}\left[\varphi(x)\right] = 1$, i.e. $2^{-n}\sum_{x\in\mathbb{F}_2^n}\varphi(x) = 1$. Equivalently, the probability mass function corresponding to φ is: $\Pr(x) = 2^{-n}\varphi(x)$.

Evidently, the uniform distribution is $\varphi \equiv 1$. The dirac delta concentrated on $(0,\ldots,0) \in \mathbb{F}_2^n$ is $\varphi_{\{0\}}(x) = 2^n \mathbb{1}_{x=0^n}$. We now prove that sampling and taking inner products are the same operations.

Proposition 4. Let φ be a PDF. Then $\mathbb{E}_{y\sim\varphi}\left[f(y)\right] = \langle\varphi, f\rangle$, where $f : \mathbb{F}_2^n \mapsto \mathbb{R}$ is any Boolean function.

Proof. Note that $\mathbb{E}_{y\sim\varphi}\left[f(y)\right] = 2^{-n}\sum_{y\in\mathbb{F}_2^n}\varphi(y)f(y) = \langle\varphi, f\rangle$. ■

The reason why probability densities arise in the study of Boolean Function Analysis is that Fourier coefficients multiply under the convolution of PDFs.
Before that, let's recall what convolutions were: Let φ, ψ be probability distributions. Sample $y \sim \varphi, z \sim \psi$ (independently), and set $x := y + z$. Now, note that the PDF of x is given by $\mathbb{E}_{y\sim\varphi}\left[\psi(x-y)\right]$. But $\mathbb{E}_{y\sim\varphi}\left[\psi(x-y)\right] = \mathbb{E}_{y\sim\mathbb{F}_2^n}\left[\varphi(y)\psi(x-y)\right]$. We thus define the *convolution* of φ, ψ to be:

$$(\varphi * \psi)(x) := \mathbb{E}_{y\sim\mathbb{F}_2^n}\left[\varphi(y)\psi(x-y)\right]$$

Standard probability theory tells us that $*$ is commutative and associative. We can now state the connection between convolutions and Fourier coefficients:

Theorem 1.9. Let f, g be PDFs. Then $\widehat{f * g}(S) := \widehat{f}(S)\widehat{g}(S)$.

Proof. Note that

$$\widehat{f * g}(S) = \langle f * g, \chi_S\rangle = \mathbb{E}_{x\sim\mathbb{F}_2^n}\left[(f*g)(x)\cdot\chi_S(x)\right] = \mathbb{E}_{x\sim\mathbb{F}_2^n}\left[\mathbb{E}_{y\sim\mathbb{F}_2^n}\left[f(y)g(x-y)\right]\cdot\chi_S(x)\right]$$

$$= \mathbb{E}_{y,z\overset{\text{i.i.d}}{\sim}\mathbb{F}_2^n}\left[f(y)g(z)\chi_S(y+z)\right]$$

Note that $\chi_S(y+z) = \chi_S(y)\chi_S(z)$. Thus

$$\widehat{f * g}(S) = \mathbb{E}_{y,z\overset{\text{i.i.d}}{\sim}\mathbb{F}_2^n}\left[f(y)\chi_S(y)g(z)\chi_S(z)\right] = \mathbb{E}_{y\sim\mathbb{F}_2^n}\left[f(y)\chi_S(y)\right]\mathbb{E}_{z\sim\mathbb{F}_2^n}\left[g(z)\chi_S(z)\right] = \widehat{f}(S)\widehat{g}(S)$$

as desired. ■

1.1. Linearity Testing

Note that if a function $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2$ is linear, then:

1. $f(x+y) = f(x) + f(y)$ for all x, y.
2. $f(x) = \sum_{i=1}^n a_i x_i$ where $a_i \in \mathbb{F}_2 = \{0,1\}$ for all $i \in [n]$, i.e. $f = \chi_S$, where $S := \{i : a_i = 1\}$.

Thus, we say that a function $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2$ is *approximately linear* if $\operatorname{dist}(f, \chi_S) = \varepsilon$ for some $S \subseteq [n]$, i.e. f is close to a character.
The notion of approximate linearity lends itself well to the setting of *property testing*, which we describe below:
Suppose we have a function $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2$, and suppose we have *oracle access* to f, i.e. we can query the value of $f(x)$, given some $x \in \mathbb{F}_2^n$. Using as few oracle queries as possible, we want to determine if f is approximately linear.
The very famous Blum-Luby-Rubinfeld (BLR) test [BLR90] provides a rather surprising resolution to this question.

Theorem 1.10 (BLR Linearity Test). Given $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2$, choose x, y from $\mathbb{F}_2^n$ in an i.i.d manner. Declare f to be linear if $f(x+y) = f(x) + f(y)$.
If the above test declares f to be linear with probability $1-\varepsilon$, then f is ε-close to some χ_S.

Remark. We repeat the above test M times to obtain an estimator for ε.

Proof. View f as a map from $\mathbb{F}_2^n$ to $\{-1, 1\}$, and thus the BLR test is `f(x)f(y)==f(x+y)`, or equivalently, we accept iff $\frac{1}{2} + \frac{1}{2} f(x)f(y)f(x+y) = 1$, and reject if the expression is 0. Thus,

$$1 - \varepsilon = \Pr(\text{BLR accepts } f) = \mathbb{E}_{x,y}\left[\frac{1}{2} + \frac{1}{2} f(x)f(y)f(x+y)\right] = \frac{1}{2} + \frac{1}{2}\mathbb{E}_{x,y}\left[f(x)f(y)f(x+y)\right]$$

$$= \frac{1}{2} + \frac{1}{2}\mathbb{E}_x\left[f(x)\cdot(f*f)(x)\right] = \frac{1}{2} + \frac{1}{2}\langle f, f*f\rangle = \frac{1}{2} + \frac{1}{2}\sum_{S\subseteq[n]} \widehat{f}(S)\widehat{f*f}(S) = \frac{1}{2} + \frac{1}{2}\sum_{S\subseteq[n]} \widehat{f}(S)^3$$

Thus,

$$\sum_{S\subseteq[n]} \widehat{f}(S)^3 = 1 - 2\varepsilon$$

However,

$$1 - 2\varepsilon = \sum_{S\subseteq[n]} \widehat{f}(S)^3 \leqslant \left(\max_{S\subseteq[n]} \widehat{f}(S)\right)\cdot \sum_{S\subseteq[n]} \widehat{f}(S)^2$$

Since f is a Boolean-valued function, by Parseval's theorem, $\sum_{S\subseteq[n]} \widehat{f}(S)^2 = 1$, and thus $\max_{S\subseteq[n]} \widehat{f}(S) \geqslant 1-2\varepsilon$, and thus there exists some $S^* \subseteq [n]$ such that $\widehat{f}(S^*) \geqslant 1-2\varepsilon$. But that implies $\widehat{f}(S^*) = \langle f, \chi_{S^*}\rangle = 1 - 2\operatorname{dist}(f, \chi_{S^*}) \geqslant 1-2\varepsilon$, and we're done. ∎

Note that even though we can determine if f is close to *some* linear function in $\mathcal{O}(1)$ queries, actually determining the linear function will take us $\geqslant n$ queries.
This naturally leads us to the next question: Suppose f is ε-close to some χ_{S^*}, but we don't know what S^*. Can we nevertheless evaluate the output of χ_{S^*} on some given input x? Note that directly querying f may not work, as x may be one of those inputs where f gives the wrong output.
Thus, consider the following algorithm:
For this algorithm, we are assured that there is some $S^* \subseteq [n]$ such that $\operatorname{dist}(f, \chi_{S^*}) \leqslant \varepsilon$. Note that this algorithm

Algorithm 1: Local-Correct

Data: $f \in \{-1, 1\}^{\mathbb{F}_2^n}, x \in \mathbb{F}_2^n$
Result: $\chi_{S^*}(x)$, correct with probability $\geqslant 1 - 2\varepsilon$
1 Pick y uniformly from $\mathbb{F}_2^n$;
2 **return** $f(x)f(x+y)$

returns the correct value of $\chi_{S^*}(x)$ (with high probability) for *every* x, while if we directly queried $f(x)$, then for some x, we would be wrong with probability 1. Thus, while averaged over $\mathbb{F}_2^n$, the success probability of Algorithm 1 and $f(x)$ is the same, Algorithm 1 gives us a pointwise guarantee which direct querying doesn't.

Proof of correctness of Algorithm 1. Since $y, x+y$ are uniformly distributed (though not independent), except with probability $\leqslant 2\varepsilon$, $f(y) = \chi_{S^*}(y), f(x+y) = \chi_{S^*}(x+y)$, and thus $f(y)f(x+y) = \chi_{S^*}(y)\chi_{S^*}(x+y) = \chi_{S^*}(x+2y) = \chi_{S^*}(x)$, as desired. ■

§2. Social Choice Theory

Observe that a function $f : \{-1,1\}^n \mapsto \{-1,1\}$ can be seen as a voting scheme of an election between two candidates, 1 and -1. For every possible voting instance in $\{-1,1\}^n$, f selects an output for that instance.
To quickly see some voting rules, we have:

1. Majority (Maj_n): Assume n is odd. Then Maj_n simply selects the majority vote. In other words, for any $x \in \{-1,1\}^n$, $\mathrm{Maj}_n(x) := \mathrm{sign}(x_1 + \cdots + x_n)$.
2. Weighted Majority: The weighted majority rule, also known as the linear thresholding rule, is defined as $\mathrm{sign}(a_0 + a_1 x_1 + \cdots + a_n x_n)$, where $a_0, \ldots, a_n \in \mathbb{R}$.
3. Dictator: Boolean functions of the form $\chi_{\{i\}} =: \chi_i$ are called dictators since their entire output depends upon a single bit x_i.
4. k-juntas: Boolean functions whose output depends upon $\leqslant k$ input bits. For example, the number of 1-juntas is $2n+2$ (2 constant functions, n dictators, and n anti-dictators (of the form $-\chi_i$)).
5. $\mathrm{OR}_n : \{-1,1\}^n \mapsto \{-1,1\}$: The OR function is -1 if and only if the input is $(-1, \ldots, -1)$.
6. $\mathrm{AND}_n : \{-1,1\}^n \mapsto \{-1,1\}$: The AND function is 1 if and only if the input is $(1, \ldots, 1)$.
7. Tribes: We define $\mathrm{Tribes}_{w,s} : \{-1,1\}^{sw} \mapsto \{-1,1\}$ as follows:

$$\mathrm{Tribes}_{w,s}(x_{1,1}, \ldots, x_{1,w}, \ldots, x_{s,1}, \ldots, x_{s,w}) := \mathrm{OR}(\mathrm{AND}(x_{1,1}, \ldots, x_{1,w}), \ldots, \mathrm{AND}(x_{s,1}, \ldots, x_{s,w}))$$

   Note that $\mathbb{E}\left[\mathrm{Tribes}_{w,s}\right] = 1 - 2(1-2^{-w})^s$. Thus, if $s \sim 2^w \ln 2$, then $\mathbb{E}\left[\mathrm{Tribes}_{w,s}\right] \approx 0$.

We say that a voting rule is *unbiased* if $\mathbb{E}[f] = 0$ (this is assuming that the co-domain of f is $\{-1,1\}$): Maj_n and χ_i are unbiased. $\mathrm{Tribes}_{w,s}$, where $s \sim 2^w \ln 2$, is also approximately unbiased.
A function $f : \{-1,1\}^n \mapsto \{-1,1\}$ is called unanimous if $f(b, \ldots, b) = b$ for $b \in \{-1,1\}$. Unanimity is an obvious requirement for any voting rule. A function $f : \{-1,1\}^n \mapsto \{-1,1\}$ is called odd if $f(-x) = -f(x)$ for all $x \in \{-1,1\}^n$.
We call a function $f : \{-1,1\}^n \mapsto \mathbb{R}$ *monotone* if $f(x'_1, x'_2, \ldots, x'_n) \geqslant f(x_1, x_2, \ldots, x_n)$ whenever $x'_1 \geqslant x_1, x'_2 \geqslant x_2, \ldots, x'_n \geqslant x_n$. Monotonicity is also a very natural requirement for a voting rule.
A function $f : \{-1,1\}^n \mapsto \{-1,1\}$ is called symmetric if $f(\sigma(x)) = f(x)$ for every permutation $\sigma \in S_n$, $x \in \{-1,1\}^n$. Maj_n is symmetric.
A function $f : \{-1,1\}^n \mapsto \{-1,1\}$ is called transitive if for any $i, j \in [n]$, there exists a permutation $\pi \in S_n$ such that $\pi(i) = j$, and $f(\pi(x)) = f(x)$ for every $x \in \{-1,1\}^n$. Symmetric functions are obviously transitive. Tribes is transitive but not symmetric.
Maj is the most common voting rule. In fact, it is the *only* voting rule satisfying a few obvious requirements:

Theorem 2.1 (May's Theorem)**.** Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a monotone, unanimous, odd and symmetric boolean function. Then $f = \mathrm{Maj}_n$.

Proof. Since f is symmetric, the output of f only depends on the number of -1s among its n input bits. Thus there exists a function $g : \{0, 1, \ldots, n\} \mapsto \{-1,1\}$ such that $f(x) = g(\mathrm{wt}(x))$, where $\mathrm{wt}(x)$ is the number of -1s among its n input bits. Since f is monotone, g is a non-increasing function. Since f is unanimous, $g(0) = 1, g(n) = -1$. It is now easy to see that since g is odd, g corresponds to the majority function, as desired. ■

We also define the very important notion of *influence*:

Definition 2.1 (Influence)**.** For any $i \in [n]$, and any Boolean function on $\mathbb{F}_2^n$, we define the influence of i on f, denoted $\mathrm{Inf}_i(f)$, to be $\Pr_{x \sim \mathbb{F}_2^n}\left(f(x) \neq f(x \oplus i)\right)$, where $x \oplus i$ is just x, with the i^{th} bit flipped. Interpreted differently, the influence of a bit i is the probability that it flips the result with its vote.

Example. We shall work out some examples of influence:

1. $\mathrm{Inf}_i(\mathrm{AND}_n)$: Note that the only time the i^{th} bit has the power to flip the result is if all the other bits have the same value. Thus $\mathrm{Inf}_i(\mathrm{AND}_n) = 2^{-(n-1)}$.

2. It is easy to see that $\mathrm{Inf}_i(\chi_j) = \mathbb{1}_{i=j}$, and $\mathrm{Inf}_i(\chi_{[n]}) = 1$.

3. Note that
$$\mathrm{Inf}_i(\mathrm{Maj}_n) = \frac{\binom{n-1}{(n-1)/2}}{2^{n-1}} = \Theta\left(\frac{1}{\sqrt{n}}\right)$$
Indeed, the i^{th} bit flips the vote if and only if the other votes are evenly split between 1 and -1.

4. The tribes function: Since Tribes is transitive, the influence of all variables on the tribe function is the same. Calculating the influence yields
$$\mathrm{Inf}_i(\mathrm{Tribes}_{w,s}) = \frac{2(2^w-1)^{s-1}}{2^{ws}}$$
Putting $n = ws$, $s \sim 2^w \ln 2$ (to make $\mathrm{Tribes}_{w,s}$ approximately unbiased), and simplifying yields $\mathrm{Inf}_i(\mathrm{Tribes}_{w,s}) = \Theta(\frac{\ln n}{n})$.

The influence of the Tribes function is essentially as low as it gets, thanks to the famous Kahn-Kalai-Linial theorem:

Theorem 2.2 (Kahn-Kalai-Linial Theorem ([KKL88]))**.** For every $f : \{-1,1\}^n \mapsto \{-1,1\}$, there exists an $i \in [n]$, such that
$$\mathrm{Inf}_i(f) \geqslant \Omega\left(\frac{\ln n}{n}\right) \cdot \mathrm{Var}(f)$$

Remark. The $\mathrm{Var}(f)$ factor is just a 'normalizing factor'. Note that if f is unbiased, i.e. $\mathbb{E}[f] = 0$, then $\mathrm{Var}(f) = 1$ (since f is Boolean valued).

