These are some concise lecture summaries of the course CS229BR – Analysis of Boolean Functions taught at Harvard University in Fall 2022. If you find any mistakes please feel free to email chinho.lee@ncsu.edu.

Contents

1.1 Introduction

This course is about Boolean functions, functions of the form

$$f\colon \{0,1\}^n \to \{0,1\}.$$

Here are some examples where Boolean functions appear:

1. Computational complexity: Given $f\colon \{0,1\}^n \to \{0,1\}$, we can study the computational resources needed to compute f. For example, we can ask whether f can be computed by a circuit of a certain size, or a communication protocol that exchanges a certain amount of information.

2. Graph property: Given a graph $G = (V, E)$, we want to know if G satisfies some graph property, e.g. if G contains a 10-clique. We can think of it as a Boolean function, where the number of input bits is the $n := \binom{|V|}{2}$ many possible edges in G, and each input bit is an indicator variable $e_{u,v}$ representing whether the edge between vertices u and v are present in G, i.e.
$$e_{u,v} = \begin{cases} 1 & \text{if } e_{u,v} \in E \\ 0 & \text{if } e_{u,v} \notin E. \end{cases}$$
In our example, one can define $f\colon \{0,1\}^n \to \{0,1\}$ as
$$f(e_{1,1}, e_{1,2}, \ldots, e_{|V|,|V|-1}) = \begin{cases} 1 & \text{if } G \text{ contains a 10-clique} \\ 0 & \text{otherwise.} \end{cases}$$

3. Learning theory: Here we think of $x_1, \ldots, x_n$ as n binary attributes, and f is a concept.

4. Social choice theory: There are 2 candidates, represented by 0 and 1, and there are n voters $x_1, \ldots, x_n$, and $f\colon \{0,1\}^n \to \{0,1\}$ can be viewed as a voting rule.

1.2 Notation

We will typically identify $\{0,1\}$ with the field $\mathbb{F}_2$ (and $\{0,1\}^n$ with $\mathbb{F}_2^n$). In particular, addition is over $\mathbb{F}_2$, which is the same as XOR. For a finite set X, we use $\boldsymbol{x} \sim X$ to denote the random variable $\boldsymbol{x}$ sampled uniformly from the set X, that is, $\mathbf{Pr}[\boldsymbol{x} = x] = \frac{1}{|X|}$ for every $x \in X$.

1.3 Fourier expansion

Proposition 1.1. *Every $f\colon \{0,1\}^n \to \mathbb{R}$ has a degree-n multilinear polynomial representation.*

Proof. This follows from interpolation. For each $a \in \{0,1\}^n$, we can write the point function $\mathbb{1}_a\colon \{0,1\}^n \to \{0,1\}$ defined by
$$\mathbb{1}_a(x) := \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a. \end{cases}$$
as a degree-n multilinear polynomial. Now write $f(x) = \sum_a f(a)\mathbb{1}_a(x)$. □

We will often work with the domain $\{-1,1\}^n$. By a simple change of variable we have the following corollary.

Corollary 1.2. *Every $f\colon \{-1,1\}^n \to \mathbb{R}$ has a degree-n multilinear polynomial representation.*

Proof. Given $a_i, x_i \in \{-1,1\}$, consider the mapping $(1 + a_i x_i)/2 \in \{0,1\}$. □

Switching between $\{0,1\}$ vs. $\{-1,1\}$. In this course we will always switch between $\{0,1\}$ and $\{-1,1\}$, and identify $1 \in \{0,1\}$ and $-1 \in \{-1,1\}$ with "True", and $0 \in \{0,1\}$ and $1 \in \{-1,1\}$ as "False". The following transformations are some convenient ways of switching between both domains.

1. $\{0,1\} \ni x \mapsto (-1)^x \in \{-1,1\}$
2. $\{0,1\} \ni x \mapsto 1 - 2x \in \{-1,1\}$ and conversely $\{-1,1\} \ni y \mapsto (1-x)/2 \in \{0,1\}$.

Let us prove the following standard fact to illustrate how these transformations can be useful.

Fact 1.3. $\mathbf{Pr}_{\boldsymbol{x}\sim\{0,1\}^n}[\sum_{i=1}^n \boldsymbol{x}_i = 1] = 1/2$.

(Recall that we identify $\{0,1\}^n$ with $\mathbb{F}_2^n$, so $\sum_i$ is the same as $\bigoplus_i$.)

Proof. Applying both transformations, for every $x \in \{0,1\}^n$ we have

$$\sum_{i=1}^n x_i = \frac{1 - \prod_{i=1}^n (-1)^{x_i}}{2}.$$

Now we take expectation on both sides, and use the independence of the $\boldsymbol{x}_i$'s and $\mathbf{E}_{\boldsymbol{x}_i\sim\{0,1\}}[(-1)^{\boldsymbol{x}_i}] = 0$. □

Another equivalent way of viewing $f\colon \{0,1\}^n \to \mathbb{R}$ is to think of it as a 2^n-dimensional vector in $\mathbb{R}^{\{0,1\}^n}$, where coordinates are indexed by $x \in \{0,1\}^n$, and the x-th coordinate of f is $f(x)$, in fact you can see that whether the domain is $\{0,1\}^n$ or $\{-1,1\}^n$ does not matter: it is just a different naming of the indices. Under this view, note that $\mathbb{1}_a : a \in \{0,1\}^n$ are simply the elementary basis, i.e. 1 at position a and 0 everywhere else. (We will soon see that the monomials of $\{-1,1\}^n$ also form another basis.) This linear-algebraic view motivates the following definition of *inner product* of two functions.

Definition 1.4. Given $f, g\colon \{-1,1\}^n \to \mathbb{R}$, we define the *inner product* of f and g by

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

The normalization 2^{-n} may seem odd at first, but it will soon be appreciated.

1.3.1 The parity functions

We will define $\chi_S\colon \{-1,1\}^n \to \{-1,1\}$ by

$$\chi_S(x) := \prod_{i\in S} x_i.$$

For ease of notation, we will also use x^S to denote the monomial $\prod_{i\in S} x_i$.

We will overload χ_S when the domain is $\{0,1\}^n \equiv \mathbb{F}_2^n$ and define $\chi_S \colon \{0,1\}^n \to \{0,1\}$ as

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

(These are also called the characters of $\mathbb{F}_2^n$.) We will often use the well-known 1-1 correspondence between a subset $S \subseteq [n]$ and an n-bit string $\alpha \in \{0,1\}^n$, where $\alpha_i = 1$ if and only if $i \in S$ and 0 otherwise. In that case, we have

$$\chi_\alpha(x) := (-1)^{\langle \alpha, x \rangle}.$$

For two subsets $S, T \subseteq [n]$, we sometimes use $S + T$ to denote their symmetric difference $S \triangle T$. Because inner product is linear in its arguments, we have

Proposition 1.5. *Let $\chi_S \colon \{0,1\}^n \to \{-1,1\}$ be a parity function. We have*

1. $\chi_{\alpha+\beta}(x) = \chi_\alpha(x) \cdot \chi_\beta(x)$ *for every* $\alpha, \beta \in \{0,1\}^n$.
2. $\chi_\alpha(x+y) = \chi_\alpha(x) \cdot \chi_\alpha(y)$ *for every* $x, y \in \{0,1\}^n$.

We now show that the representation in Corollary 1.2 is unique.

Proposition 1.6. *The 2^n parity functions $\{\chi_S\}_{S \subseteq [n]}$ are pairwise orthonormal, that is*

$$\langle \chi_S, \chi_T \rangle = \begin{cases} 1 & \text{if } S = T \\ 0 & \text{if } S \neq T. \end{cases}.$$

(Verify that this proposition indeed holds regardless of the domain of χ_S being $\{0,1\}^n$ or $\{-1,1\}^n$.)

Theorem 1.7. *Every $f \colon \{-1,1\}^n \to \mathbb{R}$ has a* unique *degree-n multilinear polynomial presentation*

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

Proof. We will show that the parity functions $\{\chi_S : S \subseteq [n]\}$ form a basis of the vector space of all functions of the form $f \colon \{-1,1\}^n \to \mathbb{R}$.

In Corollary 1.2, we showed that every $f \colon \{-1,1\}^n \to \mathbb{R}$ has a degree-n multilinear polynomial presentation. When the domain is $\{-1,1\}^n$, the monomials $\prod_{i \in S} x_i$ are actually the parity functions $\chi_S(x)$. This shows the 2^n parity functions span the space of $\{f \colon \{-1,1\}^n \to \mathbb{R}\}$.

It remains to show that $\{\chi_S : S \subseteq [n]\}$ are independent, this follows from their pairwise orthonormality in Proposition 1.6. □

The coefficients $\widehat{f}(S)$ are called the Fourier coefficients of f.

2.1 Basic Identities

Thanks to the orthonormality of the parity functions, there is an explicit formula for computing $\widehat{f}(S)$.

Proposition 2.1 (Fourier inversion formula)**.** *Suppose* $f\colon \{-1,1\}^n \to \{-1,1\}$ *has the Fourier expansion* $f(x) = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S(x)$*. Then* $\widehat{f}(S) := \langle f, \chi_S\rangle = \mathbf{E}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})]$.

Example 2.2. Compute the Fourier expansion of $\max_2\colon \{-1,1\}^2 \to \{-1,1\}$. (Note that this is just the AND_2 function in disguise.)

We now show some basic identities.

Proposition 2.3 (Parseval's identity)**.** *Let* $f\colon \{-1,1\}^n \to \mathbb{R}$*. Then* $\|f\|_2^2 := \langle f, f\rangle = \mathbf{E}[f(\boldsymbol{x})^2] = \sum_{S\subseteq[n]} \widehat{f}(S)^2$.

Note that when $f\colon \{-1,1\}^n \to \{-1,1\}$, we have $1 = \mathbf{E}[f(\boldsymbol{x})^2] = \sum_{S\subseteq[n]} \widehat{f}(S)^2$. A more general identity is the following.

Proposition 2.4 (Plancherel's identity)**.** *Let* $f, g\colon \{-1,1\}^n \to \mathbb{R}$*. Then* $\langle f, g\rangle = \mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{x})] = \sum_{S\subseteq[n]} \widehat{f}(S)\widehat{g}(S)$.

Remark 2.5. The quantity $\langle f, g\rangle$ is sometimes called the *correlation* of f and g. To see this, observe that for $f, g\colon \{-1,1\}^n \to \{-1,1\}$, we can write

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x})g(\boldsymbol{x})] &= \mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})] - \mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \\
&= \mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})] + \Big(\mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] - \mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})]\Big) - \mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \\
&= 1 - 2\,\mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \\
&= 2\,\mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})] - 1.
\end{aligned}$$

The last inequality can be derived by replacing $(\mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] - \mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})])$ with $(\mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})] - \mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})])$. By rearranging we get

$$\mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})] = \frac{1}{2} + \frac{\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{x})]}{2}.$$

In complexity theory, we often ask the following question. Given a family $\mathcal{F}$ of Boolean functions, is there a Boolean function g that is hard for $\mathcal{F}$ on average? That is, we want to minimize

$$\max_{f\in F} \mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})].$$

When $\mathcal{F}$ contains both constant functions $x \mapsto 1$ and $x \mapsto -1$, one can see that $1/2$ is always achievable. Thus sometimes it is easier to bound $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{x})]$ instead.

To state the next identity, we define the *convolution* of two functions. We define it in $\{0,1\}$-notion as it is more commonly defined. Again, one can define a $\{-1,1\}$ analogue by replacing addition over $\mathbb{F}_2$ with multiplication.

Definition 2.6. Given $f, g\colon \mathbb{F}_2^n \to \mathbb{R}$, the *convolution* of f and g, denoted $f * g\colon \mathbb{F}_2^n \to \mathbb{R}$, is

$$f * g(x) := \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\mathbb{F}_2^n}[f(x+\boldsymbol{z})g(\boldsymbol{z})].$$

To give some intuition, suppose $f, g\colon \{0,1\}^n \to [0,1]$ are two probability mass functions on $\{0,1\}^n$. Let $\boldsymbol{F}$ and $\boldsymbol{G}$ be two independent random variables sampled according to f and g, respectively. If we consider the random variable $\boldsymbol{F}+\boldsymbol{G}$ (over $\mathbb{F}_2^n$), then we have

$$\begin{aligned}\mathbf{Pr}[\boldsymbol{F}+\boldsymbol{G}=x] &= \sum_{z\in\{0,1\}^n} \mathbf{Pr}[\boldsymbol{F}=x+z]\,\mathbf{Pr}[\boldsymbol{G}=z] \\ &= 2^n \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\{0,1\}^n}[f(x+\boldsymbol{z})g(\boldsymbol{z})].\end{aligned}$$

Proposition 2.7 (Convolution)**.** $\widehat{f*g}(S) := \widehat{f}(S)\widehat{g}(S)$.

Proposition 2.8 (Mean)**.** $\mathbf{E}[f(x)] = \widehat{f}(\varnothing)$.

The meaning of $\mathbf{E}[f(\boldsymbol{x})]$ can be very different depending on whether f has outputs $\{0,1\}$ or $\{-1,1\}$. For instance, for $f\colon \{-1,1\}^n \to \{-1,1\}$, we have

$$\mathbf{E}[f(\boldsymbol{x})] = \mathbf{Pr}[f(\boldsymbol{x})=1] - \mathbf{Pr}[f(\boldsymbol{x})=-1] = 2\,\mathbf{Pr}[f(\boldsymbol{x})=1] - 1 = 1 - 2\,\mathbf{Pr}[f(\boldsymbol{x})=-1].$$

On the other hand, for $f\colon \{-1,1\}^n \to \{0,1\}$, we have

$$\mathbf{E}[f(\boldsymbol{x})] = \mathbf{E}[f(\boldsymbol{x})^2] = \mathbf{Pr}[f(\boldsymbol{x})=1].$$

Switching between $\{0,1\}$ and $\{-1,1\}$ outputs will be important and useful throughout the course.

Proposition 2.9 (Variance)**.** $\mathbf{Var}[f] = \sum_{\varnothing\neq S\subseteq[n]} \widehat{f}(S)^2$.

Proof. Recall that $\mathbf{Var}[f] := \mathbf{E}_{\boldsymbol{x}}[(f(\boldsymbol{x})-\mathbf{E}[f(\boldsymbol{x})])^2] = \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})^2] - \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})]^2$. Apply Propositions 2.3 and 2.8. □

Note that for $f\colon \{-1,1\}^n \to \{-1,1\}$, we have

$$\begin{aligned}\mathbf{Var}[f] &= 1 - \mathbf{E}[f]^2 \\ &= (1+\mathbf{E}[f])(1-\mathbf{E}[f]) \\ &= \big(1 + (2\,\mathbf{Pr}[f(\boldsymbol{x})=1]-1)\big)\big(1-(1-2\,\mathbf{Pr}[f(\boldsymbol{x})=-1])\big) \\ &= 4\,\mathbf{Pr}[f(\boldsymbol{x})=1]\,\mathbf{Pr}[f(\boldsymbol{x})=-1].\end{aligned}$$

2.2 Linearity Testing

Definition 2.10. A function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ is *linear* if

$$f(x) = \sum_{i\in S} x_i \text{ for some } S \subseteq [n]. \tag{1}$$

We now show that Equation (1) is equivalent to

$$f(x+y) = f(x) + f(y) \text{ for every } x, y \in \mathbb{F}_2^n. \tag{2}$$

Equation (2) follows easily from Equation (1). To see the other direction (that (2) implies (1)), first observe that $f(\vec{0}) = f(\vec{0}+\vec{0}) = f(\vec{0})+f(\vec{0})$ and thus $f(\vec{0}) = 0$. Now for each $i \in [n]$, define $a_i := f(e_i)$, and we can prove by induction (using (2)) on the number of 1s in x that $f(x) = \langle a, x\rangle = \sum_{i=1}^{n} a_i x_i$.

Given blackbox access to a Boolean function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$, that is, we are allowed to make queries at $x \in \mathbb{F}_2^n$ and is given $f(x)$ in return. How many queries do we need to determine if f is linear?

It is not hard to see that any correct deterministic algorithm for linearity testing requires 2^n queries. So we will consider a relaxation the problem, which is testing if f is *close* to linear. One can define various distance measures of closeness, we will use the following:

Definition 2.11. Two functions $f, g\colon \{0,1\}^n \to \{0,1\}$ are ϵ-close if $\mathbf{Pr}_{\boldsymbol{x}\sim\{0,1\}^n}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] := 2^{-n}\sum_{x\in\{0,1\}^n} \mathbb{1}(f(x) \neq g(x)) \leq \epsilon$.

In other words, f and g disagrees on ϵ fraction of the points in $\{0,1\}^n$. Now we can define the following generalization of Equation (1) of linearity.

Definition 2.12. A function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ is ϵ-close to linear if $\mathbf{Pr}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[f(\boldsymbol{x}) \neq \sum_{i\in S} \boldsymbol{x}_i] \leq \epsilon$ for some $S \subseteq [n]$.

Do we have the same equivalence between (1)*and* (2) for our relaxed notion of linearity? The answer is yes, and is given by the following result of Blum, Luby, and Rubinfeld.

Theorem 2.13 (Robust characterization of linearity)**.** *Let $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ be a Boolean function. Then*

1. *if f is ϵ-close to linear, then $\mathbf{Pr}_{\boldsymbol{x},\boldsymbol{y}\sim\mathbb{F}_2^n}[f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})] \geq 1 - 3\epsilon$, and*
2. *if f is* not *ϵ-close to linear, then $\mathbf{Pr}_{\boldsymbol{x},\boldsymbol{y}\sim\mathbb{F}_2^n}[f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})] \leq 1 - \epsilon$.*

Item 1 simply follows from the union bound, and we now prove the other direction.

2.2.1 Proof of Item 2 in Theorem 2.13

As in Fact 1.3, we will work over $\{-1,1\}$ instead of $\{0,1\}$. Define $F\colon \mathbb{F}_2^n \to \{-1,1\}$ by $F(x) = (-1)^{f(x)}$. Observe that for every $x, y \in \mathbb{F}_2^n$,

$$\begin{aligned} f(x+y) = f(x) + f(y) &\iff f(x) + f(y) + f(x+y) = 0 \\ &\iff F(x)F(y)F(x+y) = 1 \\ &\iff \frac{1}{2} + \frac{F(x)F(y)F(x+y)}{2} = 1 \end{aligned}$$

So taking expectation on both sides,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x},\boldsymbol{y}\sim\mathbb{F}_2^n}[f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})] = \frac{1}{2} + \frac{1}{2}\mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\mathbb{F}_2^n}[F(\boldsymbol{x})F(\boldsymbol{y})F(\boldsymbol{x}+\boldsymbol{y})] \tag{3}$$

The expectation on the right hand side has a very simple expression in the Fourier coefficients of F.

Claim 2.14. $\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathbb{F}_2^n}[F(\boldsymbol{x})F(\boldsymbol{y})F(\boldsymbol{x}+\boldsymbol{y})] = \sum_{S\subseteq[n]} \widehat{F}(S)^3$.

Proof. Replace $F(\boldsymbol{y})$ and $F(\boldsymbol{x}+\boldsymbol{y})$ with their Fourier expansion. □

Now recall that $\widehat{F}(S) = \langle F, \chi_S \rangle = \mathbf{E}[F(\boldsymbol{x})\chi_S(\boldsymbol{x})]$ is the correlation between F and some parity function, which is equivalent to the distance between F and χ_S,

$$\widehat{F}(S) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\big[F(\boldsymbol{x})\chi_S(\boldsymbol{x})\big] = 1 - 2\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}\big[F(\boldsymbol{x}) \neq \chi_S(\boldsymbol{x})\big] = 1 - 2\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}\Big[f(\boldsymbol{x}) \neq \sum_{i\in S}\boldsymbol{x}_i\Big].$$

Since f is *not* ϵ-close to linear, the probability on the right hand side is at least ϵ for every $S \subseteq [n]$. Hence, we have $\widehat{F}(S) \leq 1 - 2\epsilon$ for every S. Recall that $\sum_{S\subseteq[n]} \widehat{F}(S)^2 = 1$, and so

$$\sum_{S\subseteq[n]} \widehat{F}(S)^3 \leq \max_{S\subseteq[n]} \widehat{f}(S) \cdot \sum_{S\subseteq[n]} \widehat{f}(S)^2 = \max_{S\subseteq[n]} \widehat{f}(S) \leq 1 - 2\epsilon. \tag{4}$$

Plugging Equation (4) into Equation (3) completes the proof. □

2.2.2 The BLR algorithm for linearity testing

Theorem 2.13 suggests the following 3-query randomized algorithm for testing ϵ-closeness of linearity:

1. Sample $\boldsymbol{x}, \boldsymbol{y} \sim \mathbb{F}_2^n$ uniformly at random.
2. Query $f(\boldsymbol{x}), f(\boldsymbol{y})$, and $f(\boldsymbol{x} + \boldsymbol{y})$.
3. If $f(\boldsymbol{x}) + f(\boldsymbol{y}) = f(\boldsymbol{x} + \boldsymbol{y})$, accept. Otherwise, reject.

The correctness of this algorithm is an immediate corollary of Theorem 2.13

Corollary 2.15. *The BLR algorithm satisfies the following property:*

1. *(Completeness) If f is linear, then the algorithm always accepts.*
2. *(Soundness) If f is ϵ-far (i.e., not ϵ-close) to linear, then the algorithm accepts with probability $1 - \epsilon$.*

By repeating the algorithm independently for $O(\log(1/\delta)/\epsilon)$ times, we can bring down the acceptance probability in the soundness to δ.

3.1 Social choice

We will study several important concepts in Boolean function analysis. We will introduce these concepts through the lens of social theory theory. We will work with $f\colon \{-1,1\}^n \to \{-1,1\}$, where we think of the inputs $x_1, \dots, x_n$ as n voters and f as some voting rule determining the outcome of an election with 2 candidates -1 and 1.

3.1.1 Examples of voting rules

First let us consider some common examples of voting rules.

1. The *majority* function is defined as $\mathsf{Maj}(x_1, \dots, x_n) = \operatorname{sgn}(x_1 + \cdots + x_n)$. To avoid ties, we will typically assume n is odd.

2. The AND function is defined as $\mathsf{AND}_n(x_1, \dots, x_n) := \begin{cases} -1 & \text{if } x_1 = \cdots = x_n = -1 \\ 1 & \text{otherwise.} \end{cases}$

3. The OR function is defined as $\mathsf{OR}_n(x_1, \dots, x_n) := \begin{cases} 1 & \text{if } x_1 = \cdots = x_n = 1 \\ -1 & \text{otherwise.} \end{cases}$

4. A *dictator* is defined as $\chi_{\{i\}}(x_1, \dots, x_n) = x_i$ for some $i \in [n]$.

5. A *k-junta* for $k \in [n]$ is a function f only depends on $\le k$ of its n coordinates, i.e. there exist coordinates $i_1 < \cdots i_k \in [n]$ and a function $g\colon \{-1,1\}^k \to \{-1,1\}$ such that

$$f(x_1, \dots, x_n) = g(x_{i_1}, \dots, x_{i_k}).$$

6. The Tribes function is parametrized by its width w and its number of tribes s. It outputs -1 if all members in a tribe agrees on voting for -1, and 1 otherwise. (However, it may not output 1 even if all members in a tribe votes for 1.) Specifically, $\mathsf{Tribes}_{w,s}\colon \{-1,1\}^{w \cdot s} \to \{-1,1\}$ is defined as

$$\mathsf{Tribes}_{w,s}(x_{1,1}, \dots, x_{s,w}) := \bigvee_{i=1}^{s} \bigwedge_{j=1}^{w} x_{i,j},$$

   where we use $\wedge_{i=1}^{s} x_i$ to denote $\mathsf{AND}_s(x)$ and $\vee_{j=1}^{w} x_j$ to denote $\mathsf{OR}_w(x)$. Note that $\mathsf{Tribes}_{w,s}$ is a read-once width w DNF formula with s terms[1], where read-once means that every input variable appears in the formula at most once.

   Given a width w, we typically want to choose $s = s(w)$ so that $\mathsf{Tribes}_{w,s}$ is close to balanced, i.e. $\mathbf{Pr}_{\boldsymbol{x}}[\mathsf{Tribes}_{w,s}(\boldsymbol{x})] \approx 1/2$, and so we can take $s \approx 2^w \ln 2$.

[1] Sometimes the Tribes function is defined as a CNF, i.e., AND_s composed with OR_w on disjoint variables.

	not a junta	symmetric	monotone	unanimous	odd
Maj	YES	YES	YES	YES	YES
AND_n	YES	YES	YES	YES	NO
OR_n	YES	YES	YES	YES	NO
dictator	NO	NO	YES	YES	YES
k-junta	NO	NO	NO	NO	NO
Tribes	YES	NO	YES	YES	NO

Figure 1: Properties that are satisfied by the voting rules we considered.

3.1.2 Properties of voting schemes

Let us consider some desirable properties in a voting scheme.

1. f is *not a junta* if every coordinate $i \in [n]$ matters. Specifically, for every $i \in [n]$ we can find an input $x \in \{-1,1\}^n$ such that flipping its input bit changes the outcome of f, i.e. $f(x) \neq f(x^{\oplus i})$, where we use $x^{\oplus i}$ to denote $(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_n)$.

2. f is *symmetric* if the identities of the voters do not matter. Specifically, for every permutation $\pi \colon [n] \to [n]$, we have $f(x) = f(x^\pi) := f(x_{\pi(1)}, \ldots, x_{\pi(n)})$. One can verify that f only depends on $\sum_{i=1}^n x_i$, or the Hamming weight of x (after switching the value of the x_i's to $\{0,1\}$).

3. f is *monotone* if switching a bit x_i of x from -1 to 1 can change $f(x)$ to 1. Specifically, if we define the partial ordering on $\{-1,1\}^n$ by $x \leq y$ if and only if $x_i \leq y_i$ for every $i \in [n]$, then $x \leq y$ implies $f(x) \leq f(y)$.

4. f is *unanimous* if $f(1, \ldots, 1) = 1$ and $f(-1, \ldots, -1) = -1$.

5. f is *odd* if $f(-x) = -f(x)$.

Figure 1 shows which properties each of the above examples satisfy. (Verify the May's theorem, which says that the majority function is the only function that is symmetric, monotone, unanimous, and odd.) The Tribes function is monotone because the composition of monotone functions is again monotone. We can see that the $\mathsf{Tribes}_{2,2}$ is not symmetric by considering the input $x = (1, -1, 1, -1)$. However, it satisfies a weaker notion of symmetry called *transitive-symmetric*. Intuitively, a function is transitive-symmetric if every two coordinates $i, j \in [n]$ are equivalent.

Definition 3.1. f is *transitive-symmetric* if for every two coordinates $i, j \in [n]$, there exists a permutation $\pi \colon [n] \to [n]$ with $\pi(i) = j$ such that $f(x^\pi) = f(x)$ for every $x \in \{-1,1\}^n$.

Proposition 3.2. $\mathsf{Tribes}_{w,s}$ *is transitive-symmetric for every integer* w *and* s.

Proof. At a high level, this follows from AND and OR being symmetric. Given two coordinates $(i_1, j_1), (i_2, j_2) \in [s] \times [w]$, we construct a permutation $\pi \colon [s] \times [w] \to [s] \times [w]$ that permutes the tribes and the members in every tribes separately. Specifically, consider a permutation $\pi_1 \colon [s] \to [s]$ that swaps the two tribe indices i_1 and i_2, and a permutation $\pi_2 \colon [w] \to [w]$ that swaps j_1 and j_2 (in all tribes). We define $\pi(i, j) := (\pi_1(i), \pi_2(j))$. By definition, we have $\pi((i_1, j_1)) = (\pi(i_1), \pi(j_1)) = (i_2, j_2)$. Now, for every $x \in \{-1,1\}^n$,

$$f(x^\pi) = \bigvee_{i=1}^{w} \bigwedge_{j=1}^{s} x_{\pi(i),\pi(j)} = \bigvee_{i=1}^{w} \bigwedge_{j=1}^{s} x_{\pi(i),j} = \bigvee_{i=1}^{w} \bigwedge_{j=1}^{s} x_{i,j} = f(x),$$

where the second equality follows from the symmetry of AND and the third follows from the symmetry of OR. □

4.1 Influence

We will study the *influence* of a function f, one of the most fundamental concept in the analysis of Boolean functions.

