COMP 760

Analysis of Boolean Functions

Lecture Notes

Hamed Hatami

Chapter 1

Background: Basic Analysis

This chapter provides some background material from measure theory, probability theory, and functional analysis. We will not immediately need these definitions and results; the reader can skip this chapter or some parts in the first reading and return to it when we refer to the material in future chapters.

1.1 Some basic inequalities

One of the most basic inequalities in analysis concerns the arithmetic and geometric mean. It is sometimes called AM-GM inequality.

Theorem 1.1. *The geometric mean of n non-negative reals is less than or equal to their arithmetic mean: If $a_1, \ldots, a_n$ are non-negative reals, then*

$$(a_1 \ldots a_n)^{1/n} \leqslant \frac{a_1 + \ldots + a_n}{n}.$$

In 1906, Jensen founded the theory of convex functions and proved a significant extension of the AM-GM inequality. We call a subset D of a real vector space *convex* if every convex linear combination of a pair of points of D belongs to D. Equivalently, if $x, y \in D$, then $tx + (1-t)y \in D$ for every $t \in [0,1]$. Given a convex set D, we call a function $f : D \to \mathbb{R}$ *convex* if for every $t \leqslant [0,1]$,

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y).$$

If the inequality is strict for every $t \in (0,1)$, then the function is called *strictly convex*.

Note that f is a convex function if and only if $\{(x,y) \in D \times \mathbb{R} : y \geqslant f(x)\}$ is a convex set. Also note that $f : D \to \mathbb{R}$ is convex if and only if $f_{xy} : [x,y] \to \mathbb{R}$ with $f_{xy} : tx + (1-t)y \mapsto tf(x) + (1-t)f(y)$ is convex for every $x, y \in D$. By Rolle's theorem, if f_{xy} is twice differentiable, this condition is equivalent to $f''_{xy} \geqslant 0$.

A function $f : D \to \mathbb{R}$ is *concave* if $-f$ is convex. The following theorem is one of the most useful inequalities in analysis.

Theorem 1.2 (Jensen's inequality)**.** *If $f : D \to \mathbb{R}$ is a concave function, then for every $x_1, \ldots, x_n \in D$ and $t_1, \ldots, t_n \geqslant 0$ with $\sum_{i=1}^n t_i = 1$ we have*

$$t_1 f(x_1) + \ldots + t_n f(x_n) \leqslant f(t_1 x_1 + \ldots + t_n x_n).$$

Furthermore, if f is strictly concave, then the equality holds if and only if all x_i are equal.

The most frequently used inequalities in functional analysis are the Cauchy-Schwarz inequality, Hölder's inequality, and Minkowski's inequality.

Theorem 1.3 (Cauchy-Schwarz)**.** *If $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are complex numbers, then*

$$\left| \sum_{i=1}^n x_i \overline{y_i} \right| \leqslant \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} \left(\sum_{i=1}^n |y_i|^2 \right)^{1/2}.$$

Hölder's inequality is an important generalization of the Cauchy-Schwarz inequality.

Theorem 1.4 (Hölder's inequality)**.** *Let $x_1, \dots, x_n$ and $y_1, \dots, y_n$ be complex numbers, and $p, q > 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\left| \sum_{i=1}^{n} x_i \overline{y_i} \right| \leqslant \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q \right)^{1/q}.$$

The numbers p and q appearing in Theorem 1.4 are called *conjugate exponents*. Moreover, $p = 1$ and $q = \infty$ are also called conjugate exponents, and Hölder's inequality in this case becomes:

$$\left| \sum_{i=1}^{n} x_i \overline{y_i} \right| \leqslant \left(\sum_{i=1}^{n} |x_i| \right) \left(\max_{i=1}^{n} |y_i| \right).$$

Finally, let us state Minkowski's inequality, which corresponds to the triangle inequality for ℓ_p norms.

Theorem 1.5 (Minkowski's inequality)**.** *If $p \geqslant 1$ is a real number, and $x_1, \dots, x_n$ are complex numbers, then*

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p} \leqslant \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p}.$$

The case of $p = \infty$ of Minkowski's inequality is the following:

$$\max_{i=1}^{n} |x_i + y_i| \leqslant \left(\max_{i=1}^{n} |x_i| \right) + \left(\max_{i=1}^{n} |y_i| \right).$$

1.2 Measure and Probability Spaces

In this course, we will mainly work with measures over finite sets. However, to provide a reference for the interested reader and put the concepts in a broader context, we state the following definitions in a more general form.

A *σ-algebra* over a set Ω is a collection $\mathcal{F}$ of subsets of Ω that satisfies the following three properties:

- We have $\emptyset \in \mathcal{F}$.
- It is closed under taking complements. That is, if $A \in \mathcal{F}$, then $A^c \coloneqq \Omega \setminus A$ also belongs to $\mathcal{F}$.
- It is closed under any countable union of its members. That is, if $A_1, A_2, \dots$ belong to $\mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Example 1.6. Let Ω be an arbitrary set. Then $\mathcal{F} = \{\emptyset, \Omega\}$ is called the *minimal* or *trivial* σ-algebra over Ω. The power set of Ω, denoted by $\mathcal{P}(\Omega)$, is the *maximal* σ-algebra over Ω.

For two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ over Ω, if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then we say that $\mathcal{F}_2$ is *finer* than $\mathcal{F}_1$, or that $\mathcal{F}_1$ is *coarser* than $\mathcal{F}_2$. Note that the trivial σ-algebra is the coarsest σ-algebra over Ω, while the maximal σ-algebra is the finest σ-algebra over Ω.

Definition 1.7 (measure and probability spaces)**.** A *measure space* is a triple $(\Omega, \mathcal{F}, \mu)$ where $\mathcal{F}$ is a σ-algebra over Ω and the measure $\mu : \mathcal{F} \to [0, \infty) \cup \{+\infty\}$ satisfies the following two axioms:

- Null empty set: $\mu(\emptyset) = 0$.
- Countable additivity: if $\{E_i\}_{i \in \mathcal{I}}$ is a countable set of *pairwise disjoint sets* in $\mathcal{F}$, then

$$\mu(\cup_{i \in \mathcal{I}} E_i) = \sum_{i \in \mathcal{I}} \mu(E_i).$$

The function μ is called a *measure*, and the elements of $\mathcal{F}$ are called *measurable sets*.

If furthermore $\mu : \mathcal{F} \to [0, 1]$ and $\mu(\Omega) = 1$, then $(\Omega, \mathcal{F}, \mu)$ is a *probability measure*. In this case, the sets $E \in \mathcal{F}$ are called events, and $\mu(E)$ is the probability that the event E occurs.

Definition 1.8 (Counting Measure)**.** The *counting measure* on Ω is the triple $(\Omega, \mathcal{P}(\Omega), \mu)$ where the measure of a subset $S \subseteq \Omega$ is the number of elements in S. Note that $\mu(S) = \infty$ if S is infinite.

When Ω is a finite set, another natural measure is associated with Ω: the uniform probability measure, which assigns an equal weight of $\frac{1}{|\Omega|}$ to every element.

Definition 1.9 (Uniform Probability Measure)**.** The *uniform probability measure* on a *finite set* Ω is the triple $(\Omega, \mathcal{P}(\Omega), \mu)$ with $\mu(S) = \frac{|S|}{|\Omega|}$ for all $S \subseteq \Omega$.

A measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$ is called *σ-finite* if Ω is a countable union of measurable sets of finite measure. In other words, there exists sets $E_1, E_2, \ldots$ in $\mathcal{F}$ such that $\mu(E_i) < \infty$ and $\Omega = \bigcup_{i=1}^{\infty} E_i$. The class of σ-finite measures has many convenient properties.

Every measure space in this course is assumed to be σ-finite.

For many natural measure spaces $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$, it is difficult to specify the elements of the σ-algebra $\mathcal{F}$. Instead, to definite μ, one specifies μ on a subcollection $\mathcal{F}' \subseteq \mathcal{F}$ that uniquely determines $\mathcal{M}$. To make this rigorous, we need the following definition.

Definition 1.10. For a set Ω, a collection $\mathcal{A}$ of subsets of Ω is called an *algebra* if

- $\emptyset \in \mathcal{A}$.
- $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.
- $A, B \in \mathcal{A}$, then $A \setminus B \in \mathcal{A}$.

The σ-algebra generated by $\mathcal{A}$ is the minimal σ-algebra containing $\mathcal{A}$.

Example 1.11. Let $\mathcal{A}$ be the set of all *finite* unions of disjoint (open, closed, or half-open) intervals in $\mathbb{R}$. Then $\mathcal{A}$ is an algebra over $\mathbb{R}$, but it is not a σ-algebra as it is not closed under taking countable unions.

A function $\mu : \mathcal{A} \to [0, \infty) \cup \{+\infty\}$ is a *measure over an algebra* $\mathcal{A}$ if for every finite set of disjoint $E_1, \ldots, E_m \in \mathcal{A}$, we have

$$\mu(\cup_{i=1}^{m} E_i) = \sum_{i=1}^{m} \mu(E_i).$$

The following theorem shows that to define a measure space $(\Omega, \mu, \mathcal{F})$, it suffices to specify μ on an algebra $\mathcal{A}$ that generates $\mathcal{F}$. By this theorem, such a measure extends to $\mathcal{F}$ uniquely.

Theorem 1.12 (Carathéodory's extension theorem)**.** *Let $\mathcal{A}$ be an algebra of subsets of a given set Ω. One can always extend a σ-finite measure μ on $\mathcal{A}$ to the σ-algebra generated by $\mathcal{A}$; moreover, the extension is unique.*

Example 1.13 (Borel measure on $\mathbb{R}$)**.** Let $\mathcal{A}$ be the algebra on $\mathbb{R}$, defined in Example 1.11. Set the measure of an (open, closed, or half-open) interval as its length and, more generally, the measure of a finite union of disjoint intervals to be the sum of their lengths.

By Carathéodory's extension theorem, μ extends uniquely to the σ-algebra generated by $\mathcal{A}$. The generated σ-algebra on $\mathcal{R}$ is called the Borel σ-algebra on $\mathbb{R}$, and the resulting measure, the *Borel* measure.

Product Measure: Consider two σ-finite measure spaces $\mathcal{M}_1 := (\Omega_1, \mathcal{F}_1, \mu_1)$ and $\mathcal{M}_2 := (\Omega_2, \mathcal{F}_2, \mu_2)$. Let $\mathcal{F}_1 \otimes \mathcal{F}_2$ denote the σ-algebra on the Cartesian product $\Omega_1 \times \Omega_2$ generated by subsets of the form $A_1 \times A_2$ with $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. It should be noted that $\mathcal{F} \times \mathcal{G}$ is *not* the Cartesian product of the two sets $\mathcal{F}$ and $\mathcal{G}$, and instead it is the σ-algebra generated by this Cartesian product.

We define the *product measure* $\mathcal{M}_1 \times \mathcal{M}_2 := (\Omega \times \Sigma, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \times \mu_2)$ as follows: For $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$, let $\mu_1 \times \mu_2(F_1 \times F_2) := \mu_1(F_1)\mu_2(F_2)$. One can use Theorem 1.12 to show that $\mu_1 \times \mu_2$ extends uniquely to a measure over all of $\mathcal{F} \times \mathcal{G}$, as desired.

Random Variables: Let $\mathcal{P} = (\Omega, \mathcal{F}, \mu)$ be a measure space. Let $(\Sigma, \mathcal{E})$ be a pair where $\mathcal{E}$ is a σ-algebra over Σ. A function $X : \Omega \to \Sigma$ is called *measurable* if the preimage of every set in $\mathcal{E}$ belongs to $\mathcal{F}$. In other words, for every $E \in \mathcal{E}$,

$$X^{-1}(E) \coloneqq \left\{X^{-1}(a) \mid a \in E\right\} \in \mathcal{F}.$$

If $\mathcal{P}$ is a probability space, then X is called a *random variable*. In this case, the random variable X induces a probability distribution on $(\Sigma, \mathcal{E})$: for every $E \in \mathcal{E}$, we have

$$\mathbf{Pr}[X \in E] \coloneqq \mu(X^{-1}(E)).$$

Example 1.14. Let $\Omega = \{00, 01, 10, 11\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, and let μ be the uniform probability measure on Ω. Define $X : \Omega \to \mathbb{N}$ as $X(00) = 0$, $X(10) = X(01) = 1$, and $X(11) = 2$, corresponding to the number of 1's in the string. Here, $\mathbb{N}$ refers to the set of natural numbers endowed with the discrete σ-algebra $\mathcal{P}(\mathbb{N})$, which is the power set of $\mathbb{N}$. Note that X is a random variable, and for example, we have

$$\mathbf{Pr}[X \in \{1, 2\}] = \mu(\{10, 01, 11\}) = \frac{3}{4}.$$

We finish this section by stating the Borel-Cantelli theorem.

Theorem 1.15 (Borel-Cantelli)**.** *Let $(E_n)_{n=1}^{\infty}$ be a sequence of events in some probability space. If the sum of the probabilities of E_n is finite, then the probability that infinitely many of them occur is 0, that is,*

$$\sum_{n=1}^{\infty} \mathbf{Pr}[E_n] < \infty \Rightarrow \mathbf{Pr}[\limsup_{n\to\infty} E_n] = 0,$$

where

$$\limsup_{n\to\infty} E_n := \bigcap_{n=1}^{\infty} \bigcup_{k=1}^{n} E_k.$$

1.3 Normed spaces

A *metric space* is an ordered pair (M, d) where M is a set and d is a *metric* on M, that is, a function $d : M \times M \to [0, \infty)$ such that

- Non-degeneracy: $d(x, y) = 0$ if and only if $x = y$.
- Symmetry: $d(x, y) = d(y, x)$, for every $x, y \in M$.
- Triangle inequality: $d(x, z) \leqslant d(x, y) + d(y, z)$, for every $x, y, z \in M$.

A sequence $\{x_i\}_{i=1}^{\infty}$ of elements of a metric space (M, d) is called a *Cauchy sequence* if for every $\varepsilon > 0$, there exist an integer N_ε, such that for every $m, n \geqslant N_\varepsilon$, we have $d(x_m, x_n) \leqslant \varepsilon$. A metric space (M, d) is *complete* if every Cauchy sequence has a limit in M. A metric space is *compact* if every sequence has a convergent subsequence.

Next, we define a normed space, a central concept to function analysis.

Definition 1.16. A *normed space* is a pair $(V, \|\cdot\|)$, where V is a vector space over $\mathbb{R}$ or $\mathbb{C}$, and $\|\cdot\|$ is a function from V to nonnegative reals satisfying

- (non-degeneracy): $\|x\| = 0$ if and only if $x = 0$.
- (homogeneity): For every scalar λ, and every $x \in V$, $\|\lambda x\| = |\lambda|\|x\|$.
- (triangle inequality): For $x, y \in V$, $\|x + y\| \leqslant \|x\| + \|y\|$.

We call $\|x\|$, the *norm* of x. A *semi-norm* is a function similar to a norm except that it might not satisfy the non-degeneracy condition.

Example 1.17. The spaces $(\mathbb{C}, |\cdot|)$ and $(\mathbb{R}, |\cdot|)$ are respectively examples of 1-dimensional complex and real normed spaces.

Every normed space $(V, \|\cdot\|)$ has a metric space structure where the distance of two vectors x and y is $\|x-y\|$.

Consider two normed spaces X and Y. A *bounded operator* from X to Y, is a *linear* function $T : X \to Y$, such that

$$\|T\| := \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X} < \infty. \tag{1.1}$$

The set of all bounded operators from X to Y is denoted by $B(X,Y)$. Note that the *operator norm* defined in (1.1) makes $B(X,Y)$ a normed space.

A *linear functional* on a normed space X over $\mathbb{C}$ (or $\mathbb{R}$) is a bounded linear map $f : X \to \mathbb{C}$ (respectively $\mathbb{R}$), where bounded means

$$\|f\| := \sup_{x \neq 0} \frac{|f(x)|}{\|x\|} < \infty.$$

The set of all bounded linear functionals on X endowed with the operator norm, is called *the dual* of X and is denoted by X^*. So for a normed space X over complex numbers, $X^* = B(X, \mathbb{C})$, and similarly for a normed space X over real numbers, $X^* = B(X, \mathbb{R})$.

For a normed space X, the set $\mathbf{B}_X := \{x : \|x\| \leqslant 1\}$ is called the *unit ball* of X. Note that by the triangle inequality, $\mathbf{B}_X$ is a convex set, and also by homogeneity, it is symmetric around the origin (i.e., $x \in \mathbf{B}_X$ iff $-x \in \mathbf{B}_X$). The non-degeneracy condition implies that $\mathbf{B}_X$ has a non-empty interior.

Every compact symmetric convex subset of $\mathbb{R}^n$ with a non-empty interior is called a *convex body*. Convex bodies are in one-to-one correspondence with norms on $\mathbb{R}^n$. A convex body K corresponds to the norm $\|\cdot\|_K$ on $\mathbb{R}^n$, where

$$\|x\|_K := \sup\{\lambda \in [0,\infty) : \lambda x \in K\}.$$

Note that K is the unit ball of $\|\cdot\|_K$. For a set $K \subseteq \mathbb{R}^n$, define its *polar conjugate* as

$$K^\circ = \{x \in \mathbb{R}^n : \sum x_i y_i \leqslant 1,\ \forall y \in K\}. \tag{1.2}$$

The polar conjugate of a convex body K is a convex body, and furthermore $(K^\circ)^\circ = K$.

Consider a normed space X on $\mathbb{R}^n$. For $x \in \mathbb{R}^n$ define $T_x : \mathbb{R}^n \to \mathbb{R}$ as $T_x(y) := \sum_{i=1}^n x_i y_i$. It is easy to see that T_x is a linear functional on X, and every functional on X is of the form T_x for some $x \in \mathbb{R}^n$. For $x \in \mathbb{R}^n$ define $\|x\|^* := \|T_x\|$. This shows that we can identify X^* with $(\mathbb{R}^n, \|\cdot\|^*)$. Let K be the unit ball of $\|\cdot\|$. It is easy to see that K°, the polar conjugate of K, is the unit ball of $\|\cdot\|^*$.

1.4 Hilbert Spaces

Consider a vector space V over $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. An *inner product* $\langle\cdot,\cdot\rangle$ on V, is a function from $V \times V$ to $\mathbb{K}$ that satisfies the following axioms.

- Conjugate symmetry: $\langle x, y\rangle = \overline{\langle y, x\rangle}$.
- Linearity in the first argument: $\langle ax + z, y\rangle = a\langle x, y\rangle + \langle z, y\rangle$ for $a \in \mathbb{K}$ and $x, y \in V$.
- Positive-definiteness: $\langle x, x\rangle > 0$ if and only if $x \neq 0$, and $\langle 0, 0\rangle = 0$.

An *inner product space* is a vector space endowed with an inner product.

Example 1.18. Let Ω be a finite set endowed with the uniform probability measure and consider the vector space V of all functions $f : \Omega \to \mathbb{C}$. For $f, g : \Omega \to \mathbb{C}$ define

$$\langle f, g\rangle := \mathbb{E}_{x\in\Omega} f(x)\overline{g(x)} = \frac{1}{|\Omega|} \sum_{x \in |\Omega|} f(x)\overline{g(x)}.$$

Note that this is a valid inner product as it satisfies all the axioms of an inner product.

Example 1.19. More generally, consider a measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$, and let $\mathcal{H}$ be the space of measurable functions $f : \Omega \to \mathbb{C}$ such that $\int |f(x)|^2 d\mu(x) < \infty$. For two functions $f, g \in \mathcal{H}$ define

$$\langle f, g \rangle \coloneqq \int f(x)\overline{g(x)} d\mu(x).$$

An inner product naturally defines a norm on V. For a vector $x \in V$, define $\|x\| \coloneqq \sqrt{\langle x, x \rangle}$.

Lemma 1.20. *For an inner product space V, the function $\|\cdot\| : x \mapsto \sqrt{\langle x, x \rangle}$ is a norm. Furthermore, it satisfies the Cauchy-Schwarz inequality*

$$|\langle x, y \rangle| \leqslant \|x\|\|y\|.$$

Proof. First, note that for every $x \in V$, we have

$$\langle 0, x \rangle = \langle x - x, x \rangle = \langle x, x \rangle - \langle x, x \rangle = 0,$$

and similarly $\langle x, 0 \rangle = 0$.

To verify the Cauchy-Schwarz inequality, we may assume that $\langle x, y \rangle \neq 0$ and $x, y \neq 0$ as otherwise the Cauchy-Schwarz inequality is trivial. By the positive-definiteness of the inner product,

$$0 \leqslant \langle x + \lambda y, x + \lambda y \rangle = \langle x, x \rangle + |\lambda|^2 \langle y, y \rangle + \lambda \overline{\langle x, y \rangle} + \overline{\lambda} \langle y, x \rangle.$$

Now taking $\lambda \coloneqq \sqrt{\frac{\langle x,x \rangle}{\langle y,y \rangle}} \times \frac{\langle x,y \rangle}{|\langle x,y \rangle|}$ shows that

$$0 \leqslant 2\langle x, x \rangle \langle y, y \rangle - 2\sqrt{\langle x, x \rangle \langle y, y \rangle} |\langle x, y \rangle|,$$

which shows

$$|\langle x, y \rangle| \leqslant 2\sqrt{\langle x, x \rangle \langle y, y \rangle}.$$

It remains to show that $\|x\| = \sqrt{\langle x, x \rangle}$ is a norm. The non-degeneracy and homogeneity conditions are trivially satisfied. To verify the triangle inequality, note that by the Cauchy-Schwarz inequality, we have

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \leqslant \|x\|\|x\| + \|x\|\|y\| + \|y\|\|x\| + \|y\|\|y\| = (\|x\| + \|y\|)^2.$$

Therefore,

$$\|x + y\| \leqslant \|x\| + \|y\|.$$

□

A *Hilbert space* is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the norm $\|x\| \coloneqq \sqrt{\langle x, x \rangle}$. It is easy to verify that every finite-dimensional inner product space is a Hilbert space.

Example 1.21. Consider the vector space V of all functions $f : \mathbb{N} \to \mathbb{R}$ that have finite supports, meaning that $\{x : f(x) \neq 0\}$ is finite. This is a vector space over $\mathbb{R}$ and can be turned into an inner product space with the inner product

$$\langle u, v \rangle = \sum_{i \in \mathbb{N}} u_i v_i.$$

However, this is not a Hilbert space as it is not complete. For example, consider the sequence of vectors

$$u^{(k)} = (1, 2^{-1}, 2^{-2}, \ldots, 2^{-k}, 0, 0, \ldots).$$

It is easy to see that $u^{(1)}, u^{(2)}, \ldots$ is a Cauchy sequence, but it does not have a limit in V, and hence V is not a Hilbert space. However, we can complete V to a Hilbert space by extending it to include all functions $f : \mathbb{N} \to \mathbb{R}$ with

$$\sum_{i \in \mathbb{N}} |f(i)|^2 < \infty.$$

1.5 L_p spaces

Consider a measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$. For $1 \leqslant p < \infty$, the space $L_p(\mathcal{M})$ is the space of all functions $f : \Omega \to \mathbb{C}$ such that

$$\|f\|_{L_p(\mu)} \coloneqq \left(\int |f(x)|^p d\mu(x) \right)^{1/p} < \infty.$$

When μ is clear from the context, and there is no ambiguity, we write $\|f\|_p$ instead of $\|f\|_{L_p(\mu)}$

Strictly speaking, every element in $L_p(\mu)$ is an equivalent class: Two functions f_1 and f_2 are equivalent and are considered identical if they agree almost everywhere or equivalently $\|f_1 - f_2\|_{L_p(\mu)} = 0$.

Proposition 1.22. *For every measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$, the vector space $L_p(\mathcal{M})$ is a normed space.*

Proof. Non-degeneracy and homogeneity are trivial. It remains to verify the triangle inequality (or equivalently prove Minkowski's inequality). By applying Hölder's inequality:

$$\begin{aligned}
\|f+g\|_p^p &= \int |f(x)+g(x)|^p d\mu(x) = \int |f(x)+g(x)|^{p-1} |f(x)+g(x)| d\mu(x) \\
&\leqslant \int |f(x)+g(x)|^{p-1} |f(x)| d\mu(x) + \int |f(x)+g(x)|^{p-1} |g(x)| d\mu(x) \\
&\leqslant \left(\int |f(x)+g(x)|^p d\mu(x) \right)^{\frac{p-1}{p}} \|f\|_p + \left(\int |f(x)+g(x)|^p d\mu(x) \right)^{\frac{p-1}{p}} \|g\|_p \\
&= \|f+g\|_p^{p-1} (\|f\|_p + \|g\|_p),
\end{aligned}$$

which simplifies to the triangle inequality □

Another useful fact about the L_p norms is that when defined on a probability space, they are monotone increasing in the parameter p.

Theorem 1.23. *Let $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$ be a probability space, $1 \leqslant p \leqslant q \leqslant \infty$ be real numbers, and $f \in L_q(\mathcal{M})$. Then $f \in L_p(\mathcal{M})$, and*

$$\|f\|_p \leqslant \|f\|_q.$$

Proof. The case $q = \infty$ is trivial. For the case $q < \infty$, by Hölder's inequality (applied with conjugate exponents $\frac{q}{p}$ and $\frac{q}{q-p}$), we have

$$\|f\|_p^p = \int |f(x)|^p \times 1 d\mu(x) \leqslant \left(\int |f(x)|^q d\mu(x) \right)^{p/q} \left(\int 1^{\frac{q}{q-p}} d\mu(x) \right)^{\frac{q-p}{q}} = \|f\|_q^p.$$

□

Theorem 1.23 does not hold when $\mathcal{M}$ is not a probability space. For example, consider the set of natural numbers $\mathbb{N}$ with the counting measure. It is common to use the notation $\ell_p(\mathbb{N}) \coloneqq L_p(\mathbb{N})$ when we consider the counting measure on $\mathbb{N}$. In this case,

$$\|f\|_{\ell_p} = \left(\sum_{n=1}^{\infty} |f(n)|^p \right)^{1/p},$$

and it is not difficult to verify that the ℓ_p norms are actually decreasing.

Proposition 1.24. *Let $1 \leqslant p \leqslant q \leqslant \infty$ be real numbers, and $f \in \ell_p(\mathbb{N})$. Then $f \in \ell_q(\mathbb{N})$ and*

$$\|f\|_{\ell_p} \geqslant \|f\|_{\ell_q}.$$

1.6 Exercises

Exercise 1.1. Prove Proposition 1.24.

Exercise 1.2. Recall that by Hölder's inequality, if $p, q \geqslant 1$ are conjugate exponents and $a_1, \ldots, a_n, b_1, \ldots, b_n$ are complex numbers, then

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leqslant \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |b_i|^q \right)^{1/q}.$$

Deduce from this, that if $\mu(1), \ldots, \mu(n)$ are non-negative numbers with $\sum_{i=1}^{n} \mu(i) = 1$, then

$$\left| \sum_{i=1}^{n} a_i b_i \mu(i) \right| \leqslant \left(\sum_{i=1}^{n} |a_i|^p \mu(i) \right)^{1/p} \left(\sum_{i=1}^{n} |b_i|^q \mu(i) \right)^{1/q}.$$

Exercise 1.3. Let X be a probability space, and $p, q \geqslant 1$ be conjugate exponents (i.e., $\frac{1}{p} + \frac{1}{q} = 1$). Show that for every$f \in L_p(X)$, we have

$$\|f\|_p = \sup_{g : \|g\|_q = 1} |\langle f, g \rangle|.$$

Exercise 1.4. Suppose that (X, μ) is a measure space and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, for $p, q, r \geqslant 1$. Show that if $f \in L_p(X)$, $g \in L_q(X)$, and $h \in L_r(X)$, then

$$\left| \int f(x) g(x) h(x) d\mu(x) \right| \leqslant \|f\|_p \|g\|_q \|h\|_r.$$

Exercise 1.5. Suppose that X is a measure space and $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$, for $p, q, r \geqslant 1$. Show that if $f \in L_p(X)$ and $g \in L_q(X)$, then

$$\|fg\|_r \leqslant \|f\|_p \|g\|_q.$$

Exercise 1.6. Let X be a probability space. Let $\|T\|_{p \to q}$ denote the operator norm of $T : L_p(X) \to L_q(X)$. In other words

$$\|T\|_{p \to q} := \sup_{f : \|f\|_p = 1} \|Tf\|_q.$$

Recall that the adjoint of T is an operator T^*such that

$$\langle Tf, g \rangle = \langle f, T^* g \rangle,$$

for all $f, g \in L_2(X)$. Prove that for conjugate exponents $p, q \geqslant 1$, and every linear operator $T : L_2(X) \to L_2(X)$, we have

$$\|T\|_{p \to 2} = \|T^*\|_{2 \to q}.$$

Chapter 2

Fourier analysis of Finite Abelian Groups

This chapter develops the basic Fourier analysis of finite Abelian groups. Recall that the *cyclic group* $\mathbb{Z}_N$ is the Abelian group with elements $\{0, 1, \ldots, N-1\}$, where the group operation is $a + b \coloneqq a + b \pmod N$.

The fundamental theorem of finite Abelian groups states that every finite Abelian group is a direct product of cyclic groups.

Theorem 2.1. *Every finite Abelian group G is isomorphic to the group $\mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$ for some positive integers $N_1, \ldots, N_k$.*

In this course, we will mainly focus on the group $\mathbb{Z}_2^n \coloneqq \mathbb{Z}_2 \times \ldots \times \mathbb{Z}_2$ as it naturally represents the discrete cube $\{0,1\}^n$. This identification of $\{0,1\}^n$ with $\mathbb{Z}_2^n$ enables us to use the Fourier analysis as a powerful tool in the study of Boolean functions $f : \{0,1\}^n \to \{0,1\}$.

2.1 The space of functions on G

Let G be a finite Abelian group G. We endow G with the uniform probability measure, which assigns a probability of $\frac{1}{|G|}$ to each element. We define the inner product of every two functions $f, g : G \to \mathbb{C}$ accordingly as

$$\langle f, g\rangle \coloneqq \mathbb{E}_{x\in G} f(x)\overline{g(x)} = \frac{1}{|G|}\sum_{x\in G} f(x)\overline{g(x)}. \tag{2.1}$$

We denote the linear space of all functions $f : G \to \mathbb{C}$ with the above inner product by $L_2(G)$. Note that $L_2(G)$ is a $|G|$-dimensional vector space, as it is spanned by the set of functions $\mathbf{1}_a : G \to \{0,1\}$ for $a \in G$, where

$$\mathbf{1}_a(x) \coloneqq \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a \end{cases}.$$

Note that the norm defined by the above inner product is

$$\|f\|_2 = \sqrt{\langle f, f\rangle} = \sqrt{\mathbb{E}_{x\in G}|f(x)|^2}.$$

As we proved in Section 1.4, $\|f\|_2$ satisfies the axioms of a norm, and we have the Cauchy-Schwarz inequality:

$$|\langle f, g\rangle| \leqslant \|f\|_2\|g\|_2.$$

For $1 \leqslant p < \infty$, we define

$$\|f\|_p = \left(\mathbb{E}_{x\in G}|f(x)|^p\right)^{1/p},$$

and for $p = \infty$,

$$\|f\|_\infty = \max_{x\in G}|f(x)|.$$

We proved in Section 1.5 if $p \in [1, \infty]$, then $\|\cdot\|_p$ satisfies the axioms of a norm. Moreover, we have a generalization of the Cauchy-Schwarz inequality, called Hölder's inequality. It states that if $p, q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$, then

$$|\langle f, g \rangle| \leqslant \|f\|_p \|g\|_q.$$

2.2 Fourier Analysis

We start with the definition of Fourier characters.

Definition 2.2 (Fourier Character)**.** Let G be a *finite Abelian group*. A function $\chi : G \to \mathbb{C} \setminus \{0\}$ mapping the group to the non-zero complex numbers is called a *character* of G if it satisfies the following two conditions:

- $\chi(0) = 1$, where 0 is the identity of G;
- $\chi(a + b) = \chi(a)\chi(b)$ for all $a, b \in G$.

In other words, χ is a group homomorphism from G to the group $(\mathbb{C} \setminus \{0\}, \times)$.

Note that the constant function 1 is always a character, which is called the *principal character* of G. Let χ be a character of G, and consider an element $a \in G$. Since G is a finite group, every element a is of some finite order n (i.e., $na = 0$ where na refers to adding a to itself n times). Hence $1 = \chi(0) = \chi(|G|a) = \chi(a)^n$ which shows that $\chi(a)$ is an n-th root of unity. Recall that the n-th roots of unity are of the form

$$e^{2\pi i \frac{k}{n}} = \cos\left(\frac{2\pi k}{n}\right) + i \sin\left(\frac{2\pi k}{n}\right) \quad \text{where } k = 0, \ldots, n-1.$$

In particular, every character χ of G satisfies $\chi : G \to \mathbb{T}$ where $\mathbb{T}$ is the unit complex circle.

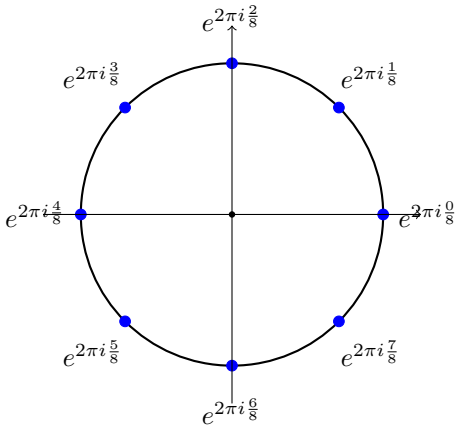

Figure 2.1: The set of 8-th roots of unity. If $a \in G$ is an element with $8a = 0$, then $\chi(a)$ is an 8-th root of unity.

Theorem 2.3 (Pontryagin dual)**.** *The set of the characters of every finite Abelian group G equipped with the point-wise product of complex-valued functions form an Abelian group $\widehat{G}$, which is called the* Pontryagin dual *of G.*

Proof. The principal character 1 is the identity of $\widehat{G}$ as we have $\chi 1 = 1\chi = \chi$ for every $\chi \in \widehat{G}$. Note that if χ and ξ are characters of G, then $\chi\xi$ is also a character. To verify this fact, note $\chi(ab)\xi(ab) = \chi(a)\xi(a)\chi(b)\xi(b)$, and $\chi(0)\xi(0) = 1 \times 1 = 1$. To check the existence of the inverse elements, note that if χ is a character, then $\chi^{-1} := \frac{1}{\chi} = \overline{\chi}$ is also a character as $\overline{\chi(0)} = \overline{1} = 1$ and $\overline{\chi(ab)} = \overline{\chi(a)\chi(b)}$. □

2.2.1 Fourier characters of $\mathbb{Z}_2^n$

First, consider the group $\mathbb{Z}_2 \equiv \{0, 1\}$. Let χ be a character of $\mathbb{Z}_2$. According to the definition of a character, we must have $\chi(0) = 1$. Furthermore, since $1 + 1 = 0$, we have $1 = \chi(1 + 1) = \chi(1)^2$, which shows that $\chi(1) = 1$ or $\chi(1) = -1$. Therefore, $\mathbb{Z}_2$ has only two characters: the principal character $\chi_0 \equiv 1$ and $\chi_1 : x \mapsto (-1)^x$ for $x \in \mathbb{Z}_2$.

Now that we have described the two characters of $\mathbb{Z}_2$, we can easily construct the characters of $\mathbb{Z}_2^n$. Indeed if χ is a character of G and ψ is a character of H, then the map $\chi \times \psi : G \times H \to \mathbb{T}$ defined as $\chi \times \psi(g,h) \coloneqq \chi(g)\psi(h)$ is a character of $G \times H$. Since $\mathbb{Z}_2^n$ is the product of n copies of $\mathbb{Z}_2$, by choosing χ_0 or χ_1, for each copy, we can construct 2^n characters for $\mathbb{Z}_2$. Let us describe these characters.

For every $a = (a_1, \ldots, a_n) \in \mathbb{Z}_2^n$, we construct a corresponding character $\chi_a : \mathbb{Z}_2^n \to \{-1, +1\}$ with

$$\chi_a(x) \coloneqq \prod_{i:a_i=1} (-1)^{x_i} = (-1)^{\sum_{i:a_i=1} x_i}.$$

The principal character is $\chi_0 \equiv 1$ where the 0 in the index refers to $(0, \ldots, 0)$, the identity element of the group. It is easy to verify that these are all the characters of $\mathbb{Z}_2^n$. In the case of $\mathbb{Z}_2^n$, the characters are real-valued (they only take values 1 and -1), but as we shall see below for all other Abelian groups, some characters take non-real values.

Since the coordinates of $a \in \mathbb{Z}_2^n$ are 0 or 1, we will sometimes identify a with the set $S = \{i : a_i = 1\} \subseteq \{1, \ldots, n\}$, and denote the characters as χ_S for $S \subseteq \{1, \ldots, n\}$. This notation is sometimes more intuitive as

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

furthermore, in future chapters, when we take a probabilistic approach to decomposing functions, this notation extends to general product spaces (where there is no group structure). In this notation, $\chi_\emptyset$ corresponds to the principal character.

2.2.2 Fourier characters of $\mathbb{Z}_N$ and $\mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$

Next, consider the group $\mathbb{Z}_N \equiv \{0, 1, \ldots, N-1\}$, where the addition is mod N. Let χ be a character of $\mathbb{Z}_N$. According to the definition of a character, we must have $\chi(0) = 1$. Moreover, as we discussed earlier

$$\chi(1)^N = \chi(1 + \ldots + 1) = \chi(0) = 1,$$

which shows that $\chi(1)$ is an N-th root of unity and therefore, it is of the form

$$\chi(1) = e^{2\pi i \frac{a}{N}},$$

for some $a \in \{0, \ldots, N-1\}$. Note further that for every $x \in \mathbb{Z}_N$, we have $\chi(x) = \chi(1)^x = e^{2\pi i \frac{ax}{N}}$.

We showed that every character of $\mathbb{Z}_N$ must be of the form

$$\chi_a : x \mapsto e^{2\pi i \frac{ax}{N}},$$

for some $a \in \{0, \ldots, N-1\}$, and on the other hand, one can easily verify that each such χ_a is a character.

Finally, now that we have described the characters of $\mathbb{Z}_N$, we can multiply these characters to obtain the characters of any Abelian group $G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$.

For $a = (a_1, \ldots, a_k) \in \{0, \ldots, N_1 - 1\} \times \ldots \times \{0, \ldots, N_k - 1\}$, we define the corresponding character $\chi_a : G \to \mathbb{T}$ as

$$\chi_a(x_1, \ldots, x_k) \coloneqq \prod_{i=1}^k \chi_{a_i}(x_i) = \prod_{i=1}^k e^{2\pi i \frac{a_i x_i}{N_i}} = e^{2\pi i \sum_{i=1}^k \frac{a_i x_i}{N_i}}. \tag{2.2}$$

2.2.3 Self-duality: $G \cong \widehat{G}$

In Theorem 2.3, we showed that the characters form an Abelian group $\widehat{G}$ under the point-wise multiplication. On the other hand, in Eq. (2.2), we gave a full description of the characters of a general finite Abelian group $G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$. Let us try to understand the structure of $\widehat{G}$ using this description of characters.

First, let us consider $G = \mathbb{Z}_N$ for simplicity. Consider $a, b \in \{0, \ldots, N-1\}$ and their corresponding characters $\chi_a(x) = e^{2\pi i \frac{ax}{N}}$ and $\chi_b(x) = e^{2\pi i \frac{bx}{N}}$.

Which χ_c corresponds to the character $\chi_a \chi_b$? Note that

$$\chi_a \chi_b(x) = \chi_a(x)\chi_b(x) = e^{2\pi i \frac{(a+b)x}{N}} = e^{2\pi i \frac{cx}{N}},$$

where $c \in \{0, \ldots, N-1\}$ is the element with $c = (a+b) \bmod N$. Note also that $\chi_0 = 1$. Indeed, $\widehat{\mathbb{Z}_N}$ is isomorphic to $\mathbb{Z}_N$ with the isomprphism $\chi_a \mapsto a$ for $a \in \mathbb{Z}_N \equiv \{0, \ldots, N-1\}$. Note that this argument easily generalizes to $\mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$. We obtain the following theorem.

Theorem 2.4. *For every finite Abelian group G, we have $G \cong \widehat{G}$.*

Remark 2.5. We emphasize that Theorem 2.4 is not necessarily true for infinite Abelian groups. However, the case of infinite Abelian groups is beyond the scope of this course.

In light of Theorem 2.4, it is convenient to index the characters of G with the elements of G and denote the characters of G by χ_a for $a \in G$.

2.2.4 Fourier Transform and Orthogonality of characters

Our next goal will be to prove that the characters form an orthonormal basis with respect to the inner product defined in Eq. (2.1). First, let us prove a simple lemma regarding the sum of a character over all elements in the group.

Lemma 2.6. *Let G be a finite Abelian group, and χ be a non-principal character of G. Then $\sum_{x \in G} \chi(x) = 0$.*

Proof. Suppose to the contrary that $\sum_{x \in G} \chi(x) \neq 0$. Consider an arbitrary $y \in G$, and note

$$\chi(y) \sum_{x \in G} \chi(x) = \sum_{x \in G} \chi(y+x) = \sum_{x \in G} \chi(x),$$

which shows that $\chi(y) = 1$. Since this is true for every $y \in G$, we conclude that χ must be the principal character, which contradicts the assumption of the lemma. □

Now, we can prove the orthogonality of the characters.

Lemma 2.7. *The characters of a finite Abelian group G are orthonormal: For $\chi, \psi \in \widehat{G}$, we have*

$$\langle \chi, \psi \rangle = \begin{cases} 1 & \text{if } \chi = \psi \\ 0 & \text{if } \chi \neq \psi \end{cases}.$$

Proof. Since the range of a character is $\mathbb{T}$, we have

$$\langle \chi, \chi \rangle = \mathbb{E}\left[|\chi(x)|^2\right] = \mathbb{E}[1] = 1.$$

It remains to verify the orthogonality. Let $\chi \neq \psi$ be two different characters. Then $\chi\overline{\psi} = \chi\psi^{-1}$ is a *non-principal* character of G. Hence by Lemma 2.6, we have

$$\langle \chi, \psi \rangle = \mathbb{E}\left[\chi(x)\overline{\psi}(x)\right] = \mathbb{E}\left[\chi\overline{\psi}(x)\right] = 0.$$

□

Since $L_2(G)$ is a $|G|$-dimensional vector space and $|\widehat{G}| = |G|$, the orthonormality of the characters implies that they must form an orthonormal basis for $L_2(G)$. Namely, in addition to being orthonormal, they span the whole space $L_2(G)$.

Theorem 2.8. *If G is a finite Abelian group, then the characters of G form an orthonormal basis for $L_2(G)$.*

Since the characters form an orthonormal basis for $L_2(G)$, every function $f : G \to \mathbb{C}$ has a unique expression as a linear combination of the characters

$$f = \sum_{a \in G} \widehat{f}(a)\chi_a.$$

The corresponding coefficients $\widehat{f}(a) \in \mathbb{C}$ are the *Fourier coefficients* of f.

Definition 2.9. The Fourier transform of a function $f : G \to \mathbb{C}$ is the function $\widehat{f} : \widehat{G} \to \mathbb{C}$ defined as

$$\widehat{f}(\chi) = \langle f, \chi \rangle = \mathbb{E} f(x)\overline{\chi(x)}.$$

For $a \in G$, we will often use the notation $\widehat{f}(a)$ to denote $\widehat{f}(\chi_a)$. Similarly, in the case of $\mathbb{Z}_2^n$, for $S \subseteq [n] \coloneqq \{1, \ldots, n\}$, we use the notation $\widehat{f}(S)$ to denote $\widehat{f}(\chi_S)$.

The formula

$$f = \sum_{a \in G} \widehat{f}(a)\chi_a,$$

that uniquely expresses f as a linear combination of characters is called the *Fourier inversion formula* as it shows how the functions f can be reconstructed from its Fourier transform.

Example 2.10. Let $f : \mathbb{Z}_2^n \to \{0, 1\}$ be the parity function $f : x \mapsto \sum_{i=1}^n x_i \pmod 2$. Then

$$\widehat{f}(\emptyset) = \mathbb{E} f(x)\chi_\emptyset = \mathbb{E} f(x) = \frac{1}{2}.$$

We also have

$$\widehat{f}([n]) = \mathbb{E} f(x)(-1)^{\sum_{j=1}^n x_j} = -\frac{1}{2},$$

since $f(x) = 1$ if and only if $\sum_{j=1}^n x_j = 1 \pmod 2$.

Next consider $\emptyset \subsetneq S \subsetneq [n]$, and let Consider $j_0 \in S$ and $j_1 \notin S$. We have

$$\widehat{f}(S) = \mathbb{E} f(x)\chi_S(x) = \frac{1}{2}\mathbb{E}\left[f(x)\chi_a(x) + f(x + e_{j_0} + e_{j_1})\chi_S(x + e_{j_0} + e_{j_1})\right],$$

where e_j denotes the vector in $\mathbb{Z}_2^n$ which has 1 at its jth coordinate and 0 everywhere else. Note that $f(x) = f(x + e_{j_0} + e_{j_1})$ and furthermore $\chi_S(x) = -\chi_S(x + e_{j_0} + e_{j_1})$. We conclude that $\widehat{f}(S) = 0$ for every $\emptyset \subsetneq S \subsetneq [n]$. The Fourier expansion of f is

$$f(x) = \frac{1}{2} - \frac{1}{2}\chi_{[n]}(x).$$

2.2.5 Basic properties of the Fourier Coefficients

The Fourier transform is a linear operator: $\widehat{\lambda f + g} = \lambda\widehat{f} + \widehat{g}$, and we have the following easy observation.

Lemma 2.11. *The Fourier transform satisfies*

$$\|\widehat{f}\|_\infty \coloneqq \max_a |\widehat{f}(a)| \leqslant \|f\|_1.$$

Proof. For every $a \in G$, we have

$$|\widehat{f}(a)| = \left|\mathbb{E} f(x)\overline{\chi_a(x)}\right| \leqslant \mathbb{E}|f(x)||\overline{\chi_a(x)}| = \mathbb{E}|f(x)| = \|f\|_1.$$

□

The *principal Fourier coefficient* $\widehat{f}(0)$ is of particular importance as

$$\widehat{f}(0) = \mathbb{E}[f(x)].$$

So if $\mathbf{1}_A$ is the indicator function of a subset $A \subseteq G$, then $\widehat{\mathbf{1}_A}(0) = \frac{|A|}{|G|}$ is the density of A.

Next, we prove Parseval's identity, a simple but extremely useful fact in Fourier's analysis.

Theorem 2.12 (Parseval)**.** *For every $f \in L_2(G)$,*

$$\|f\|_2^2 = \sum_{a \in G} |\widehat{f}(a)|^2.$$

Proof. We have

$$\|f\|_2^2 = \langle f, f\rangle = \left\langle \sum_{a\in G} \widehat{f}(a)\chi_a, \sum_{b\in G} \widehat{f}(b)\chi_b \right\rangle = \sum_{a,b\in G} \widehat{f}(a)\overline{\widehat{f}(b)}\langle \chi_a, \chi_b\rangle.$$

The identity now follows from the orthonormality of characters:

$$\langle \chi_a, \chi_b\rangle = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}.$$

□

The proof of the Parseval identity, when applied to two different functions $f, g \in L_2(G)$, implies the *Plancherel theorem*:

$$\langle f, g\rangle = \sum_{a\in G} \widehat{f}(a)\overline{\widehat{g}(a)}.$$

Let $\mathbf{1}_A$ denote the indicator function of $A \subseteq G$. In this case, Parseval's identity shows

$$\sum_{a\in G} |\widehat{\mathbf{1}_A}(a)|^2 = \mathbb{E}|\mathbf{1}_A(x)|^2 = \mathbb{E}|\mathbf{1}_A(x)| = \frac{|A|}{|G|}.$$

Recall also that

$$\widehat{\mathbf{1}_A(0)} = \frac{|A|}{|G|}.$$

2.2.6 Physical Space versus Fourier Space

Consider the space $L_2(G)$ of the functions $f : G \to \mathbb{C}$. The most natural linear basis for $L_2(G)$ is the set of functions $\mathbf{1}_a : G \to \{0, 1\}$ for $a \in G$, where

$$\mathbf{1}_a(x) \coloneqq \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a \end{cases}.$$

Note that the unique expansion of f in this linear basis is

$$f = \sum_{a\in G} f(a)\mathbf{1}_a.$$

This is the expansion of f in the "*physical space*" where we expanded f in terms of the indicator functions of the elements in G, and the coefficients are simply the values of f on those elements.

In contrast, in the *Fourier space*, we expand f as a linear combination of the characters of the group, and the coefficients are the Fourier coefficients:

$$f = \sum_{a\in G} \widehat{f}(a)\chi_a.$$

The Fourier transform is simply a change of basis from the indicator functions $\mathbf{1}_a$ to Fourier characters χ_a.

Not that with our normalization choice in defining the inner product, the characters are orthonormal, while unfortunately, the indicator functions 1_a need to be normalized to $|\sqrt{G}|\mathbf{1}_a$ to become orthonormal.

2.3 Convolution

In this section, we introduce a key notion in functional analysis called convolution.

Definition 2.13 (Convolution)**.** Let G be a finite Abelian group. Given two functions $f, g : G \to \mathbb{C}$, the convolution $f * g : G \to \mathbb{C}$ is defined as

$$f * g(x) = \mathbb{E}_{\mathbf{y}\in G}[f(x - \mathbf{y})g(\mathbf{y})].$$

Note that $f * g(x)$ is the average of $f(a)g(b)$ over all pairs a, b with $a + b = x$.

Remark 2.14. Consider a set $A \subseteq G$. Then $f * \mathbf{1}_A(x)$ is the average of f over the set $x - A \coloneqq \{x - y : y \in A\}$. For example if A is the Hamming ball[1] of radius r around 0 in $\mathbb{Z}_2^n$, then $f * \mathbf{1}_A(x)$ is the average of f over the Hamming ball of radius r around x.

Next, we list some basic properties of the convolution.

Lemma 2.15. *Consider three functions $f, g, h : G \to \mathbb{C}$.*

(a) *We have*
$$f * g = g * f.$$

(b) *We have*
$$(f * g) * h = f * (g * h).$$

(c) *We have*
$$f * (\lambda h + g) = \lambda f * h + f * g.$$

(d) *We have*
$$\|f * g\|_\infty \leqslant \|f\|_1 \|g\|_\infty.$$

(e) *More generally, if p and q are conjugate exponents (i.e., they satisfy $\frac{1}{p} + \frac{1}{q} = 1$), then*
$$\|f * g\|_\infty \leqslant \|f\|_p \|g\|_q.$$

(f) *We have*
$$\|f * g\|_1 \leqslant \|f\|_1 \|g\|_1.$$

Proof. **(a)** For every $x \in G$, we have
$$f * g(x) = \mathbb{E}_{\mathbf{y}}[f(x - \mathbf{y})g(\mathbf{y})] = \mathbb{E}_{\mathbf{y}}[f(x - \mathbf{y})g(x - (x - \mathbf{y}))] = \mathbb{E}_{\mathbf{z}}[f(\mathbf{z})g(x - \mathbf{z})] = g * f(x).$$

(b) By Part **(a)**,
$$\begin{aligned}(f * g) * h(x) &= (g * f) * h(x) = \mathbb{E}_{\mathbf{z}}\mathbb{E}_{\mathbf{y}}[g(x - \mathbf{z} - \mathbf{y})f(\mathbf{y})]h(\mathbf{z}) = \\ &= \mathbb{E}_{\mathbf{y},\mathbf{z}} g(x - \mathbf{z} - \mathbf{y})f(\mathbf{y})h(\mathbf{z}) = (h * g) * f(x) = f * (g * h)(x).\end{aligned}$$

(c) is trivial.
(d) is a special case of **(e)**.
(e) For every $x \in G$, by Hölder's inequality, we have
$$|f * g(x)| \leqslant \mathbb{E}_{\mathbf{y}\in G}|f(x - \mathbf{y})||g(\mathbf{y})| \leqslant \left(\mathbb{E}|f(x - \mathbf{y})|^p\right)^{1/p} \left(\mathbb{E}|g(\mathbf{y})|^q\right)^{1/q} = \left(\mathbb{E}|f(\mathbf{y})|^p\right)^{1/p} \|g\|_q = \|f\|_p \|g\|_q.$$

(f) We have
$$\|f * g\|_1 = \mathbb{E}_{\mathbf{x}}\,|f * g(\mathbf{x})| \leqslant \mathbb{E}_{\mathbf{x},\mathbf{y}}|f(\mathbf{x} - \mathbf{y})||g(\mathbf{y})| = \mathbb{E}_{\mathbf{z},\mathbf{y}}|f(\mathbf{z})||g(\mathbf{y})| = \mathbb{E}_{\mathbf{z}}|f(\mathbf{z})|\mathbb{E}_{\mathbf{y}}|g(\mathbf{y})| = \|f\|_1 \|g\|_1.$$

□

The following lemma states that the Fourier transform of $f * g$ is the point-wise product of the individual Fourier transforms $\widehat{f}$ and $\widehat{g}$.

Lemma 2.16. *For every $f, g : G \to \mathbb{C}$, we have*
$$\widehat{f * g} = \widehat{f} \cdot \widehat{g}.$$

[1] The Hamming ball of radius r around 0 is defined as $\left\{x \in \mathbb{Z}_2^n : \sum_{i=1}^n x_i \leqslant r\right\} \subseteq \mathbb{Z}_2^n$.

Proof. We have

$$\begin{aligned}\widehat{f*g}(a) &= \mathbb{E}_{\mathbf{x}} f*g(\mathbf{x})\overline{\chi_a(\mathbf{x})} = \mathbb{E}_{\mathbf{x}}\left(\mathbb{E}_{\mathbf{y}} f(x-\mathbf{y})g(\mathbf{y})\right)\overline{\chi_a(\mathbf{x})} = \mathbb{E}_{\mathbf{x},\mathbf{y}} f(\mathbf{x}-\mathbf{y})g(\mathbf{y})\overline{\chi_a(\mathbf{x}-\mathbf{y})\chi_a(\mathbf{y})} \\ &= \mathbb{E}_{\mathbf{z},\mathbf{y}} f(\mathbf{z})g(\mathbf{y})\overline{\chi_a(\mathbf{z})\chi_a(\mathbf{y})} = \mathbb{E}_{\mathbf{z}} f(\mathbf{z})\overline{\chi_a(\mathbf{z})}\mathbb{E}_{\mathbf{y}} g(\mathbf{y})\overline{\chi_a(\mathbf{y})} = \widehat{f}(a)\cdot\widehat{g}(a).\end{aligned}$$

□

Note that Lemma 2.16 in particular shows that

$$\mathbb{E}_{\mathbf{x}} f*g(\mathbf{x}) = \mathbb{E}[f]\mathbb{E}[g].$$

We also have the dual version of Lemma 2.16,

$$\widehat{fg}(x) = \sum_{y\in G}\widehat{f}(x-y)\widehat{g}(y). \tag{2.3}$$

2.4 Exercises

Exercise 2.1. Suppose that for $f:\mathbb{Z}_2^n\to\{0,1\}$ satisfies $\widehat{f}(S)=0$ for all $|S|\geqslant 2$ (that is $\deg_{\mathcal{F}}(f)\leqslant 1$). Show that either $f\equiv 0$, $f\equiv 1$, $f(x)=x_i$, or $f(x)=1-x_i$ for some $i\in[n]$.

Exercise 2.2. Let G be a finite Abelian group and let $f:G\to\{0,1\}$. Prove that

$$\|\widehat{f}\|_1 \coloneqq \sum_{a\in G}|\widehat{f}(a)| \leqslant \sqrt{|G|}.$$

Exercise 2.3. Let G be a finite Abelian group and let $f:G\to\{0,1\}$. Prove that

$$\|\widehat{f}\|_\infty^4 \leqslant \sum_{a\in G}|\widehat{f}(a)|^4 \leqslant \|\widehat{f}\|_\infty^2.$$

Chapter 3

An Application: Linearity Testing

Blum, Luby, and Rubinfeld [BLR90] made a beautiful observation that given a function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$, it is possible to inquire the value of f on a few random points to probabilistically distinguish between the case that f is a linear function and the case that f has to be modified on at least $\varepsilon > 0$ fraction of points to become a linear function. This result is known as the BLR linearity testing in theoretical computer science.

Inspired by the BLR test, Rubinfeld and Sudan [RS93] defined the concept of *property testing*, which is now an active area of research in theoretical computer science. Roughly speaking, to test a function for a property means to examine the value of the function on a few random points and accordingly (probabilistically) distinguish between the case that the function has the property and the case that it is not too close to any function with that property. Interestingly, and to some extent surprisingly, these tests exist for various basic properties. The first substantial investigation of property testing occurred in Goldreich, Goldwasser, and Ron [GGR98], who showed that several natural combinatorial properties are testable. Since then, significant research has been conducted on classifying testable properties in combinatorial and algebraic settings.

The BLR test is a fundamental result in computer science, and its significance goes beyond property testing. It is a crucial component of many results in coding theory, the study of pseudo-random generators, and PCP (probabilistically checkable proofs) theorems. Its proof is surprisingly elementary. It only relies on the orthogonality of the Fourier characters and the Parseval identity.

3.1 Linearity testing

In this section, we will state and analyze the BLR linearity test. We start by formally defining a linear function.

Definition 3.1. A function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ is called *linear* if $f(x + y) = f(x) + f(y)$ for all $x, y \in \mathbb{Z}_2^n$.

Note that every linear function is of the form $\ell_a : x \mapsto a_1x_1 + \ldots + a_nx_n \pmod 2$ where $a = (a_1, \ldots, a_n) \in \mathbb{Z}_2^n$.

We are interested in the following problem: given access to the truth table of a function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$, how quickly can we verify whether f is linear? Note that any method for determining exact linearity would require probing the function at every point as f could be almost linear, except corrupted on a single input. Therefore, we relax this requirement slightly: can we quickly determine whether the function is approximately linear?

The BLR test says that it is possible to query the value of a function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ on a few points, and with some non-negligible positive probability distinguish correctly between the following two cases

1. f is linear.
2. f is ε-*far* from every linear function: for every linear $\ell : \mathbb{Z}_2^n \to \mathbb{Z}_2$,

$$\mathbf{Pr}[f(x) \neq \ell(x)] \geqslant \varepsilon.$$

More precisely, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that the following holds. Given a function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$, we can query the value of f on only 3 points and output accept or reject such that the following holds.