An even stronger version of the KKL theorem holds, thanks to Talagrand:

Theorem 2.3 (Talagrand's Theorem ([Tal94]))**.** For every $f : \{-1,1\}^n \mapsto \{-1,1\}$,
$$\sum_{i=1}^{n} \frac{\mathrm{Inf}_i(f)}{\ln\left(\frac{1}{\mathrm{Inf}_i(f)}\right)} \geqslant \Omega(\mathrm{Var}(f))$$

Remark. Talagrand's theorem shows that the total amount of influence can't be too small, i.e. if the maximum influence of any variable is small, then lots of variables must have that influence.

The notion of influence is so important that it deserves a definition even in the case the co-domain of our Boolean function is $\mathbb{R}$.

Definition 2.2 (Derivative operator)**.** Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a function, and let $i \in [n]$ be some bit. Then the derivative w.r.t i, $D_i(f) : \{-1,1\}^n \mapsto \mathbb{R}$, is defined as
$$\big(D_i(f)\big)(x) := \frac{f(x^{i\mapsto 1}) - f(x^{i\mapsto -1})}{2}$$
where $x^{i\mapsto b} := (x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n)$, for $b \in \{-1,1\}$.

Note that when the co-domain of f is $\{-1,1\}$, the range of $D_i(f)$ is a subset of $\{-1,0,1\}$. Furthermore, $\left(D_i(f)\right)(x) \neq 0$ only when the i^{th} bit is *pivotal* for x, i.e. flipping the i^{th} bit changes the output of the function. Thus, an indicator variable for whether the i^{th} bit is pivotal at x is given by $\left(D_i(f)\right)(x)^2$.
Now, when the co-domain of f is $\mathbb{R}$, $\left(D_i(f)\right)(x)^2$ is not an indicator variable in general, but it continues to indicate the extent to which i is pivotal, i.e. if i is not pivotal, then $\left(D_i(f)\right)(x) = 0$, while if i is pivotal, then $\left(D_i(f)\right)(x)^2$ measures the "magnitude of the flip". Thus, we define,

Definition 2.3 (Influence for general Boolean functions)**.** Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a function. We define the influence of the i^{th} bit as

$$\mathrm{Inf}_i(f) := \mathbb{E}_x\left[\left(D_i(f)\right)(x)^2\right]$$

Remark. Note that

$$\mathrm{Inf}_i(f) := \mathbb{E}_x\left[\left(D_i(f)\right)(x)^2\right] = \langle D_i(f), D_i(f)\rangle = \|D_i(f)\|_2^2$$

If f is monotone, then $D_i(f) \geqslant 0$, which conforms to our intuition of derivative from real analysis. Also note that all of the voting rules we have seen so far ((weighted) Majority, AND, OR, Tribes), are monotone.
We prove another very useful property of derivatives.

Proposition 5. Let $f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S$ be the Fourier decomposition of f. Then $D_i(f) = \sum_{i\in S} \widehat{f}(S)\chi_{S\setminus\{i\}}$.

Proof. Since D_i is a linear operator, we're done by observing that $D_i(\chi_S) = \mathbb{1}_{i\in S}\cdot\chi_{S\setminus\{i\}}$. ■

Corollary 2.4. Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a function. Then $\mathrm{Inf}_i(f) = \sum_{i\in S} \widehat{f}(S)^2$.

For monotone Boolean-valued functions, we have a remarkable characterization of influence.

Lemma 2.5. Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a **monotone** function. Then $\mathrm{Inf}_i(f) = \widehat{f}(\{i\}) = \widehat{f}(i)$.

Proof. Since f is a monotone Boolean valued function, $D_i(f) \geqslant 0$, and thus the range of $D_i(f)$ is a subset of $\{0,1\}$. Consequently, $D_i(f)^2 = D_i(f)$, and thus $\mathbb{E}\left[D_i(f)^2\right] = \mathbb{E}\left[D_i(f)\right] = \widehat{D_i(f)}(\varnothing) = \widehat{f}(i)$, as desired. ■

Under some symmetry assumptions, we can prove even more.

Lemma 2.6. Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a monotone and transitive function. Then $\mathrm{Inf}_i(f) \leqslant \frac{1}{\sqrt{n}}$ for every $i \in [n]$.

Proof. Since f is transitive, $\widehat{f}(i) = \widehat{f}(j)$ for all $i,j \in [n]$. [1] Now, applying Parseval's theorem yields:

$$1 = \sum_{S\subseteq[n]} \widehat{f}(S)^2 \geqslant \sum_{i=1}^{n} \widehat{f}(i)^2 = n\widehat{f}(1)^2$$

as desired. ■

[1] Indeed, $\widehat{f}(i) = \langle f, \chi_i\rangle = \mathbb{E}_x[f(x)x_i] = \mathbb{E}_x[f(\sigma(x))x_{\sigma(i)}]$ for any permutation σ. Now suppose σ flips i,j, and keeps everything else fixed. Since f is transitive, $f(\sigma(x)) = f(x)$. Thus $\mathbb{E}_x[f(x)x_i] = \mathbb{E}_x[f(x)x_j] = \widehat{f}(j)$, as desired.

We now define the total influence of a function.

Definition 2.4 (Total Influence)**.** Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a function. We define the *total influence* of f to be:

$$\mathbb{I}[f] := \sum_{i=1}^{n} \mathrm{Inf}_i(f)$$

Remark. Some remarks are as follows:

1. Note that $\mathbb{I}[\chi_{[n]}] = n$, $\mathbb{I}[\chi_\varnothing] = \mathbb{I}[1] = 0$. It's easy to see that these are extremal among Boolean-valued functions, i.e. if $f : \{-1,1\}^n \mapsto \{-1,1\}$ is some function, then $0 \leqslant \mathbb{I}[f] \leqslant n$.

2. If $f : \{-1,1\}^n \mapsto \{-1,1\}$ is a monotone function, then $\mathbb{I}[f] = \sum_{i=1}^{n} \widehat{f}(i)$.

3. $\mathbb{I}[\mathrm{Maj}_n] = \Theta(\sqrt{n})$.

4. Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be any function. Note that

$$\mathbb{I}[f] = \sum_{i=1}^{n} \mathrm{Inf}_i(f) \overset{\text{Corollary 2.4}}{=} \sum_{i=1}^{n} \sum_{i \in S} \widehat{f}(S)^2 = \sum_{S \subseteq [n]} |S| \cdot \widehat{f}(S)^2$$

Furthermore, $\mathbb{I}[f] = \sum_{S \subseteq [n]} |S| \cdot \widehat{f}(S)^2 \geqslant \sum_{S \neq \varnothing} \widehat{f}(S)^2 \overset{\text{Proposition 3}}{=} \mathrm{Var}(f)$. We have thus proven what is known as *Poincaré's inequality*, and we state our conclusions below.

Theorem 2.7 (Fourier Characterization of Total Influence)**.** For any function $f : \{-1,1\}^n \mapsto \mathbb{R}$,

$$\mathbb{I}[f] = \sum_{S \subseteq [n]} |S| \cdot \widehat{f}(S)^2 = \sum_{k=0}^{n} k \cdot W^k[f]$$

Theorem 2.8 (Poincaré's Inequality)**.** For any function $f : \{-1,1\}^n \mapsto \mathbb{R}$, $\mathbb{I}[f] \geqslant \mathrm{Var}(f)$. Equality holds iff $\widehat{f}(S) = 0$ for all $|S| \geqslant 2$, i.e. $f = \widehat{f}(\varnothing) + \sum_{i=1}^{n} \widehat{f}(i) x_i$, i.e. $f - \widehat{f}(\varnothing)$ is linear.

Remark. Recall from Lemma 1.8 that if $f : \{-1,1\}^n \mapsto \{-1,1\}$ is a function, then $\mathrm{Var}(f) = 4\Pr(f(x) = 1) \cdot \Pr(f(x) \neq 1)$. Now, WLOG assume $\mathbb{E}[f] \geqslant 0$, and write $\mathcal{S} := f^{-1}(1)$.
Note that f can be viewed as a 2-coloring of the Boolean hypercube, with vertices in $\mathcal{S}$ being "colored" 1, and the other vertices being colored -1. Also, we call an edge (of the Boolean hypercube) "i-directed" if the endpoints of that edge differ in their i^{th} coordinate. Then observe that $\mathrm{Inf}_i(f) = \Pr(f(x) \neq f(x \oplus i))$ is the fraction of i-directed edges whose endpoints have different colors. Along similar lines, $\frac{1}{n}\mathbb{I}[f] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Inf}_i(f)$ is the fraction of edges whose endpoints have different colors. In other words, $\frac{1}{n}\mathbb{I}[f]$ is a (normalized) measure of the surface area of $\mathcal{S}$ since the edges with differently colored endpoints are precisely the edges emanating from $\mathcal{S}$.
Thus an inequality between $\mathbb{I}[f]$, which measures the surface area of $\mathcal{S}$, and $\mathrm{Var}(f)$, which is linked to $|\mathcal{S}| = \mathrm{vol}(S)$, is an **isoperimetric inequality** associated to the Boolean function f. This is why Poincaré's inequality is sometimes also referred to as an isoperimetric inequality. We shall explore this connection in greater detail later.

We finish our discussion with a nice social-scientific perspective on total influence.

Lemma 2.9. Let $f : \{-1, 1\}^n \mapsto \{-1, 1\}$ be a function. Then

$$\mathbb{E}_x \left[\text{Number of voters who agreed with the outcome}\right] = \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \widehat{f}(i)$$

Proof. Note that

$$\mathbb{E}_x \left[\text{Number of voters who agreed with the outcome}\right] = \mathbb{E}_x \left[\sum_{i=1}^{n} \left(\frac{1}{2} + \frac{1}{2}x_i f(x)\right)\right]$$

$$= \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_x \left[x_i f(x)\right] = \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \langle \chi_i, f\rangle = \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \widehat{f}(i)$$

■

Remark. In social choice theory, one of the objectives while designing a voting mechanism is to ensure that as many people agree with the outcome as possible, in expectation. Thus, achieving this goal is equivalent to maximizing $\sum_{i=1}^{n} \widehat{f}(i)$. Furthermore, if f is monotone (as most voting schemes are), then maximizing $\sum_{i=1}^{n} \widehat{f}(i)$ is equivalent to maximizing $\mathbb{I}[f]$.

We shall now prove that among *all* Boolean valued functions, $\sum_{i=1}^{n} \widehat{f}(i)$ is maximized by Maj_n. Thus, according to the social objective desired above, Maj_n is the "optimal" voting mechanism.

Theorem 2.10. Among all $f : \{-1, 1\}^n \mapsto \{-1, 1\}$, $\sum_{i=1}^{n} \widehat{f}(i)$ is maximized by Maj_n.

Proof. Note that

$$\sum_{i=1}^{n} \widehat{f}(i) = \sum_{i=1}^{n} \mathbb{E}_x \left[x_i f(x)\right] = \mathbb{E}_x \left[f(x) \cdot (x_1 + \cdots + x_n)\right]$$

Clearly this expression is maximized when $f(x) = \mathrm{sign}(x_1 + \cdots + x_n) = \mathrm{Maj}_n(x)$, as desired. ■

Corollary 2.11. For all monotone $f : \{-1, 1\}^n \mapsto \{-1, 1\}$, $\mathbb{I}[f] \leqslant \mathbb{I}[\mathrm{Maj}_n] = \Theta(\sqrt{n})$.

§3. Noise Stability and Arrow's Theorem

We will investigate how stable a Boolean function is to perturbations in its input. To do that, we first define a model of perturbation.

Definition 3.1 (ρ-perturbation)**.** Let $x \in \{-1,1\}^n$, and let $\rho \in [-1,1]$ be a parameter. Construct a random string $y \in \{-1,1\}^n$, as follows:

$$y_i = \begin{cases} x_i & \text{with probability } \frac{1+\rho}{2} \\ -x_i & \text{with probability } \frac{1-\rho}{2} \end{cases}$$

We write $y \sim N_\rho(x)$ to denote that y was generated through the above process.

Remark. Some remarks are as follows:

1. y should be viewed as a "noisy" version of x.
2. If $\rho = 1$ (resp. -1), then y is always equal to x (resp. $-x$). If $\rho = 0$, y is a uniformly random string in $\{-1,1\}^n$.
3. We call (x,y) a ρ-correlated random pair if x is uniformly random in $\{-1,1\}^n$, and $y \sim N_\rho(x)$. If (x,y) is a ρ-correlated random pair, then $\mathbb{E}[x_i] = \mathbb{E}[y_i] = 0$, but $\mathbb{E}[x_i y_i] = \rho$.
4. Let x be arbitrary (x is **not** random), and let $y \sim N_\rho(x)$. Then y_i and y_j are independent, for every $i \neq j$.

We can now define the main notion of interest, namely *noise stability*:

Definition 3.2. For $f \in \mathbb{R}^{\{-1,1\}^n}$, $\rho \in [-1,1]$, the *noise stability* of f, denoted $\mathrm{Stab}_\rho[f]$, is defined as:

$$\mathrm{Stab}_\rho[f] := \mathbb{E}_{(x,y)\ \rho\text{-correlated random pair}}\left[f(x)f(y)\right]$$

Remark. Note that $\mathrm{Stab}_\rho[\chi_S] = \mathbb{E}_{(x,y)\ \rho\text{-correlated random pair}}\left[\chi_S(x)\chi_S(y)\right] = \mathbb{E}_{(x,y)\ \rho\text{-correlated random pair}}\left[\prod_{i\in S} x_i y_i\right] = \prod_{i \in S}\mathbb{E}[x_i y_i] = \rho^{|S|}$, i.e. $\mathrm{Stab}_\rho[\chi_S] = \rho^{|S|}$.

Note that for any $f : \{-1,1\}^n \mapsto \{-1,1\}$, and any distribution $\mathcal{D}$ on $\{-1,1\}^n \times \{-1,1\}^n$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(x)f(y)\right] = \Pr(f(x) = f(y)) - \Pr(f(x) \neq f(y)) = 1 - 2\Pr(f(x) \neq f(y))$$

We thus define the *noise sensitivity* of f as follows:

Definition 3.3. For $f : \{-1,1\}^n \mapsto \{-1,1\}$, and a parameter $\delta \in (0,1]$, we define the *noise sensitivity* of f to be:

$$\mathrm{NS}_\delta[f] := \Pr_{(x,y)\ (1-2\delta)\text{-correlated random pair}}(f(x) \neq f(y))$$

In other words, if every bit of x is flipped with probability δ, then noise sensitivity measures the probability that the output changes.

Remark. Note that $\mathrm{NS}_\delta[f] = \frac{1}{2}\left(1 - \mathrm{Stab}_{1-2\delta}[f]\right)$.

We can prove, using the Central Limit Theorem, that,

$$\mathrm{Stab}_\rho[\mathrm{Maj}_n] \overset{n\to\infty}{\longrightarrow} \frac{2}{\pi}\arcsin(\rho) \implies \mathrm{NS}_\delta[\mathrm{Maj}_n] \overset{n\to\infty}{\longrightarrow} \frac{2}{\pi}\sqrt{\delta} + \mathcal{O}(\delta^{3/2})$$

We now make the process of adding noise an "operator".