Definition 4.1. A coordinate $i \in [n]$ is *pivotal* on x if $f(x) \neq f(x^{\oplus i})$.

The *i-th influence* of a Boolean function is the probability that the i-th voter can change the outcome when the others vote uniformly, that is, the probability of the i-th coordinate being pivotal on a uniform random $\boldsymbol{x} \sim \{-1,1\}^n$.

Definition 4.2 (i-th influence)**.** The *i-th influence* of f, denoted by $\mathsf{Inf}_i[f]$, is defined as

$$\mathsf{Inf}_i[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}\big[f(\boldsymbol{x}^{\oplus i}) \neq f(\boldsymbol{x})\big].$$

Every $f\colon \{-1,1\}^n \to \{-1,1\}$ induces a 2-coloring on the vertices in the Boolean hypercube $\{-1,1\}^n$ by coloring each x with $f(x)$. We can partition the edges in the hypercube $\{-1,1\}^n$ into n subsets $E_1, \ldots, E_n$ of equal size, where the edges in $E_i := \{(x, x^{\oplus i}) : x \in \{-1,1\}^n\}$ are in the i-th direction of the hypercube. Each E_i divides $\{-1,1\}^n$ into two equal halves $\{x \in \{-1,1\}^n : x_i = -1\}$ and $\{x \in \{-1,1\}^n : x_i = 1\}$. We call an edge $(x, x^{\oplus i})$ pivotal if its two endpoints are colored differently, i.e., $f(x^{\oplus i}) \neq f(x)$. (Consider coloring $\{-1,1\}^3$ with Maj_3.)

Fact 4.3. $\mathsf{Inf}_i[f] = \frac{\#\text{ pivotal edges in the } i\text{-th direction}}{\#\text{ edges in the } i\text{-th direction}} = \frac{\#\text{ pivotal edges in the } i\text{-th direction}}{2^{n-1}}$.

Some examples. We now look at the i-th influence of some Boolean functions.

1. The two constant functions $f(x) \equiv -1$ and $f(x) \equiv 1$ have $\mathsf{Inf}_i[f] = 0$ for every $i \in [n]$.
2. The dictators satisfy $\mathsf{Inf}_j[\chi_{\{i\}}] = 1$ if $i = j$ and 0 otherwise.
3. The majority of 3 bits Maj_3 satisfies $\mathsf{Inf}_1[f] = \mathsf{Inf}_2[f] = \mathsf{Inf}_3[f] = 1/2$.
4. For the majority of n bits, the i-th coordinate is pivotal if the rest of the $n-1$ bits contain equal number of 1s and -1s. So

$$\mathsf{Inf}_i[\mathsf{Maj}_n] = \mathbf{Pr}\big[\mathsf{Bin}(n-1,1/2) = (n-1)/2\big] = 2^{-(n-1)}\binom{n-1}{(n-1)/2} = \Theta(1/\sqrt{n}).$$

Is there a function whose individual influences are smaller than $\Theta(1/\sqrt{n})$? If turns out the Tribes function is one such example.

Fact 4.4. $\mathsf{Inf}_{(1,1)}[\mathsf{Tribes}_{w,2^w \ln 2}] = \Theta((\log n)/n)$, *where* $n = w \cdot 2^w \ln 2$ *is the number of input bits.*

Proof. The coordinate $(1,1)$ is pivotal when the AND of all but the first tribes output "False = -1", and all the rest of the bits in the first tribe are "False = -1". Thus,

$$\mathsf{Inf}_{(1,1)}[\mathsf{Tribes}_{w,2^w \ln 2}] = (1-2^{-w})^{s-1} \cdot 2^{-(w-1)} \approx 2^{-w} = \Theta((\log n)/n). \qquad \square$$

What about the influence of the other coordinates? It turns out every coordinate $i \in [n]$ has the same i-th influence as long as f is transitive-symmetric. This captures the intuition of transitive-symmetric that every $i, j \in [n]$ are "equivalent".

Proposition 4.5. *If f is transitive-symmetric, then $\mathsf{Inf}_i[f] = \mathsf{Inf}_j[f]$ for every $i, j \in [n]$.*

Proof. Given $i \neq j$, let $\pi \colon [n] \to [n]$ be the permutation with $\pi(i) = j$ and that $f(x^\pi) = f(x)$ for every $x \in \{-1,1\}^n$. Then

$$\begin{aligned}\mathsf{Inf}_i[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})] &= \mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}^\pi) \neq f((\boldsymbol{x}^\pi)^{\oplus \pi(i)})\big] \\ &= \mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}^\pi) \neq f((\boldsymbol{x}^\pi)^{\oplus j})\big] \\ &= \mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}) \neq f((\boldsymbol{x})^{\oplus j})\big] = \mathsf{Inf}_j[f],\end{aligned}$$

where the second last equality is because $\boldsymbol{x}$ and $\boldsymbol{x}^\pi$ are identically distributed. □

Therefore, we have $\mathsf{Inf}_{(i,j)}[\mathsf{Tribes}w, 2^w \ln 2] = \Theta((\log n)/n)$ for every $(i,j) \in [s] \times [w]$. The following theorem by Kahn, Kalai, and Linial says that this is actually best possible.

Theorem 4.6 (KKL theorem)**.** $\max_{i \in [n]} \mathsf{Inf}_i[f] \geq \Theta\left(\frac{\log n}{n}\right) \cdot \mathbf{Var}[f]$ *for every* $f \colon \{-1,1\}^n \to \{-1,1\}$.

We will prove Theorem 4.6 later when we cover Hypercontractivity.

4.1.1 Formula for Influences

We now express the i-th influence of f in terms of the Fourier coefficients of f. We first introduce the derivative operator $D_i f$. Then we will relate $\mathsf{Inf}_i[f]$ to $D_i f$, and relate $D_i f$ to the coefficients $\widehat{f}(S)$.

For $b \in \{-1,1\}$, let $x^{i \mapsto b} := (x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n)$.

Definition 4.7 (Derivative Operator)**.** The *i-th derivative operator* D_i on $f \colon \{-1,1\}^n \to \mathbb{R}$ is the function $D_i f \colon \{-1,1\}^n \to \mathbb{R}$ defined by

$$D_i f(x) = \frac{f(x^{i \mapsto 1}) - f(x^{i \mapsto -1})}{2} = \mathop{\mathbf{E}}_{\boldsymbol{y}_i \sim \{-1,1\}}[f(x_1, \ldots, x_{i-1}, \boldsymbol{y}_i, x_{i+1}, \ldots, x_n)\boldsymbol{y}_i].$$

Remark 4.8. Observe that actually $D_i f$ does not depend on x_i and is only a function of $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

In particular, for $f \colon \{-1,1\}^n \to \{-1,1\}$ we have

$$D_i f(x) = \begin{cases} \pm 1 & \text{if } f(x^{i \mapsto 1}) \neq f(x^{i \mapsto -1}) \\ 0 & \text{if } f(x^{i \mapsto 1}) = f(x^{i \mapsto -1}). \end{cases} \tag{5}$$

Therefore we have the following proposition.

Proposition 4.9. *For $f \colon \{-1,1\}^n \to \{-1,1\}$, we have $\mathsf{Inf}_i[f] = \mathbf{E}_{\boldsymbol{x}}[|D_i f(\boldsymbol{x})|] = \mathbf{E}_{\boldsymbol{x}}[(D_i f(\boldsymbol{x}))^2]$.*

We will use $\mathbf{E}_{\boldsymbol{x}}[(D_i f(\boldsymbol{x}))^2]$ as our definition of $\mathsf{Inf}_i[f]$ for real-valued functions f.

Definition 4.10 (i-influence of real-valued functions)**.** The *i-th influence* of $f \colon \{-1,1\}^n \to \mathbb{R}$, denoted $\mathsf{Inf}_i[f]$, is defined as $\mathsf{Inf}_i[f] := \mathbf{E}_{\boldsymbol{x}}[(D_i f(\boldsymbol{x}))^2]$.

We now look at the Fourier expansion $D_i f\colon \{-1,1\}^n \to \mathbb{R}$.

Proposition 4.11. *For $f\colon \{-1,1\}^n \to \mathbb{R}$, we have $D_i f(x) = \sum_{S \ni i} \widehat{f}(S) x^{S\setminus\{i\}}$.*

Proof. First show that D_i is a linear operator, that is, we have $D_i(f+g) = D_i(f) + D_i(g)$ for any f, g, and $D_i(\alpha f) = \alpha D_i f$ for any $\alpha \in \mathbb{R}$. Then observe that

$$D_i(x^S) = \begin{cases} x^{S\setminus\{i\}} & \text{if } i \in S \\ 0 & \text{if } i \notin S. \end{cases}$$

□

We can see that $D_i f$ operates the same as the partial derivative operator $\frac{\partial f}{\partial x_i}$ on the Fourier expansion of f. Now we are ready to express $\mathsf{Inf}_i[f]$ in terms of its Fourier coefficients $\widehat{f}(S)$.

Theorem 4.12. $\mathsf{Inf}_i[f] := \sum_{S \ni i} \widehat{f}(S)^2$.

Proof. We can expand $\mathsf{Inf}_i[f] = \mathbf{E}_{\boldsymbol{x}}[D_i f(\boldsymbol{x})^2]$ by replacing $D_i f$ with its Fourier expansion, then as before use the orthonormality of the parity functions.

Alternatively, we can apply Parseval's identity to $D_i f$ and relate $\widehat{D_i f}(T)$ to $\widehat{f}(S)$. Since the Fourier expansion of any $g\colon \{-1,1\}^n \to \mathbb{R}$ is unique, from Proposition 4.11 we see that[2]

$$\widehat{D_i f}(S \setminus \{i\}) = \begin{cases} \widehat{f}(S) & \text{if } S \ni i \\ 0 & \text{otherwise.} \end{cases}$$

So we have $\sum_{T \subseteq [n]} \widehat{D_i f}(T)^2 = \sum_{S \ni i} \widehat{f}(S)^2$. □

Note that if f is monotone, then from the definition of $D_i f$ (Equation (5)) we see that $D_i f(x) \in \{1, 0\}$ and hence we have the following simple expression for $\mathsf{Inf}_i[f]$.

Proposition 4.13. *If $f\colon \{-1,1\}^n \to \{-1,1\}$ is monotone, then $\mathsf{Inf}_i[f] = \widehat{f}(\{i\})$.*

Proof. $\mathsf{Inf}_i[f] = \mathbf{E}_{\boldsymbol{x}}[D_i f(\boldsymbol{x})] = \mathbf{E}_{\boldsymbol{x}}[D_i(\boldsymbol{x})\boldsymbol{x}_i] = \widehat{f}(\{i\})$. □

We can derive some new properties of $\mathsf{Inf}_i[f]$ from their Fourier coefficients.

Claim 4.14. *If $f\colon \{-1,1\}^n \to \{-1,1\}$ is monotone and transitive-symmetric, then $\mathsf{Inf}_i[f] \le 1/\sqrt{n}$.*

Proof. $n \cdot \widehat{f}(\{1\})^2 = \sum_{i=1}^n \widehat{f}(\{i\})^2 \le 1$. □

[2] In the lecture I mistakenly claimed that $\widehat{D_i f}(S) = \widehat{f}(S)$. Thanks to Gabriel Wu for pointing out the mistake.

5.1 Total Influence

We first define the total influence of a Boolean function, then we will give 3 different ways of interpreting this notion, and a Fourier formula for it.

Definition 5.1 (Total influence)**.** The *total influence* of $f\colon \{-1,1\}^n \to \{-1,1\}$, denoted by $\mathbf{I}[f]$, is defined as

$$\mathbf{I}[f] := \sum_{i=1}^{n} \mathsf{Inf}_i[f].$$

For example, we have $\mathbf{I}[\mathsf{Maj}_n] = \Theta(\sqrt{n})$, $\mathbf{I}[\mathsf{Tribes}_{\log n, \Theta(n/\log n)}] = \Theta(\log n)$, and $\mathbf{I}[\chi_S] = |S|$. We now give 3 different ways of interpreting $\mathbf{I}[f]$.

5.1.1 Boundary of f

The first one arises from our graph interpretation of the individual influences $\mathsf{Inf}_i[f]$. We have

$$\begin{aligned}\mathbf{I}[f] &= \sum_{i=1}^{n} \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})] \\ &= \sum_{i=1}^{n} \frac{\#\text{ pivotal edges (of } f) \text{ in the } i\text{-th direction}}{2^{n-1}} \\ &= \frac{\#\text{ pivotal edges (of } f)}{2^{n-1}}.\end{aligned}$$

One can think of $\mathbf{I}[f]$ as the *surface area* of f, as it is (up to some scaling) counting the fraction of edges that lie on the boundary of $f^{-1}\{1\}$, i.e., edges with one endpoint in $f^{-1}\{1\}$ and the other in $f^{-1}\{-1\}$. (Note: this is not the only definition of surface area of a function f.) Om the other hand, one can think of the variance $\mathbf{Var}[f]$ of f as the *volume* of f: if we let $\mu = \mathbf{Pr}[f(\boldsymbol{x}) = 1]$, then $\mathbf{Var}[f] = 4\mu(1-\mu) = \Theta(\mu)$ when $\mu \leq 1/2$ (if $\mu > 1/2$, we can look at $-f$ instead). In this course, we will prove several *isoperimetric inequalities* that give lower bounds on the surface area of f in terms of its volume.

5.1.2 Average sensitivity

An equivalent definition of $\mathbf{I}[f]$ is the *average sensitivity* of f. First we define the sensitivity of f at a point $x \in \{-1,1\}$, which counts how many pivotal coordinates $i \in [n]$ in x, i.e flipping the i-th bit of x changes the value of $f(x)$. A graph interpretation of sensitivity would be the number of neighbors of the vertex x that are colored (by f) differently from x itself.

Definition 5.2 (Sensitivity)**.** The sensitivity of f at x, denoted by $\mathsf{sens}_f(x)$, is defined as

$$\mathsf{sens}_f(x) := \sum_{i=1}^{n} \mathbb{1}\big(f(x) \neq f(x^{\oplus i})\big).$$

The *average sensitivity* is simply the average of $\mathsf{sens}_f(\boldsymbol{x})$ over a uniform random $\boldsymbol{x} \in \{-1,1\}^n$, that is,

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\mathsf{sens}_f(\boldsymbol{x})] &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\Big[\sum_{i=1}^n \mathbb{1}\big(f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})\big)\Big] \\ &= \sum_{i=1}^n \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\Big[\sum_{i=1}^n \mathbb{1}\big(f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})\big)\Big] \\ &= \mathbf{I}[f].\end{aligned}$$

5.1.3 Spectral sampling

Before giving the third interpretation of $\mathbf{I}[f]$ we have to give the Fourier formula of $\mathbf{I}[f]$. Recall that $\mathsf{Inf}_i[f] := \sum_{S\ni i} \widehat{f}(S)^2$. Summing over all $i \in [n]$ and swapping summations, we can express $\mathbf{I}[f]$ in terms of the Fourier coefficients of f.

Proposition 5.3. $\mathbf{I}[f] = \sum_{S\subseteq[n]} |S| \widehat{f}(S)^2$.

Recall that for a Boolean function $f\colon \{-1,1\}^n \to \{-1,1\}$ we have $\sum_{S\subset[n]} \widehat{f}(S)^2 = 1$. Since the $\widehat{f}(S)^2$'s are always non-negative, the square of the coefficients form a distribution on $2^{[n]}$.

Definition 5.4 (Spectral sampling)**.** The *spectral sample* for $f\colon \{-1,1\}^n \to \{-1,1\}$, denoted by $\mathcal{S}_f$, is the probability distribution on $2^{[n]}$ defined by $\mathbf{Pr}[\mathcal{S}_f = S] = \widehat{f}(S)^2$.

Given this definition, one can see that the $\mathbf{I}[f]$ is the expected size of a random subset drawn according to $\mathcal{S}_f$.

Proposition 5.5. $\mathbf{I}[f] = \mathbf{E}_{\boldsymbol{S}\sim\mathcal{S}_f}[|S|]$.

It follows from Markov's inequality that if the total influence $\mathbf{I}[f]$ is small, then most of the Fourier mass of f lies in the low-degree part of its Fourier spectrum.

Proposition 5.6. $\sum_{S\subseteq[n]:|S|>\epsilon} \widehat{f}(S)^2 = \mathbf{Pr}_{\boldsymbol{S}\sim\mathcal{S}_f}[|\boldsymbol{S}| > \epsilon] \leq \mathbf{I}[f]/\epsilon$.

By comparing the Fourier formula of $\mathbf{I}[f]$ and $\mathbf{Var}[f]$, we immediately obtain our first isoperimetric inequality in this course.

Claim 5.7 (Poincaré's inequality)**.** $\mathbf{I}[f] \geq \mathbf{Var}[f]$.

5.2 Noise

We now introduce another natural concept in computer science that plays a crucial role in Boolean function analysis. Later in the course we will see how this concept leads to various important results in theoretical computer science. We first define our model of noise.

Definition 5.8. Given $x \sim \{-1,1\}^n$ and $\rho \in [-1,1]$, define the noisy random string $\boldsymbol{y} \sim N_\rho(x)$ by setting each $\boldsymbol{y}_i$ independently to

$$\begin{aligned}\boldsymbol{y}_i &:= \begin{cases}\text{uniform} & \text{with probability } 1-\rho \\ \mathrm{sgn}(\rho)\cdot x_i & \text{with probability } \rho\end{cases} \\ &= \begin{cases}x_i & \text{with probability } \frac{1+\rho}{2} \\ -x_i & \text{with probability } \frac{1-\rho}{2}.\end{cases}\end{aligned}$$

We say $\boldsymbol{y} \sim N_\rho(x)$ is *ρ-correlated* with x. Observe that when $\rho = 0$, we have that $\boldsymbol{y} \sim N_\rho(x)$ is the uniform random string, which is uncorrelated with x, and when $|\rho| = 1$, we have that $\boldsymbol{y} \in \{x, -x\}$, which is completely correlated with x.

One may think of the parameter ρ as the ($\rho = r$)etention probability of the noise. Equivalently, one can think of $\boldsymbol{y} \sim \mathsf{BSC}_\delta(x)$ for $\delta = \frac{1-\rho}{2} \in [0, 1]$, where $\mathsf{BSC}_\delta(x)$ is the binary symmetric channel that independently flips each bit x_i to $-x_i$ with probability δ.

Definition 5.9. We say $(\boldsymbol{x}, \boldsymbol{y})$ is a ρ-correlated pair if

1. $\boldsymbol{x}$ and $\boldsymbol{y}$ are both uniform in $\{-1, 1\}^n$ (but they are not necessarily independent), and
2. $\mathbf{E}[\boldsymbol{x}_i \boldsymbol{y}_i] = \rho$ for every $i \in [n]$.

Equivalently, a ρ-correlated pair can be sampled by drawing $\boldsymbol{x} \sim \{-1, 1\}^n$ uniformly at random, and $\boldsymbol{y} \sim N_\rho(\boldsymbol{x})$.

We can now introduce the important concept of noise stability.

Definition 5.10 (Noise Stability)**.** For $f\colon \{-1, 1\}^n \to \mathbb{R}$, and $\rho \in [-1, 1]$, the noise stability of f at ρ, denoted by $\mathsf{Stab}_\rho[f]$, is defined as

$$\mathsf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n, \boldsymbol{y} \sim N_\rho(x)} [f(\boldsymbol{x}) f(\boldsymbol{y})].$$

Note that for $f\colon \{-1, 1\}^n \to \{-1, 1\}$ we have $\mathsf{Stab}_\rho[f] = 2 \mathbf{Pr}_{(\boldsymbol{x}, \boldsymbol{y})\ \rho\text{-correlated}} \mathbf{Pr}[f(\boldsymbol{x}) = f(\boldsymbol{y})] - 1$, and for $f\colon \{-1, 1\}^n \to \{0, 1\}$, we have $\mathsf{Stab}_\rho[f] = \mathbf{Pr}_{(\boldsymbol{x}, \boldsymbol{y})\ \rho\text{-correlated}}[f(\boldsymbol{x}) = 1 \wedge f(\boldsymbol{y}) = 1]$. Viewing f as an indicator of a subset A, we can see that the noise stability of f measures the probability of a uniform random $\boldsymbol{x}$ lying in A and remains in A under some perturbation of noise.

Example 5.11. We have $\mathsf{Stab}_\rho[\chi_S] = \rho^{|S|}$. One way to see this is that for every x, as long as some bit of $\boldsymbol{y}$ in S is "rerandomized" to uniform, then the expectation is 0. Otherwise, we have $\boldsymbol{y} = x$.

Example 5.12. The stability of Majority satisfies $\lim_{n\to\infty} \mathsf{Stab}_\rho(\mathsf{Maj}_n) = \frac{2}{\pi} \arcsin(\rho)$, which is roughly $\frac{2}{\pi}\rho$ when ρ is close to 0, and $1 - O(\sqrt{1-\rho})$ when ρ is close to 1. (We will not prove it in this course.)

It is often useful to look at the "opposite" of noise stability, i.e., how sensitive f is under perturbing a uniform x with noise. Recall that $\boldsymbol{y} \sim \mathsf{BSC}_\delta(x) = N_{1-2\delta}(x)$.

Definition 5.13 (Noise Sensitivity)**.** The *noise sensitivity* of $f\colon \{-1, 1\}^n \to \{-1, 1\}$, denoted by $\mathsf{NS}_\delta[f]$, is defined as

$$\mathsf{NS}_\delta[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \{-1,1\}^n, \boldsymbol{y} \sim \mathsf{BSC}_\delta(\boldsymbol{x})} [f(\boldsymbol{x}) \neq f(\boldsymbol{y})] = \frac{1}{2}\big(1 - \mathsf{Stab}_{1-2\delta}[f]\big).$$

5.2.1 Noise operator

We now derive the Fourier formula for $\mathsf{Stab}_\rho[f]$, we first look at the behavior of f when a fixed input x is perturbed by noise. This motivates the definition of the noise operator.

Definition 5.14 (Noise Operator T_ρ)**.** The *noise operator* T_ρ on $f\colon \{-1, 1\}^n \to \mathbb{R}$, is the function $T_\rho f$ defined by

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

The noise operator dampens each coefficient of f by a factor $\rho^{|S|}$:

Proposition 5.15. $T_\rho(x) = \sum_{S\subseteq[n]} \widehat{f}(S)\rho^{|S|}\chi_S(x)$.

Proof. Show that $T_\rho\chi_S = \rho^{|S|}\chi_S$, and apply linearity of expectation. □

Now we can express $\mathsf{Stab}_\rho[f]$ in terms of $\widehat{f}(S)$.

Proposition 5.16. $\mathsf{Stab}_\rho[f] = \mathbf{E}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x})T_\rho(\boldsymbol{x})] = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}(S)^2$.

Clearly, the most noise stable functions are the two constant functions $f(x) \equiv 1$ and $f(x) \equiv -1$. What if we require the function to be balanced? We now show that the most noise stable balanced functions are dictators and anti-dictators (i.e. $-\chi_{\{i\}}$ for $i \in [n]$).

Claim 5.17. *Let $\rho \in (0,1)$. Then $\mathsf{Stab}_\rho[f] \le \rho$ for every f with $\mathbf{E}[f(\boldsymbol{x})] = 0$.*

Proof. Observe that $\rho^{|S|}$ is non-increasing in $|S|$. □

From the proof of Claim 5.17, we see that to attain $\mathsf{Stab}_\rho[f] = \rho$ the function must be a function of (real) degree 1. We have the following simple claim.

Claim 5.18. *If $f\colon \{-1,1\} \to \{-1,1\}$ has degree 1, then f is a constant, dictator, or anti-dictator.*

Proof. Since f has degree 1, we can write $f(x)$ as $\widehat{f}(\varnothing) + \sum_{i=1}^n \widehat{f}(\{i\})x_i$. By Parseval's identity, $\widehat{f}(\varnothing)^2 + \sum_{i=1}^n \widehat{f}(\{i\})^2 = 1$. In Homework 1, we showed that if $\deg(f) \le 1$, $\widehat{f}(S)$ must be an integer. So only one coefficient can be nonzero and it must have magnitude 1. □

6.1 Low-degree functions

Recall that the (real) degree of a function $f\colon \{-1,1\}^n \to \mathbb{R}$ is defined as

$$\deg(f) = \max\{|S| : \widehat{f}(S) \neq 0\}.$$

We first study Boolean functions of low degree, and later relax our notion of low degree and study Boolean functions that are "close" to being low degree.

An example of degree-k functions are the class of k-juntas: as every k-junta can be written as a function on k bits, every k-junta has degree at most k. Another class of low-degree functions is the class of low-depth decision trees.

Definition 6.1 (Decision Tree)**.** A decision tree T is a rooted binary tree, in which the nodes are labeled by a bit $x_i : i \in [n]$, the edges are labeled by $\{-1,1\}$, and the leaves are labeled by $\mathbb{R}$. It computes a function $f\colon \{-1,1\}^n \to \mathbb{R}$ as follows: Given an input $x \in \{-1,1\}^n$, it traverses from its root to a leaf by repeatedly querying the bit x_i labeled by the current node, and then moving to one of its two children by taking the edge labeled by x_i. The output of T on x is the value of the leaf.

Example 6.2. Draw a decision tree that computes the function $f(x_1, x_2, x_3)$ that returns -1 if $x_1 \leq x_2 \leq x_3$ or $x_1 \geq x_2 \geq x_3$, and 0 otherwise.

We use $\mathsf{depth}(T)$ to denote the length of the longest root-to-leaf path in T, and $\mathsf{size}(T)$ to denote the number of nodes in T.

The decision tree depth of f is the smallest depth of a decision tree computing f. Likewise, the decision tree size of f is the smallest size of a decision tree computing f.

Claim 6.3. *If f has decision tree depth k, then* $\deg(f) \leq k$.

Proof. We use the interpolation idea in the first lecture: consider the indicator function for each root-to-leaf path, which can be written as a polynomial of degree at most k. □

We have seen that a k-junta must have degree at most k. We now show the converse that a degree-k function must be a $k \cdot 2^k$-junta.

Theorem 6.4 (Nisan and Szegedy)**.** *If $f\colon \{-1,1\}^n \to \{-1,1\}$ has degree at most k, then f is a $k \cdot 2^{k-1}$-junta.*

Note that this theorem is non-trivial only when $k < \log n$, because every f is an n-junta.

Proof. We will show that

1. $\mathbf{I}[f] \leq \deg(f)$, and
2. For every $i \in [n]$, either $\mathsf{Inf}_i[f] \geq 2^{-k}$ or $\mathsf{Inf}_i[f] = 0$.

From this it is immediate that f has at most $k \cdot 2^k$ nonzero influential coordinates, and thus is a $(k \cdot 2^k)$-junta. Item 1 follows immediately from the spectral sampling interpretation of $\mathbf{I}[f]$. To prove Item 2, we express $\mathsf{Inf}_i[f]$ in terms of its derivative $D_i f$. Recall that $D_i f(x) \in \{-1, 0, 1\}$. So $\mathsf{Inf}_i[f] = \mathbf{Pr}_{\boldsymbol{x} \sim \{-1,1\}^n}[D_i(\boldsymbol{x}) \neq 0]$. Note that $D_i f$ is a degree $k-1$ polynomial. We now prove the following general claim about polynomials.

Claim 6.5. *Let $p\colon \{-1,1\}^n \to \mathbb{R}$ be a polynomial of degree at most d. Then either $p \equiv 0$ or $\mathbf{Pr}_{\boldsymbol{x}\sim\{-1,1\}^n}[p(\boldsymbol{x}) \neq 0] \geq 2^{-d}$.*

Proof. We apply induction on n. Write f as

$$p(x) = \left(\frac{1+x_1}{2}\right) f(1, x_2, \ldots, x_n) + \left(\frac{1-x_1}{2}\right) f(-1, x_2, \ldots, x_n).$$

For $b \in \{-1,1\}$, let $g_b\colon \{-1,1\}^n \to \{-1,1\}$ denote $f(b, x_2, \ldots, x_n)$.