1. Always accept f if it is linear;

2. Reject f with probability at least $\delta > 0$ if f is ε-far from every linear function.

In this description, the error is *one-sided* as we always accept f if it satisfies the property. We can easily boost the probability of the success of such a test by applying it several times. More precisely, we can run the test N independent times and accept f only if all the N executions accept f. In this case, if f is ε-far from every linear function, then the test will reject it with probability at least $1 - (1-\delta)^N$, which can be made very close to 1, by setting, for example, $N = 10^3\delta^{-1}$. Note that such an error reduction makes $3N$ queries to f. Now let us finally state the BLR test:

Blum, Luby, and Rubinfeld's [BLR90] linearity test:

- Given a function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$.
- Pick two random points $x, y \in \mathbb{Z}_2^n$.
- If $f(x) + f(y) \neq f(x+y)$, then Reject, otherwise Accept.

Note that as we claimed above, if f is linear, then the BLR test always succeeds; that is, it never rejects a linear function. The main part of the analysis lies in proving that if f is ε-far from every linear function, then the test rejects f with probability at least $\delta > 0$.

3.1.1 Analysis of the BLR test

Theorem 3.2 (Blum, Luby, and Rubinfeld's [BLR90])**.** *Consider $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$.*

- *If f is linear, then the BLR test accepts with probability 1;*
- *If f is ε-far from every linear function, then the BLR test rejects with probability at least $\delta \coloneqq \varepsilon > 0$.*

Proof. First, note that if f is a linear function, the BLR test always succeeds; that is, it never rejects a linear function. We need to prove that if f is ε-far from every linear function, then f is rejected with probability at least $\delta > 0$ for some δ depending only on ε.

To analyze the test, it is more convenient to work with $g : \mathbb{Z}_2^n \to \{-1, 1\}$ with $g(x) \coloneqq (-1)^{f(x)}$ instead of f. Note that for a linear function $\ell_a : x \mapsto a_1x_1 + \ldots + a_nx_n$, we have $(-1)^{\ell_a(x)} = \chi_a(x)$, and thus

$$\mathbf{Pr}[f(x) \neq \ell_a(x)] = \mathbf{Pr}[g(x) \neq \chi_a(x)] = \mathbb{E}\left[\frac{1 - g(x)\chi_a(x)}{2}\right] = \frac{1}{2} - \frac{1}{2}\mathbb{E}[g(x)\chi_a(x)] = \frac{1}{2} - \frac{1}{2}\widehat{g}(a).$$

So if f is ε-far from every linear function, then

$$\varepsilon \leqslant \frac{1}{2} - \frac{1}{2}\max_a \widehat{g}(a),$$

or equivalently

$$\max_a \widehat{g}(a) \leqslant 1 - 2\varepsilon. \tag{3.1}$$

Next, let us analyze the probability that the BLR test does not reject f. By the definition of g, we have

$$\mathop{\mathbf{Pr}}_{x,y}[f(x) + f(y) = f(x+y)] = \mathop{\mathbf{Pr}}_{x,y}[g(x)g(y)g(x+y) = 1] = \frac{1}{2} + \frac{1}{2}\mathbb{E}_{x,y}[g(x)g(y)g(x+y)].$$

Replacing g with its Fourier expansion, we get

$$\begin{aligned}\mathbb{E}[g(x)g(y)g(x+y)] &= \mathbb{E}\left[\sum_{a,b,c} \widehat{g}(a)\widehat{g}(b)\widehat{g}(c)\chi_a(x)\chi_b(y)\chi_c(x+y)\right] \\ &= \sum_{a,b,c} \widehat{g}(a)\widehat{g}(b)\widehat{g}(c)\mathbb{E}_x[\chi_{a+c}(x)]\mathbb{E}_y[\chi_{a+b}(y)].\end{aligned}$$

Note that $a + c = 0$ if and only if $a = c$, and thus by Lemma 2.6,

$$\mathbb{E}[\chi_{a+c}(x)] = \begin{cases} 1 & \text{if } a = c \\ 1 & \text{if } a \neq c \end{cases}.$$

Similarly

$$\mathbb{E}[\chi_{b+c}(x)] = \begin{cases} 1 & \text{if } b = c \\ 1 & \text{if } b \neq c \end{cases}.$$

Therefore, the above expression for $\mathbb{E}[g(x)g(y)g(x+y)]$ simplifies to

$$\mathbb{E}[g(x)g(y)g(x+y)] = \sum_a \widehat{g}(a)^3.$$

Consequently,

$$\mathop{\mathbf{Pr}}_{x,y}[f(x)f(y) = f(x+y)] = \frac{1}{2} + \frac{1}{2}\sum_a \widehat{g}(a)^3 \leqslant \frac{1}{2} + \frac{1}{2}\left(\max_a \widehat{g}(a)\right)\sum_a \widehat{g}(a)^2. \tag{3.2}$$

Note that by Parseval identity and the fact that $g : \mathbb{Z}_2^n \to \{-1, 1\}$, we have

$$\sum_{a \in G} \widehat{g}(a)^2 = \|g\|_2^2 = \mathbb{E}g(x)^2 = 1.$$

Therefore, Eq. (3.2) shows

$$\mathop{\mathbf{Pr}}_{x,y}[f(x) + f(y) = f(x+y)] \leqslant \frac{1}{2} + \frac{1}{2}\max_a \widehat{g}(a). \tag{3.3}$$

To conclude the proof note that by (3.1) and (3.3), if f is ε-far from every character, then

$$\mathop{\mathbf{Pr}}_{x,y}[f(x) + f(y) = f(x+y)] \leqslant 1 - \varepsilon.$$

In other words, the test rejects with probability at least $\varepsilon > 0$. □

3.2 Linear functions as error-correcting codes

An *error-correcting code* is an injective map $\mathcal{C} : \{0,1\}^n \to \{0,1\}^m$ that maps a binary message of length n to a longer code-word of length m. Since $\mathcal{C}$ is injective, we can uniquely recover the original message a if we receive the intact code-word $\mathcal{C}(a)$.

However, what happens if we receive a slightly corrupted version $y \in \{0,1\}^m$ instead of the exact code-word $\mathcal{C}(a)$? The code can still correct the errors, provided the following conditions are met. If the minimum Hamming distance between the code-words $\{\mathcal{C}(a)\}_{a \in \{0,1\}^n}$ is at least $2k + 1$ and the number of corrupted bits in y is at most k, then there is a unique codeword $\mathcal{C}(a)$ that is the closest to y in Hamming distance. Thus, we can correctly recover $\mathcal{C}(a)$, and consequently a, even when the received message is corrupted in up to k bits.

A key example of an error-correcting code with strong error-correcting capabilities is the *Hadamard code*. The Hadamard code $\mathcal{H} : \{0,1\}^n \to \{0,1\}^{2^n}$ maps every element $a \in \{0,1\}^n$ to

$$\mathcal{H}(a) \coloneqq (\ell_a(x))_{x \in \{0,1\}^n} \in \{0,1\}^{2^n}.$$

In other words, we map a to the truth table of the linear function ℓ_a.

A simple observation demonstrates that the Hamming distance between any two distinct codewords in the Hadamard code is 2^{n-1}, which implies that error correction is possible as long as fewer than $1/4$ of the bits are corrupted.

Claim 3.3. *For distinct $a, b \in \mathbb{Z}_2^n$, there are exactly 2^{n-1} elements $x \in \mathbb{Z}_2^n$ with $\ell_a(x) = \ell_b(x)$.*

Proof. Note that $\ell_a(x) = \ell_b(x)$ is equivalent to $\ell_c(x) = 0$ where $c = a + b \in \mathbb{Z}_2^n$. Since $c \neq (0, \ldots, 0)$, the set of solutions to $\ell_c(x) = 0$ is a subspace of codimension 1 in $\mathbb{Z}_2^n$. The size of such a subspace is 2^{n-1}, completing the proof. □

A remarkable property of the Hadamard code is its *local decodability*. Local decodability means that we can (probabilistically) recover any individual bit of $\mathcal{C}(a)$ by querying only a small number of positions in the corrupted copy y. This holds as long as the fraction of corrupted bits in y is small.

Proposition 3.4. *Let $a \in \mathbb{Z}_2^n$ and let $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ satisfy*

$$\mathop{\mathbf{Pr}}_{x}[f(x) \neq \ell_a(x)] \leqslant \delta.$$

Then, given any $x \in \mathbb{Z}_2^n$, the probability that we can recover $\ell_a(x)$ by querying only two positions y and $x + y$ in f is

$$\mathop{\mathbf{Pr}}_{y}[f(y) + f(x + y) = \ell_a(x)] \geqslant 1 - 2\delta.$$

Proof. We have

$$\mathbf{Pr}[f(y) + f(x + y) \neq \ell_a(x)] \leqslant \mathop{\mathbf{Pr}}_{y}[f(y) \neq \ell_a(y)] + \mathop{\mathbf{Pr}}_{y}[f(x + y) \neq \ell_a(x + y)] = 2 \mathop{\mathbf{Pr}}_{y}[f(y) \neq \ell_a(y)] \leqslant 2\delta.$$

□

3.3 Exercises

Exercise 3.1. Note that every function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ of polynomial degree at most 1 satisfies $f(x) + f(x + y + z) = f(x + y) + f(x + z)$. Use this property to design a test with one-sided error for the property of being of degree (over $\mathbb{Z}_2$) at most 1. Prove that the test works correctly.

Chapter 4

An Application: Roth's theorem

Our next application of Fourier analysis concerns a problem in number theory with a rich history.

What is the largest possible size of a set $A \subseteq \{1, \dots, N\}$ without nontrivial 3-term arithmetic progressions?

In 1927, van der Waerden proved that for any given positive integers r and k, for any sufficiently large N, every colouring of the integers $\{1, \dots, N\}$ using r colours will result in a monochromatic arithmetic progression of length k. Note that in any such colouring, at least one colour class is of size at least N/r. Erdős and Turán [ET36] conjectured that this size constraint was the key reason behind the existence of monochromatic progressions in van der Waerden's theorem. They proposed a strengthening of van der Waerden's theorem that for any $k \in \mathbb{N}$ any subset $A \subseteq 1, \dots, N$ without k-term progressions must be of size $o(N)$.

In 1953, Roth [Rot53] used Fourier analysis to confirm Erdös and Turán's conjecture for 3-term arithmetic progressions, showing that any set $A \subseteq \{1, \dots, N\}$ without nontrivial 3-term progressions must have size $O\left(\frac{N}{\log \log N}\right)$. This result, now known as Roth's theorem, became a cornerstone in additive number theory and spurred decades of further research. Determining the optimal bound in Roth's theorem became one of the central problems in additive number theory.

Heath-Brown[HB87] and Szemerédi[Sze90], and later Bourgain[Bou99a] refined Roth's argument to attain the bound $O(\frac{N}{\log^c N})$ for $c = \frac{1}{2} - \varepsilon$. Heath-Brown [HB87], Szemerédi [Sze90], and Bourgain [Bou99a] refined Roth's methods, improving the bound to $O\left(\frac{N}{\log^c N}\right)$ for $c = \frac{1}{2} - \varepsilon$. Further progress by Sanders [San11a], followed by Bloom and Sisask [BS21b], led to an improved bound of $O\left(\frac{N}{\log^{1+\varepsilon} N}\right)$ for some small $\varepsilon > 0$. In a remarkable breakthrough, recently Kelley and Meka [KM23] improved the bound to $N2^{-\Omega(\log^{1/12} N)}$, which was subsequently refined to $N2^{-\Omega(\log^{1/9} N)}$ by Bloom and Sisask [BS23]. On the other hand, Behrand's classical construction [Beh46] shows that there are sets of size $N2^{-O(\log^{1/2} N)}$ that are free of non-trivial 3-progressions.

Regarding the general case of Erdös and Turán's conjecture, in 1975, Szemerédi [Sze75], using a completely new approach, proved the full conjecture and showed that for any fixed k, a set $A \subseteq \{1, \dots, N\}$ without k-term progressions must be of size $o(N)$. However, in contrast to Roth's Fourier-analytic proof, all the various known proofs of Szemerédi's theorem give much weaker bounds. Indeed, it was considered a breakthrough when Gowers [Gow01] proved an upper bound of $O\left(\frac{N}{(\log \log N)^{2^{-2^{k+9}}}}\right)$ on the size of sets of integers without k-term arithmetic progressions. Very recently, Leng, Sah, and Sawhney posted a preprint [LSS24] improving Gowers' bound to an impressive bound of $O\left(\frac{N}{2^{(\log \log N)^{c_k}}}\right)$, where $c_k > 0$ is a constant depending on k.

Finite field model: The study of many problems in additive combinatorics, such as Szemerédi's theorem on arithmetic progressions, is often made easier by first studying the problem in $\mathbb{Z}_p^n$ for some fixed small prime p. This setting is most relevant to applications in combinatorics and theoretical computer science, and it also serves as an elegant model for tackling additive problems concerning integers.

In [Mes95], Meshulam carried out Roth's argument in the case of $\mathbb{Z}_p^n$ for odd fixed prime p, and the asymptotics is as n grows to infinity. Since the particular choice of p is unimportant, we will assume $p = 3$. The rich subgroup

structure of $\mathbb{Z}_3^n$ simplifies some of the nuances in Roth's argument. Moreover this simplification leads to the stronger bound of $O(\frac{N}{\log N})$, where $N \coloneqq |\mathbb{Z}_3^n| = 3^n$. In this chapter, we will present Meshulam's proof.

4.1 Roth's theorem in $\mathbb{Z}_3^n$

Throughout this section, we denote $G \coloneqq \mathbb{Z}_3^n$ and $N \coloneqq |\mathbb{Z}_3^n| = 3^n$.

The density of 3-term progressions in a subset $A \subseteq G$ is captured by

$$t_{3\mathrm{AP}}(A) \coloneqq \mathbb{E}_{\mathbf{x},\mathbf{y}\in G}\left[A(\mathbf{x})A\left(\mathbf{x}+\mathbf{y}\right)A(\mathbf{x}+2\mathbf{y})\right], \tag{4.1}$$

where we identified A with its indicator function.

A set $A \subseteq G$ is a *cap set* if it is free of nontrivial 3-term progressions. More precisely, there are no $x, y \in G$ such that $y \neq 0$ and $x, x+y, x+2y \in A$.

Theorem 4.1. *For sufficiently large n, every cap set $A \subseteq \mathbb{Z}_3^n$ satisfies $\frac{|A|}{3^n} \leqslant \frac{16}{n}$.*

Proof. Let $\alpha \coloneqq \frac{|A|}{3^n} \leqslant \frac{16}{n}$ denote the density of A, and $N \coloneqq |G| = 3^n$ the size of the group.

Replacing A with its Fourier expansion $A(x) = \sum_{a\in G}\widehat{A}(a)\chi_a(x)$ in Eq. (4.1) yields

$$\begin{aligned}
t_{3\mathrm{AP}}(A) &= \mathbb{E}_{\mathbf{x},\mathbf{y}\in G}\left(\sum_{a\in G}\widehat{A}(a)\chi_a(\mathbf{x})\right)\left(\sum_{b\in G}\widehat{A}(b)\chi_b(\mathbf{x}+\mathbf{y})\right)\left(\sum_{c\in G}\widehat{A}(c)\chi_c(\mathbf{x}+2\mathbf{y})\right)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}\in G}\sum_{a,b,c\in G}\widehat{A}(a)\widehat{A}(b)\widehat{A}(c)\chi_a(\mathbf{x})\chi_b(\mathbf{x}+\mathbf{y})\chi_c(\mathbf{x}+2\mathbf{y})\\
&= \sum_{a,b,c\in G}\widehat{A}(a)\widehat{A}(b)\widehat{A}(c)\mathbb{E}_{\mathbf{x}\in G}\left[\chi_{a+b+c}(\mathbf{x})\right]\mathbb{E}_{\mathbf{y}\in G}\left[\chi_{b+2c}(\mathbf{y})\right]
\end{aligned}$$

Note that

$$\mathbb{E}_{\mathbf{x}\in G}\left[\chi_{a+b+c}(\mathbf{x})\right]\mathbb{E}_{\mathbf{y}\in G}\left[\chi_{b+2c}(\mathbf{y})\right] = \begin{cases}1 & \text{if } a+b+c=0 \text{ and } b+2c=0\\ 0 & \text{otherwise}\end{cases}.$$

Note that $a+b+c=0$ and $b+2c=0$ imply $a=c$ and $b=-2c$. Therefore,

$$t_{3\mathrm{AP}}(A) = \sum_{a\in G}\widehat{A}(a)^2\widehat{A}(-2a). \tag{4.2}$$

Since A is a cap set, it only contains trivial 3-term progressions, and therefore,

$$t_{3\mathrm{AP}}(A) = \mathop{\mathbf{Pr}}_{\mathbf{y}\in G}[\mathbf{y}=0]\mathop{\mathbf{Pr}}_{\mathbf{x}\in G}[\mathbf{x}\in A] = \frac{\alpha}{N}, \tag{4.3}$$

which is tiny, assuming G is a large group. On the other hand, the sum on the right-hand side of Eq. (4.2) contains at least one large term: $\widehat{A}(0)^3 = \alpha^3$. We will show that there must be at least one other large Fourier coefficient to cancel $\widehat{A}(0)$'s contribution.

We have

$$t_{3\mathrm{AP}}(A) = \sum_{a\in G}\widehat{A}(a)^2\widehat{A}(-2a) = \alpha^3 + \sum_{a\neq 0}\widehat{A}(a)^2\widehat{A}(-2a) \geqslant \alpha^3 - \left(\max_a|\widehat{A}(-2a)|\right)\sum_a|\widehat{A}(a)|^2.$$

By Parseval, we have $\sum_a|\widehat{A}(a)|^2 = \mathbb{E}_{\mathbf{x}}|A(\mathbf{x})|^2 = \alpha$, and therefore,

$$t_{3\mathrm{AP}}(A) \geqslant \alpha^3 - \alpha\max_{a\neq 0}|\widehat{A}(a)|.$$

Recalling that A is a cap set and satisfies $t_{3\mathrm{AP}}(A) \leqslant \frac{\alpha}{N} \leqslant \frac{\alpha^2}{2}$ as shown in Eq. (4.3), we have

$$\max_{a\neq 0}|\widehat{A}(a)| \geqslant \frac{\alpha^2}{2}.$$

Density increment: Let $a \neq 0$ satisfy $|\widehat{A}(a)| \geqslant \frac{\alpha^2}{2}$. We will use this assumption to show that A is significantly denser in some large affine subspace of $\mathbb{Z}_3^n$.

For $\ell \in \{0,1,2\}$, let

$$V_\ell = \{x \in \mathbb{Z}_3^n \ : \ a \cdot x = \ell \bmod 3\}.$$

The set V_0 is an $(n-1)$-dimensional linear subspace of $\mathbb{Z}_3^n$, and V_1 and V_2 are its cosets. In particular, all V_0, V_1, V_2 are affine subspaces of dimension $n-1$, and therefore, have the same linear structure as $\mathbb{Z}_3^{n-1}$. We will show that the density of A in at least one is significantly higher than α.

Denoting $\omega = e^{2\pi i/3}$, by the definition of $\chi_a(x)$, we have

$$\widehat{A}(a) = \mathbb{E}_{\mathbf{x}\in G} A(\mathbf{x})\omega^{a\cdot x} = \frac{|A\cap V_0|}{3^n} + \frac{|A\cap V_1|}{3^n}\omega + \frac{|A\cap V_1|}{3^n}\omega^2.$$

Let $\mu_\ell \coloneqq |A\cap V_\ell|/3^n$. Since $|\widehat{A}(a)| \geqslant \frac{\alpha^2}{2}$, we have

$$\frac{\alpha^2}{2} \leqslant \mu_0 + \mu_1\omega + \mu_2\omega^2.$$

On the other hand, since

$$\mu_0 + \mu_1 + \mu_2 = \alpha \text{ and } 1 + \omega + \omega^2 = 0,$$

it easily follows that, for some $\ell \in \{0,1,2\}$, we must have

$$\mu_\ell \geqslant \frac{\alpha}{3} + \frac{\alpha^2}{12}.$$

We showed that for some $\ell \in \{0,1,2\}$, we have

$$\alpha + \frac{\alpha^2}{4} \leqslant \frac{|A\cap V_\ell|}{3^{n-1}} = \frac{|A\cap V_\ell|}{|V_\ell|}.$$

Putting things together: As mentioned earlier, V_ℓ is an affine subspace of dimension $(n-1)$, and therefore it has the same linear structure as $\mathbb{Z}_3^{n-1}$. Note also that since A is a cap set, $A\cap V_\ell$ is free of 3-progressions. Therefore, we have shown the existence of a cap set in $\mathbb{Z}_3^{n-1}$ with density $\alpha + \frac{\alpha^2}{4}$. We can repeat this process $c = \frac{4}{\alpha} \leqslant \frac{n}{4}$ many times, with each repetition increasing the density by at least $\frac{\alpha^2}{4}$, to arrive at a cap set in $\mathbb{Z}_3^{n-c}$ with density at least $\alpha + c\frac{\alpha^2}{4} \geqslant 2\alpha$.

We showed that $\frac{8}{\alpha}$ repetition of the above process doubles the density from α to 2α. Now let us repeat this doubling of the density $k = \log(1/\alpha)$ many times. This results in a cap set in $\mathbb{Z}_3^m$ with

$$m = n - \frac{4}{\alpha} - \frac{4}{2\alpha} - \ldots - \frac{4}{2^k\alpha} \geqslant n - \frac{8}{\alpha} \geqslant \frac{n}{2},$$

with density $2^k\alpha > 1$. Since the density of a set cannot be larger than 1, this is a contradiction. □

An interesting consequence of the above proof is the following *counting lemma*, which states that if all the non-principal Fourier coefficients of A are small, then $t_{3\mathrm{AP}}(A) \approx \alpha^3$.

Corollary 4.2. *Let p be an odd number, and let $A \subseteq \mathbb{Z}_p^n$ be any subset with density α. We have*

$$|t_{3AP}(A) - \alpha^3| \leqslant \alpha \max_{a\neq 0} |\widehat{A}(a)|.$$

Finally, let us mention that in 2017, [CLP17, EG17] found an extremely elegant and short proof based on the polynomial method showing that cap sets in $\mathbb{Z}_3^n$ are of size at most $(3-\varepsilon)^n = N^{1-\delta}$ for some fixed $\varepsilon, \delta > 0$. Note that such a strong bound is not valid in $\mathbb{Z}_N$ due to Behrand's construction. The significance of the Fourier analytic approach lies in the integer case, where there is no known analog of the polynomial method. Moreover, compared to the Fourier analytic approach, even in $\mathbb{Z}_3^n$, the polynomial method appears to be much limited in dealing with linear structures other than arithmetic progressions.

4.1.1 Roth's original case $A \subseteq \{1, \ldots, N\}$

To study the number of occurrences of a linear pattern (e.g., 3-progressions) in a subset of the interval $\{1, \ldots, M\}$, it suffices to embed $\{1, \ldots, M\}$ in $\mathbb{Z}_N$ for a prime $N = O(M)$ chosen sufficiently large to avoid wraparound. Consequently, rather than working with the interval, one can focus on the finite Abelian group $\mathbb{Z}_N$.

Let $A \subseteq \mathbb{Z}_N$ be a cap set of density α. Similar to the $\mathbb{Z}_3^n$, one can apply Corollary 4.2 to show the existence of a large *non-principal* Fourier coefficient: $|\widehat{A}(a)| \geqslant \frac{\alpha^2}{2}$. Unlike $\mathbb{Z}_3^n$, the group $\mathbb{Z}_N$ does not have a rich collection of large subgroups. Therefore, we cannot deduce that A is significantly denser in a large coset. Instead, we need to work with some notion of an "approximate subgroup".

In the case of $\mathbb{Z}_N$, Roth's argument shows that $|\widehat{A}(a)| \geqslant \frac{\alpha^2}{2}$ implies the existence of a set $P \subseteq \mathbb{Z}_N$ such that the following conditions hold.

- (Density increment) $\frac{|A \cap P|}{|P|} \geqslant \alpha + \frac{\alpha^2}{100}$.
- (Approximate subgroup) P is an arithmetic progression of size $m := |P| \geqslant N^{1/3}$.

Note that P has the same linear structure as an interval. In particular, since $A \cap P$ is a cap set in P, we can deduce that there exists a cap set $A' \subseteq \{1, \ldots, m\}$ with density at least $\alpha + \frac{\alpha^2}{100}$. To be more precise, if $P = \{x + jy \ : \ j \in \{1, \ldots, m\}\}$, then we take

$$A' := \{j \in \{1, \ldots, m\} \ : \ x + jy \in A\}.$$

Note that each step of this proof increases the density by $\frac{\alpha^2}{100}$, which is similar to the case of $G = \mathbb{Z}_3^n$. However, this density increment comes at a higher cost of decreasing the group size from N to approximately $N^{1/3}$. This large decrease in the group size is the reason behind the extra logarithm in the denominator of Roth's bound $O\left(\frac{N}{\log \log N}\right)$.

Later improvements use more efficient notions of "approximate subgroups". In particular, Szemerédi [Sze90] uses the so-called generalized arithmetic progressions, which are sets of the form $x + j_1 y_1 + \ldots + j_d y_d$ where $x, y_1, \ldots, y_d$ are fixed and each j_i ranges over some interval $[0, k_i]$. In Bourgain[Bou99a], Bohr sets were used to develop a theory of approximate subgroups. Bohr sets are a key component of many recent improvements in the bounds of Roth's theorem.

4.2 Exercises

Exercise 4.1. Let $H = (V, E)$ be a small undirected graph. Let $A \subseteq \mathbb{Z}_2^n$. Consider

$$t_H(A) = \mathbb{E} \prod_{(u,v) \in E} A(\mathbf{x}_u + \mathbf{x}_v),$$

where $\{\mathbf{x}_u : u \in V\}$ are independent random variables taking values in $\mathbb{Z}_2^n$ uniformly at random.

In each of the following cases, express $t_H(A)$ in terms of the Fourier coefficients of A. Your formula must be as simple as possible.

1. H is a tree.
2. H is a cycle on k vertices.
3. H is the graph with vertex set $\{1, 2, 3, 4\}$ and edges $\{(1, 2), (1, 3), (1, 4), (3, 2), (4, 2)\}$. In this case, your final formula will involve two sums.
4. Similarly, for $A \subseteq \mathbb{Z}_N$, give a Fourier analytic formula for

$$\mathbb{E} A(\mathbf{x}) A(\mathbf{x} + \mathbf{y}) A(\mathbf{x} + 2\mathbf{y}) A(\mathbf{x} + 3\mathbf{y}).$$

Chapter 5

Pseudorandomness: Fourier Uniformity

Pseudo-randomness is one of the most useful concepts in computer science and several branches of mathematics. Broadly speaking, we consider a mathematical object pseudo-random if it mimics the typical behaviour of truly random objects according to specific criteria.

For instance, by the law of large numbers, a random sequence of ± 1's typically contains an approximately equal number of each. Based on this criterion, we could define a notion of pseudo-randomness, where any sequence with a roughly balanced count of ± 1's is considered pseudo-random. Interestingly, even such a basic notion can lead to deep and notoriously difficult problems in mathematics.

To be more rigorous, consider a random sequence $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots)$ where each $\mathbf{a}_i$ is chosen randomly and independently from the set $\{-1, 0, 1\}$. It follows from Hoeffding's concentration inequality (Lemma 5.1 below) that, with probability 1, we have

$$\left| \sum_{i=1}^{n} \mathbf{a}_i \right| = n^{\frac{1}{2}+o(1)}. \tag{5.1}$$

Now consider the Möbius function $\mu : \mathbb{N} \to \{-1, 0, 1\}$, defined as

$$\mu(n) = \begin{cases} 0 & p^2 | n \text{ for some prime } p; \\ (-1)^k & n = p_1 \ldots p_k \text{ for distinct primes } p_1, \ldots, p_k. \end{cases}$$

We might ask whether the Möbius function behaves similarly to a typical random function $\mathbf{f} : \mathbb{N} \to \{-1, 0, 1\}$ in regards to having a balanced count of ± 1's: *Is it true that* $|\sum_{i=1}^{n} \mu(i)| = O(n^{\frac{1}{2}+o(1)})$*?*

Remarkably, this seemingly basic question is equivalent to one of the most important unsolved problems in mathematics, the Riemann Hypothesis! The weaker statement that $|\sum_{i=1}^{n} \mu(i)| = o(n)$ is equivalent to the prime number theorem, which was first proved independently by Hadamard and Poussin in 1896.

For a notion of pseudo-randomness to be truly useful, it must ensure that a pseudo-random object behaves similarly to a random one in multiple ways beyond the specific criteria used to define it. For example, the Riemann Hypothesis is of great interest in number theory because, if true, it would show that, in many ways, the distribution of primes is similar to the numbers generated according to specific random heuristics.

In this chapter, we will discuss a notion of pseudo-randomness based on Fourier coefficients. Let us first recall Hoeffding's concentration inequality, which we will frequently apply to establish various properties of random functions.

Lemma 5.1 (Hoeffding's Inequality)**.** *Suppose that* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *are independent random variables with* $|\mathbf{x}_i| \leqslant 1$ *for each* $1 \leqslant i \leqslant n$*. Let* $\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i$*. For every* $t > 0$*,*

$$\mathbf{Pr}\left[|\mathbf{X} - \mathbb{E}[\mathbf{X}]| > t\right] < 2e^{\frac{-t^2}{2n}}.$$

5.1 Fourier Uniformity

Let G be a finite group. The simplest statistic we can use to define pseudo-randomness for functions $f : G \to \{0, 1\}$ is their average, $\mathbb{E}[f] = \widehat{f}(0)$. By Hoeffding's inequality, for a random $\mathbf{f} : G \to \{0, 1\}$, we expect $\widehat{\mathbf{f}}(0)$ to be close to $\frac{1}{2}$.

Thus, we may call f pseudorandom if $|\mathbb{E}[f] - \frac{1}{2}|$ is small. However, this notion is too weak, as functions that meet this criterion do not exhibit many interesting properties of truly random functions. We will introduce a stronger notion of pseudo-randomness, called *Fourier uniformity*. Since we wish to apply this notion to sets $A \subseteq G$ of a given fixed density $\alpha \in [0,1]$, we will discard the principal Fourier coefficient $\widehat{A}(0) = \mathbb{E}[A]$, and consider $A - \mathbb{E}[A]$.

Definition 5.2. Let G be a finite Abelian group and $\delta > 0$ be a parameter. A function $f : G \to \mathbb{R}$ is *δ-Fourier uniform* if

$$\|\widehat{f - \mathbb{E}[f]}\|_\infty = \max_{a \neq 0} |\widehat{f}(a)| \leqslant \delta.$$

Fourier uniformity measures the correlation of f with non-principal characters of G. We will show in Section 5.1.1 that, in certain aspects, a Fourier uniform function $f : G \to \{0,1\}$ with $\mathbb{E}[f] = \alpha$ behaves similar to a random function $\mathbf{f} : G \to \{0,1\}$ conditioned on $\mathbb{E}_{\mathbf{x}}[\mathbf{f}(\mathbf{x})] = \alpha$.

First, let's establish that Fourier uniformity is a meaningful measure of pseudo-randomness by showing that a truly random function $\mathbf{f} : G \to \{0,1\}$ is typically Fourier uniform.

Proposition 5.3. *Let G be a finite Abelian group of size N, and let $\mathbf{A} \subseteq G$ be a random subset of G. We have*

$$\Pr_{\mathbf{A}}\left[\max_{a \neq 0} |\widehat{\mathbf{A}}(a)| > \frac{2\sqrt{\log N}}{\sqrt{N}}\right] = o_{N \to \infty}(1). \tag{5.2}$$

Proof. By Hoeffding's inequality, for $a \in G$ with $a \neq 0$, we have

$$\Pr_{\mathbf{A}}\left[|\widehat{\mathbf{A}}(a)| > \delta\right] = \Pr_{\mathbf{A}}\left[\left|\sum_{x \in G} \mathbf{A}(x)\chi_a(x)\right| > \delta N\right] \leqslant 2e^{\frac{-\delta^2 N^2}{2N}} = 2e^{-\delta^2 N/2}$$

Then, the union bound over all non-principal characters implies

$$\Pr\left[\max_{a \neq 0} |\widehat{\mathbf{A}}(a)| > \delta\right] \leqslant 2Ne^{-\delta^2 N/2}\ .$$

Setting $\delta = \frac{2\sqrt{\log N}}{\sqrt{N}}$ establishes Eq. (5.2). □

Remark 5.4. Note that the proof of Proposition 5.3 holds even if we sample $\mathbf{A}$ by including each element independently with any fixed probability $\alpha \in [0,1]$. Moreover, one can extend Proposition 5.3 further to the case where $\mathbf{A} \subseteq G$ is a random subset of a given density $\alpha > 0$. However, in that case, since $\mathbf{A}(x)$ are not independent, the proof is slightly more involved as one cannot simply apply Hoeffding's inequality.

On the other hand, the following proposition shows that no subset $A \subseteq G$ with density bounded away from 0 and 1 can achieve Fourier uniformity with parameters significantly stronger than those of a random subset. The assumption on density is crucial, as extreme cases like the empty set $A = \emptyset$ and the entire group $A = G$ are 0-Fourier uniform.

Proposition 5.5. *Let $\varepsilon \in (0,1)$ be a fixed constant. Every set $A \subseteq G$ with density $\varepsilon < \frac{|A|}{|G|} < 1 - \varepsilon$ satisfies*

$$\max_{a \in G} |\widehat{A}(a)| \geqslant \frac{\varepsilon(1-\varepsilon)}{\sqrt{N}}.$$

Proof. Let $\alpha \coloneqq \frac{|A|}{|G|}$. By Parseval's identity, we have

$$\alpha = \|A\|_2^2 = \sum_{a \in G} |\widehat{A}(a)|^2 = |\widehat{A}(0)|^2 + \sum_{a \neq 0} |\widehat{A}(a)|^2 = \alpha^2 + \sum_{a \neq 0} |\widehat{A}(a)|^2,$$

which shows that

$$\max_{a \neq 0} |\widehat{A}(a)| \geqslant \frac{\alpha(1-\alpha)}{\sqrt{N}}.$$

□

5.1.1 Fourier Uniformity and Counting Linear Patterns

In this section, we prove that Fourier uniformity is sufficient to guarantee that a subset A of a finite Abelian group G contains the "expected" number of certain linear patterns. We have already seen this result for the particular case of 3-term progressions in the proof of Roth's theorem in Chapter 4. As the reader may recall, every set $A \subseteq \mathbb{Z}_N$ with density $\frac{|A|}{N} = \alpha$ satisfies

$$|t_{3\mathrm{AP}}(A) - \alpha^3| \leqslant \alpha \max_{a \neq 0} |\widehat{A}(a)|.$$

In particular, $t_{3\mathrm{AP}}(A) \approx \alpha^3$ if A is δ-Fourier uniform for a small δ. Note that a truly random subset $\mathbf{A} \subseteq \mathbb{Z}_N$ with density α is expected to satisfy $t_{3\mathrm{AP}}(\mathbf{A}) \approx \alpha^3$ with high probability.

We aim to generalize this result to a larger class of linear patterns.

Systems of linear forms. A *linear form* in d variables is a vector $L = (\lambda_1, \ldots, \lambda_d) \in \mathbb{Z}^d$. The linear form L defines a linear map $G^d \to G$ by $L(x_1, \ldots, x_d) \coloneqq \sum_i \lambda_i x_i$. A *system of m linear forms in d variables* is a tuple $\mathcal{L} = (L_1, \ldots, L_m)$ of linear forms. A tuple $(a_1, \ldots, a_m) \in G^m$ is called an *instance* of $\mathcal{L}$ if there exists $x \in G^d$ with $\mathcal{L}(x) \coloneqq (L_1(x), \ldots, L_m(x)) = (a_1, \ldots, a_m)$. It is called a *non-degenerate instance* if additionally $a_1, \ldots, a_m$ are all distinct.

We emphasize that the choice of the system of linear forms that defines a linear pattern is not unique. For example, 3-term arithmetic progressions are instances of the system of linear forms $(x_1, x_1 + x_2, x_1 + 2x_2)$, or alternatively, one can take the system of linear forms $\mathcal{L} \coloneqq (2x_1 - 2x_2, x_1 - x_3, 2x_2 - 2x_3)$.

We often require that $|G|$ be coprime with all the coefficients of the linear forms that define $\mathcal{L}$. Otherwise, it would be possible to have sets that do not contain instances of even a single linear form. For example, let $L(x) = 2x$ and $G = \mathbb{Z}_4^n$. Then the set $A = \{x \in G : x \bmod 2 \equiv 1\}$ does not contain any instance of $L(x)$ for $x \in G$.

Let $\mathcal{L} = (L_1, \ldots, L_m)$ be a system of linear forms in d variables and let $A \subseteq G$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_d$ be independent random variables taking values in G uniformly at random. The probability that $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_d) \in A^m$ is given by the expectation

$$t_{\mathcal{L}}(A) \coloneqq \mathbb{E}\left[\prod_{i=1}^{m} A(L_i(\mathbf{x}_1, \ldots, \mathbf{x}_d))\right].$$

Definition 5.6 (Binary Systems of Linear forms)**.** A system of linear forms $\mathcal{L} = (L_1, \ldots, L_m)$ in the d variables $x_1, \ldots, x_d$ is *binary* if every linear form in $\mathcal{L}$ is supported on exactly two variables, and moreover, no two linear forms in $\mathcal{L}$ are supported on the same two variables.

Example 5.7. The system of linear forms $(2x_1 - 2x_2, x_1 - x_3, 2x_2 - 2x_3)$ that defines 3-term arithmetic progressions is a binary system. As another example, $(x_1 + x_2, x_1 + x_3, x_1 + x_4, x_2 + x_3, x_2 + x_4, x_3 + x_4)$ is a binary system of 6-linear forms in 4 variables.

Let $\mathcal{L}$ be a system of m linear forms. A simple second-moment argument shows that if $\mathbf{A}$ is a random subset of G with density α, then with high probability $t_{\mathcal{L}}(\mathbf{A}) \approx \alpha^m$. We will show that if $\mathcal{L}$ is a binary system, then $t_{\mathcal{L}}(A) \approx \alpha^m$ for any Fourier uniform set $A \subseteq G$. We will need the following lemma.

Lemma 5.8. *Let G be a finite Abelian group and $f : G \to \mathbb{R}$. We have*

$$\max_{g,h:G\to[-1,1]} |\mathbb{E}_{\mathbf{x},\mathbf{y}\in G} f(\mathbf{x}+\mathbf{y})g(\mathbf{x})h(\mathbf{y})| \leqslant \|\widehat{f}\|_\infty.$$

Proof. By replacing f, g, h with their Fourier expansion and using the orthogonality of Fourier characters, we obtain

$$\mathbb{E}_{\mathbf{x},\mathbf{y}\in G} g(\mathbf{x}) f(\mathbf{x}+\mathbf{y}) h(\mathbf{y}) = \sum_a \widehat{f}(a)\widehat{g}(-a)\widehat{h}(-a).$$

Note that $g, h : G \to [-1, 1]$, and therefore, $\|g\|_2, \|h\|_2 \leqslant 1$. By using the Cauchy-Schwarz inequality and then Parseval,

we have

$$\begin{aligned}\sum_a \widehat{f}(a)\widehat{g}(-a)\widehat{h}(-a) &\leqslant \|\widehat{f}\|_\infty \sum_a |\widehat{g}(a)||\widehat{h}(a)| \\ &\leqslant \|\widehat{f}\|_\infty \left(\sum_a |\widehat{g}(a)|^2\right)^{1/2} \left(\sum_a |\widehat{h}(a)|^2\right)^{1/2} \\ &= \|\widehat{f}\|_\infty \|g\|_2 \|h\|_2 \leqslant \|\widehat{f}\|_\infty.\end{aligned}$$

□

Theorem 5.9. *Let $\mathcal{L} = (L_1, \dots, L_m)$ be a binary system of linear forms. Suppose G is a finite Abelian group such that $|G|$ is coprime with all the coefficients in $\mathcal{L}$. If $A \subseteq G$ is δ-Fourier uniform, then*

$$|t_{\mathcal{L}}(A) - \alpha^m| \leqslant m\delta$$

Proof. We will use induction on m. The statement is trivial for $m = 1$, since in that case, $t_{\mathcal{L}}(A) = \alpha$.

Consider $m > 1$. Suppose that the first linear form is $L_1(x_1, \dots, x_d) = \lambda_1 x_1 + \lambda_2 x_2$ for constants $\lambda_1, \lambda_2 \in \mathbb{Z}$ that are coprime with $|G|$. We may apply[1] the change of variables $x_1' = \lambda_1 x_1$ and $x_2' = \lambda_2 x_2$, and assume, without loss of generality, that $L_1(x_1, \dots, x_d) = x_1 + x_2$. We have

$$t_{\mathcal{L}}(A) \coloneqq \mathbb{E}\left[A(\mathbf{x}_1 + \mathbf{x}_2) \prod_{i=2}^m A(L_i(\mathbf{x}_1, \dots, \mathbf{x}_d))\right].$$

Denote $f \coloneqq A - \alpha$, and note that δ-Fourier uniformity of A means $\|\widehat{f}\|_\infty \leqslant \delta$. By substituting $A = \alpha + f$ in the first linear form, we obtain

$$t_{\mathcal{L}}(A) = \alpha\mathbb{E}\left[\prod_{i=2}^m A(L_i(\mathbf{x}_1, \dots, \mathbf{x}_d))\right] + \mathbb{E}\left[f(\mathbf{x}_1 + \mathbf{x}_2) \prod_{i=2}^m A(L_i(\mathbf{x}_1, \dots, \mathbf{x}_d))\right]. \tag{5.3}$$

Since $(L_2, \dots, L_m)$ is a binary system of $(m-1)$-linear forms, we can apply the induction hypothesis to the first expected value and obtain

$$\left|\mathbb{E}\left[\prod_{i=2}^m A(L_i(\mathbf{x}_1, \dots, \mathbf{x}_d))\right] - \alpha^{m-1}\right| \leqslant (m-1)\delta. \tag{5.4}$$

Next, we will study the second expected value. By the definition of a binary system, L_1 is the only linear form involving both x_1 and x_2. Let S be the set of $i \in \{2, \dots, m\}$ such that L_i involves x_1, and let $S' = \{2, \dots, m\} \setminus S$. For any choice of $x_3, \dots, x_n \in G$, we can decompose

$$\prod_{i=2}^m A(L_i(x_1, \dots, x_d)) = \prod_{i \in S} A(L_i(x_1, \dots, x_d)) \prod_{i \in S'} A(L_i(x_1, \dots, x_d)) = g_{x_3,\dots,x_n}(x_1) h_{x_3,\dots,x_n}(x_2),$$

for some functions $g_{x_3,\dots,x_n}, h_{x_3,\dots,x_n} : G \to \{0, 1\}$. With this notation, we have

$$\mathbb{E}\left[f(\mathbf{x}_1 + \mathbf{x}_2) \prod_{i=2}^m A(L_i(\mathbf{x}_1, \dots, \mathbf{x}_d))\right] = \mathbb{E}_{\mathbf{x}_3,\dots,\mathbf{x}_d} \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2} f(\mathbf{x}_1 + \mathbf{x}_2) g_{\mathbf{x}_3,\dots,\mathbf{x}_n}(\mathbf{x}_1) h_{\mathbf{x}_3,\dots,\mathbf{x}_n}(\mathbf{x}_2).$$

Since $\|\widehat{f}\|_\infty \leqslant \delta$, we can apply Lemma 5.8 to the inner expected value $\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2} f(\mathbf{x}_1 + \mathbf{x}_2) g(\mathbf{x}_1) h(\mathbf{x}_2)$ and upper-bound it by δ. We obtain

$$\mathbb{E}\left[f(\mathbf{x}_1 + \mathbf{x}_2) \prod_{i=2}^m A(L_i(\mathbf{x}_1, \dots, \mathbf{x}_d))\right] \leqslant \delta,$$

[1] Since λ_1, λ_2 are coprime with $|G|$, there exist λ_1^{-1} and λ_2^{-1} such that $x_1 = \lambda_1^{-1} x_1'$ and $x_2 = \lambda_2^{-1} x_2'$.

which combined with Eq. (5.4) and Eq. (5.3) gives

$$|t_{\mathcal{L}}(A) - \alpha^m| \leqslant \delta + \alpha\delta(m-1) \leqslant \delta m.$$

□

Remark 5.10. By more careful analysis, one can improve the assertion of Theorem 5.9 to $|t_{\mathcal{L}}(A) - \alpha^m| \leqslant m\alpha\delta$. See Exercise 5.1.

Theorem 5.9 shows that binary linear patterns are controlled by Fourier uniformity. On the other hand, many interesting linear form systems are not controlled by Fourier uniformity. For example, there are sets $A \subseteq \mathbb{Z}_N$ that are $o(1)$-Fourier uniform, but the count of 4-term arithmetic progressions in them is far from what is expected from a random set of the same density [Gow01].

5.2 Gowers Uniformity Norms

In [Gow01], Gowers introduced a way of quantifying Fourier uniformity that operates entirely in the physical space without relying on Fourier coefficients.

Definition 5.11. Let G be a finite Abelian group and $f : G \to \mathbb{C}$ be a function. The U^2 norm of f is defined as

$$\|f\|_{U^2} := \left(\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{z}\in G} f(\mathbf{x})\overline{f(\mathbf{x}+\mathbf{y})f(\mathbf{x}+\mathbf{z})}f(\mathbf{x}+\mathbf{y}+\mathbf{z})\right)^{1/4}. \tag{5.5}$$

We must establish that the $\|\cdot\|_{U^2}$ is a norm. Note that a priori, it is not even clear that the expected value in the right-hand side of (5.5) is a non-negative real number. Replacing f by its Fourier expansion and expanding, we have

$$\begin{aligned}
\|f\|_{U^2}^4 &= \mathbb{E} f(\mathbf{x})\overline{f(\mathbf{x}+\mathbf{y})f(\mathbf{x}+\mathbf{z})}f(\mathbf{x}+\mathbf{y}+\mathbf{z}) \\
&= \sum_{a,b,c,d} \widehat{f}(a)\overline{\widehat{f}(b)\widehat{f}(c)}\widehat{f}(d)\mathbb{E}\left[\chi_a(\mathbf{x})\chi_{-b}(\mathbf{x}+\mathbf{y})\chi_{-c}(\mathbf{x}+\mathbf{z})\chi_d(\mathbf{x}+\mathbf{y}+\mathbf{z})\right] \\
&= \sum_{a,b,c,d} \widehat{f}(a)\overline{\widehat{f}(b)\widehat{f}(c)}\widehat{f}(d)\mathbb{E}\left[\chi_{a-b-c+d}(\mathbf{x})\chi_{-b+d}(\mathbf{y})\chi_{-c+d}(\mathbf{z})\right].
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}\left[\chi_{a-b-c+d}(\mathbf{x})\chi_{-b+d}(\mathbf{y})\chi_{-c+d}(\mathbf{z})\right] &= \begin{cases} 1 & a-b-c+d=0, -b+d=0, -c+d=0; \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} 1 & a=b=c=d; \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}$$

we have $\|f\|_{U^2}^4 = \sum_{a\in G} |\widehat{f}(a)|^4$. Therefore,

$$\|f\|_{U^2} = \left(\sum_{a\in G} |\widehat{f}(a)|^4\right)^{1/4} = \|\widehat{f}\|_4, \tag{5.6}$$

and the U^2 norm coincides with the ℓ_4 norm of the Fourier coefficients of f. The following lemma shows that for functions $f : G \to [0,1]$ or more generally $f : G \to \mathbb{C}$ with $\|f\|_\infty \leqslant 1$, the $\|f\|_{U^2}$ and $\|\widehat{f}\|_\infty$ are within a quadratic factor of each other.

Lemma 5.12. *Let G be a finite Abelian group, and $f : G \to \mathbb{C}$ satisfy $\|f\|_\infty \leqslant 1$. Then*

$$\|\widehat{f}\|_\infty \leqslant \|f\|_{U^2} \leqslant \sqrt{\|\widehat{f}\|_\infty}.$$

Proof. By (5.6), we have

$$\|f\|_{U^2}^4 = \sum_{a\in G} |\widehat{f}(a)|^4 \geqslant \max_{a\in G} |\widehat{f}(a)|^4 = \|\widehat{f}\|_\infty^4,$$

which establishes the first inequality. To prove the second inequality, note that $\|f\|_2 \leqslant \|f\|_\infty$, and therefore by Parseval's identity,

$$\|f\|_{U^2}^4 = \sum_{a\in G} |\widehat{f}(a)|^4 \leqslant \left(\max_{a\in G} |\widehat{f}(a)|^2\right) \sum_{a\in G} |\widehat{f}(a)|^2 \leqslant \max_{a\in G} |\widehat{f}(a)|^2 = \|\widehat{f}\|_\infty^2.$$

□

In light of Lemma 5.12, a set A is Fourier uniform if and only if $\|A - \mathbb{E}[A]\|_{U^2}$ is small. The advantage of the U^2 norm over $\|\widehat{f}\|_\infty$ is that its definition does not involve the Fourier transform and can be fully described in the physical space without any reference to Fourier coefficients.

Gowers's interpretation of Fourier uniformity using the U^2 norm enabled him to generalize it to stronger notions of pseudo-randomness. In his proof [Gow01] of Szemerédi's theorem, he introduced a hierarchy of increasingly stronger notions of pseudo-randomness based on the so-called Gowers uniformity norms $\|\cdot\|_{U^k}$ for $k \geqslant 1$.

Definition 5.13 (Gowers Uniformity Norms)**.** Let G be a finite Abelian group and $f : G \to \mathbb{C}$ be a function. For $k \in \mathbb{N}$, the U^k uniformity norm of f is

$$\|f\|_{U^k} \coloneqq \left(\mathbb{E}_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k\in G} \prod_{S\subseteq[k]} \mathcal{C}^{|S|} f(\mathbf{x} + \sum_{i\in S} \mathbf{y}_i)\right)^{1/2^k},$$

where $\mathcal{C}$ denotes the complex conjugation operator. In other words, the terms with odd $|S|$ are conjugated.

He used iterated applications of the classical Cauchy–Schwarz inequality to prove

$$\left|\mathbb{E}_{\mathbf{x},\mathbf{y}} \left[A(\mathbf{x})A(\mathbf{x}+\mathbf{y})\cdots A(\mathbf{x}+(k-1)\mathbf{y})\right] - \alpha^k\right| \leqslant k\|A-\alpha\|_{U^{k-1}},$$

showing that the density of k-progressions in A is controlled by the pseudo-randomness condition $\|A-\alpha\|_{U^{k-1}} = o(1)$.

5.3 Conclusion

Green and Tao [GT10] determined the most general class of systems of linear forms that can be handled by Gowers' iterated Cauchy-Shwarz argument.

Definition 5.14 (The Cauchy–Schwarz Complexity)**.** Let $\mathcal{L} = \{L_1, \ldots, L_m\}$ be a system of linear forms. The *Cauchy-Schwarz complexity* of $\mathcal{L}$ is the minimal k such that the following holds. For every $1 \leqslant i \leqslant m$, we can partition $\{L_j\}_{j\in[m]\setminus\{i\}}$ into $k+1$ subsets, such that L_i does not belong to the linear span of any of these subsets.

In particular, Green and Tao [GT10] showed that if the Cauchy–Schwarz complexity of $\mathcal{L}$ is k, then

$$|t_{\mathcal{L}}(A) - \alpha^m| \leqslant m\|A-\alpha\|_{U^{k+1}}. \tag{5.7}$$

Remark 5.15. It is easy to see that binary systems of linear forms have CS-complexity 1. The system of linear forms $(x, x+y, \ldots, x+(k-1)y)$, which represents k-term arithmetic progressions, has CS-complexity $k-2$.

Later, in a series of articles [GW11, GW10], Gowers and Wolf initiated a systematic study of classifying the systems of linear forms controlled by the k-th Gowers uniformity norm. They defined the *true complexity* of a system of linear forms $L = (L_1, \ldots, L_m)$ as the smallest k such that the pseudo-randomness condition of $\|A-\alpha\|_{U^{k+1}} = o(1)$ implies $|t_{\mathcal{L}}(A) - \alpha^m| = o(1)$. Note that by (5.7), the true complexity is at most the CS-complexity.

In connection to Theorem 5.9, we have

$$\{\mathcal{L} \;:\; \mathcal{L} \text{ is binary}\} \subseteq \{\mathcal{L} \;:\; \mathcal{L} \text{ has CS-complexity } 1\} \subseteq \{\mathcal{L} \;:\; \mathcal{L} \text{ has true complexity } 1\}.$$

Randomness versus structure The dichotomy between pseudorandomness and structure refers to a general phenomenon that *non-pseudo-random* behaviour indicates resemblance to a highly structured object. Like many proofs in extremal and additive combinatorics, Roth's argument on 3-term progressions exploits this dichotomy. It shows that if the number of 3-progressions in $A \subseteq \mathbb{Z}_3^n$ significantly deviates from what is expected from a random set of density α, then A has a large non-principal Fourier coefficient. Equivalently, it has a notable positive correlation with an affine subspace $V \subseteq \mathbb{Z}_3^n$ of codimension 1 (a highly structured object).

5.4 Exercises

Exercise 5.1. Improve the assertion of Theorem 5.9 to $|t_{\mathcal{L}}(A) - \alpha^m| \leqslant m\alpha\delta$.

Exercise 5.2. This exercise shows that for some linear patterns, the Fourier uniformity (i.e. having small non-principal Fourier coefficients) is insufficient to guarantee that a set behaves similarly to a random set in terms of the density of the pattern. Let $n = 2m$ be an even integer, and let

$$A_n = \left\{ x \in \mathbb{Z}_2^n \ : \ \sum_{i=1}^{m} x_{2i-1}x_{2i} \equiv 0 \bmod 2 \right\}.$$

1. Directly calculate all the Fourier coefficients of A_n.
2. What are $\widehat{A_n}(0)$ and $\max_{a \neq 0} |\widehat{A_n}(a)|$?
3. Prove that
$$\lim_{n\to\infty} \left| \mathbb{E}\left[\prod_{1 \leqslant i < j < k \leqslant 6} A_n(\mathbf{x}_i + \mathbf{x}_j + \mathbf{x}_k) \right] - |\widehat{A_n}(0)|^{\binom{6}{3}} \right| \neq 0,$$
where $\mathbf{x}_1, \ldots, \mathbf{x}_6$ are independent random variables taking values uniformly in $\mathbb{Z}_2^n$.

Exercise 5.3. Extend Theorem 5.9 to systems of linear forms of CS-complexity 1.

Chapter 6

Degree and Granularity of Fourier Coefficients

By identifying $\{\text{FALSE}, \text{TRUE}\}$ with $\mathbb{Z}_2$, $\{-1, 1\}$, or $\{0, 1\}$ we obtain different representation of functions on Boolean domain.

(I) $f : \mathbb{Z}_2^n \to \{0, 1\}$. This representation allows us to consider the Fourier transform of f over the Abelian group $\mathbb{Z}_2^n$, and write

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S(x), \tag{6.1}$$

with $\chi_S(x) = (-1)^{\sum x_i}$.

(II) $f : \{-1, 1\}^n \to \{0, 1\}$. This representation will allow us to consider the unique representation of f as a multilinear polynomial

$$f(y) = \sum_{S \subseteq [n]} \widehat{f}(S) \prod_{i \in S} y_i, \tag{6.2}$$

where

$$\widehat{f}(S) = \mathbb{E}_{\mathbf{y}} \left[f(\mathbf{y}) \prod_{i \in S} y_i \right] \text{ for } S \subseteq [n].$$

To verify that the correct coefficients are indeed the Fourier coefficients $\widehat{f}(S)$, note that the change of variable $(-1)^x = y$ converts the character $\chi_S(x) = (-1)^{\sum x_i}$ to the monomial $\prod_{i \in S} y_i$.

(III) $f : \{0, 1\}^n \to \{0, 1\}$. This representation will allow us to consider the polynomial representation of f as

$$f(z) = \sum_{S \subseteq [n]} a_S \prod_{i \in S} z_i, \tag{6.3}$$

with the coefficients given by the inclusion-exclusion formula

$$a_S = \sum_{T \subseteq S} (-1)^{|S \setminus T|} f(\mathbf{1}_T),$$

where $\mathbf{1}_T \in \{0, 1\}^n$ denotes the vector that is 1 for the coordinates in T and 0 in other coordinates. We emphasize that, unlike Fourier coefficients, the coefficients a_S are integers in this representation.

The change of variable $z_i = \frac{y_i + 1}{2}$ between $z_i \in \{0, 1\}$ and $y_i \in \{-1, 1\}$ shows

$$\sum_{S \subseteq [n]} a_S \prod_{i \in S} z_i = \sum_{S \subseteq [n]} a_S \prod_{i \in S} \frac{y_i + 1}{2} = \sum_{S \subseteq [n]} \left(\sum_{T \supseteq S} 2^{-|T|} a_T \right) \prod_{i \in S} y_i.$$

Consequently, we obtain the following relations between the coefficients a_S and the Fourier coefficients:

$$\widehat{f}(S) = \sum_{T \supseteq S} 2^{-|T|} a_T. \tag{6.4}$$

and conversely

$$a_S = \sum_{T \supseteq S} 2^{|S|} (-1)^{|T \setminus S|} \widehat{f}(T). \tag{6.5}$$

6.1 Real Degree

A crucial fact about the representations Eq. (6.2) and Eq. (6.3) is that whether we define $f : \{-1,1\}^n \to \mathbb{R}$ or $f : \{0,1\}^n \to \mathbb{R}$, the degree of the polynomial remains the same: The largest S such that $a_S \neq 0$ also satisfies $\widehat{f}(S) \neq 0$ and vice versa.

Definition 6.1 (Degree)**.** The *real degree* (degree for short) of $f : \mathbb{Z}_2^n \to \mathbb{R}$ denoted by $\deg(f)$, is the largest $|S|$ such that $\widehat{f}(S) \neq 0$.