Definition 3.4 (Noise Operator)**.** Given a parameter $\rho \in [-1, 1]$, we have an operator $T_\rho : \mathbb{R}^{\{-1,1\}^n} \mapsto \mathbb{R}^{\{-1,1\}^n}$, such that

$$\big(T_\rho(f)\big)(x) := \mathbb{E}_{y\sim N_\rho(x)}\left[f(y)\right]$$

Proposition 6. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, where $f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S$,

$$T_\rho(f) = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}(S)\chi_S$$

Proof. Note that T_ρ is linear, so it suffices to show that $T_\rho(\chi_S) = \rho^{|S|}\chi_S$. Now,

$$T_\rho(\chi_S) = \mathbb{E}_{y\sim N_\rho(x)}\left[\chi_S(y)\right] = \mathbb{E}_{y\sim N_\rho(x)}\left[\prod_{i\in S} y_i\right] = \prod_{i\in S}\mathbb{E}\left[y_i\right] = \rho^{|S|}$$

■

Remark. A few remarks are in order:

1. Note that T_ρ is a "smoothening" operator: It replaces $f(x)$ by some weighted average of $f(y)$, where the y's "closer" to x are given more weightage.
2. Another way to look at the smoothening effect of T_ρ is to notice that the "high-frequency" components, namely $\widehat{f}(S)x_S$, where $|S|$ is large, are dampened more.
3. Note that T_ρ is a self-adjoint operator (i.e. the matrix representing T_ρ in the $\{\chi_S\}_{S\subseteq[n]}$ basis is symmetric), i.e. $\langle T_\rho f, g\rangle = \langle f, T_\rho g\rangle$ for any f, g.
4. By Parseval's theorem, it is easy to see that $\|T_\rho f\|_2 \leqslant \|f\|_2$. T_ρ is thus a *contractive* map. Later on, we will prove a vast generalization of this fact, namely $\|T_\rho f\|_q \leqslant \|f\|_p$ for all $\rho \in \left[0, \sqrt{\frac{p-1}{q-1}}\right]$, i.e. provided ρ is small enough, not only is T_ρ is contractive in the $\|\cdot\|_p$ norm, but it is also contractive in the $\|\cdot\|_q$ norm (recall that $\|f\|_a \geqslant \|f\|_b$ if $a > b$ for any $f \in \mathbb{R}^{\{-1,1\}^n}$), i.e. it is "hyper"-contractive.

We shall now establish a connection between noise stability and the noise operator.

Lemma 3.1. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, $\mathrm{Stab}_\rho[f] = \langle f, T_\rho(f)\rangle$.

Proof. Note that

$$\mathrm{Stab}_\rho[f] = \mathbb{E}_{(x,y)\ \rho\text{-correlated random pair}}\left[f(x)f(y)\right] = \mathbb{E}_{x\sim\{-1,1\}^n}\left[f(x)\cdot\mathbb{E}_{y\sim N_\rho(x)}\left[f(y)\right]\right]$$
$$= \mathbb{E}_x\left[f(x)\cdot(T_\rho f)(x)\right] = \langle f, T_\rho(f)\rangle$$

■

Corollary 3.2. For any $f \in \mathbb{R}^{\{-1,1\}^n}$, $\text{Stab}_\rho[f] = \sum_{S \subseteq [n]} \rho^{|S|} \widehat{f}(S)^2 = \sum_{k=0}^{n} \rho^k W^k[f]$.

Proof. Follows by applying Corollary 1.6 on Lemma 3.1. ∎

We would now like to characterize the most stable functions in terms of noise stability.

Theorem 3.3. Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be an unbiased Boolean function, i.e. $\mathbb{E}[f] = 0$. Then $\text{Stab}_\rho[f] \leqslant \text{Stab}_\rho[\chi_i] = \rho$ for any $\rho \in [0,1], i \in [n]$, i.e. *dictators have the maximum noise stability among all unbiased Boolean-valued functions*.

Proof. Note that

$$\text{Stab}_\rho[f] = \sum_{k=0}^{n} \rho^k W^k[f]$$

Since $\mathbb{E}[f] = \widehat{f}(\varnothing) = 0$, $\sum_{k=0}^{n} \rho^k W^k[f] = \sum_{k=1}^{n} \rho^k W^k[f]$, and furthermore,

$$\sum_{k=1}^{n} \rho^k W^k[f] \leqslant \rho \underbrace{\sum_{k=1}^{n} W^k[f]}_{=1 \text{ by Corollary 1.7}} = \rho$$

Since $\text{Stab}_\rho[\chi_i] = \rho$ for any $i \in [n]$, we're done. ∎

Remark. Note that the noise stability of anti-dictators, i.e. $-\chi_i$, is also ρ. Thus a more inclusive phrasing of the above theorem would say 1-juntas instead of dictators.

3.1. Arrow's Theorem

We shall now present Kalai's proof [Kal02] of Arrow's theorem, one of the crown jewels of modern social choice theory, using Boolean Function Analysis.

We first lay down some basic definitions: Suppose we have candidates a, b, c standing in an election, and suppose n voters give their (strict) preference orders over a, b, c.

The *Condorcet winner* of this election, is decided as follows: Fix some voting scheme $f : \{-1,1\}^n \mapsto \{-1,1\}$. Fix a pair, say $\{a, c\}$, and decide a winner among a, c by restricting all preference orders to $\{a, c\}$, and then applying f. Suppose the winner is a, so we say $c \prec a$. Do the same for the pairs $\{a, b\}$, and $\{b, c\}$.

To give an example, suppose we have 3 candidates, a, b, c, and $n = 5$ voters, with preference orders $a \prec c \prec b, b \prec c \prec a, b \prec a \prec c, c \prec a \prec b, c \prec b \prec a$. Further, let our voting rule f simply be Maj_5. Then, restricted to the pair $\{a, c\}$, we have the preferences $a \prec c, c \prec a, a \prec c, c \prec a, c \prec a$. Thus, if we denote $a \prec c$ by -1, and $c \prec a$ by 1, we have the input $(-1, 1, -1, 1, 1)$, and $\text{Maj}_5(-1, 1, -1, 1, 1) = 1 = c \prec a$, i.e. *a defeats c in a pairwise election*. Similarly, we note that a defeats b and b defeats c in pairwise elections too, and thus we have the order $c \prec b \prec a$, i.e. the Condorcet winner of this particular instance of preference orders is a.

However, if we get a cycle in our order, then we are doomed, since a Condorcet winner can't be consistently defined. Indeed, for the candidates a, b, c, the preference orders $a \prec b \prec c, b \prec c \prec a, c \prec a \prec b$, and the voting rule $f = \text{Maj}_3$, pairwise elections between $\{a, b\}$ yields $a \prec b$, pairwise elections between $\{b, c\}$ yields $b \prec c$, and pairwise elections between $\{c, a\}$ yields $c \prec a$. Clearly the pairwise results $a \prec b, b \prec c, c \prec a$ can't be consistently extended to a total order, and thus we can't define a Condorcet winner for this particular instance of preference orders.

If a particular preference list induces a cyclic Condorcet order, we call that preference list *irrational*. Naturally, we want voting rules f which don't have any irrational preference list, and thus a Condorcet winner can always be found. Arrow's theorem says that the only such functions are the dictators.

Theorem 3.4 (Arrow's theorem ([Arr50])). Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a unanimous function which doesn't have any irrational preference list. Then f must be a dictator.

Proof. Suppose we have n voters, and each voter picks one of the 6 preference lists on a, b, c, uniformly (and independently of others). Let $x, y, z \in \{-1,1\}^n$ be 3 strings such that $x_i = 1$ if and only if the i^{th} voter prefers a over b, and similarly, $y_i = 1$ if and only if the i^{th} voter prefers b over c, and $z_i = 1$ if and only if the i^{th} voter prefers c over a. This preference list is irrational if and only if $f(x) = f(y) = f(z)$. Now, define the "Not-All-Equals" function $\text{NAE} : \{-1,1\}^3 \mapsto \{0,1\}$, such that $\text{NAE}(-1,-1,-1) = \text{NAE}(1,1,1) = 0$, and NAE is 1 on all other inputs. It is easy to verify that

$$\text{NAE}(w_1, w_2, w_3) = \frac{3}{4} - \frac{1}{4}w_1w_2 - \frac{1}{4}w_2w_3 - \frac{1}{4}w_3w_1$$

Thus,

$$\Pr(f \text{ is rational}) = \mathbb{E}_{x,y,z}\left[\frac{3}{4} - \frac{1}{4}f(x)f(y) - \frac{1}{4}f(y)f(z) - \frac{1}{4}f(z)f(x)\right] = \frac{3}{4} - \frac{3}{4}\mathbb{E}_{x,y}\left[f(x)f(y)\right]$$

Now, note that $\Pr(x_i = 1) = \Pr(y_i = 1) = \frac{1}{2}$, since the preference orders were chosen (uniformly) randomly. Also, note that $x_i = y_i$ if and only if $c \prec b \prec a$ or $a \prec b \prec c$. Thus $\Pr(x_i = y_i) = \frac{1}{3}$, and thus (x_i, y_i) are $(-1/3)$-correlated. [2] Consequently, $\mathbb{E}_{x,y}\left[f(x)f(y)\right] = \text{Stab}_{-1/3}[f]$, and thus

$$\Pr(f \text{ is rational}) = \frac{3}{4} - \frac{3}{4}\text{Stab}_{-1/3}[f] = \frac{3}{4} - \frac{3}{4}\left(1 \cdot W^0[f] - \frac{1}{3} \cdot W^1[f] + \frac{1}{9} \cdot W^2[f] - \cdots\right)$$
$$\leqslant \frac{3}{4} - \frac{3}{4}\left(-\frac{1}{3}\right)\left(W^0[f] + W^1[f] + W^2[f] + \cdots\right) = 1$$

where the last equality follows by Corollary 1.7.
Thus, note that $\Pr(f \text{ is rational}) = 1$ if and only if $W^1[f] = 1$, and $W^0[f] = W^2[f] = W^3[f] = \cdots = 0$. Finally, observe that if $f : \{-1,1\}^n \mapsto \{-1,1\}$ is a function such that $W^1[f] = 1$, then f is either a dictator or an anti-dictator: Indeed, $W^1[f] = 1$ implies that $f = \sum_{i=1}^n a_i \chi_i$. Now, note that $a_1 = \frac{f(1,1,\ldots,1) + f(1,-1,\ldots,-1)}{2}$, and thus $a_1 \in \{-1,0,1\}$ (and similarly, $a_i \in \{-1,0,1\}$ for every $i \in [n]$). Furthermore if $|a_i| = 1$ for some $i \in [n]$, then $a_j = 0$ for every $j \neq i$, since $\sum_{t=1}^n a_t^2 = 1$.
Finally, the unanimity condition forces that f can't be an anti-dictator, and thus, f must be a dictator. ∎

Remark. Fourier analysis also yields a "robust" version of Arrow's theorem, i.e. if $\Pr(f \text{ is rational}) = 1 - \varepsilon$, then f is $\mathcal{O}(\varepsilon)$-close to a dictator/anti-dictator. This robust version was proved by Friedgut, Kalai and Naor in 2003.

[2] note that x_i, y_i are negatively correlated since had they been independent, $\Pr(x_i = y_i) = 1/2$, and $1/3 < 1/2$

§4. The Hypercontractivity Theorem

Recall that the T_ρ operator is contractive. In this chapter, we will initiate our journey towards the so-called *hypercontractive* theorem, which says that if ρ is small enough, then $\|T_\rho f\|_q \leqslant \|f\|_p$ even if $q > p$, i.e. the contractivity of T_ρ is so strong that even increasing the norm doesn't kill off the contraction.
We begin with a simpler lemma, called *Bonami's lemma*:

Lemma 4.1 (Bonami's Lemma)**.** Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a degree-k function, i.e. $\widehat{f}(S) = 0$ if $|S| > k$. Then $\mathbb{E}\left[f^4\right] \leqslant 9^k\, \mathbb{E}\left[f^2\right]^2$.

Proof. The statement is trivial for $k = 0$. Thus assume $k \geqslant 1$. We will prove the statement by double induction on n, k. Once again, the base case of $n = 0$ is trivial.
Now, note that

$$f = x_n \cdot \text{poly}_1(x_1, \ldots, x_{n-1}) + \text{poly}_2(x_1, \ldots, x_{n-1}) = x_n \cdot d + e$$

where d, e are polynomials in $x_1, \ldots, x_{n-1}$. Note that $\deg(d) \leqslant k-1$, since $\deg(f) \geqslant 1 + \deg(d)$. Now,

$$f^4 = x_n^4 d^4 + 4x_n^3 d^3 e + 6x_n^2 d^2 e^2 + 4x_n d e^3 + e^4$$

Since d, e don't involve x_n (and thus are independent of it), $\mathbb{E}\left[x_n^{j_1} d^{j_2} e^{j_3}\right] = \mathbb{E}\left[x_n^{j_1}\right] \cdot \mathbb{E}\left[d^{j_2} e^{j_3}\right]$ for any $j_1, j_2, j_3 \geqslant 0$. Furthermore, $\mathbb{E}\left[x_n^{2j+1}\right] = 0$, and $\mathbb{E}\left[x_n^{2j}\right] = 1$ for every $j \geqslant 0$, since x_n equals ± 1 with equal probability. Thus

$$\mathbb{E}\left[f^4\right] = \mathbb{E}\left[d^4\right] + 6\,\mathbb{E}\left[d^2 e^2\right] + \mathbb{E}\left[e^4\right]$$

Similarly,

$$\mathbb{E}\left[f^2\right] = \mathbb{E}\left[d^2\right] + \mathbb{E}\left[e^2\right] \implies \mathbb{E}\left[f^2\right]^2 = \mathbb{E}\left[d^2\right]^2 + 2\,\mathbb{E}\left[d^2\right]\mathbb{E}\left[e^2\right] + \mathbb{E}\left[e^2\right]^2$$

Since e has only $n-1$ variables, $9^k\,\mathbb{E}\left[e^2\right]^2 \geqslant \mathbb{E}\left[e^4\right]$. Similarly, since d has only $n-1$ variables, and $\deg(d) \leqslant k-1$, $9^{k-1}\,\mathbb{E}\left[d^2\right]^2 \geqslant \mathbb{E}\left[d^4\right]$. Finally, by Cauchy-Schwartz, $\mathbb{E}\left[d^2 e^2\right] \leqslant \sqrt{\mathbb{E}\left[d^4\right]} \cdot \sqrt{\mathbb{E}\left[e^4\right]}$. But $\sqrt{\mathbb{E}\left[d^4\right]} \cdot \sqrt{\mathbb{E}\left[e^4\right]} \leqslant \frac{9^k}{3}\,\mathbb{E}\left[e^2\right]\mathbb{E}\left[d^2\right]$, and we thus have:

$$\mathbb{E}\left[f^2\right]^2 \geqslant \frac{1}{9^{k-1}}\,\mathbb{E}\left[d^4\right] + \frac{1}{9^k}\,\mathbb{E}\left[e^4\right] + \frac{6}{9^k}\,\mathbb{E}\left[d^2 e^2\right] \geqslant \frac{1}{9^k}\left(\mathbb{E}\left[d^4\right] + \mathbb{E}\left[e^4\right] + 6\,\mathbb{E}\left[d^2 e^2\right]\right)$$

as desired. ■

Remark. A few remarks are in order:

1. Note that $(D_n f)(x) = (f(x^{n\mapsto 1}) + f(x^{n\mapsto -1}))/2 = d(x_1, \ldots, x_{n-1})$, i.e. d is the derivative of f w.r.t x_n. Furthermore, $\mathbb{E}_{x_n}[f] = e(x_1, \ldots, x_{n-1}) = f(x_1, \ldots, x_{n-1}, 0)$. Thus, $f = f(x_1, \ldots, x_{n-1}, 0) + (D_n f)(x_1, \ldots, x_{n-1}, 0)$ is basically the "Taylor expansion" of f about $x_n = 0$! Furthermore, since every function over $\{-1,1\}^n$ is a multilinear polynomial, the said Taylor expansion doesn't have any order 2 terms, since the double derivative of any function $f : \{-1,1\}^n \mapsto \mathbb{R}$ w.r.t any x_i is 0.

2. To illustrate the next point, consider a random variable X, such that $\Pr(X = 0) = 1 - 2^{-n} = 1 - \Pr(X = 1)$. Then $\mathbb{E}\left[X^2\right] = \mathbb{E}\left[X^4\right] = 2^{-n}$, and thus $\mathbb{E}\left[X^2\right]^2 / \mathbb{E}\left[X^4\right] = 2^{-n}$. Now, some reflection on why $\mathbb{E}\left[X^2\right]^2 / \mathbb{E}\left[X^4\right]$ is so small reveals that the behavior of X is not really random: Indeed, it assumes the value 0 with overwhelming probability and is thus "almost" deterministic. Thus, $\mathbb{E}\left[X^2\right]^2 / \mathbb{E}\left[X^4\right]$ is a good proxy for how random the variable is, or how close to uniform it is. Consequently, Bonami's lemma, and by extension the Hypercontractivity theorem, by providing a lower bound for $\mathbb{E}\left[X^2\right]^2 / \mathbb{E}\left[X^4\right]$, can also be viewed as *anti-concentration* results.