If $g_1 \equiv g_{-1} \equiv 0$, then there is nothing to prove. If neither of them is identically zero, then by induction

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}[p(\boldsymbol{x}) \neq 0)] = \frac{1}{2} \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim\{-1,1\}^{n-1}}[g_1(\boldsymbol{y}) \neq 0] + \frac{1}{2} \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim\{-1,1\}^{n-1}}[g_{-1}(\boldsymbol{y}) \neq 0] \geq 2^{-d}.$$

For the remaining case, assume $g_1 \equiv 0$ but g_{-1} is not. Then we have

$$f(x) = \frac{1-x_1}{2} g_{-1}(x_2, \ldots, x_n).$$

So g_{-1} must have degree $d-1$, and $\mathbf{Pr}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x}) \neq 0] = \frac{1}{2}\mathbf{Pr}_{\boldsymbol{y}\sim\{-1,1\}^{n-1}}[g(\boldsymbol{y}) \neq 0] \geq 2^{-d}$. □

□

6.2 Fourier concentration

We now relax the notion of low degree through its Fourier spectrum.

Definition 6.6 (Low-degree Fourier concentration)**.** A function $f\colon \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on degree at most k if

$$W^{>k}[f] := \sum_{|S|>k} \widehat{f}(S)^2 = \mathop{\mathbf{Pr}}_{S\sim\mathcal{S}_f}\big[|S| > k\big] \leq \epsilon.$$

Note that if f is 0-concentrated on degree $\leq k$, then $\deg(f) \leq k$. So this is indeed a generalization of having degree $\leq k$. Let us compare this notion of closeness with one we saw in linearity testing.

6.2.1 Measures of closeness

Recall in linearity testing, we say that two functions $f, g\colon \{-1,1\}^n \to \{-1,1\}$ are ϵ-close if

$$\mathrm{dist}_{L_0}(f,g) := \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \leq \epsilon.$$

We sometimes call this L_0-distance because this is equal to $\mathbf{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^0]$ (where we define $0^0 = 0$). Replacing L_0 with L_2, we have the definition of L_2-distance.

Definition 6.7 (L_2 distance)**.** Two functions $f, g\colon \{-1,1\}^n \to \mathbb{R}$ are ϵ-close in L_2-distance if

$$\mathrm{dist}_{L_2}(f,g) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\big[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2\big] \leq \epsilon.$$

The L_2-distance is the "right" metric to use in Fourier analysis because of we can express the L_2 distance between f and g in terms of their Fourier coefficients, thanks to Parseval's identity.

Fact 6.8. $\mathrm{dist}_{L_2}(f,g) = \sum_{S\subseteq[n]} (\widehat{f}(S) - \widehat{g}(S))^2$.

We now compare L_0-distance with L_2-distance and see if closeness in one notion implies closeness in the other. It turns out they are the same up to a factor of 4.

Proposition 6.9. *For every $f, g\colon \{-1,1\}^n \to \{-1,1\}$, we have $\mathrm{dist}_{L_0}(f,g) = 4\mathrm{dist}_{L_2}(f,g)$.*

Proof. Since $f(x) - g(x) \in \{-2, 0, 2\}$,

$$\mathrm{dist}_{L_0}(f,g) = \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] = 4 \mathop{\mathbf{E}}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2] = 4 \cdot \mathrm{dist}_{L_2}(f,g). \qquad \square$$

We now compare ϵ-concentration to L_2-closeness. We will show that $f\colon \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on degree $\leq k$ if and only if f is ϵ-close to some function $g\colon \{-1,1\}^n \to \mathbb{R}$ of degree at most k in L_2-distance.

Proposition 6.10. *$f\colon \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on degree $\leq k$ if and only if $\mathbf{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2] \leq \epsilon$ for some $g\colon \{-1,1\}^n \to \mathbb{R}$.*

Proof. Suppose f is ϵ-concentrated on degree $\leq k$. Define $g\colon \{-1,1\}^n \to \mathbb{R}$ by $g(x) := \sum_{|S| \leq k} \widehat{f}(S)\chi_S(x)$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2] = \sum_{S \subseteq [n]} (\widehat{f}(S) - \widehat{g}(S))^2 = \sum_{|S| > k} \widehat{f}(S)^2 \leq \epsilon.$$

Suppose f is ϵ-close to some function g of degree at most k. Then

$$\sum_{|S|>k} \widehat{f}(S)^2 = \sum_{|S|>k} (\widehat{f}(S) - \widehat{g}(S))^2 \leq \sum_{S \subseteq [n]} (\widehat{f}(S) - \widehat{g}(S))^2 \leq \epsilon. \qquad \square$$

From the proof you can see that $g := \sum_{|S| \leq k} \widehat{f}(S)\chi_S$ is the unique degree-k function closest to f in L_2 distance. What if we further require g to be Boolean? We have the following theorem by Kindler and Safra.

Theorem 6.11 (Kindler and Safra)**.** *If $f\colon \{-1,1\}^n \to \{-1,1\}$ is ϵ-concentrated on degree $\leq k$, where $\epsilon \leq \epsilon_k$ for some $\epsilon_k > 0$, then $\mathbf{E}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2] \leq 2\epsilon$ for some Boolean function $g\colon \{-1,1\}^n \to \{-1,1\}$.*

Let us now look at some functions with low-degree spectral concentration. Recall in Proposition 5.6 we showed that functions with total influence $\mathbf{I}[f]$ are ϵ-concentrated on degree at most $\mathbf{I}[f]/\epsilon$. The proof idea is to look at the spectral sampling interpretation of $\mathbf{I}[f]$. It turns out we can apply the same idea to the noise stability of f.

Proposition 6.12. *Every $f\colon \{-1,1\}^n \to \{-1,1\}$ is $2(1 - \mathsf{Stab}_\rho[f])$-concentrated on degree at most $1/(1-\rho)$.*

Note that Proposition 6.12 is interesting when the retention rate $\rho := 1 - 1/\epsilon$ is close to 1 and the $\mathsf{Stab}_\rho[f]$ is close to 1, so let's state it in terms of noise sensitivity.

Proposition 6.13. *Every $f\colon \{-1,1\}^n \to \{-1,1\}$ is $(4\mathsf{NS}_\delta[f])$-concentrated on degree at most $1/(2\delta)$.*

Proof. Recall that $\mathsf{NS}_\delta[f] = \frac{1}{2}(1 - \mathsf{Stab}_{1-2\delta}[f])$. $\square$

We now prove Proposition 6.12.

Proof. Observe that

$$\mathsf{Stab}_\rho[f] = \sum_{S \subseteq [n]} \rho^{|S|} \widehat{f}(S)^2 = \mathop{\mathbf{E}}_{\boldsymbol{S} \sim \mathcal{S}_f}[\rho^{|\boldsymbol{S}|}].$$

Since $1 = \mathbf{E}_{\boldsymbol{S} \sim \mathcal{S}_f}[1]$, we have

$$\begin{aligned}
1 - \mathsf{Stab}_\rho[f] &= \mathop{\mathbf{E}}_{\boldsymbol{S} \sim \mathcal{S}_f}[1 - \rho^{|\boldsymbol{S}|}] \\
&\geq \mathop{\mathbf{E}}_{\boldsymbol{S} \sim \mathcal{S}_f}[(1 - \rho^{|\boldsymbol{S}|}) \cdot \mathbb{1}(|\boldsymbol{S}| > t)] \\
&\geq \mathop{\mathbf{E}}_{\boldsymbol{S} \sim \mathcal{S}_f}[(1 - \rho^{t}) \cdot \mathbb{1}(|\boldsymbol{S}| > t)] \\
&= (1 - \rho^t) \mathop{\mathbf{Pr}}_{\boldsymbol{S} \sim \mathcal{S}_f}[|\boldsymbol{S}| > t].
\end{aligned}$$

Setting $t = 1/(1-\rho)$, we have $1 - \rho^t \geq 1 - e^{-1} \geq 1/2$. □

We saw in Theorem 6.4 that every degree-k Boolean function must be a $(k \cdot 2^k)$-junta. Later in the course, we will later see a robust analogue of this theorem by Friedgut. (Its proof is also closely related to the proof of Theorem 6.11). Recall that we always have $\mathbf{I}[f] \leq \deg(f)$.

Theorem 6.14 (Friedgut's junta theorem)**.** *Every $f\colon \{-1,1\}^n \to \{-1,1\}$ is ϵ-close to a $2^{O(\mathbf{I}[f]/\epsilon)}$-junta.*

7.1 Learning low-degree functions

We will show how to learn functions that are close to low-degree. The model we will be using is the PAC (Probably Approximately Correct) learning model introduced by Valiant in 1984.

The setup is as follows. Fix a concept class $\mathcal{C} \subseteq \{f\colon \{-1,1\}^n \to \{-1,1\}\}$. The learning algorithm is given "restricted access" to an unknown function $f \in \mathcal{C}$. We will focus on two access models:

1. **Random examples:** The algorithm is given random examples of the form $(\boldsymbol{x}^{(1)}, f(\boldsymbol{x}^{(1)}))$, $\ldots, (\boldsymbol{x}^{(T)}, f(\boldsymbol{x}^{(T)}))$, where the $\boldsymbol{x}^{(t)}$'s are uniform.

2. **Query access:** The algorithm makes queries on the points $x^{(1)}, \ldots, x^{(T)}$ of its choice, and receives the values $f(x^{(1)}), \ldots, f(x^{(T)})$ in return.

Note that a learning algorithm in the query access model can simulate an algorithm in the random examples model by choosing the samples uniformly at random.

Realizable vs. Agnostic learning. *Realizable* means that the value of $f(x^t)$ is always correct; *Agnostic* means that the algorithm may not receive the correct value of $f(x^{(t)})$ sometimes.

In both the random examples and query access settings, given $(x^{(1)}, f(x^{(1)})), \ldots, (x^{(T)}, f(x^{(T)}))$, the learning algorithm outputs a function $h\colon \{-1,1\}^n \to \{-1,1\}$ that is close to f. In *proper learning*, we requires $h \in \mathcal{C}$, whereas in *improper learning*, h can be an arbitrary Boolean function.

We will only talk about *realizable* and *improper learning*.

Definition 7.1. A learning algorithm learns a concept class $\mathcal{C}$ with accuracy ϵ and T samples/queries if for every $f \in \mathcal{C}$, given $(\boldsymbol{x}^{(1)}, f(\boldsymbol{x}^{(1)})), \ldots, (\boldsymbol{x}^{(T)}, f(\boldsymbol{x}^{(T)}))$ it outputs an $\boldsymbol{h}\colon \{-1,1\}^n \to \{-1,1\}$ such that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{h}}[f \text{ and } h \text{ are } \epsilon\text{-close}] \geq 9/10.$$

where the randomness of $\boldsymbol{h}$ is over the randomness of the random examples (if we are in the random examples model), and the internal randomness of the algorithm.

Theorem 7.2 (Linial–Nisan–Mansour)**.** *Let $\mathcal{C}$ be the class of functions $f\colon \{-1,1\}^n \to \{-1,1\}$ that are ϵ-concentrated on degree at most k. There is a learning algorithm that learns $\mathcal{C}$ with accuracy ϵ in time* $\mathrm{poly}(n^k, 1/\epsilon)$.

Note that in expectation all the 2^n inputs will appear in $O(n \cdot 2^n)$ random examples, and so every Boolean function can be learned using this many examples.

7.1.1 LMN algorithm

We now prove Theorem 7.2. The algorithm is based on two ideas. The first one is to estimate all the low-degree Fourier coefficients using random examples. The second one is to come up with a Boolean function that is close to f using these estimates.

Algorithm 1: Linial–Mansour–Nisan Algorithm

Input: Random examples access to an unknown $f\colon \{-1,1\}^n \to \{-1,1\}$.
Output: A Boolean function $h\colon \{-1,1\}^n \to \{-1,1\}$

1 Estimate $\widehat{f}(S)$ for every $|S| \le k$ good enough accuracy
2 Let $\widehat{g}(S)$ be our estimates of $\widehat{f}(S)$
3 Output $h(x) := \operatorname{sgn}\big(\sum_{|S|\le k} \widehat{g}(S)\chi_S(x)\big)$

Estimating $\widehat{f}(S)$. Recall that $\widehat{f}(S) = \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})]$. We will approximate this average by empirical estimation. Specifically, for each $S : |S| \le k$, we draw some T random samples $(\boldsymbol{x}^{(t)}, f(\boldsymbol{x}^{(t)})) : t \in [T]$ and estimate $\widehat{f}(S)$ with

$$\widehat{g}(S) := \frac{1}{T}\sum_{t\in[T]} f(\boldsymbol{x}^{(t)})\chi_S(\boldsymbol{x}^{(t)}).$$

By the Chernoff bound, we have

$$\mathbf{Pr}\Big[\big|\widehat{g}(S) - \widehat{f}(S)\big| > \delta\Big] \le e^{-\delta^2 T/8}.$$

Since we will be summing the square of the differences over all the low-degree coefficients, we set δ to be $\sqrt{\epsilon/\binom{n}{k}}$, and we also need to take an union bound over the $\binom{n}{k} \le n^k$ coefficients. So we pick $T = 100\binom{n}{k}(k\log n)/\epsilon$, and we have

$$\mathbf{Pr}\Big[\big(\widehat{g}(S) - \widehat{f}(S)\big)^2 \le \frac{\epsilon}{\binom{n}{k}} \text{ for every } |S| \le k\Big] \le 1/10.$$

It follows from our choice of T that the algorithm runs in time $\mathrm{poly}(n^k, 1/\epsilon)$.

Showing h is close to f. Here the key observation is that $\mathbb{1}(f(x) \ne h(x)) \le (f(x) - g(x))^2$, because $f(x) \in \{-1,1\}^n$ and $g(x)$ has the opposite sign to $f(x)$. The rest is just a simple calculation:

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}) \ne h(\boldsymbol{x})\big] &\le \mathop{\mathbf{E}}_{\boldsymbol{x}}\big[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2\big] \\
&= \sum_{S\subseteq[n]} (\widehat{f}(S) - \widehat{g}(S))^2 \\
&= \sum_{S:|S|\le k} (\widehat{f}(S) - \widehat{g}(S))^2 + \sum_{S:|S|>k} \widehat{f}(S)^2 \\
&\le 2\epsilon
\end{aligned}$$

This completes the proof of Theorem 7.2. It immediately implies learning algorithms for the following concept classes.

Corollary 7.3. *The following concept classes can be learned using random samples.*

1. $\mathcal{C} := \{f\colon \deg(f) \le t\}$ *can be learned in time* $\mathrm{poly}(n^k, 1/\epsilon)$.
2. $\mathcal{C} := \{f\colon \mathbf{I}[f] \le t\}$ *can be learned in time* $\mathrm{poly}(n^{t/\epsilon})$.
3. $\mathcal{C} := \{f\colon \mathsf{NS}_\delta[f] \le \epsilon\}$ *can be learned in time* $\mathrm{poly}(n^{1/\delta}, 1/\epsilon)$.

4. $\mathcal{C} := \{f : f \text{ is monotone}\}$ *can be learned in time* $\mathrm{poly}(n^{O(\sqrt{n}/\epsilon)})$.

Proof. Items 2 and 3 follow from ϵ-concentration of low total influence and low noise sensitivity functions (Propositions 5.6 and 6.13). Item 4 follows from Homework 1 Q8, because for monotone functions we have $\mathsf{Inf}_i[f] = \widehat{f}(\{i\})$. □

We can also consider spectral concentration on arbitrary subsets of Fourier coefficients, not just the low-degree ones.

Definition 7.4. $f : \{-1,1\}^n \to \{-1,1\}$ is ϵ-concentrated on a set of coefficients $\mathcal{F} \subseteq 2^{[n]}$ if $\sum_{S \notin \mathcal{F}} \widehat{f}(S)^2 \leq \epsilon$.

It is straightforward to see the LMN algorithm can also learn functions that are ϵ-concentrated on a "small" subset $\mathcal{F}$ of coefficients.

Corollary 7.5. *Let $\mathcal{C}$ be the class of functions $f : \{-1,1\}^n \to \{-1,1\}$ that are ϵ-concentrated on $\mathcal{F} \subseteq 2^{[n]}$. There is a learning algorithm that learns $\mathcal{C}$ with accuracy ϵ in time* $\mathrm{poly}(|\mathcal{F}|, 1/\epsilon)$.

8.1 Goldreich–Levin Theorem

Corollary 7.5 assumes the algorithm knows the subset of coefficients f is concentrated on. We now show how to identify these "heavy" coefficients efficiently.

Recall in Homework 1 Q6, we showed that given query access to an unknown function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ that is 1/8-close to linear, we can learn the linear function f is ϵ-close to in time $O(n \log n)$.

We now consider the what happens if f is $(1/2-\epsilon)$-close to linear. The first thing to notice is that there may be multiple linear functions that are $(1/2-\epsilon)$-close to f; so we will output all the linear functions that are close to f. We will identify each linear function $\sum_{i\in S} x_i$ by the subset $S \subseteq [n]$. Also, we will instead identify all the linear functions that are *correlated* to f. So new task is the following.

Given query access an unknown $f\colon \{-1,1\}^n \to \{-1,1\}$, output the subsets S such that $|\mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})]| = |\widehat{f}(S)| \geq \epsilon$.

We will prove the following theorem by Goldreich and Levin.

Theorem 8.1 (Goldreich–Levin Theorem)**.** *Given query access to an unknown $f\colon \{-1,1\}^n \to \{-1,1\}$ and a threshold parameter $\tau \in (0,1)$, there is a $\mathrm{poly}(n, 1/\tau)$-time algorithm that with probability at least $9/10$ outputs a list $L = \{S_1, \ldots, S_\ell\} \subseteq 2^{[n]}$ such that*

1. *(Completeness) If $|\widehat{f}(S)| > \tau$ then $S \in L$.*
2. *(Soundness) If $S \in L$ then $|f(S)|^2 \geq \tau/2$.*

Note that by Parseval's identity, the list size ℓ is bounded above by $4/\tau^2$.

The motivation of Theorem 8.1 comes from cryptography, and is related to hardness amplification and list-decoding. The original proof of Theorem 8.1 does not use Fourier analysis, but is also quite elegant. Here we present a Fourier-based algorithmic proof given by Kushilevitz and Mansour.

8.1.1 Kushilevitz–Mansour algorithm

High-level idea. As $\widehat{f}(S)^2 \geq 0$, if there is a subset $\mathcal{S} \subseteq 2^{[n]}$ of coefficients so that its weight $\sum_{S\in\mathcal{S}} \widehat{f}(S)^2$ is less than τ, then we know none of the coefficients in $\mathcal{S}$ belongs to our list. Otherwise, we can divide the set $\mathcal{S}$ into two halves and recur. We now describe the algorithm in detail.

Algorithm 2: Kushilevitz–Mansour Algorithm

Input: Query access to an unknown $f\colon \{-1,1\}^n \to \{-1,1\}$.
Output: A list $L = \{S_1, \ldots, S_\ell\} \in 2^{[n]}$

```
KM(γ, i):
    Compute R_γ := Σ_{β∈{0,1}^{n-i}} f̂(γ ∘ β)^2
    if R_γ < γ^2/2 then Stop
    if i = n then Output γ ∈ {0,1}^n
    KM(γ ∘ 0, i + 1)
    KM(γ ∘ 1, i + 1)
Run KM(ε, 0)
```

The depth of the recursion tree of the algorithm is at most n, and at each level of the tree we only recur into at most $4/\tau^2$ of the nodes. So the total number of calls to line 2 is at most $O(1/\tau^2)$.

Just like in the LMN algorithm, we do not compute $\sum_{\beta\in\{0,1\}^{n-i}} \widehat{f}(\gamma\circ\beta)^2$ exactly, but rather estimate this sum by expressing it as the expectation of some Boolean function.

Restrictions. Since we are estimating a sum of square of the coefficients, if there were a function f_γ such that

$$f_\gamma(y) := \sum_{\beta\in\{0,1\}^{n-i}} \widehat{f}(\gamma\circ\beta)\chi_\beta(y),$$

then by Parsevel's we have $\mathbf{E}_{\boldsymbol{y}}[f_\gamma(\boldsymbol{y})^2] = R_\gamma$, and we can estimate the sum by empirical estimation. To identify such function, let us divide the n-bit input into a prefix $x\in\{-1,1\}^i$ and a suffix $y\in\{-1,1\}^{n-i}$. We can write the Fourier expansion of f as

$$f(x,y) = \sum_{\alpha\in\{0,1\}^i}\sum_{\beta\in\{0,1\}^{n-i}} \widehat{f}(\alpha\circ\beta)\chi_\alpha(x)\chi_\beta(y).$$

Now suppose $\gamma=\vec{0}$. Then we would like apply some operation to f so that the terms corresponding to $\alpha\neq 0$ in the above sum vanish. Because each term contains a factor of $\chi_\alpha(x)$, we can simply average over the prefix x to achieve this goal, that is,

$$\mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x},y)] = \sum_{\beta\in\{0,1\}^{n-i}} \widehat{f}(0^i\circ\beta)\chi_\beta(y).$$

For general $\gamma\in\{0,1\}^n$, as in Homework 1 Q2 we can apply a "shift" by multiplying $\chi_\gamma(x)$. Therefore, we define f_γ as

$$f_\gamma(y) := \mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x},y)\chi_\gamma(\boldsymbol{x})] = \sum_{\beta\in\{0,1\}^{n-i}} \widehat{f}(\gamma\circ\beta)\chi_\beta(y).$$

So to estimate R_γ, it suffices to estimate

$$\mathop{\mathbf{E}}_{\boldsymbol{y}}\Big[\mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{x}'}[f(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{x}',\boldsymbol{y})\chi_\gamma(\boldsymbol{x}+\boldsymbol{x}')]\Big].$$

As before, we draw T independent tuples $\{(\boldsymbol{x}^{(t)},\boldsymbol{x}'^{(t)},\boldsymbol{y}^{(t)})\}_{t\in[T]}$, and compute the average of $f(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{x}',\boldsymbol{y})\chi_\gamma(\boldsymbol{x}+\boldsymbol{x}')\in\{-1,1\}$. We will ensure each of our $O(n/\tau^2)$ coefficients is within $\tau^2/4$ accuracy by taking $T=O(\log(n)/\tau^4)$. The correctness then follows from the usual Chernoff bound and union bound argument. The total running time of the algorithm is $n\log n\cdot\mathrm{poly}(1/\tau)$. This finishes the proof of Theorem 8.1.

We can now combine the two algorithms we just learned to learn the class of low-degree Boolean functions.

Corollary 8.2. *Given query access to an unknown Boolean function f of degree at most k, f can be recovered exactly using $2^{O(k)}$ samples.*

Proof. In Homework 1 Q4, we know that f contains at most 2^{2k} coefficients and each of them is an integer multiple of 2^{-k}. We can use Kushilevitz–Mansour algorithm to identify the set of non-empty coefficients, then use LMN to estimate each non-empty coefficient within an accuracy of $2^{-k}/100$, and round our estimates to the nearest integer multiple of 2^{-k}. □

So far we have talked about learning Boolean functions $f\colon\{-1,1\}^n\to\{-1,1\}$. Here we record a recent result by Eskenazis and Ivanisvili on learning low-degree bounded real-valued functions $f\colon\{-1,1\}^n\to[-1,1]$.

Theorem 8.3. *Every $f\colon \{-1,1\}^n \to [-1,1]$ of degree at most k can be learned with accuracy* 0.001 *using* $\log n \cdot 2^{\tilde{O}(d^{3/2})}$ *random samples.*

8.2 DNFs

We now look at the class of DNF formulas, a basic computational model that generalizes the class of decision trees. Unlike small-depth decision trees, a DNF formula can have degree as large as n, but we will show that its Fourier mass is concentrated on the low degree, and therefore is easy to learn.

Definition 8.4. A DNF (resp. CNF) formula over variables $x_1, \ldots, x_n$ is an OR of AND (resp. AND of OR) of literals, where each literal is either x_i or $\overline{x_i}$ for some $i \in [n]$. The width of a DNF (resp. CNF) is the maximum number of literals appearing in each OR (also called a term) (resp. AND (also called a clause)). The size of a DNF (resp. CNF) is the number of terms (resp. clauses) in it.

Let us recall the following fact.

Fact 8.5. *If f is computable by a DNF of size s and width w, then its negation $\overline{f}$ can be computed by a CNF of size s and width w.*

Proof. De Morgan's law. □

Since f and $\overline{f}$ have the same Fourier spectrum up to negation, in Fourier analysis we can focus on DNF, which is what we will do when we look at their Fourier spectrum. DNFs and CNFs are generalization of the class of decision trees in the following sense.

Fact 8.6. *If $f\colon \{0,1\}^n \to \{0,1\}$ is computable by a decision tree of size s and depth k, then f is computable by a DNF (and CNF) of size s and width k.*

Proof. To convert a decision tree to a DNF, take the OR over each computation path that leads to a leaf value 1, observe that the indicator of each path is an AND. To convert the decision tree into a CNF, first convert negation of the decision tree to a DNF, then negate the DNF. □

Example 8.7. The function $f\colon \{0,1\}^3 \to \{0,1\}$ defined by $f(x_1,x_2,x_3) = \mathbb{1}((x_1 \le x_2 \le x_3) \vee (x_1 \ge x_2 \ge x_3))$ has a decision tree of size 6 and depth 3. It has a degree-3 polynomial representation

$$\begin{aligned} f(x_1,x_2,x_3) := {} & 1 \cdot (1-x_1)(1-x_2) \\ & + 0 \cdot (1-x_1)x_2(1-x_3) \\ & + 1 \cdot (1-x_1)x_2x_3 \\ & + 1 \cdot x_1(1-x_2)(1-x_3) \\ & + 0 \cdot x_1(1-x_2)x_3 \\ & + 1 \cdot x_1x_2. \end{aligned}$$

It can be written as the following DNF of size 4 and width 3

$$f(x_1,x_2,x_3) := \overline{x_1}\,\overline{x_2} + 1 \cdot \overline{x_1}x_2x_3 + x_1\overline{x_2}\,\overline{x_3} + x_1x_2,$$

and the following CNF of size 2 and width 3

$$f(x_1,x_2,x_3) := (x_1 \vee \overline{x_2} \vee x_3) \wedge (\overline{x_1} \vee x_2 \vee \overline{x_3}).$$

Note that this conversion does not necessary give the DNF of the smallest width. Indeed, one can verify $x_1x_2 \wedge \overline{x_2}\,\overline{x_3} \wedge \overline{x_1}x_3$ is another DNF computing f.

As a result we have the following corollary.

Corollary 8.8. *Every $f\colon \{-1,1\}^n \to \{-1,1\}$ has a width-n DNF (and CNF) of size at most 2^n.*

8.2.1 Total influence of DNF

By considering the AND_n function, we can see that a DNF can have degree as large as n. We will show that every DNF is ϵ-concentrated on its low degree part. By Proposition 5.6, it suffices to show that the total influence of a width-w DNF is bounded by $2w$. As the proof does not use any Fourier analysis, let us stick with $\{0,1\}$ for ease of understanding.

Claim 8.9. *Let $f\colon \{0,1\}^n \to \{0,1\}$, then $\mathbf{I}[f] \le 2w$.*

Remark 8.10. Recall that any Boolean function on w bits can be computed by a width-w DNF and the parity on w bits has total influence w.

Proof. We first claim that

$$\mathsf{Inf}_i[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})\big] = 2\mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}) = 1 \wedge f(\boldsymbol{x}^{\oplus i}) = 0\big].$$

To see this, observe that $\boldsymbol{x}$ and $\boldsymbol{x}^{\oplus i}$ are identically distributed and so $\mathbf{Pr}_{\boldsymbol{x}}[f(\boldsymbol{x}) = 0 \wedge f(\boldsymbol{x}^{\oplus i}) = 1] = \mathbf{Pr}_{\boldsymbol{x}}[f(\boldsymbol{x}) = 1 \wedge f(\boldsymbol{x}^{\oplus i}) = 0]$. We will instead look at the average sensitivity of f. We have

$$\mathbf{I}[f] = 2\sum_{i=1}^{n} \mathop{\mathbf{Pr}}_{\boldsymbol{x}}\big[f(\boldsymbol{x}) = 1 \wedge f(\boldsymbol{x}^{\oplus i}) = 0\big] = 2\mathop{\mathbf{E}}_{\boldsymbol{x}}\Big[\#\{i \in [n] : f(x) = 1 \wedge f(x^{\oplus i}) = 0\}\Big].$$

Now, fix any $x \in \{0,1\}^n$ such that $f(x) = 1$. Without loss of generality assume the first term is satisfied and it contains the variables $x_1, \dots, x_w$. Flipping the bits $x_i : i > w$ does not change the value of the first term and therefore does not change the outcome. Hence there can be at most w sensitive coordinates for each $x \in f^{-1}\{1\}$. □

Corollary 8.11. *If $f\colon \{-1,1\}^n \to \{-1,1\}$ is computable by a width-w DNF, then it is ϵ-concentrated on degree $\le 2w/\epsilon$.*

Later we will improve the $2w/\epsilon$ to $O(w \log(1/\epsilon))$.