Remark 6.2. In Definition 6.1, we called $\deg(f)$ the "real" degree to differentiate it from the degree of polynomials $p : \mathbb{Z}_2^n \to \mathbb{Z}_2$. For example, the parity function as a polynomial from $\mathbb{Z}_2^n$ the $\mathbb{Z}_2$ is of degree 1 since $\textsc{Parity}(x_1, \ldots, x_n)$ is the sum (over $\mathbb{Z}_2$) of the coordinates. However, as the Example 6.3 below shows, the real degree of the parity function is n.

Example 6.3. Consider the function $\textsc{Parity} : \mathbb{Z}_2^n \to \{0,1\}$ defined as

$$\textsc{Parity}(x) = \sum_{i=1}^{n} x_i \bmod 2.$$

We have

$$\textsc{Parity}(x) = \frac{1}{2} + \frac{1}{2}\chi_{[n]}(x),$$

and therefore, $\deg(\textsc{Parity}) = n$.

6.2 Granularity of Fourier Coefficients

The following theorem shows that non-zero Fourier coefficients of a low-degree function $f : \mathbb{Z}_2^n \to \{0,1\}$ are large in magnitude.

Theorem 6.4 (Granularity of Fourier Coefficients)**.** *Consider $f : \mathbb{Z}_2^n \to \{0,1\}$ and let $d := \deg(f)$. Every Fourier coefficient of f is of the form $\widehat{f}(S) = \frac{b_S}{2^d}$ for some integer $b_S \in \mathbb{Z}$. In particular, non-zero Fourier coefficients satisfy $|\widehat{f}(S)| \geqslant \frac{1}{2^d}$.*

Proof. In Eq. (6.4), we have $a_S \in \mathbb{Z}$ and $a_T = 0$ if $|T| > \deg(f)$. The assertion immediately follows. □

Remark 6.5. Theorem 6.4 shows that for $G = \mathbb{Z}_2^n$, every non-zero Fourier coefficient of a Boolean function $f : G \to \{0,1\}$ must satisfy $|\widehat{f}(S)| \geqslant \frac{1}{|G|}$. Such a strong lower bound is not true for general finite Abelian groups. In particular, there exist Boolean functions $f : \mathbb{Z}_N \to \{0,1\}$ that have non-zero Fourier coefficients with $|\widehat{f}(a)| \leqslant \frac{O(1)}{2^N}$. See Exercise 6.1.

It is also possible to prove Theorem 6.4 directly, without referring to the polynomial representation of f as a real function on $\{0,1\}^n$.

Define the *discrete derivative* of $f : \mathbb{Z}_2^n \to \{0,1\}$ in direction e_i as $\partial_i f : x \mapsto \frac{f(x) - f(x+e_i)}{2}$. Since

$$\chi_T(x + e_i) = \begin{cases} -\chi_T(x) & \text{if } i \in T \\ \chi_T(x) & \text{if } i \notin T \end{cases},$$

we have

$$\partial_i f = \sum_{T: i \in T} \widehat{f}(T)\chi_T.$$

Note that ∂_i is a linear operator on the space of function $f : \mathbb{Z}_2^n \to \mathbb{R}$. Given $S = \{i_1, \ldots, i_k\} \subseteq [n]$, let $\partial_S \coloneqq \partial_{i_1} \circ \ldots \circ \partial_{i_k}$ and note

$$\partial_S f = \partial_{i_1} \circ \ldots \circ \partial_{i_k} f = 2^{-|S|} \sum_{T \subseteq S} (-1)^{|S \setminus T|} f(x + \sum_{i \in T} e_i) \tag{6.6}$$

and

$$\partial_S f = \sum_{T \supseteq S} \widehat{f}(T)\chi_T. \tag{6.7}$$

A second proof of Theorem 6.4. Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be a polynomial of degree d. We prove the statement by induction on $|S|$ with the base case $|S| = d$. To verify the base of induction, consider $S \subseteq [n]$ with $|S| = d$. Since $\widehat{f}(T) = 0$ for any T whose size is larger than d, Equation (6.7) shows

$$\partial_S f(0) = \widehat{f}(S)\chi_S(0) = \widehat{f}(S).$$

Therefore, $\widehat{f}(S) = \partial_S f(0)$, which verifies the base of induction since by Equation (6.6), we have $\partial_S f(0) \in 2^{-|S|} \cdot \mathbb{Z}$.

Next, consider $S \subseteq [n]$ with $|S| < d$. By Equation (6.7), we have

$$\partial_S f(0) = \sum_{T \supseteq S} \widehat{f}(S)\chi_T(0) = \sum_{T \supseteq S} \widehat{f}(T),$$

and therefore,

$$\widehat{f}(S) = \partial_S f(0) - \sum_{T \supsetneq S} \widehat{f}(T).$$

By the induction hypothesis, for every $T \supsetneq S$, we have $\widehat{f}(T) = \frac{b_T}{2^d}$ with $b_T \in \mathbb{Z}$. Consequently, $\widehat{f}(S) = \frac{b_S}{2^d}$ with $b_S \in \mathbb{Z}$. □

6.3 Low-degree functions are dictators and juntas

Theorem 6.4 shows that the non-zero Fourier coefficients of a low-degree function $f : \mathbb{Z}_2^n \to \{0,1\}$ are large in magnitude. On the other hand, by the Parseval identity, we have $\sum_S |\widehat{f}(S)|^2 = \mathbb{E}_{\mathbf{x}} f(\mathbf{x})^2 \leqslant 1$. Therefore, a low-degree Boolean function cannot have many non-zero Fourier coefficients. This observation allows us to obtain certain classifications of low-degree Boolean functions.

The following two definitions are crucial in the analysis of Boolean functions. They describe functions that are "local" in that a few variables determine their values.

Definition 6.6 (Dictator)**.** Let $f : \mathbb{Z}_2^n \to \{0,1\}$ is a *dictator* if there exists $i \in [n]$, such that one of the following cases hold:

- $f(x) = x_i$ for all $x \in \mathbb{Z}_2^n$.
- $f(x) = 1 - x_i$ for all $x \in \mathbb{Z}_2^n$.

In other words, the value of $f(x)$ is dictated by only one variable. More generally, we can consider the functions that depend on a few variables.

Definition 6.7 (Junta)**.** Let $f : \mathbb{Z}_2^n \to \mathbb{R}$ and $J = \{j_1, \ldots, j_k\} \subseteq [n]$. We call f a *J-junta* if there exists $g : \mathbb{Z}_2^k \to \mathbb{R}$ such that $f(x) = g(x_{j_1}, \ldots, x_{j_k})$ for all $x \in \mathbb{Z}_2^n$. We call f a *k-junta* if f is a J-junta for a set J of size at most k.

We have the following characterization of low-degree Boolean functions.

Proposition 6.8. *Let $f : \mathbb{Z}_2^n \to \{0,1\}$.*

1. *If $\deg(f) = 0$, then $f \equiv 0$ or $f \equiv 1$.*

2. *If* $\deg(f) = 1$*, then* f *is a dictator.*

3. *If* $\deg(f) = d$*, then* f *is a* $d2^{2d}$*-junta.*

Proof. The first item is trivial. To prove the second item, suppose $\deg(f) = 1$. Since $f \not\equiv 0$ and $f \not\equiv 1$, we have $0 < \widehat{f}(0) < 1$, and therefore, by Theorem 6.4, we have $\widehat{f}(0) = \frac{1}{2}$. By Parseval, we have $\sum_{S\emptyset} |\widehat{f}(S)|^2 = \mathbb{E}[f^2] - \mathbb{E}[f] = \frac{1}{4}$. Then Theorem 6.4, implies that there exists exactly one non-zero Fourier coefficient $\widehat{f}(\{i\}) \neq 0$ and it satisfies $\widehat{f}(\{i\}) = \pm\frac{1}{2}$. Depending on the sign, either $f(x) = x_i$ or $f(x) = 1 - x_i$.

To prove the third item, note that by Theorem 6.4, all the non-zero Fourier coefficients satisfy $|\widehat{f}(S)| \geqslant \frac{1}{2^d}$. Combined with Parseval $\sum |\widehat{f}(S)|^2 = \mathbb{E}[f^2] \leqslant 1$, we conclude that there are at most 2^{2d} non-zero Fourier coefficients. Then $f = \sum_S \widehat{f}(S)\chi_S$ is a J-junta with

$$J = \bigcup_{S:\widehat{f}(S)\neq 0} S,$$

which satisfies $|J| \leqslant d2^{2d}$, where we used the fact that χ_S is an S-Junta. □

6.4 Exercises

Exercise 6.1. TO BE COMPLETED.

Chapter 7

Degree, decision trees, and sensitivity

In Chapter 6, we obtained a characterization of low-degree boolean functions: The real degree of a k-junta is at most k, and we prove that, conversely, a boolean function with degree at most k is a $k2^{2k}$-junta. In particular, $\deg(f) = O(1)$ if and only if f is a $O(1)$-junta.

In this chapter, we will obtain a more refined characterization of low-degree boolean that does not suffer the exponential loss of the degree versus junta characterization. This new characterization shows that the real degree is polynomially equivalent to the decision tree complexity.

$$\deg(f) \leqslant \mathrm{dt}(f) \leqslant 2^8 \deg(f)^6.$$

7.1 Decision trees

We define the decision tree complexity of boolean functions $f : \{0,1\}^n \to \{0,1\}$.

Definition 7.1 (decision tree)**.** A *decision tree* over variables $x_1, \ldots, x_n$ is a binary tree where each internal node has two children, left and right. Moreover, each internal node is labelled with a variable, and each leaf is labelled with a value of 0 or 1. To evaluate a decision tree at a point $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, we start from the root, and at each internal node with label x_i we *query* the value of x_i, go left if $x_i = 0$ and right if $x_i = 1$ until we reach a leaf. The leaf's value is the decision tree's output on x. For a boolean function $f : \{0,1\}^n \to \{0,1\}$, we let $\mathrm{dt}(f)$ denote the *smallest depth of a decision tree* computing f.

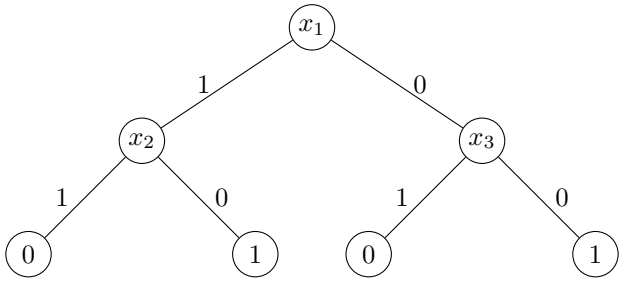

Figure 7.1: A depth 2 decision tree computing $f : \{0,1\}^3 \to \{0,1\}$ with $f(x_1, x_2, x_3) \coloneqq (\neg x_1 \wedge \neg x_3) \vee (x_1 \wedge \neg x_2)$.

The following proposition shows that the decision tree complexity is an upper bound on the real degree.

Proposition 7.2. *For every $f : \{0,1\}^n \to \{0,1\}$, we have $\deg(f) \leqslant \mathrm{dt}(f)$.*

Proof. Recall that the real degree is equal to the degree of $f : \{0,1\}^n \to \{0,1\}$ as a multilinear polynomial $\{0,1\}^n \to \mathbb{R}$. Consider a decision tree of f with depth $\mathrm{dt}(f)$. Let $\mathcal{L}$ be the set of the leaves of this tree, and let $\mathcal{L}_1$ be the set of leaves with label 1.

For every $x \in \{0,1\}^n$, let $\ell(x) \in \mathcal{L}$ denote the leaf reached by the tree when computing $f(x)$. Note

$$f(x) = \sum_{l \in \mathcal{L}_1} \mathbf{1}_{[\ell(x)=l]}.$$

For every $l \in \mathcal{L}_1$, consider the path from the root to l, and let T_l be the indices of variables on this path that returned the value 1, and F_l be the indices of the variables that returned the value 0. We have

$$\left(\prod_{i \in T_l} x_i\right)\left(\prod_{i \in F_l} (1 - x_i)\right) = 1 \iff \ell(x) = l,$$

and the degree of $\left(\prod_{i \in T_l} x_i\right)\left(\prod_{i \in F_l}(1-x_i)\right)$ is $|T_l| + |F_l| \leqslant \mathrm{dt}(f)$. Therefore,

$$f(x) = \sum_{l \in \mathcal{L}_1} \mathbf{1}_{[\ell(x)=l]} = \sum_{l \in \mathcal{L}_1} \left(\prod_{i \in T_l} x_i\right)\left(\prod_{i \in F_l} (1 - x_i)\right),$$

which shows $\deg(f) \leqslant \mathrm{dt}(f)$. □

7.2 Certificate complexity

Next, we will discuss a complexity measure closely related to decision trees.

Definition 7.3 (Certificate)**.** A certificate for an input $x \in \{0,1\}^n$ to a boolean function $f : \{0,1\}^n \to \{0,1\}$ is a set $S \subseteq [n]$ of indices such that f is constant on all inputs that match x on S.

- We denote by $C_f(x)$ the size of a smallest certificate for x.
- The *certificate complexity* of f is $C(f) \coloneqq \max_x C_f(x)$.

The following theorem shows that certificate and decision tree complexities are polynomially equivalent.

Theorem 7.4. *For every boolean function $f : \{0,1\}^n \to \{0,1\}$, we have*

$$C(f) \leqslant \mathrm{dt}(f) \leqslant C(f)^2.$$

Proof. To prove $C(f) \leqslant \mathrm{dt}(f)$, consider a decision tree of depth at most $\mathrm{dt}(f)$ for f. The set of the variables queried by the decision tree on an input x determines the value of $f(x)$ and, therefore, forms a certificate of size at most $\mathrm{dt}(f)$ for x.

Next, we prove $\mathrm{dt}(f) \leqslant C(f)^2$. Observe that if S is a certificate for $x \in f^{-1}(0)$ and T is a certificate for $y \in f^{-1}(1)$, then $x|_S$ cannot be compatible with $y|_T$, and therefore, $S \cap T \neq \emptyset$. Define

$$C^0(f) \coloneqq \max_{x \in f^{-1}(0)} C_f(x) \leqslant C(f) \quad \text{and} \quad C^1(f) \coloneqq \max_{x \in f^{-1}(1)} C_f(x) \leqslant C(f).$$

Pick any input x with $f(x) = 0$. If such an x does not exist, $\mathrm{dt}(f) = 0$, and the theorem trivially follows. Let S be the smallest certificate for x. Construct a partial decision tree of depth $|S| \leqslant C^0(f)$ by querying all the variables in S. Since S intersects all the certificates for inputs $y \in f^{-1}(1)$, each leaf ℓ of this partial tree corresponds to a function f_ℓ with $C^1(f_\ell) \leqslant C^1(f) - 1$ and $C^0(f_\ell) \leqslant C^0(f)$. By induction, f_ℓ has a decision tree of depth at most $C^0(f)(C^1(f) - 1)$, and therefore

$$\mathrm{dt}(f) \leqslant C^0(f) + C^0(f)(C^1(f) - 1) = C^0(f)C^1(f) \leqslant C(f)^2.$$

□

7.3 Degree of univariate polynomials and symmetrization

Our primary tool for proving lower bounds on the real degree of boolean functions is the following lower bound on the degree of univariate polynomials.

Theorem 7.5. *Let $q : \mathbb{R} \to \mathbb{R}$ be a univariate polynomial that satisfies*

(i) $q(0) = 0$ *and* $q(1) = 1$*;*

(ii) $|q(k)| \leqslant 1$ *for every* $k \in \{0, \ldots, m\}$*.*

Then $\deg(q) \geqslant \sqrt{m/2}$*.*

To prove Theorem 7.5, we need the following classical theorem from approximate theory.

Theorem 7.6 (Markov)**.** *Let $q : \mathbb{R} \to \mathbb{R}$ be a univariate polynomial of degree d such that any real number $x \in [a_1, a_2]$ satisfies $q(x) \in [b_1, b_2]$. Then for all $x \in [a_1, a_2]$, the derivative of q satisfies $|q'(x)| < d^2 \frac{b_2 - b_1}{a_2 - a_1}$.*

Proof of Theorem 7.5. Let $d = \deg(q)$. By the mean value theorem, there exists a point $x \in [0, 1]$ with $|q'(x)| \geqslant 1$. Let $c = \max_{x \in [0,m]} |q'(x)| \geqslant 1$. The mean value theorem implies that every real $x \in [0, m]$ satisfies

$$-\frac{c}{2} \leqslant q(x) \leqslant 1 + \frac{c}{2}.$$

Therefore, by Theorem 7.6, we have

$$c \leqslant d^2 \frac{1 + c}{m},$$

or equivalently,

$$d \geqslant \sqrt{\frac{cm}{1 + c}} \geqslant \sqrt{\frac{m}{2}},$$

where the last inequality uses $c \geqslant 1$. □

We can *symmetrize* a multivariate polynomial $p : \mathbb{R}^n \to \mathbb{R}$ by averaging it over all permutations of the variables

$$p^{\mathrm{sym}}(x_1, \ldots, x_n) \coloneqq \frac{1}{n!} \sum_{\pi \in S_n} p(x_{\pi_1}, \ldots, x_{\pi_n}),$$

where S_n denotes the set of all permutations of $\{1, \ldots, n\}$. Note that $\deg(p) \leqslant \deg(p^{\mathrm{sym}})$.

The following theorem, often attributed to Minsky and Paper [MP88], shows that $p^{\mathrm{sym}}(x_1, \ldots, x_n)$ corresponds to a univariate polynomial evaluated at $x_1 + \ldots + x_n$.

Theorem 7.7. *For every multilinear polynomial $p : \mathbb{R}^n \to \mathbb{R}$, there is a univariate polynomial $q : \mathbb{R} \to \mathbb{R}$ with* $\deg(q) \leqslant \deg(p)$ *such that*

$$p^{\mathrm{sym}}(x_1, \ldots, x_n) = q(x_1 + \ldots + x_n),$$

for all $(x_1, \ldots, x_n) \in \{0, 1\}^n$*.*

Proof. Define

$$P_k(x) \coloneqq \sum_{S \in \binom{[n]}{k}} \prod_{i \in S} x_i$$

as the sum of all degree k monomials. Let d be the degree of p. Using the symmetry of p^{sym}, we have

$$p^{\mathrm{sym}}(x) = c_0 + c_1 P_1(x) + \cdots + c_d P_d(x),$$

where $c_i \in \mathbb{R}$. Notice that $P_k(x)$ can be written as $\binom{x_1 + \ldots + x_n}{k}$, and define the univariate polynomial $q(t)$ as

$$q(t) \coloneqq c_0 + c_1 \binom{t}{1} + \cdots + c_d \binom{t}{d}$$

sp that $p^{\mathrm{sym}}(x) = q(x_1 + \ldots + x_n)$. Finally, note that $\deg(q) \leqslant \deg(\binom{t}{d}) = d$. □

Symmetrization combined with Theorem 7.5 provides a method for proving lower bounds on the real degree of certain boolean functions.

Corollary 7.8. *Let* $f : \{0,1\}^n \to \{0,1\}$ *satisfy* $f(0,\ldots,0) = 0$ *and* $f(e_1) = f(e_2) = \ldots = f(e_n) = 1$. *Then* $\deg(f) \geqslant \sqrt{n/2}$.

Proof. Let $p : \mathbb{R}^n \to \mathbb{R}$ be the unique representation of f as a multilinear polynomial of degree $\deg(f)$. Let $q : \mathbb{R} \to \mathbb{R}$ be the univariate polynomial provided by Theorem 7.7. Note that

$$q(0) = p^{\mathrm{sym}}(0,\ldots,0) = 0,$$

and

$$q(1) = p^{\mathrm{sym}}(e_1) = \frac{\sum_{i=1}^n p(e_i)}{n} = 1.$$

Moreover, $|q(k)| \leqslant 1$ for all $k \in \{0,\ldots,n\}$. Theorem 7.5 shows $\deg(q) \geqslant \sqrt{n/2}$. □

7.4 Sensitivity

We define another important parameter in the study of boolean functions, which measures the sensitivity of a function $f : \{0,1\}^n \to \{0,1\}$ to the changes in the input bits at a given point x. Note that given $x \in \{0,1\}^n$ and $i \in [n]$, $x \oplus e_i$ corresponds to flipping the i-th bit of x.

Definition 7.9 (Sensitivity)**.** The *sensitivity* of $f : \{0,1\}^n \to \{0,1\}$ at a point x, denoted by $s_f(x)$, is the number of bits in x such that flipping any one of these bits changes the value of the function. More formally,

$$s_f(x) \coloneqq |\{i \in [n] : f(x) \neq f(x \oplus e_i)\}|.$$

The *sensitivity* of f is

$$s(f) \coloneqq \max_{x \in \{0,1\}^n} s_f(x).$$

The function f in Corollary 7.8 satisfies $s(f) = s_f(0,\ldots,0) = n$. Corollary 7.8 easily generalizes to the following lower bound on the degree.

Theorem 7.10 (Nisan and Szegedy [NS94])**.** *Every* $f : \{0,1\}^n \to \{0,1\}$ *satisfies* $\deg(f) \geqslant \sqrt{s(f)/2}$.

Proof. Let $y \in \{0,1\}^n$ be such that $s \coloneqq s_f(y) = s(f)$, and let $i_1,\ldots,i_s \in [n]$ be the sensitive coordinates at y. Without loss of generality, we may assume that $f(y) = 0$ since replacing f with $1 - f$ does not change the degree or the sensitivity.

Define $g : \{0,1\}^s \to \{0,1\}$ as $g(z) \coloneqq f(y \oplus z_1 e_{i_1} \oplus \ldots \oplus z_s e_{i_s})$. The multilinear polynomial representation of g is given by substituting

$$x_i = \begin{cases} y_i & \text{if } i \notin \{i_1,\ldots,i_s\} \\ z_i & \text{if } i \in \{i_1,\ldots,i_s\} \text{ and } y_i = 0 \\ 1 - z_i & \text{if } i \in \{i_1,\ldots,i_s\} \text{ and } y_i = 1 \end{cases},$$

in the polynomial representation of $f(x_1,\ldots,x_n)$. Each x_i is a polynomial of degree 1 or 0 in z_i. Therefore, $\deg(g) \leqslant \deg(f)$.

On the other hand, g satisfies the assumption of Corollary 7.8, and therefore, $\deg(g) \geqslant \sqrt{s/2}$. □

7.5 Block Sensitivity

To relate the decision trees to the degree. we need to generalize Theorem 7.10 further. Given $B \subseteq [n]$, let $\mathbf{1}_B \in \{0,1\}^n$ denote the indicator vector of B.

Definition 7.11 (Block Sensitivity)**.** The *block sensitivity* of $f : \{0,1\}^n \to \{0,1\}$ at a point x, denoted by $\mathrm{bs}_f(x)$, is the maximum number of *disjoint* subsets $B_1,\ldots,B_k \subseteq [n]$ such that $f(x) \neq f(x \oplus \mathbf{1}_{B_i})$ for all $i = 1,\ldots,k$. The *block sensitivity* of f is

$$\mathrm{bs}(f) \coloneqq \max_{x \in \{0,1\}^n} \mathrm{bs}_f(x).$$

Note that $\mathrm{bs}(f) \geqslant s(f)$ since $s(f)$ corresponds to block sensitivity with blocks of size 1. Finally, we are ready to show that low-degree boolean functions have small decision tree complexity.

Theorem 7.12 (Nisan and Szegedy [NS94]). *Every $f : \{0,1\}^n \to \{0,1\}$ satisfies $\deg(f) \geqslant \sqrt{\mathrm{bs}(f)/2}$.*

Proof. The proof is almost identical to the proof of Theorem 7.10. Let $y \in \{0,1\}^n$ be such that $s \coloneqq \mathrm{bs}_f(y) = \mathrm{bs}(f)$, and let $B_1, \ldots, B_s \in [n]$ be disjoint sensitive blocks at y. Again, without loss of generality, we may assume that $f(y) = 0$ since replacing f with $1 - f$ does not change the degree or the block sensitivity.

The theorem follows by applying Corollary 7.8 to $g : \{0,1\}^s \to \{0,1\}$, defined as

$$g(z) \coloneqq f(y \oplus z_1 \mathbf{1}_{B_1} \oplus \ldots \oplus z_s \mathbf{1}_{B_s}).$$

□

Theorem 7.13 (Nisan [Nis91]). *Every function $f : \{0,1\}^n \to \{0,1\}$ satisfies*

$$C(f) \leqslant \mathrm{bs}(f) s(f).$$

Consequently,

$$\mathrm{dt}(f) \leqslant \mathrm{bs}(f)^4 \leqslant 2^8 \deg(f)^8.$$

Proof. First note that by Theorem 7.4, we have $\mathrm{dt}(f) \leqslant C(f)^2$, and by Theorem 7.12, we have $s(f) \leqslant \mathrm{bs}(f) \leqslant 4\deg(f)^2$. These facts verify the second inequality in the assertion. It remains to prove $C(f) \leqslant \mathrm{bs}(f)s(f)$.

Let x be any input, and consider a maximal set of disjoint blocks $B_1, \ldots, B_s \subseteq [n]$ such that $f(x) \neq f(x \oplus \mathbf{1}_{B_i})$ for all i. By the definition of block sensitivity $s \leqslant \mathrm{bs}(f)$.

Without loss of generality, we may assume that B_i is minimal, in the sense that $f(x) = f(x \oplus \mathbf{1}_{\widetilde{B}_i})$ for every $\widetilde{B}_i \subsetneq B_i$, as otherwise, we may replace B_i with $\widetilde{B}_i$.

Since B_i is minimal, for every $j \in B_i$, we have

$$f(x \oplus \mathbf{1}_{B_i \setminus \{j\}}) = f(x) \neq f(x \oplus \mathbf{1}_{B_i}),$$

and therefore, on the input $x \oplus \mathbf{1}_{B_i}$, f is sensitive to all the bits $j \in B_i$, and we must have $|B_i| \leqslant s(f)$. We conclude that $B \coloneqq \bigcup_{i=1}^{s} B_i$ is of size at most $\mathrm{bs}(f)s(f)$.

Recall that $B_1, \ldots, B_s \subseteq [n]$ is a *maximal* set of disjoint sensitive blocks at x. The maximality implies that B is a certificate for x and fixing the variables in B to $x|_B$ must determine the value of f to be $f(x)$; otherwise, we could find another sensitive block for x that is disjoint from B.

□

Theorem 7.13 and Proposition 7.2 show that $\deg(f)$ and $\mathrm{dt}(f)$ are polynomially equivalent

$$\deg(f) \leqslant \mathrm{dt}(f) \leqslant 2^8 \deg(f)^8.$$

7.6 Approximate degree and randomized decision trees

A *randomized decision tree* $\mathbf{T}$ of depth d is a random variable that takes values in the set of decision trees of depth at most d.

The *randomized decision tree complexity* of $f : \{0,1\}^n \to \{0,1\}$, denoted by $\mathrm{rdt}(f)$, is the smallest d such that there exists a randomized decision tree $\mathbf{T}$ of depth d with

$$\Pr_{\mathbf{T}}[\mathbf{T}(x) \neq f(x)] \leqslant \frac{1}{3} \qquad \forall x \in \{0,1\}^n . \tag{7.1}$$

Consider such a randomized decision tree, and let μ be the distribution of $\mathbf{T}$. Define $g : \{0,1\}^n \to \mathbb{R}$ as

$$g(x) \coloneqq \Pr_{\mathbf{T}}[\mathbf{T}(x) = 1] = \mathbb{E}_{\mathbf{T}} \mathbf{T}(x) = \sum_{T} \mu(T) T(x),$$

where the sum is over all decision trees of depth at most d. Since each $T(x)$ is a polynomial of degree at most d, we have $\deg(g) \leqslant d$. On the other hand, by Eq. (7.1), we have $\|f - g\|_\infty \leqslant \frac{1}{3}$.

Definition 7.14 (Approximate degree)**.** The *approximate degree* of f, denoted by $\widetilde{\deg}(f)$, is the smallest d such that there exists $g : \{0,1\}^n \to \mathbb{R}$ with $\deg(g) \leqslant d$ and $\|f - g\|_\infty \leqslant \frac{1}{3}$.

The discussion above shows that the approximate degree is a lower bound on the randomized decision tree complexity.

Proposition 7.15. *Every $f : \{0,1\}^n \to \{0,1\}$ satisfies $\widetilde{\deg}(f) \leqslant \mathrm{rdt}(f)$.*

The proof of Theorem 7.12 easily extends to the approximate degree and shows that the block sensitivity also provides a strong lower bound for this parameter.

Theorem 7.16 (Nisan and Szegedy [NS94])**.** *Every $f : \{0,1\}^n \to \{0,1\}$ satisfies $\widetilde{\deg}(f) \geqslant \sqrt{\mathrm{bs}(f)/6}$.*

Proof. Recall Theorem 7.5 about univariate polynomials. If we replace Theorem 7.5 (i) with $q(0) \leqslant \frac{1}{3}$ and $q(1) \geqslant \frac{2}{3}$, then there exists $x \in [0,1]$ with $|q'(x)| \geqslant \frac{1}{6}$. Now, the application of Markov's inequality shows $\deg(q) \geqslant \sqrt{m/6}$. Using this lower bound in the proof of Theorem 7.12 yields the desired result. □

Combining these facts with the bound $\mathrm{dt}(f) \leqslant \mathrm{bs}(f)^4$ from Theorem 7.13, we conclude

$$\frac{\mathrm{dt}(f)^{1/8}}{\sqrt{6}} \leqslant \sqrt{\mathrm{bs}(f)/6} \leqslant \widetilde{\deg}(f) \leqslant \mathrm{rdt}(f) \leqslant \mathrm{dt}(f).$$

Therefore, randomized decision tree complexity is polynomially equivalent to the deterministic decision tree complexity! Similarly, the approximate real degree is polynomially equivalent to the real degree!

7.7 Conclusion

In this chapter, we proved that several complexity measures of boolean functions $f : \{0,1\}^n \to \{0,1\}$ are polynomially equivalent:

- real degree $\deg(f)$;
- approximate real degree $\widetilde{\deg}(f)$;
- decision tree complexity $\mathrm{dt}(f)$;
- randomized decision tree complexity $\mathrm{rdt}(f)$;
- certificate complexity $C(f)$;
- block sensitivity $\mathrm{bs}(f)$.

We also discussed the sensitivity of f and showed that $s(f) \leqslant \mathrm{bs}(f)$. In [NS94], Nisan and Szegedy conjectured that $s(f)$ is also polynomially equivalent to the abovementioned parameters. Their conjecture remained open for more than three decades. Scott Aaronson, in his blog, writes:

> Ever since it was posed by Nisan and Szegedy in 1989, this conjecture has stood as one of the most frustrating and embarrassing open problems in all of combinatorics and theoretical computer science. It seemed so easy, and so similar to other statements that had 5-line proofs. But a lot of the best people in the field sank months into trying to prove it.

Finally, in 2019, Hao Huang [Hua19] proved the longstanding sensitivity conjecture in a short and beautiful paper. We will discuss Huang's proof in the next chapter.

7.8 Exercises

Exercise 7.1. Prove that every $f : \{0,1\}^n \to \{0,1\}$ satisfies

$$\mathrm{dt}(f) \leqslant C(f)\deg(f) \leqslant 16\deg(f)^5.$$

Exercise 7.2. Consider the stronger model of a *parity* decision tree, where at every internal node of the decision tree, we can query $\bigoplus_{i\in S} x_i$ for any $S \subseteq [n]$, and branch left or right accordingly. Let $\mathrm{dt}^{\oplus}(f)$ and $\mathrm{rdt}^{\oplus}(f)$ denote the deterministic and randomized decision tree complexities of $f : \{0,1\}^n \to \{0,1\}$, respectively. Note $\mathrm{dt}^{\oplus}(f) \leqslant \mathrm{dt}(f)$ and $\mathrm{rdt}^{\oplus}(f) \leqslant \mathrm{rdt}(f)$.

Given any $a \in \{0,1\}^n$, let the point mass $\mathbf{1}_a : \{0,1\}^n \to \{0,1\}$ be defined as $\mathbf{1}_a(x) = 1$ iff $x = a$.

1. Prove that $\mathrm{dt}^{\oplus}(\mathbf{1}_a) = n$ for any $a \in \{0,1\}^n$.
2. Prove that $\mathrm{rdt}^{\oplus}(\mathbf{1}_a) \leqslant 10$ for any $a \in \{0,1\}^n$. .

Exercise 7.3. Consider the stronger model of an AND-decision tree, where at every internal node of the decision tree, we can query $\bigwedge_{i\in S} x_i$ for any $S \subseteq [n]$, and branch left or right accordingly. Let $\mathrm{dt}^{\wedge}(f)$ and $\mathrm{rdt}^{\wedge}(f)$ denote the deterministic and randomized AND-decision tree complexities of $f : \{0,1\}^n \to \{0,1\}$, respectively.

Let $t : \{0,1\}^n \to \{0,1\}$ be defined as $t(x) = 1$ iff $\sum x_i \leqslant n-1$.

1. Prove that $\mathrm{dt}^{\wedge}(t) = n$.
2. Prove that $\mathrm{rdt}^{\wedge}(t) \leqslant 10$.

Chapter 8

The sensitivity theorem

This chapter presents Huang's elegant and short proof of the longstanding sensitivity conjecture, showing that sensitivity is polynomially equivalent to the real degree.

In Chapter 7, we proved $s(f) \leqslant \mathrm{bs}(f) \leqslant 2\deg(f)^2$. Conversely, the sensitivity theorem provides a lower bound on the sensitivity in terms of the degree.

Theorem 8.1 (Sensitivity Theorem [Hua19])**.** *Every* $f : \{0,1\}^n \to \{0,1\}$ *satisfies* $s(f) \geqslant \sqrt{\deg(f)}$.

The proof of Theorem 8.1 uses spectral techniques to show that every large induced subgraph of the hypercube contains a vertex of large (graph) degree. Before presenting proof of the sensitivity conjecture, let us recall the definition of the hypercube and state a few simple facts about it.

8.1 The hypercube graph

Definition 8.2 (hypercube)**.** For $n \in \mathbb{N}$, the n-dimensional hypercube Q_n is the undirected graph with vertex set $\{0,1\}^n$, where two vertices are connected by an edge iff they differ by exactly one bit. See Figure 8.1.

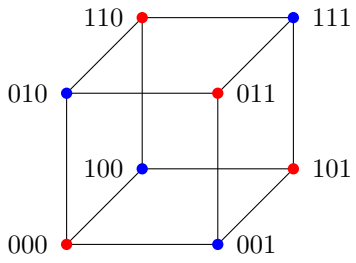

Figure 8.1: The hypercube Q_3. The colours red and blue represent the bipartition of Q_3.

The hypercube Q_n is an n-regular graph as every vertex $x \in \{0,1\}^n$ has exactly n neighbours $x \oplus e_1, \ldots, x \oplus e_n$. Note also that hypercube is a *bipartite* graph with the bipartition

$$V_{\text{even}} \coloneqq \left\{ x \in \{0,1\}^n \ : \ \sum_{i=1}^{n} x_i \equiv 0 \bmod 2 \right\} \quad \text{and} \quad V_{\text{odd}} \coloneqq \left\{ x \in \{0,1\}^n \ : \ \sum_{i=1}^{n} x_i \equiv 1 \bmod 2 \right\}.$$

We can alternatively construct Q_n by taking two disjoint copies Q_{n-1} and adding a perfect matching connecting each vertex in one copy of Q_{n-1} to the corresponding vertex in the other copy. Consequently, the following recursive formula describes the adjacency matrix A_n of Q_n.

$$A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad A_n = \begin{bmatrix} A_{n-1} & \mathbf{I} \\ \mathbf{I} & A_{n-1} \end{bmatrix}. \tag{8.1}$$

8.2 Two theorems from matrix theory

For a symmetric matrix $A \in \mathbb{R}^{m\times m}$, we denote the i-th largest eigenvalue of A by $\lambda_i(A)$ so that

$$\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \geqslant \lambda_m(A).$$

The following well-known theorem bounds $\max_{i=1}^m |\lambda_i|$ by the maximum L_1-norm of the rows of A.

Theorem 8.3. *Every eigenvalue λ of a symmetric matrix $A \in \mathbb{R}^{m\times m}$ satisfies*

$$|\lambda| \leqslant \max_i \sum_{j=1}^m |A_{ij}|.$$

Proof. Let $u = (u_1, \ldots, u_n)$ be an eigenvector corresponding to λ so that $Au = \lambda u$. Let $i^* = \arg\max_i |u_i|$. We have

$$|\lambda u_{i^*}| = |(Au)_{i^*}| = \sum_{j=1}^m A_{i^*j}u_j \leqslant \left(\max_j |u_j|\right)\left(\sum_{j=1}^m |A_{i^*j}|\right) = |u_{i^*}|\left(\sum_{j=1}^m |A_{i^*j}|\right),$$

which shows

$$|\lambda| \leqslant \sum_{j=1}^m |A_{i^*j}| \leqslant \max_i \sum_{j=1}^m |A_{ij}|.$$

□

Finally, let us recall the Cauchy interlacing theorem, a useful fact from matrix theory.

Theorem 8.4 (Cauchy interlacing theorem)**.** *Let $A \in \mathbb{R}^{m\times m}$ be a symmetric matrix, and let $B \in \mathbb{R}^{k\times k}$ be a principal submatrix of A. Then denoting $r \coloneqq m - k$, we have*

$$\lambda_i(A) \geqslant \lambda_i(B) \geqslant \lambda_{i+r}(A) \quad \text{for every } i = 1, \ldots, k.$$

8.3 Proof of the sensitivity theorem

Given an undirected graph $G = (V, E)$ and a subset $T \subseteq V$, let $G[T]$ denote the *subgraph induced* by G on T. Denote the largest degree of a vertex in an undirected graph G by $\Delta(G)$. The following theorem lies at the core of Huang's proof of the sensitivity conjecture.

Theorem 8.5 (Huang [Hua19])**.** *For every $T \subseteq \{0,1\}^n$ with $|T| > 2^{n-1}$, the subgraph $H \coloneqq Q_n[T]$ induced by the hypercube Q_n on T satisfies*

$$\Delta(H) \geqslant \sqrt{n}.$$

Proof. Define the $2^n \times 2^n$ symmetric matrices $\widetilde{A}_n$ recursively as

$$\widetilde{A}_1 \coloneqq \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \widetilde{A}_n \coloneqq \begin{bmatrix} \widetilde{A}_{n-1} & \mathbf{I} \\ \mathbf{I} & -\widetilde{A}_{n-1} \end{bmatrix}. \tag{8.2}$$

By comparing these formulas to Eq. (8.1), note that

$$|\widetilde{A}_n(x,y)| = A_n(x,y) \text{ for all } x, y \in \{0,1\}^n,$$

where A_n is the adjacency matrix of the hypercube Q_n.

Let B denote the $T \times T$ principal submatrix of $\widetilde{A}_n$. By Theorem 8.3, we have

$$\lambda_1(B) \leqslant \max_{x\in T} \sum_{y\in T} |B(x,y)| = \max_{x\in T} \sum_{y\in T} |\widetilde{A}_n(x,y)| = \max_{x\in T} \sum_{y\in T} A_n(x,y) \leqslant \Delta(H).$$

Next, we wish to determine the eigenvalues of $\widetilde{A}_n$. An easy induction shows that $\widetilde{A}_n^2 = n\mathbf{I}_{2^n}$, and therefore, all the eigenvalues of $\widetilde{A}_n$ are of the form $\pm\sqrt{n}$. On the other hand, since $\mathrm{Tr}(\widetilde{A}_n) = 0$, each of $\sqrt{n}$ and $-\sqrt{n}$ have multiplicity 2^{n-1}.

Since $|T| > 2^{n-1}$, by the Cauchy interlacing theorem, we have

$$\lambda_1(B) \geqslant \lambda_{1+2^n-|T|}(\widetilde{A}_n) \geqslant \lambda_{2^{n-1}}(\widetilde{A}_n) = \sqrt{n}.$$

□

The sensitivity conjecture easily follows from Theorem 8.5, through the following reductions observed by [GL92].

Proof of Theorem 8.1. Without loss of generality, we may assume that $\deg(f) = n$. Otherwise, pick any monomial of maximum degree with a non-zero coefficient in the polynomial representation of f and assign 0 to the variables *not* involved in this monomial. The restricted function has the same degree as f and, by induction, has sensitivity at least $\sqrt{\deg(f)}$, and consequently $s(f) \geqslant \sqrt{\deg(f)}$. We will assume $\deg(f) = n$ in the sequel.

We will change the range of f to ± 1. Define $g : \mathbb{Z}_2^n \to \{-1, 1\}$ as $g(x) \coloneqq 1 - 2f(x)$ and note $\deg(g) = \deg(f) = n$ and $s(g) = s(f)$. Therefore, it suffices to prove $s(g) \geqslant \sqrt{n}$.

Since $\deg(g) = n$, we have $\widehat{g}([n]) \neq 0$ in the Fourier expansion of g:

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

Define $h : \mathbb{Z}_2^n \to \{-1, 1\}$ as $h(x) \coloneqq g(x)\chi_{[n]}(x) = g(x)(-1)^{\sum_{i=1}^n x_i}$. We make the following observations.

(i) We have $h(x) = \sum_{S \subseteq [n]} \widehat{g}([n] \setminus S)\chi_S(x)$, and therefore $\mathbb{E}[h] = \widehat{g}([n]) \neq 0$.

(ii) We have $g(x) \neq g(x + e_i) \Leftrightarrow h(x) = h(x + e_i)$, and therefore,

$$s_g(x) = |\{i : h(x) = h(x + e_i)\}|.$$

In other words, $s_g(x)$ is the (graph) degree of x in $Q_n[T]$ where $T = \{y : h(y) = h(x)\}$.

Let T be the larger of the two sets $T^+ = h^{-1}(1)$ and $T^- = h^{-1}(-1)$. By (i), these two sets are not of equal size and therefore $|T| > 2^{n-1}$. By Theorem 8.5, the largest vertex degree in $Q_n[T]$ is at least $\sqrt{n}$, which by (ii), shows $s(g) \geqslant \sqrt{n}$. □

Chapter 9

Influences, Isoperimetry, and Efron-Stein inequality

This chapter introduces another fundamental concept in studying Boolean functions: the notion of *influence*.

Let (X, μ) be a probability space. Given a function $f : (X, \mu) \to \mathbb{R}$, the variance of f provides a natural measure of how sensitive the output of f is to changes in the input. For instance, if the variance of f is zero, f is constant, meaning that the input does not influence the value of $f(x)$. Conversely, a larger variance implies that the output of f varies more significantly, and therefore, the input has a greater influence on $f(x)$.

More generally, consider a function $f : (X^n, \mu^n) \to \mathbb{R}$. We wish to measure the influence of a single variable x_i on $f(x_1, \ldots, x_n)$. Fixing $x_{[n]\setminus\{i\}} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ in $f(x_1, \ldots, x_n) = f(x_{[n]\setminus\{i\}}, x_i)$, reduces it to a function of the single variable x_i. We can then apply the variance to quantify the influence of x_i. Finally, by taking the expected value of this variance, we obtain a natural way to measure the overall influence of the individual variable x_i on $f(x_1, \ldots, x_n)$.

Definition 9.1 (Influence)**.** Let (X, μ) be a probability space, and let $f : (X^n, \mu^n) \to \mathbb{R}$. The *influence* of the ith variable on f is defined as

$$I_i(f) \coloneqq \mathbb{E}_{\mathbf{x}_{[n]\setminus\{i\}} \sim \mu^{n-1}} \operatorname*{Var}_{\mathbf{x}_i \sim \mu}[f(\mathbf{x}_{[n]\setminus\{i\}}, \mathbf{x}_i)].$$

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

$$I_f = \sum_{i=1}^{n} I_i(f).$$

One can express the influences in terms of the following notion of Laplacian.

Definition 9.2 (ith coordinate Laplacian)**.** Let (X, μ) be a probability space. The ith coordinate Laplacian of the function $f : (X, \mu)^n \to \mathbb{R}$ is $\partial_i f \coloneqq f - \mathbb{E}_{\mathbf{x}_i} f$.

Note that we have

$$I_i(f) = \mathbb{E}_{\mathbf{x}_{[n]\setminus\{i\}}} \mathbb{E}_{\mathbf{x}_i} (f - \mathbb{E}_{\mathbf{x}_i} f)^2 = \|\partial_i f\|_2^2.$$

Uniform measure on $\{0,1\}^n$: When $f : \{0,1\}^n \to \mathbb{R}$, where $\{0,1\}^n$ is endowed with the uniform probability measure, we have

$$\partial_i f(x) = f(x) - \mathbb{E}_{\mathbf{x}_i} f(x) = f(x) - \frac{f(x) + f(x \oplus e_i)}{2} = \frac{f(x) - f(x + e_i)}{2}.$$

If we further assume that $f : \{0,1\}^n \to \{0,1\}$ is Boolean valued, then

$$I_i(f) = \frac{1}{4} \mathbf{Pr}_{\mathbf{x}}[f(\mathbf{x}) \neq f(\mathbf{x} \oplus e_i)],$$

and

Example 9.3. For the parity function

$$\textsc{Parity} : x \mapsto x_1 + \ldots + x_n (\text{mod } 2),$$

we have $I_i(\textsc{Parity}) = \frac{1}{4}$ for all i, and $I_{\textsc{Parity}} = n/4$.

9.1 Sensitivity and influences

Recall that we defined the sensitivity of f as $s(f) = \max_x s_f(x)$. It is also natural to consider the *average sensitivity* of f defined as $s^{\text{avg}}(f) \coloneqq \mathbb{E}_{\mathbf{x}} s_f(\mathbf{x})$. The following simple observation connects the total influence to average sensitivity.

Proposition 9.4. *Every $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfies*

$$I_f = \frac{s^{\text{avg}}(f)}{4}.$$

where $s^{\text{avg}}(f) \coloneqq \mathbb{E}_{\mathbf{x}} s_f(\mathbf{x})$.

Proof. We have

$$I_f = \frac{1}{4}\sum_{i=1}^{n} \mathbf{Pr}_{\mathbf{x}}[f(\mathbf{x}) \neq f(\mathbf{x}+e_i)] = \frac{1}{4}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}} \mathbf{1}_{[f(\mathbf{x})\neq f(\mathbf{x}+e_i)]} = \frac{1}{4}\mathbb{E}_{\mathbf{x}} \sum_{i=1}^{n} \mathbf{1}_{[f(\mathbf{x})\neq f(\mathbf{x}+e_i)]} = \frac{1}{4}\mathbb{E}_{\mathbf{x}} s_f(\mathbf{x}).$$

□

9.2 Isoperimetric Inequalities for the Hypercube

Isoperimetric problems in mathematics ask for the minimum possible "boundary size" of a set with a given "size," where the precise definitions of these terms depend on the specific problem. The classic example is minimizing the perimeter among all shapes in the plane with area 1. The solution to this problem – that the circle is optimal – was known to the Ancient Greeks, but the first rigorous proof was found by Karl Weierstrass in the 1870s.

In the discrete setting, isoperimetric inequalities are studied within graphs, where the goal is to find the minimum edge or vertex boundary among subsets of vertices of a given size.

Definition 9.5 (Edge and vertex boundary)**.** Let $G = (V, E)$ be an undirected graph, and let $A \subseteq V$. The *edge boundary* of A is the set $E(A, A^c)$ of edges with one endpoint is A and one in $A^c = V \setminus A$. The *vertex boundary* of A is the set

$$\partial A = \{v \in V \setminus A \ : \ v \text{ has a neighbour in } A\}.$$

Recall from Definition 8.2 that the hypercube Q_n is the undirected graph with vertex set $\mathbb{Z}_2^n$ where two vertices are adjacent if they differ in exactly one coordinate. The following proposition shows that the size of the edge boundary in a subset of the hypercube is equivalent to its total influence.

Observation 9.6. *Give a set of vertices $A \subseteq V(Q_n) = \mathbb{Z}_2^n$ in the hypercube, let $f = \mathbf{1}_A$. We have*

$$I_f = \frac{|E(A, A^c)|}{2^{n+1}}.$$

Proof. We have

$$|E(A, A^c)| = \sum_{x\in A} s_f(x) = \sum_{x\in A^c} s_f(x),$$

and therefore, by Proposition 9.4,

$$2|E(A, A^c)| = \sum_{x\in\mathbb{Z}_2^n} s_f(x) = 2^n s^{\text{avg}}(f) = 2^{n+2} I_f.$$

□

The edge isoperimetric inequality on the hypercube states that subcubes have the smallest edge boundaries.

Theorem 9.7 (Harper's edge isoperimetric inequality)**.** *Every subset A of the vertices of the hypercube Q_n satisfies*

$$|E(A, A^c)| \geqslant |A| \log \frac{2^n}{|A|},$$

with equality when A is a subcube. Equivalently, every $f : \mathbb{Z}_2^n \to \{0, 1\}$ satisfies

$$I_f \geqslant \frac{1}{2}\mathbb{E}[f] \log \frac{1}{\mathbb{E}[f]}. \tag{9.1}$$

Proof. The proof is straightforward using induction on n. The base case, $n = 1$, is easily verified, so we directly move to the induction step. Partition Q_n into two disjoint subcubes Q^1_{n-1} and Q^2_{n-1} of dimension $n - 1$ each. Similarly partition A into two sets $A_1 = A \cap V(Q^1_{n-1})$ and $A_2 = A \cap V(Q^2_{n-1})$. Let $a_1 = |A_1|$ and $a_2 = |A_2|$. Without loss of generality, assume $a_1 = a_2 + t$ for $t \geqslant 0$. The edge boundary of S will have edges from the boundary of A_1 in A^1_{n-1}, edges from the boundary of A_2 in A^2_{n-1}, and also at least t edges that must go between the two subcubes. Using the induction hypothesis, we have

$$\begin{aligned}
|E(A, A^c)| &\geqslant a_1(n - 1 - \log a_1) + a_2(n - 1 - \log a_2) + t \\
&= a_1 n - a_1 - a_1 \log a_1 + a_2 n - a_2 - a_2 \log a_2 + t \\
&= a_1 n - a_1 \log a_1 + a_2 n - a_2 \log a_2 - 2a_2.
\end{aligned}$$

Since $|A| \log \frac{2^n}{|A|} = a_1 n + a_2 n - (a_1 + a_2) \log(a_1 + a_2)$, it suffices to show

$$a_1 \log a_1 + a_2 \log a_2 + 2a_2 \leqslant (a_1 + a_2) \log(a_1 + a_2),$$

for $a_2 \leqslant a_1$. This inequality can be easily proved using simple manipulations. □

On the other hand, Harper's theorem [Har66] states that Hamming balls have the smallest vertex boundaries. Given $0 \leqslant r \leqslant n$ and $x \in \mathbb{Z}_2$, let $B_r(x)$ denote the Hamming ball of radius r centered at x. Note that $|B_r(x)| = \sum_{i=0}^{r} \binom{n}{r} = \binom{n}{\leqslant r}$.

Theorem 9.8 (Harper's vertex isoperimetric theorem)**.** *Every set A of vertices in the hypercube Q_n with $\binom{n}{\leqslant r} \leqslant |A| < \binom{n}{\leqslant r+1}$ satisfies $|A \cup \partial A| \geqslant \binom{n}{\leqslant r+1}$.*

We will refer the reader to [FF81] for a short proof of Harper's vertex isoperimetric theorem.

9.3 Fourier expansion and Influences

Let $f : \mathbb{Z}_2^n \to \mathbb{R}$. One can easily describe the Fourier expansion of $\partial_i f$ from the Fourier expansion of f. Since the characters satisfy

$$\chi_S(x + e_i) = \begin{cases} -\chi_S(x) & \text{if } i \in S \\ \chi_S(x) & \text{if } i \notin S \end{cases},$$

we have[1]

$$\partial_i f = \sum_{S : i \in S} \widehat{f}(S)\chi_S. \tag{9.2}$$

Equation (9.2) allows us to express the influences and the total influence in terms of the Fourier coefficients.

Proposition 9.9. *For every $f : \mathbb{Z}_2^n \to \mathbb{R}$, we have*

$$I_i(f) = \|\partial_i f\|_2^2 = \sum_{S : i \in S} |\widehat{f}(S)|^2,$$

[1] Compare this formula to $\frac{\partial}{\partial x_i} f = \sum_{S : i \in S} a_S \prod_{j \in S \setminus \{i\}} x_j$, which holds for multilinear polynomials $f(x_1, \ldots, x_n) = \sum_{S \subseteq [n]} a_S \prod_{i \in S} x_i$.

and consequently,

$$I_f = \sum_{S \subseteq [n]} |S| |\widehat{f}(S)|^2.$$

Since $\mathrm{Var}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = \sum_{S:S\neq\emptyset} |\widehat{f}(S)|^2$, Proposition 9.9 immediately implies the following Poincaré inequality

$$I_f \geqslant \sum_{S\neq\emptyset} |\widehat{f}(S)|^2 = \mathrm{Var}[f].$$

From its proof, it is apparent that this inequality is only tight for Boolean functions of degree at most 1, namely, the constant functions and the dictators. Note that Harper's edge isoperimetric inequality Equation (9.1) provides a stronger lower bound for I_f in terms of $\mathbb{E}[f]$.

9.4 General product spaces

Consider the Fourier expansion of a function $f : \mathbb{Z}_2^n \to \mathbb{R}$,

$$f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S. \tag{9.3}$$

Many key properties of this expansion stem from the product structure of $\mathbb{Z}_2^n$. For example, the property that $F_S = \widehat{f}(S)\chi_S$ is a S-junta since it only depends on the variables in S, is inherently about the product structure. In [Hoe48], Hoeffding introduced an analogous expansion for functions on arbitrary product spaces, which shares many of the key features of Eq. (9.3). Fourier-Walsh expansion often called the *Fourier-Walsh expansion*, is particularly useful for analyzing functions on domains such as $[0,1]^n$ where there is no group structure to define a Fourier expansion.

Let (X,μ) be a probability space on a finite set X, and let μ^n be the corresponding product measure on X^n. We only assume X is finite to avoid discussing measurability and integrability conditions. However, if we require that $f : X^n \to \mathbb{R}$ is integrable, the following discussion easily generalizes to infinite settings such as $[0,1]$ or the Gaussian measure on $\mathbb{R}$.

Notation: For $S \subseteq [n]$, we will sometimes denote the complement of S by $\overline{S} \coloneqq S^c = [n] \setminus S$. For $x \in X^n$ and $S \subseteq [n]$, let $x_S \in X^S$ denote the restriction of x to the coordinates in S. For disjoint sets $S, T \subseteq X$, and $y \in X^S$ and $z \in X^T$, let (y,z) denote the unique element in $X^{S\cup T}$ satisfying $(y,z)_S = y$ and $(y,z)_T = z$. In the sequel, by an abuse of notation, we sometimes identify a function $f : X^S \to \mathbb{R}$ with its corresponding S-junta on X^n defined as $x \mapsto f(x_S)$ for $x \in X^n$.

Given a function $f : X^n \to \mathbb{R}$ and $S \subseteq X$, we use the notation $\mathbb{E}_S f$ to denote the function $\mathbb{E}_S f : X^n \to \mathbb{R}$ with

$$(\mathbb{E}_S f)(y) = \mathbb{E}_{\mathbf{x}_S} f(\mathbf{x}_S, y_{\overline{S}}) = \int_{X^S} f(x_S, y_{\overline{S}}) d\mu^S(x_S).$$

For example,

$$(\mathbb{E}_i f)(y) = \mathbb{E}_{\mathbf{x}_i} f(y_1, \ldots, y_{i-1}, \mathbf{x}_i, y_{i+1}, \ldots, y_n) = \int_X f(x_i, y_{[n]\setminus\{i\}}) d\mu(x_i).$$

Note that $\mathbb{E}_S f$ is a $\overline{S}$-junta.

9.4.1 Hoeffding's Fourier-Walsh Expansion

Given a function $f : X^n \to \mathbb{R}$, we define the Fourier-Walsh expansion $f = \sum_{S\subseteq[n]} F_S$ based on two key properties of the functions F_S.

Definition 9.10 (Hoeffding [Hoe48])**.** The Fourier-Walsh expansion of $f : X^n \to \mathbb{R}$ is the unique decomposition $f = \sum_{S\subseteq[m]} F_S$ that satisfies the following two properties.

(i) For every $S \subseteq [n]$, the function F_S is an S-junta, meaning it depends only on the coordinates in S.

(ii) $\mathbb{E}_i F_S \equiv 0$ for every $S \subseteq [n]$ and $i \in S$.

The two properties (i) and (ii) uniquely determined the functions F_S from f. Notably, by (i) and (ii), we have

$$\mathbb{E}_T f = \sum_{S \subseteq \overline{T}} F_S.$$

For example, taking $T = [n]$ shows $F_\emptyset = \mathbb{E}[f]$, representing the mean of f similar to the principal Fourier coefficient. More generally, More generally, this yields the following inclusion-exclusion formula for F_S:

$$F_S = \sum_{T \subseteq S} (-1)^{|S \setminus T|} \mathbb{E}_{\overline{T}} f.$$

Remark 9.11. Note that for $f : \mathbb{Z}_2^n \to \mathbb{R}$, we have $F_S = \widehat{f}(S)\chi_S$ in the Hoeffding expansion $f = \sum_{S \subseteq [n]} F_S$. More generally, for $f : \mathbb{Z}_k^n \to \mathbb{R}$, we have $F_S = \sum_{a:\mathrm{supp}(a)=S} \widehat{f}(a)\chi_a$, as it satisfies (i) and (ii).

Proposition 9.12 (Orthogonality of Hoeffding terms)**.** *The Hoeffding expansion $f = \sum_S F_S$ of $f : X^n \to \mathbb{R}$ satisfies*

$$\langle F_S, F_T \rangle \coloneqq \mathbb{E}_{\mathbf{x}} F_S(\mathbf{x}) F_T(\mathbf{x}) = \begin{cases} 0 & \text{if } S \neq T \\ \|F_S\|_2^2 & \text{if } S = T \end{cases}.$$

In particular, we have Parseval's identity

$$\mathbb{E} f^2 = \sum_S \|F_S\|_2^2.$$

Proof. If $S \neq T$, there exists an element $i \in (S \setminus T) \cup (T \setminus S)$. Then by Definition 9.10 (ii), we have $\langle F_S, F_T \rangle = 0$. □

9.5 The Efron-Stein Inequality

As an application of Fourier-Walsh expansion, we will prove the Efron-Stein inequality [ES81], which extends the Poincaré inequality $I_f \geqslant \mathrm{Var}[f]$ to general product spaces.