Corollary 4.2. For any $f : \{-1,1\}^n \mapsto \mathbb{R}$ with $\deg(f) \leqslant k$, $\|f\|_2 \leqslant \|f\|_4 \leqslant \sqrt{3}^k \|f\|_2$, where $\|f\|_p := \mathbb{E}\left[f^p\right]^{1/p}$ for any $p \geqslant 1$.

Corollary 4.3. For any $f : \{-1,1\}^n \mapsto \mathbb{R}$, $\|T_{\frac{1}{\sqrt{3}}} f^{=k}\|_4 \leqslant \|f^{=k}\|_2$, where $f^{=k} := \sum_{|S|=k} \widehat{f}(S)\chi_S$.

Proof. Recall that $T_{\frac{1}{\sqrt{3}}} f^{=k} = \sum_{|S|=k} \left(\frac{1}{\sqrt{3}}\right)^k \widehat{f}(S)\chi_S$, from which the result follows. ∎

Remark. Note that $\|T_{\frac{1}{\sqrt{3}}} f\|_2^2 = \langle T_{\frac{1}{\sqrt{3}}} f, T_{\frac{1}{\sqrt{3}}} f\rangle = \sum_{S \subseteq [n]} \frac{1}{3^k} \widehat{f}(S)^2 = \mathrm{Stab}_{\frac{1}{3}}[f]$. Thus, $\|T_{\frac{1}{\sqrt{3}}} f\|_2$ has a combinatorial meaning too.

Although the inequality as stated works for only homogenous polynomials, we can obtain a "free" upgrade.

Theorem 4.4 ((2,4)-hypercontractivity theorem)**.** For any $f : \{-1,1\}^n \mapsto \mathbb{R}$, $\|T_{\frac{1}{\sqrt{3}}} f\|_4 \leqslant \|f\|_2$.

Proof. We mimic the proof of Lemma 4.1. We have $f = x_n d + e \implies Tf = x_n \cdot \frac{1}{\sqrt{3}} Td + Te$, where $T = T_{\frac{1}{\sqrt{3}}}$. Thus

$$\mathbb{E}\left[(Tf)^4\right] = \left(\frac{1}{\sqrt{3}}\right)^4 \mathbb{E}\left[(Td)^4\right] + \left(\frac{1}{\sqrt{3}}\right)^2 \cdot 6\,\mathbb{E}\left[(Td)^2(Te)^2\right] + \mathbb{E}\left[(Te)^4\right] \leqslant \mathbb{E}\left[(Td)^4\right] + 2\,\mathbb{E}\left[(Td)^2(Te)^2\right] + \mathbb{E}\left[(Te)^4\right]$$

$$\overset{\text{Cauchy-Schwartz}}{\leqslant} \mathbb{E}\left[(Td)^4\right] + 2\sqrt{\mathbb{E}\left[(Td)^4\right]}\sqrt{\mathbb{E}\left[(Te)^4\right]} + \mathbb{E}\left[(Te)^4\right] = \left(\sqrt{\mathbb{E}\left[(Td)^4\right]} + \sqrt{\mathbb{E}\left[(Te)^4\right]}\right)^2$$

$$\overset{\text{induction hypothesis}}{\leqslant} \left(\mathbb{E}\left[d^2\right] + \mathbb{E}\left[e^2\right]\right)^2 = \mathbb{E}\left[f^2\right]^2$$

as desired. ∎

We can easily prove another hypercontractivity result from the above, using Hölder's inequality.

Theorem 4.5 ((4/3,2)-hypercontractivity theorem)**.** For any $f : \{-1,1\}^n \mapsto \mathbb{R}$, $\|T_{\frac{1}{\sqrt{3}}} f\|_2 \leqslant \|f\|_{\frac{4}{3}}$.

Proof. Observe that

$$\|T_{\frac{1}{\sqrt{3}}} f\|_2^2 = \langle f, T_{\frac{1}{\sqrt{3}}} T_{\frac{1}{\sqrt{3}}} f\rangle \overset{\text{Hölder}}{\leqslant} \|f\|_{\frac{4}{3}} \cdot \|T_{\frac{1}{\sqrt{3}}} T_{\frac{1}{\sqrt{3}}} f\|_4 \overset{\text{Theorem 4.4}}{\leqslant} \|f\|_{\frac{4}{3}} \cdot \|T_{\frac{1}{\sqrt{3}}} f\|_2$$

as desired. ∎

Remark. Since $\|T_{\frac{1}{\sqrt{3}}} f\|_2^2 = \mathrm{Stab}_{1/3}[f]$, we have $\mathrm{Stab}_{1/3}[f] \leqslant \|f\|_{4/3}^2$.

To further emphasize the "anti-concentration" nature of Bonami's lemma, we frame it in a corollary.

Corollary 4.6. Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a non-constant function of degree $\leqslant k$. Write $\mu := \mathbb{E}[f], \sigma := \sqrt{\mathrm{Var}(f)}$. Then

$$\Pr_{x\sim\{-1,1\}^n}\left(|f(x)-\mu| \geqslant \frac{\sigma}{2}\right) > \frac{9^{1-k}}{16}$$

Proof. Write $g := (f-\mu)/\sigma$. Then $\|g\|_2 = 1$, and thus by Lemma 4.1, $\mathbb{E}\left[g^4\right] \leqslant 9^k$. Then

$$\Pr\left(|f-\mu| \geqslant \frac{\sigma}{2}\right) = \Pr\left(|g| \geqslant \frac{1}{2}\right) = \Pr\left(|g| \geqslant \frac{1}{2}\|g\|_2\right) = \Pr\left(g^2 \geqslant \frac{1}{4}\mathbb{E}\left[g^2\right]\right)$$

We now invoke the Paley-Zygmund inequality, to get that

$$\Pr\left(g^2 \geqslant \frac{1}{4}\mathbb{E}\left[g^2\right]\right) > \frac{9}{16}\frac{\mathbb{E}\left[g^2\right]^2}{\mathbb{E}\left[g^4\right]} \geqslant \frac{9^{1-k}}{16}$$

as desired. ■

Remark. For a proof of the Paley-Zygmund inequality, note that $Z = \mathbb{E}\left[Z\cdot\mathbb{1}_{Z<\theta\,\mathbb{E}[Z]}\right] + \mathbb{E}\left[Z\cdot\mathbb{1}_{Z\geqslant\theta\,\mathbb{E}[Z]}\right] < \theta\,\mathbb{E}[Z] + \mathbb{E}\left[Z\cdot\mathbb{1}_{Z\geqslant\theta\,\mathbb{E}[Z]}\right]$, and now $\mathbb{E}\left[Z\cdot\mathbb{1}_{Z\geqslant\theta\,\mathbb{E}[Z]}\right] \leqslant \sqrt{\mathbb{E}[Z^2]}\cdot\sqrt{\mathbb{E}\left[\mathbb{1}_{Z\geqslant\theta\,\mathbb{E}[Z]}\right]} = \Pr(Z\geqslant\theta\,\mathbb{E}[Z])\cdot\sqrt{\mathbb{E}[Z^2]}$, i.e.

$$Z < \theta\,\mathbb{E}[Z] + \Pr(Z\geqslant\theta\,\mathbb{E}[Z])\cdot\sqrt{\mathbb{E}[Z^2]} \implies (1-\theta)\,\mathbb{E}[Z] < \Pr(Z\geqslant\theta\,\mathbb{E}[Z])\cdot\sqrt{\mathbb{E}[Z^2]}$$

and we get that $\Pr(Z\geqslant\theta\,\mathbb{E}[Z]) > (1-\theta)^2\,\mathbb{E}[Z]^2/\mathbb{E}\left[Z^2\right]$.

Even though Bonami's lemma is a very special case of the Hypercontractivity theorem, it is already sufficient to prove the Friedgut-Kalai-Naor theorem.

Theorem 4.7 (Friedgut-Kalai-Naor Theorem)**.** Suppose we have $f : \{-1,1\}^n \mapsto \{-1,1\}$ such that $W^1[f] := \sum_{i=1}^n \widehat{f}(i)^2 = 1-\delta$. Then f is $\mathcal{O}(\delta)$-close to some $\pm\chi_i$ for some $i\in[n]$.

Proof. Let $\ell = f^{=1} = \sum_{i=1}^n \widehat{f}(i)\chi_i$, and thus $\mathbb{E}\left[\ell^2\right] = W^1[f] = 1-\delta$. Towards our goal, we first show that $\mathrm{Var}(\ell^2) \leqslant 6400\delta$ [3]. Indeed, by Corollary 4.6,

$$\Pr\left(|\ell^2-(1-\delta)| \geqslant \frac{1}{2}\sqrt{\mathrm{Var}(\ell^2)}\right) > \frac{9^{1-2}}{16} = \frac{1}{144}$$

Now, assume for the sake of contradiction that $\mathrm{Var}(\ell^2) > 6400\delta$, and also assume WLOG that $\delta \leqslant \frac{1}{1600}$. Then $\Pr\left(|\ell^2-(1-\delta)| > 40\sqrt{\delta}\right) \leqslant \Pr\left(|\ell^2-1| > 39\sqrt{\delta}\right)$. Now, since $|f| = 1$ and $\delta \leqslant 1/1600$, $|\ell^2-1| > 39\sqrt{\delta} \implies (\ell-f)^2 \geqslant 169\delta$, and thus $\mathbb{E}\left[(\ell-f)^2\right] \geqslant \frac{169\delta}{144} > \delta$, which is a contradiction since $\mathbb{E}\left[(\ell-f)^2\right] = 1-W^1[f] = \delta$.
Now,

$$\mathbb{E}\left[\ell^4\right] = \sum_{i=1}^n \widehat{f}(i)^4 + 6\sum_{1\leqslant i<j\leqslant n}\widehat{f}(i)^2\widehat{f}(j)^2$$

$$\mathbb{E}\left[\ell^2\right]^2 = \sum_{i=1}^n \widehat{f}(i)^4 + 2\sum_{1\leqslant i<j\leqslant n}\widehat{f}(i)^2\widehat{f}(j)^2$$

[3] note that if $f = \pm\chi_i$, then $\ell^2 \equiv 1$ and $\mathrm{Var}(\ell^2) = 0$

Thus

$$\frac{1}{2}\operatorname{Var}(\ell^2) = 2\sum_{1\leqslant i<j\leqslant n}\widehat{f}(i)^2\widehat{f}(j)^2 = \left(\sum_{i=1}^{n}\widehat{f}(i)^2\right)^2 - \sum_{i=1}^{n}\widehat{f}(i)^4 = (1-\delta)^2 - \sum_{i=1}^{n}\widehat{f}(i)^4 \geqslant (1-2\delta) - \sum_{i=1}^{n}\widehat{f}(i)^4$$

$$\implies (1-2\delta) - \sum_{i=1}^{n}\widehat{f}(i)^4 \leqslant 3200\delta \implies 1-3202\delta \leqslant \sum_{i=1}^{n}\widehat{f}(i)^4 \leqslant \left(\max_{i\in[n]}\widehat{f}(i)^2\right)\cdot\sum_{i=1}^{n}\widehat{f}(i)^2 \leqslant \max_{i\in[n]}\widehat{f}(i)^2 \leqslant \max_{i\in[n]}|\widehat{f}(i)|$$

as desired. ■

Remark. We immediately have a "robust" version of Arrow's theorem, i.e. if $\Pr(f \text{ is rational}) = 1-\delta$, then f is $\mathcal{O}(\delta)$-close to a dictator.

We will now begin our journey towards the KKL theorem, once again by deriving some corollaries of Lemma 4.1 [4].

Corollary 4.8. Let $A \subseteq \{-1,1\}^n$ have volume α, i.e. $|A| = \alpha\cdot 2^n$. Then $\operatorname{Stab}_{1/3}[\mathbb{1}_A] \leqslant \alpha^{3/2}$.

Proof. We know that $\operatorname{Stab}_{1/3}[f] \leqslant \|f\|_{4/3}^2$ for any $f:\{-1,1\}^n \mapsto \mathbb{R}$. Thus set $f = \mathbb{1}_A$, and observe that

$$\|\mathbb{1}_A\|_{4/3}^2 = \left(\mathbb{E}\left[\mathbb{1}_A^{4/3}\right]^{3/4}\right)^2 = \mathbb{E}\left[\mathbb{1}_A^{4/3}\right]^{3/2} = \mathbb{E}\left[\mathbb{1}_A\right]^{3/2} = \alpha^{3/2}$$

as desired. ■

Exactly similar to the above proof, we have the proof of the following result:

Proposition 7. If $g:\{-1,1\}^n \mapsto \{-1,0,1\}$ is a function such that $\alpha = \Pr(|g|=1)$, then $\operatorname{Stab}_{1/3}[g] \leqslant \alpha^{3/2}$.

We are now finally ready to prove the Kahn-Kalai-Linial theorem:

Theorem 4.9 (Kahn-Kalai-Linial Theorem ([KKL88]))**.** For any $f:\{-1,1\}^n \mapsto \{-1,1\}$,

$$\max_{i\in[n]}\operatorname{Inf}_i[f] \geqslant \Omega\left(\frac{\log n}{n}\right)\cdot\operatorname{Var}(f)$$

Proof. If $f:\{-1,1\}^n \mapsto \{-1,1\}$ is a function, and $i\in[n]$, then $g := D_i f$ has co-domain $\{-1,0,1\}$. Now, using the facts that $\operatorname{Stab}_{1/3}[h] = \|T_{1/\sqrt{3}}h\|_2^2 = \sum_{S\subseteq[n]}(1/3)^{|S|}\widehat{h}(S)^2$, and $D_i f = \sum_{i\in S}\widehat{f}(S)\chi_{S\setminus\{i\}}$ (see Proposition 5), we get that $\operatorname{Stab}_{1/3}[g] = \sum_{i\in S}(1/3)^{|S|-1}\widehat{f}(S)^2$.