So far we have showed that small-width DNFs have low-degree concentration. Is the same true for DNF of small size? We now show well-known fact that a size-s DNF is ϵ-close to a width-$O(\log(s/\epsilon))$ DNF.

Proposition 8.12. *Every $f\colon \{0,1\}^n \to \{0,1\}$ computable by size-s DNF is ϵ-close to a DNF of width $O(\log(s/\epsilon))$.*

Proof. The observation is that a term with many literals is unlikely to be satisfied, so removing it from the DNF only affects its expectation very little. Let $F = T_1 \vee \cdots \vee T_s$ be the size-s DNF computing f. Let $S := \{i \in [n] :\ T_i \text{ has more than } \log(s/\epsilon) \text{ literals}\}$. Define $F' = \bigwedge_{i \in S} T_i$. Observe that $F'(x) \le F(x)$ for every $x \in \{0,1\}^n$, and

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}}[F(\boldsymbol{x}) = 1 \wedge F(\boldsymbol{x}) = 0] \le \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[F(\boldsymbol{x}) = 1 \wedge F(\boldsymbol{x}) = 0] \le \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[T_i(\boldsymbol{x}) = 1 \text{ for some } i \in S] \le \epsilon. \quad \square$$

Recall in Proposition 6.10 we showed that ϵ-closeness in L_2-distance and ϵ-spectral concentration are equivalent. So we have the following corollary.

Corollary 8.13. *Every size-s DNF is ϵ-concentrated on degree at most $O(\log(s/\epsilon)/\epsilon)$.*

As a result, we have the following learning algorithm for DNFs.

Corollary 8.14. *Every size-s DNF can be learned using random examples within* 0.01*-accuracy in time* $n^{O(\log s)}$.

We end this lecture by recording a longstanding conjecture by Mansour about the Fourier spectrum of DNFs.

Conjecture 8.15 (Mansour's Conjecture)**.** *Every width-w DNF is* 0.01*-concentrated on* $2^{O(w)}$ *coefficients.*

Thus, not only the heavy coefficients are concentrated in the low-degree part, but they are also conjectured to be quite sparse. Note that this also implies every $\mathrm{poly}(n)$-size DNF is 0.01-concentrated on $\mathrm{poly}(n)$ many coefficients.

9.1 Random restrictions

We introduce the concept of random restrictions. First let us define restrictions properly.

Definition 9.1. For a subset $J \subseteq [n]$ and $z \in \{-1,1\}^{\overline{J}}$, the restriction of f to J with z is the function $f_{J,z} \colon \{-1,1\}^J \to \{-1,1\}$ defined by

$$f_{J,z}(x) := f(x \circ z).$$

Recall from the last lecture and Homework 1 that we compute at the Fourier coefficients $\widehat{f_{J,z}}$ with $\overline{J} = [k]$ by first writing down the Fourier expansion of f as

$$f(x \circ z) = \sum_{S \subseteq J} \sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T) \chi_S(x) \chi_T(x) = \sum_{S \subseteq J} \Big(\sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T) \chi_T(x) \Big) \chi_T(x).$$

Using the uniqueness of Fourier expansion we immediately have the following formula for $\widehat{f_{J,z}}(S)$.

Proposition 9.2. $\widehat{f_{J,z}}(S) = \mathbb{1}(S \subseteq J) \sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T) \chi_T(z)$.

Since $f_{J,z}$ is defined on $\{-1,1\}^J$, the indicator $\mathbb{1}(S \subset J)$ is redundant. However, we will often think of $f_{J,z}$ as a function on $x \in \{0,1\}^n$ but ignoring the $\overline{J}$ portion of x. So we will keep the indicator as is.

We often look at the $f_{J,\boldsymbol{z}}$ for a random $\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}$, and we have the following formulas.

Proposition 9.3. *We have*

1. $\mathbf{E}_{\boldsymbol{z} \sim \{-1,1\}^n}[\widehat{f_{J,\boldsymbol{z}}}(S)] = \mathbb{1}(S \subseteq J) \widehat{f}(S)$
2. $\mathbf{E}_{\boldsymbol{z} \sim \{-1,1\}^n}[\widehat{f_{J,\boldsymbol{z}}}(S)^2] = \mathbb{1}(S \subseteq J) \sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T)^2$.

We can of course choose J at random as well. This leads to the concept of *random restrictions*, a fundamental concept in theoretical computer science, which we will many applications of it in the upcoming lectures.

Definition 9.4 (Random restriction)**.** Let $\rho \in [0,1]$. A ρ-random restriction, denoted by R_ρ, is a pair $(\boldsymbol{J}, \boldsymbol{z})$ sampled as follows.

1. $\boldsymbol{J} \sim_\rho [n]$: include each $i \in [n]$ to $\boldsymbol{J}$ with probability ρ independently
2. $\boldsymbol{z} \sim \{-1,1\}^n$ is uniform.

Given $x \in \{0,1\}^n$, we can think of $\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(x)$ (as a function on $\{0,1\}^n$) as setting each coordinate x_i to uniform independently with probability $1-\rho$. So $\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(x)$ is identically distributed as $f(\boldsymbol{y})$, where $\boldsymbol{y} \sim N_\rho(x)$ is the noisy random string defined in Definition 5.8.

Proposition 9.5. *Let R_ρ be a ρ-random restriction. Then*

1. $\mathbf{E}_{(\boldsymbol{J},\boldsymbol{z}) \sim R_\rho}[\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(S)] = \rho^{|S|} \cdot \widehat{f}(S)$
2. $\mathbf{E}_{(\boldsymbol{J},\boldsymbol{z}) \sim R_\rho}[\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(S)^2] = \sum_{U \subseteq [n]} \widehat{f}(U)^2 \, \mathbf{Pr}[U \cap \boldsymbol{J} = S]$.

Proof. We prove Item 2.

$$\mathop{\mathbf{E}}_{\boldsymbol{J}\sim_\rho[n]}\left[\mathbb{1}(S\subseteq \boldsymbol{J})\sum_{T\subseteq\overline{\boldsymbol{J}}}\widehat{f}(S\cup T)^2\right]=\sum_{U\subseteq[n]}\widehat{f}(U)^2\mathop{\mathbf{E}}_{\boldsymbol{J}\sim_\rho[n]}\big[\mathbb{1}(U\cap\boldsymbol{J}=S)\big].$$ □

We can often relate a measure of a typical $f_{\boldsymbol{J},\boldsymbol{z}}$ to the same measure of f, and vice versa. The total influence is one example.

Proposition 9.6. $\mathbf{E}_{(\boldsymbol{J},z)\sim R_\rho}[\mathbf{I}[f_{\boldsymbol{J},\boldsymbol{z}}]]=\rho\cdot\mathbf{I}[f]$.

Proof. Recall that $\mathbf{I}[f]=\sum_{S\subseteq[n]}|S|\widehat{f}(S)^2$. So

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\boldsymbol{J},\boldsymbol{z})\sim R_\rho}\left[\sum_{S\subseteq[n]}|S|\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(S)^2\right]&=\sum_{S\subseteq[n]}|S|\mathop{\mathbf{E}}_{(\boldsymbol{J},\boldsymbol{z})\sim R_\rho}\left[\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(S)^2\right]\\
&=\sum_{S\subseteq[n]}|S|\sum_{U\subseteq[n]}\widehat{f}(U)^2\mathop{\mathbf{Pr}}_{\boldsymbol{J}\sim_\rho[n]}[U\cup\boldsymbol{J}=S]\\
&=\sum_{U\subseteq[n]}\widehat{f}(U)^2\sum_{S\subseteq[n]}|S|\mathop{\mathbf{Pr}}_{\boldsymbol{J}\sim_\rho[n]}[U\cup\boldsymbol{J}=S]\\
&=\sum_{U\subseteq[n]}\widehat{f}(U)^2\mathop{\mathbf{E}}_{\boldsymbol{J}\sim_\rho[n]}\big[|U\cup\boldsymbol{J}|\big]\\
&=\sum_{U\subseteq[n]}\widehat{f}(U)^2\rho|U|=\rho\cdot\mathbf{I}[f].
\end{aligned}$$ □

(The following two lectures have no Fourier analysis. We will study the Fourier implications of these lemmas after these lectures.)

10.1 Switching Lemma

We study the random restrictions of a DNF formula. Recall that while the negation of a DNF can be computed by a CNF of the same size and width, the DNF itself could require a CNF of exponential size to compute. We will show that if we apply a random restriction to a DNF, then a typical restricted DNF can be computed by a CNF of a similar complexity. Recall that a depth-k decision tree can be computed by a width-k CNF.

Lemma 10.1. *If $f\colon \{0,1\}^n \to \{0,1\}$ is computable by a DNF of width-w, then*

$$\Pr_{\boldsymbol{\tau}\sim R_\rho}[\mathsf{depth}_{DT}(f|_{\boldsymbol{\tau}}) \geq k] \leq (10\rho w)^k.$$

Suppose $\rho = 1/(100w)$, then the probability that $f|_{\boldsymbol{\tau}}$ cannot be computed by a depth-k is exponential small in k. Note that the bound on the right hand side is independent on the number of variables n.

To prove this lemma we first consider a decision tree of a special form that we call *canonical decision trees*. We will describe the canonical decision tree using the following query algorithm.

Algorithm 3: Query algorithm of a canonical decision tree (CDT)

Input: $x \in \{0,1\}^n$, with access to $f|_\tau$
Output: $\{0,1\}$

Set $j = 0$.
while $f|_{\tau\circ\pi_1\cdots\pi_j}$ *is not constant* **do**
  Set $j = j + 1$
  Let T_{i_j} be its first non-constant term in $f|_{\tau\circ\pi_1\cdots\pi_{j-1}}$
  Query the variables in T_{i_j}
  Let π_j be the answer, and σ_j be the satisfying assignment of T_{i_j}
  if $\pi_j = \sigma_j$, *i.e.,* $f_{\tau\pi_1\cdots\pi_j} \equiv 1$ **then return** 1.
end
return 0

A canonical decision tree queries all the variables in T_{i_j} at once, which is not necessarily the case for a general decision tree. Since a canonical decision tree is also a decision tree, we have that

$$\Pr_{\boldsymbol{\tau}\sim R_\rho}[\mathsf{depth}_{DT}(f|_{\boldsymbol{\tau}}) \geq k] \leq \Pr_{\boldsymbol{\tau}\sim R_\rho}[\mathsf{depth}_{CDT}(f|_{\boldsymbol{\tau}}) \geq k].$$

Let us call a restriction $\tau \in R_\rho$ *bad* if $\mathsf{depth}_{CDT}(f|_\tau) \geq k$. For each bad restriction τ, we can find a path of length k in CDT. Let us take the first such π to witness τ is bad. Observe that given a description of π, there is a 1-1 correspondence between τ and $\tau \circ \pi$, because we can use π to undo the restriction π in $\tau \circ \pi$ to get back τ. For a fixed witness π, we have

$$\mathbf{Pr}[R_\rho = \tau \circ \pi] = \mathbf{Pr}[R_\rho = \tau] \cdot \left(\frac{1-\rho}{2}\right)^k \cdot \left(\frac{1}{\rho}\right)^k \tag{6}$$

because in the restriction τ the k variables in π were retained by R_ρ with probability ρ^k, and the same k variables are randomly fixed by R_ρ in $\tau \circ \pi$ with probability $(\frac{1-\rho}{2})^k$. Since for a fixed ρ there is a bijection between τ and $\tau \circ \pi$, we have

$$\sum_{\substack{\text{bad } \tau \\ \text{with witness } \pi}} \mathbf{Pr}[R_\rho = \tau] = \sum_{\substack{\text{bad } \tau \\ \text{with witness } \pi}} \mathbf{Pr}[R_\rho = \tau \circ \pi] \cdot \left(\frac{2\rho}{1-\rho}\right)^k. \tag{7}$$

Summing over all the witnesses π, we have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\tau} \sim R_\rho}[\boldsymbol{\tau} \text{ is bad}] \leq \sum_{\pi} \sum_{\substack{\text{bad } \tau \\ \text{with witness } \pi}} \mathbf{Pr}[R_\rho = \tau] \leq (\#\text{ of witnesses } \pi) \cdot \left(\frac{2\rho}{1-\rho}\right)^k.$$

We can encode each witness π with $([n] \times \{0,1\})^k$, specifying the indices of the variables and their restricted value in π. Hence the number of π's can be bounded above by $(2n)^k$. However, our bound in Lemma 10.1 has no dependence on n, and we would like the n be replaced by w. We will show that by a slight tweak of the argument we can encode π with $([w] \times \{0,1\})^k$.

10.1.1 Succinct encoding of witnesses

We now explain how to tweak our previous argument to improve our encoding of the witnesses.

Warm up. Let us first consider the case $w = 1$, which captures the main idea. A width-1 DNF is simply an OR function. Suppose f is the OR function of 4 variables $x_1, \ldots, x_4$, and τ fixes x_2 to 0. It should be clear that $\mathsf{depth}_{CDT}(f|_\tau) \geq 3$. The restriction $\pi = \pi_1\pi_2 = (x_1 = 0, x_3 = 0)$ is a witness of $\mathsf{depth}_{CDT}(f|_\tau) \geq 2$, and there is a bijection between τ and $(\tau \circ \pi, \pi)$, where we encode the witness by $\pi \in ([n] \times \{0,1\})^2$. We now show that there is another bijection that admits a more succinct encoding of π.

Observe that no terms in $f|_\tau, f|_{\tau\pi_1}$ can be satisfied, for otherwise π is not a valid witness. We are going to use this observation to use a different restriction σ to encode the indices i_j of the terms T_{i_j} containing each of x_1, x_3.

Consider the pair $(\tau \circ \sigma, \pi)$, where $\sigma = \sigma_1\sigma_2 = (x_1 = 1, x_3 = 1)$. We claim that in this case, we do not even need π to recover π, that is, there is a bijection between $(\tau \circ \pi, \pi)$ and $\tau \circ \sigma$. To see this, given $f|_{\tau\circ\sigma_1\sigma_2} = 1 \vee 0 \vee 1 \vee x_4$, we look for the first term (= variable) that is satisfied. This reveals x_1 as the term that was restricted, and since $f|_{\tau\pi_1}$ cannot be satisfied, we must have $\pi_1 = (x_1 = 0)$. So now we consider $f|_{\tau\circ\pi_1\sigma_2} = 0 \vee 0 \vee 1 \vee x_4$ and again look for the first term that is not satisfied. This reveals x_2 as the restricted variable in σ_2 and π_2, and so by the same reasoning we must have $\pi_2 = (x_2 = 0)$.

The same idea extends to an arbitrary width-w DNF as follows. Recall the σ_j defined in the Algorithm 3. We claim that there is a bijection between τ and $(\tau \circ \sigma, \pi)$, where we encode π with elements in $([w] \times \{0,1\})^{t}$[3] , with $[w]$ being the relative indices of π_j within a term in f.[4] We describe how to decode in Algorithm 4.

[3]To be more precise, we also need to specify which variable is the last one in a term $T_{i,j}$ because distinct indices may share the same relative index across terms, but we can encode each index by $[w] \times \{\text{non-last}, \text{last}\}$, which only costs a factor of 2 in the base.

[4]Thanks to Aaron (Louie) Putterman for pointing out that this should be f and not $f|_\tau$, because in the decoding algorithm we do not know $f|_\tau$.

Algorithm 4: Decoding $(\pi\sigma, \pi)$ to τ

Input: $\tau\sigma$ and $\pi \in ([w] \times \{0,1\})^k$, with access to a width-w DNF f
Output: τ

Let $\sigma = \sigma_1 \cdots \sigma_s$ and $\pi = \pi_1 \cdots \pi_s$, where $\pi \in ([w] \times \{0,1\})^k$. **for** $j = 1, \ldots, s$ **do**
  Let T_{i_j} be the first term in $f|_{\tau \circ \pi_1 \cdots \pi_{j-1} \sigma_j \cdots \sigma_s}$ with $T_i = 1$
  Use π_j to locate indices of the restricted variables
  Replace σ_j with π_j
end
Undo the restriction π.

Finally, observe that $\mathbf{Pr}[R_\rho = \tau \circ \sigma] = \mathbf{Pr}[R_\rho = \tau \circ \pi]$ because both restrictions fix the same number of variables. So we can replace the former by the latter in Equation (6). Then in Equation (7), the number of witnesses can be bounded above by $(2w)^k$, giving us

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\tau} \sim R_\rho}[\boldsymbol{\tau} \text{ is bad}] \leq \left(\frac{4\rho w}{1-\rho}\right)^k.$$

10.1.2 An example

Consider

$$f(x_1, \ldots, x_8) := \underbrace{x_1 \overline{x_2} x_4}_{T_1} \vee \underbrace{x_1 x_2 \overline{x_5}}_{T_2} \vee \underbrace{\overline{x_4} x_6}_{T_3} \vee \underbrace{x_7 x_8}_{T_4}.$$

Let $\tau = (x_1 = 1)$, and so

$$f|_\tau(x_1, \ldots, x_8) = \overline{x_2} x_4 \vee x_2 \overline{x_5} \vee \overline{x_4} x_6 \vee x_7 x_8.$$

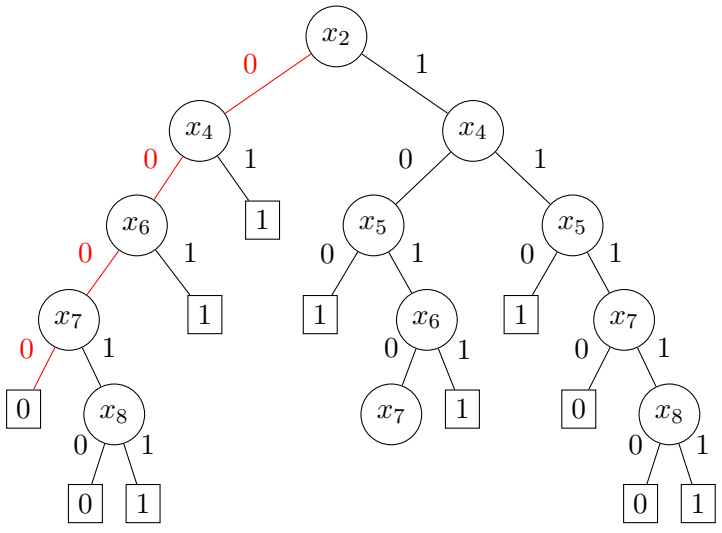

Figure 2: A canonical decision tree computing f. The red path of length-4 is our witness.

We have a witness $\pi = \pi_1\pi_2\pi_3$ for $\mathsf{depth}_{CDT}(f) \geq 4$, where $\pi_1 = (x_2 = 0, x_4 = 0)$, $\pi_2 = (x_6 = 0)$, and $\pi_3' = (x_7 = 0)$. We also have $\sigma = \sigma_1\sigma_2\sigma_3$, where $\sigma_1 = (x_1 = 0, x_4 = 1)$, $\sigma_2 = (x_6 = 1)$, and $\sigma_3 = (x_7 = 1)$. We encode each x_j being restricted by their relative index in term T_i being

looked at. So we encode π_1 as $(y_2 = 0, y_3 = 0)$, π_2 as $(y_2 = 0)$ and π_3' as $(y_1 = 0)$. Note that $\mathbf{Pr}[R_\rho = \tau] = \left(\frac{1-\rho}{2}\right) \cdot \rho^7$, and $\mathbf{Pr}[R_\rho = \tau \circ \pi] = \mathbf{Pr}[R_\rho = \tau \circ \sigma] = \left(\frac{1-\rho}{2}\right)^5 \cdot \rho^3$. We also have

$$f|_{\tau\sigma_1\sigma_2\sigma_3'} = (1 \cdot 1 \cdot 1) \vee (1 \cdot 0 \cdot x_5) \vee (0 \cdot 1) \vee (1 \cdot x_8) = 1 \vee 0 \vee 0 \vee x_8$$
$$f|_{\tau\pi_1\sigma_2\sigma_3'} = (1 \cdot 1 \cdot 0) \vee (1 \cdot 0 \cdot x_5) \vee (1 \cdot 1) \vee (1 \cdot x_8) = 0 \vee 0 \vee 1 \vee x_8$$
$$f|_{\tau\pi_1\pi_2\sigma_3'} = (1 \cdot 1 \cdot 0) \vee (1 \cdot 0 \cdot x_5) \vee (1 \cdot 0) \vee (1 \cdot x_8) = 0 \vee 0 \vee 0 \vee x_8.$$

11.1 Multi-switching lemma

Definition 11.1 (Partial decision tree)**.** We say m functions $f_1, \dots, f_m \colon \{0,1\}^n \to \{0,1\}$ are computable by a *q-partial decision tree of depth d* if there is a depth-d decision tree T such that for every root-to-leaf path τ in T, every $f_i|_\tau$ is computable by a decision tree of depth at most q.

In other words, the query algorithm for a q-partial depth-d decision tree can

1. *globally* query d bits of the input x, and then
2. for each of the restricted functions (induced by the global queries) $f_i|_\tau$, it can further *locally* query another q bits to compute its output. Note that it can make different local queries for different f_i's.

Lemma 11.2 (Multi-switching lemma (Impagliazzo–Matthews–Paturi 2012, Håstad 2014)**.** *Suppose $f_1, \dots, f_m$ are computable by DNFs of width-w. Then*

$$\mathbf{Pr}\Big[f_1, \dots, f_m \text{ do not have a } q\text{-partial decision tree of depth} < d\Big] \le m^{\lceil d/q \rceil} \cdot (100\rho w)^d.$$

Proof attempt. For each "bad" restriction τ, we find a path γ of length d in the "canonical partial decision tree" and encode it succinctly. We will consider a path γ in the partial decision tree by going through as few DNFs as possible as follows.

Specifically, we look at the first DNF f_{ℓ_1} with $\mathsf{depth}_{DT}(f_{\ell_1}|_\tau) =: k_1 > q$, then find a restriction π_1 of length k_1 that witnesses $\mathsf{depth}_{CDT}(f_{\ell_1}|_\tau) > q$. Then we apply the restriction π_1, and look at the first DNF f_{ℓ_2} with $\mathsf{depth}_{DT}(f_{\ell_1}|_{\tau\pi_1}) =: k_2 > q$, and find a restriction π_2 that further fixes another $k_2 > q$ variables. Continuing this way, we can find a restriction $\pi := \pi_1 \cdots \pi_t$ of d variables by going through $t \le \lceil d/q \rceil$ many DNFs.

Our final witness for a bad restriction would be a list ℓ of the indices ℓ_i's of the DNFs we looked at, and the encoding of the restriction $\pi = \pi_1 \cdots \pi_t$. As in the proof of the switching lemma, we encode π succinctly in $([w] \times \{0,1\})^d$, by considering the bijection between τ and $(\tau\sigma_1 \cdots \sigma_t, L, \pi_1 \cdots \pi_t)$, where σ_i is another restriction on the same k_i variables as π_i that helps us identify the terms in f_{ℓ_i} that contains the variables in π_i.

There are at most $m^{\lceil d/q \rceil}$ different ways of choosing $\lceil d/q \rceil$ out of m DNFs for our list ℓ, and so we can conclude that the probability R_ρ is bad is at most $m^{\lceil d/q \rceil} \cdot (10\rho w)^d$.

Issue. The argument above fails because the local restriction π_1 can globally affect the complexity of the rest of the DNFs: it could be the case that after fixing π_1, the DNF $f_1|_{\tau\pi}, \dots, f_m|_{\tau\pi_1}$ have a q-partial decision tree of small depth. The statement "$f_1|_\tau, \dots, f_m|_\tau$ do not have a q-partial decision tree of depth d" only means that for every decision tree of depth d, there is a restriction γ and $\ell \in [m]$ such that $\mathsf{depth}_{DT}(f_\ell|_\tau) > q$. There is no reason to believe π_1 is the "right prefix" of γ.

Solution. To get around this issue, we will use a global witness γ_1 together with the local witness π_1. The global witness γ_1 restricts the *same* set of variables I_1 as π_1 but assigns them to some different values in order to witness the *global* condition, namely, the DNFs $f_1|_{\tau\circ\gamma_1}, \dots, f_m|_{\tau\circ\gamma_1}$ have no q-partial decision of depth $> d - |\gamma_1|$. (There is always such a choice of γ_1, otherwise $f_1|_\tau, \dots, f_m|_\tau$ have a q-partial decision tree of depth $(d - |\gamma_1|) + |\gamma_1| = d$.) The local witness π is then used to succinctly encode the indices of the variables in γ using the switching lemma argument.

Overall argument. If $\vec{f}|_\tau := (f_1|_\tau, \dots, f_m|_\tau)$ do not have a q-partial decision tree of depth d, then there exists a DNF $f_{\ell_1}|_\tau$ with $\mathsf{depth}_{CDT}(f_{\ell_1}|_\tau) =: k_1 > q$ and thus it contains $k > q$ unfixed variables. We fix all these variables by some γ_1 so that $\vec{f}|_{\tau\gamma_1}$ do not have a q-partial decision tree of depth $d - |\gamma_1|$. Then we find a DNF $f_{\ell_2}|_{\tau\circ\gamma_1}$ with $\mathsf{depth}_{CDT}(f_{\ell_2}|_{\tau\circ\gamma_1}) =: k_2 > q$, and fix all these k_2 variables by some γ_2 so that $\vec{f}_{\tau\circ\gamma_1\gamma_2}$ do not have a q-partial decision tree of depth $d - |\gamma_1| - |\gamma_1|$. Repeating the same argument, we have a restriction $\gamma_1, \dots, \gamma_t$ on d variables for some $t \le \lceil d/q \rceil$. Given $\gamma := \gamma_1, \dots, \gamma_t \in [n]^{\le \lceil d/q \rceil}$, there is a bijection between τ and $\tau \circ \gamma_1 \cdots \gamma_t$ and so we have the relation

$$\sum_{\substack{\text{bad } \tau \text{ with} \\ \text{witness } \gamma}} \mathbf{Pr}[R_\rho = \tau] \le \sum_{\substack{\text{bad } \tau \text{ with} \\ \text{witness } \gamma}} \mathbf{Pr}[R_\rho = \tau \circ \gamma_1 \cdots \gamma_t] \left(\frac{2\rho}{1-\rho}\right)^d \le 1 \cdot \left(\frac{2\rho}{1-\rho}\right)^d.$$

We now show how to encode γ succinctly. Recall in the proof of the switching lemma that if $\mathsf{depth}_{CDT}(f_{\ell_1}|_{\tau\circ\gamma_1\cdots\gamma_{i-1}}) = k_i$, then there is a restriction π_i on the same k variables as γ_i witnessing a path of length k_i in its canonical decision tree, and there is another restriction σ_i on the same k variables so that given $(\tau \circ \gamma_1 \cdots \gamma_{i-1}\sigma_i, \pi_i)$, where $\pi_i \in ([w] \times \{0,1\})^{k_i}$, we can identify the variables in π_i.

So given $\ell_i \in [m], \pi_i \in ([w] \times \{0,1\})^k$ and $\gamma_i \in \{0,1\}^k$, we can identity the variables in π_i and recover $\tau \circ \gamma_1 \cdots \gamma_i$ from $\tau \circ \gamma_1 \cdots \gamma_{i-1} \circ \sigma_i$. Repeating this iteratively, we can recover $\tau \circ \gamma_1 \cdots \gamma_t$ from $\tau \circ \sigma_1 \cdots \sigma_t$, given $\ell := (\ell_1, \dots, \ell_t) \in [m]^t$, $\pi := \pi_1, \dots, \pi_t \in ([w] \times \{0,1\}^t)$ and $\gamma := \gamma_1, \dots, \gamma_t \in \{0,1\}^t$. In particular, we can identify the variables in $\gamma_1, \dots, \gamma_t$, which allows us to recover τ. So we have the relation

$$\sum_{\substack{\text{bad } \tau \text{ with} \\ \text{witness } \ell, \pi, \gamma}} \mathbf{Pr}[R_\rho = \tau] \le \sum_{\substack{\text{bad } \tau \text{ with} \\ \text{witness } \ell, \pi, \gamma}} \mathbf{Pr}[R_\rho = \tau \circ \sigma_1 \cdots \sigma_t] \left(\frac{2\rho}{1-\rho}\right)^d \le 1 \cdot \left(\frac{2\rho}{1-\rho}\right)^d,$$

and we can now bound above the number of witnesses by $m^{\lceil d/q \rceil} \cdot (2w)^d \cdot 2^d \le m^{\lceil d/q \rceil} \cdot (100\rho w)^d$. [5]

We now go through the details. First we explain how to construct our canonical partial decision tree in Algorithm 5. Note that the ℓ_i's in Algorithm 5 may not be distinct, because unlike π_i, the restriction γ_i does not necessarily make f_{ℓ_i} constant. We choose the π_i's so that each of them restricts at least $\mathsf{depth}_{CDT}(f_{\ell_i}|_{\tau\circ\gamma_1,\dots,\gamma_{i-1}}) > q$ variables.