Theorem 9.13 (Efron-Stein [ES81])**.** *Let (X, μ) be a probability space. Every integrable function $f : X^n \to \mathbb{R}$, satisfies*

$$\mathrm{Var}[f] \leqslant I_f.$$

Proof. Consider the Fourier-Walsh expansion $f = \sum_{S \subseteq [n]} F_S$. By Definition 9.10 (i) and (ii), we have

$$\mathbb{E}_i f = \sum_{S : i \notin S} F_S \text{ and } \partial_i f = f - \mathbb{E}_i f = \sum_{S : i \in S} F_S.$$

In particular, by Parseval,

$$I_i(f) = \|\partial_i f\|_2^2 = \sum_{S : i \in S} \|F_S\|_2^2,$$

which summing over all i gives

$$I_f = \sum_{i=1}^n I_i(f) = \sum_{S \subseteq [n]} |S| \|F_S\|_2^2.$$

On the other hand,

$$\mathrm{Var}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = \sum_{S \neq \emptyset} \|F_S\|_2^2.$$

□

9.6 Fourier levels

The discussions in this chapter, particularly the formula for the total influence and the proof of the Efron-Stein inequality, suggest a decomposition of functions over product spaces into different levels.

Given a function $f : X^n \to \mathbb{R}$ represented by its Hoeffding expansion $f = \sum_{S\subseteq[n]} F_S$, we define the *k-th level* projection of f as

$$f^{=k} := \sum_{S:|S|=k} F_S.$$

Note that for $f : \mathbb{Z}_2^n \to \mathbb{R}$, the k-th level is given by

$$f^{=k} := \sum_{S:|S|=k} \widehat{f}(S)\chi_S,$$

and our formula for the total influences, given in Proposition 9.9, translates to

$$I_f = \sum_{k=0}^{n} k\|f^{=k}\|_2^2.$$

When considering the characters χ_S of $\mathbb{Z}_2^n$ it is helpful to think of $|S|$ as an analogue of the *frequency* $\frac{a}{N}$ for characters $\chi_a(x) = e^{2\pi axi/N}$ of $\mathbb{Z}_{\mathbb{N}}$. As the frequency $\frac{a}{N}$ increases, $\chi_a(x)$ becomes more sensitive to the changes in x (see Figure 9.1). Similarly, larger $|S|$ indicates a higher sensitivity to the input bits for χ_S.

Finally, we also introduce the notations

$$f^{\leqslant k} := \sum_{S:|S|\leqslant k} \widehat{f}(S)\chi_S,$$

and

$$f^{\geqslant k} := \sum_{S:|S|\geqslant k} \widehat{f}(S)\chi_S.$$

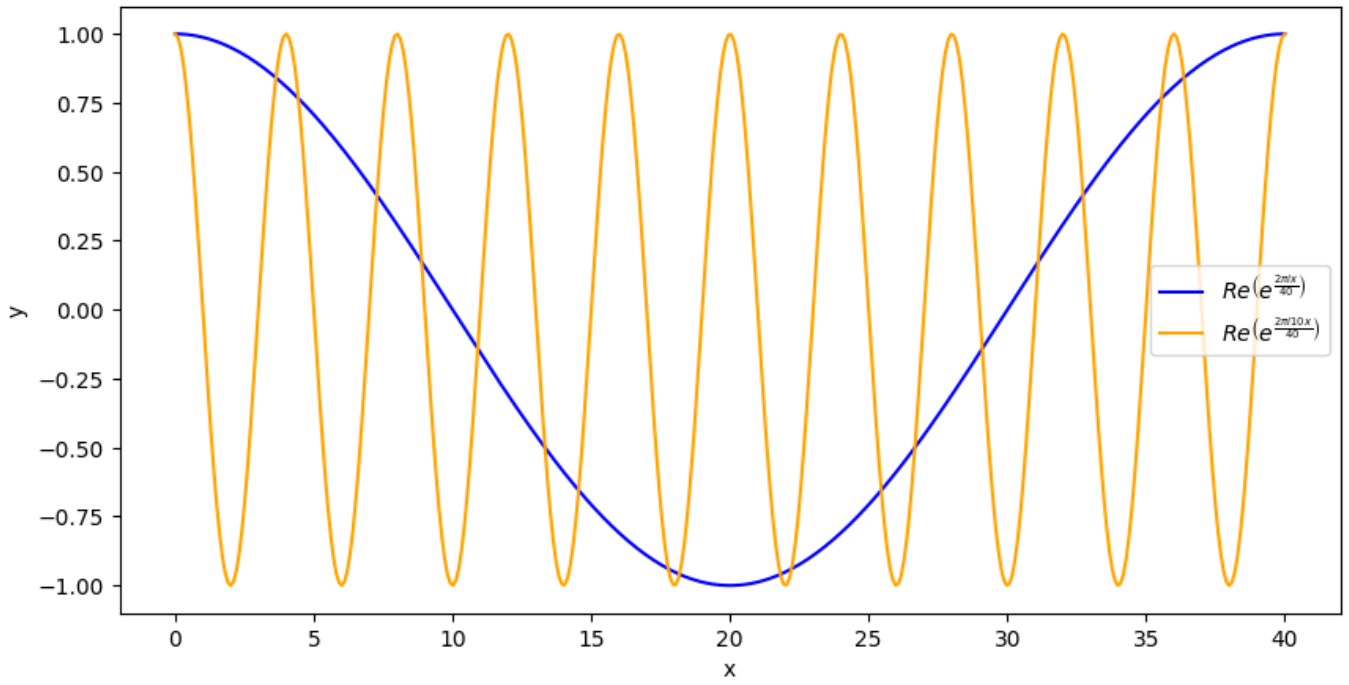

Figure 9.1: Blue illustrates $\Re(e^{2\pi i\frac{x}{40}})$ which has frequency $\frac{1}{40}$. Red illustrates $\Re(e^{2\pi i\frac{10x}{40}})$ which has frequency $\frac{1}{4}$. The character with a smaller frequency is more stable under small changes in x.

Chapter 10

Introduction to hypercontractivity

This chapter explores an important phenomenon in functional analysis known as hypercontractivity. For functions defined on the discrete cube $\{0,1\}^n$, hypercontractivity was first proved by Bonami [Bon70] and later developed further by Beckner [Bec75] and Gross [Gro75].

Consider the vector space of functions $f : \mathbb{Z}_2^n \to \mathbb{R}$ and recall that $\|f\|_p \leqslant \|f\|_q$ if $1 \leqslant p \leqslant q \leqslant \infty$. The difference in L_p norms is closely connected to the concentration of $|f|$. If f is a constant function, all the L_p norms coincide. More generally, if f is nearly constant, we expect that $\|f\|_p$ does not increase significantly as p increases. Conversely, moment concentration inequalities (e.g. Chebyshev's inequality) suggest that when two distinct L_p norms are close, then $|f|$ is somewhat concentrated[1].

To "smooth" a function $f : \mathbb{Z}_2^n \to \mathbb{R}$ and increase its concentration, we can average f over a neighbourhood of each point x such as a Hamming ball of a fixed radius centred at x. More specifically, we will consider a smoothing operator that averages f over noisy versions of the inputs x. Define $Tf(x) \coloneqq \mathbb{E}_{\mathbf{z}} f(\mathbf{z})$ where $\mathbf{z}$ is obtained from x by independently flipping each bit of x with probability γ for some fixed $\gamma \in [0,1]$. Since T is a stochastic operator, the convexity of norms implies that it is contractive under the L_p norms, i.e., $\|Tf\|_p \leqslant \|f\|_p$. However, the smoothing effect of the averaging operator allows us to make an even stronger assertion: The operator T is *hypercontractive*, meaning that $\|Tf\|_q \leqslant \|f\|_p$ for some $q > p$ depending on γ.

10.1 Hypercontractivity in dimension one

We introduce the noise operator in dimension one, i.e. on the space of functions $f : \mathbb{Z}_2 \to \mathbb{R}$. Let μ_γ denote the Bernoulli distribution with success probability γ (i.e., $\mu_\gamma(1) = \gamma$ and $\mu_\gamma(0) = 1-\gamma$).

Definition 10.1 (The 1-dimensional noise operator)**.** Let $0 \leqslant \rho \leqslant 1$ be a parameter, and let $\gamma \coloneqq \frac{1}{2}(1-\rho)$. Given a function $f : \mathbb{Z}_2 \to \mathbb{R}$, define $T_\rho f : \mathbb{Z}_2 \to \mathbb{R}$ by

$$T_\rho f(x) \coloneqq \mathbb{E}_{\mathbf{y} \sim \mu_\gamma} f(x + \mathbf{y}).$$

Remark 10.2. In Definition 10.1, the value $p = \frac{1}{2}(1-\rho)$ is chosen so that $\mathbb{E}_{\mathbf{y}\sim\mu_\gamma}(-1)^{\mathbf{y}} = \rho$.

Remark 10.3. We have $T_\rho f(x) = \mathbb{E}_{\mathbf{z}}[f(\mathbf{z})]$, where the random variable $\mathbf{z}$ is a noisy copy of x with corruption probability p. More formally,

$$\mathbf{z} \coloneqq \begin{cases} x & \text{with probability } 1-\gamma \\ 1-x & \text{with probability } \gamma \end{cases}. \tag{10.1}$$

The operator T_ρ is linear

$$T_\rho(f + \lambda g) = T_\rho f + \lambda T_\rho g,$$

and since

$$\begin{bmatrix} T\rho f(0) \\ T\rho f(1) \end{bmatrix} = \begin{bmatrix} (1-\gamma)f(0) + \gamma f(1) \\ (1-\gamma)f(1) + \gamma f(0) \end{bmatrix} = \begin{bmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \end{bmatrix},$$

[1] Since the L_p norm of f and $|f|$ are equal, the values of $\|f\|_p$ can only imply concentration for $|f|$.

its corresponding matrix is

$$T_\rho = \begin{bmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{bmatrix}. \tag{10.2}$$

Since $\mathbb{E}_{\mathbf{y}\sim\mu_\gamma}\chi_1(x+\mathbf{y}) = \mathbb{E}_{\mathbf{y}\sim\mu_\gamma}(-1)^{x+\mathbf{y}} = \rho(-1)^x = \rho\chi_1(x)$, the two characters χ_0 and χ_1 of $\mathbb{Z}_2$ are the eigenvectors of T_ρ with corresponding eigenvalues 1 and ρ:

$$T_\rho\chi_0 = \chi_0 \text{ and } T_\rho\chi_1 = \rho\chi_1 \tag{10.3}$$

In particular, for every $f : \mathbb{Z}_2 \to \mathbb{R}$ with Fourier expansion $f = \widehat{f}(0) + \widehat{f}(1)\chi_1$, we have

$$T_\rho f = \widehat{f}(0) + \rho\widehat{f}(1)\chi_1.$$

Next, we show that since T_ρ is an averaging operator, it is a contraction for the L_p norm.

Theorem 10.4 (contractivity in dimension one)**.** *For every $f : \mathbb{Z}_2 \to \mathbb{R}$, we have*

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

Proof. Let $\mu \coloneqq \mu_{\frac{1}{2}(1-\rho)}$. Given $y \in \mathbb{Z}_2$, define the y-translation of f, $f_y : \mathbb{Z}_2 \to \mathbb{R}$ as $f_y(x) = f(x+y)$. By convexity of norms, we have

$$\|T_\rho f\|_p = \|\mathbb{E}_{\mathbf{y}\sim\mu} f_{\mathbf{y}}(x)\|_p \leqslant \mathbb{E}_{\mathbf{y}\sim\mu}\|f_{\mathbf{y}}(x)\|_p = \mathbb{E}_{\mathbf{y}\sim\mu}\|f\|_p = \|f\|_p .$$

□

Remark 10.5. The proof of Theorem 10.4 generalizes to any norm that satisfies $\|f\| = \|f_y\|$ for all y, i.e., any translation invariant norm.

The operator T_ρ satisfies a stronger property called *hypercontractivity*.

Theorem 10.6 (Hypercontractivity in dimension one)**.** *Let $1 < p \leqslant q < \infty$ and $f : \mathbb{Z}_2 \to \mathbb{R}$. We have*

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

Proof. Consider $f : \mathbb{Z}_2 \to \mathbb{R}$ and set $\gamma = \frac{1}{2}(1-\rho)$. Denoting $a = |f(0)|$ and $b = |f(1)|$, using standard methods from calculus, for $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$, we have

$$\begin{aligned}
\|T_\rho f\|_q &= \left(\mathbb{E}_{\mathbf{x}}\left|\mathbb{E}_{\mathbf{y}\sim\mu_\gamma} f(\mathbf{x}+\mathbf{y})\right|^q\right)^{1/q} \\
&= \left(\frac{1}{2}\big((1-\gamma)a + \gamma b\big)^q + \frac{1}{2}\big(\gamma a + (1-\gamma)b\big)^q\right)^{1/q} \\
&\leqslant \left(\frac{1}{2}a^p + \frac{1}{2}b^p\right)^{1/p} = \|f\|_p .
\end{aligned}$$

□

10.2 Hypercontactivity

In Section 10.1, we discussed the noise operator and hypercontractivity in dimension one. In this section, we will generalize these concepts to arbitrary dimensions.

Let μ_γ^n denote the product probability measure on $\{0,1\}^n$ defined by the Bernoulli measure μ_γ. More formally, for $y \in \{0,1\}^n$, we have $\mu_\gamma^n(y) = \gamma^{\sum y_i}(1-\gamma)^{n-\sum y_i}$.

Definition 10.7 (Noise operator in arbitrary dimension)**.** Let $0 \leqslant \rho \leqslant 1$ and set $\gamma = \frac{1}{2}(1-\rho)$. Given a function $f : \mathbb{Z}_2^n \to \mathbb{R}$, define $T_\rho f : \mathbb{Z}_2^n \to \mathbb{R}$ as

$$T_\rho f(x) \coloneqq \mathbb{E}_{\mathbf{y}\sim\mu_\gamma^n} f(x+\mathbf{y}).$$

We can write $T_\rho f(x) = \mathbb{E}_{\mathbf{z}}[f(\mathbf{z})]$, where

$$\mathbf{z}_i \coloneqq \begin{cases} x_i & \text{with probability } 1-\gamma \\ 1-x_i & \text{with probability } \gamma \end{cases}, \tag{10.4}$$

independently for each i. In other words, $\mathbf{z}$ is a noisy copy of x, where each coordinate is flipped with probability γ. Similar to the one dimension, T_ρ is a linear operator. We have

$$T_\rho(f + \lambda g) = T_\rho f + \lambda T_\rho g,$$

and the corresponding matrix of this operator is the the nth tensor power of the 2×2 matrix in (10.2).

The operator T_ρ has a smoothing effect. When $\rho = 1$, we have $T_\rho f = f$, but as one decreases ρ, the function $T_\rho f$ approaches to the constant $\mathbb{E}[f]$ and when $\rho = 0$, we have $T_\rho f = \mathbb{E}[f]$. Note $T_\rho f(x)$ takes the average of f evaluated at points sampled according to $\mathbf{z}$. When $\rho = 1$, the random variable $\mathbf{z}$ is concentrated on the original point x, and we obtain $f(x)$. As ρ decreases, the random variable $\mathbf{z}$ increasingly spreads over the whole group $\mathbb{Z}_2^n$. Finally, at $\rho = 0$, we lose the information about x and $\mathbf{z}$ is distributed uniformly over all points in $\mathbb{Z}_2^n$. Therefore, $\mathbb{E}_{\mathbf{z}} f(\mathbf{z}) = \mathbb{E}[f]$ in this case.

Let us now investigate the effect of the noise operator on the Fourier expansion.

Lemma 10.8. *Given $f : \mathbb{Z}_2^n \to \mathbb{R}$, we have*

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

Proof. Since T_ρ is linear, it suffices to show that for every $S \subseteq [n]$, we have

$$T_\rho \chi_S = \rho^{|S|} \chi_S.$$

Namely, χ_S are the eigenvectors of T_ρ with corresponding eigenvalues $\rho^{|S|}$. We have

$$\begin{aligned} T_\rho \chi_S(x) &= \mathbb{E}_{\mathbf{y} \sim \mu_\gamma^n} \chi_S(x + \mathbf{y}) = \chi_S(x) \mathbb{E}_{\mathbf{y} \sim \mu_\gamma^n} \chi_S(\mathbf{y}) = \chi_S(x) \mathbb{E}_{\mathbf{y} \sim \mu_\gamma^n} \prod_{i \in S} (-1)^{\mathbf{y}_i} \\ &= \chi_S(x) \prod_{i \in S} \mathbb{E}_{\mathbf{y}_i \sim \mu_\gamma} (-1)^{\mathbf{y}_i} = \chi_S(x) \rho^{|S|}. \end{aligned}$$

□

Lemma 10.8 shows that the noise operator dampens the Fourier coefficients, and the dampening effect increases exponentially as a function of $|S|$.

Remark 10.9. For every $x \in \mathbb{Z}_2^n$, we have

$$\mathbb{E}_{\mathbf{y} \sim \mu_\gamma^n}[f(x + \mathbf{y})] = 2^n \mathbb{E}_{\mathbf{y} \in \mathbb{Z}_2^n}\left[f(x + \mathbf{y}) \mu_\gamma^n(\mathbf{y})\right] = 2^n f * \mu_\gamma^n(x).$$

where in the second expectation, $\mathbf{y} \in \mathbb{Z}_2^n$ is chosen according to the uniform distribution. The Fourier expansion of μ_γ^n is given by $\widehat{\mu_\gamma^n}(S) = \frac{\rho^{|S|}}{2^n}$, which also implies $\widehat{T_\rho f}(S) = 2^n \widehat{\mu_\gamma^n}(S) \widehat{f}(S) = \rho^{|S|} \widehat{f}(S)$.

In light of Lemma 10.8, we can extend the definition of the noise operator T_ρ to include arbitrary values of $\rho \in \mathbb{R}$ beyond the interval $\rho \in [0, 1]$.

Definition 10.10. For $\rho \in \mathbb{R}$ and $f : \mathbb{Z}_2^n \to \mathbb{R}$, define

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

In the proof of Lemma 10.8, we used the fact that the noise operator acts independently on each coordinate. This approach applies to many results involving the noise operator: we first analyze the effect of noise on a single coordinate and then extend this analysis to the entire space using its product structure.

Theorem 10.11 (contractivity)**.** *For $1 \leqslant p \leqslant \infty$, the operator T_ρ is a contractive operator from L_p to L_p. That is,*

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

Proof. The proof of contractivity from Theorem 10.4 remains valid for the general setting of functions $f : \mathbb{Z}_2^n \to \mathbb{R}$. □

Next, we will show that T_ρ is hypercontractive. Before stating this theorem and presenting its proof, we introduce some notation.

For a distribution μ over X and a distribution ν over Y, consider the product probability distribution $\mu \times \nu$. For a function $f : (X \times Y, \mu \times \nu) \to \mathbb{R}$, let $\|f\|_{L_p(\nu)}$ denote the function $x \mapsto \|f_x\|_{L_p(\nu)}$ with $f_x = f(x, \cdot)$. Similarly, let $\|f\|_{L_p(\mu)}$ denote the function $y \mapsto \|f_y\|_{L_p(\mu)}$, where $f_y = f(\cdot, y)$.

Given a subset $S \subset [n]$, we can view a function $f : \mathbb{Z}_2^n \to \mathbb{R}$ as a function $f : \mathbb{Z}_2^S \times \mathbb{Z}_2^{\bar{S}} \to \mathbb{R}$. It is instructive to see how $\left\| \|f\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})}$ expands. We have

$$\left\| \|f\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} = \left(\mathbb{E}_{y \in \mathbb{Z}_2^{\bar{S}}} \left| \|f_y\|_{L_p(\mathbb{Z}_2^S)} \right|^q \right)^{1/q} = \left(\mathbb{E}_{y \in \mathbb{Z}_2^{\bar{S}}} \left| \left(\mathbb{E}_{x \in \mathbb{Z}_2^S} |f_y(x)|^p \right)^{1/p} \right|^q \right)^{1/q} . \tag{10.5}$$

Since $f_y(x) = f(x, y)$, we have

$$\|f\|_q = \left\| \|f\|_{L_q(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} . \tag{10.6}$$

As in the proof of Theorem 10.4, convexity of the norms show

$$\left\| \|f\|_{L_1(\mu)} \right\|_{L_p(\nu)} = \|\mathbb{E}_{x \sim \mu} |f(x, \cdot)|\|_{L_p(\nu)} \leqslant \mathbb{E}_{x \sim \mu} \| \, |f(x, \cdot)| \, \|_{L_p(\nu)} = \left\| \|f\|_{L_p(\nu)} \right\|_{L_1(\mu)} .$$

This is a special case of a more general inequality:

Theorem 10.12 (Generalized Minkowski's Inequality)**.** *For $1 \leqslant p \leqslant q \leqslant \infty$, we have*

$$\left\| \|f\|_{L_p(\nu)} \right\|_{L_q(\mu)} \leqslant \left\| \|f\|_{L_q(\mu)} \right\|_{L_p(\nu)} .$$

Now, we have all the tools to prove the hypercontractivity of T_ρ.

Theorem 10.13 (Hypercontractivity)**.** *Let $1 < p \leqslant q < \infty$ and $f : \mathbb{Z}_2^n \to \mathbb{R}$. We have*

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

Proof. The proof is by induction on n. We have already verified the inequality for $n = 1$ in Theorem 10.6. Next, we exploit the product structure to prove the inequality for arbitrary n.

Consider $f : \mathbb{Z}_2^n \to \mathbb{R}$. For $S \subseteq [n]$, let T_ρ^S denote the noise operator applied only to the coordinates in S. More formally, T_ρ^S is an operator on the function $f(\cdot, x_{\bar{S}})$, where $x_{\bar{S}}$ denotes the variables x_i for $i \notin S$. Let $S = \{1\}$. By Equation (10.5), we have

$$\begin{aligned}
\|T_\rho f\|_q &= \left\| T_\rho^S T_\rho^{\bar{S}} f \right\|_q \\
&= \left\| \left\| T_\rho^S T_\rho^{\bar{S}} f \right\|_{L_q(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} && \text{(Equation (10.6))} \\
&\leqslant \left\| \left\| T_\rho^{\bar{S}} f \right\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} && \text{(Induction Hypothesis)} \\
&\leqslant \left\| \left\| T_\rho^{\bar{S}} f \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} \right\|_{L_p(\mathbb{Z}_2^S)} && \text{(Generalized Minkowski)} \\
&\leqslant \left\| \|f\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_p(\mathbb{Z}_2^S)} && \text{(Induction Hypothesis)} \\
&= \|f\|_p && \text{(Equation (10.6)).}
\end{aligned}$$

□

10.3 Degree and hypercontractivity

In most applications, we will apply the hypercontractivity in the following form.

Theorem 10.14 (hypercontractivity and projection to low degrees)**.** *Let $f : \mathbb{Z}_2^n \to \mathbb{R}$ be a function and $k > 0$ an integer.*

(i) For $2 \leqslant q < \infty$, we have

$$\left\|f^{\leqslant k}\right\|_q \leqslant (q-1)^{\frac{k}{2}} \|f\|_2 .$$

(ii) For $1 \leqslant p \leqslant 2$,

$$\left\|f^{\leqslant k}\right\|_2 \leqslant e^{(\frac{2}{p}-1)k} \|f\|_p .$$

Proof. **Proof of (i):** Let $\rho = \frac{1}{\sqrt{q-1}}$ and let $g \coloneqq T_{\sqrt{q-1}} f = \sum \sqrt{q-1}^{|S|} \widehat{f}(S)\chi_S$ so that $f = T_\rho g$. By hypercontractivity (Theorem 10.13), we have

$$\left\|f^{\leqslant k}\right\|_q = \left\|T_\rho g^{\leqslant k}\right\|_q \leqslant \left\|g^{\leqslant k}\right\|_2 = \left(\sum_{|S|\leqslant k} \sqrt{q-1}^{2|S|} \widehat{f}(|S|)^2\right)^{1/2} \leqslant (q-1)^{k/2} \left(\sum_S \widehat{f}(|S|)^2\right)^{1/2} = (q-1)^{k/2} \|f\|_2 .$$

Proof of (ii): Instead of directly applying the hypercontractivity, we first use the duality of norms and apply the hypercontractivity to some $q > 2$.

Let $\varepsilon > 0$ be a parameter and set $q \coloneqq 2+\varepsilon$. Let θ be the solution to $\frac{1}{2} = \frac{\theta}{p} + \frac{1-\theta}{q}$, which is $\theta = \frac{p(q-2)}{2(q-p)} = \frac{\varepsilon p}{2(2+\varepsilon-p)}$. By Hölder's inequality and Part (i), we have

$$\left\|f^{\leqslant k}\right\|_2 = \sqrt{\langle f^{\leqslant k}, f\rangle} \leqslant \left\|f^{\leqslant k}\right\|_q^{1-\theta} \|f\|_p^{\theta} \leqslant (q-1)^{\frac{k(1-\theta)}{2}} \left\|f^{\leqslant k}\right\|_2^{1-\theta} \|f\|_p^{\theta},$$

which simplifies to

$$\left\|f^{\leqslant k}\right\|_2 \leqslant (q-1)^{\frac{k(1-\theta)}{2\theta}} \|f\|_p = (1+\varepsilon)^{\frac{k(1-\theta)}{2\theta}} \|f\|_p \leqslant e^{\frac{k\varepsilon(1-\theta)}{2\theta}} \|f\|_p .$$

Substituting $\theta = \frac{\varepsilon p}{2(2+\varepsilon-p)}$ and taking the limit as $\varepsilon \to 0$ yields the desired bound. □

Remark 10.15. In Theorem 10.14 (ii), it is more straightforward to show a weaker bound. Let q satisfy $\frac{1}{p} + \frac{1}{q} = 1$, which is equivalent to $q - 1 = \frac{1}{p-1}$. By Hölder's inequality and Theorem 10.14 (i),

$$\left\|f^{\leqslant k}\right\|_2^2 = \left\langle f^{\leqslant k}, f\right\rangle = \left\|f^{\leqslant k}\right\|_q \|f\|_p \leqslant (q-1)^{\frac{k}{2}} \left\|f^{\leqslant k}\right\|_2 \|f\|_p = (p-1)^{-k/2} \left\|f^{\leqslant k}\right\|_2 \|f\|_p ,$$

which simplifies to $\left\|f^{\leqslant k}\right\|_2 \leqslant (p-1)^{-k/2} \|f\|_p$. However, this bound deteriorates rapidly when p tends to 1, and it becomes meaningless at $p = 1$.

10.3.1 Equivalence of norms for low degree polynomials

If f is a constant function, all the L_p norms coincide. The following immediate consequence of Theorem 10.14 states that similarly, for low-degree functions, all the L_p norms are within constant factors of each other.

Corollary 10.16 (equivalence of norms for low degree polynomials)**.** *Let $f : \mathbb{Z}_2^n \to \mathbb{R}$ have degree at most k.*

(i) For $2 \leqslant q < \infty$, we have

$$\|f\|_2 \leqslant \|f\|_q \leqslant (q-1)^{\frac{k}{2}} \|f\|_2 .$$

(ii) For $1 \leqslant p \leqslant 2$,

$$e^{(1-\frac{2}{p})k} \|f\|_2 \leqslant \|f\|_p \leqslant \|f\|_2 .$$

By setting $f = \sum_i a_i \chi_{\{i\}}$ to be a degree 1 function, we immediately obtain the so-called Khintchine inequality.

Theorem 10.17 (Khintchine inequality)**.** *Let $a_1, a_2, \ldots, a_n \in \mathbb{R}$, and let $\varepsilon_1, \ldots, \varepsilon_n$ be ± 1-valued i.i.d. random variables with $\mathbf{Pr}[\varepsilon_i = 1] = 1/2$. For $q \geqslant 2$, we have*

$$\left(\sum_i |a_i|^2\right)^{1/2} \leqslant \left(\mathbb{E}\left|\sum_i \varepsilon_i a_i\right|^q\right)^{1/q} \leqslant \sqrt{q-1}\left(\sum_i |a_i|^2\right)^{1/2},$$

and for $1 \leqslant p \leqslant 2$,

$$e^{(1-\frac{2}{p})}\left(\sum_i |a_i|^2\right)^{1/2} \leqslant \left(\mathbb{E}\left|\sum_i \varepsilon_i a_i\right|^p\right)^{1/p} \leqslant \left(\sum_i |a_i|^2\right)^{1/2},$$

10.4 Noise and hypercontractivity for general distributions

Let (X, μ) be a probability space, and consider the product space (X^n, μ^n). Given a parameter $\rho \in [0,1]$, the *ρ-equal* copy of $x \in X^n$ is the random variable $\mathbf{y}$ that is sampled from X^n through the following process: for each $i \in [n]$, with probability ρ, set $\mathbf{y}_i = x_i$ and with probability $1-\rho$, sample $\mathbf{y}_i$ from (X, μ).

Definition 10.18 (Noise Operator)**.** Let (X, μ) be a probability space, and $\rho \in [0,1]$ a parameter. Given a function $f : (X^n, \mu^n) \to \mathbb{R}$, define $T_\rho f : (X^n, \mu^n) \to \mathbb{R}$ as

$$T_\rho f(x) \coloneqq \mathbb{E}_{\mathbf{y}} f(\mathbf{y}),$$

where $\mathbf{y}$ is the ρ-equal copy of x.

Recall from Section 9.4.1 that for every probability distribution (X, μ), every function $f : (X, \mu)^n \to \mathbb{R}$ has a unique Fourier-Walsh expansion $f = \sum_{S \subseteq [n]} F_S$, where the functions F_S are S-juntas, and satisfy $\mathbb{E}_i F_S \equiv 0$ for every $i \in S$. By the definition of the noise operator, we have

$$T_\rho F_S(x) = \sum_{T \subseteq S} \rho^{|T|}(1-\rho)^{|S|-|T|}\mathbb{E}_{S \setminus T} F_S(x).$$

Since $\mathbb{E}_{S \setminus T} F_S(x) \equiv 0$ unless $S = T$, we get $T_\rho F_S(x) = \rho^{|S|} F_S(x)$ and

$$T_\rho f = \sum_{S \subseteq [n]} \rho^{|S|} F_S,$$

similar to the uniform distribution on $\mathbb{Z}_2^n$.

It is straightforward to verify that T_ρ is a contractive operator on the L_p spaces. In Theorem 10.13, we saw that for the uniform distribution on $\mathbb{Z}_2^n$, we have hypercontractivity $\|T_\rho f\|_q \leqslant \|f\|_p$ for $1 < p < q < \infty$, when $\rho = \sqrt{\frac{p-1}{q-1}}$. Naturally, one may ask whether a similar hypercontractive inequality holds for general product distributions μ^n. However, for general product measures, it turns out that one cannot choose a $\rho > 0$ depending solely on p and q. When μ contains atoms with small probability masses, ρ must be significantly smaller for hypercontractivity to hold. Unfortunately, this means that for continuous spaces such as $[0,1]^n$, hypercontractivity does not hold unless $\rho = 0$.

We will only state the general form of hypercontractivity for the case where one of the norms is the L_2 norm, as this is the most relevant form for applications.

Theorem 10.19 (General hypercontractivity [**?**])**.** *Let (X, μ) be a probability space, and let $\lambda \coloneqq \min_{a \in X} \mu(a)$ be the minimum probability of any outcome according to μ. For every $f : (X^n, \mu^n) \to \mathbb{R}$, the following holds.*

(i) For $2 \leqslant q < \infty$, we have

$$\|T_\rho f\|_q \leqslant \|f\|_2 \qquad \text{for every } 0 \leqslant \rho \leqslant \frac{1}{\sqrt{q-1}}\lambda^{\frac{1}{2}-\frac{1}{q}}.$$

(ii) For $1 < p \leqslant 2$,

$$\|T_\rho f\|_2 \leqslant \|f\|_p \qquad \text{for every } 0 \leqslant \rho \leqslant \sqrt{p-1}\lambda^{\frac{1}{p}-\frac{1}{2}}.$$

We get the following corollary, analogues to Corollary 10.16. To keep the exposition simple, we do not optimize it for when p is close to or equals 1.

Corollary 10.20. *Let (X, μ) be a probability space, and let $\lambda \coloneqq \min_{a \in X} \mu(a)$ be the minimum probability of any outcome according to μ. Let $f : (X^n, \mu^n) \to \mathbb{R}$ be a function and $k > 0$ an integer.*

(i) For $2 \leqslant q < \infty$, we have

$$\left\| f^{\leqslant k} \right\|_q \leqslant (q-1)^{\frac{k}{2}} \lambda^{k\left(\frac{2}{q}-1\right)} \|f\|_2 .$$

(ii) For $1 \leqslant p \leqslant 2$,

$$\left\| f^{\leqslant k} \right\|_2 \leqslant (p-1)^{-\frac{k}{2}} \lambda^{k\left(1-\frac{2}{p}\right)} \|f\|_p .$$

Proof. To prove (i), let $\rho = \frac{1}{\sqrt{q-1}} \lambda^{\frac{1}{2}-\frac{1}{q}}$ and let $g \coloneqq T_{\frac{1}{\rho}} f = \sum \rho^{-|S|} F_S$ so that $f = T_\rho g$. By Theorem 10.19, we have

$$\left\| f^{\leqslant k} \right\|_q = \left\| T_\rho g^{\leqslant k} \right\|_q \leqslant \left\| g^{\leqslant k} \right\|_2 = \left(\sum_{|S| \leqslant k} \rho^{-2|S|} \|F_S\|_2^2 \right)^{1/2} \leqslant \rho^{-k} \left(\sum_S \|F_S\|_2^2 \right)^{1/2} = \rho^{-k} \|f\|_2 .$$

Part (ii) easily follows from (i) for $\frac{1}{p} + \frac{1}{q} = 1$ by duality. □

Remark 10.21. Note that this dependency of Corollary 10.20 on $\lambda \coloneqq \min_{a \in X} \mu(a)$ is essential, and it is of the right order of magnitude. For example, consider the p-biased distribution[2] with $p = \lambda$, and define $f : (\{0,1\}, \mu_\lambda^n)^n \to \{0,1\}$ as $f(x) = x_1$. Even though f is a 1-junta, we have $\|f\|_q = \lambda^{1/q}$ and $\|f\|_2 = \lambda^{1/2}$, which shows that the bounds in Corollary 10.20 have the opetimal dependency on λ.

10.5 Exercises

[2] The p-biased Bernoulli distribution, denoted by μ_p, is the probability distribution on $\{0,1\}$ with $\mu_p(1) = p$ and $\mu_p(0) = 1 - p$.

Chapter 11

Level-k inequality, Chang's lemma, and the FKN theorem

In this chapter, we study three theorems in the analysis of Boolean functions whose proofs rely on hypercontractivity. We begin with the *level-k inequality*, which shows that Boolean functions that have *small density* assign a small weight to the low-degree Fourier levels. In contrast, when the density is larger, as illustrated by the dictator functions and juntas, the Fourier mass can be entirely concentrated on lower-degree coefficients.

Afterwards, we turn our attention to a key tool in additive combinatorics introduced by Chang [Cha02]. This theorem, which is closely related to the level-1 inequality, states that the *large* Fourier coefficients of a Boolean function must all reside in a low-dimensional subspace. Chang's lemma has numerous applications in additive combinatorics, and in particular, many recent quantitative improvements over Roth's theorem on 3-term arithmetic progressions utilize this lemma.

Finally, we study a result of Friedgut, Kalai, and Naor [FKN02], known as the FKN theorem. In Chapter 6, we showed that Boolean functions of degree at most one are either constant functions or dictators. The FKN theorem provides a robust version of this result by showing that if a Boolean function is close to a degree-1 real-valued function, then it must be close to a degree-1 Boolean function, i.e., a constant function or a dictator.

11.1 Level-k inequality

Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be a function with mean $\mathbb{E}[f] = \alpha$. By Parseval's identity, the Fourier mass of f satisfies $\sum_a |\widehat{f}(a)|^2 = \alpha$. The level-$k$ inequality states that when α is small, the contribution of the lower levels to this sum is very small. In contrast, functions such as dictators and juntas illustrate that when α is relatively large, the Fourier mass can be entirely on the lower levels.

Theorem 11.1 (Level-k inequality)**.** *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ have mean $\mathbb{E}[f] = \alpha \leqslant \frac{1}{10}$, and let k be a positive integer. We have*

$$\|f^{\leqslant k}\|_2^2 \leqslant e^2 \ln(1/\alpha)^k \alpha^2.$$

Proof. Assume $\|f^{\leqslant k}\|_2 \neq 0$ since otherwise the theorem is trivial. Let $1 < p \leqslant 2$ to be determined, and let $q \geqslant 2$ be its conjugate exponent satisfying $\frac{1}{p} + \frac{1}{q} = 1$. By applying Hölder's inequality and hypercontractivity, we have

$$\|f^{\leqslant k}\|_2^2 = \left\langle f^{\leqslant k}, f \right\rangle \leqslant \|f^{\leqslant k}\|_q \|f\|_p \leqslant (q-1)^{\frac{k}{2}} \|f^{\leqslant k}\|_2 \|f\|_p,$$

which shows

$$\|f^{\leqslant k}\|_2 \leqslant (q-1)^{\frac{k}{2}} \|f\|_p = (q-1)^{\frac{k}{2}} \alpha^{1/p} \leqslant q^{\frac{k}{2}} \alpha^{1-\frac{1}{q}}.$$

Thus

$$\|f^{\leqslant k}\|_2^2 \leqslant q^k \alpha^{-\frac{2}{q}} \alpha^2.$$

Substituting $q = \ln(1/\alpha) \geqslant 2$ yields the desired result. □

11.2 Chang's lemma

In Chapter 5, we discussed a notion of pseudo-randomness based on Fourier uniformity. We showed that functions that do not have significant non-principal Fourier coefficients mimic the behaviour of random functions in that they contain the "expected" number of certain linear patterns.

Let $\varepsilon > 0$ be a parameter, and consider a Boolean function $f : \mathbb{Z}_2^n \to \{0,1\}$ with density $\mathbb{E}[f] = \alpha$. Note that every Fourier coefficient of f satisfies $|\widehat{f}(a)| = |\mathbb{E}[f(x)\chi_a(x)]| \leqslant \mathbb{E}[f] = \alpha$. Let us denote the set of large Fourier coefficients of f by

$$\operatorname{Spec}_\varepsilon(f) \coloneqq \left\{a \ : \ |\widehat{f}(a)| \geqslant \varepsilon\alpha\right\}.$$

The non-principal characters in $\operatorname{Spec}_\varepsilon(f)$ are the obstacles to the Fourier uniformity of f. In many applications, one wishes to extract some structure about f from this set or to eliminate these large Fourier coefficients by restricting to a subgroup (or approximate subgroup).

First, note that by Parseval's identity, we have

$$\alpha = \mathbb{E}[f^2] = \sum_a |\widehat{f}(a)|^2 \geqslant |\operatorname{Spec}_\varepsilon(f)|(\varepsilon\alpha)^2,$$

showing

$$|\operatorname{Spec}_\varepsilon(f)| \leqslant \frac{1}{\varepsilon^2\alpha}.$$

Even though $\operatorname{Spec}_\varepsilon(f)$ can be of size $\Omega(1/\alpha)$, Chang, in her work [Cha02] on Frieman's theorem, proved that all these significant Fourier coefficients must lie within a subspace of dimension at most $O(\log(\frac{1}{\alpha}))$. In other words, large Fourier coefficients cannot be arbitrarily scattered, and their distribution has some inherent structure. Chang's lemma has been used in several central works in additive combinatorics. In particular, it is a key step in Sander's quasi-polynomial bound on Frieman's theorem [San12, Lov15], and most of the recent improvements in the bounds for Roth's theorem [San11a, KM23].

Theorem 11.2 (Chang's lemma [Cha02])**.** *Let $\varepsilon > 0$ be a parameter, and let $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfy $\mathbb{E}[f] = \alpha \leqslant \frac{1}{10}$. The linear span of $\operatorname{Spec}_\varepsilon(f)$ in $\mathbb{Z}_2^n$ has dimension at most $\frac{e}{\varepsilon^2}\ln(1/\alpha)$.*

Proof. Let $a_1, \ldots, a_d \in \operatorname{Spec}_\varepsilon(f)$ be a maximal set of linearly independent elements in $\operatorname{Spec}_\varepsilon(f)$. We can apply a change of coordinates for $\mathbb{Z}_2^n$ and, without loss of generality, assume that $a_i = e_i$ for $1 \leqslant i \leqslant d$.

After this change of variables, $\widehat{f}(a_1)\chi_{a_1}, \ldots, \widehat{f}(a_d)\chi_{a_d}$ are all in level-one, and therefore, contribute at least $d(\varepsilon\alpha)^2$ to $\|f^{\leqslant 1}\|_2^2$. By applying the level-1 inequality (Theorem 11.1 with $k = 1$), we have

$$d(\varepsilon\alpha)^2 \leqslant \|f^{\leqslant 1}\|_2^2 \leqslant e^2\ln(1/\alpha)\alpha^2,$$

which shows

$$d \leqslant \frac{e^2}{\varepsilon^2}\ln(1/\alpha).$$

□

Remark 11.3. Alternatively, we can prove Chang's lemma with a direct probabilistic argument. Define $g = \sum_{i=1}^d \chi_{a_i}$. We wish to upper-bound and lower-bound $\mathbb{E}[fg]$. By Parseval, the lower bound is $\varepsilon\alpha d$. For the upper bound, first note that the linear independence of $a_1, a_2, \ldots, a_d$ translates to probabilistic independence for the corresponding characters, i.e., when $\mathbf{x}$ is chosen uniformly at random, $g(\mathbf{x})$ is a sum of d i.i.d. ± 1-valued random variables. Consequently, $g(\mathbf{x})$ is strongly concentrated, and one can use the concentration of $g(\mathbf{x})$ to upper-bound $\mathbb{E}[fg]$.

11.3 The FKN dictator theorem

In Proposition 6.8, we showed that a Boolean function $f : \mathbb{Z}_2^n \to \{0,1\}$ with degree at most one is either a constant function or a dictator. In this section, we examine the robustness of this statement. Instead of requiring the degree to be at most one, we assume that the Fourier mass is highly concentrated on the characters of degree at most one. Friedgut, Kalai, and Naor [FKN02] proved that every such function must be close to either a constant function or a dictator.

As it is more convenient, we will state the proof of the FKN theorem for functions $f : \{-1,1\}^n \to \{-1,1\}$. To translate Theorem 11.4 to a statement about functions $f : \{-1,1\}^n \to \{0,1\}$, note that $2f - 1 : \{-1,1\}^n \to \{-1,1\}$ satisfies $\|(2f-1)^{>1}\|_2^2 = 2\|f^{>1}\|_2^2$.

Theorem 11.4 (FKN theorem [FKN02])**.** *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ satisfy $\|f^{>1}\|_2^2 = \delta$. There exists $g : \{-1,1\}^n \to \{-1,1\}$ such that g is either a constant function or a dictator, and it satisfies*

$$\mathbf{Pr}[f(\mathbf{x}) \neq g(\mathbf{x})] \leqslant 10^3\delta.$$

Proof. Denoting the coefficients as $f^{\leqslant 1} = a_0 + \sum_{i=1}^n a_i x_i$ and let $h \coloneqq (f^{\leqslant 1})^2$. Note that

$$\mathbb{E}[h] = \sum_{i=0}^n a_i^2 = \|f^{\leqslant 1}\|_2^2 = \|f\|_2^2 - \|f^{\leqslant 1}\|_2^2 = 1 - \delta.$$

We will show that h is almost a constant function, i.e., it has a small variance. Since $\deg(h) \leqslant 2$, by hypercontractivity (Theorem 10.14 with $p = 1$), we have

$$\sqrt{\mathrm{Var}[h]} = \|h - \mathbb{E}[h]\|_2 \leqslant e^2 \mathbb{E}|h - \mathbb{E}[h]|.$$

By $f^2 \equiv 1$, the Cauchy-Schwarz inequality, and the triangle inequality, we have

$$\begin{aligned}
\mathbb{E}|h - \mathbb{E}[h]| &\leqslant \mathbb{E}|h - f^2| + \mathbb{E}[|1 - \mathbb{E}[h]|] \\
&= \mathbb{E}\left|(f^{\leqslant 1} - f)(f^{\leqslant 1} + f)\right| + \delta \\
&\leqslant \|f^{\leqslant 1} - f\|_2 \|f^{\leqslant 1} + f\|_2 + \delta \\
&\leqslant \|f^{>1}\|_2 (\|f^{\leqslant 1}\|_2 + \|f\|_2) + \delta = 2\sqrt{\delta} + \delta \leqslant 3\sqrt{\delta}.
\end{aligned}$$

Therefore,

$$\sqrt{\mathrm{Var}[h]} \leqslant 3e^2\sqrt{\delta} \leqslant 30\sqrt{\delta}.$$

On the other hand,

$$h = \left(a_0 + \sum_{i=1}^n a_i x_i\right)^2 = \sum_{i=0}^n a_i^2 + \sum_{i=1}^n 2a_0 a_i x_i + \sum_{i<j} 2a_i a_j x_i x_j,$$

which shows

$$\mathrm{Var}[h^2] = 4a_0^2 \sum_{i=1}^n a_i^2 + \sum_{i<j} 4a_i^2 a_j^2 = 2\left(\left(\sum_{i=0}^n a_i^2\right)^2 - \sum_{i=0}^n a_i^4\right) \geqslant 2\left(\sum_{i=0}^n a_i^2\right)^2 - 2\left(\sum_{i=0}^n a_i^2\right) \max_{i\in\{0,\ldots,n\}} |a_i|^2.$$

Substituting $\sum_{i=0}^n a_i^2 = 1 - \delta$ and using $\mathrm{Var}[h^2] \leqslant (30\sqrt{\delta})^2 \leqslant 10^3\delta$, we have

$$2(1-\delta)^2 - 2 \max_{i\in\{0,\ldots,n\}} |a_i|^2 \leqslant 10^3\delta.$$

Therefore, there exists $m \in \{0,\ldots,n\}$ with $|a_m| \geqslant |a_m|^2 \geqslant 1 - 10^3\delta$.

If $m = 0$, let $g : \{-1,1\}^n \to \{-1,1\}$ be the constant function $g \equiv \mathrm{sgn}(a_m)$. If $m \neq 0$, let $g(x) \coloneqq \mathrm{sgn}(a_m)x_m$. We have

$$\mathbf{Pr}[f(\mathbf{x}) \neq g(\mathbf{x})] = \frac{1}{4}\|f - g\|_2^2 = \frac{1}{4}\left(\mathbb{E}[f^2] + \mathbb{E}[g^2] - 2\mathbb{E}[fg]\right) = \frac{1}{4}\left(2 - 2|a_m|\right) \leqslant 10^3\delta.$$

□

11.4 Exercises

Exercise 11.1. Prove Chang's lemma using the argument sketched in Remark 11.3.

Chapter 12

Junta Theorem and KKL Inequality

In this chapter, we study two central results in the analysis of Boolean functions: the KKL inequality of Kahn, Kalai, and Linial [KKL88], and Friedgut's junta theorem [Fri98]. The KKL inequality is the first result in discrete mathematics and theoretical computer science to use hypercontractivity. This inequality asserts that every balanced Boolean function must have at least one influential variable.

Building on the techniques from KKL, Friedgut showed that a Boolean function with a small total influence is essentially a junta. Note that conversely, every junta has a small total influence.

We will present the proof of Friedgut's junta theorem in Section 12.2, and then present the KKL inequality in Section 12.3.

12.1 A rule of thumb for applying hypercontractivity

The proofs of Friedgut's junta theorem, the KKL inequality, as well as many other applications of hypercontractivity in the analysis of Boolean functions use the fact the L_p norms of *sparse* Boolean functions grow very rapidly as p increases. To see this, note that if f is a $\{0,1\}$-valued function with $\mathbb{E}[f] = \alpha$, then $\|f\|_p = (\mathbb{E}[f^p])^{1/p} = (\mathbb{E}[f])^{1/p} = \alpha^{1/p}$, which for small α, grows very rapidly as p increases. On the other hand, hypercontractivity implies that for low-degree functions $g : \mathbb{Z}_2^n \to \mathbb{R}$, all the L_p norms are within constant factors of each other. Consequently, sparse Boolean functions cannot have significant weights on low-degree Fourier coefficients.

Proposition 12.1. *Consider $f : \mathbb{Z}_2^n \to \{0,1\}$ with $\mathbb{E}[f] = \alpha$. For every integer k, we have*

$$\|f^{\leqslant k}\|_2 \leqslant (3^k \alpha^{1/4})\|f\|_2.$$

Proof. By applying hypercontractivity (Remark 10.15) with $p = 4/3$, we have[1]

$$\|f^{\leqslant k}\|_2 \leqslant 3^k \|f\|_{4/3} = 3^k \mathbb{E}[f]^{3/4} = 3^k \mathbb{E}[f]^{1/4}\mathbb{E}[f]^{1/2} = 3^k \alpha^{1/4}\|f\|_2.$$

□

In the proof of Proposition 12.1, we used the fact that for Boolean functions $\mathbb{E}[f^p] \leqslant \mathbb{E}[f]$. The following proposition shows a similar statement for the ith coordinate Laplacian when $1 \leqslant p \leqslant 2$.

Proposition 12.2. *Let (X, μ) be a probability space and $f : (X, \mu)^n \to \{0,1\}$. For every $p \in [1,2]$, we have*

$$\mathbb{E}|\partial_i f|^p \leqslant \mathbb{E}|\partial_i f| = 2\mathbb{E}|\partial_i f|^2 = 2I_i(f),$$

where $\partial_i f \coloneqq f - \mathbb{E}_{\mathbf{x}_i} f$.

[1] The choice of $p = 4/3$ is quite arbitrary, and we can use any $1 < p < 2$. Optimizing the value of p will only affect the hidden constants in $2^{O(I_f/\varepsilon)}$ in the assertion of the theorem.

Proof. For every $z \in (X,\mu)^{[n]\setminus\{i\}}$, let $\alpha_z = \mathbf{Pr}_{\mathbf{y}_i\sim\mu}[f(z,\mathbf{y}_i) = 1]$. Since f is Boolean, for $z \in (X,\mu)^{[n]\setminus\{i\}}$ and $y_i \in (X,\mu)$, we have

$$|\partial_i f(z,y_i)| = \begin{cases} 1-\alpha_z & \text{if } f(z,y_i)=1 \\ -\alpha_z & \text{if } f(z,y_i)=0 \end{cases}.$$

Therefore,

$$\mathbb{E}|\partial_i f|^2 = \mathbb{E}_{\mathbf{z}}\mathbb{E}_{\mathbf{y}_i}|\partial_i f(\mathbf{z},\mathbf{y}_i)|^2 = \mathbb{E}_{\mathbf{z}}[\alpha_{\mathbf{z}}(1-\alpha_{\mathbf{z}})^2 + (1-\alpha_{\mathbf{z}})\alpha_{\mathbf{z}}^2] = \mathbb{E}_{\mathbf{z}}\alpha_{\mathbf{z}}(1-\alpha_{\mathbf{z}}),$$

and

$$\mathbb{E}|\partial_i f| = \mathbb{E}_{\mathbf{z}}[\alpha_{\mathbf{z}}(1-\alpha_{\mathbf{z}}) + (1-\alpha_{\mathbf{z}})\alpha_{\mathbf{z}}] = 2\mathbb{E}_{\mathbf{z}}\alpha_{\mathbf{z}}(1-\alpha_{\mathbf{z}}),$$

which shows $\mathbb{E}|\partial_i f| = 2\mathbb{E}|\partial_i f|^2$. Let $\theta \in [0,1]$ be such that $\frac{1}{p} = \frac{\theta}{1} + \frac{1-\theta}{2}$. By Hölder's inequality

$$\|\partial_i f\|_p^p \leqslant \|\partial_i f\|_1^{p\theta}\|\partial_i f\|_2^{p(1-\theta)} \leqslant \|\partial_i f\|_1^{p\theta}\|\partial_i f\|_1^{p(1-\theta)/2} = \|\partial_i f\|_1.$$

□

12.2 Friedgut's Junta Theorem

Recall from Chapter 9 that the influence of the ith variable on a function $f : \mathbb{Z}_2^n \to \{0,1\}$ is

$$I_i(f) = \frac{1}{4}\mathbf{Pr}_{\mathbf{x}}[f(\mathbf{x}) \neq f(\mathbf{x}+e_i)] = \|\partial_i f\|_2^2 = \sum_{S:i\in S}|\widehat{f}(S)|^2.$$

The total influence of f is defined as follows.

$$I_f = \sum_i I_i(f) = \sum_{S\subseteq[n]}|S||\widehat{f}(S)|^2.$$

If f is a k-junta, then $I_f \leqslant k$. Friedgut's theorem gives a partial converse to this statement. It states that Boolean functions with small total influences are nearly juntas.

Theorem 12.3 (Friedgut's junta theorem [Fri98])**.** *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be a Boolean function. There exists a $2^{O(I_f/\varepsilon)}$-junta $g : \mathbb{Z}_2^n \to \{0,1\}$ such that*

$$\mathbf{Pr}[f(\mathbf{x}) \neq g(\mathbf{x})] \leqslant \varepsilon.$$

Proof. Let J be the set of most influential variables of f, i.e., $J \coloneqq \{i \in [n] \,:\, I_i(f) \geqslant \delta\}$ for some parameter $\delta > 0$ to be determined. It is natural to try to find a g that depends only on the variables in J. The function

$$h \coloneqq \sum_{S\subseteq J}\widehat{f}(S)\chi_S.$$

depends only on the variables in J, but it is not necessarily a Boolean function. We shall round h to a Boolean function g as follows.

$$g(x) \coloneqq \begin{cases} 1 & \text{if } h(x) \geqslant \frac{1}{2} \\ 0 & \text{if } h(x) < \frac{1}{2} \end{cases}.$$

Note $f(x) \neq g(x)$ implies $|f(x)-h(x)|^2 \geqslant 1/4$, and therefore,

$$\mathbf{Pr}[f(\mathbf{x}) \neq g(\mathbf{x})] = \|f-g\|_2^2 \leqslant 4\|f-h\|_2^2.$$

Therefore, our task was reduced to showing

$$\|f-h\|_2^2 = \sum_{S\not\subseteq J}\widehat{f}(S)^2 \leqslant \frac{\varepsilon}{4}. \tag{12.1}$$

We will divide the sum in (12.1) into the low-degree and high-degree parts and deal with them separately.

Bounding high-degree coefficients: Intuitively, if the total influence is small, we cannot have a large L_2 mass on high-degree characters since high-degree characters have large total influences. We can easily formalize this intuition. Set $k \coloneqq 8I_f/\varepsilon$. We have

$$I_f = \sum_S |S|\widehat{f}(S)^2 \geqslant 8k \sum_{|S|\geqslant k} \widehat{f}(S)^2,$$

which implies

$$\sum_{S:|S|\geqslant k} \widehat{f}(S)^2 \leqslant \frac{I_f}{8k} \leqslant \frac{\varepsilon}{8}.$$

Thus,

$$\|f-h\|_2^2 \leqslant \frac{\varepsilon}{8} + \sum_{\substack{S:|S|<k \\ S\not\subseteq J}} \widehat{f}(S)^2. \tag{12.2}$$

Bounding low-degree coefficients: We will use hypercontractivity to bound the low-degree part. Every $S \not\subseteq J$ in Equation (12.1) contains at least one non-influential variable $i \notin J$. Therefore,

$$\sum_{\substack{S:|S|<k \\ S\not\subseteq J}} \widehat{f}(S)^2 = \sum_{i\notin J} \sum_{\substack{S:|S|<k \\ i\in S}} \widehat{f}(S)^2 = \sum_{i\notin J} \left\|\partial_i f^{<k}\right\|_2^2. \tag{12.3}$$

Denote $f_i(x) \coloneqq \partial_i f(x)$ and recall $I_i(f) = \mathbb{E}|f_i|^2$. By applying hypercontractivity (Remark 10.15) with $p = 4/3$, we have

$$\left\|f_i^{<k}\right\|_2 \leqslant 3^k \left\|f_i\right\|_{4/3},$$

and by Proposition 12.2, we have $\mathbb{E}|f_i(\mathbf{x})|^{4/3} \leqslant 2I_i(f)$. We obtain

$$\left\|f_i^{<k}\right\|_2^2 \leqslant 3^k \left(\mathbb{E}_x|f_i(x)|^{4/3}\right)^{3/2} = 3^k(2I_i(f))^{3/2} \leqslant 3^{k+2} I_i(f)^{3/2}.$$

For $i \notin J$, we have $I_i(f) < \delta$, and therefore,

$$\left\|f_i^{<k}\right\|_2^2 \leqslant 3^{k+2}\delta^{1/2} I_i(f).$$

Equivalently,

$$\sum_{\substack{S:|S|<k \\ i\in S}} \widehat{f}(S)^2 \leqslant 3^{k+2}\delta^{1/2} I_i(f).$$

Going back to (12.3), we have

$$\sum_{i\notin J} \sum_{\substack{S:|S|<k \\ i\in S}} \widehat{f}(S)^2 \leqslant \sum_{i\notin J} 3^{k+2}\delta^{1/2} I_i(f) \leqslant 3^{k+2}\delta^{1/2} I_f.$$

Substituting in (12.2), yields

$$\|f-h\|_2^2 \leqslant \frac{\varepsilon}{8} + 3^{k+2}\delta^{1/2} I_f.$$

Since $I_f = k\varepsilon/8$, by setting $\delta = 1/3^{3k+100}$, we have

$$\|f-h\|_2^2 \leqslant \frac{\varepsilon}{8} + 3^k\delta^{1/2} I_f \leqslant \frac{\varepsilon}{8} + \frac{\varepsilon}{8} = \frac{\varepsilon}{4},$$

which verifies (12.1). With this choice of δ, we have

$$|J| \leqslant \frac{I_f}{\delta} \leqslant 2^{O(I_f/\varepsilon)}.$$

□

12.3 KKL inequality

A Boolean function f is called *balanced* function if $\mathbb{E}[f] = 1/2$. In this section, we will prove the Kahn-Kalai-Linial (KKL) inequality [KKL88] that says every balanced function $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfies

$$\max_i I_i(f) = \Omega\left(\frac{\log n}{n}\right).$$

The proofs of the KKL inequality and Friedgut's junta theorem share many similarities. Historically, the KKL inequality was proven first, and Friedgut later applied ideas from the KKL result to establish the junta theorem.