At the same time, by the definition of influence, $\Pr(|g|=1) = \operatorname{Inf}_i[f]$. Thus

$$\sum_{i\in S}\left(\frac{1}{3}\right)^{|S|-1}\widehat{f}(S)^2 \leqslant \operatorname{Inf}_i[f]^{3/2} \tag{4.1}$$

[4] Lemma 4.1 captures a significant section of the power of the Hypercontractivity theorem, it is thus unsurprising that it has so many powerful consequences

Summing up Eq. (4.1) for all $i \in [n]$, we get:

$$\sum_{i=1}^{n} \sum_{i \in S} \left(\frac{1}{3}\right)^{|S|-1} \widehat{f}(S)^2 \leqslant \sum_{i=1}^{n} \mathrm{Inf}_i[f]^{3/2}$$

Now, set

$$M := \max_{i \in [n]} \mathrm{Inf}_i[f]$$

to obtain $\sum_{i=1}^n \mathrm{Inf}_i[f]^{3/2} \leqslant M^{1/2} \sum_{i=1}^n \mathrm{Inf}_i[f] = M^{1/2} \cdot \mathbb{I}[f]$. Now,

$$\sum_{i=1}^{n} \sum_{i \in S} \left(\frac{1}{3}\right)^{|S|-1} \widehat{f}(S)^2 = \sum_{S \neq \varnothing} |S| \cdot \left(\frac{1}{3}\right)^{|S|-1} \widehat{f}(S)^2 \geqslant 3 \sum_{S \neq \varnothing} \left(\frac{1}{3}\right)^{|S|} \widehat{f}(S)^2 = 3\,\mathrm{Stab}_{1/3}[f]$$

We now define the so-called "spectral sample" $\mathscr{S}$, which is a probability distribution supported on $2^{[n]} \setminus \{\varnothing\}$, where each S is sampled with probability $\widehat{f}(S)^2 / \mathrm{Var}(f)$. Then note that $\sum_{S \neq \varnothing} \left(\frac{1}{3}\right)^{|S|} \widehat{f}(S)^2 = \mathrm{Var}(f) \cdot \mathbb{E}_{\mathscr{S}}\left[3^{-|S|}\right]$. Applying Jensen's inequality (since $x \mapsto 3^{-x}$ is convex) yields $\mathbb{E}_{\mathscr{S}}\left[3^{-|S|}\right] \geqslant 3^{-\mathbb{E}_{\mathscr{S}}[|S|]}$. But

$$\mathbb{E}_{\mathscr{S}}\left[|S|\right] = \sum_{S \neq \varnothing} |S| \cdot \frac{\widehat{f}(S)^2}{\mathrm{Var}(f)} = \frac{\mathbb{I}[f]}{\mathrm{Var}(f)} =: \widetilde{\mathbb{I}}[f]$$

Thus

$$M^{1/2} \cdot \mathbb{I}[f] \geqslant 3\,\mathrm{Var}(f) \cdot 3^{-\widetilde{\mathbb{I}}[f]} \implies M^{1/2} \geqslant \frac{3^{1-\widetilde{\mathbb{I}}[f]}}{\widetilde{\mathbb{I}}[f]} \implies M \geqslant \frac{9^{1-\widetilde{\mathbb{I}}[f]}}{\widetilde{\mathbb{I}}[f]^2}$$

Now, if $\widetilde{\mathbb{I}}[f] \geqslant 0.3 \log_2 n$, then $\mathbb{I}[f] \geqslant 0.3 \cdot \mathrm{Var}(f) \log n \implies M \geqslant 0.3 \cdot \mathrm{Var}(f) \log n/n$. Otherwise, note that $9^{1-x}/x^2$ is a decreasing function of x, and thus

$$M \geqslant \frac{9^{1-0.3\log_2 n/n}}{0.09 \log_2^2 n/n^2} = \widetilde{\Omega}(n^{-0.3\log_2 9}) = \widetilde{\Omega}(n^{-0.951}) \gg \mathrm{Var}(f) \cdot \Omega\left(\frac{\log n}{n}\right)$$

as desired. ■

Remark. Recall that the unbiased Tribes function has $\mathrm{Var} \approx 1$, and maximum influence equal to $\Theta(\log n/n)$, thus showing that the KKL inequality is sharp.

4.1. General Hypercontractivity

The general hypercontractivity theorem is as stated below:

Theorem 4.10 (General Hypercontractivity ((p,q)-hypercontractivity))**.** Let $f \in \mathbb{R}^{\{-1,1\}^n}$ and assume $1 \leqslant p \leqslant q \leqslant \infty$. Then $\|T_\rho f\|_q \leqslant \|f\|_p$ for all $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$.

We shall prove a weaker version of general hypercontractivity, namely $(p,2)$-hypercontractivity. The strategy for proving the general hypercontractivity theorem is very similar to the strategy for proving $(p,2)$-hypercontractivity. We first show that $(p,2)$-hypercontractivity immediately implies $(2,q)$-hypercontractivity:

Proposition 8. Let $p \in [1,2]$ and $\rho \in [0,1]$ be such that $\|T_\rho f\|_2 \leqslant \|f\|_p$ for all $f \in \mathbb{R}^{\{-1,1\}^n}$. Then for q such that $\frac{1}{p} + \frac{1}{q} = 1$, and for all $f \in \mathbb{R}^{\{-1,1\}^n}$, we have $\|T_\rho f\|_q \leqslant \|f\|_2$.

Proof. We use the following fact: For any $v \in \mathbb{R}^{\{-1,1\}^n}$, we have

$$\|v\|_q = \sup_{\|g\|_p=1} \langle v, g \rangle$$

This fact simply follows from the sharpness of Hölder's inequality.
Since T_ρ is self adjoint, we have that $\langle T_\rho f, g\rangle = \langle f, T_\rho g\rangle$. Thus, for any f, $\langle T_\rho f, g\rangle = \langle f, T_\rho g\rangle \leqslant \|f\|_2 \cdot \|T_\rho g\|_2 \leqslant \|f\|_2 \cdot \|g\|_p$, and thus $\|T_\rho f\|_q \leqslant \|f\|_2$, as desired. ■

We now state the $(p, 2)$-hypercontractivity theorem:

Theorem 4.11 ($(p, 2)$-hypercontractivity)**.** For any $p \in [1, 2], \rho \in [0, \sqrt{p-1}]$, and $f \in \mathbb{R}^{\{-1,1\}^n}$, we have $\|T_\rho f\|_2 \leqslant \|f\|_p$.

We first prove this theorem for $n = 1$, where it is an exercise in one variable calculus:

Lemma 4.12. For any $p \in [1, 2], \rho \in [0, \sqrt{p-1}]$, and $f : \{-1, 1\} \mapsto \mathbb{R}$, we have $\|T_\rho f\|_2 \leqslant \|f\|_p$.

Proof. It suffices to prove the above statement for $f \geqslant 0$: Indeed, if it is true for $f \geqslant 0$, then for any arbitrary g, we have $\|T_\rho g\|_2 \leqslant \|T_\rho |g|\|_2 \leqslant \||g|\|_p = \|g\|_p$, as desired.
Note that $f(x) = a + bx$ for $x \in \{-1, 1\}$ for some $a, b \in \mathbb{R}$. Since $f \geqslant 0$, we have $a + b \geqslant 0, a - b \geqslant 0$, implying that $a \geqslant 0$. If $a = 0$, the statement is obvious, so assume $a > 0$. Note that multiplying f by a scalar doesn't affect the inequality, so WLOG assume $a = 1$. Then $f = 1 + bx$, and $f \geqslant 0$ implies $b \in [-1, 1]$.
Thus $f = 1 + bx$ with $b \in [-1, 1]$. Then $T_\rho f = 1 + \rho bx$, and consequently,

$$\|T_\rho f\|_2^{2p} = \frac{1}{2^p}\left((1-\rho b)^2 + (1+\rho b)^2\right)^p$$

On the other hand,

$$\|f\|_p^{2p} = \frac{1}{4}\left((1-b)^2 + (1+b)^2\right)^2$$

The inequality then follows by some routine calculus. ■

We now wish to extend this theorem to all n via induction. Now, it may seem natural to write the one function analog of the above statement as the induction statement, but that fails to work: Intuitively speaking, even if we start with a single function f, we end up with two different functions f and $T_\rho f$ the moment we apply the T_ρ operator.
Consequently, it is more natural to write a "two-function" analog of the induction statement. Before doing that, we convert Lemma 4.12 to two functions:

Proposition 9. Write $\rho = \sqrt{rs}$, where $r, s \in [0, 1]$. Then for any $f, g : \{-1, 1\} \mapsto \mathbb{R}$, we have:

$$\mathbb{E}_{y\sim N_\rho(x)}\left[f(x)g(y)\right] \leqslant \|f\|_{1+r} \cdot \|g\|_{1+s}$$

Proof. Note that if $y \sim N_\rho(x)$, then $\mathbb{E}[xy] = \rho$, while $\mathbb{E}[x] = \mathbb{E}[y] = 0$. Consequently, if $f(x) = a_1 + b_1x, g(y) = a_2 + b_2y$, then

$$\mathbb{E}_{y\sim N_\rho(x)}\left[f(x)g(y)\right] = \mathbb{E}_{y\sim N_\rho(x)}\left[a_1a_2 + a_2b_1x + a_1b_2y + b_1b_2xy\right] = a_1a_2 + b_1b_2\rho = a_1a_2 + (\sqrt{r}b_1)\cdot(\sqrt{s}b_2)$$

Note that $(T_{\sqrt{r}}f)(x) = a_1 + \sqrt{r}b_1x, (T_{\sqrt{s}}g)(y) = a_2 + \sqrt{s}b_2y$, and thus $\langle T_{\sqrt{r}}f, T_{\sqrt{s}}g\rangle = a_1a_2 + (\sqrt{r}b_1)\cdot(\sqrt{s}b_2)$.
Consequently,

$$\mathbb{E}_{y\sim N_\rho(x)}\left[f(x)g(y)\right] = \langle T_{\sqrt{r}}f, T_{\sqrt{s}}g\rangle \overset{\text{Cauchy-Schwartz}}{\leqslant} \|T_{\sqrt{r}}f\|_2 \cdot \|T_{\sqrt{s}}g\|_2 \overset{\text{Lemma 4.12}}{\leqslant} \|f\|_{1+r} \cdot \|g\|_{1+s}$$ ■

We now extend the above theorem to n variables via a very simple induction.

Lemma 4.13 (Weak Two-Function Hypercontractivity Theorem). Write $\mu = \sqrt{rs}$, where $r, s \in [0,1]$. Then for any $f, g : \{-1,1\}^n \mapsto \mathbb{R}$ (where $n \geqslant 1$), we have:

$$\mathbb{E}_{y \sim N_\mu(x)}\left[f(x)g(y)\right] \leqslant \|f\|_{1+r} \cdot \|g\|_{1+s}$$

Remark. The above theorem is "weak" since the restriction $r, s \leqslant 1$ can be removed.

Proof. Write $x = (x', x_n), y = (y', y_n)$. Note that if $y \sim N_\mu(x)$, then $y' \sim N_\mu(x')$, and $y_n \sim N_\mu(x_n)$. Also assume the theorem is true for functions on $n-1$ variables. Thus

$$\mathbb{E}_{y \sim N_\mu(x)}\left[f(x)g(y)\right] = \mathbb{E}_{y_n \sim N_\mu(x_n)}\left[\mathbb{E}_{y' \sim N_\mu(x')}\left[f(x)g(y)\right]\right]$$

Write $f_\alpha(\cdot) := f(\cdot, \alpha)$, where $\alpha \in \{-1, 1\}$. Then

$$\mathbb{E}_{y_n \sim N_\mu(x_n)}\left[\mathbb{E}_{y' \sim N_\mu(x')}\left[f(x)g(y)\right]\right] = \mathbb{E}_{y_n \sim N_\mu(x_n)}\left[\mathbb{E}_{y' \sim N_\mu(x')}\left[f_{x_n}(x')g_{y_n}(y')\right]\right] \leqslant \mathbb{E}_{y_n \sim N_\mu(x_n)}\left[\|f_{x_n}\|_{1+r} \cdot \|g_{y_n}\|_{1+s}\right]$$

where the inequality follows from the induction hypothesis. Now, let F be the function given by $x_n \mapsto \|f_{x_n}\|_{1+r}$, and define G analogously. Then

$$\mathbb{E}_{y_n \sim N_\mu(x_n)}\left[\|f_{x_n}\|_{1+r} \cdot \|g_{y_n}\|_{1+s}\right] = \mathbb{E}_{y_n \sim N_\mu(x_n)}\left[F(x_n)G(y_n)\right] \leqslant \|F\|_{1+r} \cdot \|G\|_{1+s}$$

Now,

$$\|F\|_{1+r}^{1+r} = \mathbb{E}_{x_n}\left[|F(x_n)|^{1+r}\right]$$

But $F(x_n) = \|f_{x_n}\|_{1+r}$, and thus $|F(x_n)|^{1+r} = \mathbb{E}_{x'}\left[|f_{x_n}(x')|^{1+r}\right] = \mathbb{E}_{x'}\left[|f(x)|^{1+r}\right]$, and thus $\mathbb{E}_{x_n}\left[|F(x_n)|^{1+r}\right] = \mathbb{E}_{x_n}\left[\mathbb{E}_{x'}\left[|f(x)|^{1+r}\right]\right] = \mathbb{E}_x\left[|f(x)|^{1+r}\right] = \|f\|_{1+r}^{1+r}$, i.e. $\|F\|_{1+r}^{1+r} = \|f\|_{1+r}^{1+r} \implies \|F\|_{1+r} = \|f\|_{1+r}$, as desired. ■

Proof of Theorem 4.11. Take $f = g$, and $r = s = \rho^2$ in Lemma 4.13. Indeed, note that

$$\mathbb{E}_{y \sim N_\mu(x)}\left[f(x)f(y)\right] = \sum_{S \subseteq [n]} \widehat{f}(S)^2 \mu^{|S|} = \|T_{\sqrt{\mu}} f\|_2^2$$

But $\mu = \sqrt{rs} = \rho^2$, and thus $\|T_{\sqrt{\mu}} f\|_2^2 = \|T_\rho f\|_2^2$. Meanwhile, $\|f\|_{1+\rho^2} \cdot \|f\|_{1+\rho^2} = \|f\|_{1+\rho^2}^2$. Since $\rho \leqslant \sqrt{p-1}$, we have $1 + \rho^2 \leqslant p$, and thus $\|f\|_{1+\rho^2}^2 \leqslant \|f\|_p^2$, as desired. ■

4.2. Hypercontractivity for General Random Variables

The proof of hypercontractivity of general random variables isn't very different from the proof of $(p, 2)$-hypercontractivity. Consequently, we just state the theorems and leave them as it is:

Theorem 4.14 (General Hypercontractivity for Product spaces). Let $\Omega_1, \ldots, \Omega_n$ be finite sets, and let $\pi_1, \ldots, \pi_n$ be probability distributions on $\Omega_1, \ldots, \Omega_n$ respectively. Let λ be the minimum probability mass placed by any distribution, i.e.

$$\lambda := \min_{i \in [n]} \min_{\omega \in \operatorname{supp}(\pi_i)} \pi_i(\omega)$$

Fix $q > 2$, and let p be the Hölder conjugate of q, i.e. $p^{-1} + q^{-1} = 1$. Also suppose

$$0 \leqslant \rho \leqslant \frac{\lambda^{1/2 - 1/q}}{\sqrt{q-1}}$$

Then we have

$$\|T_\rho f\|_q \leqslant \|f\|_2, \|T_\rho f\|_2 \leqslant \|f\|_p$$

Remark. Note that we define

$$\|f\|_p^p := \mathbb{E}_{x\sim\pi_1\otimes\cdots\otimes\pi_n}\left[|f(x)|^p\right]$$

A version of hypercontractivity even holds for general random variables on discrete spaces:

Definition 4.1. Fix $1 \leqslant p \leqslant q \leqslant \infty$, and $0 \leqslant \rho < 1$. Let X be a real-valued random variable such that $\|X\|_q := \mathbb{E}\left[|X|^q\right]^{1/q} < \infty$. We say X is (p, q, ρ)-hypercontractive if for all $a, b \in \mathbb{R}$, we have

$$\|a + \rho bX\|_q \leqslant \|a + bX\|_p$$

Theorem 4.15. Let X be a mean 0 discrete random variable, and let $\lambda < \frac{1}{2}$ be the least value of its probability mass function. Write $u = \ln\frac{1-\lambda}{\lambda}$. Fix $q > 2$, and let p be the Hölder conjugate of q. Then X is $(2, q, \rho)$-hypercontractive, and $(p, 2, \rho)$-hypercontractive for

$$0 \leqslant \rho \leqslant \sqrt{\frac{\sinh(u/q)}{\sinh(u/p)}}$$

4.3. Applications of Hypercontractivity

We shall see some applications of hypercontractivity:

Lemma 4.16. Suppose $f : \{-1, 1\}^n \mapsto \mathbb{R}$ has degree $\leqslant k$, i.e. $\widehat{f}(S) = 0$ for all S such that $|S| > k$. Then:

1. For $q \geqslant 2$, we have $\|f\|_q \leqslant (q-1)^{k/2}\|f\|_2$.
2. For $p \leqslant 2$, we have $\|f\|_2 \leqslant (e^{2/p} - 1)^k\|f\|_p$.

Remark. Note that if we didn't have the low-degree information about f, then bounds upper bounding ℓ_q-norms in terms of ℓ_2-norms would have involved n, which is the number of variables. Thus, the power of this inequality comes from the fact that it is independent of n.