Let $L = (\ell_1, \dots, \ell_t)$, $\pi = \pi_1 \cdots \pi_t$, $\gamma = \gamma_1 \cdots \gamma_t$, and $\sigma = \sigma_1 \cdots \sigma_t$. Note that all 3 restrictions $\pi_i, \gamma_i, \sigma_i$ restrict the same set of variables but to different values. Given L, π and γ, we show in Algorithm 6 that the map $\tau \leftrightarrow \tau\sigma$ is bijective.[6]

We can encode L by $[m]^{\lceil d/q \rceil}$, $\pi \in ([w] \times \{0,1\})^d$, and $\gamma \in \{0,1\}^d$. Therefore,

$$\sum_{\substack{\text{bad } \tau \text{ with} \\ \text{witness } L, \pi, \gamma}} \mathbf{Pr}[R_\rho = \tau] \le \sum_{\substack{\text{bad } \tau \text{ with} \\ \text{witness } L, \pi, \gamma}} \mathbf{Pr}[R_\rho = \tau\sigma] \left(\frac{2\rho}{1-\rho}\right)^d \le 1 \cdot \left(\frac{2\rho}{1-\rho}\right)^d.$$

So

$$\mathbf{Pr}[R_\rho \text{ is bad}] \le m^{\lceil d/q \rceil} \cdot (2w)^d \cdot 2^d \cdot \left(\frac{2\rho}{1-\rho}\right)^d \le m^{\lceil d/q \rceil} \cdot (100\rho w)^d. \quad \square$$

[5] More precisely, we encode $\ell \in ([m] \cup \{*\})^{\lceil d/q \rceil}$ because the number of DNFs is at most but not exactly $\lceil d/q \rceil$, and encode $\pi \in ([w] \times \{\text{non-last}, \text{last}\} \times \{0,1\})^d$ to indicate whether the restricted variable being looked at is the last one so that we can move on to the next term. We are ignoring these technicalities since it only affect by a factor of $O(1)^d$.

[6] More precisely, we do not know s, but we can handle this using the encoding in Footnote 5.

Algorithm 5: Query algorithm of a canonical partial decision tree

Input: A global string $x \in \{0,1\}^n$, with access to an auxiliary local string $y \in \{0,1\}^n$, the restriction τ, and $f_1, \ldots, f_m$

Output: A partial assignment γ

Set $t = 0$
while $|\gamma| < d$ **do**
  if $\mathsf{depth}_{DT}(f_\ell|_{\tau\circ\gamma_1\cdots\gamma_t}) \le q$ *for each* $\ell \in [m]$ **then** STOP
  Set $t = t + 1$
  Let ℓ_t be the first ℓ with $\mathsf{depth}_{CDT}(f_\ell|_{\tau\circ\gamma_1\cdots\gamma_{t-1}}) > q$
  Set $s = 0$.
  while $f_{\ell_t}|_{\tau\circ\gamma_1\cdots\gamma_{t-1}\circ\pi_{t,1}\cdots\pi_{t,s}}$ *is not constant and* $|\gamma_1\cdots\gamma_{t-1}\circ\pi_{t,1}\cdots\pi_{t,s}| < d$ **do**
    Set $s = s + 1$
    Let $T_{i_{t,s}}$ be its first non-constant term, and $I_{t,s}$ be the indices of variables in $T_{i_{t,s}}$
    Query the local variables $y_{I_{t,s}}$
    Let $\pi_{t,s}$ be the answer, and $\sigma_{t,s}$ be the satisfying assignment of $T_{i_{t,s}}$
    if $\pi_{t,s} = \sigma_{t,s}$ **then** STOP and return ERROR
  end
  Let $I_t := \cup_{u=1}^{s} I_{t,u}$
  Query the global variables x_{I_t}
  Let γ_t be the answer
end
return $\gamma_1\cdots\gamma_t$

Algorithm 6: Decoding $\tau\sigma$ to τ given L, τ, γ

Input: $\tau \circ \sigma_1 \cdots \sigma_t$, given $L = (\ell_1, \ldots, \ell_t) \in [m]^{\le \lceil d/q \rceil}$, $\pi = \pi_1 \cdots \pi_t \in ([w] \times \{0,1\})^d$ and $\gamma = \gamma_1 \cdots \gamma_t \in \{0,1\}^d$ and description of $f_1, \ldots, f_m$.

Output: τ

for $i = 1, \ldots, t$ **do**
  Let $\pi_i = (\pi_{i,1}, \ldots, \pi_{i,s})$
  for $j = 1, \ldots, s$ **do**
    Let T_{i_j} be the first satisfied term in $f_{\ell_i}|_{\tau\circ\gamma_1\cdots\gamma_{i-1}\circ\pi_{i,1}\cdots\pi_{i,j-1}\sigma_{i,j}\cdots\sigma_{i,s}}$
    Replace $\sigma_{i,j}$ with $\pi_{i,j}$
  end
  Replace π_i with γ_i
end
Undo the restriction γ

12.1 Spectral concentration of DNFs

We look at some Fourier implication of the switching lemma and multi-switching lemma. We first prove a sharp bound on the low-degree concentration result for DNFs, improving the one we proved before using the bound on the total influence of DNFs. We first set up a definition for convenience.

Definition 12.1. The *Fourier tail of f at level k*, denoted $W^{\geq k}[f]$, is

$$W^{\geq k}[f] := \sum_{|S|\geq k} \widehat{f}(S)^2 = \mathop{\mathbf{Pr}}_{\boldsymbol{S}\sim\mathcal{S}}\big[|S| \geq k\big].$$

Lemma 12.2. *If $f\colon \{-1,1\}^n \to \{-1,1\}$ is computable by a width-w DNF, then $W^{\geq k}[f] \leq 2^{-\Omega(k/w)}$. In other words, f is ϵ-concentrated on degree at most $O(w\log(1/\epsilon))$.*

Recall the switching lemma says that with high probability over a ρ-random restriction on a width-w DNF f, the restricted function can be computed by a small-depth decision tree, which is 0-concentrated on the low degree. We will relate the tail of f to the tail of the random restriction of f.

Claim 12.3. $W^{\geq k}[f] \leq 2\,\mathbf{E}_{\boldsymbol{\tau}\sim R_\rho}[W^{\geq \frac{\rho k}{2}}[f|_{\boldsymbol{\tau}}]]$.

Proof. Recall from Proposition 9.5 that for a ρ-random restriction $(\boldsymbol{J},\boldsymbol{z}) \sim R_\rho$, we have

$$\mathop{\mathbf{E}}_{(\boldsymbol{J},\boldsymbol{z})\sim R_\rho}\big[\widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(S)^2\big] = \sum_{U\subseteq[n]} \widehat{f}(U)^2 \mathop{\mathbf{Pr}}_{\boldsymbol{J}\sim\rho[n]}[U\cap\boldsymbol{J} = S].$$

So

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\boldsymbol{J},\boldsymbol{z})\sim R_\rho}\Big[\sum_{|S|\geq k} \widehat{f_{\boldsymbol{J},\boldsymbol{z}}}(S)^2\Big] &= \sum_{U\subseteq[n]} \widehat{f}(U)^2 \sum_{|S|\geq k} \mathop{\mathbf{Pr}}_{\boldsymbol{J}\sim\rho[n]}\big[U\cap\boldsymbol{J} = S\big] \\
&= \sum_{U\subseteq[n]} \widehat{f}(U)^2 \mathop{\mathbf{Pr}}_{\boldsymbol{J}\sim\rho[n]}[|U\cap\boldsymbol{J}| \geq k] \\
&= \sum_{U\subseteq[n]} \widehat{f}(U)^2\, \mathbf{Pr}[\mathsf{Bin}(|U|,\rho) \geq k] \\
&\geq \frac{2}{3}\sum_{|U|\geq 2k/\rho} \widehat{f}(U)^2 = \frac{2}{3}\cdot W^{\geq 2k/\rho}[f],
\end{aligned}$$

where the inequality is because whenever $|U| \geq 2k/\rho$, we have $\mathbf{Pr}[\mathsf{Bin}(|U|,\rho) \geq k] \geq 2/3$. □

Proof of Lemma 12.2. We apply Claim 12.3 and the switching lemma with $\rho = 1/(20w)$,

$$\begin{aligned}
\frac{1}{2}W^{\geq k}[f] &\leq \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\big[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\big] \\
&\leq \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\big[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(\mathsf{depth}_{DT}(f_{\boldsymbol{\tau}}) < \rho k/2)\big] + \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\big[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(\mathsf{depth}_{DT}(f_{\boldsymbol{\tau}}) \geq \rho k/2)\big] \\
&\leq (10\rho w)^{\rho k/2} \leq 2^{-\Omega(k/w)}.
\end{aligned}$$

□

13.1 Spectral concentration of small-depth circuits

We now use the multi-switching lemma to extend our low-degree concentration result for the class of constant-depth circuits, a generalization of DNF formulas.

Definition 13.1. A depth-D circuit on n variables $x_1, \ldots, x_n$ is a layered directed acyclic graph with $D+1$ layers. The bottom layer 0 has $2n$ nodes representing x_i and $\overline{x_i}$ and the top layer D has exactly 1 node. Each edge goes from some $(j-1)$-th layer to the j-th layer. For layers $j \geq 1$, all nodes in the same layer have the same label $\wedge$ or $\vee$. Nodes in adjacent layers have different labels, so the labels alternate between layers. Each node computes the function labeled by the node with inputs being the functions computed by their incoming nodes.

The *size* of a circuit is the number of nodes in layers ≥ 1, i.e., the number of AND/OR gates in the graph.

Note that CNFs and DNFs are depth-2 circuits. We use AC^0 to denote the class of constant-depth circuits. (AC stands for alternating circuits; the superscript 0 stands for depth $O((\log n)^0)$. Sometimes AC^0 circuits are assumed to have size polynomial in n, but here we will always specify the size of the circuits when we talk about them. The class AC^0 was first introduced by Furst, Saxe and Sipser (1981), and Ajtai (1982), who showed that the parity function on n bits cannot be computed by circuits of constant depth and polynomial size. Furst, Saxe and Sipser observed a connection between AC^0 and the polynomial hierarchy (PH) which can be used to give oracle separation between PH and PSPACE. Shortly after, Yao (1985) and Håstad (1986) strengthened their results and gave exponential lower bounds on the size of AC^0 circuits computing the parity function. The result by Håstad also implicitly gave *correlation bounds* of AC^0 and the parity function. Specifically, it showed that every function f computable by a depth-D circuit of size $m \leq 2^{o(1/n^{\frac{1}{D-1}})}$ must satisfy

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}}[f(\boldsymbol{x}) = \mathsf{Parity}_n(\boldsymbol{x})] \leq 1/2 + 2^{-\Omega(n^{\frac{1}{D-1}})}.$$

Finally, Impagliazzo, Matthews, Paturi (2012) and Håstad (2014) gave tight correlation bounds. They showed that every function f computable by a depth-D size-m circuit must satisfy

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}}[f(\boldsymbol{x}) = \mathsf{Parity}_n(\boldsymbol{x})] \leq 1/2 + 2^{-\frac{n}{O(\log m)^{D-1}}}.$$

It is tight in the sense that a depth-D circuit of size $2^{n^{O(\frac{1}{D-1})}}$ can compute the parity function on n bits exactly.

Let us show how to construct a depth-3 circuit of size $2^{O(\sqrt{n})}$ that computes the parity function; the same idea generalizes to larger d straightforwardly. Recall that every Boolean function on n bits can be trivially computed by a decision tree of size at most $2 \cdot 2^n$ (by exhausting the 2^n possible inputs), and thus can be computed by both a CNF and a DNF of size at most $2 \cdot 2^n$. We divide the n input bits into blocks of $\sqrt{n}$ bits, then compute the parity of the parities of the $\sqrt{n}$ blocks. The parity of each of the $\sqrt{n}$ blocks can be computed by a DNF of size $2 \cdot 2^{\sqrt{n}}$, and the parity of the $\sqrt{n}$ outputs can be computed by a CNF of size $2 \cdot 2^{\sqrt{n}}$. So we have an AND of ORs of ORs of ANDs, and merging the two levels of ORs gives us a depth-3 circuit of size $O(\sqrt{n}) \cdot 2^{\sqrt{n}} = 2^{O(\sqrt{n})}$.

Note that this lower bound remains state-of-the-art for depth-3 circuits.

Open Problem 13.2. *Give an explicit function f (say polynomial-time computable) such that every depth-3 circuit computing f has size $2^{\omega(\sqrt{n})}$.*

We will prove the following result by Tal (2015) which generalizes the above correlation bound.

Theorem 13.3. *If f is computable by a depth-D circuit of size m, then*

$$W^{\geq k}[f] \leq 2 \cdot 2^{-\frac{k}{O(\log m)^{D-1}}}.$$

In other words, f is ϵ-concentrated on degree at most $O(\log^{D-1}(m) \cdot \log(1/\epsilon))$.

Note that by setting $k = n$ we recover the above correlation bound, because $W^{\geq n}[f] = \widehat{f}([n])^2 = \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})\chi_{[n]}(\boldsymbol{x})]^2$. Also note that Theorem 13.3 is also a tight strengthening of the Linial–Mansour–Nisan result in O'Donnell's book (Chapter 4.5), which showed that depth-D circuits are ϵ-concentrated on degree $O(\log^D(m/\epsilon))$.

13.1.1 Proof of Theorem 13.3

We now prove Theorem 13.3. We will prove a more general bound which depends on the bottom fan-in of the circuit and its *effective size*, defined to be the number of gates in layers ≥ 2.

Lemma 13.4. *Suppose f is computable by a depth-D circuit of effective size m and bottom fan-in w. Then*

$$W^{\geq k}[f] \leq 2 \cdot 2^{-\frac{k}{w \cdot O(\log m)^{D-2}}}.$$

To see how Lemma 13.4 implies Theorem 13.3, observe that by adding a dummy layer of fan-in 1 AND (or OR) gates between the inputs and the gates at the bottom layer of a depth-D circuit of size m, we obtain a depth-$(D+1)$ circuit of effective size m with bottom fan-in 1. Now applying Lemma 13.4 gives us the bound we want.

We will prove Lemma 13.4 for the special case of AND of DNFs (a depth $D = 3$ circuit) by reducing it to a CNF (a depth $D = 2$ circuit) using random restrictions; the general case can be proved by induction on D.

Like the last lecture, we are going to relate the tail of the circuit to (the average of) the tail of its random restrictions. Suppose the bottom two layers of the circuit are ORs of ANDs (i.e. DNFs). We will apply a random restriction to switch the restricted DNFs into CNFs, so that the circuit becomes an AND of ANDs of ORs, and we can collapse the two adjacent layers of ANDs, resulting a depth-2 circuit, for which we can apply our Fourier tail bound for DNFs/CNFs (Lemma 12.2) proved in the last lecture.

There are at most m DNFs at the bottom two levels (because the circuit has effective size at most m). Let us call them $g_1, \ldots, g_{m'}$ for some $m' \leq m$. We first apply a random restriction $\boldsymbol{\tau}$, so that with high probability over $\boldsymbol{\tau}$, the restricted DNFs can be computed by a depth $(\log m)$-partial depth-d decision tree. Conditioned on this "good" event, if we further fix any path π in the partial decision tree, each of the restricted DNFs $g_i|_{\boldsymbol{\tau} \circ \pi}$ can be computed by a depth $(\log m)$-decision tree, and therefore can be computed by a CNF of width $\log m$. So the restricted circuit (under $\boldsymbol{\tau} \circ \pi$) is a depth-2 circuit of effective size at most m and bottom fan-in $\log m$.

We now prove it formally. The following lemma follows from choosing $q = \log m$ in the multi-switching lemma (Lemma 11.2), and observing that $m^{\lceil d/q \rceil} \leq m^{d/q+1} \leq m \cdot 2^d$.

Lemma 13.5. *Suppose $f_1, \ldots, f_m$ be width-w CNFs. Then*

$$\Pr_{\boldsymbol{\tau} \sim R_\rho}\left[f_1|_{\boldsymbol{\tau}}, \ldots, f_m|_{\boldsymbol{\tau}} \text{ do not have a } (\log m)\text{-partial decision tree of depth } d\right] \leq m \cdot (200\rho w)^d.$$

Let us restate our claim that relates the tail of f to (the average of) the tail of its random restrictions.

Claim 13.6. $W^{\geq k}[f] \leq 2\,\mathbf{E}_{\boldsymbol{\tau}\sim R_\rho}[W^{\geq \frac{\rho k}{2}}[f|_{\boldsymbol{\tau}}]]$.

We also need the following claim to relate the tail of f to its (deterministic) restrictions.

Claim 13.7. *Let T be a depth-d (partial) decision tree. If for every root-to-leaf path π we have $W^{\geq k-d}[f|_\pi] \leq \epsilon$, then $W^{\geq k}[f] \leq \epsilon$.*

We will prove Claim 13.7 in Homework 3.

Proof of Lemma 13.4. Let f be computable by an AND of DNFs of effective size m and bottom fan-in w. Let $g_1, \ldots, g_{m'}$ be the $m' \leq m$ DNFs in the bottom 2 layers. Let E be the event "$g_1|_{\boldsymbol{\tau}}, \ldots, g_{m'}|_{\boldsymbol{\tau}}$ have a $(\log m)$-partial decision tree of depth d." We have

$$\begin{aligned}\frac{1}{2}W^{\geq k}[f] &\leq \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\right] \\ &\leq \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(E)\right] + \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(\overline{E})\right].\end{aligned}$$

Conditioned on E, let T be the depth-d partial decision tree computing $g_1|_{\boldsymbol{\tau}}, \ldots, g_{m'}|_{\boldsymbol{\tau}}$. By the definition of partial decision tree, for every fixed path π in T, each $g_i|_{\boldsymbol{\tau}\pi}$ can be computed by a CNF of width $\log m$. So $f|_{\boldsymbol{\tau}\pi}$ can be computed by an AND of CNFs, which by collapsing the top two layers of ANDs, is a CNF of size at most m. (This is where we use effective size, as the switch may introduce more gates in the bottom layer.) So by Claim 13.7, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\right] \leq \max_{\pi\in T} W^{\geq \rho k/2-d}[f|_{\boldsymbol{\tau}\circ\pi}].$$

We choose $d = \rho k/4$ and $\rho = 1/(400w)$. Then by Lemma 12.2 the above is at most $2^{-\frac{\rho k}{O(\log m)}} \leq 2^{-\frac{k}{w\cdot O(\log m)}}$.

It remains to bound above the probability of the bad event. We have

$$\mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(\overline{E})\right] \leq \mathbf{Pr}[\overline{E}] \leq m\cdot(100\rho w)^d \leq m\cdot 2^{-k/O(w)} \leq 2\cdot 2^{-\frac{k}{w\cdot O(\log m)}}.$$

So $\mathbf{E}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(E)\right] + \mathbf{E}_{\boldsymbol{\tau}\sim R_\rho}\left[W^{\geq \rho k/2}[f|_{\boldsymbol{\tau}}]\cdot \mathbb{1}(\overline{E})\right] \leq 2\cdot 2^{-\frac{k}{w\cdot O(\log m)}}$, as desired.

□

14.1 Bonami's lemma

Consider the random variable $\boldsymbol{X}$, where $\boldsymbol{X} = n$ with probability $1/2$ and $-n$ otherwise. This random variable has mean 0, but it puts no mass near the point 0; so knowing its mean reveals very little information about about $\boldsymbol{X}$. In many scenarios we would like a random variable to possess some nice properties such as concentration and anti-concentration. We will look at some conditions for a random variable to satisfy these properties. One condition is the following.

Definition 14.1. A real random variable $\boldsymbol{X}$ is B-reasonable if $\mathbf{E}[\boldsymbol{X}^4] \leq B \cdot \mathbf{E}[\boldsymbol{X}^2]^2$.

We can state the definition in terms of the *norm* of $\boldsymbol{X}$.

Definition 14.2 (norms of a random variable)**.** The q-norm of a real random variable $\boldsymbol{X}$, denoted by $\|\boldsymbol{X}\|_q$ is defined as

$$\|\boldsymbol{X}\|_q = \mathbf{E}[|\boldsymbol{X}|^q]^{1/q}.$$

So, $\boldsymbol{X}$ is B is reasonable if $\|\boldsymbol{X}\|_4 \leq B^{1/4}\|\boldsymbol{X}\|_2$.
Let us consider some examples.

1. $\boldsymbol{x} \sim \{-1, 1\}$. We have $\mathbf{E}[\boldsymbol{x}^4] = \mathbf{E}[\boldsymbol{x}^2]^2 = 1$ and thus $\boldsymbol{x}$ is 1-reasonable.
2. $\boldsymbol{g} \sim \mathcal{N}(0, 1)$, where $\mathcal{N}(0, 1)$ is the standard Gaussian. The $2k$-th moment $\mathbf{E}[\boldsymbol{g}^{2k}]$ of $\boldsymbol{g}$ is at most $(2k)!! = (2k-1)(2k-3)\cdots 1$. So $\mathbf{E}[\boldsymbol{g}^4] = 3$ and $\mathbf{E}[\boldsymbol{g}^2]^2 = 1$ and thus $\boldsymbol{g}$ is 3-reasonable.
3. $\boldsymbol{u} \sim [-1, 1]$. We have $\mathbf{E}[\boldsymbol{u}^4] = \frac{1}{2}\int_{-1}^{1} u^4 du = \frac{1}{5}$, and $\mathbf{E}[\boldsymbol{u}^2] = \frac{1}{2}\int_{-1}^{1} u^2 du = \frac{1}{3}$. So $\boldsymbol{u}$ is $9/5$-reasonable.
4. Consider the $\mathsf{AND}\colon \{0,1\}^n \to \{0,1\}$ function. We have $\mathbf{E}_{\boldsymbol{x}\sim\{0,1\}^n}[\mathsf{AND}(\boldsymbol{x})^4] = \mathbf{E}_{\boldsymbol{x}\sim\{0,1\}^n}[\mathsf{AND}(\boldsymbol{x})^2] = 2^{-n}$. So the random variable $\mathsf{AND}(\boldsymbol{x})$, where $\boldsymbol{x} \sim \{0,1\}^n$ is 2^n-reasonable.

Soon we will look at Boolean function $f\colon \{-1, 1\} \to \mathbb{R}$ such that $f(\boldsymbol{x})$ (on a uniform $\boldsymbol{x} \sim \{-1, 1\}^n$) is reasonable. Already in the last example, we saw that $\mathsf{AND}(\boldsymbol{x})$, which has degree-$n$, is not really reasonable as 2^n is large. The goal of this lecture is showing that $f(\boldsymbol{x})$ is reasonable when f has low degree.

But first let us show that reasonable random variables give concentration and anti-concentration bounds. Recall the proof of Chebyshev's inequality in basic probability. When we have a bound on the higher moment of a random variable, we can apply a similar argument to get some concentration bound.

Proposition 14.3. *Suppose $\boldsymbol{X} \not\equiv 0$ is B-reasonable. Then $\mathbf{Pr}[|\boldsymbol{X}| \geq t\|\boldsymbol{X}\|_2] \leq B/t^4$.*

Proof. By Markov's inequality, $\mathbf{Pr}[\boldsymbol{x}^4 \geq t^4\,\mathbf{E}[\|\boldsymbol{X}\|_2^4] \leq \frac{\mathbf{E}[X^4]}{t^4\cdot\|\boldsymbol{X}^4\|} \leq \frac{B}{t^4}$. □

Reasonable random variables also satisfy the following anti-concentration property.

Proposition 14.4. *Suppose $\boldsymbol{X} \not\equiv 0$ is B-reasonable. Then $\mathbf{Pr}[|\boldsymbol{X}| > t \cdot \|\boldsymbol{X}\|_2] \geq \frac{(1-t^2)^2}{B}$ for every $t \in [0, 1]$.*

Proof. This is similar to how one proves the Paley–Zygmund inequality (also called the 2nd moment method). We consider $\mathbf{Pr}[\boldsymbol{X}^2 > t^2 \cdot \|\boldsymbol{X}\|_2^2]$. The idea is to relate the probability with the expectation of $\boldsymbol{X}^2$ as follows:

$$\mathbf{E}[\boldsymbol{X}^2] = \mathbf{E}\big[\boldsymbol{X}^2 \cdot \mathbb{1}(\boldsymbol{X}^2 > t^2\|\boldsymbol{X}\|_2^2)\big] + \mathbf{E}\big[\boldsymbol{X}^2 \cdot \mathbb{1}(\boldsymbol{X}^2 \le t^2\|\boldsymbol{X}\|_2^2)\big].$$

The second term on the right hand side can be bounded above by $t^2\|\boldsymbol{X}\|_2^2$; the first term can be bounded using Cauchy–Schwarz by

$$\mathbf{E}[|\boldsymbol{X}|^4]^{1/2} \cdot \mathbf{E}\big[\mathbb{1}(\boldsymbol{X}^2 > t^2\|\boldsymbol{X}\|_2^2)^2\big]^{1/2} = \mathbf{E}[\boldsymbol{X}^4]^{1/2} \cdot \mathbf{Pr}\big[\boldsymbol{X}^2 > t^2\|\boldsymbol{X}\|_2^2\big]^{1/2}.$$

Rearranging gives

$$\mathbf{Pr}\big[\boldsymbol{X}^2 > t^2\|\boldsymbol{X}\|_2^2\big]^{1/2} \ge \frac{\mathbf{E}[\boldsymbol{X}^2] - t^2\|\boldsymbol{X}\|_2^2}{\mathbf{E}[\boldsymbol{X}^4]^{1/2}} \ge (1-t^2)\frac{\mathbf{E}[\boldsymbol{X}^2]}{\mathbf{E}[\boldsymbol{X}^4]^{1/2}} \ge \frac{1-t^2}{\sqrt{B}}.$$

Squaring both sides completes the proof. □

One condition for a random variable to be reasonable is that it has low min-entropy.

Proposition 14.5. *Let $\boldsymbol{X}$ be a real random variable such that $\min_{x\in\mathrm{Supp}(\boldsymbol{X})} \mathbf{Pr}[\boldsymbol{X} = x] \ge \lambda$. Then $\boldsymbol{X}$ is $(1/\lambda)$-reasonable.*

Proof. Let $M := \max_{x\in\mathrm{Supp}(\boldsymbol{X})}|x|$. Then $\mathbf{E}[\boldsymbol{X}^2] \ge \lambda M^2$, and so

$$\mathbf{E}[\boldsymbol{X}^4] \le \mathbf{E}[\boldsymbol{X}^2 \cdot \boldsymbol{X}^2] \le M^2\,\mathbf{E}[\boldsymbol{X}^2] \le \frac{1}{\lambda}\,\mathbf{E}[\boldsymbol{X}^2]^2.$$ □

We can see that for the proof to go through, we only need the weaker condition $\mathbf{Pr}[\boldsymbol{X} = M] \ge \lambda$.

14.1.1 Low-degree polynomials are reasonable

Given $\boldsymbol{x} \sim \{-1,1\}^n$, we would like to characterize the functions $f\colon \{-1,1\}^n \to \mathbb{R}$ such that $f(\boldsymbol{x})$ is reasonable. The Bonami's lemma says that having low degree is a sufficient condition.