Note that a balanced function satisfies $\mathrm{Var}[f] = \frac{1}{4}$, and therefore, the Poincare inequality $\mathrm{Var}[f] \leqslant I_f$ immediately implies $I_{\max} \coloneqq \max_i I_i(f) \geqslant \frac{1}{4n}$. To prove the KKL inequality, we will prove a stronger upper bound on the variance, which is of independent interest. More precisely, we will prove

$$\mathrm{Var}(f) \lesssim \frac{I_f}{\log(1/I_{\max})},$$

which shows that when $I_{\max}$ is small, I_f must be large. Since $I_f \leqslant I_{\max} n$, it easily follows that $I_{\max} = \Omega\left(\mathrm{Var}[f]\frac{\log n}{n}\right)$.

Theorem 12.4 (KKL inequality [KKL88])**.** *Consider $f : \mathbb{Z}_2^n \to \{0,1\}$ and denote $I_{\max} \coloneqq \max_i I_i(f)$. We have*

$$\mathrm{Var}(f) \leqslant \frac{20 I_f}{\log(1/I_{\max})},$$

Consequently,

$$I_{\max} = \Omega\left(\mathrm{Var}(f)\frac{\log n}{n}\right).$$

Proof. We have

$$\sum_{S\neq\emptyset} |\widehat{f}(S)|^2 = \mathrm{Var}(f).$$

Bounding high-degree coefficients: Let $k \approx \log \frac{1}{I_{\max}}$ to be determined later. Since

$$I_f = \sum_S |S||\widehat{f}(S)|^2 \geqslant k \sum_{|S|>k} |\widehat{f}(S)|^2,$$

we have

$$\sum_{|S|>k} |\widehat{f}(S)|^2 \leqslant \frac{I_f}{k} \lesssim \frac{I_f}{\log(1/I_{\max})}.$$

Bounding low-degree coefficients: To handle the low degree part, let $f_i(x) \coloneqq \partial_i f(x)$. By applying hypercontractivity (Remark 10.15) to f_i with $p = 4/3$, we have

$$\sum_{1\leqslant|S|\leqslant k} |\widehat{f}(S)|^2 \leqslant \sum_{i=1}^n \sum_{\substack{i\in S\\ |S|\leqslant k}} |\widehat{f}(S)|^2 = \sum_{i=1}^n \left\|f_i^{\leqslant k}\right\|_2^2 \leqslant \sum_{i=1}^n 3^k \|f_i\|_{4/3}^2.$$

By Proposition 12.2,, we have

$$\sum_{i=1}^n 3^k \|f_i\|_{4/3}^2 = 3^k \sum_{i=1}^n (2I_i(f))^{3/2} \leqslant 3^{k+2} I_{\max}^{1/2} \sum_{i=1}^n I_i(f) = 3^{k+2} I_{\max}^{1/2} I_f.$$

Combining the bounds for low-degree and high-degree parts, we get

$$\mathrm{Var}(f) = \sum_{S:|S|\geqslant 1} |\widehat{f}(S)|^2 \leqslant \frac{I_f}{k} + 3^{k+2} I_{\max}^{1/2} I_f.$$

Setting $k = \frac{1}{10}\log\frac{1}{I_{\max}}$ shows

$$\mathrm{Var}(f) \leqslant \frac{10I_f}{\log(1/I_{\max})} + 9I_{\max}^{\frac{1}{2}-\frac{1}{5}} I_f \leqslant \frac{20I_f}{\log(1/I_{\max})}.$$

Combining with $I_f \leqslant nI_{\max}$, a straightforward calculation implies

$$I_{\max} = \Omega\left(\mathrm{Var}(f)\frac{\log n}{n}\right).$$

□

We have the following immediate corollary, which is, for example, applicable when $f : \mathbb{Z}_2^n \to \{0,1\}$ is invariant under a transitive action on its variables.

Corollary 12.5. *If a balanced function $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfies $I_1(f) = I_2(f) = \cdots = I_n(f)$, then $I_f \gtrsim \log n$.*

Remark 12.6. Bourgain and Kalai [BK99] show that the KKL inequality will significantly improve under strong symmetry assumptions. For instance, if f is a symmetric function, i.e., f's output only depends on the Hamming weight of the input, then $I_f \gtrsim \sqrt{n}$ and $\max_i I_i(f) \gtrsim \frac{1}{\sqrt{n}}$.

12.3.1 Tribes function

The KKL inequality is tight, which can be seen by considering the *tribes* function, an important example in the analysis of Boolean functions. Let

$$f(x) = \bigvee_{i=1}^{m}\bigwedge_{j=1}^{k} x_{i,j},$$

where $k = \log n - \log\ln n$ and $m = n/k$. To study the influences of variables, consider one of the variables, say, $x_{1,1}$. For $x_{1,1}$ to be able to change the output, all other variables in the clause $\bigwedge_{j=1}^{k} x_{1,j}$ must be 1, and all other clauses must evaluate to 0. Therefore, the influence of the variable $x_{1,1}$ is at most

$$\frac{1}{4}(1-2^{-k})^{m-1}\cdot 2^{-k+1} = 2^{-k-1}(1-2^{-k})^{m-1} = \frac{\ln n}{2n}\left(1-\frac{\ln n}{n}\right)^{m-1} = \frac{\ln n}{2n}(1-o(1)).$$

12.3.2 Monotone functions

A Boolean function $f : \{0,1\}^n \to \{0,1\}$ is *monotone* if $f(x) \leqslant f(y)$ whenever $x_i \leqslant y_i$ for all i. Consider a monotone $f : \{0,1\}^n \to \{0,1\}$ and a variable, say, x_1. Since f is monotone, we have

$$I_1(f) = \frac{1}{4}\mathbb{E}[f(1,\mathbf{x}_2,\ldots,\mathbf{x}_n) - f(0,\mathbf{x}_2,\ldots,\mathbf{x}_n)] = \frac{\widehat{f}(\emptyset) - \widehat{f}(\{1\})}{4} - \frac{\widehat{f}(\emptyset) + \widehat{f}(\{1\})}{4} = -\frac{\widehat{f}(\{1\})}{2}.$$

Proposition 12.7. *For every monotone function $f : \{0,1\}^n \to \{0,1\}$, the influence of the ith variable on f satisfies*

$$I_i(f) = -\frac{\widehat{f}(\{i\})}{2}.$$

Consequently,

$$I_f \leqslant \sqrt{n}.$$

Proof. We have already showed $I_i(f) = -\frac{\widehat{f}(\{i\})}{2}$. This identity and the Cauchy-Schwarz inequality imply

$$I_f \leqslant \sum_i |\widehat{f}(\{i\})| \leqslant \sqrt{n}\left(\sum_i |\widehat{f}(\{i\})|^2\right)^{1/2} \leqslant \sqrt{n}\|f\|_2^2 \leqslant \sqrt{n}.$$

□

Note that for non-monotone functions, it is possible to have $I_f = \Omega(n)$ (e.g., $f = \textsc{Parity}$). The bound $I_f = O(\sqrt{n})$ for monotone functions is tight since $I_{\textsc{Maj}} = \Theta(\sqrt{n})$, where $\textsc{Maj}$ denotes the majority function:

$$\textsc{Maj}(x) \coloneqq \begin{cases} 1 & \text{if } \sum_i x_i \geqslant n/2 \\ 0 & \text{otherwise} \end{cases}.$$

Next, we will discuss an application of the KKL inequality. The following corollary shows that for every monotone balanced function, some coalition of $o(n)$ variables can collectively shift the expected value of f to be close to either 0 or 1.

Corollary 12.8. *Let $f : \{0,1\}^n \to \{0,1\}$ be a monotone balanced function. There is a set $J \subseteq [n]$ of size $O_\varepsilon\left(\frac{n}{\log n}\right)$ such that*

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_J = \vec{1}\right] \geqslant 1 - \varepsilon,$$

and

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_J = \vec{0}\right] \leqslant \varepsilon.$$

Proof Sketch. Let $i \in [n]$ have the highest influence. Then, by the KKL inequality, fixing $x_i = 1$ will increase the average of f by at least $\Omega(\frac{\log n}{n})$. We can repeatedly apply this argument, each time increasing the expected value by $\Omega(\frac{\log n}{n})$, to obtain a set J_1 of size $O_\varepsilon(\frac{n}{\log n})$ with

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_{J_1} = \vec{1}\right] \geqslant 1 - \varepsilon.$$

Note that the variance remains at least $\varepsilon(1-\varepsilon)$ through these iterations, and at least step we can find a variable with influence $\Omega\left(\varepsilon(1-\varepsilon)\frac{\log n}{n}\right)$.

Repeating the same process but setting the variables to 0 leads to another set J_2 of size $O_\varepsilon\left(\frac{n}{\log n}\right)$ with

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_{J_2} = \vec{0}\right] \leqslant \varepsilon.$$

The set $J := J_1 \cup J_2$ satisfies the desired properties. □

Ajtai and Linial constructed examples to show that there are functions for which Corollary 12.8 cannot be improved significantly in any direction.

Theorem 12.9 ([AL93])**.** *There exists a balanced monotone function $f : \{0,1\}^n \to \{0,1\}$ such that for every set J of size $o\left(\frac{n}{\log^2 n}\right)$, we have*

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_J = \vec{1}\right] = \frac{1}{2} + o(1),$$

and

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_J = \vec{0}\right] = \frac{1}{2} - o(1).$$

12.4 A few open problems

It is believed that in Theorem 12.9, the bound $o(\frac{n}{\log^2(n)})$ can be improved to close to $o(\frac{n}{\log(n)})$ matching the bound in Corollary 12.8

Conjecture 12.10. *The bound in Theorem 12.9 can be improved to $O\left(\frac{n}{\log^{1+\varepsilon} n}\right)$ for every $\varepsilon > 0$.*

Friedgut [Fri04] conjectures that an analogue of Corollary 12.8 is true over the continuous domain $[0,1]^n$.

Conjecture 12.11 (Friedgut [Fri04])**.** *Let $f : [0,1]^n \to \{0,1\}$ be an increasing function and $\varepsilon > 0$ be a parameter. There exists a subset $J \subseteq [n]$ with $|J| = o_\varepsilon(n)$ such that*

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_J = \vec{0}\right] \leqslant \varepsilon \quad \text{or} \quad \mathbb{E}\left[f(\mathbf{x})|\mathbf{x}_J = \vec{1}\right] \geqslant 1 - \varepsilon.$$

12.4.1 The Aaronson-Ambainis conjecture

In Theorem 7.13, we showed that for Boolean functions $f : \{0,1\}^n \to \{0,1\}$, the degree and decision complexity are polynomially equivalent. The following conjecture extends this idea and speculates that low-degree real-valued bounded functions can be well-approximated by low-complexity decision trees.

Conjecture 12.12 (folklore)**.** *Suppose $f : \{0,1\}^n \to [0,1]$ have degree at most d. For every $\varepsilon > 0$, there exists a decision tree T of depth at most $\mathrm{poly}(d, 1/\varepsilon)$ such that*

$$\|f - T\|_2^2 \leqslant \varepsilon.$$

Conjecture 12.12 is particularly significant in the context of quantum query complexity. Suppose a quantum algorithm Q makes t queries to a Boolean input string x. Can a classical algorithm, making only $\mathrm{poly}(t)$ queries to x, approximate Q's acceptance probability for most inputs x? Conjecture 12.12 would imply a positive answer to this question.

Aaronson and Ambainis [AA14] showed that Conjecture 12.12 is equivalent to the following conjecture about maximum influence.

Conjecture 12.13 (Aaronson-Ambainis conjecture [AA14])**.** *If $f : \{0,1\}^n \to [0,1]$ have degree at most d, then*

$$\max_i \|\partial_i f\|_2^2 \geqslant \mathrm{poly}(1/d, \mathrm{Var}[f]).$$

Chapter 13

Phase transition and influences

One major motivation for studying the influences of variables on Boolean functions is a deep connection to the threshold phenomenon, discovered independently by Margulis [Mar74] (in Russian) and later by Russo [Rus81]. This phenomenon refers to the rapid transition of a property from being highly unlikely to hold to highly likely as a parameter p increases. This abrupt change is observed in various contexts. It is analogous to a phase transition, a concept from statistical physics that captures the sudden shifts in the behaviour of physical systems as parameters, like temperature, change.

One of the main questions in studying phase transitions is determining the speed at which the phase transition occurs. *How sharp is the threshold?* In other words, how narrow is the *critical interval* in which the transition occurs?

As an example, consider the Erdös-Rényi random graph $G(n,p)$, where each of the possible $\binom{n}{2}$ edges is included independently with probability p. Erdös and Rényi [EoR59] proved that when $p = (1-\varepsilon)\frac{\ln(n)}{n}$, with high probability, $G(n,p)$ contains several isolated vertices and is therefore disconnected. On the other hand, when $p = (1+\varepsilon)\frac{\ln(n)}{n}$, then, with high probability, the random graph is connected. This illustrates a sharp phase transition: within a narrow interval around $p = \frac{\ln(n)}{n}$, the random graph rapidly shifts from being very unlikely to be connected to very likely.

Margulis [Mar74] and Russo [Rus81] showed that the total influence of the underlying Boolean function controls the sharpness of such thresholds. To characterize properties that do not exhibit sharp thresholds, one must understand the structure of Boolean functions with small total influences.

13.1 The p-biased distribution

To study random objects such as the Erdös-Rényi random graph $G(n,p)$, we often need to work with the *p-biased distribution* on the Boolean cube, which models the randomness in such structures.

Definition 13.1 (p-biased distribution)**.** The p-biased Bernoulli distribution, denoted by μ_p, is the probability distribution on $\{0,1\}$ with $\mu_p(1) = p$ and $\mu_p(0) = 1-p$. The *p-biased distribution* on $\{0,1\}^n$ is the product distribution μ_p^n, where each coordinate is sampled independently from μ_p.

Recall from Section 9.4.1 that for every probability distribution (X,μ), every function $f : (X,\mu)^n \to \mathbb{R}$ has a unique Fourier-Walsh expansion $f = \sum_{S\subseteq[n]} F_S$, where the functions F_S are S-juntas and satisfy $\mathbb{E}_i F_S \equiv 0$ for every $i \in S$.

Fourier-Walsh expansion for the p-biased distribution: Consider the function $w : \{0,1\} \to \mathbb{R}$ with

$$w(x) \coloneqq \begin{cases} \sqrt{\frac{p}{1-p}} & \text{if } x = 0 \\ -\sqrt{\frac{1-p}{p}} & \text{if } x = 1 \end{cases},$$

which satisfies $\mathbb{E}_{\mu_p} w(\mathbf{x}) = 0$ and $\mathbb{E}_{\mu_p} w(\mathbf{x})^2 = 1$. For $S \subseteq [n]$, define $w_S : (\{0,1\}^n, \mu_p^n) \to \mathbb{R}$ as

$$w_S(x) = \prod_{i\in S} w(x_i),$$

with the convention $w_\emptyset \equiv 1$. It follows easily from $\mathbb{E}_{\mu_p} w(\mathbf{x}) = 0$ and $\mathbb{E}_{\mu_p} w(\mathbf{x})^2 = 1$ that

$$\langle w_S, w_T \rangle_{\mu_p} \coloneqq \mathbb{E}_{\mathbf{x} \sim \mu_p^n} w_S(\mathbf{x}) w_T(\mathbf{x}) = \begin{cases} 0 & S \neq T \\ 1 & S = T \end{cases}.$$

We conclude the following proposition.

Proposition 13.2. *The functions $\{w_S\}_{S \subseteq [n]}$ form an orthonormal bases for the space of functions $f : (\{0,1\}^n, \mu_p^n) \to \mathbb{R}$ with the inner product $\langle \cdot, \cdot \rangle_{\mu_p}$.*

By Proposition 13.2, we can write the Fourier-Walsh expansion of $f : (\{0,1\}^n, \mu_p^n) \to \mathbb{R}$ as

$$f = \sum_{S \subseteq [n]} \langle f, w_S \rangle_{\mu_p} w_S.$$

13.2 Phase transitions

We can represent an n-vertex undirected graph G with a vector $x_G \in \{0,1\}^{\binom{n}{2}}$ possible edges, and the value indicates whether the edge is present in G. A *graph property* can be viewed as a subset of all finite graphs, specifically consisting of those graphs that satisfy a certain condition or characteristic. Given a graph property $\mathcal{P}$ and a positive integer n, we denote the set of n-vertex graphs in $\mathcal{P}$ by $\mathcal{P}_n$.

We can identify $\mathcal{P}_n$ with Boolean function $f : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$ defined as $f(x_G) = 1$ iff $G \in \mathcal{P}_n$. Therefore, for example, we have

$$\mathbf{Pr}[G(n,p) \in \mathcal{P}] = \mathbb{E}_{\mathbf{x} \sim \mu_p^{\binom{n}{2}}} [f(\mathbf{x})].$$

Here, $\mu_p^{\binom{n}{2}}$ is the p-biased product distribution over the set of edges in the graph, and the expectation is taken over random graphs generated under this distribution.

In the study of phase transition, we focus on *monotone graph properties*, i.e., adding edges to a graph that satisfies the property results in another graph that also satisfies it. We can represent a monotone graph property by a sequence of monotone functions $f : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$.

More generally, we will discuss sequences of monotone functions $f : \{0,1\}^n \to \{0,1\}$. We will shorthand

$$\mu_p(f) \coloneqq \mathbb{E}_{\mu_p^n}[f].$$

Definition 13.3 (critical probability)**.** The *critical probably* of a non-constant monotone function $f : \{0,1\}^n \to \{0,1\}$ is the unique number $p_c \in [0,1]$ with

$$\mu_{p_c}(f) = \frac{1}{2}.$$

The following theorem due to Bollobás and Thomason [BT87] shows that every non-constant monotone function exhibits a *threshold behaviour*, meaning that in the *sub-critical regime* $p = o(p_c)$, we have

$$\mathop{\mathbf{Pr}}_{\mu_p^n}[f(\mathbf{x}) = 1] = o(1),$$

and in the *super-critical regime* $p = \Omega(p_c)$, we have

$$\mathop{\mathbf{Pr}}_{\mu_p^n}[f(\mathbf{x}) = 1] = 1 - o(1).$$

Theorem 13.4 (Bollobás and Thomason [BT87])**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a sequence of non-constant monotone functions, and let p_c be the critical probability so that $\mu_{p_c}(f) = \frac{1}{2}$.*

(i) For $p = o(p_c)$, we have

$$\mu_p(f) = o(1).$$

(ii) For $p = \Omega(p_c)$, we have

$$\mu_p(f) = 1 - o(1).$$

Proof. We only prove (i) as (ii) follows from a similar argument. To prove (i), we show that for every $\varepsilon > 0$, there exists $m \in \mathbb{N}$ such that $p_m \coloneqq p_c/m$ satisfies

$$\mu_{p_m}(f) \leqslant \varepsilon.$$

Let m be the smallest natural number with

$$\frac{1}{2} < 1 - (1-\varepsilon)^m. \tag{13.1}$$

Note that m does not depend on n. Sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \sim (\{0,1\}^n, \mu_{p_m}^n)$ independently and let $\widetilde{\mathbf{x}} = \mathbf{x}^{(1)} \vee \ldots \vee \mathbf{x}^{(m)}$. We have $\widetilde{\mathbf{x}} \sim \mu_q^n$ where $q = 1 - (1-p_m)^m \leqslant p_m m \leqslant p_c$. Since $q \leqslant p_c$, we have

$$\frac{1}{2} \geqslant \mu_q(f) = \mathbf{Pr}[f(\widetilde{\mathbf{x}}) = 1] \geqslant \mathbf{Pr}[f(\mathbf{x}^{(1)}) = 1 \vee \ldots \vee f(\mathbf{x}^{(m)}) = 1] = 1 - (1 - \mu_{p_m}(f))^m.$$

Therefore, by (13.1), we have must have $\mu_{p_m}(f) \leqslant \varepsilon$ as desired.

□

13.2.1 Sharpness of threshold: The Margulis-Russo formula

While Theorem 13.4 shows that the phase transition occurs in an interval of length $O(p_c)$ around the critical probably, for many natural properties (e.g., connectivity, 3-colourability, satisfiability of a random instance of 3SAT), this interval is much smaller.

Definition 13.5 (Critical interval)**.** Let $\varepsilon > 0$ be a fixed number. The *ε-critical interval* for a non-constant monotone function $f : \{0,1\}^n \to \{0,1\}$ is $[p_0, p_1]$ where $\mu_{p_0}(f) = \varepsilon$ and $\mu_{p_1}(f) = 1 - \varepsilon$.

Definition 13.6 (Sharp threshold)**.** A sequence of non-constant monotone functions $f : \{0,1\}^n \to \{0,1\}$ with critical probability $p_c = p_c(n)$ exhibits *sharp threshold* if for every $\varepsilon > 0$, the ε-critical interval is of length $o(p_c)$.

In other words, a sharp threshold indicates that the phase transition occurs in the interval $[p_c(1-o(1)), p_c(1+o(1)]$. Not every graph property has a sharp threshold. For example, "local properties" such as containing a triangle, do not exhibit sharp thresholds.

Since the speed at which $\mu_p(f)$ changes is captured by the derivative $\frac{d\mu_p(f)}{dp}$, we have the following easy observation.

Proposition 13.7 (Coarse threshold)**.** *If a sequence of non-constant monotone functions $f : \{0,1\}^n \to \{0,1\}$ with critical probability $p_c = p_c(n)$ does* not *exhibit a sharp threshold, then there exists fixed constants $C, \eta, \varepsilon > 0$ and a parameter $p = p(n) \in [(1-\eta)p_c, (1+\eta)p_c]$ such that*

$$\varepsilon < \mu_p(f) < 1 - \varepsilon \quad \text{and} \quad \frac{d\mu_p(f)}{dp} \leqslant \frac{C}{p}.$$

On the other hand, the following theorem of Margulis [Mar74] and Russo [Rus81] relates $\frac{d\mu_p(f)}{dp}$ to the total influence of f with respect to μ_p^n.

Theorem 13.8 (Margulis-Russo formula)**.** *Every monotone function $f : (\{0,1\}^n, \mu_p^n) \to \{0,1\}$ satisfies*

$$p(1-p)\frac{d\mu_p(f)}{dp} = I_f,$$

where $\mu_p(f) = \mathbb{E}_{\mu_p^n}[f]$.

Proof. Since

$$\mu(x) = p^{|x|}(1-p)^{n-|x|} \quad \text{and} \quad \mu_p(f) = \sum_{x \in \mathrm{supp}(f)} p^{|x|}(1-p)^{n-|x|},$$

we have

$$\begin{aligned}\frac{d\mu_p(f)}{dp} &= \sum_{x\in\mathrm{supp}(f)} |x|\frac{\mu(x)}{p} - \sum_{x\in\mathrm{supp}(f)} (n-|x|)\frac{\mu(x)}{1-p} \\ &= \sum_{i=1}^{n}\sum_{x\in\mathrm{supp}(f)} \frac{\mu(x)}{p}\mathbf{1}_{[x_i=1]} - \sum_{i=1}^{n}\sum_{x\in\mathrm{supp}(f)} \frac{\mu(x)}{1-p}\mathbf{1}_{[x_i=0]} \\ &= \sum_{i=1}^{n}\mathbb{E}[f|\mathbf{x}_i=1] - \sum_{i=1}^{n}\mathbb{E}[f|\mathbf{x}_i=0].\end{aligned}$$

On the other hand, since a non-constant Boolean function on $(\{0,1\},\mu_p)$ has variance $p(1-p)$, we have

$$I_i(f) = p(1-p)\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mu_p^n}[f(\mathbf{x}) \neq f(\mathbf{x}\oplus e_i)].$$

Since f is monotone,

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mu_p^n}[f(\mathbf{x}) \neq f(\mathbf{x}\oplus e_i)] &= \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mu_p^n}[f(\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},1,\mathbf{x}_{i+1},\ldots,\mathbf{x}_n) = 1 \wedge f(\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},0,\mathbf{x}_{i+1},\ldots,\mathbf{x}_n) = 0] \\ &= \mathbb{E}[f|\mathbf{x}_i=1] - \mathbb{E}[f|\mathbf{x}_i=0],\end{aligned}$$

which shows $I_i(f) = p(1-p)\left(\mathbb{E}[f|\mathbf{x}_i=1] - \mathbb{E}[f|\mathbf{x}_i=0]\right)$ and completes the proof. □

Theorem 13.8 and Proposition 13.7 show that if a sequence of monotone functions $f : \{0,1\}^n \to \{0,1\}$ does not exhibit a sharp threshold, then for some p in the critical interval, we have $\mathrm{Var}_{\mu_p}(f) = \Omega(1)$ and $I_f = O(1)$. Therefore, characterizing properties that do not exhibit sharp thresholds is equivalent to characterizing the monotone Boolean functions with total influence $O(1)$ in the p-biased setting.

13.2.2 KKL and the length of the critical interval

By applying generalized hypercontractivity for the p-biased distribution in the proof of the KKL theorem, one can obtain the following generalization of the KKL inequality due to Talagrand [Tal94].

Theorem 13.9 (Talagrand's generalization of KKL inequality [Tal94])**.** *Let $f : (\{0,1\}^n,\mu_p^n) \to \{0,1\}$ be such that $\mathbb{E}[f] = \alpha$. Denoting $I_{\max} \coloneqq \max_i I_i(f)$ and $p' = \min(p,1-p)$, we have*

$$I_{\max} = \Omega\left(\frac{1}{\log(1/p')}\mathrm{Var}[f]\frac{\log n}{n}\right).$$

Proof. We assume $p \leqslant 1/2$ and $p' = p$; The case where $p > 1/2$ can be reduced to the case $p \leqslant 1/2$ by interchanging the roles of 0's and 1's.

The proof is identical to the proof of the KKL inequality, except we will apply the generalization of the hypercontractivity for the p-biased distribution (Corollary 10.20).

Consider the Fourier-Walsh expansion $f = \sum_S F_S$, and recall that

$$\mathrm{Var}[f] = \sum_{|S|>0}\|F_S\|_2^2.$$

Bounding high frequencies: Let k to be determined later. Since

$$I_f = \sum_S |S|\|F_S\|_2^2 \geqslant k\sum_{|S|>k}\|F_S\|_2^2,$$

we have

$$\sum_{|S|>k}\|F_S\|_2^2 \leqslant \frac{I_f}{k} \leqslant \frac{nI_{\max}}{k}.$$

Bounding low frequencies: To handle the low degree part, let $f_i(x) \coloneqq \partial_i f(x)$. By applying hypercontractivity (Corollary 10.20 with $\lambda = \min(p, 1-p) = p$) to f_i for the $\|\cdot\|_{4/3}$, we have

$$\sum_{1 \leqslant |S| \leqslant k} \|F_S\|_2^2 \leqslant \sum_{i=1}^{n} \sum_{\substack{i \in S \\ |S| \leqslant k}} \|F_S\|_2^2 = \sum_{i=1}^{n} \left\| f_i^{\leqslant k} \right\|_2^2 \leqslant \sum_{i=1}^{n} (3/p)^k \|f_i\|_{4/3}^2 .$$

By Proposition 12.2,, we have

$$\sum_{i=1}^{n} (3/p)^k \|f_i\|_{4/3}^2 = (3/p)^k \sum_{i=1}^{n} (2I_i(f))^{3/2} \leqslant 4(3/p)^k n I_{\max}^{3/2}.$$

Combining the bounds for low-degree and high-degree parts, we get

$$\mathrm{Var}[f] = \sum_{S : |S| \geqslant 1} \|F_S\|_2^2 \leqslant \frac{n I_{\max}}{k} + 4(3/p)^k n I_{\max}^{3/2} = \left(\frac{n}{k} + 4(3/p)^k n I_{\max}^{1/2} \right) I_{\max}.$$

Taking $k = c \log(1/p) \log(n)$ for sufficiently small $c > 0$ implies the theorem. □

Let $f : (\{0,1\}, \mu_p)^{\binom{n}{2}} \to \{0,1\}$ represent a graph property. Since f is invariant under graph automorphisms, all the variables have the same influence and therefore, by Theorem 13.9, we have

$$I_f = \binom{n}{2} I_{\max} = \Omega\left(\frac{1}{\log(1/p')} \mathrm{Var}(f) \log(n) \right),$$

where $p' = \min(p, 1-p)$. In combination with the Margulis-Russo formula, we have

$$\frac{d\mu_p(f)}{dp} = \frac{I_f}{p(1-p)} = \Omega\left(\frac{1}{p' \log(1/p')} \mathrm{Var}(f) \log(n) \right) = \Omega\left(\mathrm{Var}(f) \log(n) \right). \tag{13.2}$$

Note that in the ε-critical interval, we have $\mathrm{Var}[f] \geqslant \varepsilon(1-\varepsilon) = \Omega_\varepsilon(1)$ and therefore $\frac{d\mu_p(f)}{dp} = \Omega_\varepsilon(\log(n))$. It follows that for monotone graph properties, the ε-critical interval is always of length $O_\varepsilon\left(\frac{1}{\log(n)}\right)$, which is a theorem due to Friedgut and Kalai [FK96]. In fact, Equation (13.2) implies the following stronger theorem.

Theorem 13.10. *Let $\varepsilon > 0$ be a fixed constant. If $\mathcal{P}$ is a monotone property graph property with $\mu_{p_0}(\mathcal{P}) = \varepsilon$ and $\mu_{p_1}(\mathcal{P}) = 1 - \varepsilon$, then*

$$p_1 = p_0 + O_\varepsilon\left(\frac{p_1 \log(2/p_1)}{\log(n)} \right).$$

Proof. If $p_1 \geqslant \frac{1}{2}$, then the assertion is equivalent to

$$p_1 = p_0 + O_\varepsilon\left(\frac{1}{\log(n)} \right),$$

which follows from $\frac{d\mu_p(f)}{dp} = \Omega_\varepsilon(\log(n))$. Therefore, assume $p_1 < \frac{1}{2}$. In this case, by Equation (13.2), for every $p \in [p_0, p_1]$, we have

$$\frac{d\mu_p(f)}{dp} = \Omega_\varepsilon\left(\frac{\log(n)}{p_1 \log(1/p_1)} \right),$$

and the theorem follows. □

Remark 13.11. If $\log(2/p_1) = o(\log(n))$, then by Theorem 13.10, the montone graph property $\mathcal{P}$ must exhibit a sharp threshold.

Chapter 14

Low-degree Fourier-Walsh expansions

In Section 10.4, we discussed an extension of hypercontractivity that applies to functions over arbitrary product spaces (X^n, μ^n). However, this general form of hypercontractivity requires the noise parameter $\rho > 0$ to depend on $\lambda = \min_{a \in X} \mu(a)$. Unfortunately, this dependence weakens the result when λ is small, which is often the case when studying phase transitions. Moreover, this general form of hypercontractivity becomes entirely ineffective for continuous spaces such as $[0,1]^n$.

This chapter will present an inequality due to Bourgain that is closely related to hypercontractivity. Crucially, Bourgain's inequality remains applicable when the original hypercontractivity breaks down, as it imposes no dependency on the underlying probability distribution (X, μ).

Theorem 14.1 (Bourgain [Bou80])**.** *Let (X, μ) be a probability space. Consider $f : (X^n, \mu^n) \to \mathbb{R}$ and its Fourier-Walsh expansion $f = \sum_{S \subseteq [n]} F_S$. Let $1 < p \leqslant 2 \leqslant q < \infty$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$.*

(i) For every integer $k \geqslant 0$, we have

$$\left\| \left(\sum_{S : |S| \leqslant k} F_S^2 \right)^{1/2} \right\|_q \leqslant 2^{5qk} \|f\|_q \quad \text{and} \quad \left\| \left(\sum_{S : |S| \leqslant k} F_S^2 \right)^{1/2} \right\|_p \leqslant 2^{5qk} \|f\|_p .$$

(ii) We also have the following inequalities in the opposite direction about the norms of $f^{\leqslant k}$:

$$\|f^{\leqslant k}\|_q \leqslant 2^{5qk} \left\| \left(\sum_{S : |S| \leqslant k} F_S^2 \right)^{1/2} \right\|_q \text{ and } \|f^{\leqslant k}\|_p \leqslant 2^{5qk} \left\| \left(\sum_{S : |S| \leqslant k} F_S^2 \right)^{1/2} \right\|_p .$$

(iii) Consequently,

$$\|f^{\leqslant k}\|_q \leqslant 2^{10qk} \|f\|_q \text{ and } \|f^{\leqslant k}\|_p \leqslant 2^{10qk} \|f\|_p .$$

In Chapter 15, we will apply Theorem 14.1 to prove a fundamental theorem about sharp thresholds. We will present the proof of Theorem 14.1 in Section 14.2. First, we prove some preliminary lemmas for the proof of Theorem 14.1.

14.1 Preliminary lemmas

We start with a simple technical inequality.

Lemma 14.2. *For every even integer $q > 0$, and $0 \leqslant \rho \leqslant 2^{-3q}$, the inequality*

$$(1 - \rho x)^q + q\rho x \leqslant (1 + x)^q - qx$$

holds for all real $x \in \mathbb{R}$.

Proof. We have

$$(1+x)^q - qx = 1 + \sum_{r=2}^{q} \binom{q}{r} x^r \geqslant 1 + \sum_{r=2}^{q} \binom{q}{r} (-|x|)^r = (1-|x|)^q + q|x|. \tag{14.1}$$

We claim that for $\tau = 2^{-2q}$, and every positive $t \geqslant 0$,

$$(1-t)^q + qt \geqslant 1 + \tau(t^2 + t^q).$$

To prove this claim not that the inequality is trivial for $q = 2$, so we assume $q \geqslant 4$. We verify the inequality in three different intervals.

- For $t \geqslant 2$, we have $t - 1 \geqslant t/2$, which shows $(1-t)^q + qt \geqslant 1 + (1-t)^q \geqslant 1 + \tau(t^2 + t^q)$.
- For $t \in [1/2, 2]$, $qt \geqslant 4t \geqslant 1 + 2t \geqslant 1 + \tau(t^2 + t^q)$.
- For $t \in [0, 1/2]$, define $f(t) \coloneqq (1-t)^q + qt - 1 - \tau(t^2 + t^q)$. We have

  $$f'(t) = -q(1-t)^{q-1} + q - \tau(2t + qt^{q-1})$$

  and

  $$f''(t) = q(q-1)\left[(1-t)^{q-2} - \tau t^{q-2}\right] - 2\tau.$$

  Note that $f(0) = f'(0) = 0$, and for every $t \in [0, 1/2]$, we have

  $$f''(t) = q(q-1)\left[(1-t)^{q-2} - \tau t^{q-2}\right] - 2\tau \geqslant 2(2^{2-q} - \tau 2^{2-q}) - 2\tau > 0.$$

  Therefore, $f(t) \geqslant 0$ in $[0, 1/2]$ as desired.

Combining with Equation (14.1), for $\tau = 2^{-2q}$, we have

$$(1+x)^q - qx \geqslant (1-|x|)^q + q|x| \geqslant 1 + \tau(x^2 + x^q).$$

On the other hand, since for every $x \in \mathbb{R}$ and $2 \leqslant r \leqslant q$, we have $x^r \leqslant x^2 + x^q$ (recall that q is even), we conclude that for $0 \leqslant \rho \leqslant 2^{-q}\tau$,

$$\begin{aligned}(1-\rho x)^q + q\rho x = 1 + \sum_{r=2}^{q} \binom{q}{r} (-\rho x)^r &\leqslant 1 + \sum_{r=2}^{q} \binom{q}{r} (\rho^2 x^2 + \rho^q x^q) \\ &\leqslant 1 + 2^q(\rho^2 x^2 + \rho^q x^q) \leqslant 1 + \tau(x^2 + x^q).\end{aligned}$$

□

Let (X, μ) be a probability space, and consider $f : (X, \mu) \to \mathbb{R}$. Note that the Fourier-Walsh expansion of f is $f = F_\emptyset + F_{\{1\}}$ where $F_\emptyset = \mathbb{E}[f]$ and $F_{\{1\}} = f - \mathbb{E}[f]$. Recall the noise operator T_ρ for general probability spaces from Section 10.4. We have

$$T_\rho f \coloneqq F_\emptyset + \rho F_{\{1\}}.$$

Lemma 14.3 (A dimension-one inequality)**.** *Let (X, μ) be a probability space, and consider $f : (X, \mu) \to \mathbb{R}$.*

(i) For every $1 \leqslant p \leqslant \infty$ and $0 \leqslant \rho \leqslant 1$, we have

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

(ii) Suppose $\frac{1}{p} + \frac{1}{q}$ for $1 < p \leqslant 2 \leqslant q < \infty$. For every $0 \leqslant \rho \leqslant 2^{-4q}$, we have

$$\|T_{-\rho} f\|_q \leqslant \|f\|_q,$$

and

$$\|T_{-\rho} f\|_p \leqslant \|f\|_p.$$

Proof. When $\rho \in [0,1]$, the noise operator T_ρ is an averaging operator, which easily implies that T_ρ is contracting. We give the formal proof below.

Proof of (i): Since

$$T_\rho f = \mathbb{E}[f] + \rho(f - \mathbb{E}[f]) = \rho f + (1-\rho)\mathbb{E}[f],$$

we have

$$\|T_\rho f\|_p \leqslant \rho\|f\|_p + (1-\rho)\|\mathbb{E}[f]\|_p \leqslant \rho\|f\|_p + (1-\rho)\|f\|_p \leqslant \|f\|_p.$$

Proof of (ii): If $\mathbb{E}[f] = 0$, then $\|T_{-\rho} f\|_q = \rho\|f\|_q$ and the theorem follows. Otherwise, we can normalize f to $\frac{f}{\mathbb{E}[f]}$ and assume, without loss of generality, $\mathbb{E}[f] = 1$. Denote $F \coloneqq F_{\{1\}} = f - 1$.

First, we verify the statement for *even* integers $q \geqslant 2$. By Lemma 14.2, for every $0 \leqslant \rho \leqslant 2^{-3q}$, we have

$$(1 - \rho F(x))^q + q\rho F(x) \leqslant (1 + F(x))^q - qF(x),$$

for all $x \in X$. Taking the expected value over $\mathbf{x} \sim (X, \mu)$, and using $\mathbb{E}F(\mathbf{x}) = 0$, we get

$$\mathbb{E}(1 - \rho F(\mathbf{x}))^q \leqslant \mathbb{E}(1 + F(\mathbf{x}))^q,$$

which shows

$$\|T_{-\rho} f\|_q = \|1 - \rho F\|_q \leqslant \|1 + F\|_q = \|f\|_q,$$

as desired.

Next we consider arbitrary $q \in [2, \infty)$. Let $q_0 \geqslant 2$ be the largest even integer satisfying $q_0 \leqslant q$, and let $q_1 = q_0 + 2$. For $\rho \leqslant 2^{-4q} \leqslant 2^{-3q_1} \leqslant 2^{-3q_0}$, we have

$$\|T_{-\rho} f\|_{q_0} \leqslant \|f\|_{q_0},$$

and

$$\|T_{-\rho} f\|_{q_1} \leqslant \|f\|_{q_1},$$

for all f. Since $q_0 \leqslant q \leqslant q_1$, it follows from the Riesz–Thorin's interpolation theorem that

$$\|T_{-\rho} f\|_q \leqslant \|f\|_q.$$

It remains to handle the case $1 < p < 2$. In this case, we have

$$\|T_{-\rho} f\|_p = \sup_{g:\|g\|_q \leqslant 1} \langle T_{-\rho} f, g\rangle = \sup_{g:\|g\|_q \leqslant 1} \langle f, T_{-\rho} g\rangle \leqslant \|f\|_p \sup_{g:\|g\|_q \leqslant 1} \|T_{-\rho} g\|_q \leqslant \|f\|_p.$$

□

We have the following corollary, which, together with the classical hypercontractivity, are the main ingredients in the proof of Theorem 14.1.

Corollary 14.4. *Consider $f : (X, \mu)^n \to \mathbb{R}$ and its Fourier-Walsh expansion $f = \sum_{S \subseteq [n]} F_S$, and suppose $1 < p \leqslant 2 \leqslant q < \infty$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. For $y \in \{-1,1\}^n$ and $S \subseteq [n]$, let $w_S(y) \coloneqq \prod_{i \in S} y_i$.*

(i) For every $0 \leqslant \rho \leqslant 2^{-4q}$ and every $y \in \{-1,1\}^n$, we have

$$\left\| \sum_{S \subseteq [n]} \rho^{|S|} w_S(y) F_S(x) \right\|_q \leqslant \|f\|_q \quad \text{and} \quad \left\| \sum_{S \subseteq [n]} \rho^{|S|} w_S(y) F_S(x) \right\|_p \leqslant \|f\|_p.$$

(ii) For every $0 \leqslant \rho \leqslant 2^{-4q}$ and every $y \in \{-1,1\}^n$, we have

$$\left\| \sum_{S \subseteq [n]} \rho^{|S|} F_S(x) \right\|_q \leqslant \left\| \sum_{S \subseteq [n]} w_S(y) F_S(x) \right\|_q \quad \text{and} \quad \left\| \sum_{S \subseteq [n]} \rho^{|S|} F_S(x) \right\|_p \leqslant \left\| \sum_{S \subseteq [n]} w_S(y) F_S(x) \right\|_p.$$

Proof. For $i \in [n]$, let $T_\rho^{(i)}$ be the noise operator applied only to the ith coordinate of f. In other words,

$$T_\rho^{(i)} f = \sum_{S: i \notin S} F_S + \rho \sum_{S: i \in S} F_S.$$

By Lemma 14.3 (i), we have

$$\|T_\rho^{(i)} f\|_q \leqslant \|f\|_q,$$

and by Lemma 14.3 (ii), we have

$$\|T_{-\rho}^{(i)} f\|_q \leqslant \|f\|_q.$$

Since

$$\sum_{S \subseteq [n]} \rho^{|S|} w_S(y) F_S(x) = T_{y_1\rho}^{(1)} \circ \ldots \circ T_{y_n\rho}^{(n)} f,$$

the above two inequalities show that for every $y \in \{-1,1\}^n$, we have

$$\left\| \sum_{S \subseteq [n]} \rho^{|S|} w_S(y) F_S(x) \right\|_q \leqslant \|f\|_q.$$

The same proof applies to $\|\cdot\|_p$.

Finally, note that

$$\sum_{S \subseteq [n]} \rho^{|S|} F_S(x) = \sum_{S \subseteq [n]} \rho^{|S|} w_S(y) \times w_S(y) F_S(x),$$

and therefore, (ii) follows from applying (i) to $\sum_{S \subseteq [n]} w_S(y) F_S(x)$. □

14.2 Proof of Theorem 14.1

Recall $1 < p \leqslant 2 \leqslant q < \infty$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$ and we have $\rho = 2^{-4q}$. It is clear that (iii) follows from (i) and (ii). We present the proofs of (i) and (ii) below.

Proof of Theorem 14.1 (i): By Corollary 14.4 (i), for every $y \in \{-1,1\}^n$, we have

$$\left\| \sum_{S \subseteq [n]} \rho^{|S|} w_S(y) F_S(x) \right\|_{L_q(x)} \leqslant \|f\|_q,$$

which,by taking the $\|\cdot\|_{L_q(y)}$ norm, shows

$$\left\| \sum_{S \subseteq [n]} \rho^{|S|} w_S(y) F_S(x) \right\|_{L_q(x,y)} \leqslant \|f\|_q.$$

By the classical hypercontractivity for $\|\cdot\|_{L_q(y)}$ on $\{-1,1\}^n$, i.e, Theorem 10.14 (i), we have

$$\begin{aligned}
\left\|\sum_{S\subseteq[n]}\rho^{|S|}w_S(y)F_S(x)\right\|_{L_q(x,y)} &= \left\|\left\|\sum_{S\subseteq[n]}\rho^{|S|}F_S(x)w_S(y)\right\|_{L_q(y)}\right\|_{L_q(x)} \\
&\geqslant \left\|\left\|\sum_{S\subseteq[n]}\left(\frac{\rho}{\sqrt{q-1}}\right)^{|S|}F_S(x)w_S(y)\right\|_{L_2(y)}\right\|_{L_q(x)} \\
&= \left\|\left(\sum_{S\subseteq[n]}\left(\frac{\rho}{\sqrt{q-1}}\right)^{2|S|}F_S^2\right)^{1/2}\right\|_q \geqslant (q-1)^{-k/2}\rho^k\left\|\left(\sum_{|S|\leqslant k}F_S^2\right)^{1/2}\right\|_q.
\end{aligned}$$

Therefore,

$$(q-1)^{-k/2}\rho^k\left\|\left(\sum_{|S|\leqslant k}F_S^2\right)^{1/2}\right\|_q \leqslant \|f\|_q, \tag{14.2}$$

which finishes the proof of Part (i) for $\|\cdot\|_q$ since $\sqrt{q-1}\rho^{-1}\leqslant 2^{5q}$. The proof for $\|\cdot\|_p$ is identical.

Proof of Theorem 14.1 (ii): By Corollary 14.4 (ii), for every $y\in\{-1,1\}^n$, we have

$$\left\|f^{\leqslant k}\right\|_q \leqslant \left\|\sum_{|S|\leqslant k}\rho^{-|S|}w_S(y)F_S\right\|_q$$

Taking the L_q norm on y and applying the classical hypercontractivity, we have

$$\begin{aligned}
\left\|f^{\leqslant k}\right\|_q &\leqslant \left\|\left\|\sum_{|S|\leqslant k}\rho^{-|S|}w_S(y)F_S\right\|_{L_q(y)}\right\|_{L_q(x)} \\
&\leqslant (q-1)^{k/2}\left\|\left\|\sum_{|S|\leqslant k}\rho^{-|S|}w_S(y)F_S\right\|_{L_2(y)}\right\|_{L_q(x)} \\
&\leqslant (q-1)^{k/2}\left\|\left(\sum_{|S|\leqslant k}\rho^{-2|S|}F_S^2\right)^{1/2}\right\|_q \\
&\leqslant (q-1)^{k/2}\rho^{-k}\left\|\left(\sum_{|S|\leqslant k}F_S^2\right)^{1/2}\right\|_q
\end{aligned}$$

which finishes the proof of Part (ii) for $\|\cdot\|_q$ since $\sqrt{q-1}\rho^{-1}\leqslant 2^{5q}$. The proof for $\|\cdot\|_p$ is identical.

Chapter 15

Friedgut-Bourgain's threshold theorems

The problem of finding general conditions under which a sharp threshold occurs is first investigated by Russo [Rus81, Rus82]. In Chapter 13, we showed that if the critical probability p_c of a monotone graph property satisfies $p_c = 2^{-o(\log(n))}$, then the property exhibits a sharp threshold. However, the critical probability p_c tends to be much smaller for many interesting graph properties. For instance, consider the well-studied case of graph connectivity, where the critical probability is $p_c = \frac{\ln(n)}{n}$. While the transition for connectivity is known to be sharp [EoR59], this sharpness does not follow from the earlier discussed results of Chapter 13 such as Theorem 13.10.

As we discussed earlier, the Margulis-Russo formula (Theorem 13.8) states that

$$p(1-p)\frac{d\mu_p(f)}{dp} = I_f,$$

and therefore, having a *coarse* threshold is equivalent to $I_f = O(1)$ for some p in the critical interval.

In the setting of the p-biased distribution, when p is not too small, the works of Talagrand [Tal94], Friedgut and Kalai [FK96], Bourgain and Kalai [BK99], and Friedgut [Fri98] provide a satisfactory understanding of the functions with small total influences. Intuitively, these results say that the total influence of f is large unless the value of $f(x)$ is determined only by "local information" about x, such as a small number of coordinates. However, as the following simple example illustrates, these results lose relevance when p is small, particularly when $\log \frac{1}{p} \sim \log n$, which is often the case in applications.

Example 15.1. Set $p = n^{-1}$, and define $f : \{0,1\}^n \to \{0,1\}$ as $f(x) = 1$ if and only if $x \neq (0,\ldots,0)$. Then $I_1(f) = \ldots = I_n(f) \leqslant p(1-p)$, and so $I_f \leqslant 1$. However, f does not depend only on a small set of coordinates. Indeed for every constant size set $J \subseteq [n]$, we have

$$\mathbb{E}_{\mu_p^n}[f(\mathbf{x})|\mathbf{x}_J = (0,\ldots,0)] = 1 - (1-p)^{n-|J|} = 1 - \frac{1}{e} \pm o(1),$$

Since

$$\mathbf{Pr}[\mathbf{x}_J = (0,\ldots,0)] \geqslant 1 - |J|p = 1 - o(1),$$

we have $\|f - g\|_1 \geqslant \frac{1}{e} - o(1)$ for every function g that depends only on the coordinates in J.

No significant progress on the case of small p was made until Friedgut's breakthrough work [Fri99], which characterized graph and hypergraph properties with small total influences. Friedgut used his theorem to show that the satisfiability of a random k-CNF Boolean formula exhibits a sharp threshold. Friedgut's sharp threshold theorem is now indispensable for studying threshold behaviour. We refer the reader to [Fri05]) for many applications of this theorem in establishing sharp thresholds for various graph and constraint satisfaction properties.

Given a set $\mathcal{H}$ of graphs, let the upper set of $\mathcal{H}$ be the set of all graphs that contain some $H \in \mathcal{H}$ as a (not necessarily induced) subgraph:

$$\mathcal{H}^{\uparrow} \coloneqq \{G : \exists H \in \mathcal{H} \text{ s.t. } H \subseteq G\}$$

Roughly speaking, Friedgut's theorem says that a monotone graph property with total influence $O(1)$ can be approximated by some $\mathcal{H}^{\uparrow}$ where graphs in $\mathcal{H}$ are all of size $O(1)$.

Theorem 15.2 (Friedgut [Fri99]). *For every integer $I > 0$ and real $\varepsilon > 0$, there exists a constant $k(I, \varepsilon) > 0$ such that the following holds. Let $p > 0$ and let $\mathcal{P}$ be a monotone graph property of n-vertex graphs with total influence at most I. There exists a collection $\mathcal{H}$ of graphs such that the following holds.*

- *Every $H \in \mathcal{H}$ has at most $k(I, \varepsilon)$ edges.*
- *We have $\mu_p(\mathcal{P} \Delta \mathcal{H}^{\uparrow}) \leqslant \varepsilon$, where $\mathcal{P} \Delta \mathcal{H}^{\uparrow}$ denotes the symmetric difference between $\mathcal{P}$ and $\mathcal{H}^{\uparrow}$.*

Since a graph property is invariant under graph isomorphisms, its corresponding Boolean function is invariant under all permutations of the coordinates corresponding to permuting the graph's vertices. Friedgut's proof leverages this symmetry extensively and in many steps of his proof. Nevertheless, he conjectured [Fri99, Conjecture 1.5] that his theorem holds without requiring symmetry assumptions. To state his conjecture, we recall the definition of a DNF.

A DNF (Disjunctive Normal Form) is a Boolean formula consisting of a disjunction (logical OR) of conjunctions (logical AND) of Boolean variables or their negations (NOTs). A DNF is a *monotone DNF* if it does not involve any negated variable. The *width* of a DNF is the size of the largest conjunction (AND-clause) in the DNF, where the size is the number of variables in the clause. A DNF of width k is called a k-DNF. For example $(x_1 \wedge x_2 \wedge x_4) \vee (x_2 \wedge x_3)$ is a monotone 3-DNF.

Proposition 15.3. *If $f : (\{0,1\}^n, \mu_p^n) \to \{0,1\}$ is representable by a k-DNF, then*

$$I_f \leqslant k.$$

Proof. For every $x \in \{0,1\}^n$, let $s_{1\to 0}(x)$ denote the number of coordinates $i \in [n]$ such that $f(x) = 1$ and $f(x \oplus e_i) = 0$. If $f(x) = 1$, then x satisfies at least one clause C. If $f(x \oplus e_i) = 0$, then C must involve x_i or $\neg x_i$, and since there are at most k literals in C, we have $s_{1\to 0}(x) \leqslant k$ for every x. Therefore,

$$I_f = \sum_{i=1}^{n} p(1-p)\,\mathbf{Pr}[f(\mathbf{x}) \neq f(\mathbf{x} \oplus e_i)] \leqslant \sum_{i=1}^{n} \mathbf{Pr}[f(\mathbf{x}) = 1 \wedge f(\mathbf{x} \oplus e_i) = 0] = \mathbb{E}_{\mathbf{x}} s_{1\to 0}(\mathbf{x}) \leqslant k.$$

□

Friedgut conjectures that every monotone function with total influence $O(1)$ is approximately a monotone DNF of width $O(1)$.

Conjecture 15.4 (Friedgut [Fri99, Conjecture 1.5]). *For every integer $I > 0$ and real $\varepsilon > 0$, there exists a constant $k(I, \varepsilon) > 0$ such that the following holds. Let $p > 0$ and $f : (\{0,1\}^n, \mu_p^n) \to \{0,1\}$ have total influence at most I. There exists $g : \{0,1\}^n \to \{0,1\}$ that is representable by a monotone DNF of width at most $k(I, \varepsilon)$ and satisfies*

$$\mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mu_p^n}[f(\mathbf{x}) \neq g(\mathbf{x})] \leqslant \varepsilon.$$

15.1 Bourgain's theorem

While Conjecture 15.4 is still open, [Bou99b] and [Hat12] have made some progress towards describing functions with a small total influence on the p-biased distribution. In particular, Bourgain's theorem, published as an appendix to Friedgut's paper [Fri99], suffices for all known applications where the goal is to prove that some concrete property exhibits a sharp threshold. In this section, we will state and prove Bourgain's theorem.

Recall from Section 9.4 that given a function $f : (X, \mu)^n \to \mathbb{R}$ and a set $S \subseteq [n]$, the notation $\mathbb{E}_S f$ denotes the function $\mathbb{E}_S f : X^n \to \mathbb{R}$ with

$$(\mathbb{E}_S f)(y) = \mathbb{E}_{\mathbf{x}_S} f(\mathbf{x}_S, y_{\overline{S}}).$$

For $S \subseteq [n]$, denoting the complement of S by $\overline{S} \coloneqq S^c = [n] \setminus S$, the Fourier-Walsh expansion of f is given by $f = \sum_{S \subseteq [n]} F_S$, where

$$F_S = \sum_{T \subseteq S} (-1)^{|S \setminus T|} \mathbb{E}_{\overline{T}} f.$$

Theorem 15.5 (Bourgain's sharp threshold theorem)**.** *Let (X,μ) be a probability space. Consider $f:(X,\mu)^n\to\{0,1\}$ with $\mathrm{Var}(f)=\varepsilon>0$ and Fourier-Walsh expansion $f=\sum F_S$. For $k=\frac{4I_f}{\varepsilon}$, we have*

$$\mathbb{E}_{\mathbf{x}}\max_{S:|S|\leqslant k}|F_S(\mathbf{x})|\geqslant 2^{-O(I_f^2/\varepsilon^2)}. \tag{15.1}$$

In particular,

$$\mathbb{E}_{\mathbf{x}}\max_{S:|S|\leqslant k}|\mathbb{E}_{\overline{S}}[f(\mathbf{x})]-\mathbb{E}[f]|\geqslant 2^{-O(I_f^2/\varepsilon^2)},$$

which shows there is $S\subseteq[n]$ with $|S|\leqslant k$ and $x\in X^n$ with

$$|\mathbb{E}_{\overline{S}}[f(x)]-\mathbb{E}[f]|\geqslant 2^{-O(I_f^2/\varepsilon^2)}.$$

Proof. First, note that since

$$|F_S|=\left|\sum_{T\subseteq S}(-1)^{|S\setminus T|}\mathbb{E}_{\overline{T}}f\right|=\left|\sum_{T\subseteq S}(-1)^{|S\setminus T|}\mathbb{E}_{\overline{T}}[f-\mathbb{E}[f]]\right|\leqslant 2^{|S|}\max_{T\subseteq S}|\mathbb{E}_{\overline{T}}[f-\mathbb{E}[f]]|,$$

the second assertion follows from Equation (15.1) as claimed.

It remains to prove Equation (15.1). We have

$$\varepsilon=\mathrm{Var}(f)=\sum_{S\neq\emptyset}\|F_S\|_2^2.$$

Bounding high frequencies: Since $k=\frac{4I_f}{\varepsilon}$ and

$$I_f=\sum_S|S|\|F_S\|_2^2\geqslant k\sum_{|S|>k}\|F_S\|_2^2,$$

we have

$$\sum_{|S|>k}\|F_S\|_2^2\leqslant\frac{I_f}{k}\leqslant\frac{\varepsilon}{4},$$

and therefore,

$$\mathbb{E}\left[\sum_{\substack{S\neq\emptyset\\|S|\leqslant k}}F_S^2\right]\geqslant\frac{3\varepsilon}{4}. \tag{15.2}$$

Dealing with low frequencies: Let real $\delta>0$ and $M\in\mathbb{N}$ be parameters to be determined later. Define

$$\eta_i(x):=\begin{cases}1 & \sum_{\substack{i\in S\\|S|\leqslant k}}F_S(x)^2\geqslant\delta\\ 0 & \text{otherwise}\end{cases},$$

where we say that the coordinate $i\in[n]$ is activated by x if $\eta_i(x)=1$. Let

$$\xi(x):=\begin{cases}1 & \sum_{i\in[n]}\eta_i(x)\leqslant M\\ 0 & \text{otherwise}\end{cases}$$

be the indicator of the event that at most M variables are activated by x. By Equation (15.2), we have

$$\begin{aligned}\frac{3\varepsilon}{4} \leqslant \mathbb{E}\sum_{\substack{S\neq\emptyset\\|S|\leqslant k}} F_S^2 \leqslant \mathbb{E}\left[\sum_{i=1}^n (1-\eta_i)\sum_{\substack{S:i\in S\\|S|\leqslant k}} F_S^2\right] & \qquad (S \text{ contains an inactive coordinate})\\ + \mathbb{E}\left[(1-\xi)\sum_{|S|\leqslant k} F_S^2\right] & \qquad (\text{more than } M \text{ variables are activated})\\ + \mathbb{E}\sum_{\substack{S\neq\emptyset\\|S|\leqslant k}} |F_S|^2\left(\prod_{i\in S}\eta_i\right)\xi & \qquad (\text{all variables in } S \text{ are activated, and } \textstyle\sum \eta_i \leqslant M) \qquad (15.3)\end{aligned}$$

We will show that the first two expectations on the right-hand side are small, and the main contribution comes from the third expectation.