Proof. Let q be the Hölder conjugate of p, i.e. $p^{-1} + q^{-1} = 1$. Now,

$$\|f\|_q = \|T_{1/\sqrt{q-1}}T_{\sqrt{q-1}}f\|_q$$

Note that we have $(p, 2)$-hypercontractivity for $\rho = \sqrt{p-1} = \frac{1}{\sqrt{q-1}}$. We thus have $(2, q)$-hypercontractivity for $\rho = \frac{1}{\sqrt{q-1}}$ by Proposition 8. Consequently, $\|T_{1/\sqrt{q-1}}T_{\sqrt{q-1}}f\|_q \leqslant \|T_{\sqrt{q-1}}f\|_2$. But

$$\|T_{\sqrt{q-1}}f\|_2^2 = \langle T_{\sqrt{q-1}}f, T_{\sqrt{q-1}}f\rangle = \sum_{S\subseteq[n]}(q-1)^{|S|}\widehat{f}(S)^2 = \sum_{|S|\leqslant k}(q-1)^{|S|}\widehat{f}(S)^2 \leqslant (q-1)^k\sum_{|S|\leqslant k}\widehat{f}(S)^2 = (q-1)^k\|f\|_2^2$$

Thus, $\|f\|_q \leqslant (q-1)^{k/2}\|f\|_2$.
Now, let $\theta \in (0, 1)$ be the solution to $\frac{1}{2} = \frac{\theta}{p} + \frac{1-\theta}{2+\varepsilon}$, for some $\varepsilon > 0$. Then by generalized Hölder's inequality, we have

$$\|f\|_2 \leqslant \|f\|_{2+\varepsilon}^{1-\theta}\cdot\|f\|_p^\theta \leqslant (1+\varepsilon)^{k(1-\theta)/2}\|f\|_2^{1-\theta}\cdot\|f\|_p^\theta$$

Assuming $\|f\|_2 \neq 0$ (because otherwise the inequality is trivial), we have

$$\|f\|_2^\theta \leqslant (1+\varepsilon)^{k(1-\theta)/2}\|f\|_p^\theta \implies \|f\|_2 \leqslant (1+\varepsilon)^{k(1-\theta)/(2\theta)}\|f\|_p$$

Taking $\varepsilon \searrow 0$ yields the desired result. ■

Remark. In the setting of Theorem 4.14, we have a similar result stating that for $q \geqslant 2$, we have $\|f\|_q \leqslant (\sqrt{q-1} \cdot \lambda^{1/2-1/q})^k \|f\|_2$. Similarly, for $p = 1$ (for other p in $[1, 2]$ similar results follow), we have $\|f\|_2 \leqslant c(\lambda)^k \|f\|_1$, where

$$c(\lambda) := \left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{2(1-2\lambda)}}$$

Corollary 4.17 (Khintchine's inequality)**.** Suppose $X_1, \ldots, X_n$ are mean 0 i.i.d ± 1-valued random variables. Also suppose $a_1, \ldots, a_n \in \mathbb{R}$ are any real numbers. Then for any $p \in [1, \infty)$, there exist constants $0 < c_p \leqslant C_p$ such that

$$c_p \left(\sum_i a_i^2\right)^{1/2} \leqslant \left(\mathbb{E}\left[\left|\sum_i a_i X_i\right|^p\right]\right)^{1/p} \leqslant C_p \left(\sum_i a_i^2\right)^{1/2}$$

Proof. Define the function $f(x) := \sum_{i=1}^n a_i x_i$. Note that $\left(\mathbb{E}\left[\left|\sum_i a_i X_i\right|^p\right]\right)^{1/p} = \|f\|_p$, and $\left(\sum_i a_i^2\right)^{1/2} = \|f\|_2$, for any $a \in [1, \infty)$. Now, if $p \in [1, 2]$, we may take $C_p = 1$ and $c_p = (e^{2/p} - 1)^{-1}$.
For $p > 2$, we may take $c_p = 1$ and $C_p = \sqrt{p-1}$. ■

We can have concentration for general polynomials too:

Theorem 4.18. Suppose $f : \{-1, 1\}^n \mapsto \mathbb{R}$ has degree $\leqslant k$, and also assume f is not identically 0. Then for any $t \geqslant (2e)^{k/2}$ we have:

$$\Pr_{x \sim \{\pm 1\}^n} \left(|f(x)| \geqslant t\|f\|_2\right) \leqslant \exp(-kt^{2/k}/(2e))$$

Remark. A few remarks are due:

1. Note that for $k = 1$, this recovers the Chernoff bound (upto quantitative factors). Thus, the above theorem can be seen as a generalization of Chernoff's bound from sums to low-degree polynomials.
2. In the setting of Theorem 4.14, we have for any $t \geqslant (2e/\lambda)^{k/2}$,

$$\Pr\left(|f(x)| \geqslant t\|f\|_2\right) \leqslant \lambda^k \exp(-k\lambda t^{2/k}/(2e))$$

Proof. Note that

$$\Pr(|f(x)| \geqslant t\|f\|_2) = \Pr(|f(x)|^q \geqslant t^q \|f\|_2^q) \overset{\text{Markov}}{\leqslant} \frac{\mathbb{E}\left[|f|^q\right]}{t^q \|f\|_2^q} \overset{\text{Lemma 4.16}}{\leqslant} \frac{(q-1)^{kq/2}}{t^q} < \frac{q^{kq/2}}{t^q}$$

Taking $q = t^{2/k}/e$ proves the desired theorem. ■

We can also use hypercontractivity to obtain anti-concentration results (generalizing Corollary 4.6)!

Theorem 4.19. Suppose $f : \{-1, 1\}^n \mapsto \mathbb{R}$ is a non-constant function with degree $\leqslant k$. Then:

$$\Pr_{x \sim \{\pm 1\}^n} \left(f(x) \geqslant \mathbb{E}[f]\right) \geqslant \frac{(e^2-1)^{-k}}{4}$$

Proof. We can replace f by $f - \mathbb{E}[f]$ to assume WLOG that $\mathbb{E}[f] = 0$. Now, write $g(x) := f(x) \cdot \mathbb{1}_{f(x)>0}$, and let $h = g - f$. Then $g, h \geqslant 0$. Also, $0 = \mathbb{E}[f] = \mathbb{E}[g] - \mathbb{E}[h] = \|g\|_1 - \|h\|_1$. At the same time, $\|g\|_1 + \|h\|_1 = \|f\|_1$. Thus $\|g\|_1 = \frac{\|f\|_1}{2}$. Thus

$$\frac{\|f\|_1^2}{4} = \mathbb{E}[g]^2 = \mathbb{E}\left[f(x) \cdot \mathbb{1}_{f(x)>0}\right]^2 \overset{\text{Cauchy-Schwartz}}{\leqslant} \mathbb{E}\left[f^2\right] \cdot \mathbb{E}\left[\mathbb{1}_{f(x)>0}^2\right] \overset{\text{Lemma 4.16}}{\leqslant} (e^2-1)^k \|f\|_1^2 \cdot \Pr(f(x) > 0)$$ ■

Remark. In the setting of Theorem 4.14, we have $\Pr\left(f(x) \geqslant \mathbb{E}[f]\right) \geqslant (15/\lambda)^{-k}$.

We turn our attention to something called small-set expansion: Basically, if f is the indicator of a "small set", i.e. if $f = \mathbb{1}_A$, where $|A|/2^n \ll 1$, then most of the mass of f lies in high frequency components. Another consequence is that the "noisy hypercube" is a small set expander.

Theorem 4.20 (Noisy Hypercube is a small-set expander)**.** Suppose $A \subseteq \{-1,1\}^n$ has volume $\alpha \cdot 2^n$, i.e. $\mathbb{E}[\mathbb{1}_A] = \alpha$. Then for any $\rho \in [0,1]$,

$$\text{Stab}_\rho(\mathbb{1}_A) = \Pr_{x \sim \{\pm 1\}^n, y \sim N_\rho(x)}(x \in A, y \in A) \leqslant \alpha^{2/(1+\rho)}$$

Equivalently, for $\alpha \neq 0$,

$$\Pr_{x \sim A, y \sim N_\rho(x)}(y \in A) \leqslant \alpha^{(1-\rho)/(1+\rho)}$$

Proof. Note that

$$\Pr_{x \sim \{\pm 1\}^n, y \sim N_\rho(x)}(x \in A, y \in A) = \langle \mathbb{1}_A, T_\rho \mathbb{1}_A \rangle = \|T_{\sqrt{\rho}} \mathbb{1}_A\|_2^2 \overset{\text{Theorem 4.11}}{\leqslant} \|\mathbb{1}_A\|_{1+\rho}^2 = \alpha^{2/(1+\rho)}$$ ■

Remark. Why do we use the term "small-set expander"? Suppose $\alpha \ll 1$, and also suppose $\rho \approx 0$, i.e. the hypercube is very noisy. Then $\frac{1-\rho}{1+\rho} \approx 1$, and thus $\Pr_{x \sim A, y \sim N_\rho(x)}(y \in A) \leqslant \alpha^{(1-\rho)/(1+\rho)} \approx \alpha \ll 1$, i.e. with very high probability, one step of the random walk escapes the set.

Theorem 4.21 (Generalized Small-Set Expansion Theorem)**.** We have a generalized product space analog of Lemma 4.13. Applying that to this setting yields the following result: Suppose $A, B \subseteq \{-1,1\}^n$ are such that their normalized volumes are $\exp(-a^2/2), \exp(-b^2/2)$ respectively. Suppose $\rho \in [0,1]$ is such that $0 \leqslant \rho a \leqslant b \leqslant a$. Then

$$\Pr_{y \sim N_\rho(x)}(x \in A, y \in B) \leqslant \exp\left(-\frac{a^2 - 2\rho ab + b^2}{2(1-\rho^2)}\right)$$

Remark. The hypercontractivity theorem can be "reversed". In particular, consider the same setting as the above theorem, except that we don't have the restriction $0 \leqslant \rho a \leqslant b \leqslant a$, i.e. $\rho \in [0,1]$ is arbitrary. Then

$$\Pr_{y \sim N_\rho(x)}(x \in A, y \in B) \geqslant \exp\left(-\frac{a^2 + 2\rho ab + b^2}{2(1-\rho^2)}\right)$$

Theorem 4.22 (Level k-Inequalities)**.** Suppose we have $f = \mathbb{1}_A$ for some $A \subseteq \{-1,1\}^n$, and suppose $\mathbb{E}[f] = \alpha$. Also suppose $k \leqslant 2\ln(1/\alpha)$. Then

$$W^{\leqslant k}[f] \leqslant \left(\frac{2e}{k}\ln(1/\alpha)\right)^k \alpha^2$$

Proof. We know that $\mathrm{Stab}_\rho[f] = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}(S)^2$ by Corollary 3.2. Thus $W^{\leqslant k}[f] \leqslant \rho^{-k}\,\mathrm{Stab}_\rho[f]$ for $\rho \leqslant 1$. By Theorem 4.20, we also know that $\mathrm{Stab}_\rho(\mathbb{1}_A) \leqslant \alpha^{2/(1+\rho)}$. Thus $W^{\leqslant k}[f] \leqslant \rho^{-k}\alpha^{2/(1+\rho)}$ for all $\rho \in [0,1]$. Setting $\rho = \frac{k}{2\ln(1/\alpha)}$ proves the desired claim. ■

§5. p-biased Analysis

We often come across scenarios where we have Bernoulli Random variables, with parameters that aren't necessarily $1/2$. For example, take the Erdős-Rényi model $G(n, p)$, or Bernoulli percolation on lattices in general.
We would thus like to develop a theory of Fourier analysis where all $x \in \{-1, 1\}^n$ aren't sampled uniformly; rather, x is sampled with probability $p^k(1-p)^{n-k}$, where k is the number of entries of x that are -1.
To that end, let π_p denote the distribution of the Bernoulli/Rademacher random variable, which assumes -1 with probability p, and 1 with probability $1-p$, i.e.

$$\mu_i := \mathbb{E}_{x_i \sim \pi_p}[x_i] = 1 - 2p$$

$$\sigma_i := \sqrt{\mathrm{Var}(x_i)} = 2\sqrt{p(1-p)}$$

For simplicity, we will assume that all our variables $x_1, \ldots, x_n$ are i.i.d sampled from π_p, i.e. x is sampled from $\pi_p^{\otimes n}$. Now, define

$$\phi(x_i) := \left(\frac{x_i - \mu}{\sigma}\right)$$

We have $\phi(1) = \sqrt{p/(1-p)}, \phi(-1) = -\sqrt{(1-p)/p}$. Furthermore, $\mathbb{E}[\phi] = 0, \mathbb{E}[\phi^2] = 1$. Now, for any $S \subseteq [n]$, define

$$\phi_S(x) := \prod_{i \in S} \phi(x_i)$$

Note that if $p \neq 1/2$, then $\phi_S \cdot \phi_T \neq \phi_{S \oplus T}$ in general. However, $\mathbb{E}_{x \sim \pi_p^{\otimes n}}\left[\phi_S(x) \cdot \phi_T(x)\right] = 0$, since any $i \in (S \setminus T) \cup (T \setminus S)$ factorizes out and gives 0 expectation.
Now, for any $f : \{-1, 1\}^n \mapsto \mathbb{R}$, we can define our Fourier coefficients as:

$$\widehat{f}_p(S) := \mathbb{E}_{x \sim \pi_p^{\otimes n}}\left[f(x)\phi_S(x)\right]$$

Since the $\{\phi_S\}_{S \subseteq [n]}$ is still an orthonormal basis under the modified inner product space, our Fourier decomposition $f = \sum_{S \subseteq [n]} \widehat{f}_p(S)\phi_S$ continues to hold.
We now define the derivative and influence operators: Note that $\frac{\partial}{\partial \phi_i} = \frac{\partial}{\partial x_i} \cdot \frac{\partial x_i}{\partial \phi_i} = \sigma \cdot \frac{\partial}{\partial x_i}$. Thus the new derivative operator is defined as:

$$D_{i,p} f := \sigma D_i f = 2\sqrt{p(1-p)} D_i f$$

Thus the p-biased derivative is just a rescaled version of the "original" derivative operator. Similar to Proposition 5, we have the equality

$$D_{i,p}(f) = \sum_{i \in S} \widehat{f}_p(S)\phi_{S \setminus i}$$

Using the above definition of derivatives, and simplifying $\mathrm{Inf}_{i,p}[f] := \mathbb{E}_{x \sim \pi_p^{\otimes n}}\left[(D_{i,p} f)^2\right]$ yields:

$$\mathrm{Inf}_{i,p}[f] := \sigma^2 \Pr_{x \sim \pi_p^{\otimes n}}(f(x) \neq f(x^{\oplus i}))$$

Also define:

$$\mathbb{I}_p[f] := \sum_{i=1}^{n} \mathrm{Inf}_{i,p}[f]$$

Furthermore, similar to Lemma 2.5, if f is monotone, we have

$$\mathrm{Inf}_{i,p}[f] = \sigma \widehat{f}_p(i)$$

The reason p-biased Fourier analysis is so important is that it allows us to track the behavior of the system as we change p, which is a proxy for the expectation of the system. For example, consider the Boolean function $f_{\mathsf{conn}} : \{-1, 1\}^{\binom{n}{2}} \mapsto \{-1, 1\}$, which tracks whether the graph described by $\{-1, 1\}^{\binom{n}{2}}$ is connected or not. Clearly then, tracking phase transitions in f_{conn} as we vary the parameter p is the theory of Erdős-Rényi random graphs!
One of the most fundamental lemmata in this regard is the *Margulis-Russo* formula:

Lemma 5.1 (Margulis-Russo formula)**.** Let $f : \{-1,1\}^n \mapsto \mathbb{R}$ be a function, let $p \in [0,1]$, and set $\mu = 1-2p$. Then

$$\frac{d}{d\mu}\left(\mathbb{E}_{x\sim\pi_p^{\otimes n}}\left[f(x)\right]\right) = \frac{1}{\sigma}\cdot\sum_{i=1}^{n}\widehat{f}_p(i)$$