Lemma 14.6 (Bonami's lemma)**.** *Suppose $f\colon \{-1,1\}^n \to \mathbb{R}$ has degree k. Then $f(\boldsymbol{x})$ is 9^k-reasonable for $\boldsymbol{x} \sim \{-1,1\}^n$. That is, $\mathbf{E}[f(\boldsymbol{x})^4] \le 9^k \cdot \mathbf{E}[f(\boldsymbol{x})^2]^2$.*

We will decompose f into two parts. Once we have done that, the proof is a simple induction on the number of variables n.

Recall in Definition 4.7 we defined the derivative operator $D_i f$ as

$$D_i f(x) = \frac{f(x^{i\mapsto 1}) - f(x^{i\mapsto -1})}{2} = \sum_{S\ni i} \widehat{f}(S) x^{S\setminus\{i\}}.$$

We define another operator called the *expectation operator*, denoted $E_i f$, defined as

$$E_i f(x) = \mathop{\mathbf{E}}_{\boldsymbol{x}_i}[f(x_1,\dots,x_{i-1},\boldsymbol{x}_i,x_{i+1},\dots,x_n)] = \frac{f(x^{i\mapsto 1}) + f(x^{i\mapsto -1})}{2} = \sum_{S\not\ni i} \widehat{f}(S) x^S.$$

Observe that both $D_i f(x)$ and $E_i f(x)$ do not depend on x_i. (So it is a function on the $n-1$ variables $x_1,\dots,x_{i-1},x_{i+1},\dots,x_n$.) Also, if f has degree $k \ge 1$ then $D_i f(x)$ has degree at most $k-1$. It is straightforward to verify that $f(x) = x_i D_i(x) + E_i f(x)$ for every $i \in [n]$. We now prove Lemma 14.6.

Proof of Lemma 14.6. When $n = 0$, then the lemma is clear. Writing $f := f(\boldsymbol{x}), d := D_n(\boldsymbol{x}), e := E_n(\boldsymbol{x})$, and using $x_n^2 = 1$ and $\mathbf{E}[\boldsymbol{x}_n] = 0$, we have

$$\begin{aligned}\mathbf{E}[f^4] &= \mathbf{E}[(x_n d + e)^4] = \mathbf{E}[d^4] + 6\,\mathbf{E}[d^2 e^2] + \mathbf{E}[e^4] \\ \mathbf{E}[f^2] &= \mathbf{E}[(x_n d + e)^2] = \mathbf{E}[d^2] + \mathbf{E}[e^2].\end{aligned}$$

By Cauchy–Schwarz, $\mathbf{E}[d^2 e^2] \le \mathbf{E}[d^4]^{1/2}\,\mathbf{E}[e^4]^{1/2}$, and so

$$\begin{aligned}\mathbf{E}[f^4] &\le \mathbf{E}[d^4] + 6\,\mathbf{E}[d^4]^{1/2}\,\mathbf{E}[e^4]^{1/2} + \mathbf{E}[e^4] \\ &\le 9^{k-1} + 6 \cdot 9^{\frac{k-1}{2}}\,\mathbf{E}[d^2] \cdot 9^{\frac{k}{2}}\,\mathbf{E}[e^2] + 9^k\,\mathbf{E}[e^2]^2 \\ &\le 9^k \Big(\mathbf{E}[d^2]^2 + 2\,\mathbf{E}[d^2]\,\mathbf{E}[e^2] + \mathbf{E}[e^2]^2\Big) \\ &= 9^k \Big(\mathbf{E}[d^2] + \mathbf{E}[e^2]\Big)^2 \\ &= 9^k\,\mathbf{E}[f^2]^2.\end{aligned}$$

□

15.1 Hypercontractivity

Recall the Bonami lemma says that for Boolean functions $f\colon \{-1,1\}^n \to \mathbb{R}$ of degree at most k we have

$$\mathbf{E}[f(\boldsymbol{x})^4] \le 9^k \,\mathbf{E}[f(\boldsymbol{x})^2]^2.$$

Equivalently, we have $\|f\|_4 \le \sqrt{3}^k \|f\|_2$. We can generalize this inequality from 4-norm to any $(q \ge 2)$-norm.

Lemma 15.1. *$\|f\|_q \le \sqrt{q-1}^k \|f\|_2$ for every $f\colon \{-1,1\}^n \to \mathbb{R}$ of degree at most k.*

We omit the proof but remark that the proof for even q follows from a similar inductive argument to the $q = 2$ case. What about $q < 2$? Note that for $q = 1$, we can use the Cauchy–Schwarz to show that

$$\|f\|_1 = \mathbf{E}\big[|f|\big] = \mathbf{E}\big[1 \cdot |f|\big] \le \|1\|_2 \|f\|_2 = \|f\|_2.$$

In general we have $\|f\|_q \le \|f\|_p$ whenever $q \le p$.

Proposition 15.2 (Norm inequality)**.** *$\|f\|_q \le \|f\|_p$ for every $1 \le q \le p \le \infty$,*

To prove this, we will use the following fundamental inequality in analysis, which is a generalization of the Cauchy–Schwarz inequality.

Proposition 15.3 (Hölder's inequality)**.** *For every $1 \le r \le s \le \infty$ such that $\frac{1}{r} + \frac{1}{s} = 1$, we have*

$$\mathbf{E}\big[|f(x) \cdot g(x)|\big] \le \|f\|_r \|f\|_s.$$

Note that we have the Cauchy–Schwarz inequality by setting $r = s = 2$. We will see many applications of this inequality in this lecture. First let us see how to use it to prove the norm inequality.

Proof of Proposition 15.2. Let $r := p/q \ge 1$ and $s := 1/(1 - 1/r)$. Then by Hölder's inequality,

$$\|f\|_q = \mathbf{E}\big[1 \cdot |f(\boldsymbol{x})|^q\big]^{1/q} \le \mathbf{E}\big[1^s\big]^{1/s} \cdot \mathbf{E}\big[|f(\boldsymbol{x})^{qr}|\big]^{1/qr} = \|f\|_p. \qquad \square$$

As Lemma 15.1 gives us a bound on the higher moments of $f(\boldsymbol{x})$, we can prove a concentration inequality using the same approach in Proposition 14.3.

Theorem 15.4. *Let $f\colon \{0,1\}^n \to \mathbb{R}$ be a function of degree at most k. Then $\mathbf{Pr}[|f(\boldsymbol{x})| \ge t\|f\|_2] \le 4 \cdot e^{-t^{2/k}/2}$.*

We first make some remarks. First, note that without loss of generality we may assume $\mathbf{E}[f] = 0$ as otherwise we can consider $f' = f - \mathbf{E}[f]$ which does not change the degree. So indeed Theorem 15.4 is bounding the deviation of $f(\boldsymbol{x})$ from the mean. Also, when $\mathbf{E}[f] = 0$, $\|f\|_2 = \mathbf{E}[f^2]^{1/2}$ is the standard deviation of $f(\boldsymbol{x})$. In particular, if f is a linear function, then Theorem 15.4 is simply the Hoeffding's inequality. Therefore, we can think of this theorem as a generalization of tail bound for sums of independent random variables to low-degree polynomial of independent random variables.

Proof of Theorem 15.4. As before we take powers of both sides of $f(\boldsymbol{x}) \geq t\|f\|_2$ and then apply Markov's inequality and Bonami's lemma. For every $q \geq 2$ we have

$$\begin{aligned}
\mathbf{Pr}\big[|f(\boldsymbol{x})| \geq t\|f\|_2\big] &= \mathbf{Pr}\big[|f(\boldsymbol{x})|^q \geq t\|f\|_2^q\big] \\
&\leq \frac{\mathbf{E}\big[|f(\boldsymbol{x})|^q\big]}{t^q \cdot \|f\|_2^q} \\
&= \frac{1}{t^q} \cdot \frac{\|f\|_q^q}{\|f\|_2^q} \\
&\leq \frac{1}{t^q} \cdot (\sqrt{q-1})^{kq} \\
&= \left(\frac{(q-1)^{k/2}}{t}\right)^q.
\end{aligned}$$

We choose q to optimize this quantity. Note that if $t^{2/k}/2 < 1$, then the bound in the theorem is ≥ 1 which is trivial. So we can assume $t^{2/k}/2 \geq 1$. Taking $q = \frac{t^{2/k}}{2} \geq 1$, this is at most $e^{-\frac{t^{2/k}}{2}}$. □

The Bonami lemma bounds above the 4-norm of f in terms of its 2-norm. What if we want to bound its 2-norm in terms of its 1-norm? We can use the following trick.

Claim 15.5. *Let $f\colon \{-1,1\}^n \to \mathbb{R}$ be a polynomial of degree at most k. Then $\|f\|_2 \leq 3^k\|f\|_1$.*

The case $k = 1$ is called the Khintchine inequality.

Proof. The trick is the following:

$$\|f\|_2^2 = \mathbf{E}[f(\boldsymbol{x}) \cdot f(\boldsymbol{x})] = \mathbf{E}\big[|f(\boldsymbol{x})|^{4/3} \cdot |f(\boldsymbol{x})|^{2/3}\big].$$

Now applying Hölder's inequality with $r = 3$ and $s = 3/2$ and Bonami's lemma, we get

$$\|f\|_2^2 \leq \mathbf{E}\big[|f(\boldsymbol{x})|^4\big]^{1/3} \cdot \mathbf{E}\big[|f(\boldsymbol{x})|\big]^{2/3} \leq 9^k \cdot \mathbf{E}\big[f(\boldsymbol{x})^2\big]^{2/3} \cdot \mathbf{E}[|f(\boldsymbol{x})|]^{2/3} = 9^k \cdot \|f\|_2^{4/3} \cdot \|f\|_1^{2/3}.$$

The claim then follows from dividing both sides by $\|f\|_2^{4/3}$. □

15.1.1 Hypercontractivity

Recall in Definition 5.8, we defined $\boldsymbol{y} \sim N_\rho(x)$ as the random string obtained by independently setting each bit $\boldsymbol{y}_i$ to uniform with probability $1-\rho$ and $\mathrm{sgn}(\rho)x_i$ otherwise. We also define the noise operator T_ρ on f (Definition 5.14) as

$$T_\rho f(x) := \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N_\rho(x)}[f(\boldsymbol{y})] = \sum_{S \subseteq [n]} \widehat{f}(S) \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N_\rho(x)}\big[\chi_S(\boldsymbol{y})\big] = \sum_{S \subseteq [n]} \widehat{f}(S)\rho^{|S|}\chi_S(x).$$

The noise operator T_ρ is *contractive* as it does not increase the norm of a function.

Fact 15.6. *For every $\rho \in [-1,1]$, we have $\|T_\rho f\|_2 \leq \|f\|_2$ for every $f\colon \{-1,1\}^n \to \mathbb{R}$.*

The hypercontractivity theorem says that not only T_ρ is contractive, but it is *hypercontractive* in the following sense.

Theorem 15.7 ((2,4)-Hypercontractivity theorem)**.** *$\|T_{1/\sqrt{3}} f\|_4 \leq \|f\|_2$ for every $f\colon \{-1,1\}^n \to \mathbb{R}$.*

We will not prove this theorem, as this can be proved in the same inductive manner as proving the Bonami lemma. The intuition is that the dampening factor of $\rho^{|S|}$ introduced by the noise in T_ρ cancels the factor 9^k in the Bonami lemma. For example, if we let $f^{=k}(x) := \sum_{|S|=k} \widehat{f}(S)\chi_S(x)$ be the degree-k homogenous part of f, then

$$\begin{aligned}\mathbf{E}\left[\left(T_{\frac{1}{\sqrt{3}}}f(\boldsymbol{x})\right)^4\right] &= \mathbf{E}\left[\left(\sum_{|S|=k}\widehat{f}(S)\left(\frac{1}{\sqrt{3}}\right)^k\chi_S(\boldsymbol{x})\right)^4\right] \\ &= \left(\frac{1}{\sqrt{3}}\right)^{4k}\mathbf{E}\left[\left(\sum_{|S|=k}\widehat{f}(S)\chi_S(\boldsymbol{x})\right)^4\right] \\ &= 9^{-k}\cdot\mathbf{E}[f(\boldsymbol{x})^4].\end{aligned}$$

By Bonami's lemma, this is at most $\mathbf{E}[f(\boldsymbol{x})^2]^2$. (Note that we cannot simply sum over k to prove the theorem.)

We now state the general Hypercontractivity without proof.

Theorem 15.8 ((p,q)-Hypercontractivity theorem)**.** *For every $f\colon\{-1,1\}^n\to\mathbb{R}$, $1\le p\le q$ and $0\le\rho\le\sqrt{\frac{p-1}{q-1}}$, we have $\|T_\rho f\|_q\le\|f\|_p$.*

Setting to $p=2$, we have $\|T_{\frac{1}{\sqrt{p-1}}}f\|_q\le\|f\|_2$ for $q\ge 2$. Likewise, setting $q=2$, we have $\|T_{\sqrt{p-1}}f\|_2\le\|f\|_p$ for every $p\le 2$.

15.1.2 Small-set expansion of the noisy hypercube

Let $G=(V,E,w)$ be a weighted (undirected) graph, where for every $x\in V$, the weights $\{w(x,y)\}_{y\in V}$ on its adjacent edges form a distribution on V. We define the *expansion* of a subset $S\subseteq V$ in G as

$$\Phi_G(S) := \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim S\\ \boldsymbol{y}\sim_w\boldsymbol{x}}}[\boldsymbol{y}\notin S].$$ [7]

In words, we draw a uniformly random vertex in S and ask what is the probability of a random neighbor of x (sampled according the weight function w) leaving S. We also define the expansion of G as

$$\Phi(G) := \min_{S:|S|\le|V|/2}\Phi_G(S).$$

(Note that without the condition $|S|\le|V|/2$ we can take $S=V$ and this quantity is always 0.) We say that G is an *c-expander* if $\Phi(G)\ge c$. It is not hard to see that $\Phi(G)$ is at most a constant and cannot be arbitrarily close to 1. (To see this, take the union of a subset $S\subseteq V$ of size $|V|/4$ and some appropriately chosen size-$(|V|/4)$ subset of the neighbor of S.) We now look at a weaker notion of expansion in which we only require expansion over *small* subsets of V.

Definition 15.9 (Small-set expander)**.** G is an (δ,ϵ)-small-set expander if $\Phi_G(S)\ge 1-\epsilon$ for every subset S of size at most δn.

Let (V,E) be the Boolean hypercube $\{-1,1\}^n$. We will show that for $\rho\in[-1,1]$, the ρ-noisy hypercube $G=(V,E,w_\rho)$, where the weight $w_\rho(x,y)$ is defined by

$$w_\rho(x,y) := \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim N_\rho(x)}[\boldsymbol{y}=y],$$

[7] This is usually called the *conductance* of the cut $(S,\overline{S})$ in G

is a small-set expander. For intuition, we can think of $\boldsymbol{y} \sim N_\rho(x)$ as taking a random walk of length roughly $\frac{(1-\rho)n}{2}$ starting from x in $\{-1,1\}^n$, as $\frac{(1-\rho)n}{2}$ is the expected number of bits flipped in x.

Claim 15.10. *For $\rho = 1/\sqrt{3}$, the ρ-noisy hypercube is a $(\delta, \delta^{1/4})$-small-set expander.*

Proof. Fix a subset $S \subseteq \{-1,1\}^n$ of size $\delta 2^n$. Let $f\colon \{-1,1\}^n \to \{0,1\}$ be the indicator $f(x) := \mathbb{1}(x \in S)$. We have $\mathbf{E}[f(\boldsymbol{x})] = \delta$. We will relate the escape probability to the stability of f. Recall from Definition 5.10 that the stability of f is defined as

$$\mathsf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n, \boldsymbol{y}\sim N_\rho(x)}[f(\boldsymbol{x})f(\boldsymbol{y})] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\Big[f(\boldsymbol{x})\cdot \mathop{\mathbf{E}}_{\boldsymbol{y}\sim N_\rho(x)}[f(\boldsymbol{y})]\Big] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\big[f(\boldsymbol{x})\cdot T_\rho f(\boldsymbol{x})\big].$$

We will bound above the inner product on the right hand side using Hypercontractivity. Observe that $\mathbf{E}[f(\boldsymbol{x})^q] = \mathbf{E}[f(\boldsymbol{x})]$ for any $q > 0$ and thus $\|f\|_q = \mathbf{E}[f(\boldsymbol{x})]^{\frac{1}{q}}$. By Hölder's inequality and Hypercontractivity,

$$\mathbf{E}[f(\boldsymbol{x})\cdot T_\rho(\boldsymbol{x})] \le \|f\|_{4/3}\cdot\|T_\rho f\|_4 \le \mathbf{E}[f(\boldsymbol{x})]^{3/4}\cdot\mathbf{E}[f(\boldsymbol{x})]^{1/2} = \mathbf{E}[f(\boldsymbol{x})]^{5/4}.$$

Now,

$$\mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim S\\ \boldsymbol{y}\sim N_\rho(\boldsymbol{x})}}[\boldsymbol{y}\in S] = \frac{1}{\mathbf{E}[f(\boldsymbol{x})]}\mathsf{Stab}_\rho[f] \le \frac{1}{\mathbf{E}[f(\boldsymbol{x})]}\cdot\mathbf{E}[f(\boldsymbol{x})]^{5/4} = \mathbf{E}[f(\boldsymbol{x})]^{1/4}.$$

So $\Phi_G(S) \ge 1 - \mathbf{E}[f(\boldsymbol{x})]^{1/4} = 1-\delta^{1/4}$. □

16.1 The Fourier spectrum of small sets

Let $S \subseteq \{-1,1\}^n$ be a subset of size $\delta 2^n$. We will study the Fourier spectrum of the indicator function $f(x) := \mathbb{1}(x \in S)$.

As a warm-up let us give a lower bound on its degree. Recall that in Claim 6.5 we showed that for every $g\colon \{-1,1\}^n \to \mathbb{R}$ of degree at most d, we must have $\mathbf{Pr}[g(\boldsymbol{x}) \neq 0] \geq 2^{-d}$. Applying this claim to f shows that f must have degree at least $\Omega(\log(1/\delta))$.

We now strengthen this observation by showing that most of the Fourier weight of f is on subsets of size at least some $\Omega(\log(1/\delta))$.

Claim 16.1. *Let $f\colon \{-1,1\}^n \to \{-1,0,1\}$ with $\mathbf{Pr}[f(\boldsymbol{x}) \neq 0] \leq \delta$. Then $W^{\leq d}[f] \leq 3^d \cdot \delta^{3/2}$.*

Letting $d = \log(1/\delta)/4$ we can conclude that

$$W^{\leq \log(1/\delta)/4}[f] \leq 3^{\log(1/\delta)/4} \cdot \delta^{3/2} = \delta^{(\log 3)/4} \cdot \delta^{3/2} \ll \delta = \mathbf{E}[f(\boldsymbol{x})^2] = \sum_{S \subseteq [n]} \widehat{f}(S)^2.$$

Therefore, most of the Fourier weight is on subsets S of size at least $\Omega(\log(1/\delta))$.

Proof of Claim 16.1. Let $f^{\leq d}(x) := \sum_{|S| \leq d} \widehat{f}(S)\chi_S(x)$ be the "low-degree part" of f. The key observation is that by Plancherel's identity we have

$$W^{\leq d}[f] = \mathbf{E}\big[f^{\leq d}(\boldsymbol{x}) \cdot f^{\leq d}(\boldsymbol{x})\big] = \mathbf{E}\big[f(\boldsymbol{x}) \cdot f^{\leq d}(\boldsymbol{x})\big].$$

Now we apply Hölder inequality with $r = 4/3$ and $s = 4$, and then the Bonami lemma (to $f^{\leq d}$) to conclude that

$$\mathbf{E}\big[f(\boldsymbol{x}) \cdot f^{\leq d}(\boldsymbol{x})\big] \leq \|f\|_{4/3} \cdot \|f^{\leq d}\|_4 \leq \delta^{3/4} \cdot \sqrt{3}^d \|f^{\leq d}\|_2 = \sqrt{3}^d \cdot \delta^{3/4} \cdot W^{\leq d}[f]^{1/2}.$$

So $W^{\leq d}[f] \leq 3^d \cdot \delta^{3/2}$. □

We now show that the exponent $3/2$ of δ can be improved when d is small.

Lemma 16.2 (Level-k inequality)**.** *Let $f\colon \{-1,1\}^n \to \{-1,0,1\}$ with $\mathbf{Pr}[f(\boldsymbol{x}) \neq 0] = \mathbf{E}[|f(\boldsymbol{x})|] \leq \delta$. Then $W^{\leq k}[f] \leq \delta^2 \cdot (100 \log(2/\delta))^k$.*

Note that the bounds in Claim 16.1 and Lemma 16.2 are incomparable. If we set $k = c\log(2/\delta)$ for some $c > 0$, then using Lemma 16.2, we only get an upper bound of $W^{\leq c\log(2/\delta)}[f] \leq \delta^2 \left(c\log(2/\delta)\right)^{c\log(2/\delta)} = \delta^2(1/\delta)^{\Theta(\log\log(1/\delta))} \not\ll \delta$, and thus we cannot even conclude that f has degree at least $\Theta(\log(1/\delta))$.

Indeed, Lemma 16.2 is not tight and can be improved to

$$W^{\leq k}[f] \leq \delta^2 \cdot \left(100 \log(2/\delta^{1/k})\right)^k.$$

By considering the AND function one can verify that this is close to tight.

We will only prove Lemma 16.2. One can prove this using a similar argument to the proof of Claim 16.1 using the q-norm version of Bonami's lemma (Lemma 15.1), but we will give a different proof below to demonstrate a use of tail bound for low-degree polynomials (Theorem 15.4).

Proof of Lemma 16.2. As in the previous proof, let $f^{\le k}(x) := \sum_{|S|\le k} \widehat{f}(S)\chi_S(x)$ be the low-degree part of f and write $W^{\le k}[f] = \mathbf{E}[f(\boldsymbol{x}) \cdot f^{\le k}(\boldsymbol{x})]$. Note that $f^{\le k}(x)$ is not bounded, and so we will decompose this expectation into two parts depending on the magnitude of $f^{\le k}(x)$, and use the tail bound for low-degree polynomials to argue that $|f^{\le k}(\boldsymbol{x})|$ cannot be too large for most $\boldsymbol{x} \in \{-1,1\}^n$.

To proceed, let E be the event "$f^{\le k}(\boldsymbol{x}) \le T\|f^{\le k}\|_2$" for some T that will be chosen later. We have

$$\mathbf{E}\big[f(\boldsymbol{x}) \cdot f^{\le k}(\boldsymbol{x})\big] = \mathbf{E}\Big[|f(\boldsymbol{x})| \cdot |f^{\le k}(\boldsymbol{x})| \cdot \mathbb{1}(E)\Big] + \mathbf{E}\Big[|f(\boldsymbol{x})| \cdot |f^{\le k}(\boldsymbol{x})| \cdot \mathbb{1}(\overline{E})\Big].$$

The first term can be easily bounded by $T\|f^{\le k}\|_2 \cdot \mathbf{E}[|f(x)|] \le \delta \cdot T \cdot \|f^{\le k}\|_2$. To bound the second term, we apply Cauchy–Schwarz and then Theorem 15.4 to obtain

$$\begin{aligned}
\mathbf{E}\Big[|f(\boldsymbol{x})| \cdot |f^{\le k}(\boldsymbol{x})| \cdot \mathbb{1}(\overline{E})\Big] &\le \mathbf{E}\Big[|f(\boldsymbol{x})|^2 \cdot |f^{\le k}(\boldsymbol{x})|^2\Big]^{1/2} \cdot \mathbf{Pr}[\overline{E}]^{1/2}\Big] \\
&= \mathbf{E}\Big[|f^{\le k}(\boldsymbol{x})|^2\Big]^{1/2} \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}\Big[f^{\le k}(\boldsymbol{x}) > T\|f^{\le k}\|_2\Big]^{1/2} \\
&\le \|f^{\le k}\|_2 \cdot 4 \cdot e^{-\frac{T^{2/k}}{2}} \qquad \text{(Theorem 15.4)}.
\end{aligned}$$

Hence, we have

$$W^{\le k}[f] \le \|f^{\le k}\|_2 \cdot \left(\delta \cdot T + 4 \cdot e^{-\frac{T^{2/k}}{2}}\right) = W^{\le k}[f]^{1/2} \cdot \left(\delta \cdot T + 4 \cdot e^{-\frac{T^{2/k}}{2}}\right).$$

Choosing $T = (50\log(1/\delta))^{k/2}$ completes the proof. □

16.2 FKN theorem

We now prove the Friedgut–Kalai–Naor (FKN) theorem, which is a robust version of the following claim.

Claim 16.3. *Suppose $f\colon \{-1,1\}^n \to \{-1,1\}$ satisfies $W^{=1}[f] = \sum_{|S|=1} \widehat{f}(S)^2 = 1$. Then f must be a dictator (i.e. $\chi_{\{i\}}$ for some $i \in [n]$) or an anti-dictator (i.e. $-\chi_{\{i\}}$ for some $i \in [n]$).*

This can be proved by observing that f has degree 1 and therefore from Homework 1 we know that its coefficients must be a multiple of 1. Here we give a more direct proof.

Proof. First of all note that by Parseval, $\{\widehat{f}(\{i\}) : i \in [n]\}$ are the only nonzero coefficients. Let $a_i := \widehat{f}(\{i\})$ and we have $f(x) = \sum_i a_i x_i$. By choosing $x_i = \operatorname{sgn}(a_i) \in \{-1,1\}$ we have $\sum_{i=1}^n |a_i| = f(x) = 1$. Now by Parseval we have

$$1 = \sum_{i=1}^n a_i^2 \le \max_i |a_i| \sum_i |a_i| = \max_i |a_i|.$$

So we must have $|a_i| = 1$ for some $i \in [n]$ and the rest of the $a_j : j \ne i$ must be 0. □

If instead of $W^{=1}[f] = 1$ we have $W^{\le 1}[f] = 1$, we claim that f is a dictator, an anti-dictator, or a constant, by reducing it to the former case using the following trick.

Consider $f'\colon \{-1,1\}^{n+1} \to \{-1,1\}$ defined by $f'(x_0, x_1, \ldots, x_n) = \widehat{f}(\varnothing)x_0 + f(x_1, \ldots, x_n)$. Observe that $f'(1, x_1, \ldots, x_n) = f(x_1, \ldots, x_n)$ and $f'(-1, -x_1, \ldots, -x_n) = -f(x_1, \ldots, x_n)$ and thus the range of f' is indeed $\{-1,1\}$. Now observe that $W^{=1}[f'] = W^{\le 1}[f]$.

Theorem 16.4. *Suppose $f\colon \{-1,1\}^n \to \{-1,1\}$ satisfies $W^{\le 1}[f] \ge 1-\delta$. Then f is $O(\delta)$-close to a dictator, an anti-dictator, or a constant.*

By setting $\delta = 0$, we recover Claim 16.3.

Proof. Without loss of generality, we can assume $W^{=1}[f] = 1-\delta$ using the trick above. Let $\ell(x) := f^{=1}(x) = \sum_{i=1}^n \widehat{f}(\{i\})x_i$ be the degree-1 part of f. Most of our effort will go into showing that $\mathbf{Var}[\ell(\boldsymbol{x})^2] = O(\delta)$. Let us first see how it implies the theorem.

We first look at $\mathbf{Var}[\ell^2]$ in terms of the Fourier coefficients of ℓ. Let us first look at the Fourier expansion of $\ell(x)^2$. We have

$$\ell(x)^2 = \sum_{i,j} \widehat{f}\{i\}\widehat{f}\{j\}x_ix_j = \sum_{i\neq j} \widehat{f}\{i\}\widehat{f}\{j\}x_ix_j + \sum_{i=1}^n \widehat{f}\{i\}^2.$$

As $\mathbf{Var}[g] = \sum_{S\neq\varnothing} \widehat{g}(S)^2$ for any $g\colon \{-1,1\}^n \to \mathbb{R}$,

$$\begin{aligned}
\mathbf{Var}[\ell^2] &= \sum_{i\neq j} \widehat{f}\{i\}^2\widehat{f}\{j\}^2 \\
&= \Big(\sum_{i=1}^n \widehat{f}\{i\}^2\Big)^2 - \sum_{i=1}^n \widehat{f}\{i\}^4 \\
&= \mathbf{E}\big[\ell(\boldsymbol{x})^2\big]^2 - \sum_i \widehat{f}\{i\}^4 \\
&= (1-\delta)^2 - \sum_i \widehat{f}\{i\}^4.
\end{aligned}$$

Hence

$$\max_{i\in[n]} \widehat{f}\{i\}^2 \sum_{i\in[n]} \widehat{f}\{i\}^2 \ge \sum_i \widehat{f}\{i\}^4 \ge (1-\delta)^2 - \mathbf{Var}[\ell^2] \ge (1-\delta)^2 - O(\delta) \ge 1 - O(\delta),$$

and so

$$\max_i |\widehat{f}\{i\}| \ge \left(\frac{1-O(\delta)}{1-\delta}\right)^{1/2} \ge 1 - O(\delta). \qquad \square$$

It remains to show that $\mathbf{Var}[\ell^2] \le O(\delta)$.