When S contains an inactive coordinate: For every $i \in n$, by the definition of η_i, we have

$$\mathbb{E}(1-\eta_i)\sum_{\substack{S:i\in S\\|S|\leqslant k}} F_S^2 \leqslant \delta^{1/3}\mathbb{E}\left(\sum_{\substack{S:i\in S\\|S|\leqslant k}} F_S^2\right)^{2/3} = \delta^{1/3}\left\|\left(\sum_{\substack{i\in S\\|S|\leqslant k}} F_S^2\right)^{1/2}\right\|_{4/3}^{4/3}.$$

On the other hand, since

$$\partial_i f = \sum_{S:i\in S} F_S,$$

by applying Bourgain's inequality on low degree Fourier-Wash expansions (Theorem 14.1 with $p = 4/3$ and $q = 4$), and then using Proposition 12.2, we have

$$\left\|\left(\sum_{\substack{i\in S\\|S|\leqslant k}} F_S^2\right)^{1/2}\right\|_{4/3}^{4/3} \leqslant 2^{\frac{4}{3}\times 20k}\|\partial_i f\|_{4/3}^{4/3} \leqslant 2^{\frac{80k}{3}} 2I_i(f) \leqslant 2^{30k} I_i(f).$$

Therefore, by taking $\delta = 2^{-100\frac{I_f}{\varepsilon}}$,

$$\mathbb{E}\left[\sum_{i=1}^n (1-\eta_i)\sum_{\substack{i\in S\\|S|\leqslant k}} F_S^2\right] \leqslant \sum_{i=1}^n \delta^{1/3} 2^{30k} I_i(f) \leqslant \delta^{1/3} 2^{30k} I_f \leqslant \frac{\varepsilon}{4}. \qquad (15.4)$$

More than M variables are activated: By the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[(1-\xi)\sum_{|S|\leqslant k} F_S^2\right] \leqslant \left\|\sum_{|S|\leqslant k} F_S^2\right\|_2 \|1-\xi\|_2 = \left\|\left(\sum_{|S|\leqslant k} F_S^2\right)^{\frac{1}{2}}\right\|_4^2 \|1-\xi\|_2.$$

Note

$$\mathbb{E}(1-\xi) \leqslant \frac{1}{M}\mathbb{E}\sum_{i=1}^n \eta_i \leqslant \frac{1}{M}\mathbb{E}\sum_{i=1}^n \delta^{-1}\sum_{\substack{S:i\in S\\|S|\leqslant k}} F_S^2 \leqslant \frac{\delta^{-1}k\varepsilon}{M},$$

which since $1-\xi$ is $\{0,1\}$-valued, shows

$$\|1-\xi\|_2 \leqslant \sqrt{\frac{\delta^{-1}k\varepsilon}{M}}.$$

Again, by applying Bourgain's inequality low degree Fourier-Wash expansions (Theorem 14.1 with $q=4$), we have

$$\left\| \left(\sum_{|S|\leqslant k} F_S^2 \right)^{\frac{1}{2}} \right\|_4 \leqslant 2^{20k} \|f\|_4 \leqslant 2^{20k}.$$

Therefore, by taking $M = 2^{10^3 \frac{I_f}{\varepsilon}}$, we have

$$\mathbb{E}\left[(1-\xi) \sum_{|S|\leqslant k} F_S^2 \right] \leqslant \frac{\delta^{-1}k\varepsilon}{M} 2^{40k} \leqslant \frac{\varepsilon}{4}. \tag{15.5}$$

Concluding the desired inequality: By plugging Equation (15.4) and Equation (15.5) in Equation (15.3), we obtain

$$\frac{\varepsilon}{2} - \frac{\varepsilon}{4} - \frac{\varepsilon}{4} \leqslant \mathbb{E} \sum_{\substack{S\neq\emptyset \\ |S|\leqslant k}} |F_S|^2 \left(\prod_{i\in S} \eta_i \right) \xi \leqslant \binom{M}{\leqslant k} \mathbb{E} \max_{\substack{S\neq\emptyset \\ |S|\leqslant k}} |F_S|^2 \leqslant 2^{\Omega(I_f^2/\varepsilon^2)} \mathbb{E} \max_{\substack{S\neq\emptyset \\ |S|\leqslant k}} |F_S|^2.$$

□

We obtain the following corollary for monotone functions over the p-biased distribution.

Corollary 15.6 ([Bou99b, Proposition 1])**.** *The following holds for every $\varepsilon > 0$ and $I > 0$. Given any sequence of monotone functions $f : (\{0,1\}^n, \mu_p^n) \to \{0,1\}$ with $\mathrm{Var}(f) \geqslant \varepsilon$, the total influence $I_f \leqslant I$, and $p = o(1)$, there exists $S \subseteq [n]$ with $|S| \leqslant 4I/\varepsilon$ such that*

$$\mathbb{E}[f(\mathbf{x})|\mathbf{x}_S = (1,\dots,1)] \geqslant \mathbb{E}[f] + 2^{\Omega(I^2/\varepsilon)}.$$

Proof. By Theorem 15.5, there exists $|S| \leqslant 4I_f/\varepsilon$ and $y \in \{0,1\}^n$ such that

$$|\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})|\mathbf{x}_S = y_S] - \mathbb{E}[f]| = |\mathbb{E}_{\overline{S}}[f(y)] - \mathbb{E}[f]| \geqslant 2^{-O(I_f^2/\varepsilon^2)}.$$

Since f is monotone, it follows that either

$$\mathbb{E}[f(\mathbf{x})|\mathbf{x}_S = (0,\dots,0)] \leqslant \mathbb{E}[f] - 2^{\Omega(I_f^2/\varepsilon)},$$

or

$$\mathbb{E}[f(\mathbf{x})|\mathbf{x}_S = (1,\dots,1)] \geqslant \mathbb{E}[f] + 2^{\Omega(I_f^2/\varepsilon)}.$$

However,

$$\mathbb{E}[f] \leqslant \mathbb{E}[f(\mathbf{x})|\mathbf{x}_S = (0,\dots,0)] + (1-(1-p)^{|S|}) = \mathbb{E}[f(\mathbf{x})|\mathbf{x}_S = (0,\dots,0)] + p|S|,$$

and since $p|S| = O(pI_f/\varepsilon) = o(1)$, the first case cannot hold. □

15.2 Sharp threshold for graph properties

For a given graph H on t vertices and a graph G on n vertices (think of t as fixed and n large), let $G \cup H^*$ denote the union of G with a copy of H planted on t randomly chosen vertices of G.

In the case of monotone graph properties, Corollary 15.6 implies the existence of a graph H with at most $4I/\varepsilon$ edges such that

$$\mathbf{Pr}[\boldsymbol{G}(n,p) \cup H^* \in \mathcal{P}] \geqslant \mathbf{Pr}[\boldsymbol{G}(n,p) \in \mathcal{P}] + 2^{\Omega(I^2/\varepsilon)}.$$

In other words, planting H in $\boldsymbol{G}(n,p)$ significantly increases the probability that it satisfies $\mathcal{P}$.

We will show that one can choose H with the additional property that

$$\mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)] \geqslant 2^{-O(I_f^2/\varepsilon^2)}.$$

Corollary 15.7 ([Fri99, Bou99b]). *The following holds for every $\varepsilon > 0$ and $I > 0$. If a sequence of monotone graph properties of n-vertex graphs $\mathcal{P}$ satisfies $\mathrm{Var}_{\mu_p}(\mathcal{P}) \geqslant \varepsilon$ and $\frac{d\mu_p(\mathcal{P})}{dp} \leqslant Ip$ for some $p = o(1)$, there exists a graph H such that the following holds*

(i) H has at most $8I/\varepsilon$ edges.

(ii) $\mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)] \geqslant 2^{-O(I^2/\varepsilon^2)}$.

(iii)

$$\mathbf{Pr}[\boldsymbol{G}(n,p) \cup H \in \mathcal{P}] \geqslant \mathbf{Pr}[\boldsymbol{G}(n,p) \in \mathcal{P}] + 2^{\Omega(I^2/\varepsilon)}.$$

Proof. Let f be the corresponding Boolean function. By the Margulis-Russo formula, we have $I_f = \frac{1}{p(1-p)}\frac{d\mu_p(\mathcal{P})}{dp} \leqslant 2I$. Let $\mathcal{H}_k$ be the set of all graphs with at most $k := 4I_f/\varepsilon$ edges. Since every such graph has at most $2k$ vertices, $|\mathcal{H}_k| \leqslant \binom{2k}{2}^k \leqslant k^{3k} \leqslant 2^{3k^2}$. For every $H \in \mathcal{H}_k$, let

$$\phi(H) := \mathbf{Pr}[\boldsymbol{G}(n,p) \cup H \in \mathcal{P}] - \mathbf{Pr}[\boldsymbol{G}(n,p) \in \mathcal{P}] \geqslant 0.$$

While Theorem 15.5 shows

$$\mathbb{E}_{\mathbf{x}} \max_{S:|S|\leqslant k} |\mathbb{E}_{\overline{S}}[f(\mathbf{x})] - \mathbb{E}[f]| \geqslant 2^{-O(I^2/\varepsilon^2)}, \tag{15.6}$$

the proof of Corollary 15.6 only uses the fact that

$$\exists x \max_{S:|S|\leqslant k} |\mathbb{E}_{\overline{S}}[f(x)] - \mathbb{E}[f]| \geqslant 2^{-O(I^2/\varepsilon^2)}.$$

By using Equation (15.6) instead, we obtain

$$\mathbb{E} \max_{\substack{H\in\mathcal{H}_k \\ H\subseteq\boldsymbol{G}(n,p)}} \phi(H) \geqslant 2^{-O(I^2/\varepsilon^2)}.$$

Since

$$\mathbb{E} \max_{\substack{H\in\mathcal{H}_k \\ H\subseteq\boldsymbol{G}(n,p)}} \phi(H) \leqslant \sum_{H\in\mathcal{H}} \mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)]\phi(H) \leqslant |\mathcal{H}_k| \max_H \mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)]\phi(H),$$

there must exist $H \in \mathcal{H}_k$ that satisfies (ii) and (iii). □

Remark 15.8. To illustrate the importance of (ii), consider the monotone graph property of being non-3-colourable, which is known [AF99] to have a sharp threshold at the critical probability $p_c = \Theta(1/n)$. Suppose we aim to apply Equation (15.6) to prove that non-3-colourability has a sharp threshold by showing that no H can satisfy (i)-(iii).

Note that if we take $H = K_4$ to be the complete graph on 4 vertices, then $\boldsymbol{G}(n,p) \cup H$ is non-3-colourable with probability 1, and therefore, both (i) and (iii) are satisfied. However, K_4 is too dense and $\mathbf{Pr}[K_4 \subseteq \boldsymbol{G}(n,p)] = o(1)$ at $p = \Theta(1/n)$, and (ii) is not satisfied. Therefore, (ii) is essential in such applications.

Combined with the Margulis-Russo lemma, we obtain the following theorem.

Theorem 15.9 ([Fri05]). *There exist functions $k(\varepsilon,\alpha)$ and $\tau(\varepsilon,\alpha)$ such that the following holds. Let $\mathcal{P}$ be a graph property with a coarse threshold. Specifically let $\varepsilon > 0$, $\alpha > 0$, and $p = p(n)$ be such that*

$$\alpha < \mathbf{Pr}[\boldsymbol{G}(n,(1-\varepsilon)p) \in \mathcal{P}] < \mathbf{Pr}[\boldsymbol{G}(n,p) \in \mathcal{P}] < 1 - 2\alpha.$$

Then there exists a graph H with no more than $k(\varepsilon,\alpha)$ vertices such that

$$\mathbf{Pr}[\boldsymbol{G}(n,p) \cup H^* \in \mathcal{P}] > \mathbf{Pr}[\boldsymbol{G}(n,p) \in \mathcal{P}] + \tau(\varepsilon,\alpha).$$

Furthermore, H is a "reasonable" graph:

$$\mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)] > \tau(\varepsilon,\alpha).$$

Proof. Exercise. □

Remark 15.10. In fact, Friedgut's proof from [Fri99] shows that in Theorem 15.9, one can have a stronger guarantee that

$$\mathbf{Pr}[\boldsymbol{G}(n,p) \cup H^* \in \mathcal{P}] > 1-\alpha.$$

Finally, we mention another corollary that characterizes the critical probability values for graph properties with a coarse threshold.

Corollary 15.11 (Friedgut [Fri99])**.** *If a sequence of monotone graph properties of n-vertex graphs $\mathcal{P}$ satisfies $\mathrm{Var}_{\mu_p}(\mathcal{P}) \geqslant \varepsilon$ and $\frac{d\mu_p(\mathcal{P})}{dp} \leqslant Ip$, then $p = \Theta(n^{-a/b})$ for positive integers a and b that are bounded from above by some function of I and ε.*

Proof. Let H be the graph provided by Corollary 15.7.

Note if $\mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)] = 1 - o(1)$, then Corollary 15.7 (iii) cannot hold. Moreover, by Corollary 15.7 (ii), we have $\mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)] \geqslant 2^{-O(I_f^2/\varepsilon^2)} = \Omega(1)$. Therefore, there exists $\delta = \delta(I,\varepsilon) > 0$ such that

$$\delta < \mathbf{Pr}[H \subseteq \boldsymbol{G}(n,p)] < 1-\delta.$$

Since H has at most $O(I/\varepsilon)$ edges, we have $p = \Theta(n^{-a/b})$ for a and b are positive integers depending on H. □

By Corollary 15.11, coarse thresholds only happen near rational powers of n. Corollary 15.11 immediately implies, for example, the well-known fact that connectivity has a sharp threshold as the critical probability for connectivity is $\Theta\left(\frac{\ln(n)}{n}\right)$.

Chapter 16

Expansion of small sets in the noisy cube

Recall that given a correlation parameter $\rho \in [0,1]$, and $x \in \{0,1\}^n$, the ρ-equal copy of x is the random variable $\mathbf{y}$ that is sampled from $\{0,1\}^n$ through the following process: for each $i \in [n]$, with probability ρ, set $\mathbf{y}_i = x_i$ and with probability $1-\rho$, sample $\mathbf{y}_i$ uniformly at random from $\{0,1\}$.

Definition 16.1 (noisy hypercube)**.** The *ρ-noisy hypercube graph* is the undirected weighted complete graph with the vertex set $\{0,1\}^n$, where the weight of the edge (x,y) is $\mathbf{Pr}[\mathbf{y}' = y]$ where $\mathbf{y}'$ is the ρ-equal copy of x.

Note that if $\mathbf{y}$ is a ρ-equal copy of x, then $\mathbf{Pr}[\mathbf{y}_i = x_i] = \frac{1+\rho}{2}$ and $\mathbf{Pr}[\mathbf{y}_i \neq x_i] = \frac{1-\rho}{2}$ for every $i \in [n]$. Therefore, the weight of an edge (x,y) in the ρ-noisy hypercube graph is given by the formula

$$w(x,y) = w(y,x) = \left(\frac{1+\rho}{2}\right)^{n-d_H(x,y)} \left(\frac{1-\rho}{2}\right)^{d_H(x,y)},$$

where $d_H(x,y)$ denotes the hamming distance between x and y.

In this section, we are interested in the expansion properties of small subsets of the noisy cube. Let $\alpha, \beta \in (0,1)$ be small constants. Given sets $A, B \subseteq \{0,1\}^n$ with relative densities α, β, we wish to analyze

$$\mathbf{Pr}[\mathbf{y} \in B | \mathbf{x} \in A], \tag{16.1}$$

when $\mathbf{x}$ is uniformly selected from $\{0,1\}^n$ and $\mathbf{y}$ is the ρ-equal copy of $\mathbf{x}$. The two questions of *minimizing* and *maximizing* this quantity in terms of α and β are important in many applications.

16.1 Small-set expansion in noisy cube

In this section, we will study the small-set expansion properties of the noisy hypercube, which is related to maximizing Equation (16.1) for $A = B$.

Let $\mathbf{x}$ be uniform sampled from $\{0,1\}^n$ and $\mathbf{y}$ be the ρ-equal copy of $\mathbf{x}$. Define the *ρ-noise stability* of a function $f : \{0,1\}^n \to \{0,1\}$ as

$$\mathrm{Stab}_\rho(f) \coloneqq \mathbb{E}\left[f(\mathbf{x})f(\mathbf{y})\right] = \langle T_\rho f, f\rangle = \sum_{S \subseteq [n]} \rho^{|S|} \widehat{f}(S)^2. \tag{16.2}$$

If $A \subseteq \{0,1\}^n$ has density α, then

$$\mathrm{Stab}_\rho(A) = \mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in A] = \alpha\, \mathbf{Pr}[\mathbf{y} \in A | \mathbf{x} \in A].$$

Dictators and, more generally, k-juntas for small k have large stability. The small-set expansion in the noisy cube states when α is small, the set cannot be stable, and the probability that a random ρ-noisy neighbour $\mathbf{y}$ of a random vertex $\mathbf{x} \in A$ belongs to A is small, i.e., the noisy hypercube graph has large expansion for small sets A.

Proposition 16.2 (Small-set expansion in noisy cube)**.** *Let $A \subseteq \{0,1\}^n$ have density $\alpha > 0$. We have*

$$\mathrm{Stab}_\rho(A) \leqslant \alpha^{\frac{2}{1+\rho}} \text{ and } \mathbf{Pr}[\mathbf{y} \in A | \mathbf{x} \in A] \leqslant \alpha^{\frac{1-\rho}{1+\rho}},$$

where $\mathbf{x}$ *is uniform and* $\mathbf{y}$ *is a* ρ*-equal copy of* $\mathbf{x}$*.*

Proof. By hypercontractivity, for $p \coloneqq \rho + 1$, we have

$$\mathrm{Stab}_\rho(A) = \sum_{S \subseteq [n]} \rho^{|S|} \widehat{A}(S)^2 = \|T_{\sqrt{\rho}} A\|_2^2 \leqslant \|A\|_p^2 = \alpha^{\frac{2}{1+\rho}},$$

and $\mathbf{Pr}[\mathbf{y} \in A | \mathbf{x} \in A] = \frac{\mathrm{Stab}_\rho(A)}{\alpha} = \alpha^{\frac{1-\rho}{1+\rho}}$. □

Using a similar proof, we can also obtain a two-set version of Proposition 16.2

Proposition 16.3. *Let* $A, B \subseteq \{0,1\}^n$ *have respective densities* $\alpha, \beta > 0$*. We have*

$$\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] \leqslant \alpha^{\frac{1}{1+\rho^2}} \beta^{1/2}.$$

where $\mathbf{x}$ *is uniform and* $\mathbf{y}$ *is a* ρ*-equal copy of* $\mathbf{x}$*.*

Proof. By applying the Cauchy-Schwarz inequality and then hypercontractivity with $p = 1 + \rho^2$, we have

$$\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] = \langle T_\rho A, B \rangle \leqslant \|T_\rho A\|_2 \|B\|_2 \leqslant \|A\|_p \|B\|_2 = \alpha^{1/p} \beta^{1/2} = \alpha^{\frac{1}{1+\rho^2}} \beta^{1/2}.$$

□

16.2 Reverse hypercontractivity and sparse pairs

Let us turn to the problem of minimizing $\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B]$. Since $\mathbf{x}$ and $\mathbf{y}$ are correlated, to minimize $\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B]$, picking two antipodal hamming balls seems a natural candidate. In this case, we can upper-bound $\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B]$ using the following lemma whose proof we omit.

Lemma 16.4 ([MOR+06])**.** *Fix* $a, b > 0$ *and let* $A, B \subseteq \{0,1\}^n$ *be defined as*

$$A \coloneqq \left\{ x \ : \ \sum x_i \leqslant \frac{n}{2} - a\sqrt{n} \right\},$$
$$B \coloneqq \left\{ x \ : \ \sum x_i \geqslant \frac{n}{2} + b\sqrt{n} \right\}.$$

We have

$$\lim_{n\to\infty} \frac{|A|}{2^n} = \frac{1}{\sqrt{2\pi a}} e^{-a^2/2},$$
$$\lim_{n\to\infty} \frac{|B|}{2^n} = \frac{1}{\sqrt{2\pi b}} e^{-b^2/2}.$$

and for a uniform $\mathbf{x}$ *and its* ρ*-equal copy* $\mathbf{y}$*, we have*

$$\lim_{n\to\infty} \mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] \leqslant \frac{\sqrt{1-\rho^2}}{2\pi a(\rho a + b)} \exp\left(-\frac{a^2 + b^2 + 2\rho ab}{2(1-\rho^2)} \right).$$

The main term in the above upper bound is the exponential term. We will establish a lower-bound in Theorem 16.9 that almost matches the upper-bound of Lemma 16.4.

First, we show that the naive spectral gap method provides a weak lower bound for $\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B]$.

$$\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] = \langle T_\rho A, B \rangle = \rho^{|S|} \sum \widehat{A}(S) \widehat{B}(S) \geqslant \alpha\beta - \rho \sum_{S \neq \emptyset} \left| \widehat{A}(S) \widehat{B}(S) \right|$$
$$\geqslant \alpha\beta - \rho \left(\sum_{S \neq \emptyset} \left| \widehat{A}(S) \right|^2 \right)^{1/2} \left(\sum_{S \neq \emptyset} \left| \widehat{B}(S) \right|^2 \right)^{1/2} = \alpha\beta - \rho\sqrt{\alpha - \alpha^2}\sqrt{\beta - \beta^2}.$$

When α and β are small, the second term on the right-hand side is larger than the first term, and the bound is negative (and trivial) unless ρ is very small. Therefore, we need a deeper approach to prove a more meaningful lower bound for $\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B]$.

The key tool to obtaining an effective lower bound is an extension of the hypercontractivity to $\|\cdot\|_p$ for $p < 1$, called reverse hypercontractivity. Unlike the original hypercontractivity, the reverse hypercontractivity only applies to non-negative functions. The next four theorems and lemmas all require the functions to be non-negative.

Theorem 16.5 (Reverse Hölder Inequality)**.** *If $f, g \geqslant 0$ are functions on a measure space, then*

$$\langle f, g\rangle \geqslant \|f\|_p \|g\| q,$$

where $p, q \in (-\infty, 1)$ and $\frac{1}{p} + \frac{1}{q} = 1$.

Remark 16.6. When $p < 1$, the function

$$\|f\| \coloneqq (\mathbb{E}|f|^p)^{1/p}$$

is not a norm. In fact, for $-\infty < p < 1$ and $f, g \geqslant 0$, we have the reversed triangle inequality:

$$\|f + g\| p \geqslant \|f\|_p + \|g\|_p.$$

To prove this, note that by the reversed Hölder Inequality

$$\begin{aligned} \|f+g\|_p^p &= \mathbb{E}(f+g)^p = \mathbb{E}(f+g)^{p-1} f + \mathbb{E}(f+g)^{p-1} g \\ &\geqslant (\mathbb{E}(f+g)^p)^{\frac{p-1}{p}} \|f\|_p + (\mathbb{E}(f+g)^p)^{\frac{p-1}{p}} \|g\|_p \\ &= \|f+g\|_p^{p-1} (\|f\|_p + \|g\|_p). \end{aligned}$$

which yields the reversed triangle inequality.

Theorem 16.7 (Reverse Hypercontractivity inequality [MOR+06])**.** *Let $f : \{0,1\}^n \to [0, \infty)$, then*

$$\|T_\rho f\|_q \geqslant \|f\|_p,$$

for $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$ and $-\infty < q \leqslant p < 1$.

The proof is similar to the classical hypercontractivity. First, one proves it for the 1-dimensional case, and then an induction establishes the general case. We have the following corollary.

Corollary 16.8. *Let $f, g : \{0,1\}^n \to [0, \infty)$ and consider a uniform $\mathbf{x} \in \{0,1\}^n$ and a ρ-equal $\mathbf{y}$ copy of $\mathbf{x}$. For every $0 < \rho \leqslant \sqrt{(1-p)(1-q)} \leqslant 1$ and $p, q < 1$, we have*

$$\mathbb{E} f(\mathbf{x}) g(\mathbf{y}) \geqslant \|f\|_p \|g\|_q.$$

Proof. Let $p' = \frac{p}{p-1}$, so that p and p' are conjugate exponents. We use the reverse Hölder's inequality and then apply the inverse hypercontractivity inequality,

$$\mathbb{E} f(\mathbf{x}) g(\mathbf{y}) = \mathbb{E} f(\mathbf{x}) T_\rho g(\mathbf{x}) \geqslant \|f\|_p \|T_\rho g\|_{p'} \geqslant \|f\|_p \|g\|_q,$$

where the last inequality requires $0 < \rho \leqslant \sqrt{\frac{1-q}{1-p'}} = \sqrt{(1-p)(1-q)}$. □

We are ready to prove the strong lower bound on $\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B]$.

Theorem 16.9 ([MOR+06])**.** *Suppose $A, B \subseteq \{0,1\}^n$ have relative densities*

$$\frac{|A|}{2^n} = e^{-a^2/2} \quad \text{and} \quad \frac{|B|}{2^n} = e^{-b^2/2},$$

and let $\mathbf{x} \in \{0,1\}^n$ *be uniform and* $\mathbf{y}$ *be a* ρ*-equal copy of* $\mathbf{x}$*. Then*

$$\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] \geqslant \exp\left(-\frac{a^2 + b^2 + 2\rho ab}{2(1-\rho^2)}\right).$$

Proof. Let $p, q < 1$ be such that $\rho^2 = (1-p)(1-q)$. By corollary 16.8, we have that

$$\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] = \mathbb{E}A(\mathbf{x})B(\mathbf{y}) \geqslant \|A\|_p \|B\|_q = e^{-\frac{a^2}{2p} - \frac{b^2}{2q}}.$$

Our task is to optimize p to maximize the right-hand side, which is equivalent to minimizing $\frac{a^2}{2p} + \frac{b^2}{2q}$. To simplify the calculations, write $p = 1 - \rho r$ and $q = 1 - \frac{\rho}{r}$ with

$$r = \frac{1-p}{\rho} = \frac{\rho}{1-q} > 0.$$

Then

$$\frac{a^2}{2p} + \frac{b^2}{2q} = \frac{a^2}{2(1-\rho r)} + \frac{b^2}{2(1-\frac{\rho}{r})},$$

is minimized when

$$r = \frac{\frac{b}{a} + \rho}{1 + \rho\frac{b}{a}}.$$

Using this value of r gives

$$\frac{a^2}{2p} + \frac{b^2}{2q} = \frac{a^2 + b^2 + 2\rho ab}{2(1-\rho^2)}.$$

□

We obtain the following corollary from Theorem 16.9 by parametrizing the densities differently.

Corollary 16.10. *Let* $A, B \subseteq \{0,1\}^n$ *with relative densities* $\alpha > 0$ *and* $\alpha^\sigma > 0$ *respectively, where* $\sigma > 0$*. Let* $\mathbf{x} \in \{0,1\}^n$ *be uniform and* $\mathbf{y}$ *be a* ρ*-equal copy of* $\mathbf{x}$*. Then*

$$\mathbf{Pr}[\mathbf{x} \in A, \mathbf{y} \in B] \geqslant \alpha\alpha^{\frac{(\sqrt{\sigma}+\rho)^2}{1-\rho^2}}.$$

In particular, if $|A| = |B|$*, this probability is at least* $\alpha^{\frac{1+\rho}{1-\rho}}$*.*

Another interesting corollary of reverse hypercontractivity is that it allows us to quantify how T_ρ "smooths" the "peaks" of the function f. In other words, it provides an upper bound on $\mathbf{Pr}[T_\rho f(\mathbf{x}) > 1 - \delta]$.

Theorem 16.11 ([MOO10, Theorem 4.5])**.** *Consider* $f : \{0,1\}^n \to [0,1]$ *with* $\mathbb{E}f = \alpha$*. For every* $0 < \rho < 1$ *and* $0 \leqslant \varepsilon \leqslant 1 - \alpha$ *we have*

$$\mathbf{Pr}[T_\rho f(\mathbf{x}) > 1 - \delta] < \varepsilon$$

provided that $0 \leqslant \delta < \varepsilon^{\rho^2/(1-\rho^2)+O(\kappa)}$*, where* $\kappa = \sqrt{\frac{\alpha \ln(e/(1-\alpha))}{1-\rho}}$*.*

Proof. Define indicator functions

$$g : x \to \begin{cases} 1 & \text{if } T_\rho f(x) > 1 - \delta \\ 0 & \text{otherwise} \end{cases}$$

$$h : x \to \begin{cases} 1 & \text{if } f(x) > b \\ 0 & \text{otherwise} \end{cases},$$

where $b = \frac{1}{2}(1+\alpha)$. We need to show that $\varepsilon' := \mathbb{E}g \leqslant \varepsilon$. By the first moment method,

$$\alpha = \mathbb{E}f \geqslant (1 - \mathbb{E}h)b,$$

which shows

$$\mathbb{E}h > 1 - \frac{\alpha}{b} = \frac{1-\alpha}{1+\alpha},$$

and therefore, the support of h is not very small.

By the definition of h, we have $(1-b)h(x) \leqslant 1 - f(x)$, and therefore, when $g(x) = 1$, we have

$$T_\rho[(1-b)h(x)] \leqslant T_\rho(1 - f(x)) \leqslant \delta.$$

Hence, $g(x) = 1$ implies

$$T_\rho[h(x)] \leqslant \frac{\delta}{1-b},$$

and we have

$$\mathbb{E}[gT_\rho h] \leqslant \frac{\delta\varepsilon'}{1-b} = \frac{2\delta\varepsilon'}{1-\alpha}. \tag{16.3}$$

Meanwhile, by Corollary 16.10,

$$\mathbb{E}[gT_\rho h] \geqslant \varepsilon' \cdot \varepsilon'^{\frac{(\sqrt{\beta}+\rho)^2}{1-\rho^2}} \tag{16.4}$$

where $\beta = \frac{\log \mathbb{E}h}{\log \varepsilon'}$. Finally, Equation (16.3) and Equation (16.4) together with our assumption on δ implies the desired bound $\varepsilon' \leqslant \varepsilon$. □

Chapter 17

Gaussian Spaces

Many of the concepts discussed in this course—such as noise stability, isoperimetric inequalities, and hypercontractivity—were originally developed within the geometric framework of Gaussian probability spaces. The continuous and symmetric nature of Gaussian space often leads to more elegant results and proofs that avoid some of the technical challenges of the discrete setting of the hypercube. For this reason, Gaussian space can serve as an elegant and intuitive framework for studying the properties of functions on the discrete cube. Moreover, tools like the invariance principle of Mossel, O'Donnell, and Oleszkiewicz [MOO10] and the more recent global hypercontractivity theorem [**?**] enable us to systematically translate certain results from the Gaussian setting to the discrete hypercube.

17.1 Gaussian probability space

In this section, we will define Gaussian random variables in $\mathbb{R}^n$ and outline some of their basic properties. We then introduce the Ornstein-Uhlenbeck Gaussian noise operator and discuss its hypercontractivity.

Definition 17.1 (One-dimensional standard Gaussian)**.** The standard normal distribution on $\mathbb{R}$ is the probability distribution γ_1 on $\mathbb{R}$ with the density function

$$\phi(x) \coloneqq \frac{e^{-x^2/2}}{\sqrt{2\pi}},$$

which means that for any interval $[a, b]$,

$$\gamma_1([a, b]) \coloneqq \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

Figure 17.1: Standard normal distribution

A random variable $\boldsymbol{g}$ with distribution γ_1 is called a *standard Gaussian*. It satisfies

$$\mathbb{E}\boldsymbol{g} = 0 \text{ and } \mathbb{E}\boldsymbol{g}^2 = 1.$$

Definition 17.2. For $n \in \mathbb{N}$, the multivariate Gaussian distribution γ_n is the product probability distribution defined by γ_1 on $\mathbb{R}^n$. Specifically,

$$\gamma_n(\{x \in \mathbb{R}^n | a_i \leqslant x_i \leqslant b_i\}) = \prod_{i=1}^n \gamma_1([a_i, b_i]). \tag{17.1}$$

The corresponding density function is:

$$\phi_n(x) \coloneqq \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{\frac{-\|x\|^2}{2}},$$

where $\|x\|$ denotes the length of x. For any measurable set $A \in \mathbb{R}^n$, we have $\gamma_n(A) \coloneqq \int_A \phi_n(x)dx$. A random variable $\boldsymbol{g}$ with distribution γ_n is called a *standard Gaussian vector* in $\mathbb{R}^n$.

The definition of the Gaussian measure on $\mathbb{R}^n$ as the product space in (17.1) might mistakenly suggest that the Gaussian probability distribution depends on the specific coordinate system. However, this impression is incorrect. The density function $\phi_n(x)$ depends only on the length of x and, therefore, is independent of the particular choice of the coordinate system. In particular, if $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n \in \mathbb{R}$ are i.i.d. standard Gaussians and $a \in \mathbb{R}^n$ has length $c = \|a\|$, then $\sum_{i=1}^n a_i \boldsymbol{g}_i$ has the same distribution as $c\boldsymbol{g}$, where $\boldsymbol{g}$ is a standard Gaussian.

17.2 Hermite polynomials

The Hermite polynomials are a classical orthogonal polynomial sequence for the space $L_2(\mathbb{R}, \gamma_1)$. Given $k \in \mathbb{Z}_{\geqslant 0}$, the Hermite polynomial of degree k is given by

$$\mathrm{He}_k(x) = (-1)^k e^{\frac{x^2}{2}} \frac{d^k}{dx^k} e^{-\frac{x^2}{2}}.$$

They satisfy the relations

$$\mathrm{He}_{k+1}(x) \coloneqq x\mathrm{He}_k(x) - \mathrm{He}_k'(x) = x\mathrm{He}_k(x) - k\mathrm{He}_{k-1}(x), \tag{17.2}$$

with the base case $\mathrm{He}_0(x) = 1$. The Hermite polynomials form a complete orthogonal basis for $L_2(\mathbb{R}, \gamma_1)$. It will be useful to normalize them to obtain an orthonormal basis. For $k \in \mathbb{Z}_{\geqslant 0}$, define the corresponding *normalized Hermite* polynomial as

$$h_k(x) \coloneqq \frac{\mathrm{He}_k}{\|\mathrm{He}_k\|} = \frac{\mathrm{He}_k}{\sqrt{k!}}.$$

Alternatively, we can construct the polynomials $h_k(x)$ by applying the Gram–Schmidt process to the monomials x^k. We have

$$h_0(x) \coloneqq 1, \qquad h_1(x) \coloneqq x, \qquad h_2(x) \coloneqq \frac{z^2 - 1}{\sqrt{2}}, \qquad h_3(x) \coloneqq \frac{x^3 - 3x}{\sqrt{6}}, \qquad \ldots,$$

where each polynomial $h_k(x)$ is obtained by taking the orthogonal projection of x^k to $\operatorname{span}\{h_0, \ldots, h_{k-1}\}^\perp$ and then normalizing it to have norm 1:

$$h_k(x) \coloneqq \frac{x^k - \sum_{i=0}^{k-1} \langle x^k, h_i \rangle h_i}{\|x^k - \sum_{i=0}^{k-1} \langle x^k, h_i \rangle h_i\|}$$

Given an $f \in L_2(\mathbb{R}, \gamma_1)$, we can write the Hermite expansion of f as the infinite sum

$$f = \sum_{k=0}^{\infty} \langle f, h_k \rangle h_k,$$

which converges in the L_2 norm.

More generally, we can use the Hermine polynomials to construct an orthonormal base for $L_2(\mathbb{R}^n, \gamma_n)$. For a multi-index $\alpha \in \mathbb{Z}_{\geqslant 0}^n$, define $h_\alpha : \mathbb{R}^n \to \mathbb{R}$ as

$$h_\alpha(x_1, \ldots, x_n) \coloneqq h_{\alpha_1}(x_1) \ldots h_{\alpha_n}(x_n).$$

These polynomials form a complete orthonormal basis for $L_2(\mathbb{R}^n, \gamma_n)$, and therefore, every $f \in L_2(\mathbb{R}^n, \gamma_n)$ has a unique *Hermite expansion*

$$f = \sum_{\alpha \in \mathbb{Z}_{\geqslant 0}^n} \widehat{f}(\alpha) h_\alpha,$$

where $\widehat{f}(\alpha) \coloneqq \langle f, h_\alpha \rangle$ are called the *Hermite coefficients* of f. Gain, the convergence is in the L_2, meaning that

$$\lim \left\| f - \sum_{\alpha : \sum \alpha_i \leqslant k} \widehat{k \to \infty} f(\alpha) h_\alpha \right\|_2 = 0.$$

Remark 17.3. Since $h_0(x) \equiv 1$ and $h_1(x) = x$, when $\alpha = \mathbf{1}_S$ for $S \subseteq [n]$, then $h_\alpha = \prod_{i \in S} x_i$. In other words, the set

of functions h_α includes all the multilinear monomials $\prod_{\in S} x_i$. Therefore, for *multilinear* polynomials

$$f(x_1, \ldots, x_n) = \sum_{S \subseteq [n]} \lambda_S \prod_{i \in S} x_i,$$

the Hermite expansion coincides with the polynomial representation.

17.3 Gaussian noise and hypercontractivity

We start by defining the Ornstein-Uhlenbeck noise operator.

Definition 17.4. Given $\rho \in [-1, 1]$, the corresponding Ornstein-Uhlenbeck operator U_ρ acting on $L_2(\mathbb{R}^n, \gamma_n)$ is defined as

$$U_\rho f(x) \coloneqq \mathbb{E} f(\rho x + \sqrt{1-\rho^2}\boldsymbol{g}),$$

where $\boldsymbol{g} \sim (\mathbb{R}^n, \gamma_n)$ is a standard Gaussian vector.

If $\mathbf{x}$ and $\mathbf{g}$ are independent standard Gaussian, then since $\sqrt{\rho^2 + (1-\rho^2)} = 1$, the random variable

$$\boldsymbol{y} \coloneqq \rho \boldsymbol{x} + \sqrt{1-\rho^2}\boldsymbol{g}$$

is also a standard Gaussian. In particular, $\mathbb{E}_{\gamma_n}[U_\rho f] = \mathbb{E}_{\gamma_n}[f]$.

Furthermore, the correlation of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\mathbb{E}\boldsymbol{x}\boldsymbol{y} = \mathbb{E}[\rho \boldsymbol{x}^2 + \sqrt{1-\rho^2}\boldsymbol{x}\boldsymbol{g}] = \rho.$$

We have the following theorem regarding the action of U_ρ on the set of Hermite polynomials.

Theorem 17.5. *For every $\alpha \in \mathbb{Z}^n_{\geqslant 0}$ and $\rho \in [-1, 1]$, denoting $|\alpha| = \sum \alpha_i$, we have*

$$U_\rho h_\alpha = \rho^{|\alpha|} h_\alpha.$$

Proof. □

The following theorem, due to Nelson [**?**], shows that the Ornstein-Uhlenbeck operator U_ρ is hypercontractive.

Theorem 17.6 (Hypercontractivity in Gaussian spaces)**.** *Let $1 \leqslant p \leqslant q \leqslant \infty$ and $f \in L_p(\mathbb{R}^n, \gamma_n)$. We have $\|U_\rho f\|_q \leqslant \|f\|_p$ for $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$.*

17.3.1 Comparison to the Fourier-Walsh expansion

The Gaussian space $(\mathbb{R}^n, \gamma_n)$ is a product space, and therefore, as we discussed in Section 9.4, every integrable $f : (\mathbb{R}^n, \gamma_n) \to \mathbb{R}$ has a Fourier-Walsh expansion $f = \sum_{S \subseteq [n]} F_S$. Let

$$f = \sum_{\alpha \in \mathbb{Z}^n_{\geqslant 0}} \widehat{f}(\alpha) h_\alpha$$

be the Hermite expansion of f. It is not difficult to verify that

$$F_S = \sum_{\alpha : \mathrm{supp}(\alpha) = S} \widehat{f}(\alpha) h_\alpha.$$

Recall that we defined the noise operator T_ρ as

$$T_\rho f(x) = \mathbb{E}_{\mathbf{y}} f(\mathbf{y}),$$

where with probably ρ, we set $\mathbf{y} = x$, and with probability $1 - \rho$, we sample $\mathbf{y}$ from the distribution γ_n. We showed $T_\rho f = \sum_{S \subseteq [n]} \rho^{|S|} F_S$.

Note that T_ρ and U_ρ are not the same operators, as for example, $T_\rho h_\alpha = \rho^{|\mathrm{supp}(\alpha)|}$ while $U_\rho h_\alpha = \rho^{|\alpha|}$ for every normalized Hermite polynomial.

Note also that since γ_n is a continuous probability space, unlike the U_ρ operator, T_ρ is *not* hypercontractive.

17.4 Noise stability in Gaussian Space

In Section 16.1, we defined the notion of noise stability for subsets of the discrete cube $\{0,1\}^n$. In this section, we study this notion for the Gaussian space.

Let $\boldsymbol{g}$ and $\boldsymbol{h}$ be a pair of ρ-correlated standard Gaussian vector in $\mathbb{R}^n$. Define the *ρ-noise stability* of a function $f : (\mathbb{R}^n, \gamma_n) \to \{0,1\}$ as

$$\mathrm{Stab}_\rho(f) \coloneqq \mathbb{E}\left[f(\boldsymbol{g})f(\boldsymbol{h})\right] = \langle U_\rho f, f\rangle. \tag{17.3}$$

We are interested in characterizing the most stable subsets of $\mathbb{R}^n$.

Theorem 17.7 (Noise Stability of Homogenous Halfspaces)**.** *If $H : (R^n, \gamma_n) \to \{0,1\}$ is the indicator function of a homogeneous half-space*[1]*, then we have*

$$\mathrm{Stab}_\rho(H) = \frac{1}{4} + \frac{\arcsin(\rho)}{2\pi}.$$

Proof. Since the Gaussian measure is invariant under rotations, we can assume without loss of generality that $H(x) \coloneqq \mathbf{1}_{[x_1 \geqslant 0]}$. In this case,

$$\mathrm{Stab}_\rho(H) = \mathbf{Pr}[\boldsymbol{g} \geqslant 0 \text{ and } \rho\boldsymbol{g} + \sqrt{1-\rho^2}\boldsymbol{h} \geqslant 0],$$

where $\boldsymbol{g}$ and $\boldsymbol{h}$ are independent standard Gaussians. Let

$$A = \left\{(x,y) \in \mathbb{R}^2 \ : \ x \geqslant 0 \text{ and } \rho x + \sqrt{1-\rho^2}y \geqslant 0\right\},$$

as illustrated in Figure 17.2.

Figure 17.2: Stability of a homogeneous half-space

We have

$$\mathrm{Stab}_\rho(H) = \mathbf{Pr}[(\boldsymbol{g},\boldsymbol{h}) \in A].$$

Since $(\boldsymbol{g},\boldsymbol{h})$ is a standard Gaussian vector and hence its distribution is invariant under rotations, we have

$$\mathrm{Stab}_\rho(H) = \frac{\arctan}{2\pi} = ?????.$$

□

The *Gaussian Rearrangement* A^* of a set $A \subset \mathbb{R}^n$ is the interval (t,∞) with $\gamma_1(t,\infty) = \gamma_n(A)$. Note that A^* corresponds to the halfspace with the same Gaussian measure as A. In particular, if $\gamma_n(A) = \frac{1}{2}$, then A^* corresponds to a homogenous half-space.

The following theorem due to Borell [**?**] from 1983 shows that half-spaces are the extremal sets for stability.

Theorem 17.8 (Borell [**?**])**.** *Let $A, B \subseteq \mathbb{R}^n$. Then for any $0 \leqslant \rho \leqslant 1$ and $q \geqslant 1$ we have:*

$$\mathbb{E}(U_\rho A)^q B \leqslant \mathbb{E}(U_\rho A^*)^q B^*$$

In particular,

$$\mathrm{Stab}_\rho(A) = \mathbb{E}AU_\rho A \leqslant \mathrm{Stab}_\rho(A^*).$$

[1] A half-space is *homogeneous* if the hyperplane that defines it contains the origin.

17.5 The Berry–Esseen Theorem

In this section, we explore a classical example of an invariance theorem, specifically the Berry–Esseen theorem. This theorem provides a quantitative version of the Central Limit Theorem for finite sums of independent random variables, giving a bound on how closely the distribution of a sum of random variables approximates a Gaussian distribution.

Let $\mathbf{x}$ be a random variable with mean zero and unit variance, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. copies of $\mathbf{x}$. The Berry-Esseen theorem states that if $\mathbb{E}[|\mathbf{x}|^3]$ is not too large, the cumulative distribution function of $S_n := \frac{\sum_{i=1}^n \mathbf{x}_i}{\sqrt{n}}$ is close to the cumulative distribution function of a standard Gaussian.

Theorem 17.9 (Berry-Esseen)**.** *Let $\mathbf{x}$ be a random variable with mean zero and unit variance, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. copies of $\mathbf{x}$. Let $S_n := \frac{\sum_{i=1}^n \mathbf{x}_i}{\sqrt{n}}$ and $\boldsymbol{g}$ be a standard Gaussian. We have*

$$|\operatorname{\mathbf{Pr}}[S_n \leqslant t] - \operatorname{\mathbf{Pr}}[\boldsymbol{g} \leqslant t]| \leqslant O\left(\frac{\mathbb{E}[|\mathbf{x}|^3]}{\sqrt{n}}\right) \quad \text{for every } t \in \mathbb{R}.$$

We will not give a proof of Theorem 17.9. Instead we prove a slightly weaker bound Theorem 17.11.

There are several established methods for proving the Berry–Esseen theorem. Common approaches include the Fourier method, the moment method, Stein's method, and the Lindeberg swapping trick. For a more comprehensive discussion of these proofs, we recommend Terry Tao's lecture note on the Central Limit Theorem.

In these notes, we will focus on a proof using the Lindeberg swapping trick, also known as the *replacement method* or the *hybrid method* in theoretical cryptography. In this method, we swap the variables with independent Gaussians *one by one* and carefully control the changes at each step.

The replacement method is a powerful and versatile technique. As we will see in **??** and **??**, this method plays a central role in proving two fundamental results in the analysis of Boolean functions: *global hypercontractivity* and the *invariance principle*.

We will start by proving a technical form of the Berry-Esseen theorem.

Theorem 17.10. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a three time differentiable function with $|\psi'''(x)| < B$ for all $x \in \mathbb{R}$. Let $\mathbf{x}$ be a random variable with mean zero and unit variance, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. copies of $\mathbf{x}$. Let $S := \frac{\sum_{i=1}^n \mathbf{x}_i}{\sqrt{n}}$ and $\boldsymbol{g}$ be a standard Gaussian. We have*

$$|\mathbb{E}[\psi(S)] - \mathbb{E}[\psi(\boldsymbol{g})]| \leqslant \frac{B}{\sqrt{n}}\mathbb{E}[|\mathbf{x}|^3].$$

Proof. Let $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ be independent standard Gaussians, and for $0 \leqslant i \leqslant n$, define the corresponding partially swapped version of S as

$$S_i := \frac{\mathbf{x}_1 + \ldots + \mathbf{x}_i + \boldsymbol{g}_{i+1} + \ldots + \boldsymbol{g}_n}{\sqrt{n}}.$$

Since $S_n = S$ and $S_0 \sim \gamma_1$, it suffices to show that for every $i \in [n]$,

$$|\mathbb{E}[\psi(S_{i-1})] - \mathbb{E}[\psi(S_i)]| = \frac{1}{n^{3/2}} B \mathbb{E}[|\mathbf{x}|^3]. \tag{17.4}$$

Denoting the common part of S_{i-1} and S_i as

$$A = \frac{\mathbf{x}_1 + \ldots + \mathbf{x}_i + \boldsymbol{g}_{i+2} + \ldots + \boldsymbol{g}_n}{\sqrt{n}},$$

we have

$$S_{i-1} = A + \frac{\boldsymbol{g}_i}{\sqrt{n}} \quad \text{and} \quad S_i = A + \frac{\mathbf{x}_i}{\sqrt{n}}.$$

By writing the Taylor expansion of ψ,

$$\psi(S_{i-1}) = \psi(A) + \psi'(A)\frac{\boldsymbol{g}_i}{\sqrt{n}} + \frac{1}{2}\psi''(A)\frac{\boldsymbol{g}_i^2}{n} + R,$$

for an error term R satisfying $|R| \leqslant \frac{|\boldsymbol{g}_i|^3}{6n^{3/2}} \sup_{x \in \mathbb{R}} |\psi'''(x)|$. We also obtain a similar formula for $\psi(S_{i-1})$, where $\boldsymbol{g}_i$ are

replaced by $\mathbf{x}_i$. Since $\mathbf{x}_i$ and $\boldsymbol{g}_i$ both have mean zero and unit variance, we have

$$|\mathbb{E}[\psi(S_{i-1})] - \mathbb{E}[\psi(S_i)]| \leqslant \frac{1}{6n^{3/2}} \frac{\mathbb{E}\left[|\boldsymbol{g}_i|^3 + |\mathbf{x}_i|^3\right]}{n^{3/2}} \sup_{x \in \mathbb{R}} |\psi'''(x)|.$$

Since $\boldsymbol{g}_i$ is a standard Gaussian,

$$\mathbb{E}|\boldsymbol{g_i}|^3 = 3\sqrt{\frac{2}{\pi}},$$

which combined with

$$\mathbb{E}|\mathbf{x}_i|^3 = \mathbb{E}|\mathbf{x}|^3 \geqslant (\mathbb{E}\mathbf{x}^2)^{3/2} \geqslant 1,$$

shows

$$\mathbb{E}|\boldsymbol{g_i}|^3 + \mathbb{E}|\mathbf{x}_i|^3 \leqslant 6\mathbb{E}|\mathbf{x}|^3.$$

We conclude Equation (17.4) as desired. □

Theorem 17.11 (Berry-Esseen weak form)**.** *Let $\mathbf{x}$ be a random variable with mean zero and unit variance, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. copies of $\mathbf{x}$. Let $S_n := \frac{\sum_{i=1}^n \mathbf{x}_i}{\sqrt{n}}$ and $\boldsymbol{g}$ be a standard Gaussian. We have*

$$|\operatorname{\mathbf{Pr}}[S_n \leqslant t] - \operatorname{\mathbf{Pr}}[\boldsymbol{g} \leqslant t]| \leqslant O\left(\frac{\mathbb{E}[|\mathbf{x}|^3]}{\sqrt{n}}\right)^{1/4} \quad \textit{for every } t \in \mathbb{R}.$$

Proof. We will approximate $1_{[x \leqslant t]}$ with a three times differentiable function $\psi(x)$, and apply Theorem 17.11 to ψ.

Pick a parameter $\varepsilon > 0$ and let $\psi : \mathbb{R} \to [0,1]$ be any function equal to 1 in $(\infty, t]$, vanishing on $[t+\varepsilon, \infty)$, and with $|\psi'''(x)| = O(\varepsilon^{-3})$ in $[t, t+\varepsilon]$. We have

$$\mathbb{E}[\psi(\boldsymbol{g})] = \operatorname{\mathbf{Pr}}[\boldsymbol{g} \leqslant t] + O(\varepsilon).$$

By Theorem 17.11, we have

$$|\mathbb{E}[\psi(S)] - \mathbb{E}[\psi(\boldsymbol{g})]| \leqslant O\left(\frac{\varepsilon^{-3}}{\sqrt{n}} \mathbb{E}[|\mathbf{x}|^3]\right),$$

which shows

$$|\operatorname{\mathbf{Pr}}[S \leqslant t] - \operatorname{\mathbf{Pr}}[\boldsymbol{g} \leqslant t]| \leqslant O(\varepsilon) + O\left(\frac{\varepsilon^{-3}}{\sqrt{n}} \mathbb{E}[|\mathbf{x}|^3]\right).$$

Optimizing ε, gives

$$|\operatorname{\mathbf{Pr}}[S_n \leqslant t] - \operatorname{\mathbf{Pr}}[\boldsymbol{g} \leqslant t]| \leqslant O\left(\frac{\mathbb{E}[|\mathbf{x}|^3]}{\sqrt{n}}\right)^{1/4}.$$

□

Chapter 18

Draft: Hypercontractivity for global functions

As we have observed in earlier chapters, the hypercontractivity of the noise operator on the discrete cube $\{0,1\}^n$ is a central tool in Boolean function analysis. However, when considering more general product spaces, functions such as dictators and juntas demonstrate that the noise operator does not maintain such strong hypercontractive properties.

For example, over the discrete cube $\{0,1\}^n$, hypercontractivity implies that every degree-1 function satisfies

$$\frac{\|f\|_4}{\|f\|_2} \leqslant \sqrt{3} = O(1).$$

In contrast, consider a large alphabet size m, and the dictator function $f : [m]^n \to \{0,1\}$, defined as $f(x) = \mathbf{1}_{[x_1=1]}$. Here, f is of degree 1, yet we have

$$\frac{\|f\|_4}{\|f\|_2} = m^{\frac{1}{4}},$$

which shows a significant gap between the two norms as m grows.

In Theorem 10.19, we established a weaker form of hypercontractivity for general product probability spaces $(\Omega,\mu)^n$. However, as illustrated above, when μ contains atoms with small probability masses, the noise parameter ρ must be very small for hypercontractivity to hold. Unfortunately, the dependency on μ limits the applicability of Theorem 10.19 for generalizing key results from the discrete cube to other domains $(\Omega,\mu)^n$.

Global hypercontractivity: In a breakthrough, Keevash, Lifshitz, Long, and Minzer [**?**] showed that, essentially, junta-like behaviour is the only obstacle to the strong hypercontractivity of the noise operator. More precisely, they extended the hypercontractive inequality to general discrete product measures under the additional assumption that the function is *global*, meaning it is not significantly affected by restricting a small set of coordinates. This class of functions naturally arises in results like Bourgain's sharp threshold theorem, which states that global functions exhibit sharp thresholds. Later, Keller, Lifshitz, and Marcus [**?**] later proved a sharp version of this hypercontractive inequality.

The discovery of hypercontractivity for global functions and its various applications has been one of the most fruitful research directions in Boolean function analysis in recent years, leading to several significant advances. To name a few applications:

- A stronger version of Bourgain's sharp threshold theorem [**?**];
- Progress on the inverse problem for the isoperimetric inequality on the Boolean cube [**?**]
- A shorter proof [**?**] of the breakthrough result of Khot, Minzer and Safra [**?**] on the expansion of the Grassmann graph, which was the main mathematical ingredient in the proof of the 2-to-2 games conjecture in complexity theory;
- Hypercontractivity and level-d inequality in the symmetric group [**?**], etc;

- Applications to the intersecting families of permutations [**?**];

18.1 Statement of global hypercontractivity

For the reader's convenience, we recall some definitions from Section 9.4.1 and Section 10.4. Let (Ω, μ) be a finite probability space and consider a function $f : (\Omega, \mu)^n \to \mathbb{R}$ and its Fourier-Walsh expansion $f = \sum_{S \subset [n]} F_S$.

Given a parameter $\rho \in [0, 1]$, the *ρ-equal* copy of $x \in \Omega^n$ as the random variable $\mathbf{y}$ that is sampled from Ω^n through the following process: for each $i \in [n]$, with probability ρ, set $\mathbf{y}_i = x_i$ and with probability $1 - \rho$, sample $\mathbf{y}_i$ from (Ω, μ). The noise operator T_ρ is defined as $T_\rho f(x) \coloneqq \mathbb{E}_{\mathbf{y}} f(\mathbf{y})$, where $\mathbf{y}$ is the ρ-equal copy of x. It satisfies

$$T_\rho f = \sum_{S \subseteq [n]} \rho^{|S|} F_S.$$

Definition 18.1 (Global functions)**.** Let (Ω, μ) be a finite probability space. Given $S \subseteq [n]$ and $x \in (\Omega, \mu)^n$, let $f_{S \to x} : (\Omega, \mu)^{[n] \setminus S} \to \mathbb{R}$ be the restriction of f defined as $f_{S \to x}(y) = f(x, y)$. We say that $f : (\Omega, \mu)^n \to \mathbb{R}$ is *r-global* for $r \geqslant 1$ if for every $S \subseteq [n]$ and $x \in \Omega^S$, we have

$$\|f_{S \to x}\|_2 \leqslant r^{|S|} \|f\|_2.$$

We think of $f : (\Omega, \mu)^n \to \mathbb{R}$ as a global function if it is r-global for some $r = O(1)$. In this sense, for growing m, the dictator function $f : [m]^n \to \{0, 1\}$ defined as $f(x) = \mathbf{1}_{[x_1 = 1]}$ in the introduction is not global. We have $\|f\|_2 = \frac{1}{\sqrt{m}}$ while $\|f_{\{1\} \to 1}\|_2 = \|1\|_2 = 1$, and therefore, f is not r-global for any $r < \sqrt{m}$.

Theorem 18.2 (Global Hypercontractivity)**.** *Consider $1 < p \leqslant 2 \leqslant q$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, and let (Ω, μ) be a finite probability space. If $r \geqslant 1$ and $f : (\Omega^n, \mu^n) \to \mathbb{R}$ is an r-global function, then for every*

$$0 \leqslant \rho \leqslant \frac{\log q}{32 r q},$$

we have

$$\|T_\rho f\|_p \geqslant \|f\|_2 \quad \text{and} \quad \|T_\rho f\|_q \leqslant \|f\|_2.$$

18.1.1 Proof of global hypercontractivity

Chapter 19

Draft: Invariance Principle and Majority is Stablest

The invariance principle, proved by Mossel, O'Donnell, and Oleszkiewicz in [MOO10], is a useful generalization of the Central Limit Theorem that bridges the worlds of Gaussian analysis and Boolean function analysis. It has been a crucial tool in translating many results from the Gaussian to the setting of Boolean functions on the discrete cube. One can view the invariance principle as a generalization of the Berry-Esseen theorem (Theorem 17.9) to multilinear polynomials.

Consider a multilinear polynomial $f = \sum \widehat{f}(S) \prod_{i \in S} x_i$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent random variables, each with mean zero and unit variance, and let $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ be independent standard Gaussians. Roughly speaking, the invariance principle states that when f does not depend too much on individual coordinates, then the distribution of $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is similar to the distribution of $f(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$.

Similar to the proof of the Barry-Esseen theorem presented for Theorem 17.9, Mossel et al.'s proof of the invariance principle uses the replacement method. Their proof later inspired [**?**] to use the replacement method to prove the hypercontractivity for global functions. In fact, global hypercontractivity also implies a version of the invariance principle [**?**].