In particular, if $f : \{-1,1\}^n \mapsto \{-1,1\}$ is monotone, then

$$\frac{d}{dp}\left(\Pr_{x\sim\pi_p^{\otimes n}}(f(x)=-1)\right) = \frac{1}{\sigma^2}\cdot\mathbb{I}_p[f]$$

Proof. Let $f = \sum_{S\subseteq[n]}\widehat{f}(S)\chi_S$ be the ordinary Fourier expansion of f. Then

$$\mathbb{E}_{x\sim\pi_p^{\otimes n}}[f] = \sum_{S\subseteq[n]}\widehat{f}(S)\cdot\mathbb{E}_{x\sim\pi_p^{\otimes n}}[\chi_S] = \sum_{S\subseteq[n]}\widehat{f}(S)\cdot\mu^{|S|}$$

Thus

$$\frac{d}{d\mu}\left(\mathbb{E}_{x\sim\pi_p^{\otimes n}}\left[f(x)\right]\right) = \frac{d}{d\mu}\left(\sum_{S\subseteq[n]}\widehat{f}(S)\cdot\mu^{|S|}\right) = \sum_{S\neq\varnothing}\widehat{f}(S)\cdot|S|\mu^{|S|-1} = \sum_{i=1}^{n}\sum_{i\in S}\widehat{f}(S)\mu^{|S|-1}$$

At the same time, from the definition of $\widehat{f}_p(S)$, we have:

$$\widehat{f}_p(S) = \sum_{T\subseteq[n]}\widehat{f}(T)\mathbb{E}_{x\sim\pi_p^{\otimes n}}[\chi_T\phi_S]$$

Now, if $i \in S\setminus T$, then $\mathbb{E}[\chi_T\phi_S] = 0$, since $\mathbb{E}\left[\phi(x_i)\right]$ will factorize out and become 0. On the other hand, if $S\subseteq T$, then the expectation becomes $\mu^{|T|-|S|}$. Thus

$$\widehat{f}_p(S) = \sum_{S\subseteq T}\widehat{f}(T)\mu^{|T|-|S|} \implies \widehat{f}_p(i) = \sum_{i\in T}\widehat{f}(T)\mu^{|T|-1}$$

and we're done. ■

Note that one of the central objectives of random graph theory, percolation theory, and many other fields throughout probability theory and computer science, is to show that the emergence of some property *exhibits a phase transition*, i.e. there is a certain parameter 'p', and a certain threshold p_c, such that if $p < p_c$, then the probability of the phenomenon is ≈ 0, while if $p > p_c$, then the probability of the phenomenon is ≈ 1. Furthermore, once we have established a phase transition, we are also interested in showing that the transition is sharp, i.e. for every $\varepsilon > 0$, there exists a small quantity $\delta = \delta(\varepsilon) > 0$, such that if $p < p_c - \delta$, then the probability of our phenomenon is $< \varepsilon$, while if $p > p_c + \delta$, then the probability of our phenomenon is $> 1-\varepsilon$.

Indeed, such phase transitions have been rigorously proven in many contexts: For example, using standard probability theory, it is not too difficult to establish that the emergence of a clique, or the graph becoming connected, in the Erdős-Rényi model $G(n,p)$, exhibits a sharp phase transition. In an infinite context, the emergence of infinite clusters in random lattices exhibit a sharp phase transition.

The following theorem, by Friedgut and Kalai ([FK96]), proves a sharp phase transition for *all monotone properties* [5]!

Before we state and prove the theorem, we fix a bit of notation: Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a function. Define a function $\nu : [0,1] \mapsto [0,1]$, where:

$$\nu(p) := \Pr_{x\sim\pi_p^{\otimes n}}(f(x) = -1)$$

Recall that our convention was that '-1' is the 'true' value.

Also, we shall need a small lemma for p-biased cubes, which we shall not prove. Instead, we present a proof for the unbiased case.

[5] we need our property to be monotone, so that when we increase our parameter p, the probability of the property increases

Corollary 5.2. There is an absolute constant $c > 0$ such that for every function $f : \{-1,1\}^n \mapsto \{-1,1\}$, there is a variable $i \in [n]$ such that $\mathrm{Inf}_{i,p}[f] \geqslant c\eta' \log n/n$, where $\eta' := \min(\nu(p), 1-\nu(p))$, for every $p \in [0,1]$.

Proof for $p = 1/2$. By Lemma 1.8, $\mathrm{Var}(f) = 4\nu(p)(1-\nu(p)) \geqslant 2\eta'$. The result now follows from Theorem 4.9. ■

Theorem 5.3 (Monotone properties exhibit sharp thresholds)**.** Let $f : \{-1,1\}^n \mapsto \{-1,1\}$ be a symmetric monotone Boolean-valued function such that $\nu(\cdot)$ is a strictly increasing function. Then there exists an absolute constant $c > 0$ such that if $1/2 \geqslant \nu(p_0) > \varepsilon$, then

$$\nu\left(p_0 + c\frac{\log(1/2\varepsilon)}{\log n}\right) > 1 - \varepsilon$$

Remark. Recall that a function $f : \{-1,1\}^n \mapsto \mathbb{R}$ is called symmetric, if for any permutation σ of $[n]$, and any $(x_1, \ldots, x_n) \in \{-1,1\}^n$, we have $f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$.

Proof. Since f is symmetric, the influence of every variable is the same, and thus, by Lemma 5.1,

$$\frac{d\nu}{dp} = \frac{n}{4p(1-p)}\mathrm{Inf}_{1,p}[f]$$

At the same time, by Corollary 5.2, $\mathrm{Inf}_{1,p}[f] \geqslant c'\nu(p)\log n/n$ for some absolute $c' > 0$, for every p such that $\nu(p) \leqslant 1/2$. Thus, for $p \leqslant \nu^{-1}(1/2)$,

$$\frac{d\nu}{dp} \geqslant \frac{c'\nu\log n}{4p(1-p)} \geqslant c'\nu\log n \implies \frac{d\nu}{\nu} \geqslant c'p\log n \implies d(\log\nu) \geqslant c'p\log n \tag{5.1}$$

Now, let $\nu(p_0) > \varepsilon$, $p^* = \nu^{-1}(\frac{1}{2})$. Then integrating Eq. (5.1) yields:

$$\log\frac{1}{2} - \log\varepsilon > c'\left(p^* - p_0\right)\log n \implies p_0 + c'\frac{\log(1/2\varepsilon)}{\log n} > p^*$$

Thus, by symmetry, at $q = p_0 + 2c'\frac{\log(1/2\varepsilon)}{\log n}$, we will have $\nu(q) > 1 - \varepsilon$, as desired. ■

§6. Håstad's Hardness of Approximation Results

We will now use Fourier analysis to prove some hardness of approximation results. Let us first review some standard terminology.

Problem (MAX-E3-SAT). Consider a CNF formula, where every clause has exactly 3 literals, all of which correspond to distinct variables (i.e. clauses like $x_1 \vee \neg x_1 \vee x_2$ are not allowed). Our job is to find a Boolean assignment of variables that maximizes the number of clauses satisfied.

Problem (MAX-E3-LIN). Consider a system of linear equations over $\mathbb{F}_2$, where each equation looks like $x_{i_1} + x_{i_2} + x_{i_3} = b_i$, where i_1, i_2, i_3 are distinct, and $b_i \in \mathbb{F}_2$. Once again, we have to find an assignment $\{x_1, \ldots, x_n\} \mapsto \mathbb{F}_2$ that maximizes the fraction of equations satisfied.

Both of these problems are NP-complete, so we look for approximation algorithms for them. In particular, we will look for (α, β)-approximation algorithms, i.e. on any instance of the problem with optimum value $\geqslant \beta$, our algorithm should output $\geqslant \alpha$ [6].
There is a very easy $7/8$-approximation algorithm for MAX-E3-SAT: Assign each variable a(n uniformly) random Boolean value. Each clause is satisfied with probability $7/8$, and thus in expectation, $7/8$ fraction of the clauses is satisfied. Furthermore, this randomized algorithm can be efficiently derandomized.
Similarly, there is a trivial $1/2$-approximation algorithm for MAX-E3-LIN, which assigns x_i either 0 or 1 with equal probability.
Now, a landmark result of Håstad [H01] in complexity theory says that it is NP-hard to obtain a $(7/8 + \delta, 1)$-approximation of MAX-E3-SAT, and it is also NP-hard to obtain a $(1/2 + \delta, 1 - \delta)$-approximation of MAX-E3-SAT, i.e. the trivial approximation algorithms are the best we can hope for.
We will now start our journey towards Håstad's results. While Håstad's results are quite technical, we can still establish a somewhat weaker version, which says that assuming the Unique Games Conjecture, it is NP-hard to $(7/8 + \delta, 1 - \delta)$-approximate MAX-E3-SAT, or $(1/2 + \delta, 1 - \delta)$-approximate MAX-E3-LIN. While this is not completely satisfactory (since Håstad's results are completely unconditional, while the Unique Games Conjecture is still a conjecture), the ideas used in proving this more or less capture Håstad's main ideas; indeed, with a little more technical wrangling on top of the ideas described here, even Håstad's results can be established.
We first define an "attenuated" version of influence, inspired by the T_ρ operator.

Definition 6.1. For $f : \{-1, 1\}^n \mapsto \mathbb{R}$, $\rho \in [0, 1]$ and $i \in [n]$, we define the ρ-stable influence of f at i to be:

$$\mathrm{Inf}_i^{(\rho)}[f] = \mathrm{Stab}_\rho[D_i f] = \sum_{i \in S} \rho^{|S|-1} \widehat{f}(S)^2$$

We also define $\mathbb{I}^{(\rho)}[f] := \sum_{i=1}^n \mathrm{Inf}_i^{(\rho)}[f]$.

Remark. A few remarks are in order:

1. Recall that this expression arose in the proof of the KKL theorem too.
2. Clearly, $\mathrm{Inf}_i^{(1)}[f] = \mathrm{Inf}_i[f]$.
3. One can easily verify that $\mathbb{I}^{(\rho)}[f] = \frac{d}{d\rho} \mathrm{Stab}_\rho[f] = \sum_{k=1}^n k\rho^{k-1} \cdot W^k[f]$.

[6] if our algorithm is randomized, then the expectation of the random variable outputted by our algorithm should be $\geqslant \alpha$

We define a "noise-stable" version of small influence using the definition above:

Definition 6.2. For $f : \{-1,1\}^n \mapsto \mathbb{R}$, $\rho \in [0,1]$ and $i \in [n]$, we say that i is (ε,δ)-notable if $\mathrm{Inf}_i^{(1-\delta)}[f] > \varepsilon$.

Remark. Note that as $\varepsilon \to 1, \delta \to 0$, (ε,δ)-notability captures the notion of dictatorship.

As it turns out, proving hardness-of-approximation is intimately connected to "dictatorship tests".

Definition 6.3 ((α,β)-Dictator vs. No-notables test)**.** A (α,β)-Dictator vs. No-notables test is a local tester for functions $f : \{-1,1\}^n \mapsto \{-1,1\}$ such that:

1. If f is a dictator, then the test accepts with probability $\geqslant \beta$.
2. If f has no $(\varepsilon,\varepsilon)$-notable coordinates, i.e. if $\mathrm{Inf}_i^{(1-\varepsilon)}[f] \leqslant \varepsilon$ for every $i \in [n]$, then the test accepts with probability $\leqslant \alpha + o_\varepsilon(1)$.

Remark. A local tester queries the function $\mathcal{O}(1)$ many times.

We will now design a dictatorship test for the MAX-E3-LIN problem. Naturally, our test must involve linear equations with 3 variables to maintain a connection with the MAX-E3-LIN problem.
We will modify the BLR test (Theorem 1.10) to accept dictators with high probability, and reject "egalitarian" functions with probability $\sim 1/2$.
Firstly, the "BLR"-way to design a test would simply check if $f(x)f(y)f(x \circ y) = 1$, where $x \circ y \in \{-1,1\}^n$ is the pointwise product of x, y. Clearly, dictators pass this test [7]. However, the constant function $f \equiv 1$ passes this test too, despite having no notable coordinates. Thus, to eliminate the constant 1 function, we introduce a "global flip" b. However, even with the flip, we still accept χ_S for odd $|S|$. Furthermore, if $|S|$ is large, then no coordinate in χ_S is notable. Thus, to eliminate χ_S, we introduce some noise to $z = b \cdot (x \circ y)$. The noise destroys the careful parity balance between the LHS and RHS, while still accepting dictators with high enough probability.
Thus, let's now formalize the above design intuitions:

Algorithm 2: Håstad$_\delta$-test

Data: $f : \{-1,1\}^n \mapsto \{-1,1\}$
1. Pick x, y independently and uniformly from $\{-1,1\}^n$;
2. Pick b uniformly from $\{-1,1\}$, and set $z = b \cdot (x \circ y)$ ($x \circ y \in \{-1,1\}^n$ is the pointwise product of x, y);
3. Choose $z' \sim N_{1-\delta}(z)$;
4. Accept if $f(x)f(y)f(z') = b$

Proof of correctness of Algorithm 2. Note that

$$\Pr(\text{Håstad}_\delta \text{ accepts } f | b = 1) = \mathbb{E}_{x,y,z'}\left[\frac{1}{2} + \frac{1}{2} f(x)f(y)f(z')\right] = \frac{1}{2} + \frac{1}{2} \cdot \mathbb{E}_{x,y}\left[f(x)f(y) \cdot \mathbb{E}_{z'}\left[f(z')|x,y\right]\right]$$

But note that given x, y and $z' \sim N_{1-\delta}(t)$, where $t = b \cdot (x \circ y) = x \circ y$, we have $\mathbb{E}_{z'}\left[f(z')|x,y\right] = (T_{1-\delta} f)(t) = (T_{1-\delta} f)(x \circ y)$. Thus

$$\Pr(\text{Håstad}_\delta \text{ accepts } f | b = 1) = \frac{1}{2} + \frac{1}{2} \cdot \mathbb{E}_{x,y}\left[f(x)f(y) \cdot (T_{1-\delta} f)(x \circ y)\right] = \frac{1}{2} + \frac{1}{2} \cdot \mathbb{E}_{x}\left[f(x) \cdot (f * (T_{1-\delta} f))(x)\right]$$

[7] although note that anti-dictators fail

$$= \frac{1}{2} + \frac{1}{2}\sum_{S\subseteq[n]} \widehat{f}(S)\widehat{f * T_{1-\delta}f}(S) = \frac{1}{2} + \frac{1}{2}\sum_{S\subseteq[n]} (1-\delta)^{|S|}\widehat{f}(S)^3$$

Similarly,

$$\Pr(\text{Håstad}_\delta \text{ accepts } f|b=-1) = \mathbb{E}_{x,y,z'}\left[\frac{1}{2} - \frac{1}{2}f(x)f(y)f(z')\right]$$

where $z' \sim N_{1-\delta}(-(x \circ y)) = N_{-(1-\delta)}(x \circ y)$, where we note that $N_\rho(-x) = N_{-\rho}(x)$. Thus

$$\Pr(\text{Håstad}_\delta \text{ accepts } f|b=-1) = \frac{1}{2} - \frac{1}{2}\sum_{S\subseteq[n]} (-1)^{|S|}(1-\delta)^{|S|}\widehat{f}(S)^3$$

Consequently,

$$\Pr(\text{Håstad}_\delta \text{ accepts } f) = \frac{1}{2} + \frac{1}{2}\sum_{|S| \text{ odd}} (1-\delta)^{|S|}\widehat{f}(S)^3$$

If f is a dictator, i.e. $f = \chi_i$, then the aforementioned probability is $1 - \delta/2$ [8]. On the other hand,

$$\sum_{|S| \text{ odd}} (1-\delta)^{|S|}\widehat{f}(S)^3 \leqslant \left(\max_{|S| \text{ odd}} (1-\delta)^{|S|}\widehat{f}(S)\right) \cdot \sum_{|S| \text{ odd}} \widehat{f}(S)^2 \leqslant \max_{|S| \text{ odd}} (1-\delta)^{|S|}\widehat{f}(S) \leqslant \sqrt{\max_{|S| \text{ odd}} (1-\delta)^{2|S|}\widehat{f}(S)^2}$$

$$\leqslant \sqrt{\max_{|S| \text{ odd}} (1-\delta)^{|S|-1}\widehat{f}(S)^2} \leqslant \sqrt{\max_{i\in[n]} \mathrm{Inf}_i^{(1-\delta)}[f]}$$

Thus, if f has no (δ, δ)-notable coordinates, then

$$\Pr(\text{Håstad}_\delta \text{ accepts } f) \leqslant \frac{1}{2} + \frac{1}{2}\sqrt{\max_{i\in[n]} \mathrm{Inf}_i^{(1-\delta)}[f]} \leqslant \frac{1}{2} + \frac{1}{2}\sqrt{\delta}$$

as desired. ■

Thus, we have a $(1/2 + \delta, 1 - \delta)$-dictatorship test for MAX-E3-LIN. At this point we're done, by simply invoking a black-box result from [KKMO07]:

Theorem 6.1. Suppose for each $n \in \mathbb{N}$, we have a (α, β)-Dictator vs. No-notables tester for $f : \{-1, 1\}^n \mapsto \{-1, 1\}$, such that our tester uses predicates of some CSP Ψ. Then, assuming the Unique Games Conjecture, a $(\alpha + \delta, \beta - \delta)$-approximation to MaxCSP(Ψ) is NP-hard.