16.2.1 Bounding the variance of ℓ^2

First observe that we have

$$\mathbf{E}[\ell(\boldsymbol{x})^2] = \mathbf{E}[f(\boldsymbol{x})\ell(\boldsymbol{x})] = 1-\delta \quad \text{and} \qquad \mathbf{E}[(f(x)-\ell(x))^2] = \delta, \tag{8}$$

Using the definition of variance, we have

$$\mathbf{Var}[\ell^2] = \mathbf{E}\Big[\big|\ell^2 - \mathbf{E}[\ell^2]\big|^2\Big] = \mathbf{E}\Big[\big|\ell^2 - (1-\delta)\big|^2\Big].$$

The trick is a clever use of the fact that $f(x)^2 = 1$. By Khintchine's inequality (Claim 15.5),

$$\begin{aligned}
\mathbf{E}\Big[\big|\ell^2 - (1-\delta)\big|^2\Big] &\le 9\cdot \mathbf{E}\Big[\big|\ell^2 - (1-\delta)f^2\big|\Big]^2 \\
&= 9\,\mathbf{E}\big[(\ell - \sqrt{1-\delta}f)\cdot(\ell+\sqrt{1-\delta}f)\big] \\
&\le 9\,\mathbf{E}\big[(\ell-\sqrt{1-\delta}f)^2\big]^{1/2}\cdot \mathbf{E}\big[(\ell+\sqrt{1-\delta}f)^2\big]^{1/2}.
\end{aligned}$$

It follows from Equation (8) that

$$\mathbf{E}\big[(\ell - \sqrt{1-\delta}f)^2\big]^{1/2} \le O(\delta) \quad \text{and} \qquad \mathbf{E}\big[(\ell + \sqrt{1-\delta}f)^2\big]^{1/2} \le 4.$$

So $\mathbf{Var}[\ell^2] \le \mathbf{E}[|\ell^2 - (1-\delta)|^2] \le O(\delta)$.

17.1 KKL theorem

We will prove the Kahn–Kalai–Linial (KKL) theorem.

Theorem 17.1. *Every function $f\colon \{-1,1\}^n \to \{-1,1\}$ must contain an influential coordinate $i \in [n]$ such that $\mathsf{Inf}_i[f] \geq \mathbf{Var}[f] \cdot \Omega((\log n)/n)$.*

Before we do that, it would be useful to recall the key ingredients that go into proving Theorem 6.4, that functions of degree k are $(k \cdot 2^{k-1})$-juntas. The first ingredient is that $\mathbf{I}[f] \leq \deg(f) = k$ because $\mathbf{I}[f] = \mathbf{E}_{\boldsymbol{S} \sim \mathcal{S}_f}[|\boldsymbol{S}|]$ is the "average degree" of f. The second ingredient is to show that for each coordinate $i \in [n]$, it must be the case that $\mathsf{Inf}_i[f] = 0$ or $\mathsf{Inf}_i[f] \geq 2^{1-k}$. This uses the fact that the derivative $D_i f$ has degree $k-1$ and $\mathbf{Pr}[g(\boldsymbol{x}) \neq 0] \geq 2^{1-k}$ for nonzero $g\colon \{-1,1\}^n \to \mathbb{R}$ of degree at most k.

So far we have established the robust analogue of these two ingredients by arguing about the concentration of the Fourier coefficients of f. We have seen in Proposition 5.6 that every function is ϵ-concentrated on degree at most $\mathbf{I}[f]/\epsilon$, i.e., $W^{> \frac{\mathbf{I}[f]}{\epsilon}}[f] \leq \epsilon$. In the last lecture, we showed in Claim 16.1 that "small sets" have their Fourier weight concentrated on the high-degree part, i.e. $W^{\leq \frac{1}{4}\log(1/\mathbf{E}[|g|])} \ll \mathbf{E}[|g|]$ for functions $g\colon \{-1,1\}^n \to \{-1,0,1\}$. We can argue about the second ingredient above using this language: Letting $g(x) := D_i f(x) \in \{-1,0,1\}$, we see that if $\mathsf{Inf}_i[f] = \mathbf{E}[|D_i f(\boldsymbol{x})|]$ is small, then $D_i f$ must have most of its Fourier weight on high degree, which contradicts the fact that $\deg(f)$ is small.

We will prove Theorem 17.1 using a similar idea. First of all, note that the bound in the theorem is only interesting when $\mathbf{I}[f] \leq c \cdot \mathbf{Var}[f] \cdot \log n$ for any sufficiently small constant c; for otherwise the theorem follows by simply averaging over the n coordinates. In other words, most of the Fourier weight of f is concentrated on degree at most $O(\log n)$. Now assume towards a contradiction that every individual influence is $\mathsf{Inf}_i[f] = \mathbf{E}[|D_i f(\boldsymbol{x})|] \leq O(\log n)/n$. This implies that there is very little Fourier mass on degree $\leq \frac{1}{4}\log(1/\mathbf{E}[|D_i f(\boldsymbol{x})|]) \leq \Theta(\log n)$ in the Fourier spectrum of $D_i f$. But both $\sum_i D_i f$ and f have similar Fourier spectrum in the low-degree part, and so we get a contradiction.

We now give the formal proof.

Proof of Theorem 17.1. Without loss of generality, assume $\mathbf{I}[f] \leq c \cdot \mathbf{Var}[f] \cdot \log n$ for a sufficiently small constant $c > 0$. By Proposition 5.6 we know that f is $(\mathbf{Var}[f]/10)$-concentrated on degree $10c \log n$, that is, $W^{>10c\log n}[f] \leq \mathbf{Var}[f]/10$. Assume towards a contradiction that $\mathsf{Inf}_i[f] = \mathbf{E}[|D_i f(\boldsymbol{x})|] \leq c \cdot \mathbf{Var}[f] \cdot (\log n)/n \leq 1/n^{40c}$ for every $i \in [n]$. Since $\log(\frac{1}{\mathbf{E}[|D_i f(\boldsymbol{x})|]}) \geq 40c \log n$, by Claim 16.1 we have

$$W^{\leq 10c\log n}[D_i f] \leq \left(\mathbf{Var}[f] \cdot \frac{\log n}{n}\right)^{1.01},$$

and so

$$\begin{aligned}\sum_{0<|S|\le 10c\log n}\widehat{f}(S)^2 &\le \sum_{|S|\le 10c\log n}|S|\cdot\widehat{f}(S)^2\\ &= \sum_{i=1}^{n}\sum_{\substack{S\ni i\\ |S|\le 10c\log n}}\widehat{f}(S)^2 \qquad (9)\\ &= \sum_{i=1}^{n} W^{\le 10c\log n}[D_i f]\\ &\le n\cdot\left(\mathbf{Var}[f]\cdot\frac{\log n}{n}\right)^{1.01} \le \mathbf{Var}[f]/10.\end{aligned}$$

Putting the two bounds together, we have

$$\mathbf{Var}[f] = W^{>0}[f] \le \sum_{0<|S|\le 10c\log n}\widehat{f}(S)^2 + W^{>10c\log n}[f] \le \mathbf{Var}[f]/10 + \mathbf{Var}[f]/10 \le \mathbf{Var}[f]/2,$$

a contradiction. □

Recall that the $\mathsf{Tribes}_{w,2^w\ln 2}$ function is almost-balanced and satisfies $\mathsf{Inf}_i[f] \ge \Theta((\log n)/n)$ every $i \in [n]$. The KKL theorem can be slightly improved as follows.

Theorem 17.2 (Talagrand)**.** *There is a universal constant C such that every function $f\colon \{-1,1\}^n \to \{-1,1\}$ satisfies*

$$C\sum_{i=1}^{n}\frac{\mathsf{Inf}_i[f]}{\log(1/\mathsf{Inf}_i[f])} \ge \mathbf{Var}[f].$$

To see that Theorem 17.2 implies Theorem 17.1, note that

$$C\frac{1}{\log(1/\max_i \mathsf{Inf}_i[f])}\cdot\mathbf{I}[f] = C\sum_{i=1}^{n}\frac{\mathsf{Inf}_i[f]}{\log(1/\max_i \mathsf{Inf}_i[f])} \ge C\sum_{i=1}^{n}\frac{\mathsf{Inf}_i[f]}{\log(1/\mathsf{Inf}_i[f])} \ge \mathbf{Var}[f].$$

So after rearranging we have $\max_i \mathsf{Inf}_i[f] \ge e^{-C\frac{\mathbf{I}[f]}{\mathbf{Var}[f]}}$. Now, if $\mathbf{I}[f] \ge (1/2C)\,\mathbf{Var}[f]\cdot\log n$, then $\max_i \mathsf{Inf}_i[f] \ge \frac{1}{\sqrt{n}} \ge \mathbf{Var}[f]\cdot(\log n)/n$. Otherwise, the KKL theorem follows from averaging.

Proof of Theorem 17.2. The improvement comes from the following equality.

$$\mathbf{Var}[f] = W^{>0}[f] = \sum_{i=1}^{n}\sum_{S\ni i}\frac{1}{|S|}\cdot\widehat{f}(S)^2.$$

Now, let $g_i(x) := \sum_{S\ni i}\frac{1}{\sqrt{|S|}}\widehat{f}(S)\chi_S(x)$, and d_i be some threshold on the degree that will be determined later in the proof. We have $\|g_i\|_2^2 = W^{\le d_i}[g_i] + W^{>d_i}[g_i]$. We can bound the first term by

$$W^{\le d_i}[g_i] = \sum_{\substack{S\ni i\\ |S|\le d_i}}\frac{1}{|S|}\widehat{f}(S)^2 \le \sum_{\substack{S\ni i\\ |S|\le d_i}}\widehat{f}(S)^2 = W^{\le d_i}[D_i f],$$

and the second term by

$$W^{>d_i}[g_i] \leq \sum_{\substack{S \ni i \\ |S|>d_i}} \frac{1}{|S|}\widehat{f}(S)^2 \leq \frac{1}{d_i}\sum_{S\ni i}\widehat{f}(S)^2 \leq \frac{1}{d_i}\mathsf{Inf}_i[f]$$

Choosing $d_i = \log(1/\mathsf{Inf}_i[f])/4$, we have

$$\begin{aligned}
\mathbf{Var}[f] &\leq \sum_{i=1}^{n} \|g_i\|_2^2 \\
&\leq \sum_{i=1}^{n}\left(W^{\leq d_i}[D_i f] + \frac{1}{d_i}\mathsf{Inf}_i[f]\right) \\
&\leq \sum_{i=1}^{n}\left(\mathbf{E}[|D_i f(\boldsymbol{x})|]^{1.01} + \frac{\mathsf{Inf}_i[f]}{\log(4/\mathsf{Inf}_i[f])}\right) \\
&= \sum_{i=1}^{n}\left(\mathsf{Inf}_i[f]^{1.01} + \frac{\mathsf{Inf}_i[f]}{\log(4/\mathsf{Inf}_i[f])}\right) \\
&= C\sum_{i=1}^{n}\frac{\mathsf{Inf}_i[f]}{\log(1/\mathsf{Inf}_i[f])}.
\end{aligned}$$

□

18.1 Friedgut Junta Theorem

Our last application of Hypercontractivity is the Friedgut Junta Theorem.

Theorem 18.1. *Every Boolean function $f\colon \{-1,1\}^n \to \{-1,1\}$ is ϵ-close to a $2^{O(\mathbf{I}[f])/\epsilon)}$-junta.*

Recall that the Nisan–Szegedy theorem (Theorem 6.4) says that a Boolean function f of degree k is a $(k \cdot 2^{k-1})$-junta that depends on the coordinates in $J = \{i \in [n] : \mathsf{Inf}_i[f] \geq 2^{1-k}\}$. Thus we can write f as

$$f(x) := \sum_{\substack{S \subseteq J \\ |S| \leq k}} \widehat{f}(S)\chi_S(x).$$

We are going to construct the junta using a similar strategy. Note that f is ϵ-concentrated on degree at most $\mathbf{I}[f]/\epsilon$, so in this case the set J would collect the coordinates whose individual influence is at least $2^{-\mathbf{I}[f]/10\epsilon}$, and our junta is defined by removing all the coefficients from f except the low-degree coefficients $\widehat{f}(S)$ with respect to subsets S that completely lie in J. To show that the junta is close to f, we look at the Fourier weight contributed by the removed coefficients. Note that the coefficients that are removed are the ones which either have high degree, in which case they contribute ϵ weight, or contain a non-influential coordinate $i \notin J$; in this case we can bound their weight using the fact that $\mathsf{Inf}_i[f] \leq 2^{-10\mathbf{I}[f]/\epsilon}$ and Claim 16.1.

Proof of Theorem 18.1. Let $J := \{i \in [n] : \mathsf{Inf}_i[f] \geq 2^{-10\mathbf{I}[f]/\epsilon}\}$. Define $G\colon \{-1,1\}^n \to \mathbb{R}$ as

$$G(x) := \sum_{\substack{S \subseteq J \\ |S| \leq I[f]/\epsilon}} \widehat{f}(S)\chi_S(x).$$

Note that G may not be a Boolean function, but as taking sign does not increase the number of coordinates it depends on, we can apply the same trick as in Linial–Mansour–Nisan, by defining $g\colon \{-1,1\}^n \to \{-1,1\}$ by $g(x) := \mathrm{sgn}(G(x))$.

Now, as $\{2,0\} \ni |f(x) - g(x)| \leq 2|G(x) - f(x)|$, we have $\|f - g\|_2^2 \leq 4\|f - G\|_2^2$. Moreover,

$$\begin{aligned}\mathbf{E}[(G(\boldsymbol{x}) - f(\boldsymbol{x}))^2] &= \sum_{S \subseteq [n]} (\widehat{G}(S) - \widehat{f}(S))^2 \\ &= \sum_{|S| > \mathbf{I}[f]/\epsilon} \widehat{f}(S)^2 + \sum_{|S| > \mathbf{I}[f]/\epsilon} (\widehat{G}(S) - \widehat{f}(S))^2 \quad \leq \epsilon + \sum_{|S| > \mathbf{I}[f]/\epsilon} (\widehat{G}(S) - \widehat{f}(S))^2.\end{aligned}$$

For the second term we have

$$\begin{aligned}\sum_{|S| \leq \mathbf{I}[f]/\epsilon} (\widehat{G}(S) - \widehat{f}(S))^2 \leq \sum_{i \notin J} \sum_{\substack{S \ni i \\ |S| \leq \mathbf{I}[f]/\epsilon}} \widehat{f}(S)^2 &\leq \sum_{i \notin J} W^{\leq \mathbf{I}[f]/\epsilon}[D_i f] \leq \sum_{i \notin J} \mathsf{Inf}_i[f]^{1.01} \\ &\leq 2^{-\Omega(\mathbf{I}[f]/\epsilon)} \cdot \sum_{i \notin J} \mathsf{Inf}_i[f] \leq \mathbf{I}[f] \cdot 2^{-\Omega(\mathbf{I}[f])/\epsilon)} \leq \epsilon.\end{aligned}$$

Putting the two bounds together we have $\|f - g\|_2^2 \leq 8\epsilon$. □

18.2 Pseudorandom generators

The final topic of this course is about pseudorandomness. We will show how to use Fourier analysis to analyze *pseudorandom generators* for space-bounded computation.

Definition 18.2. A distribution D on $\{0,1\}^n$ *fools* a family of functions $\mathcal{F} \subseteq \{f\colon \{0,1\}^n \to \mathbb{R}\}$ with error ϵ if for every $f \in \mathcal{F}$, we have

$$\left| \mathop{\mathbf{E}}_{x \sim D}[f(\boldsymbol{x})] - \mathop{\mathbf{E}}_{x \sim \{0,1\}^n}[f(\boldsymbol{x}) = 1] \right| \leq \epsilon.$$

A *pseudorandom generator (PRG)* is a sparse distribution that fools some family $\mathcal{F}$.

Definition 18.3. A function $G\colon \{0,1\}^s \to \{0,1\}^n$ is a pseudorandom generator (PRG) for $\mathcal{F}$ with error ϵ if $G(\boldsymbol{x}) : \boldsymbol{x} \sim \{0,1\}^s$ fools $\mathcal{F}$ with error ϵ. We call s the seed length of G.

It is easy to construct a PRG with seed length n or error 1. In fact, there always exists a PRG with seed length $s = O(\log\log(|\mathcal{F}|) + \log(1/\epsilon))$.

Claim 18.4. *For every $\mathcal{F} \subseteq \{f\colon \{0,1\}^n \to \{0,1\}\}$, there exists a PRG for $\mathcal{F}$ with error ϵ and seed length $s = \log\log(|\mathcal{F}|) + 2\log(1/\epsilon) + O(1)$.*

Proof. Given a family $\mathcal{F}$, let s be a parameter to be chosen later. We use the probabilistic method to show that a random function $\boldsymbol{G}\colon \{0,1\}^s \to \{0,1\}^n$ is a good PRG. For a fixed $f \in \mathcal{F}$ we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{0,1\}^n}[f(\boldsymbol{x})] = 2^{-s} \sum_{u \in \{0,1\}^s} \mathop{\mathbf{E}}_{\boldsymbol{G}}\big[f(\boldsymbol{G}(u))\big] = \mathop{\mathbf{E}}_{\boldsymbol{G}}\Big[2^{-s} \sum_{u \in \{0,1\}^s} \big[f(\boldsymbol{G}(u))\big]\Big].$$

As $\boldsymbol{G}(u) : u \in \{0,1\}^n$ are independent, by the Chernoff bound, we have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{G}}\left[\left|2^{-s} \sum_{u \in \{0,1\}^s} f(\boldsymbol{G}(u)) - \mathop{\mathbf{E}}_{\boldsymbol{x} \in \{0,1\}^n}[f(\boldsymbol{x})]\right| > \epsilon\right] \leq 2^{-\Omega(2^s \epsilon^2)}.$$

By a union bound over all $f \in \mathcal{F}$, the probability that $\boldsymbol{G}$ fails to fool $f \in \mathcal{F}$ is at most $|\mathcal{F}| \cdot 2^{-\Omega(2^s\epsilon^2)}$. Choosing $s = \log\log(|\mathcal{F}|) + 2\log(1/\epsilon) + O(1)$, this is strictly less than 1, and so some $G\colon \{0,1\}^s \to \{0,1\}^n$ must fool $\mathcal{F}$. □

Given that G always exists, the next question to ask is whether we can construct one *explicitly*, that is, whether we can find G efficiently, say in time polynomial in n. From now on when we talk about PRGs we assume the need to be explicit.

We now introduce one of the most fundamental tools in derandomization and algorithm design.

Definition 18.5 (k-wise independent distributions)**.** A distribution D on $\{0,1\}^n$ is k-wise independent if D is uniform on any k of the n coordinates. Equivalently, D fools all k-juntas with 0 error, that is, $\mathbf{E}_{\boldsymbol{x} \sim D}[f(\boldsymbol{x})] = \mathbf{Pr}_{\boldsymbol{x} \sim \{0,1\}^n}[f(\boldsymbol{x})]$ for every $f\colon \{0,1\}^n \to \{0,1\}$ that depends only at most k coordinates.

k-wise independent distributions are also known as universal hash functions. Using Fourier analysis, it is straightforward to see that an equivalent definition for a distribution to be k-wise independent is that it fools all parity functions of size k. We now give a explicit construction of k-wise independent distribution.

Claim 18.6. *There exists a k-wise independent distribution on $\{0,1\}^n$ that can be efficiently sampled using a seed of length $O(k \log n)$.*

Proof. We will instead show how to use k uniformly random elements in the field $\mathbb{F}_{2^{\log n}}$ to generate a k-wise independent distribution D on $\mathbb{F}_{2^{\log n}}^n$, i.e. D is uniform on every k coordinates of $\mathbb{F}_{2^{\log n}}^n$, then taking the first bit of the element in the n coordinates gives us a k-wise independent distribution on $\mathbb{F}_2^n$. In more detail, the k random elements $\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{k-1} \sim \mathbb{F}_{2^{\log n}}$ are used to specify a random degree-$(k-1)$ polynomial $p_{\boldsymbol{a}_0,\ldots,\boldsymbol{a}_{k-1}} \colon \mathbb{F}_{2^{\log n}}^n \to \mathbb{F}_{2^{\log n}}$ defined by

$$p_{\boldsymbol{a}_0,\ldots,\boldsymbol{a}_{k-1}}(x) := \sum_{i=0}^{k-1} \boldsymbol{a}_i x^i,$$

and the n coordinates of the distribution D are obtained by evaluating $p_{\boldsymbol{a}_0,\ldots,\boldsymbol{a}_{k-1}}$ on the n elements in $\mathbb{F}_{2^{\log n}}$. To see that D is k-wise independent, we use the fact that a degree-$(k-1)$ polynomial is uniquely determined by its evaluation on any k distinct points. One way to see this is to use a similar argument that we saw in HW3 Q2. Given k coordinates $(y_{i_1}, \ldots y_{i_k})$ of D, to determine $\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{k-1}$ we can form the linear system

$$\begin{bmatrix} 1 & x_{i_1} & \cdots & x_{i_1}^{k-1} \\ 1 & x_{i_2} & \cdots & x_{i_2}^{k-1} \\ \vdots & & \cdots & \\ 1 & x_{i_k} & \cdots & x_{i_k}^{k-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{bmatrix} = \begin{bmatrix} y_{i_1} \\ y_{i_2} \\ \vdots \\ y_{i_k} \end{bmatrix}.$$

As the matrix is a $k \times k$ Vandermonde matrix, it has full rank and the system has a unique solution. So the probability that $(\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{k-1})$ is equal to this solution is exactly n^{-k}. □

19.1 Bounded independence plus noise

In this lecture we study the power of adding noise to k-wise independence. We will show that it fools a family of tests called *product tests*, and we will briefly mention its connection to derandomizing space-bounded computation.

Definition 19.1 (Product tests)**.** A function $f\colon \{0,1\}^{n=m\cdot d} \to [-1,1]$ is a (m,d)-product if it can be written as

$$f(x_1, x_2, \ldots, x_m) := \prod_{i=1}^{m} f_i(x_i),$$

where each $f_i\colon \{0,1\}^d \to [-1,1]$ is an arbitrary Boolean function.

Let $\mathcal{F}$ be the class of (m,d)-products. It is not hard to see that $\mathcal{F}$ is not fooled by even $(n-1)$-independence, because if we let $f_i(x_i) = \chi_{[d]}(x_i)$ then $f(x) = \chi_{[n]}(x)$ is simply the parity function on n bits. Moreover, it is easy to check that the distribution $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \sum_{i=1}^{n-1} \boldsymbol{x}_i)$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}$ are uniform bits, is $(n-1)$-independent, and we have that $\chi_{[n]}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \sum_i \boldsymbol{x}_i) = 1$ always, whereas $\mathbf{E}[\chi_{[n]}(\boldsymbol{x})] = 0$.

We now show that adding noise to a k-wise independent distribution fools product tests, where the noise is the distribution $N_\rho(x)$ we saw earlier in Definition 5.8, and for $\rho > 0$ this is

$$N_\rho(x)_i := \begin{cases} \text{uniform} & \text{with probability } 1-\rho \\ x_i & \text{with probability } \rho. \end{cases}$$

We will focus on $\rho = 1/2$, and will instead write $N_{1/2}(x)$ as $x+N$, where $N = N_{1/2}(\vec{0})$. For notational simplicity we use $\mathbf{E}[f(D)]$ to denote $\mathbf{E}_{\boldsymbol{x}\sim D}[f(\boldsymbol{x})]$ and U_n to denote the uniform distribution over $\{0,1\}^n$ (we omit the subscript n when it is clear). We will prove the following theorem.

Theorem 19.2 (Bounded independence plus noise fools products)**.** *Let D be a $2(d+k-1)$-wise independent distribution on $\{0,1\}^n$. Then $D+N$ fools any (m,d)-product f with error $m\cdot 2^{-k}$, that is*

$$\left|\mathbf{E}[f(D+N)] - \mathbf{E}[f(U)]\right| \le m \cdot 2^{-k}.$$

Setting $k = O(d + \log(m/\epsilon))$, we get distribution that fools (m,d)-products with error ϵ.

The intuition behind proving this theorem is to use D to fool the low-degree part of f, and N to fool the high-degree part of f. Specifically, we know that if $f\colon \{0,1\}^n \to \mathbb{R}$ is a degree-k function then

$$\mathbf{E}[f(D)] = \sum_{\alpha:|\alpha|\le k} \widehat{f}(\alpha)\,\mathbf{E}[\chi_\alpha(D)] = \sum_{\alpha:|\alpha|\le k} \widehat{f}(\alpha)\,\mathbf{E}[\chi_\alpha(U)] = \widehat{f}(\varnothing) = \mathbf{E}[f(U)].$$

On the other hand, we have $\mathbf{E}[\chi_\alpha(N)] = 2^{-|\alpha|}$, which is small when $|\alpha|$ is large.

Of course, $D+N$ cannot fool every function, so we have to rely on the fact the f is a product. Indeed, we have the following decomposition lemma.

Lemma 19.3. *Every (m,d)-product $f\colon \{0,1\}^n \to [-1,1]$ can be written as*

$$f(x_1, \ldots, x_n) = f^{\le k-1}(x_1, \ldots, x_n) + \sum_{i=1}^{m} h_i(x_1, \ldots, x_i) f_{>i}(x_{i+1}, \ldots, x_n),$$

where

1. $f_{\le k-1}(x) := \sum_{|\alpha|\le k} \widehat{f}(\alpha)\chi_\alpha(x)$,
2. $f_{>i}\colon \{0,1\}^{(n-i)\cdot d} \to [-1,1]$ *is defined as* $f_{>i}(x_{i+1},\dots,x_n) = \prod_{j=i+1}^{n} f(x_j)$, *and*
3. $h_i\colon \{0,1\}^{i\cdot d} \to \mathbb{R}$ *satisfies* $\mathbf{E}[h_i(U_{i\cdot d})^2] \le 1$ *and if* $\widehat{h_i}(\alpha) \ne 0$ *then* $|\alpha| \in [k, k+d-1]$.

We first show how to prove Theorem 19.2 using Lemma 19.3.

Proof of Theorem 19.2. First, because $N + D$ is also $2(k+d-1)$-wise independence, we have

$$\mathbf{E}[f^{\le k-1}(D+N)] = \widehat{f}(\varnothing) = \mathbf{E}[f(U)].$$

It remains to show that $|\mathbf{E}[(h_i f_{>i})(D+N)]| \le 2^{-k}$ for each $i \in [n]$. We have

$$\begin{aligned}
&\left|\mathop{\mathbf{E}}_{D,N}[h_i(D+N)\cdot f_{>i}(D+N)]\right| \\
&\quad\le \mathop{\mathbf{E}}_{D}\left[\left|\mathop{\mathbf{E}}_{N}[h_i(D+N)]\right|\cdot\left|\mathop{\mathbf{E}}_{N}[f_{>i}(D+N)]\right|\right] \\
&\quad\le \mathop{\mathbf{E}}_{D}\left[\left|\mathop{\mathbf{E}}_{N}[h_i(D+N)]\right|\right] && (|f_{>i}(x)| \le 1) \\
&\quad\le \mathop{\mathbf{E}}_{D}\left[\mathop{\mathbf{E}}_{N}[h_i(D+N)]^2\right]^{1/2} && (\text{Cauchy–Schwarz}) \\
&\quad\le \mathop{\mathbf{E}}_{U}\left[\mathop{\mathbf{E}}_{N}[h_i(U+N)]^2\right]^{1/2} && (\mathop{\mathbf{E}}_{N}[h_i(x+N)]^2 \text{ has degree} \le 2(k+d-1)) \\
&\quad\le \left(\sum_{d\le|\alpha|\le d+k-1} 2^{-|\alpha|}\cdot\widehat{h_i}(\alpha)^2\right)^{1/2} && (\widehat{h_i}(\alpha)\ne 0 \implies |\alpha|\ge d) \\
&\quad\le 2^{-k}\cdot\mathbf{E}[h(U)^2]^{1/2} \\
&\quad\le 2^{-k} && (\mathbf{E}[h(U)^2] \le 1).
\end{aligned}$$

□

We now prove the decomposition lemma (Lemma 19.3).