One of the original applications of Mossel, O'Donnell, and Oleszkiewicz in [MOO10] for the invariance principle was proving a conjecture of Subhash Khot about the noise stability of large subsets of the hypercube that do not have influential variables. Khot's conjecture, which was known as the "majority is stablest conjecture", has important applications in the area of hardness of approximation. We will discuss the proof of this conjecture in ??.

19.1 Invariance principle

Theorem 19.1 (Invariance Principal I [**?**])**.** *Let $Q(x_1, \ldots, x_n) = \sum_{S \leqslant [n]} \alpha_S \prod_{i \in S} x_i$ be a multilinear polynomial with real coefficients satisfying the following three conditions.*

$$\begin{aligned} &\deg(Q) \leqslant d \\ &\sum_{|S|>0} \alpha_S^2 = 1 \\ I_i \coloneqq &\sum_{S : i \in S} \alpha_S^2 \leqslant \tau \qquad\qquad \text{for all } i \in [n] \end{aligned}$$

Then for i.i.d. ± 1 uniform random variables $(\varepsilon) = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$ and independent standard Gaussians $\boldsymbol{g} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$, we have

$$\sup_{t \in \mathbb{R}} |\mathbf{Pr}[Q(\boldsymbol{\varepsilon}) \leqslant t] - \mathbf{Pr}[Q(\boldsymbol{g}) \leqslant t]| \leqslant O(d\tau^{\frac{1}{8d}}).$$

Similar to the case of the Berry-Esseen theorem, Theorem 19.1 follows from the more technical version of the

invariance principle by approximating the step function with an appropriate four times differentiable function. We omit the details.

Theorem 19.2 (Invariance Principal II)**.** *Suppose Q satisfies the assumptions of Theorem 19.1 and let $\psi : \mathbb{R} \to \mathbb{R}$ be a four times differentiable function with $\sup_t |\psi^{(4)}(t)| < B$. We have*

$$|\mathbb{E}[\psi(Q(\boldsymbol{\varepsilon}))] - \mathbb{E}[\psi(Q(\boldsymbol{g}))]| \leqslant O(d9^d B\tau).$$

Proof. Let $Z_i = Q(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_i, \boldsymbol{\varepsilon}_{i+1}, \ldots, \boldsymbol{\varepsilon}_n)$. We claim that

$$|\mathbb{E}\psi(Z_{i-1}) - \mathbb{E}\psi(Z_i)| \leqslant O(B9^d I_i^2). \tag{19.1}$$

First, we show that the theorem can be extracted from this claim. Indeed,

$$\begin{aligned}
|\mathbb{E}\psi(Z_0) - \mathbb{E}\psi(Z_n)| &\leqslant \sum_{i=1}^{n} |\mathbb{E}\psi((Z_{i-1}) - \mathbb{E}\psi(Z_i)| \leqslant O(B9^d) \sum_{i=1}^{n} I_i^2 \\
&= O(B9^d)(\max I_i) \sum I_i \leqslant O(B9^d\tau) \sum I_i \\
&= O(B9^d\tau) \sum_{|S|>0} |S|\alpha_S^2 \leqslant O(dB9^d\tau) \sum_{|S|>0} \alpha_S^2 = O(\tau B9^d d).
\end{aligned}$$

To prove the claim, we separate the monomials according to whether they contain x_i and write

$$\begin{aligned}
Q(x_1, \ldots, x_n) &= \sum_{S: i \notin S} \alpha_S \prod_{j \in S} x_j \;+\; x_i \sum_{S: i \in S} \alpha_S \prod_{j \in S \setminus \{i\}} x_j \\
&= R(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) + x_i S(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)
\end{aligned}$$

Let

$$\boldsymbol{r} \coloneqq r(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{i-1}, \boldsymbol{\varepsilon}_{i+1}, \ldots, \boldsymbol{\varepsilon}_n) \text{ and } \boldsymbol{s} \coloneqq S(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{i-1}, \boldsymbol{\varepsilon}_{i+1}, \ldots, \boldsymbol{\varepsilon}_n).$$

We have $Z_{i-1} = \boldsymbol{r} + \boldsymbol{\varepsilon}_i \boldsymbol{s}$ and $Z_i = \boldsymbol{r} + \boldsymbol{g}_i \boldsymbol{s}$. By Taylor's theorem, we have

$$\begin{aligned}
|\mathbb{E}\psi(Z_{i-1}) - \mathbb{E}\psi(Z_i)| \quad \leqslant \quad & \Big| \mathbb{E}\psi(\boldsymbol{r}) + \boldsymbol{\varepsilon}_i \boldsymbol{s}\psi'(\boldsymbol{r}) + \frac{(\boldsymbol{\varepsilon}_i \boldsymbol{s})^2}{2}\psi''(\boldsymbol{r}) + \frac{(\boldsymbol{\varepsilon}_i \boldsymbol{s})^3 \psi^{(3)}(\boldsymbol{r})}{6} + E_1 \\
& - \mathbb{E}\psi(\boldsymbol{r}) - \boldsymbol{g}_i \boldsymbol{s}\psi'(\boldsymbol{r}) - \frac{(\boldsymbol{g}_i \boldsymbol{s})^2}{2}\psi''(\boldsymbol{r}) - \frac{(\boldsymbol{g}_i \boldsymbol{s})^3 \psi^{(3)}(\boldsymbol{r})}{6} - E_2 \Big|,
\end{aligned}$$

where $|E_1| \leqslant \frac{\sup_t |\psi^{(4)}(t)| (\boldsymbol{\varepsilon}_i \boldsymbol{s})^4}{24} \leqslant \frac{B(\boldsymbol{\varepsilon}_i \boldsymbol{s})^4}{24}$, and similarly, $|E_2| \leqslant \frac{B(\boldsymbol{g}_i \boldsymbol{s})^4}{24}$. All terms cancel except E_1 and E_2. So the expression is bounded by:

$$\mathbb{E}\left|\frac{B(\boldsymbol{\varepsilon}_i \boldsymbol{s})^4}{24}\right| + \mathbb{E}\left|\frac{B(\boldsymbol{g}_i \boldsymbol{s})^4}{24}\right| \leqslant \frac{B}{24}\mathbb{E}\boldsymbol{s}^4 + \frac{3B}{24}\mathbb{E}\boldsymbol{s}^4 \leqslant \frac{B}{6}\mathbb{E}\boldsymbol{s}^4.$$

Since $\boldsymbol{s}$ is a multilinear polynomial of degree at most d in variables $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{i-1}, \boldsymbol{\varepsilon}_{i+1}, \ldots, \boldsymbol{\varepsilon}_n$, by *hypercontractivity*, we have

$$\frac{B}{6}\mathbb{E}\boldsymbol{s}^4 \leqslant \frac{B9^d}{6}(\mathbb{E}\boldsymbol{s}^2)^2 = \frac{B9^d}{6}\sum_{i \in S} \alpha_S^2 = \frac{B9^d}{6} I_i^2,$$

which completes the proof of Equation (19.1). □

19.2 The Majority is Stablest Theorem

In this section, we focus on the noise stability for subsets of the discrete cube $\{0,1\}^n$. Recall that the *ρ-noise stability* of a function $f : \{0,1\}^n \to \{0,1\}$ is defined as

$$\mathrm{Stab}_\rho(f) \coloneqq \langle T_\rho f, f \rangle = \sum_{S \subseteq [n]} \rho^{|S|} \widehat{f}(S)^2.$$

We are interested in understanding the Boolean functions that have large noise stability. In Proposition 16.2, we used hypercontractivity to establish the small-set expansion property of the noisy cube, which states that the sets with small density are not stable: $\mathrm{Stab}_\rho(A) \leqslant \mathbb{E}[A]^{\frac{2}{1+\rho}}$.

In this chapter, we focus on large sets. For example, one might ask among functions $f : \{0,1\}^n \to \{0,1\}$ with $\mathbb{E}[f] = \frac{1}{2}$ what is the largest possible value of $\mathrm{Stab}_\rho(f)$? If we use the spectral gap approach by separating the principal coefficient and upper bounding $\rho^{|S|}$ by ρ, we get

$$\mathrm{Stab}_\rho(f) = \sum_{S\subseteq[n]} \rho^{|S|}|\hat{f}(S)|^2 \leqslant |\mathbb{E}f|^2 + \rho\sum_{S\neq\emptyset} |\hat{f}(S)|^2 = \frac{1}{4} + \frac{\rho}{4}.$$

This bound is indeed sharp as achieved by half-cubes; for example, if $f(x) = x_1$, then $f = \frac{1}{2} + \frac{1}{2}\chi_{\{i\}}$, and therefore,

$$\mathrm{Stab}_\rho(f) = \frac{1}{4} + \frac{\rho}{4}.$$

In general, if the value of the function f depends only on a few coordinates, then the function will become stable under noise as with some non-negligible probability x, and its correlated copy ρ will be the same on those coordinates. It turns out that the question becomes more interesting if we avoid these examples by assuming that all the variables have small influences.

Half-cubes are stable because their Fourier coefficients are concentrated in the first level. In [**?**], Bourgain proved that if a Boolean function $f : \{0,1\}^n \to \{0,1\}$ is not close to being a junta, then it must have a significant Fourier mass of at least $d^{-1/2-o(1)}$ on $\|f^{\geqslant d}\|_2^2$. Bourgain's bound was later sharpened by Kindler and O'Donnell by first proving a sharp bound in the Gaussian setting and then translating it to the discrete cube using an invariance principle.

Theorem 19.3 (Kindler and O'Donnell [**?**])**.** *If $f : \{0,1\}^n \to \{0,1\}$ is balanced and $I_i \leqslant 10^{-d}$ for all $i \in [n]$, then*

$$\|f^{\geqslant d}\|_2^2 = \sum_{|S|\geqslant d} |\hat{f}(S)|^2 \geqslant d^{-\frac{1}{2}-o\left(\sqrt{\frac{\ln\ln d}{\ln d}}\right)} = d^{-1/2-o(1)}.$$

Theorem 19.3, whose proof is highly nontrivial, provides an upper bound on noise stability

Corollary 19.4. *If $f : \{0,1\}^n \to \{0,1\}$ is balanced and $I_i(f) = 2^{-O(1/\varepsilon)}$ for all $i \in [n]$, then*

$$\mathrm{Stab}_{1-\varepsilon}(f) \leqslant \frac{1}{2} - \varepsilon^{1/2+o(1)}$$

While Corollary 19.4 is an improvement over the spectral gap upper bound of $\frac{1}{2} - \frac{\varepsilon}{4}$, using the tail bound on the Fourier coefficients does not seem to be the optimal approach for upper-bounding the noise stability. To illustrate this, consider the majority function $\textsc{Maj}_n : \{0,1\}^n \to \{0,1\}$, defined as $\textsc{Maj}_n(x) = 1$ iff $\sum x_i \geqslant \frac{n}{2}$. While the tail bound of Theorem 19.3 is sharp for $\textsc{Maj}_n$, the following theorem shows that $\mathrm{Stab}_\rho(\mathrm{Maj}_n)$ is much smaller than the upper bound of Corollary 19.4.

Theorem 19.5. *The noise stability of the majority function satisfies*

$$\lim_{n\to\infty} \mathrm{Stab}_\rho(\textsc{Maj}_n) = \frac{1}{4} + \frac{\arcsin\rho}{2\pi}.$$

Note that $\frac{1}{4} + \frac{\arcsin\rho}{2\pi}$ is the Gaussian noise sensitivity of homogenous halfspaces as shown in Theorem 17.7. As we discussed in Theorem 17.8, Borell proved the analogous statement in the Gaussian setting in 1983, where no condition on the influences is necessary. Mossel, O'Donnell, and Oleszkiewicz used their invariance principle to resolve Khot's conjecture and deduce the following theorem from Borell's result.

Theorem 19.6 (Majority is Stablest [**?**])**.** *For $0 < \rho < 1$, if $f : \{-1,1\}^n \to \{0,1\}$ is balanced and $I_i(f) \leqslant \varepsilon$ for all $i \in [n]$, then*

$$\mathrm{Stab}_\rho(f) \leqslant \frac{1}{4} + \frac{\arcsin\rho}{2\pi} + O\left(\frac{\log\log 1/\varepsilon}{\log 1/\varepsilon}\right) = \frac{1}{4} + \frac{\arcsin\rho}{2\pi} + o(\varepsilon).$$

Proof. Consider the polynomial representation $f = \sum \widehat{f(S)} \prod_{i \in S} x_i$. Let $(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$ be an independent standard Gaussian. We have

$$\mathrm{Stab}_\rho(f) = \sum \rho^{|S|} |\widehat{f}(S)|^2 = \mathrm{Stab}_\rho(f(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)).$$

We want to apply the invariance principle to replace the ± 1-valued random variables with Gaussians. However, since the degree of f can be large, we cannot apply the invariance principle directly to f. Instead, we apply a *smoothed* version of the theorem, which can be applied to $T_\beta f$ for $\beta < 1$. Let $\rho = \rho' \beta^2$ where $\beta < 1$ is a parameter very close to 1 to be determined later.

$$\mathrm{Stab}_\rho(f) = \sum \rho^{|S|} |\widehat{f}(S)|^2 = \sum (\rho' \beta^2)^{|S|} |\widehat{f}(S)|^2 = \mathrm{Stab}_{\rho'}(T_\beta f(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)).$$

Now using the smoothed invariance, $T_\beta f(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$ is close in distribution to $T_\beta f(\varepsilon_1, \ldots, \varepsilon_n)$ and hence it cannot be far from being in $[-1, 1]$. To make this precise, we define function ξ as follows:

$$\xi : t \to \begin{cases} 0 & |t| \leqslant 1 \\ (|t| - 1)^2 & |t| > 1 \end{cases}$$

Note that ξ measures the L_2-distance of t from its truncated value in $[-1, 1]$. By the invariance principle applied to random variables $\boldsymbol{r} = T_\beta f(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$ and $\boldsymbol{s} = T_\beta f(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$, we have $|\mathbb{E}\xi(\boldsymbol{r}) - \mathbb{E}\xi(\boldsymbol{s})| \leqslant \tau^{\Omega(1-\beta)}$. Let $\boldsymbol{s}'$ be the truncation of $\boldsymbol{s}$ to the interval $[-1, 1]$:

$$\boldsymbol{s}' = \begin{cases} \boldsymbol{s} & |\boldsymbol{s}| \leqslant 1 \\ 1 & \boldsymbol{s} > 1 \\ -1 & \boldsymbol{s} < -1 \end{cases}.$$

By assumption, $f(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) \in [-1, 1]$ and since T_β is an averaging operator, $T_\beta f(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) \in [-1, 1]$ and hence $\xi(R) = 0$. Thus,

$$\mathbb{E}|\xi(\boldsymbol{s})| = \mathbb{E}(\boldsymbol{s} - \boldsymbol{s}')^2 \leqslant \tau^{\Omega(1-\beta)}$$

which shows

$$\begin{aligned} |\operatorname{Stab}_{\rho'}(\boldsymbol{s}) - \operatorname{Stab}_{\rho'}(\boldsymbol{s}')| &= |\mathbb{E}\boldsymbol{s} U_{\rho'} \boldsymbol{s} - \mathbb{E}\boldsymbol{s}' U_{\rho'} \boldsymbol{s}'| \leqslant |\mathbb{E}\boldsymbol{s} U_{\rho'} \boldsymbol{s} - \mathbb{E}\boldsymbol{s}' U_{\rho'} \boldsymbol{s}| + |\mathbb{E}\boldsymbol{s}' U_{\rho'} \boldsymbol{s} - \mathbb{E}\boldsymbol{s}' U_{\rho'} \boldsymbol{s}'| \\ &\leqslant \|\boldsymbol{s} - \boldsymbol{s}'\|_2 \|U_{\rho'} \boldsymbol{s}\|_2 + \|\boldsymbol{s}'\|_2 \|U_{\rho'}(\boldsymbol{s} - \boldsymbol{s}')\|_2 \leqslant \|\boldsymbol{s} - \boldsymbol{s}'\|_2 \|\boldsymbol{s}\|_2 + \|\boldsymbol{s}'\|_2 \|\boldsymbol{s} - \boldsymbol{s}'\|_2 \\ &\leqslant \tau^{\Omega(1-\beta)}. \end{aligned}$$

By Borell's theorem (Theorem 17.8), $\mathrm{Stab}_{\rho'}(\boldsymbol{s}') \leqslant \mathrm{Stab}_{\rho'}(\mathbf{1}_{x \leqslant t_0})$ where t_0 is chosen so that $\mathbb{E}\mathbf{1}_{\boldsymbol{g} \leqslant t_0} = \mathbb{E}\boldsymbol{s}'$ for a standard Gaussian $\boldsymbol{g}$.

It remains to show that $\frac{1}{2} \approx \mathbb{E}\boldsymbol{s}'$, which would imply $t_0 \approx 0$. We have

$$\left| \frac{1}{2} - \mathbb{E}\boldsymbol{s}' \right| = |\mathbb{E}\boldsymbol{s} - \boldsymbol{s}'| \leqslant \|\boldsymbol{s} - \boldsymbol{s}'\|_2 \leqslant \tau^{\Omega(1-\beta)}.$$

It follows that

$$|\operatorname{Stab}_{\rho'}(\mathbf{1}_{x \leqslant 0}) - \operatorname{Stab}_{\rho'}(\mathbf{1}_{x \leqslant t_0})| \leqslant O\left(\frac{1-\beta}{1-\rho}\right).$$

Therefore,

$$\mathrm{Stab}_\rho(f) = \mathrm{Stab}_\rho(\mathbf{1}_{x \geqslant 0}) + O\left(\tau^{\Omega(1-\beta)} + \frac{1-\beta}{1-\rho}\right).$$

The theorem follows by optimizing over β. □

19.3 Arrows Theorem and Majority is stablest

Condorcet Method for Ranking 3 Candidates : In an election with n voters and 3 candidates, A, B and C, each voter submits 3 bits representing their preferences. The first bit indicates whether they prefer A to B; The second and the third bits indicate, respectively, their preference between B and C, and between C and A. These preferences

are aggregated into 3 strings $x, y, z \in (-1,1)^n$. A Boolean function $f : \{-1,1\}^n \mapsto \{-1,1\}$ is applied to x, y and z and the aggregated preference is represented by $(f(x), f(y), f(z))$.

Condorcet Paradox: If f is the Majority function, it is possible to have an irrational outcome, in which all 3 aggregated bits are 1 or all are -1 representing preferences $A < B < C < A$ or $A > B > C > A$.

Definition 19.7. A triple $(a,b,c) \in \{-1,1\}^3$ is called rational if it corresponds to a non-cyclic ordering.

Theorem 19.8 (Arrow's Impossibility Theorem)**.** *The only functions f that never give irrational outcomes are dictator functions $f(x) = x_i$ or $f(x) = 1 - x_i$ for some i.*

Note that every voter has 6 possible rational rankings. Suppose that every voter votes independently at random from the 6 possible choices. Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \{-1,1\}^n$ be the corresponding random string. Note

$$\mathbf{1}_{[a_1=a_2=a3]} = \frac{1}{4} + \frac{1}{4}a_1a_2 + \frac{1}{4}a_1a_3 + \frac{1}{4}a_2a_3,$$

and therefore,

$$\begin{aligned}\mathbf{Pr}[(f(\mathbf{x}), f(\mathbf{y}), f(\mathbf{z}))] &= 1 - \mathbb{E}1_{[f(\mathbf{x})=f(\mathbf{y})=f(\mathbf{z})]} = \frac{3}{4} - \frac{1}{4}\mathbb{E}f(\mathbf{x})f(\mathbf{y}) - \frac{1}{4}\mathbb{E}f(\mathbf{x})f(\mathbf{z}) - \frac{1}{4}\mathbb{E}f(\mathbf{y})f(\mathbf{z}) \\ &= \frac{3}{4} - \frac{3}{4}\mathbb{E}f(\mathbf{x})f(\mathbf{y}) = \frac{3}{4} - \frac{3}{4}\sum_{S,T}\widehat{f}(S)\widehat{f}(T)\mathbb{E}\chi_S(\mathbf{x})\chi_T(\mathbf{y}).\end{aligned}$$

Furthermore,

$$\mathbb{E}\chi_S(\mathbf{x})\chi_T(y) = (\prod_{i\in S\cap T}\mathbb{E}\mathbf{x}_i\mathbf{y}_i)(\prod_{i\in S\setminus T}\mathbb{E}\mathbf{x}_i)(\prod_{i\in T\setminus S}\mathbb{E}\mathbf{y}_i)$$

Since $\mathbb{E}\mathbf{y}_i = \mathbb{E}\mathbf{x}_i = 0$ and $\mathbb{E}\mathbf{x}_i\mathbf{y}_i = \frac{2}{6} - \frac{4}{6} = -\frac{1}{3}$, we have

$$\mathbb{E}\chi_S(\mathbf{x})\chi_T(\mathbf{y}) = \begin{cases} 0 & S \neq T \\ (\frac{-1}{3})^{|S|} & S = T \end{cases}.$$

Hence,

$$\mathbf{Pr}[(f(\mathbf{x}), f(\mathbf{y}), f(\mathbf{z})) \text{ is rational}] = \frac{3}{4} + \frac{3}{4}\sum\left(\frac{-1}{3}\right)^{|S|}\widehat{|f(S)|}^2 \leqslant \frac{3}{4} + \frac{3}{4}\operatorname{Stab}_{\frac{-1}{3}}(f)$$

We conclude the following theorem due to Kalai.

Theorem 19.9. *If $f : \{-1,1\}^n \to \{-1,1\}$ satisfies $I_i(f) = o_n(1)$ and $\mathbb{E}f = 0$, then assuming that all the voters vote independently and randomly from the six possible rational votes,*

$$\mathbf{Pr}[\textit{output of } f \textit{ is rational}] \leqslant 0.9123 + o_n(1).$$

Chapter 20

Learning via Fourier Coefficients

In this chapter, we study two applications of Fourier analysis to computational learning theory.

20.1 PAC learning under uniform distribution

We begin with an overview of the PAC (Probably Approximately Correct) learning framework from computational learning theory, introduced by Leslie Valiant [Val84].

A *binary concept class* over a *domain* X is simply a set $\mathcal{C}$ of functions $f : X \to \{0,1\}$. Here, the term *binary* signifies that the range is the two-element set $\{0,1\}$. The elements of $\mathcal{C}$ are called *concepts*.

In the learning problem, a concept $f \in \mathcal{C}$ and a distribution μ on X are unknown to the learner. The learner who knows $\mathcal{C}$ but not f or μ is trying to *learn* f by observing its values on a few i.i.d. samples drawn from μ. More precisely, the learner will receive a batch of samples of the form $(x, f(x))$ where $x \sim \mu$ are drawn independently, and they must produce a *hypothesis* $h : X \to \{0,1\}$ as a predictor for f.

The quality of h is measured by its *population loss*,

$$\mathcal{L}_\mu(h) \coloneqq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mu}[h(\mathbf{x}) \neq f(\mathbf{x})].$$

We emphasize that h does not need to be in the concept class $\mathcal{C}$; it simply needs to predict f well on examples from μ.

In this section, we will be only interested in the case where μ is the *uniform distribution*. While the uniform distribution may not reflect real-world scenarios, this setting has theoretical applications and has been extensively studied. In particular, a substantial body of research addresses the problem of learning concept classes, such as juntas, under uniform distribution [MOS03].

Consider $f : \mathbb{Z}_2^n \to \{0,1\}$. Let us explore what information we can learn about the Fourier spectrum of f from uniform samples. Consider a fixed $a \in \mathbb{Z}_2^n$, and recall that

$$\widehat{f}(a) = \langle f, \chi_a \rangle = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\chi_a(\mathbf{x})].$$

Since $|f(x)\chi_a(x)| \leqslant 1$, by Chernoff bound, if $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{Z}_2^n$ are sampled uniformly and independently, then with high probability the empirical estimate

$$\widetilde{f}(a) = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)\chi_a(\mathbf{x}_i),$$

will be very close to the actual expected value $\widehat{f}(a) = \mathbb{E}[f(\mathbf{x})\chi_a(\mathbf{x})]$. Thus, a few samples typically suffice to accurately estimate $\widehat{f}(a)$. The following lemma immediately follows from Chernoff bound.

Lemma 20.1. *Let $f : \mathbb{Z}_2^n \to \{0,1\}$, $a \in \mathbb{Z}_2^n$ and $\delta, \varepsilon \in (0, \frac{1}{2})$. For $m = O(\log(1/\delta)\varepsilon^{-2})$, if $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are independently and uniformly sampled from $\mathbb{Z}_2^n$, then the empirical estimate*

$$\widetilde{\boldsymbol{f}}(\boldsymbol{a}) \coloneqq \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)\chi_a(\mathbf{x}_i)$$

satisfies

$$\mathbf{Pr}\left[\left|\widetilde{\boldsymbol{f}}(\boldsymbol{a}) - \widehat{f}(a)\right| > \varepsilon\right] \leqslant \delta.$$

Let us return to the problem of learning an unknown $f : \mathbb{Z}_2^n \to \{0,1\}$ in a concept class $\mathcal{C}$. By Lemma 20.1, we can estimate individual coefficients well from a few samples. However, since there are 2^n coefficients S, estimating all of them will not be computationally efficient. Moreover, the probabilistic errors in these estimates can accumulate to a large error unless we take ε and δ to be exponentially small. Indeed, one should not expect to learn a generic function from only a few random samples, as for such a function, the values at sampled points provide no information about values at unsampled points.

Suppose now that we have the additional information about the Fourier spectrum of the functions in $\mathcal{C}$ that the mass of their Fourier coefficients is concentrated within a small set $\mathcal{S} \subseteq \mathbb{Z}_2^n$, meaning

$$\sum_{a \notin \mathcal{S}} |\widehat{f}(a)|^2 \leqslant \varepsilon.$$

In this case, we can limit our focus to estimating only the coefficients $\widehat{f}(a)$ for $a \in \mathcal{S}$ and obtain an accurate estimate of f as

$$\sum_{a \in \mathcal{S}} \widetilde{f}(a)\chi_a \approx f.$$

Theorem 20.2 (Fourier concentration and PAC learning [LMN93])**.** *Let $\mathcal{C}$ be a class of functions $f : \mathbb{Z}_2^n \to \{0,1\}$. Suppose there is a subset $\mathcal{S} \subseteq \mathbb{Z}_2^n$ of size m such that every $f \in \mathcal{C}$ satisfies*

$$\sum_{a \notin \mathcal{S}} |\widehat{f}(a)|^2 \leqslant \varepsilon.$$

There is a randomized algorithm that using at most $O(\varepsilon^{-1} m \log(m/\delta))$ uniform samples from an unknown $f \in \mathcal{C}$, outputs a Boolean function $h : \mathbb{Z}_2^n \to \{0,1\}$ such that with probability at least $1 - \delta$, we have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}}[f(\mathbf{x}) \neq h(\mathbf{x})] \leqslant 8\varepsilon.$$

Proof. By Lemma 20.1, we can use $O\left(\log(m/\delta)\left(\frac{\sqrt{m}}{\sqrt{\varepsilon}}\right)^2\right) = O\left(\varepsilon^{-1} m \log(m/\delta)\right)$ samples to estimate $\widetilde{\boldsymbol{f}}(\boldsymbol{a}) \approx \widehat{f}(a)$ for each $a \in \mathcal{S}$ such that

$$\mathbf{Pr}\left[|\widehat{f}(a) - \widetilde{\boldsymbol{f}}(\boldsymbol{a})| > \frac{\sqrt{\varepsilon}}{\sqrt{m}}\right] \leqslant \frac{\delta}{m}.$$

Hence, by the union bound,

$$\mathbf{Pr}\left[\exists a \in \mathcal{S} \text{ such that } |\widehat{f}(a) - \widetilde{\boldsymbol{f}}(\boldsymbol{a})| > \frac{\sqrt{\varepsilon}}{\sqrt{m}}\right] \leqslant \delta.$$

Let

$$g \coloneqq \sum_{a \in \mathcal{S}} \widetilde{f}(a)\chi_a(x).$$

We will show that if our estimates are successful, which happens with probability at least $1 - \delta$, then g is a good estimate of f. Indeed, by Parseval's identity, we have

$$\mathbb{E}|f(\mathbf{x}) - g(\mathbf{x})|^2 = \sum_{a \in \mathcal{S}} |\widehat{f}(a) - \widetilde{f}(a)|^2 + \sum_{a \notin \mathcal{S}} |\widehat{f}(a)|^2 \leqslant m\left(\frac{\sqrt{\varepsilon}}{\sqrt{m}}\right)^2 + \varepsilon \leqslant 2\varepsilon.$$

Let h be the Boolean rounding of g defined as $h(x) = 0$ if $g(x) < 1/2$ and otherwise $h(x) = 1$. Since f is Boolean, we always have $|f(x) - h(x)| \leqslant 2|f(x) - g(x)|$. Therefore,

$$\mathbf{Pr}[f(\mathbf{x}) \neq h(\mathbf{x})] = \mathbb{E}|f(\mathbf{x}) - h(\mathbf{x})|^2 \leqslant 4\mathbb{E}|f(\mathbf{x}) - g(\mathbf{x})|^2 \leqslant 8\varepsilon.$$

□

Example 20.3 (Small total influence)**.** Let $\mathcal{C}$ be the class of functions $f : \{0,1\}^n \to \{0,1\}$ with $I_f \leqslant k$. Then, we have

$$k = \sum_{S \subseteq [n]} |S| \widehat{f}(a)^2,$$

which shows

$$\sum_{|S| \geqslant k/\varepsilon} |\widehat{f}(a)|^2 \leqslant \varepsilon.$$

Therefore, we can apply Theorem 20.2 with $\mathcal{S} = \{S \subseteq [n] \ : \ |S| \leqslant k/\varepsilon\}$, which is of size at most $n^{k/\varepsilon}$.

20.2 Goldreich and Levin: Learning via queries

In Section 20.1, we discussed a learning algorithm that can learn an unknown $f : \mathbb{Z}_2^n \to \{0,1\}$ in a class $\mathcal{C}$ if the Fourier mass of every function in $\mathcal{C}$ is concentrated on a small fixed set $\mathcal{S}$ of characters. The discussed algorithm, which has prior knowledge of $\mathcal{C}$ and therefore knows $\mathcal{S}$, estimates the Fourier coefficients of f only for characters in $\mathcal{S}$.

This section considers classes where the Fourier mass of every $f \in \mathcal{C}$ is concentrated on some small set of characters $\mathcal{S}_f$ that varies with f and is thus unknown to the learner. Goldreich and Levin [GL89] proved it is possible to learn such classes if the learner can query the values of $f(x)$ at any x they choose. The query model allows the learner to identify and estimate the significant Fourier coefficients without knowing $\mathcal{S}_f$ in advance.

The Goldreich-Levin algorithm first detects the large Fourier coefficients of f by partitioning the Fourier coefficients according to their prefix. For $k \in [n]$ and $\alpha \in \mathbb{Z}_2^k$, define $f_\alpha : \mathbb{Z}_2^{n-k} \to \mathbb{R}$ as

$$f_\alpha(x) = \mathbb{E}_{\mathbf{y} \in \mathbb{Z}_2^k} f(\mathbf{y}, x) \chi_\alpha(\mathbf{y}).$$

Note

$$f_\alpha(x) = \sum_{(z_1,z_2) \in \mathbb{Z}_2^n} \widehat{f}(z_1,z_2)\chi_{z_2}(x)\mathbb{E}_{\mathbf{y}}\left[\chi_{z_1+\alpha}(\mathbf{y})\right] = \sum_{(\alpha,z_2) \in \mathbb{Z}_2^n} \widehat{f}(\alpha,z_2)\chi_{z_2}(x),$$

where in the first sum $z_1 \in \mathbb{Z}_2^k$ and $z_2 \in \mathbb{Z}_2^{n-k}$. By Parseval's identity,

$$\|f_\alpha\|_2^2 = \mathbb{E}_{\mathbf{x}}[f_\alpha(\mathbf{x})^2] = \sum_{(\alpha,z_2) \in \mathbb{Z}_2^n} |\widehat{f}(\alpha,z_2)|^2.$$

We will show that with oracle access to the values of $\|f_\beta\|_2$, the algorithm described in Algorithm 1 can find all characters a that satisfy $|\widehat{f}(a)| \geqslant \tau$.

Claim 20.4. *The procedure described in Algorithm 1 returns the set of all $a \in \mathbb{Z}_2^n$ that have prefix α and satisfy $|\widehat{f}(a)| \geqslant \tau$. Furthermore, the algorithm inspects the values of $\|f_\beta\|_2$ for at most $\frac{2n}{\tau^2}$ strings β.*

Proof. Observe that for $k = n$, we have $\|f_\alpha\|_2 = |\widehat{f}(\alpha)|$, and moreover, if β is a prefix of a, then $|\widehat{f}(a)| \leqslant \|f_\beta\|_2$. Therefore, the procedure correctly returns all the desired a.

To bound the number of queries $\|f_\beta\|_2$, note that for every $k \in [n]$, we have $\sum_{\beta \in \mathcal{Z}_2^k} \|f_\beta\|_2^2 = \|f\|_2^2$, and therefore, there are at most $\frac{1}{\tau^2}$ prefixes $\beta \in \mathbb{Z}_2^k$ with $\|f_\beta\|_2 \geqslant \tau$. The bound on the number of queries of the form $\|f_\beta\|_2$ follows. □

Algorithm 1 The Goldreich-Levin algorithm returns the set of characters a with prefix $\alpha \in \{0,1\}^k$ that satisfy $|\widehat{f}(a)| \geqslant \tau$. The algorithm assumes oracle access to the values of $\|f_\beta\|_2$ for any β.

procedure FIND_LARGE_FOURIER(α, k, τ)
  if $\|f_\alpha\|_2 < \tau$ **then**
    return $\emptyset$
  else
    if $k = n$ **then**
      return $\{\alpha\}$
    else
      $\alpha_0 \leftarrow (\alpha, 0) \in \mathbb{Z}_2^{k+1}$
      $\alpha_1 \leftarrow (\alpha, 1) \in \mathbb{Z}_2^{k+1}$
      return FIND_LARGE_FOURIER($\alpha_0, k+1, \tau$)$\cup$ FIND_LARGE_FOURIER($\alpha_1, k+1, \tau$)
    end if
  end if
end procedure

While we cannot compute the exact values of $\|f_\beta\|_2$ from a few queries, the following claim shows we can estimate them.

Claim 20.5. *For every $\lambda > 0$, given $\beta \in \mathbb{Z}_2^k$, we can make $3N$ queries to f and returns a value ρ_β such that with probability at least $2e^{-\lambda^2 N/2}$, we have $|\rho_\beta^2 - \|f_\beta\|_2^2| \leqslant \lambda$.*

Proof. We have

$$\|f_\beta\|_2^2 = \mathbb{E}_{\mathbf{x}}[f_\beta(\mathbf{x})^2] = \mathbb{E}_{\mathbf{x}\in\mathbb{Z}_2^{n-k}}\left(\mathbb{E}_{\mathbf{y}\in\mathbb{Z}_2^k} f(\mathbf{y},\mathbf{x})\chi_\beta(\mathbf{y})\right)^2 = \mathbb{E}_{\mathbf{x}\in\mathbb{Z}_2^{n-k}}\mathbb{E}_{\mathbf{y}_1,\mathbf{y}_2\in\mathbb{Z}_2^k}\left[f(\mathbf{y}_1,\mathbf{x})\chi_\beta(\mathbf{y}_1)f(\mathbf{y}_2,\mathbf{x})\chi_\beta(\mathbf{y}_2)\right].$$

Since $|f(y_1,x)\chi_\beta(y_1)f(y_2,x)\chi_\beta(y_2)| \leqslant 1$, the Chernoff bound implies that by taking many random points x, y_1, y_2, we can obtain an accurate estimate for $\|f_\beta\|_2^2$. More precisely, similar to Lemma 20.1, by Chernoff bound, we can average over N random triples $\mathbf{x}^{(i)}, \mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}$, and obtain the estimate

$$\rho_\beta^2 \coloneqq \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{y}_1^{(i)}, \mathbf{x}^{(i)}\right)\chi_\beta\left(\mathbf{y}_1^{(i)}\right) f\left(\mathbf{y}_2^{(i)}, \mathbf{x}^{(i)}\right)\chi_\beta\left(\mathbf{y}_2^{(i)}\right) \approx \|f_\beta\|_2^2$$

such that

$$\mathbf{Pr}\left[\left|\rho_\beta^2 - \|f_\beta\|_2^2\right| \geqslant \lambda\right] \leqslant 2e^{-\lambda^2 N/2}.$$

□

By using the estimates $\rho_\beta \approx \|f_\beta\|_2$ from Claim 20.5 in Algorithm 1, we obtain the following claim.

Claim 20.6. *Let $\tau > 0$ and $\delta > 0$ be parameters. There is a procedure that after querying the value of $f(x)$ for $O\left(\tau^{-6}\log(n)\log(1/\delta)\right)$ points, with probability $1-\delta$ it returns a set $S \subseteq \mathbb{Z}_2^n$ of size at most $|S| \leqslant \frac{16}{\tau^2}$ satisfying*

$$\left\{a \in \mathbb{Z}_2^n \;:\; |\widehat{f}(a)| \geqslant \tau\right\} \subseteq S.$$

Proof. By taking $\lambda = \frac{\tau^2}{16}$, for each β, we can produce an estimate $\rho_\beta \approx \|f_\beta\|_2$ such that with probability at least $1 - 2e^{-\lambda^2 N/2}$,

$$|\rho_\beta - \|f_\beta\|_2| \leqslant \sqrt{|\rho_\beta^2 - \|f_\beta\|_2^2|} \leqslant \sqrt{\lambda} = \frac{\tau}{4}. \tag{20.1}$$

We run the procedure of Algorithm 1 with the threshold parameter $\tau/2$ but using our estimates ρ_β instead of the actual values $\|f_\beta\|_2$. If all our estimates satisfy the accuracy of Equation (20.1), then the output S satisfies

$$\left\{a \;:\; |\widehat{f}(a)| - \frac{\tau}{4} \geqslant \tau/2\right\} \subseteq S \subseteq \left\{a \;:\; |\widehat{f}(a)| + \frac{\tau}{4} \geqslant \tau/2\right\}.$$

In particular, S would include all a with $|\widehat{f}(a)| \geqslant \tau$, and moreover every $a \in S$, it would satisfy $|\widehat{f}(a)| \geqslant \frac{\tau}{4}$ and therefore, $|S| \leqslant \frac{16}{\tau^2}$.

Since the algorithm estimates $\|f_\beta\|_2$ for at most $O(n/\tau^2)$ strings β, the probability all these estimates satisfy Eq. (20.1) is at least $1 - O(n\tau^{-2}e^{-\lambda^2 N/2}) = 1 - O(n\tau^{-2}e^{-\frac{\tau^4 N}{2^{10}}}) \geqslant 1 - \delta$ for $N = O\left(\tau^{-6}\log(n)\log(1/\delta)\right)$. □

Finally, we are ready to state the main theorem.

Theorem 20.7. *Suppose that for every function $f : \mathbb{Z}_2^n \to \{0,1\}$ in a concept class $\mathcal{C}$, there exists $\mathcal{S}_f \subseteq \mathbb{Z}_2^n$ of size m such that*

$$\sum_{a \notin \mathcal{S}_f} |\widehat{f}(a)|^2 \leqslant \varepsilon.$$

There is an algorithm that queries the value of f on $\mathrm{Poly}(m, \log(1/\delta), \varepsilon^{-1})$ points and produces a function $h : \mathbb{Z}_2^n \to \{0,1\}$ such that $\mathbf{Pr}[f(\mathbf{x}) - h(\mathbf{x})] \leqslant 12\varepsilon$ with probability at least $1 - \delta$.

Proof. Set $\tau := \sqrt{\varepsilon/m}$ and run the procedure Claim 20.6 to produce a set $\mathcal{S} \subseteq \mathbb{Z}_2^n$ of size at most $K := \frac{16}{\tau^2} = O(m/\varepsilon)$ such that with probability $1 - \frac{\delta}{2}$,

$$\left\{a \in \mathbb{Z}_2^n \ : \ |\widehat{f}(a)| \geqslant \tau\right\} \subseteq \mathcal{S}.$$

The number of queries made so far is $O\left(\tau^{-6}\log(n)\log(1/\delta)\right)$. Let $\mathcal{E}_1$ be the event that this step is successful.

As in the proof of Theorem 20.2, we can use an extra $O\left(\varepsilon^{-1}K\log(K/\delta)\right)$ many samples, to obtain an estimate $\widetilde{\boldsymbol{f}}(\boldsymbol{a}) \approx \widehat{f}(a)$ for each $a \in \mathcal{S}$. By Lemma 20.1, for every $a \in \mathcal{S}$, we have

$$\mathbf{Pr}\left[|\widehat{f}(a) - \widetilde{\boldsymbol{f}}(\boldsymbol{a})| > \frac{\sqrt{\varepsilon}}{\sqrt{K}}\right] \leqslant \frac{\delta}{2K},$$

and hence, by the union bound,

$$\mathbf{Pr}\left[\forall a \in \mathcal{S},\, |\widehat{f}(a) - \widetilde{\boldsymbol{f}}(\boldsymbol{a})| \leqslant \frac{\sqrt{\varepsilon}}{\sqrt{K}}\right] \geqslant 1 - \frac{\delta}{2}.$$

Let $\mathcal{E}_2$ denote the event that this step is successful.

Let $g = \sum_{a \in \mathcal{S}} \widehat{f}(a)\chi_a$. The probability that both $\mathcal{E}_1$ and $\mathcal{E}_2$ occur is at least $1 - \delta$, and in that case,

$$\begin{aligned}\|f - g\|_2^2 &= \sum_{a \in \mathcal{S}} |\widehat{f}(a) - \widetilde{f}(a)|^2 + \sum_{a \notin \mathcal{S}_f} |\widehat{f}(a)|^2 + \sum_{a \in \mathcal{S}_f \setminus \mathcal{S}} |\widehat{f}(a)|^2 \\ &\leqslant K\left(\frac{\sqrt{\varepsilon}}{\sqrt{K}}\right)^2 + \varepsilon + \tau^2|\mathcal{S}_f| \leqslant 3\varepsilon.\end{aligned}$$

Finally, define $h : \mathbb{Z}_2^n \to \{0,1\}$ as

$$h(x) := \begin{cases} 1 & g(x) \geqslant \frac{1}{2} \\ 0 & g(x) < \frac{1}{2} \end{cases}.$$

We have $\mathbf{Pr}[f(\mathbf{x}) \neq h(\mathbf{x})] \leqslant 4\|f - g\|_2^2 \leqslant 12\varepsilon$. □

Example 20.8 (Functions with small Fourier spectral norm)**.** Let $\mathcal{C}$ be the set of all functions $f : \mathbb{Z}_2^n \to \{0,1\}$ whose Fourier spectral norm $\|\widehat{f}\|_1$ satisfies

$$\|\widehat{f}\|_1 := \sum_a |\widehat{f}(a)| \leqslant M.$$

If we let $\mathcal{S}_f := \left\{a \ : \ |\widehat{f}(a)| \leqslant \frac{\varepsilon}{M}\right\}$, then

$$\sum_{a \notin \mathcal{S}_f} |\widehat{f}(a)|^2 \leqslant \frac{\varepsilon}{M}\|f\|_A \leqslant \varepsilon.$$

Moreover, since

$$1 \geqslant \sum_{a \in \mathcal{S}_f} |\widehat{f}(a)|^2 \geqslant |\mathcal{S}_f|\left(\frac{\varepsilon}{M}\right)^2,$$

we have $|\mathcal{S}_f| \leqslant \frac{M^2}{\varepsilon^2}$. Therefore, $\mathcal{C}$ satisfies the assumption of Theorem 20.7 with $m = \frac{M^2}{\varepsilon^2}$.

Chapter 21

Bounded depth circuits

In 1949, Shannon [Sha49] proposed using the size of Boolean circuits to measure a function's computational difficulty. Circuits are closely related in computational power to Turing machines, and thus, they provide a nice framework for understanding time complexity. On the other hand, their especially simple definition makes them amenable to various combinatorial, algebraic, and analytic methods.

A *Boolean circuit* is a directed acyclic graph. The vertices of in-degree 0 are called *inputs*. Each input is labelled with a variable x_i or a constant 0 or 1. The vertices of in-degree $k > 0$ are called *gates*, and each such gate is labelled with a k-ary Boolean function. In the context of circuits, the in-degrees and out-degrees of vertices, respectively, are called their *fan-in*s and *fan-out*s. One of the circuit nodes is designated the *output* node, and with this, the circuit represents a Boolean function naturally. Sometimes, we allow multiple output nodes to represent functions $f : \{0,1\}^n \to \{0,1\}^m$. The size of a circuit is the number of its gates[1].

Example 21.1. Figure 21.1 illustrates a simple circuit with 3 inputs and six gates. It computes a function $f : \{0,1\}^3 \to \{0,1\}$. For example, as illustrated in the picture, $f(0,1,0) = 1$.

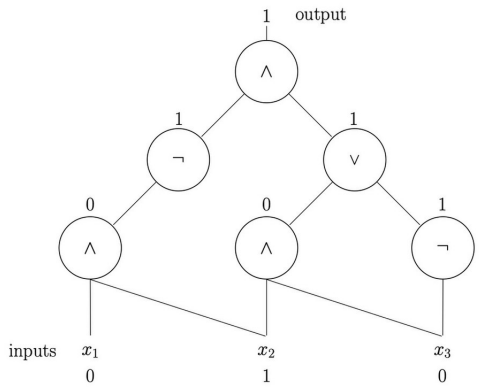

Figure 21.1: A circuit with 3 inputs and six gates.

As a more general example, recall that a formula is in disjunctive normal form (abbreviated to DNF) if it is a disjunction (i.e. $\vee$) of clauses, where a clause is a conjunction (i.e. $\wedge$) of literals (i.e. x_i or $\neg x_i$). A k-DNF is a DNF where each clause consists of at most k literals.

Therefore, DNFs are circuits with gates $\{\neg, \vee, \wedge\}$ of arbitrary fan-in. Every Boolean function $f : \{0,1\}^n \to \{0,1\}$ can be expressed as an n-DNF:

$$f(x) = \bigvee_{y : f(y)=1} C_y(x), \tag{21.1}$$

[1] In some texts, the input nodes are counted towards the circuit's size.

where the clause C_y corresponding to y is

$$C_y(x) := \left(\bigwedge_{i:y_i=1} x_i\right) \wedge \left(\bigwedge_{i:y_i=0} \neg x_i\right).$$

Note that $C_y(x) = 1$ if and only if $x = y$.

We can break the $\vee$ and $\wedge$ gates in a DNF further to use only binary $\vee$ and $\wedge$'s, and therefore, every Boolean function has a circuit with gates $\{\neg, \vee, \wedge\}$ of fan-in at most 2.

Definition 21.2. The *circuit complexity of a function* f is the size of the smallest circuit of fan-in 2 that computes f.

A simple counting argument shows that most functions require exponential circuits of fan-in 2. Roughly speaking, there are 2^{2^n} Boolean functions $f : \{0,1\}^n \to \{0,1\}$, while the number of small circuits is much smaller.

Theorem 21.3 (Shannon [Sha49])**.** *Almost every Boolean function* $f : \{0,1\}^n \to \{0,1\}$ *requires fan-in* 2 *circuits of size* $\Omega(2^n/n)$.

Proof. There are exactly 2^{2^n} Boolean functions $f : \{0,1\}^n \to \{0,1\}$. The number of circuits with t gates can be upper-bounded as follows: Since the number of fan-in 2 gates is $2^{2^2} = 16$, there are 16^t choices for assigning gates to nodes. There are $(n+2+t)^2$ choices for the two incoming wires of a gate: The n input variables, the two constant inputs 0 and 1, or the t other nodes. Finally, we must designate one of the t gates as the output gate. Hence, the number of circuits of size t with fan-in 2 is at most

$$16^t(t+n+2)^{2t}t.$$

If $t = 2^n/20n$, then

$$\lim_{n\to\infty} \frac{16^t(t+n+2)^{2t}t}{2^{2^n}} = 0.$$

Thus, almost every function has a circuit complexity larger than $2^n/20n$. □

On the other hand, we know from the DNF representation that every function $f : \{0,1\}^n \to \{0,1\}$ can be computed by a fan-in 2 circuit of size $O(n2^n)$. In fact, with some extra work (proved by Lupanov [Lup58]), one can improve this bound to $O(2^n/n)$, which matches the lower bound of Theorem 21.3.

Theorem 21.3 has a major shortcoming. It does not provide any *explicit* examples of functions that require large circuits. Also, unfortunately, it does not prove the existence of functions in NP that require circuits of super-polynomial size. Note that any function on n bits that depends on all its inputs requires fan-in 2 circuits of size at least $n-1$ just to read the inputs. Despite the incredible research on circuit complexity lower bounds, the strongest known bounds for explicit functions are extremely weak. In 1984, Blum gave an example of a function that requires fan-in 2 circuits of size $3n - o(n)$. Recently, Blum's lower bound has been improved to $(3 + \frac{1}{86})n - o(n)$ by [FGHK16].

The main open problem of circuit complexity is beating this linear lower bound for natural problems (say, in NP).

Problem 21.4. *Find an explicit function* $f : \{0,1\}^n \to \{0,1\}$ *with circuit complexity* $\omega(n)$.

21.1 Bounded depth alternating circuits

Considering our inability to prove lower bounds on the circuit complexity of explicit Boolean functions, we need to impose substantial restrictions on the circuits to be able to prove meaningful lower bounds. We will start by restricting to bounded depth circuits. The *depth* of a circuit is the longest distance from the input nodes to the output node.

While the size of a circuit essentially measures the time required to compute a function using a single simple processor, the depth of a polynomial-size circuit corresponds to the amount of time it takes a parallel algorithm to compute it.

Let us start by defining our constant depth circuits. We will be interested in the model where we are restricted to gates $\wedge$, $\vee$, $\neg$. Note that by De Morgan's laws

$$\neg(p_1 \vee \ldots \vee p_k) = (\neg p_1) \wedge \ldots \wedge (\neg p_k),$$

and

$$\neg(p_1 \wedge \ldots \wedge p_k) = (\neg p_1) \vee \ldots \vee (\neg p_k),$$

we can assume that

- There are no $\neg$ gates in the circuit, and instead, the inputs are either of the form x_i or $\neg x_i$ for variables x_i, or constants 0 and 1.
- We shall consider circuits whose depths are much smaller than n, the number of inputs. Hence, we need to allow arbitrary fan-in so the circuit can access the entire input.
- We will assume that the circuits are of the special form where all $\wedge$ and $\vee$ gates are organized into alternating levels with edges only between adjacent levels. Any circuit can be converted into this form without increasing the depth and by, at most, squaring the size.

These circuits are called *alternating circuits*. The *depth* of an alternating circuit is defined as the distance from the output node to the input nodes.

The alternating circuits of depth 2 are particularly important. Note that because of the "alternation" condition, there are two different types of depth 2 alternating circuits. They correspond to *conjunctive normal form* and *disjunctive normal form* formulas.

We have already discussed the DNFs. Similarly, a formula is in conjunctive normal form (abbreviated to CNF) if it is a conjunction (i.e. $\wedge$) of clauses, where a clause is a disjunction (i.e. $\vee$) of literals (i.e. x_i or $\neg x_i$). A k-CNF is a CNF where each clause consists of at most k literals.

For example, $(x_1 \vee x_2) \wedge (\neg x_1 \vee x_2 \vee x_3)$ is a formula in conjunctive normal form. By changing the roles of 0 and 1's in Equation (21.1), we can write an n-CNF representation for every $f : \{0,1\}^n \to \{0,1\}$.

$$f(x) = \bigwedge_{y : f(y)=0} C_y(x), \tag{21.2}$$

where the $\vee$-clause C_y corresponding to y is

$$C_y(x) := \left(\bigvee_{i : y_i = 0} x_i \right) \vee \left(\bigvee_{i : y_i = 1} \neg x_i \right).$$

Note that $C_y(x) = 0$ if and only if $x = y$.

We record these observations for future reference.

Observation 21.5. *Every function $f : \{0,1\}^n \to \{0,1\}$ has an n-DNF and an n-CNF representation, each with at most 2^n clauses.*

21.2 Håstad's Switching lemma

The first strong lower bounds for bounded depth circuits were given by Ajtai [Ajt83] in 1983 and Furst, Saxe, Sipser [FSS84] in 1984. They established a superpolynomial lower bound for constant depth circuits computing the parity function. Later, Yao [Yao85] gave a sharper exponential lower bound. In 1986, Håstad [Has86a] further strengthened and simplified this argument and obtained near-optimal bounds.

The basic idea of Ajtai [Ajt83] and Furst, Saxe, Sipser [FSS84] for proving lower-bounds on bounded depth AC circuits was to assign random values to a random subset of variables. This will simplify a small size AC[d] circuit greatly. Consider a gate at level 1 (that is, a gate directly connected to inputs x_i and $\neg x_i$'s). Noting that the gate is either $\wedge$ or $\vee$, if it has a large fan-in, there is a high chance that a random assignment of values to a random subset of variables will determine the value of the gate. Indeed, an $\wedge$ gate only needs one 0 input to be set to 0, and an $\vee$ gate only needs one 1 on its inputs to be set to 1.

Definition 21.6 (restrictions)**.** Let $X = \{x_1, \ldots, x_n\}$ be the input variables to a circuit C computing a function f. A *restriction* ρ is a function $\rho : X \to \{0, 1, \star\}$.

A restriction ρ sets the values of the variables assigned 0 or 1 and leaves those assigned stars alive. Under ρ, we may simplify C by eliminating gates whose values become determined. Call this the *induced circuit* C_ρ computing the *induced function* f_ρ.

As mentioned earlier, Håstad further explored these ideas. The core of his proof is an important lemma known as the switching lemma, a key tool for proving lower bounds on the size of the constant-depth Boolean circuits. It states that random restrictions with a few stars significantly decrease the decision tree complexity of small alternating circuits.

Lemma 21.7 (Håstad's switching lemma)**.** *Let f be given by a t-CNF formula. Choose a random restriction ρ by setting every variable independently to $\star$ with probability p, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then for every $s \in \mathbb{N}$,*

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s] \leqslant (5pt)^s.$$

In particular for $p = \frac{1}{10t}$,

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s] \leqslant 2^{-s}.$$

Remark 21.8. Note that the bound in the switching lemma does not depend on the number of clauses in the CNF. The only parameter about the CNF that appears in the assertion is its width t.

We will prove the switching lemma by induction on the number m of clauses; however, since the bound does not depend on m, we cannot afford to lose anything in the induction step: Starting with the bound $(5pt)^s$ for t-CNF's with $m-1$ clauses, we must conclude the same bound for t-CNFs with m clauses. The general proof strategy is simple. Consider the first clause, and without loss of generality, assume that this clause is $(x_1 \vee \ldots \vee x_t)$.

Case 1: If the random restriction assigns any 1's to this clause, then this clause evaluates to 1, and we can remove it and apply the induction hypothesis to the remaining $m-1$ clauses.

Case 2: If the random restriction assigns 0's to all of $x_1, \ldots, x_t$, then the clause evaluates to 0, and as a result $f_\rho \equiv 0$, which satisfies $\mathrm{dt}(f_\rho) = 0$.

Case 3: The remaining case is when ρ assigns some $\star$'s (and no 1's) to $x_1, \ldots, x_t$. Let T be the subset of the variables in this clause that receive $\star$'s. In this case, it suffices to find a decision tree of depth $s - |T|$ for the remaining $m-1$ clauses, as we can extend such a decision tree to a decision tree of depth s by always querying the values of the variables in T. The induction hypothesis tells us that the probability that the remaining clauses do not have such a decision tree is at most $(5pt)^{s-|T|}$. This bound is worse than our goal $(5pt)^s$, but fortunately, $\star$'s are generally unlikely, and the probability that all the variables in T receive $\star$'s is at most $p^{|T|}$. Putting these together and taking a union bound over all possibilities of T gives us the desired bout $(5pt)^{|T|}$.

We are going to prove the switching lemma by induction. In the sketched proof above, we assumed that what happens in the rest of the $m-1$ clauses is independent of the variables $x_1, \ldots, x_t$. However, this is not the case, for example, in Case 1. To deal with this technical issue, we need to strengthen the statement of the lemma.

Lemma 21.9 (Håstad's switching lemma, stronger version)**.** *Let f be given by a t-CNF formula. Choose a random restriction ρ by setting every variable independently to $\star$ with probability p, and to 0 and 1 each with probability $\frac{1-p}{2}$. For every $s \in \mathbb{N}$, and every function $F : \{0,1\}^n \to \{0,1\}$, we have*

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s | F_\rho \equiv 1] \leqslant (5pt)^s, \tag{21.3}$$

where $F_\rho \equiv 1$ is the event when F_ρ is the constant 1 function.