Remark. When we say our tester uses "predicates of Ψ", we mean that our tester evaluates it's queries using allowed predicates. For example, Algorithm 2 made it's decision by composing $f(x), f(y), f(z')$ using a MAX-E3-LIN predicate.

Thus, under the Unique Games Conjecture, it is NP-hard to find a $(1/2 + \delta, 1 - \delta)$-approximation for MAX-E3-LIN. As mentioned earlier, Håstad proved this result unconditionally, but nevertheless, we have managed to convey the key ideas of his proof.

[8] note that if $f = -\chi_i$, i.e. an anti-dictator, then the above probability is $\delta/2$. Thus, the Håstad$_\delta$-test accepts dictators only, not 1-juntas

§7. A Brief Taste of Additive Combinatorics over $\mathbb{F}_2^n$

Let $A \subseteq \mathbb{F}_2^n$ be a non-empty set. Define $A + A := \{a + a' : a, a' \in A\}$. Also, define $\mu(A) := |A|/2^n$ to be the measure of A. Finally, define the sum-set expansion of A to be

$$e(A) := \frac{\mu(A + A)}{\mu(A)}$$

Clearly, since $a + A \subseteq A + A$ for any $a \in A$, and since $\mu(a + A) = \mu(A)$, we have $e(A) \geqslant 1$ for any non-empty set A. Also note that for any non-empty A, $0 = a + a \in A + A$ since we're working over $\mathbb{F}_2$.

Proposition 10. $A = A + A$ if and only if A is a subspace of $\mathbb{F}_2^n$.

Proof. If $A = A + A$, then A is closed under addition and is thus a subspace of $\mathbb{F}_2^n$: Indeed, linear combinations in $\mathbb{F}_2$ are just ordinary sums, and $0 \in A + A$, so A is a subspace. Conversely, if A is a subspace then $A = A + A$. ■

Proposition 11. $e(A) = 1$ if and only if $A = a + H$, where H is a subspace of $\mathbb{F}_2^n$.

Proof. Fix arbitrary $a_0 \in A$. Clearly, we must have $a_0 + A = A + A$. Now, if $v_1, v_2 \in A + A$, then there exist $a_1, a_2 \in A$ such that $v_i = a_0 + a_i, i \in \{1, 2\}$. Thus $v_1 + v_2 = 2a_0 + a_1 + a_2 = a_1 + a_2 \in A + A$. Thus $A + A$, being closed under additions, and having 0, is a subspace. Consequently, $A = a_0 + (A + A)$ is a translation of a subspace, also known as an *affine subspace*. The converse is obvious. ■

Proposition 12. Let H be a subspace of $\mathbb{F}_2^n$, and let $A \subseteq H$ be an arbitrary subset of H such that $\mu(A) > \mu(H)/2$. Then $A + A = H$.

Proof. Choose an arbitrary $h \in H$. Since $\mu(A) > \mu(H)/2$, $A \cap (A + h) \neq \varnothing$. Thus there exist $a, a' \in A$ such that $a = a' + h \implies h = a + a' \implies h \in A + A \implies H \subseteq A + A \implies H = A + A$. ■

Thus, if $e(A) > 1$, A doesn't need to have any *a priori* structure, since we could take a subspace H, take an arbitrary subset A with $\mu(A)/\mu(H) = \nu > 1/2$, and we would have $e(A) = 1/\nu$.
However, can we make statements of the sort "$e(A)$ is close to $1 \implies A + A$ is a subspace"? Turns out we can: Freiman proved that if $e(A) < 3/2$, then $A + A$ is a subspace. Green-Tao further strengthened the statement to hold for $e(A) < 1.75$.
Now, what about the case when $e(A) = 1000$, i.e. $e(A)$ has a constant-factor, but large, sum-set expansion? Let us explore 3 sets in this regard:

1. A is a random subset of $\mathbb{F}_2^n$ with measure $1/1000$: Fix $a_0 \in A$. For any $v \in \mathbb{F}_2^n$, $\Pr(v - a_0 \notin A) = 0.999$, and thus $\Pr(v \notin A + A) \approx 0.999^{2^n/1000} = o_n(1)$, i.e. with high probability $A + A = \mathbb{F}_2^n$, and consequently, $e(A) = 1000$.

2. A is a subspace of $\mathbb{F}_2^n$ with co-dimension 10, i.e. $\dim(A) = n - 10$: In this case, $e(A) = 1$, even though $\mu(A) = 2^{-10} \approx 1/1000$. Just based on these two examples, one may be tempted to conjecture that "random" sets have high expansion, while "structured" sets have low expansion. The next example disproves this conjecture.

3. $A := \{x \in \mathbb{F}_2^n : \mathrm{wt}(x) \leqslant n/2 - 3\sqrt{n}/2\}$, i.e. A is the Hamming ball of radius $n/2 - 3\sqrt{n}/2$ centered at the origin, an unarguably "structured" set. Using standard Hamming weight estimates, $\mu(A) \approx 1/1000$. Also, it is easy to see that $A + A = \{x : \mathrm{wt}(x) \leqslant n - 3\sqrt{n}\}$, and thus $\mu(A + A) \approx 1$, leading to $e(A) \approx 1000$.

Unsurprisingly, it is hard to get a handle from A based on $e(A)$ alone. However, in all 3 examples, even when $A+A$ was not a subspace (as is the case when A was the Hamming ball), $A+A$ contains a large chunk of some large subspace. We formalize this observation through a conjecture.

Conjecture ($A+A$ conjecture)**.** $\mu(A) \geqslant \alpha \implies$ there exists a subspace H, of co-dimension $\mathcal{O}(\log(1/\alpha))$, such that $\mu(H \cap (A+A)) \geqslant 0.99\mu(H)$, i.e. $A+A$ contains a significant chunk of some large subspace.

This conjecture implies the following result: If $e(A) \leqslant 1/\alpha$, then $A+A+A+A$ contains a subspace H such that $\mu(H) \geqslant \text{poly}(\alpha) \cdot \mu(A)$. This statement is also known as the *polynomial Bogolyubov conjecture*.
The polynomial Bogolyubov conjecture also implies the following statement: If $e(A) \leqslant 1/\alpha$, then there exists an affine subspace $x+H$ such that $\mu(x+H) \leqslant \text{poly}(1/\alpha) \cdot \mu(A)$, and $\mu((x+H) \cap A) \geqslant \text{poly}(\alpha) \cdot \mu(A)$. This statement is also known as the ~~polynomial Freiman-Ruzsa conjecture~~: The Freiman-Ruzsa conjecture, also known as Marton's conjecture, was resolved very recently (Nov 2023) by Gowers, Green, Manners and Tao [GGMT23], which was a huge breakthrough, since Freiman-Ruzsa/Marton's conjecture was commonly considered to be the most important open problem in additive combinatorics.
The Freiman-Ruzsa-Marton ~~conjecture~~ theorem is equivalent to a rather innocent-looking property testing problem:

Conjecture. Suppose there exists a series of functions $f_m : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ such that $\Pr_{x,y \sim \mathbb{F}_2^n}(f(x)+f(y) = f(x+y)) \geqslant \varepsilon$, for every $m \in \mathbb{N}$. Then there exists a series of linear functions $g_m : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ such that f_m is $\text{poly}(\varepsilon)$-close to g_m.

The main reason a statement like this is interesting is that even as m shoots to ∞, the distance between f_m, g_m remains independent of m.
So while proving any of the above conjectures would be amazing, we now divert our attention to something that has been proven and is very interesting in its own right.

7.1. A Theorem of Sanders

Theorem 7.1. Suppose $A \subseteq \mathbb{F}_2^n$ is such that $\mu(A) = \alpha$. Then $A+A+A$ contains an affine subspace of co-dimension $\leqslant 1/\alpha$.

Proof. If $A+A+A = \mathbb{F}_2^n$, then we're done. Otherwise, choose $x \notin A+A+A$. Then $\Pr_{a,b,c \in A}(a+b+c = x) = 0$, where a, b, c are chosen independently, and uniformly, from A. Now, let $\phi_A := \mathbb{1}_A/\alpha$ be the PDF associated with A. Then $a+b+c$ is distributed according to $\varphi_A := \phi_A * \phi_A * \phi_A$, and thus $\varphi_A(x) = 0$. Now, let $\varphi_A := \sum_{\gamma \in \mathbb{F}_2^n} \widehat{\varphi_A}(\gamma)\chi_\gamma$ be the Fourier expansion of φ_A in $\mathbb{F}_2^n$, where $\chi_\gamma(x) := (-1)^{\sum_{i=1}^n \gamma_i x_i} = (-1)^{\langle \gamma, x \rangle}$. We know that $\widehat{\varphi_A}(\gamma) = \widehat{\phi_A}(\gamma)^3$, and thus

$$0 = \sum_{\gamma \in \mathbb{F}_2^n} \widehat{\phi_A}(\gamma)^3 (-1)^{\langle \gamma, x \rangle} = \widehat{\phi_A}(0)^3 + \sum_{\gamma \neq 0} \widehat{\phi_A}(\gamma)^3 (-1)^{\langle \gamma, x \rangle} \geqslant \widehat{\phi_A}(0)^3 - \sum_{\gamma \neq 0} |\widehat{\phi_A}(\gamma)|^3$$

But $\widehat{\phi_A}(0) = \mathbb{E}[\phi_A] = 1$, and thus

$$0 \geqslant 1 - \sum_{\gamma \neq 0} |\widehat{\phi_A}(\gamma)|^3 \geqslant 1 - \left(\max_{\gamma \neq 0} |\widehat{\phi_A}(\gamma)| \right) \cdot \sum_{\gamma \neq 0} \widehat{\phi_A}(\gamma)^2$$

But $\sum_{\gamma \neq 0} \widehat{\phi_A}(\gamma)^2 = \text{Var}(\phi_A) = \mathbb{E}\left[\phi_A^2\right] - \mathbb{E}\left[\phi_A\right]^2 = (1-\alpha)/\alpha$. Thus, there exists $\gamma^* \in \mathbb{F}_2^n \setminus \{0\}$ such that $|\widehat{\phi_A}(\gamma^*)| \geqslant \alpha/(1-\alpha)$. Now, note that $\widehat{\phi_A}(\gamma^*) = \mathbb{E}\left[\langle \phi_A, \chi_{\gamma^*} \rangle\right] = \mathbb{E}_{a \sim A}\left[\chi_{\gamma^*}(a)\right]$. Also, let $F_- := \{x \in \mathbb{F}_2^n : \chi_{\gamma^*}(x) = -1\} = \langle \gamma^* \rangle^\perp$, and $F_+ := \{x \in \mathbb{F}_2^n : \chi_{\gamma^*}(x) = 1\} = \mathbb{F}_2^n \setminus F_-$, and let $A_\dagger := A \cap F_\dagger$, where $\dagger \in \{-, +\}$. Then

$$\mathbb{E}_{a \sim A}\left[\chi_{\gamma^*}(a)\right] = \frac{|A_+| - |A_-|}{|A|} \implies \left|\frac{|A_+| - |A_-|}{|A|}\right| \geqslant \frac{\alpha}{1-\alpha}$$

Now WLOG assume $|A_+| \geqslant |A_-|$. Then

$$\left|\frac{|A_+| - |A_-|}{|A|}\right| = \frac{2|A_+| - |A|}{|A|} \geqslant \frac{\alpha}{1-\alpha} \implies \frac{|A_+|}{|A|} \geqslant \frac{1}{2(1-\alpha)} \implies \mu(A_+) \geqslant \frac{\alpha}{2(1-\alpha)}$$

Thus, $\mu_{F_+}(A_+) := |A_+|/|F_+| = 2\mu(A_+) \geqslant \frac{\alpha}{1-\alpha} > \alpha$.
Since $F_+ \cong \mathbb{F}_2^{n-1}$, we can set $A_1 := A_+$, and repeat the above process. Once again, we either obtain $A_1 + A_1 + A_1 \cong \mathbb{F}_2^{n-1}$, or we can "bump" up the relative density of A_1 in some $(n-2)$-dimensional (affine) subspace of F_+.
Now, note that $x/(1-x)$ is an increasing function, and furthermore,

$$\frac{\alpha/(1-t\alpha)}{1 - (\alpha/(1-t\alpha))} = \frac{\alpha}{1-(t+1)\alpha},\ \forall\ t \in \mathbb{N}$$

Thus, if the above process repeats k times, we get a subset $A_k \subseteq A$, such that the relative density of A_k in some affine subspace of co-dimension k is $\geqslant \alpha/(1-k\alpha)$. Clearly, $k \leqslant 1/\alpha$, and thus there is some $A' \subseteq A$ such that $A' + A' + A'$ contains an affine subspace of co-dimension at most $1/\alpha$, as desired. ■

Remark. This argument is an example of a *density increment argument*.

References

[Arr50] Kenneth J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950. URL: http://www.jstor.org/stable/1828886.

[BLR90] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. In *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*, STOC '90, pages 73–83, New York, NY, USA, 1990. Association for Computing Machinery. doi:10.1145/100216.100225.

[FK96] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. 1996. URL: https://api.semanticscholar.org/CorpusID:9721337.

[GGMT23] W. T. Gowers, Ben Green, Freddie Manners, and Terence Tao. On a conjecture of marton, 2023. arXiv:2311.05762.

[H01] Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, jul 2001. doi:10.1145/502090.502098.

[Kal02] Gil Kalai. A fourier-theoretic perspective on the condorcet paradox and arrow's theorem. *Advances in Applied Mathematics*, 29(3):412–426, 2002. URL: https://www.sciencedirect.com/science/article/pii/S0196885802000234, doi:https://doi.org/10.1016/S0196-8858(02)00023-4.

[KKL88] Jeff D. Kahn, Gil Kalai, and Nathan Linial. The influence of variables on boolean functions. [*Proceedings 1988*] *29th Annual Symposium on Foundations of Computer Science*, pages 68–80, 1988. URL: https://api.semanticscholar.org/CorpusID:2347606.

[KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? *SIAM Journal on Computing*, 37(1):319–357, 2007. arXiv:https://doi.org/10.1137/S0097539705447372, doi:10.1137/S0097539705447372.

[O'D12] Ryan O'Donnell. Analysis of boolean functions at cmu. https://www.youtube.com/playlist?list=PLm3J0oaFux3YypJNaF6sRAf2zC1QzMuTA, 2012. Accessed 3rd December, 2023.

[O'D21] Ryan O'Donnell. Analysis of boolean functions. *CoRR*, abs/2105.10386, 2021. URL: https://arxiv.org/abs/2105.10386, arXiv:2105.10386.

[Tal94] Michel Talagrand. On Russo's Approximate Zero-One Law. *The Annals of Probability*, 22(3):1576 – 1587, 1994. doi:10.1214/aop/1176988612.