Proof of Lemma 19.3. We write

$$f(x_1,\dots,x_n) = f^{\le k-1}(x_1,\dots,x_n) + \sum_{i=1}^{m} h_i(x_1,\dots,x_i) f_{>i}(x_{i+1},\dots,x_n), \tag{10}$$

where $h_i\colon \{0,1\}^{i\cdot d} \to \mathbb{R}$ is defined as

$$h_i(x_1,\dots,x_i) = \sum_{\substack{(\alpha_1,\dots,\alpha_i)\in\{0,1\}^{i\cdot d}\\ \text{the } k\text{-th 1 lies in } \alpha_i}} \widehat{f_{\le i}}(\alpha)\chi_\alpha(x_1,\dots,x_i).$$

First it should be clear that if $\widehat{f_i}(\alpha) \ne 0$ then $|\alpha| \ge k$. Next, note that if the k-th 1 lies in the last block α_i of $(\alpha_1,\dots,\alpha_i)$, then $|\alpha| \le k+d-1$ as α can have at most $d-1$ many 1s after the k-th 1. By Parseval's we have $\mathbf{E}[h_i(U)^2] \le \mathbf{E}[f_{\le i}(U)^2] \le 1$.

It remains to argue that Equation (10) is a valid decomposition. For each α, we will show that $\widehat{f}(\alpha)$ appears uniquely in the decomposition. Here we use the property of products, where $\widehat{f}(\alpha_1,\dots,\alpha_m) = \widehat{f_1}(\alpha_1)\cdots\widehat{f_m}(\alpha_m)$. When $|\alpha| \le k-1$, clearly $\widehat{f}(\alpha)$ appears as a coefficient in $f^{\le k-1}$. Suppose $|\alpha| \ge k$. Let α_i be the block that contains the k-th 1. It follows that $\widehat{f_{\le i}}(\alpha_1,\dots,\alpha_i)$ appears in h_i, and $\widehat{f_{>i}}(\alpha_{i+1},\dots,\alpha_m)$ clearly appears in $f_{>i}$. So $\widehat{f}(\alpha)$ appears in $h_i f_{>i}$. □

19.1.1 Connection to space-bounded computation

A (non-uniform) streaming algorithm that uses s bits of space can be modeled by a *read-once branching program* of width $w = 2^s$. Its computation can be captured using a layered graph with $n+1$ layers, each consisting of w vertices representing the 2^s possible memory states of the algorithm. The program reads the n input bits one at a time and updates its current state in its current layer to some state to the next layer depending on the value of the bit. In the final layer, the states are partitioned into accept and reject states.

Definition 19.4 (Read-once branching programs)**.** A *read-once branching program* B of length n and width w computes a function $B\colon \{-1,1\}^n \to \{0,1\}$. It starts at a fixed start state $v_1 \in [w]$. Then for $t = 1, \ldots, n$, it reads the next input bit x_t and updates its state according to a transition function $B_t\colon [w] \times \{-1,1\} \to [w]$ by taking $v_{t+1} := B_t(v_t, x_t)$. Note that the transition function B_t can differ at each time step. The program has a fixed set of accept states $V_{\mathsf{acc}} \subseteq [w]$, and $B(x) = \mathbb{1}(v_{n+1} \in V_{\mathsf{acc}})$.

We can model the transition between two adjacent layers by a 1-bit matrix-valued function $B_i\colon \{0,1\} \to \{0,1\}^{w\times w}$, where

$$B_i(x_i)_{u,v} = \begin{cases} 1 & \text{if the program moves from state } u \text{ to state } v \text{ on } x_i \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, replacing the range $[-1,1]$ in products with $w \times w$ Boolean matrices $\{0,1\}^{w\times w}$, we see that a $(n,1)$-matrix-product captures a read-once branching programs of width w. Indeed, we can extend Theorem 19.2 to matrix products using essentially the same proof, by generalizing some notions we saw in Fourier analysis to matrix-valued functions,

As in the case for scalar-valued functions, every matrix-valued Boolean function $B\colon \{0,1\}^n \to \{0,1\}^{w\times w}$ has a Fourier expansion

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

where $\widehat{B}(S) \in \mathbb{R}^{w\times w}$ can be computed by

$$\widehat{B}(S) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{0,1\}^n}\big[B(\boldsymbol{x})\chi_S(\boldsymbol{x})\big].$$

To introduce a matrix analogue of Parseval's identity, we need to define the *Frobenius inner-product*.

Definition 19.5 (Frobenius inner-product)**.** Given two matrices $N, M \in \mathbb{R}^{w\times w}$, we define the *Frobenius inner-product* of N and M, denoted $\langle N, M\rangle_{Fr}$ by

$$\langle N, M\rangle_{Fr} := \operatorname{tr}(N^T M) = \sum_{i,j\in[w]} N_{i,j}M_{i,j}.$$

The Frobenius norm of a matrix $N \in \mathbb{R}^{w\times w}$, denoted by $\|N\|_{Fr}$, is

$$\|N\|_{Fr} = \langle N, N\rangle^{1/2} = \Big(\sum_{i,j} N_{i,j}^2\Big)^{1/2}.$$

Proposition 19.6 (Parseval's identity)**.** $\mathbf{E}_{\boldsymbol{x}}\left[\|B(\boldsymbol{x})\|_{Fr}^2\right] = \sum_{S\subseteq[n]}\|\widehat{B}(S)\|_{Fr}^2$ *for every* $B\colon \{0,1\}^n \to \mathbb{R}^{w\times w}$.

Proof. We have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\|B(\boldsymbol{x})\|_{Fr}^2\right] &= \mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\left\langle \sum_{S\subseteq[n]} \widehat{B}(S)\chi_S(\boldsymbol{x}), \sum_{T\subseteq[n]} \widehat{B}(T)\chi_T(\boldsymbol{x})\right\rangle\right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{S,T\subseteq[n]} \left\langle \widehat{B}(S), \widehat{B}(T)\right\rangle \chi_S(\boldsymbol{x})\chi_T(\boldsymbol{x})\right] \\
&= \sum_{S\subseteq[n]} \left\langle \widehat{B}(S), \widehat{B}(S)\right\rangle \\
&= \sum_{S\subseteq[n]} \|\widehat{B}(S)\|_{Fr}^2.
\end{aligned}$$

□

We summarize the discussion by stating the matrix analogue of Theorem 19.2.

Corollary 19.7 (Bounded independence plus noise fools matrix-valued products)**.** *Let* D *be a* $2k$*-wise independent distribution on* $\{0,1\}^n$*. Then* $D+N$ *fools any* $(m,1)$*-matrix-valued product* B*, where* $\mathbf{E}_{\boldsymbol{x}}[\|B_i(\boldsymbol{x})\|_{Fr}^2] \le w$ *for every* $i\in[n]$*, with error* $w\cdot m\cdot 2^{-k}$*, that is*

$$\left|\mathbf{E}[f(D+N)] - \mathbf{E}[f(U)]\right| \le w\cdot m\cdot 2^{-k}.$$

20.1 Polarizing random walk

In the last lecture, we showed that a noisy $O(d+\log(m/\epsilon))$-wise independent distribution $D+N$ fools (m,d)-products with error ϵ. In general sampling the distribution N is costly and hence it is not clear how to construct a PRG from this theorem. In this lecture, we will see one way of converting the noisy k-wise independent distribution to a PRG.

One alternative way of stating Theorem 19.2 is that k-wise independence fools smoothed products. Recall that $D+N$ is the same as $N_{1/2}(D)$, and so we can write $f(D+N)$ as $f(N_{1/2}(D))$. Averaging over noise, we have that D fools the smoothed function $T_{1/2}f(x) = \mathbf{E}_{\boldsymbol{y}\sim N_{1/2}(x)}[f(\boldsymbol{y})]$.

Yet another alternative way stating Theorem 19.2 is the following. Recall that the Fourier expansion of $T_{1/2}f(x)$ is

$$T_{1/2}f(x) := \sum_{S\subseteq[n]} 2^{-|S|}\widehat{f}(S)\chi_S(x).$$

Also recall that in Homework 1 Q3, we extend the domain of f to the solid cube $[-1,1]^n$ by defining

$$f(x) := \sum_{S\subseteq[n]} \widehat{f}(S)\prod_{i\in S} x_i.$$

It is easy to verify that $T_{1/2}f(x) = f(x/2)$, and so the distribution $D/2$ fools (m,d)-products. Note that the distribution $D/2$ is supported on $\{-1/2,1/2\}^n$ and not $\{-1,1\}^n$. In general, it is easier to construct a distribution on $[-1,1]^n$ that fools any classes of functions. Indeed, note that $f(\vec{0}) = \mathbf{E}[f(U)]$ and so the constant $\vec{0}$ fools every f. We now give a procedure that converts $D/2$ to a PRG.

Theorem 20.1. *Let D be a k-wise independent distribution on $\{-1,1\}^n$. If ρD fools a family of function $\mathcal{F}$, where $\mathcal{F}$ is closed under restrictions, with error ϵ, then there exists a PRG that fools $\mathcal{F}$ with error $\epsilon\cdot\log(n/\epsilon)/\rho^2$ and seed length $\log(n/\epsilon)/\rho^2\cdot O(k\log n)$.*

For now let us assume $\rho = 1/2$ and at the end we will mention where the dependence on ρ appears in the proof.

The idea is to take independent copies of $D/2$ and use them to take a random walk in $[-1,1]^n$. Let $\boldsymbol{X}^1,\dots,\boldsymbol{X}^T$ be i.i.d. copies of $D/2\in\{-1/2,1/2\}^n$. Consider the random walk $\boldsymbol{Y}^1,\dots,\boldsymbol{Y}^T\in[-1,1]^n$, where for each $i\in[n]$ we define

$$\boldsymbol{Y}_i^t := \begin{cases} \boldsymbol{X}_i^1 & \text{when } t=1 \\ \boldsymbol{Y}_i^{t-1} + (1-|\boldsymbol{Y}_i^{t-1}|)\cdot\boldsymbol{X}_i^t & \text{when } t>1 \end{cases}$$

So at each step and at each coordinate, we use $\boldsymbol{X}_i^t$ take a random walk in largest interval centered at $\boldsymbol{Y}_i^{t-1}$ in $[-1,1]$. Note that $\boldsymbol{Y}^t$ is a deterministic function of $\boldsymbol{X}^1,\dots,\boldsymbol{X}^t$.

We will show that (1) each step of the walk induces an error of ϵ to the error of the final PRG, and (2) the walk converges to the $\{-1,1\}^n$ in a few steps.

We first prove (1).

Claim 20.2. *For every $t\in[T]$ and every $f\in\mathcal{F}$, we have*

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{X}^1,\dots,\boldsymbol{X}^t}[f(\boldsymbol{Y}^t)] - \mathop{\mathbf{E}}_{\boldsymbol{X}^1,\dots,\boldsymbol{X}^{t-1}}[f(\boldsymbol{Y}^{t-1})]\right| \le \epsilon.$$

Proof. Given $\boldsymbol{Y}^{t-1} = y \in [-1,1]^n$, we will show how to "recenter" y using random restrictions. Consider the random restriction $R_y \colon [-1,1]^n \to [-1,1]^n$, where for each $i \in [n]$,

$$R_y(z)_i := \begin{cases} \operatorname{sgn}(y_i) & \text{with probability } |y_i| \\ z_i & \text{with probability } 1 - |y_i|. \end{cases}$$

Observe that $\mathbf{E}_{R_y}[R_y(z)_i] = y_i + (1-|y_i|)z_i$, and so by multilinearity of f and linearity of expectation, we have

$$f(y + (1-|y|)\boldsymbol{X}^t) = \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_y}[f|_{\boldsymbol{\tau}}(\boldsymbol{X}^t)]$$

(where the $|\cdot|$ on y is coordinate-wise). Since $\mathcal{F}$ is closed under restriction, we have $f_{\boldsymbol{\tau}} \in \mathcal{F}$, and so

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{X}^t}[f(y + (1-|y|)\boldsymbol{X}^t)] - f(y) \right| \le \left| \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_y}[f_{\boldsymbol{\tau}}(\boldsymbol{X}^t)] - f_{\boldsymbol{\tau}}(\vec{0}) \right| \le \epsilon.$$

Averaging over $\boldsymbol{Y}^{t-1}$ completes the proof. □

Therefore $f(\vec{0}) \approx_\epsilon \mathbf{E}[f(\boldsymbol{Y}^1)] \approx_\epsilon \cdots \approx_\epsilon \mathbf{E}[f(\boldsymbol{Y}^T)]$ and a T-step random walk would induce an error of $T\epsilon$. We now show (2) that $\boldsymbol{Y}^T$ gets exponentially close to a vertex in $\{-1,1\}^n$.

Claim 20.3. $\mathbf{E}[(\boldsymbol{Y}_i^T)^2] \ge 1 - 2^{-\Omega(T)}$.

We will pick T to be some $O(\log(n/\epsilon))$ so that $\mathbf{E}[(\boldsymbol{Y}_i^T)^2] \ge 1 - \epsilon/n$, and let our PRG be $\boldsymbol{Y} = Y(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^T)$ on $\{-1,1\}^n$, where $\boldsymbol{Y}_i = \operatorname{sgn}(\boldsymbol{Y}_i^T)$ for each $i \in [n]$. Let us see how this gives us a PRG and then we will prove Claim 20.3.

Claim 20.4. $|\mathbf{E}[f(\boldsymbol{Y})] - \mathbf{E}[f(\boldsymbol{Y}^T)]| \le \epsilon$.

Proof. Given $\boldsymbol{Y}^T$, consider the random variable $Z \sim \{-1,1\}^n$, where for each $i \in [n]$,

$$\boldsymbol{Z}_i = \begin{cases} \operatorname{sgn}(\boldsymbol{Y}_i^T) & \text{with probability } \frac{1}{2} + \frac{|\boldsymbol{Y}_i^T|}{2} \\ -\operatorname{sgn}(\boldsymbol{Y}_i^T) & \text{with probability } \frac{1}{2} - \frac{|\boldsymbol{Y}_i^T|}{2}. \end{cases}$$

We have $\mathbf{E}[\boldsymbol{Z}] = \boldsymbol{Y}^T$, which by linearity of expectation, implies $\mathbf{E}_{\boldsymbol{Z}}[f(\boldsymbol{Z})] = f(\boldsymbol{Y}^T)$. Hence,

$$\begin{aligned} \left| f(\boldsymbol{Y}^T) - f(\boldsymbol{Y}) \right| &\le \left| \mathop{\mathbf{E}}_{\boldsymbol{Z}}[f(\boldsymbol{Z})] - f(\operatorname{sgn}(\boldsymbol{Y}^T)) \right| \\ &\le 2 \cdot \mathbf{Pr}\big[\boldsymbol{Z}_i \ne \operatorname{sgn}(\boldsymbol{Y}_i^T) \text{ for some } i \in [n]\big] \\ &\le 2 \cdot \sum_{i=1}^n \frac{1}{2}\big(1 - |\boldsymbol{Y}_i^T|\big) \\ &\le \sum_{i=1}^n 1 - |\boldsymbol{Y}_i^T| \\ &\le \sum_{i=1}^n 1 - (\boldsymbol{Y}_i^T)^2. \end{aligned}$$

Averaging over $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^T$, we have

$$\left| \mathbf{E}[f(\boldsymbol{Y})] - \mathbf{E}[f(\boldsymbol{Y}^T)] \right| \le \sum_{i=1}^n \Big(1 - \mathbf{E}[(\boldsymbol{Y}_i^T)^2]\Big) \le \epsilon.$$ □

We now prove Claim 20.3.

Proof of Claim 20.3. Our goal is to show that $\boldsymbol{Y}^T$ gets closer and closer to $\{-1,1\}$. So a natural progress measure would be $1-|\boldsymbol{Y}_i^t|$ for $t \in [T]$. We now show that $1-|\boldsymbol{Y}_i^t| \leq (1-|\boldsymbol{Y}_i^{t-1}|)(1-\boldsymbol{X}_i^t)$. For notational simplicity, let us write $y' := \boldsymbol{Y}_i^t, y := \boldsymbol{Y}_i^{t-1}$ and $x := \boldsymbol{X}_i^t$; so we have

$$1-|y'| = 1-|y+(1-|y|)x|.$$

By symmetry we can assume $y > 0$, and we can further $x < 0$ as otherwise it only gets closer to 1. We have two cases:

(1) If $y > |(1-y)x|$, then

$$1-|y+(1-|y|)\cdot x| = 1-(y+(1-y)\cdot x) = (1-y)(1-x);$$

(2) if $y \leq |(1-y)x|$, then $y+(1-|y|)\cdot x \leq 0$ and so

$$1-|y+(1-|y|)\cdot x| = 1+(y+(1-|y|)\cdot x) \leq 1-(y+(1-|y|)\cdot x) = (1-y)(1-x).$$

Given that $1-|\boldsymbol{Y}_i^t| \leq (1-|\boldsymbol{Y}_i^{t-1}|)(1-\boldsymbol{X}_i^t)$ and $\boldsymbol{Y}^{t-1}$ is independent of $\boldsymbol{X}^t$, a naïve approach would be to iterate and obtain

$$1-\mathbf{E}\big[|\boldsymbol{Y}_i^T|\big] \leq \prod_{t=1}^{T}(1-\mathbf{E}\big[\boldsymbol{X}_i^t\big]).$$

However, we have $\mathbf{E}[\boldsymbol{X}_i^t] = 0$ and so it gives nothing this way. Looking closer, the $\boldsymbol{X}^t$'s are i.i.d. copies of $D/2$, so we expect half of them to be negative and the other half of them to be positive, in which case we get

$$\prod_{t=1}^{T}(1-\boldsymbol{X}_i^t) = (1+|D|/2)^{T/2}(1-|D|/2)^{t/2} = (1-(D/2)^2)^{T/2} \leq e^{-\Omega(T)},$$

which is what we want. We can prove this formally as follows. Taking square root of both sides, we have

$$\mathbf{E}\big[(1-|\boldsymbol{Y}_i^t|)^{1/2}\big] \leq \mathbf{E}\big[(1-|\boldsymbol{Y}_i^{t-1}|)^{1/2}\big]\cdot\mathbf{E}\big[(1-\boldsymbol{X}_i^t)^{1/2}\big]$$

and hence

$$\begin{aligned}\mathbf{E}\big[(1-|\boldsymbol{Y}_i^T|)^{1/2}\big] &\leq \prod_{t=1}^{T}\mathbf{E}\big[(1-\boldsymbol{X}_i^t)^{1/2}\big]\\ &\leq \prod_{t=1}^{T}\Big(1-\mathbf{E}[(\boldsymbol{X}_i^t)^2]\Big)\\ &\leq e^{-\Omega(T)}\end{aligned} \tag{11}$$

where the second last step follows from $\mathbf{E}[(\boldsymbol{X}_i^t)^k] = 0$ for odd k and the Taylor expansion of $(1-\boldsymbol{X}_i^t)^{1/2}$.

By Markov we have $(1-|\boldsymbol{Y}_i^T|)^{1/2} \geq e^{-\Omega(T)}$ with probability $e^{-\Omega(T)}$ and this random variable is always bounded by 1. So $\mathbf{E}[(\boldsymbol{Y}_i^T)^2] \geq 1-e^{-\Omega(T)}$. □

We have proved Theorem 20.1 assuming $\rho = 1/2$. Going over the proof again, we can see that the dependence on ρ would appear only in Equation (11), where we would get a bound of $e^{-\Omega(\rho^2 T)}$ instead of $e^{-\Omega(T)}$, and so we need to take T to be $O(\log(n/\epsilon)/\rho^2)$ for general ρ.

In general we can replace the distribution ρD with an object called *fractional PRG*, which is a PRG with $[-1,1]^n$ outputs.

Definition 20.5 (Fractional Pseudorandom Generators (fPRG))**.** A function $G\colon \{-1,1\}^s \to [-1,1]^n$ is a p-noticeable *fractional pseudorandom generator* for $\mathcal{F}$ with error ϵ if $|\mathbf{E}[f(U)] - \mathbf{E}[f(G(U))]| = |\mathbf{E}[f(U)] - f(\vec{0})| \leq \epsilon$ for every $f \in \mathcal{F}$ and $\mathbf{E}[D_i^2] \geq p$ for every $i \in [n]$.

You can verify that the argument we gave so far also gives a generic way to convert a p-noticeable fPRG for a family $\mathcal{F}$ that is closed under restrictions with seed length s and error ϵ to a PRG with seed length $O(\log(n/\epsilon)/p) \cdot s$ and error $O(\log(n/\epsilon)/p) \cdot \epsilon$.

Combining Theorems 19.2 and 20.1 we have the following PRG for product tests.

Corollary 20.6. *There exists a PRG that fools (m,d)-products with error ϵ and seed length $O(\log(n/\epsilon))(d + \log(m/\epsilon)) \log n$.*

21.1 Fourier Growth

Given a family of functions, we can ask for what ρ and k such that the ρ-noisy k-wise independent distribution ρD fools $\mathcal{F}$. One sufficient condition is the following measure of the Fourier spectrum.

Definition 21.1 (Fourier Growth)**.** A family of functions $\mathcal{F}$ has bounded Fourier growth if there exists an integer $b \ll \sqrt{n}$ such that for every $k \in [n]$ and $f \in \mathcal{F}$, we have

$$L_{1,k}[f] := \sum_{|S|=k} |\widehat{f}(S)| \leq b^k.$$

Note that by Cauchy–Schwarz we always have $\sum_{|S|=k} |\widehat{f}(S)| \leq \binom{n}{k}^{1/2} \leq n^{k/2}$. So this definition is indeed non-trivial only when $b \ll \sqrt{n}$.

Claim 21.2. *Let D be a $\log(1/\epsilon)$-wise distribution on $\{-1,1\}^n$. Let $\mathcal{F}$ be a family of functions such that $L_{1,k}[f] \leq b^k$ for every $f \in \mathcal{F}$ and $k \in [n]$. Then D/b^2 fools $\mathcal{F}$ with error ϵ.*

Proof. This follows straightforwardly from the intuition that D fools the low-degree part and the "noise" dampens the high-degree part of f.

$$\begin{aligned}
\left|\mathbf{E}[f(D/(2b)] - \widehat{f}(\varnothing)\right| &= \left|\sum_{|S|>\log(1/\epsilon)} \left(\frac{1}{2b}\right)^{|S|} \widehat{f}(S)\, \mathbf{E}[\chi_S(D)]\right| \\
&\leq \sum_{k>\log(1/\epsilon)} \left(\frac{1}{2b}\right)^{k} \sum_{|S|=k} |\widehat{f}(S)| \\
&\leq \sum_{k>\log(1/\epsilon)} 2^{-k} \leq \epsilon.
\end{aligned}$$

□

One sufficient condition for a family of functions to have bounded Fourier growth is that they get simplified to a low degree function under a typical random restriction.

Lemma 21.3. *Suppose $\mathbf{Pr}_{\boldsymbol{\tau}\sim R_\rho}[\deg(f|_{\boldsymbol{\tau}}) = d] \leq (\rho t)^d$ for every $\rho > 0$ and $d \in [n]$. Then $L_{1,k}[f] \leq (8t)^k$ for every $k \in [n]$.*

Recall that the degree of a Boolean function is at most its decision tree depth. Therefore from Lemma 10.1, we can conclude that $L_{1,k}[f] \leq O(w)^k$ for every width-w DNF f.

Proof of Lemma 21.3. We are going to relate the $L_{1,k}$ of f to the $L_{1,k}$ of its restrictions. It is easy to verify that

$$L_{1,k}[f] = \frac{1}{\rho^k} \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[L_{1,k}[f|_{\boldsymbol{\tau}}]\right].$$

Let $L_1[f] := \sum_{d=1}^n L_{1,d}[f]$. Recall in HW1 Q4 we showed that a degree-d Boolean function g has at most 4^{d-1} many non-zero coefficients, and so $L_1[g] \leq 4^{d-1}$. Hence,

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[L_1[f|_{\boldsymbol{\tau}}]\right] &= \sum_{d=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{\tau}\sim R_\rho}\left[L_1[f|_{\boldsymbol{\tau}}] \mid \deg(f|_{\boldsymbol{\tau}}) = d\right] \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{\tau}\sim R_\rho}\left[\deg(f|_{\boldsymbol{\tau}}) = d\right] \\
&\leq \sum_{d=1}^{n} 4^{d-1} \cdot (\rho t)^d.
\end{aligned}$$

Therefore, setting $\rho = 1/(8t)$, we have

$$\begin{aligned} L_{1,k}[f] &= \frac{1}{\rho^k} \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_\rho} \left[L_{1,k}[f_{\boldsymbol{\tau}}] \right] \\ &\leq \frac{1}{\rho^k} \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_\rho} \left[L_1[f_{\boldsymbol{\tau}}] \right] \\ &\leq (8t)^k \sum_{d=1}^{n} 4^{d-1} \cdot \left(\frac{t}{8t} \right)^d \leq (8t)^k. \end{aligned}$$

□

We can also bound the Fourier growth of AC^0 circuits by showing that the exponential tail bound in Theorem 13.3 implies degree shrinkage under restrictions.

Lemma 21.4. *Suppose $W^{\leq k}[f] \leq e^{-k/t}$ for every $k \in [n]$. Then $\mathbf{Pr}_{\boldsymbol{\tau} \sim R_\rho}[\deg(f|_{\boldsymbol{\tau}}) = d] \leq (\rho t)^d$.*

Recall that Theorem 13.3 says that if f is computable by a size-m depth-d circuit, then

$$W^{\geq k}[f] \leq 2 \cdot 2^{-\frac{k}{O(\log m)^{D-1}}}.$$

So Lemmas 21.3 and 21.4 together imply that $L_{1,k}[f] \leq O(\log m)^{(D-1)k}$.

Proof of Lemma 21.4. We bound $\mathbf{E}_{\boldsymbol{\tau} \sim R_\rho}[W^{=k}[f|_{\boldsymbol{\tau}}]$ in two ways. First we have

$$\begin{aligned} \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_\rho} [W^{=k}[f|_{\boldsymbol{\tau}}] &= \sum_{d=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_\rho} \left[W^{=k}[f|_{\boldsymbol{\tau}}] \mid \deg(f|_{\boldsymbol{\tau}}) = d \right] \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{\tau} \sim R_\rho} \left[\deg(f|_{\boldsymbol{\tau}}) = d \right] \\ &\geq \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_\rho} \left[W^{=k}[f|_{\boldsymbol{\tau}}] \mid \deg(f|_{\boldsymbol{\tau}}) = k \right] \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{\tau} \sim R_\rho} \left[\deg(f|_{\boldsymbol{\tau}}) = k \right] \\ &\geq 4^{-k} \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{\tau} \sim R_\rho} \left[\deg(f|_{\boldsymbol{\tau}}) = k \right], \end{aligned}$$

where the last step again uses the fact the coefficients of degree-k Boolean function are integer multiple of $2^{-(1-k)}$. Now, we relate the $W^{=k}[f]$ to $W^{=k}[f|_{\boldsymbol{\tau}}]$ as in the proof of Claim 12.3. We have

$$\begin{aligned} \mathop{\mathbf{E}}_{\boldsymbol{\tau} \sim R_\rho} \left[W^{=k}[f|_{\boldsymbol{\tau}}] \right] &= \sum_{U \subseteq [n]} \widehat{f}(U)^2 \, \mathbf{Pr}[\mathsf{Bin}(d, \rho) = k] \\ &\leq \sum_{d \geq k} W^{=d}[f] \binom{|U|}{k} \rho^k \\ &= \rho^k \cdot \mathbf{I}^k[f] \\ &\leq (\rho t)^k, \end{aligned}$$

where $\mathbf{I}^k[f] := \sum_{d \geq k} \binom{d}{k} W^{=d}[f]$ is the degree-k total influence of f introduced in HW2 Q1, and the last step follows from part (d) of the question. Combining the two parts we get that $L_{1,k}[f] \leq (4\rho t)^k$. □