Proof. Set $\alpha := 5pt$, and suppose that $f = \wedge_{i=1}^m C_i$ where C_i's are clauses of size at most t. We prove this statement by induction on m, the number of clauses in f. If $m = 0$, then $f \equiv 1$, and the lemma is obvious. For the induction step, let us study what happens to C_1, the first clause in the circuit. First note that by possibly changing the role of 0's and 1's for some variables, we can assume without loss of generality that there are no negated literals in C_1 and hence

$$C_1 = \bigvee_{i \in T} x_i,$$

for a subset $T \subseteq \{1,\ldots,n\}$, $|T| \leqslant t$. First, we split the left-hand side of Eq. (21.3) into two terms based on whether C_1 receives a 1 from the restriction:

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s | F_\rho \equiv 1] = \mathbf{Pr}[\mathrm{dt}(f_\rho) > s, \rho_T \notin \{0,\star\}^T | F_\rho \equiv 1] + \mathbf{Pr}[\mathrm{dt}(f_\rho) > s,\ \rho_T \in \{0,\star\}^T | F_\rho \equiv 1].$$

Hence in order to prove (21.3), it suffices to show both

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s | F_\rho \equiv 1,\ \rho_T \notin \{0,\star\}^T] \leqslant \alpha^s, \tag{21.4}$$

and

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s | F_\rho \equiv 1,\ \rho_T \in \{0,\star\}^T] \leqslant \alpha^s, \tag{21.5}$$

as then we would have

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s | F_\rho \equiv 1] \leqslant \mathbf{Pr}[\rho_T \notin \{0,\star\}^T | F_\rho \equiv 1]\alpha^s + \mathbf{Pr}[\rho_T \in \{0,\star\}^T | F_\rho \equiv 1]\alpha^s = \alpha^s.$$

To prove (21.4), note that for $g = \wedge_{i=2}^m C_i$ (which has only $m-1$ clauses),

$$\text{L.H.S of (21.4)} \quad = \quad \mathbf{Pr}[\mathrm{dt}(g_\rho) > s | F_\rho \equiv 1,\ \rho_T \notin \{0,\star\}^T] = \mathbf{Pr}[\mathrm{dt}(g_\rho) > s \mid (F \wedge C_1)_\rho \equiv 1] \leqslant \alpha^s,$$

where in the last inequality, we used the induction hypothesis applied to g and $F \wedge C_1$. It remains to prove (21.5). We break (21.5) into $2^{|T|}$ terms based on which coordinates in T are $\star$'s and which ones are 0's:

$$\begin{aligned}
\text{L.H.S of (21.5)} \quad &= \quad \sum_{Y \subseteq T} \mathbf{Pr}[\mathrm{dt}(f_\rho) > s,\ \rho_Y = \vec{\star},\ \rho_{T-Y} = \vec{0} \mid F_\rho \equiv 1,\ \rho_T \in \{0,\star\}^T] \\
&\leqslant \quad \sum_{Y \subseteq T} \mathbf{Pr}[\rho_Y = \vec{\star}, \rho_{T-Y} = \vec{0} \mid F_\rho \equiv 1,\ \rho_T \in \{0,\star\}^T] \times \\
&\qquad\qquad \mathbf{Pr}[\mathrm{dt}(f_\rho) > s \mid F_\rho \equiv 1,\ \rho_Y = \vec{\star},\ \rho_{T-Y} = \vec{0},\ \rho_T \in \{0,\star\}^T] \\
&\leqslant \quad \sum_{Y \subseteq T} \mathbf{Pr}\left[\rho_Y = \vec{\star} \middle| F_\rho \equiv 1, \rho_T \in \{0,\star\}^T\right] \times \mathbf{Pr}\left[\mathrm{dt}(f_\rho) > s \middle| F_\rho \equiv 1, \rho_Y = \vec{\star}, \rho_{T-Y} = \vec{0}\right].
\end{aligned}$$

First note that if $Y = \emptyset$, then $\rho_T = \vec{0}$, and thus C_1 is not satisfied and $f_\rho \equiv 0$, and consequently $\mathrm{dt}(f_\rho) = 0$. Hence, we can remove the corresponding term from the above calculation and obtain the following:

$$\text{L.H.S of (21.5)} \leqslant \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} \mathbf{Pr}\left[\rho_Y = \vec{\star} \middle| F_\rho \equiv 1, \rho_T \in \{0,\star\}^T\right] \times \mathbf{Pr}\left[\mathrm{dt}(f_\rho) > s \middle| F_\rho \equiv 1, \rho_Y = \vec{\star}, \rho_{T-Y} = \vec{0}\right]. \tag{21.6}$$

We bound the two terms in the product separately.

First observation (bounding $\mathbf{Pr}\left[\rho_Y = \vec{\star} \middle| F_\rho \equiv 1, \rho_T \in \{0,\star\}^T\right]$): Since setting variables in Y to $\star$ cannot increase the probability that $F_\rho \equiv 1$, we have

$$\mathbf{Pr}[F_\rho \equiv 1 \mid \rho_Y = \vec{\star},\ \rho_T \in \{0,\star\}^T] \leqslant \mathbf{Pr}[F_\rho \equiv 1 \mid \rho_T \in \{0,\star\}^T],$$

Hence using $\mathbf{Pr}[A|B]\,\mathbf{Pr}[B] = \mathbf{Pr}[A \wedge B]$ we have

$$\begin{aligned}
\mathbf{Pr}[\rho_Y = \vec{\star} \mid F_\rho \equiv 1,\ \rho_T \in \{0,\star\}^T] \quad &= \quad \frac{\mathbf{Pr}[F_\rho \equiv 1 \mid \rho_Y = \vec{\star},\ \rho_T \in \{0,\star\}^T]}{\mathbf{Pr}[F_\rho \equiv 1 \mid \rho_T \in \{0,\star\}^T]} \mathbf{Pr}[\rho_Y = \vec{\star} \mid \rho_T \in \{0,\star\}^T] \\
&\leqslant \quad \mathbf{Pr}[\rho_Y = \vec{\star} \mid \rho_T \in \{0,\star\}^T] = \left(\frac{2p}{1+p}\right)^{|Y|} \leqslant (2p)^{|Y|}
\end{aligned}$$

Second observation: (bounding $\mathbf{Pr}\left[\mathrm{dt}(f_\rho) > s \middle| F_\rho \equiv 1, \rho_Y = \vec{\star}, \rho_{T-Y} = \vec{0}\right]$): Note that the variables in Y can contribute by at most $|Y|$ to the decision tree depth, or more precisely if for every $\sigma \in \{0,1\}^{|Y|}$, we have $\mathrm{dt}(f_{\sigma\rho}) \leqslant s - |Y|$, then $\mathrm{dt}(f_\rho) \leqslant s$. Indeed to verify this, note that we can always build a decision tree of depth at most $|Y| + \max_\sigma \mathrm{dt}(f_{\sigma\rho})$ as follows: In the first $|Y|$ levels, we query all the variables x_i for $i \in Y$ to obtain a $\sigma \in \{0,1\}^Y$.

Then we follow a decision tree of depth $\mathrm{dt}(f_{\sigma\rho})$ afterwards. Hence for $Y \neq \emptyset$, recalling that $g = \wedge_{i=2}^{m} C_i$, we have

$$\begin{aligned}
\mathbf{Pr}[\mathrm{dt}(f_\rho) > s \mid F_\rho \equiv 1, \rho_Y = \vec{\star}, \rho_{T-Y} = \vec{0}] &\leqslant \mathbf{Pr}\left[\exists \sigma \in \{0,1\}^{|Y|},\ \mathrm{dt}(f_{\sigma\rho}) > s - |Y| \;\middle|\; F_\rho \equiv 1, \rho_{T-Y} = \vec{0}\right] \\
&\leqslant \sum_{\sigma \in \{0,1\}^{|Y|}} \mathbf{Pr}[\mathrm{dt}(f_{\sigma\rho}) > s - |Y| \mid F_\rho \equiv 1, \rho_{T-Y} = \vec{0}] \\
&= \sum_{\sigma \in \{0,1\}^{|Y|}} \mathbf{Pr}[\mathrm{dt}(f_{\sigma\rho}) > s - |Y| \mid (F \wedge \wedge_{i \in T \setminus Y} \overline{x_i})_\rho \equiv 1] \\
&= \sum_{\sigma \in \{0,1\}^{|Y|}} \mathbf{Pr}[\mathrm{dt}(g_{\sigma\rho}) > s - |Y| \mid (F \wedge \wedge_{i \in T \setminus Y} \overline{x_i})_\rho \equiv 1] \\
&\leqslant \sum_{\sigma \in \{0,1\}^{|Y|}} \alpha^{s-|Y|} \leqslant 2^{|Y|} \alpha^{s-|Y|}.
\end{aligned}$$

where we applied the union bound and then the induction hypothesis.

Combining the two observations with (21.6), we finish the proof:

$$\begin{aligned}
\text{L.H.S of (21.5)} &\leqslant \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} 2^{|Y|} \alpha^{s-|Y|} (2p)^{|Y|} = \alpha^s \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} \left(\frac{4p}{\alpha}\right)^{|Y|} = \alpha^s \left(\left(1 + \frac{4p}{\alpha}\right)^{|T|} - 1\right) \\
&= \alpha^s \left(\left(1 + \frac{4}{5t}\right)^{t} - 1\right) \leqslant \alpha^s (e^{\frac{4}{5}} - 1) \leqslant \alpha^s.
\end{aligned}$$

□

Remark 21.10. Since the negation of a CNF is a DNF of similar size and vice versa, the switching lemma can be used to convert a t-DNF formula to an s-CNF in the same way as Lemma 21.9.

Corollary 21.11. *Let f be a Boolean function computed by an* AC *circuit of size M and depth d. Choose a random restriction ρ by setting every variable independently to $\star$ with probability $p = \frac{1}{10^d s^{d-1}}$, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\mathbf{Pr}[\mathrm{dt}(f_\rho) > s] \leqslant M 2^{-s}.$$

Proof. We sample the restriction ρ by first sampling a random restriction ρ_0 with $\mathbf{Pr}[\star] = 1/10$, and then sampling $d-1$ consecutive restrictions $\rho_1, \ldots, \rho_{d-1}$ each with $\mathbf{Pr}[\star] = \frac{1}{10s}$.

Assume without loss of generality that the bottom gates are $\vee$. We claim that, with high probability, after the restriction ρ_0, all the remaining bottom fan-ins are at most s. To see this, consider two cases for each gate at the bottom level of the original circuit:

1. The original fan-in is at least $2s$. In this case, the probability that the gate was not eliminated by ρ_0, that is, no input to this gate got assigned a 1 is at most $(0.55)^{2s} < 2^{-s}$.

2. The original fan-in is at most $2s$. In this case, the probability that at least s inputs got assigned a $\star$ by ρ_0 is at most $\binom{2s}{s}(1/10)^s \leqslant 2^{-s}$.

Thus, the probability of failure after the first restriction is at most $m_1 2^{-s}$, where m_1 is the number of gates at the bottom level.

We now apply the next $d-2$ restrictions, each with $\mathbf{Pr}[\star] = \frac{1}{10s}$. After each of these, we use Håstad's switching lemma (see Remark 21.10) to convert the lower two levels from CNF to DNF (or vice versa), collapse the second and third levels (from the bottom) to one level, reducing the depth by one. For each gate of distance two from the inputs, the probability that it corresponds to a function g with $\mathrm{dt}(g_{\rho_i}) > s$, is hence bounded by $(5\frac{1}{10s}s)^s \leqslant 2^{-s}$. The probability that a particular gate fails to satisfy the desired property is no more than 2^{-s}. Since the top gate is $\wedge$, after these $d-2$ stages, we are left with a CNF formula of bottom fan-in at most s. We now apply the last restriction, and by the switching lemma, we get a function f_ρ with $\mathrm{dt}(f_\rho) \leqslant s$. The probability of failure at this stage is at most 2^{-s}. To compute the total probability of failure, we observe that each gate of the original circuit contributes 2^{-s} to the probability of failure, and hence applying the union bound yields the desired bound. □

Since a restriction ρ of the party function PARITY with m starts is either a copy of PARITY_m or $1 - \text{PARITY}_m$, we have $\text{dt}(\text{PARITY}_\rho) \geqslant m$. By combining this fact with Corollary 21.11, we obtain a strong lower bound on the size of any depth-d AC circuit computing PARITY.

Theorem 21.12 ([Has86b])**.** *Any depth-d AC circuit that computes* PARITY *is of size $2^{\Omega(n^{1/d})}$.*

If in the proof of Corollary 21.11, we stop before applying the last restriction ρ_{d-1}, we can obtain the following statement, which uses a larger value for p.

Corollary 21.13. *Let f be a Boolean function computed by an AC circuit of size M and depth $d \geqslant 2$ whose output gate is $\wedge$. Choose a random restriction ρ by setting every variable independently to $\star$ with probability $p = \frac{1}{10^{d-1}s^{d-2}}$, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\mathbf{Pr}[f_\rho \text{ does not have a CNF with fan-in} \leqslant s] \leqslant M2^{-s}.$$

Similarly, if the output gate of the original circuit is $\vee$, then the probability that f_ρ does not have a DNF with fan-in $\leqslant s$ is bounded by $M2^{-s}$.

In the next section, we will show that this improvement implies a better lower bound of $2^{\Omega(n^{1/(d-1)})}$ for PARITY, as well as a lower bound for the Majority function.

21.3 Influences in bounded depth circuits

Our next goal is to show that the total influence of low depth small AC circuit cannot be large. First, we consider the CNF and the DNF circuits with small clauses.

Lemma 21.14. *Let f be a CNF or a DNF formula where all the clauses are of size at most t. Then $I_f \leqslant t$.*

Proof. We prove the lemma for the DNF case, and the CNF case follows by replacing f with $1 - f$. We prove the lemma for the DNF case, and the CNF case follows by replacing f with $1 - f$. For every $x \in \{0,1\}^n$, let $s_{1\to 0}(x)$ denote the number of coordinates $i \in [n]$ such that $f(x) = 1$ and $f(x \oplus e_i) = 0$. If $f(x) = 1$, then x satisfies at least one clause C. If $f(x \oplus e_i) = 0$, then C must involve x_i or $\neg x_i$, and since there are at most t literals in C, we have $s_{1\to 0}(x) \leqslant t$ for every x. Therefore,

$$I_f = \sum_{i=1}^{n} \frac{1}{4} \mathbf{Pr}[f(\mathbf{x}) \neq f(\mathbf{x} \oplus e_i)] \leqslant \sum_{i=1}^{n} \mathbf{Pr}[f(\mathbf{x}) = 1 \wedge f(\mathbf{x} \oplus e_i) = 0] = \mathbb{E}_{1\to 0}(\mathbf{x}) \leqslant t.$$

□

Boppana [Bop97] used Häastad's switching lemma to prove that small-size low-depthAC circuits have small total influences.

Theorem 21.15 (Boppana [Bop97])**.** *Let f be a Boolean function computed by an AC circuit of depth d and size M (including the input gates), then*

$$I_f \leqslant (20 \log M)^d.$$

Proof. Note $n \leqslant M$ since we are counting the input gates when calculating the circuit size. Applying Corollary 21.11 with $s = 2 \log M$ and $p = \frac{1}{10^d s^{d-1}}$, and combining it with the fact that $I_g \leqslant \text{dt}(g)$ for all g, shows

$$\mathbf{Pr}[I_{f_\rho} \geqslant s] \leqslant M2^{-s} \leqslant \frac{1}{M} \leqslant \frac{1}{n}.$$

Therefore,

$$\mathbb{E}_\rho[I_{f_\rho}] \leqslant \mathbf{Pr}[I_{f_\rho} > s]n + s \leqslant \frac{1}{n}n + s \leqslant s + 1 \leqslant 2s.$$

On the other hand, for every i,

$$\mathop{\mathbf{Pr}}_{\rho,x}[f_\rho(x) \neq f_\rho(x \oplus e_i)] = p \mathop{\mathbf{Pr}}_{x}[f(x) \neq f(x \oplus e_i)],$$

where p is the probability that the ith variable is not fixed by ρ. Therefore,

$$\mathbb{E}_\rho[I_{f_\rho}] = pI_f.$$

We conclude

$$I_f \leqslant \frac{2s}{p} \leqslant 2s \cdot (10s)^{d-1} \leqslant (20 \log M)^d.$$

□

One can improve the bound slightly by using Corollary 21.13 and Lemma 21.14 instead of Corollary 21.11.

Theorem 21.16 (Boppana [Bop97])**.** *Let f be a Boolean function computed by an* AC *circuit of depth d and size M, then*

$$I_f \leqslant (20 \log M)^{d-1}.$$

The majority function MAJ is defined as $\text{MAJ}(x) := 1$ if and only if $\sum x_i \geqslant n/2$. It is straightforward verify $I_{\text{MAJ}} = \Theta(\sqrt{n})$. Recall also that the total influence of PARITY is $\frac{n}{4}$. We conclude the following lower bounds on the AC circuit size of MAJ and PARITY.

Corollary 21.17. *Any depth-d* AC *circuit that computes* PARITY *is of size $2^{\Omega(n^{1/(d-1)})}$. Any depth-d* AC *circuit that computes* MAJ *is of size $2^{\Omega(n^{1/(2d-2)})}$.*

To this day, Håstad's bound for parity remains the strongest explicit known lower bound against small-depth circuits for any function, even for $d = 3$. The special case of depth-3 has received significant attention as one of the simplest restricted models where our understanding is lacking. The following open problem is one of the frontiers of circuit complexity.

Problem 21.18. *Find an explicit function $f : \{0,1\}^n \to \{0,1\}$ that requires circuit size $2^{\omega(\sqrt{n})}$ for* AC *circuits of depth* 3.

Remark 21.19. It would be interesting to prove an inverse for Boppana's Theorem 21.15. In [BKS99], Benjamini, Schramm, and Kalai conjectured a very strong inverse statement that every monotone function f can be approximated by a circuit of size $e^{O(I_f^{1/d-1})}$ for some positive integer d. However, this was disproved by O'Donnell and Wimmer [OW07] using an example consisting of $\vee$ of a DNF and a CNF (hence a depth 3-circuit) with total influence $O(\log n)$.

Chapter 22

LMN and Razborov-Smolensky

In this chapter, we will discuss two fundamental results from circuit complexity about approximating small-size, low-depth alternating circuits with polynomials. The first theorem, due to Linial, Mansour, and Nisan [LMN93], provides a strong approximation with a low-degree *real-valued* function g, where the approximation quality is measured in the L_2 norm. In the second result, proved independently by [Raz87] and [Smo87], the quality of the approximation is measured using the distance $\mathbf{Pr}[f(\mathbf{x}) \neq g(\mathbf{x})]$.

Both theorems have numerous applications in complexity theory and the theory of pseudo-random generators.

22.1 LMN: Fourier tail of low-depth circuits

By Theorem 21.16, if f is computable by an AC circuit of depth d and size M, then we have the following upper bound on the Fourier tail of f:

$$\left\|f - f^{<t}\right\|_2^2 = \left\|f^{\geqslant t}\right\|_2^2 \leqslant \frac{I_f}{t} \leqslant \frac{(20 \log M)^{d-1}}{t} \quad \text{for all } t \in [n].$$

In this section, we prove a stronger upper bound on this quantity. The switching lemma shows that under random restriction, a function f computable by a low-depth small-size AC circuit is likely to simplify to a function with small decision tree complexity. Since small height decision trees are of low degree, this observation suggests that such an f must not have a large mass on higher levels. Linial, Mansour, and Nisan [LMN93] turned this intuition into a theorem, which we will discuss below.

First, note that since the degree of a decision tree is bounded by its depth, we have the following corollary to the switching lemma.

Corollary 22.1. *Let f be a Boolean function computed by an AC circuit of size M and depth d. Choose a random restriction ρ by setting every variable independently to $\star$ with probability $p = \frac{1}{10^d s^{d-1}}$, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\mathbf{Pr}[\deg(f_\rho) > s] \leqslant M 2^{-s}.$$

The following result, sometimes called *the LMN theorem*, is the main result of this section.

Theorem 22.2 (Linial, Mansour, Nisan [LMN93])**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function computed by an AC circuit of depth d and size M, and let t be any integer. Then*

$$\|f^{>t}\|_2^2 \leqslant 2M 2^{-t^{1/d}/20}.$$

Proof. Consider a random restriction $\rho \in \{-1, 1, \star\}^n$ with $\mathbf{Pr}[\star] = p \leqslant \frac{1}{10^d k^{d-1}}$ for a value of k to be determined later. We sample ρ in two steps. First, we pick $T \subseteq [n]$ corresponding to the positions not assigned a $\star$. Then we pick $x_T \in \{0,1\}^T$ uniformly at random, and ρ is defined as $\rho \coloneqq (x_T, \vec{\star})$. Set $f_{x_T} \coloneqq f_\rho = f(x_T, \cdot)$. Since for $a \in \mathbb{Z}_2^n$,

$$\chi_a(x) = \chi_{a_T}(x_T)\chi_{a_{\overline{T}}}(x_{\overline{T}}),$$

we have

$$f(x) = \sum_{a \in \mathbb{Z}_2^n} \widehat{f}(a)\chi_a(x) = \sum_{a \in \mathbb{Z}_2^n} \widehat{f}(a)\chi_{a_T}(x_T)\chi_{a_{\overline{T}}}(x_{\overline{T}}) = \sum_{\beta \in \mathbb{Z}_2^{\overline{T}}} \left(\sum_{\alpha \in \mathbb{Z}_2^T} \widehat{f}(\alpha, \beta)\chi_\alpha(x_T) \right) \chi_\beta(x_{\overline{T}}),$$

Therefore, the Fourier expansion of $f_{x_T} : \{0,1\}^{\overline{T}} \to \{0,1\}$ is $f_{x_T}(y) = \sum_{\beta \in \mathbb{Z}_2^{\overline{T}}} \widehat{f_{x_T}}(\beta)\chi_\beta(y)$ where

$$\widehat{f_{x_T}}(\beta) = \sum_{\alpha \in \mathbb{Z}_2^T} \widehat{f}(\alpha, \beta)\chi_\alpha(x_T).$$

Hence, by the Parseval identity, we have

$$\mathbb{E}_{\mathbf{x}_T} \left| \widehat{f_{\mathbf{x}_T}}(\beta) \right|^2 = \sum_{\alpha \in \mathbb{Z}_2^T} |\widehat{f}(\alpha, \beta)|^2,$$

which shows that

$$\mathbb{E}_{\mathbf{x}_T} \left\| f_{\mathbf{x}_T}^{>k} \right\|_2^2 = \mathbb{E}_{\mathbf{x}_T} \sum_{\substack{\beta \in \mathbb{Z}_2^{\overline{T}} \\ |\beta| > k}} \left\| \widehat{f_{\mathbf{x}_T}}(\beta) \right\|_2^2 = \sum_{\substack{\beta \in \mathbb{Z}_2^{\overline{T}} \\ |\beta| > k}} \sum_{\alpha \in \mathbb{Z}_2^T} |\widehat{f}(\alpha, \beta)|^2 = \sum_{S : |S \cap \overline{T}| > k} |\widehat{f}(S)|^2,$$

where on the right-hand side, we used the set notation to denote the Fourier coefficients.

Now, we use the randomness in T. Since $f_{x_T}^{>k} = 0$ if $\deg(f_\rho) \leqslant k$, and that always $\left\| f_{x_T}^{>k} \right\|_2^2 \leqslant \|f_{x_T}\|_2^2 \leqslant 1$, we have

$$\mathbb{E}_{\boldsymbol{T}} \left[\sum_{S : |S \cap \overline{\boldsymbol{T}}| > k} |\widehat{f}(S)|^2 \right] = \mathbb{E}_{\boldsymbol{\rho}} \left\| f_{\boldsymbol{\rho}}^{>k} \right\|_2^2 \leqslant \mathbf{Pr}[\deg(f_{\boldsymbol{\rho}}) > k] \leqslant M 2^{-k}, \tag{22.1}$$

where the last inequality follows from Corollary 22.1 since we have chosen $\mathbf{Pr}[\star] = p \leqslant \frac{1}{10^d k^{d-1}}$. Moreover, we can bound the left-hand side of (22.1) from below:

$$\text{L.H.S. of (22.1)} = \sum_{S \subseteq [n]} \mathop{\mathbf{Pr}}_{\boldsymbol{T}}[|S \cap \overline{\boldsymbol{T}}| > k] |\widehat{f}(S)|^2 \geqslant \sum_{|S| > t} \mathop{\mathbf{Pr}}_{\boldsymbol{T}}[|S \cap \overline{\boldsymbol{T}}| > k] |\widehat{f}(S)|^2.$$

Taking $p = \frac{1}{10t^{(d-1)/d}}$ and $k = t^{1/d}/20$, we have $p \leqslant \frac{1}{10^d k^{d-1}}$, and therefore, by the Chernoff bound, for $|S| > t$, the probability of $|S \cap \overline{\boldsymbol{T}}| > k = pt/2$ is at least $1 - 2e^{\frac{-pt}{12}} \geqslant \frac{1}{2}$. Hence, by (22.1), we have

$$\sum_{S : |S| > t} \frac{1}{2} |\widehat{f}(S)|^2 \leqslant M 2^{-t^{1/d}/20}.$$

□

Remark 22.3. Theorem 21.15 and Theorem 22.2 show that the Fourier spectrum of small low-depth AC circuits is concentrated on the lower levels. In particular, these functions satisfy the assumption of the PAC learning algorithm of Theorem 20.2.

Theorem 22.2 is also a key ingredient in many results regarding the pseudo-random generators against low-depth circuits. For example, Braverman's celebrated result [MR209] that k-wise independent fools constant depth AC circuits hinges on Theorem 22.2.

22.2 Razborov-Smolensky

Theorem 22.2 shows that every low-depth, small-size circuit can be approximated by a low-degree function in the L_2 distance.

The next theorem by Razborov [Raz87] and Smolensky [Smo87] shows a different approximation of such circuits with low-degree functions. In this theorem, the low-degree polynomial equals f on most elements in $\{0,1\}^n$. However,

when the two functions disagree, they can be very far apart.

Theorem 22.4 ([Raz87, Smo87]). *Let $f : \{0,1\}^n \to \{0,1\}$ be computed by an AC circuit of depth d and size M. For every s, there is a polynomial $g : \{0,1\}^n \to \mathbb{R}$ with degree $r \leqslant (s \log M)^d$ such that*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}}[f(\mathbf{x}) \neq g(\mathbf{x})] \leqslant \left(1 - \frac{1}{2e}\right)^s M,$$

where $\mathbf{x}$ is chosen randomly and uniformly from $\{0,1\}^n$. In particular, taking $s = 100 \log(M)$, there is a polynomial of degree $r \leqslant (100 \log M)^{2d}$, such that

$$\mathop{\mathbf{Pr}}_{\mathbf{x}}[f(\mathbf{x}) \neq g(\mathbf{x})] \leqslant \frac{1}{100}.$$

Proof. The key is approximating the $\wedge$ and $\vee$ gates with low-degree polynomials. The function g is constructed inductively. We will show how to make a step with an $\wedge$ gate. Since the whole construction is symmetric concerning 0 and 1, the step also holds for an $\vee$ gate. Let

$$f = \wedge_{i=1}^{k} f_i$$

where $k < M$. For convenience, assume that $k = 2^\ell$ is a power of 2. For every $j = 1, \ldots, \ell$, pick s random subsets of $\{1, \ldots, k\}$ by including every element in the subset independently with probability $p = 2^{-j}$. We obtain a collection of sets $S_1, \ldots, S_t$ with $t \coloneqq s\ell \leqslant s \log M$. Let $g_1, \ldots, g_k$ be the approximating functions for $f_1, \ldots, f_k$ provided by the previous inductive step. We set

$$g \coloneqq \prod_{i=1}^{t} (1 - |S_i| + \sum_{j \in S_i} g_j).$$

By the induction assumption, the degree of each g_j is at most $(s \log M)^{d-1}$. Hence, the degree of f is bounded by $t(s \log M)^{d-1} \leqslant (s \log M)^d$. Next, we bound the probability of $f(\mathbf{x}) \neq g(\mathbf{x})$ conditioned on the event that all of the inputs $f_1, \ldots, f_k$ are approximated correctly. Consider any x such that $g_j(x) = f_j(x)$ for all j. We have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_t}[f(x) \neq g(x)] = \mathop{\mathbf{Pr}}_{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_t}\left[\prod_{i=1}^{t} \left(1 - |\boldsymbol{S}_i| + \sum_{j \in \boldsymbol{S}_i} f_j(x)\right) \neq \prod_{j=1}^{k} f_j(x)\right].$$

To bound this, we fix a vector of specific values $f_1(x), \ldots, f_k(x)$ and calculate the probability that an error occurs over the possible choices of the random sets $\boldsymbol{S}_i$.

- If all the $f_j(x)$'s are 1, then the value of $f(x) = 1$ is calculated correctly with probability 1.
- Suppose that $f(x) = 0$, and thus at least one of the f_j's is 0. Note that for the product

  $$\prod_{i=1}^{t} \left(1 - |\boldsymbol{S}_i| + \sum_{j \in \boldsymbol{S}_i} f_j(x)\right)$$

  to evaluate to 0, it suffices to have one of the terms $1 - |\boldsymbol{S}_i| + \sum_{j \in \boldsymbol{S}_i} f_j$ to be 0. Let $1 \leqslant z \leqslant k$ be the number of zeros among $f_1(x), \ldots, f_k(x)$, and $\alpha \in \mathbb{Z}$ be such that $2^\alpha \leqslant z < 2^{\alpha+1}$. Let $\boldsymbol{S}$ be a random set with parameter $p = 2^{-\alpha-1}$. Our approximation will be correct if $\boldsymbol{S}$ hits *exactly* one 0 among the z zeros of $f_1(x), \ldots, f_k(x)$, as in this case, we would get $1 - |\boldsymbol{S}| - \sum_{j \in \boldsymbol{S}} f_j = 0$, making the whole product 0. The probability of this event is

  $$zp(1-p)^{z-1} \geqslant \frac{1}{2}(1-p)^{\frac{1}{p}-1} > \frac{1}{2e}.$$

  Therefore, the probability that all the s sets that are chosen with parameter $p = 2^{-\alpha-1}$ fail is at most $(1 - \frac{1}{2e})^s$ and

  $$\mathop{\mathbf{Pr}}_{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_t}\left[\prod_{i=1}^{t} \left(1 - |\boldsymbol{S}_i| + \sum_{j \in \boldsymbol{S}_i} f_j(x)\right) \neq \prod_{j=1}^{k} f_j(x)\right] < \left(1 - \frac{1}{2e}\right)^s.$$

By making the same probabilistic argument at every node and applying the union bound over all the $\leqslant M$ gates in the circuit, we conclude that the probability that an error occurs is at most $M\left(1-\frac{1}{2e}\right)^s$. Therefore, the low-degree polynomial $\boldsymbol{g}$ that we have probabilistically constructed satisfies: For every $x \in \{0,1\}^n$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}}[f(x) \neq \boldsymbol{g}(x)] \leqslant M\left(1-\frac{1}{2e}\right)^s.$$

Since this holds for every x, we have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g},\mathbf{x}}[f(\mathbf{x}) \neq \boldsymbol{g}(\mathbf{x})] \leqslant M\left(1-\frac{1}{2e}\right)^s,$$

which shows that there is a fixed low-degree polynomial g_0 satisfying

$$\mathop{\mathbf{Pr}}_{x}[f(x) \neq g_0(x)] \leqslant M\left(1-\frac{1}{2e}\right)^s.$$

□

22.3 The entropy-influence conjecture

Theorem 22.2 shows that the Fourier mass of every function computable by an AC circuit of polynomial size and constant depth is concentrated on the first $t = \log^{O(1)}(n)$ levels. Theorem 22.2 shows that the Fourier coefficients of every function computable by an AC circuit of polynomial size and constant depth are highly concentrated on the first $t = \log^{O(1)}(n)$ levels. There are roughly $\binom{n}{\leqslant t} \leqslant n^t$ such coefficients. While this is not an exponential number, it is still super-polynomial. The following conjecture, due to Mansour, speculates that for DNF, one can pinpoint the significant coefficients to a polynomial-size set.

Conjecture 22.5 (Mansour [Man95])**.** *Let f be computable by a DNF with at most t terms. For every ε, there exists a subset $\mathcal{S} \subseteq \mathcal{P}([n])$ of size $t^{O(\log 1/\varepsilon)}$ such that*

$$\sum_{S \notin \mathcal{S}} |\widehat{f}(S)|^2 \leqslant \varepsilon.$$

Let us mention another elegant conjecture regarding the concentration of the Fourier mass. The *entropy-influence conjecture*, due to Friedgut and Kalai [FK96], speculates that the entropy of the squares of the Fourier coefficients $\left\{\widehat{f}(S)^2\right\}_{S\subseteq[n]}$ is bounded by $O(I_f)$.

Conjecture 22.6 (Entropy-influence conjecture [FK96])**.** *There is a universal $C > 0$ such that every $f : \{0,1\}^n \to \{0,1\}$ satisfies*

$$H(\widehat{f}) \coloneqq \sum_{S\subseteq[n]} |\widehat{f}(S)|^2 \log\left(\frac{1}{|\widehat{f}(S)|^2}\right) \leqslant CI_f.$$

Since $I_f = \sum_{S\subseteq[n]} |S||\widehat{f}(S)|^2$, the Fourier mass is concentrated on the first $k = O(I_f)$ levels. Furthermore, by Friedgut's junta theorem (Theorem 12.3), the Fourier mass is only concentrated on a set J of $2^{O(I_f)}$ influential variables. Therefore, the Fourier mass is on a set of size at most $\binom{|J|}{\leqslant k} = 2^{O(I_f^2)}$, which yields the bound $H(\widehat{f}) = O(I_f^2)$.

Note also that, if true, Conjecture 22.6 implies that the Fourier mass of f is concentrated on the set

$$\mathcal{S} = \left\{S \;:\; |\widehat{f}(S)|^2 \geqslant 2^{\frac{-CI_f}{\varepsilon}}\right\},$$

which is of size at most $2^{\frac{CI_f}{\varepsilon}}$. In particular, if $\mathrm{Var}[f] = \Omega(1)$, then Conjecture 22.6 would imply

$$\text{Conjecture:} \quad \max_{S\neq\emptyset} |\widehat{f}(S)| \geqslant 2^{-O(I_f)}.$$

In contrast, Friedgut's junta theorem shows that every f with $\mathrm{Var}[f] = \Omega(1)$ satisfies

$$\max_{S \neq \emptyset} |\widehat{f}(S)| \geqslant 2^{-O(I_f^2)}. \tag{22.2}$$

Recently, Kelman, Kindler, Lifshitz, Minzer, and Safra [KKL+20] made significant progress toward resolving the entropy-influence conjecture. They proved the following theorem.

Theorem 22.7 ([KKL+20])**.** *There is a universal $C > 0$ such that every $f : \{0,1\}^n \to \{0,1\}$ and every $k > 0$, we have*

$$\sum_{|S| \leqslant k} |\widehat{f}(S)|^2 \log\left(\frac{1}{|\widehat{f}(S)|^2}\right) \leqslant C I_f + C \sum_{|S| \leqslant k} |S| \left(1 + \log(|S|)\right) |\widehat{f}(S)|^2.$$

Since $I_f = \sum_{S \subseteq [n]} |S| |\widehat{f}(S)|^2$, Theorem 22.7 falls short of proving the Fourier entropy conjecture by just a factor of $\log(|S|)$. Regarding the largest non-principal Fourier coefficient, Theorem 22.7 implies that if $\mathrm{Var}(f) = \Omega(1)$, we have

$$\max_{S \neq \emptyset} |\widehat{f}(S)| \geqslant 2^{-O(I_f (1 + \log I_f))},$$

which is a significant improvement over Equation (22.2).

Question 22.8. *Recall Equation (15.1) from Bourgain's sharp threshold theorem. Can one improve this lower bound using the ideas from the work of Kelman, Kindler, Lifshitz, Minzer, and Safra [KKL+20]?*

Chapter 23

Fourier Algebra Norm

Let G be a finite Abelian group. The sum of the absolute values of the Fourier coefficients is called the *Fourier algebra norm* or *spectral norm* of $f : G \to \mathbb{R}$ and is denoted by

$$\|f\|_A \coloneqq \|\widehat{f}\|_1 = \sum_{\chi \in \widehat{G}} |\widehat{f}(\chi)|.$$

The term *algebra* norm is explained by the easy-to-prove inequality $\|fg\|_A \leqslant \|f\|_A \|g\|_A$, which shows that the algebra of the functions $f : G \to \mathbb{R}$ (with point-wise addition and multiplication) endowed with the norm $\|\cdot\|_A$ is a Banach algebra.

The spectral norm arises naturally in theoretical computer science in connection to learning theory. It has been studied for several complexity classes of Boolean functions [STV17, KM93, TWXZ13, GTW21, Tal17, MRT19]. These studies are often motivated by the existence of efficient learning algorithms for the classes of Boolean functions that have small algebra norms [KM93]. Furthermore, in recent years, tail bounds in the Fourier L_1 norm have also become essential in constructing pseudo-random generators [CHHL19a, RSV13, FK18] and separating quantum and classical computation [RT19, Tal20, BS21a]. The algebra norm is also closely related to the parity decision tree complexity, a strengthening of the decision tree complexity.

The problem of characterizing functions with small Fourier algebra norms is also fundamental in harmonic analysis. Let G be a locally compact group, and let $\widehat{G}$ be its Pontryagin dual (also a locally compact Abelian group). Let $\mathrm{M}(G)$ be the algebra of all bounded regular Borel measures on G, where multiplication corresponds to convolution. Let $B(\widehat{G})$ denote the Fourier–Stieltjes algebra of $\widehat{G}$, which is the set of all $\widehat{\mu} : \widehat{G} \to \mathbb{C}$ for all $\mu \in \mathrm{M}(G)$ endowed with the norm $\|\widehat{\mu}\|_{B(\widehat{G})} \coloneqq \|\mu\|$. This norm is well-defined since the choice of μ is unique. If $\widehat{G}$ is a *finite* Abelian group, then $B(\widehat{G})$ is the set of all functions on $\widehat{G}$, and $\|\cdot\|_{B(\widehat{G})}$ coincides with the algebra norm: $\|f\|_{B(\widehat{G})} = \|f\|_A$.

If $\mu \in \mathrm{M}(G)$ is idempotent (i.e. $\mu * \mu = \mu$), then $\widehat{\mu}^2 = \widehat{\mu}$, so $\widehat{\mu}(\chi) \in \{0, 1\}$ for all $\chi \in \widehat{G}$. Hence, the problem of characterizing all idempotent measures in $\mathrm{M}(G)$ is equivalent to finding all subsets $A \subseteq \widehat{G}$ with $\mathbf{1}_A \in B(\widehat{G})$.

In 1940, Kawada-Itô [KI40, Theorem 3] characterized idempotent *probability measures* on compact groups as the normalized Haar measures of compact subgroups. When G is a finite group, their theorem translates to the statement that $f : G \to \{0, 1\}$ satisfies $\|f\|_A = 1$ iff f is the indicator function of a coset in G (see Theorem 23.12 below).

The Kawada-Itô theorem was rediscovered independently by Wendel [Wen54] in the context of harmonic analysis. Later, Rudin [Rud59a, Rud59b], trying to extend this result to all idempotent measures on locally compact Abelian groups, showed that any such measure is concentrated on a compact subgroup. Finally, Cohen [Coh60], building on the works of Helson [Hel53] and Rudin [Rud59a], obtained a full description of idempotent measures on locally compact Abelian groups. Cohen received the Bôcher Memorial Prize in mathematical analysis in 1964 for this result.

Numerous extensions and refinements of Cohen's theorem have been discovered since [Lef72, Hos86, GS08c, Run07, San11b, San20, San21]. We will discuss the qualitative version of Cohen's idempotent theorem for the group $\mathbb{Z}_2^n$ in Section 23.4.

23.1 Decision trees and Fourier algebra norm

Suppose $C : \{0,1\}^n \to \{0,1\}$ is defined by a single AND-clause $C(x) = \bigwedge_{j\in J}(x_j = b_j)$ for some $J \subseteq [n]$ and $b \in \{0,1\}^J$. We can easily write the Fourier expansion of C as

$$C(x) = \sum_{S\subseteq J} 2^{-|J|}\chi_S(b)\chi_S(x),$$

which shows all the non-zero Fourier coefficients have magnitude $2^{-|J|}$ and

$$\|C\|_A = 1.$$

Let T be a decision tree computing a function $f : \{0,1\}^n \to \{0,1\}$. Consider a leaf ℓ and let P be the path from the root to ℓ. Suppose $(x_{i_1}, \ldots, x_{i_k})$ is the sequence of the variables queried on this path, and let $(b_{i_1}, \ldots, b_{i_k}) \in \{0,1\}^k$ be the assignment of the values to these variables consistent with P. In other words, an input $x \in \{0,1\}^n$ follows the computational path P from the root to ℓ iff $(x_{i_1}, \ldots, x_{i_k}) = (b_1, \ldots, b_k)$. Let $C_\ell(x) = \wedge_{j=1}^k (x_{i_j} = b_j)$ be the AND-clause corresponding to the leaf ℓ. Denoting by $\mathcal{L}$ the set of all leaves of T, we have

$$f(x) = \sum_{\ell:\mathrm{label}(\ell)=1} C_\ell(x).$$

Since $\|C_\ell\|_A = 1$, we immediately obtain the following statement.

Proposition 23.1. *If $f : \{0,1\}^n \to \{0,1\}$ is computable by a decision tree with M leaves, then $\|f\|_A \leqslant M$. In particular,*

$$\|f\|_A \leqslant 2^{\mathrm{dt}(f)}.$$

23.2 Parity decision trees

Let us recall some basic facts about linear and affine subspaces of $\mathbb{Z}_2^n$. If $V \subseteq \mathbb{Z}_2^n$ is a linear subspace of co-dimension d, there exists linearly independent $a_1, \ldots, a_d \in \mathbb{Z}_2^n$ such that

$$V = \{x : \forall i \ \langle a_i, x\rangle = 0 \pmod 2\},$$

or equivalently

$$V = \{x : \forall i \ \chi_{a_i}(x) = 1\}.$$

Then $V^\perp = \mathrm{span}\{a_1, \ldots, a_d\}$, and the Fourier expansion of $\mathbf{1}_V$ is

$$\mathbf{1}_V = \sum_{v\in V^\perp} \frac{1}{2^d}\chi_v.$$

Consequently,

$$\|\mathbf{1}_V\|_A = 2^d \frac{1}{2^d} = 1.$$

More generally, consider an affine subspace $W = V + b$ for some $b \in \mathbb{Z}_2^n$. Then

$$\mathbf{1}_W = \sum_{v\in V^\perp} \frac{\chi_v(b)}{2^d}\chi_v,$$

and therefore, $\mathbf{1}_W$ contains 2^d non-zero Fourier coefficients, each with magnitude 2^{-d}, and we have $\|\mathbf{1}_W\|_A = 1$. We established that every affine subspace of $\mathbb{Z}_2^n$ has Fourier algebra norm 1. As we shall see in Theorem 23.12, these are the only sets with Fourier algebra norm 1. We will use these facts to show that small parity decision trees have a small Fourier algebra norm.

Definition 23.2 (Parity Decision tree)**.** A parity decision tree, denoted as $\oplus$-decision tree, is a labelled binary tree. Each internal node of the tree is labelled with a non-empty subset $S \subseteq [n]$, and each leaf by a bit $b \in \{0,1\}$. Given

an input $x \in \{0,1\}^n$, a computation over the tree is executed as follows: Starting at the root, stop if it is a leaf, and output its label. Otherwise, query $\oplus_S(x) \coloneqq \oplus_{i \in S} x_i$. If $\oplus_S(x) = 1$, then recursively evaluate the left subtree, and if $\oplus_S(x) = 0$, evaluate the right subtree.

Parity decision trees are generalizations of decision trees since querying $\oplus_{\{i\}}(x)$ corresponds to querying a single variable x_i.

Consider a leaf ℓ of a $\oplus$-decision tree, and suppose $(\oplus_{i \in S_1} x_i, \ldots, \oplus_{i \in S_k} x_i)$ is the sequence of the variables queried on this path, and let $(b_1, \ldots, b_k) \in \{0,1\}^k$ be the assignment of the values to these variables consistent with P. For $i = 1, \ldots, k$, let $a_i = \mathbf{1}_{S_i} \in \mathbb{Z}_2^n$ be the indicator vector of S_i. The set L_ℓ of all x whose computational path leads to ℓ is a coset of the subspace

$$\{a_1, \ldots, a_k\}^\perp \coloneqq \{x : \chi_{a_i}(x) = 1 \ \ \forall 1 \leqslant i \leqslant k\}.$$

So similar to the case of the decision tree, since $\|\mathbf{1}_{L_\ell}\|_A \leqslant 1$, we conclude that $\{f\}_A$ is at most the number of the leaves of the tree.

Proposition 23.3. *Let f be a Boolean function computed by a $\oplus$-decision tree. Then $\|f\|_A$ is bounded from above by the number of leaves of the tree. In particular, $\|f\|_A \leqslant 2^{\mathrm{pdt}(f)}$, where $\mathrm{pdt}(f)$ denotes the smallest depth of a parity decision tree computing f.*

The converse of Proposition 23.3 is not true. For example, the indicator function of the single point $\vec{0} \in \{0,1\}^n$ satisfies $\left\|\mathbf{1}_{\{\vec{0}\}}\right\|_A = 1$ while $\mathrm{pdt}\left(\mathbf{1}_{\{\vec{0}\}}\right) = n$.

While the above example shows that $\|\cdot\|_A$ does not imply small parity decision tree complexity, as we will show in Theorem 23.4, it implies small *randomized* parity decision tree complexity.

Randomized L_1 sampling: A *randomized parity decision tree* of depth at most d is a probability distribution $\mathcal{T}$ over parity decision trees of depth at most d. We say $\mathcal{T}$ computes f with error ε if

$$\Pr_{\boldsymbol{T} \sim \mathcal{T}}[\boldsymbol{T}(x) = f(x)] \geqslant 1 - \varepsilon \ \text{ for all } x \in \{0,1\}^n.$$

The *randomized parity decision tree complexity*, denoted by is the smallest depth of a randomized parity decision tree computing f with error $\varepsilon = 1/3$.

Theorem 23.4. *For every $f : \{0,1\}^n \to \{0,1\}$ and $\varepsilon > 0$, there is a randomized parity decision tree of depth $O\left(\log(1/\varepsilon)\|f\|_A^2\right)$ that computes f with error at most ε.*

Proof. Let $\psi_T(x) = \mathrm{sgn}(\widehat{f}(T))\chi_T(x) \in \{-1, 1\}$ for $T \subseteq [n]$. Pick $\boldsymbol{T} \subseteq [n]$ randomly according to the probability distribution

$$\mathbf{Pr}[\boldsymbol{T} = S] \coloneqq \frac{|\widehat{f}(S)|}{\|f\|_A} \ \text{ for all } S \subseteq [n].$$

For every $x \in \{0,1\}^n$, we have

$$\mathbb{E}_{\boldsymbol{T}}[\psi_{\boldsymbol{T}}(x)] = \sum_{S \subseteq [n]} \frac{\widehat{f}(S)}{\|f\|_A} \chi_S(x) = \frac{f(x)}{\|f\|_A}.$$

Let $N \coloneqq 8\|f\|_A^2 \log(4/\varepsilon)$. Let $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_N$ be i.i.d. copies of $\boldsymbol{T}$ and define $\widetilde{\boldsymbol{f}} = \frac{\|f\|_A}{N} \sum_{i=1}^N \psi_{\boldsymbol{T}_i}$.

For every $x \in \{0,1\}^n$, by applying Hoeffding's inequality (Lemma 5.1), we have

$$\mathbf{Pr}\left[\left|\widetilde{\boldsymbol{f}}(x) - f(x)\right| \geqslant \frac{1}{2}\right] < 2\exp\left(-\frac{8}{4N(\|f\|_A/N)^2}\right) \leqslant \varepsilon,$$

where the last inequality is by the choice of N.

Let $\boldsymbol{g}(x)$ be the Boolean rounding of $\widetilde{\boldsymbol{f}}(x)$, that is, $\boldsymbol{g}(x) = 1$ iff $\widetilde{\boldsymbol{f}}(x) \geqslant \frac{1}{2}$. We have

$$\mathbf{Pr}[\boldsymbol{g}(x) \neq f(x)] \leqslant \mathbf{Pr}\left[\left|\widetilde{\boldsymbol{f}}(x) - f(x)\right| \geqslant \frac{1}{2}\right] \leqslant \varepsilon.$$

Finally, note that we can compute $\widetilde{\boldsymbol{f}}(x)$ and $\boldsymbol{g}(x)$ by making the N parity queries $\chi_{\boldsymbol{T}_1}(x), \ldots, \chi_{\boldsymbol{T}_N}(x)$. □

Remark 23.5. The proof of Theorem 23.4 is quite robust and is applicable in any situation where $f(x) = \sum_{i=1}^{m} \lambda_i g_i(x)$ with $|g_i(x)| \leqslant 1$ for all i and we have a strong upper bound on the L_1 sum $\sum_{i=1}^{m} |\lambda_i|$. By sampling and querying a few g_i randomly according to the probability distribution $\mu(i) = \frac{|\lambda_i|}{\sum_{i=1}^{m} |\lambda_i|}$, we can produce a good prediction for the value of $f(x)$.

23.3 Matrix lower bounds for the Fourier algebra norm

In this section, we will discuss the relation between the Fourier algebra norm and the two well-known matrix norms, the trace and the factorization norms. The goal is to use the factorization norm to prove lower bounds on the Fourier algebra norm.

Recall that the singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_r$ of a matrix $M \in \mathbb{C}^{m \times n}$ are the square roots of eigenvalues of MM^*, where M^* is the conjugate transpose of M and $r = \mathbf{rk}(M)$.

For $p \in [1, \infty]$, we denoting by $\|M\|_p \coloneqq \max_{x \in \mathbb{C}^n} \frac{\|Mx\|_p}{\|x\|_p}$ the operator norm of M as a linear operator $M : \ell_p \to \ell_p$. For $p = 2$, the corresponding matrix norm $\|M\|_2$ is called the *spectral norm*, and it equals the *largest singular value* of the matrix.

$$\|M\|_2 = \sigma_{\max}(M).$$

The *trace norm* of M is the sum of its singular values:

$$\|M\|_{\mathrm{tr}} \coloneqq \sum_{i=1}^{r} \sigma_i.$$

It will be more convenient to normalize this norm and define

$$\|M\|_{\mathrm{ntr}} = \frac{1}{\sqrt{mn}} \|M\|_{\mathrm{tr}}.$$

Let us now define the γ_2 factorization norm of a matrix. There are a few equivalent ways to define this norm.

Proposition 23.6 (γ_2-factorization norm)**.** *For $M \in \mathbb{C}^{m \times n}$, the following definitions of the γ_2-factorization norm are equivalent.*

1. *We have*
$$\|M\|_{\gamma_2} = \min_{A,B:AB=M} \|A\|_{\mathrm{row}} \|B\|_{\mathrm{col}},$$
where the minimum is taken over any pair of matrices A and B satisfying $AB = M$, and $\|A\|_{\mathrm{row}}$ and $\|B\|_{\mathrm{col}}$ denote the largest ℓ_2-norm of a row in A and the largest ℓ_2 norm of a column in B, respectively.

2. *$\|M\|_{\gamma_2}$ is the minimum $c \geqslant 0$ such that there exists $d \in \mathbb{N}$ and vectors $a_i, b_j \in \mathbb{R}^d$ with $M_{ij} = \langle a_i, b_j \rangle$ and $|a_i||b_j| \leqslant c$ for all i, j.*

3. *Denoting by $\|M\|_{\ell_p \to \ell_q} \coloneqq \max_{x \in \mathbb{C}^n} \frac{\|Mx\|_q}{\|x\|_p}$ the operator norm of M as a linear operator $M : \ell_p \to \ell_q$, we have*
$$\|M\|_{\gamma_2} = \min_{AB=M} \|A\|_{\ell_\infty \to \ell_2} \|B\|_{\ell_1 \to \ell_2}.$$

4. *We have*
$$\|M\|_{\gamma_2} = \max_{A} \frac{\|M \circ A\|_2}{\|A\|_2},$$
where where $M \circ A$ is the entry-wise product *(a.k.a.* Schur *or* Hadamard *product) of M and A.*

5. *We have*
$$\|M\|_{\gamma_2} = \max_{u,v:\|u\|_2,\|v\|_2 \leqslant 1} \left\|M \circ vu^T\right\|_{\mathrm{tr}}$$

Proof. Exercise. □

The term *factorization* in γ_2 factorization norm refers to the factoring of M as the product of two operators A and B, and the index 2 in γ_2 refers to the fact that the factorization goes through the ℓ_2 space. The next proposition lists some key properties of the γ_2 norm.

Proposition 23.7. *The γ_2 norm satisfies the following properties.*

1. *(Norm axioms) For every $M, M_1, M_2 \in \mathbb{C}^{m\times n}$ and $\lambda \in \mathbb{C}$, we have*

   - $\|M\|_{\gamma_2} = 0$ *iff* $M = 0$.
   - $\|\lambda M\|_{\gamma_2} = |\lambda| \, \|M\|_{\gamma_2}$.
   - $\|M_1 + M_2\|_{\gamma_2} \leqslant \|M_1\|_{\gamma_2} + \|M_2\|_{\gamma_2}$.

2. *(Banach Algebra) For every $M_1, M_2 \in \mathbb{C}^{m\times n}$, we have*

   $$\|M_1 \circ M_2\|_{\gamma_2} \leqslant \|M_1\|_{\gamma_2} \|M_2\|_{\gamma_2}, \tag{23.1}$$

   Therefore, the γ_2 norm turns the algebra of matrices $\mathbb{C}^{m\times n}$ with the matrix addition and Schur product into a Banach algebra.

3. *$\|M\|_{\gamma_2}$ is invariant under rearranging, duplicating, or negating rows or columns of M.*

4. *For every submatrix M' of M, we have*

   $$\|M'\|_{\gamma_2} \leqslant \|M\|_{\gamma_2}.$$

   In particular, $\|\gamma_2\| \geqslant \max_{i,j} |M_{ij}|$.

5. *We have*

   $$\|M\|_{\mathrm{ntr}} \leqslant \|M\|_{\gamma_2}.$$

Proof. Exercise. □

Let G be a finite Abelian group. Given $f : G \to \mathbb{C}$, consider the matrix $L_f \in \mathbb{C}^{G\times G}$ with entries

$$L_f(x, y) \coloneqq f(x - y) \text{ for all } x, y \in G.$$

The matrix L_f corresponds to the convolution with f. More precisely, for every $g : G \to \mathbb{C}$, we have $L_f g = |G| f * g$ since

$$L_f g(x) = \sum_{y\in G} f(x - y)g(y) = |G| f * g(x).$$

In particular, for characters $\chi \in \widehat{G}$, we have

$$L_f \chi = |G| f * \chi = |G| \widehat{f}(\chi)\chi.$$

Therefore, every character χ is an eigenvector of L_f with the corresponding eigenvalue $|G|\widehat{f}(\chi)$.

Theorem 23.8. *Let G be a finite Abelian group. For every $f : G \to \mathbb{C}$, we have*

$$\|f\|_A = \|L_f\|_{\gamma_2} = \|L_f\|_{\mathrm{ntr}}.$$

In particular, if M is a submatrix of L_f, then

$$\|M\|_{\gamma_2} \leqslant \|f\|_A.$$

Proof. Since the eigenvalues of L_f are $|G|\widehat{f}(\chi)$ for $\chi \in \widehat{G}$, we have

$$\|L_f\|_{\mathrm{tr}} = \sum_{\chi\in\widehat{G}} |G||\widehat{f}(\chi)| = |G| \, \|f\|_A,$$

which shows $\|f\|_A = \|L_f\|_{\mathrm{ntr}}$. Let us now examine the γ_2 norm of L_f. Since, $f(x-y) = \sum_\chi \widehat{f}(\chi)\chi(x)\overline{\chi}(y)$, we have

$$L_f = \sum_{\chi \in \widehat{G}} \widehat{f}(\chi)\chi \otimes \overline{\chi}.$$

By the definition of the γ_2 norm, we have $\|\chi \otimes \overline{\chi}\|_{\gamma_2} = 1$. Therefore, $\|\chi \otimes \overline{\chi}\|_{\gamma_2} = 1$, and we have

$$\|L_f\|_{\gamma_2} \leqslant \sum_{\chi \in \widehat{G}} |\widehat{f}(\chi)| \, \|\chi \otimes \overline{\chi}\|_{\gamma_2} \leqslant \sum_{\chi \in \widehat{G}} |\widehat{f}(\chi)| = \|f\|_A = \|L_f\|_{\mathrm{ntr}}.$$

On the other hand, by Proposition 23.7, we have $\|L_f\|_{\gamma_2} \geqslant \|L_f\|_{\mathrm{ntr}}$. □

Remark 23.9. The discussion of this section easily generalizes to finite non-Abelian groups. Let G be any finite group, and let $f : G \to \mathbb{C}$. Define $L_f \in \mathbb{C}^{G \times G}$ as $L_f(x,y) = f(xy^{-1})$. The Fourier algebra norm of f is *defined as* $\|f\|_A = \|L_f\|_{\mathrm{ntr}}$. Furthermore, the equality $\|L_f\|_{\gamma_2} = \|f\|_A \coloneqq \|L_f\|_{\mathrm{ntr}}$ remains valid in this setting.

23.4 Quanitative Cohen's idempotent theorem

Let G be a finite Abelian group and consider the algebra of functions $f : G \to \mathbb{C}$ defined by the point-wise addition and multiplication. f is called an *idempotent* of this algebra if it satisfies $f^2 = f$. Therefore, the idempotents of this algebra are precisely the Boolean functions $f : G \to \{0,1\}$.

In this section, we will discuss a complete characterization of all Boolean functions $f : G \to \{0,1\}$ that have a small Fourier algebra norm. Given a subgroup $H \subseteq G$ and an element $a \in G$, the set $H + a$ is called a coset. When $G = \mathbb{Z}_2^n$, we can identify G with the n-dimensional vector space $\mathbb{F}_2^n$ over the two-element field $\mathbb{F}_2$. In this case, the subgroups of G are the linear subspaces, and the cosets are affine subspaces.

If $H + a$ is a coset, then $\|\mathbf{1}_{H+a}\|_A = 1$. Therefore, if $f : G \to \{0,1\}$ can be expressed as

$$f = \sum_{i=1}^{L} \pm \mathbf{1}_{H_i + a_i} \tag{23.2}$$

for some cosets $H_i + a_i$, then $\|f\|_A \leqslant L$. The notation of (23.2) means that each coefficient is $+1$ or -1. Cohen's celebrated idempotent theorem [Coh60] states the Fourier algebra norm of a Boolean function f on a locally compact Abelian group is finite iff f can be expressed as (23.2) for some finite L and a collection of open cosets $H_i + a_i$. We refer the interested readers to [Rud90, Chapter 3] for more details.

Cohen's theorem left open whether the number of terms L is *uniformly* bounded from above by a function of the Fourier algebra norm of f. Moreover, Cohen's original theorem gave no information for finite groups since we can always write $f = \sum_{a \in f^{-1}(1)} \mathbf{1}_{\{a\}}$. However, one can ask whether it is possible to uniformly bound L in terms of $\|f\|_A$.

Five decades later, Green and Sanders [GS08c, GS08a], using modern tools from additive combinatorics, proved a stronger quantitative version of Cohen's theorem that resolved the uniformity question. Their result can be applied to finite groups as well. It states that in (23.2), we can choose $L \leqslant \ell(\|f\|_A)$, where $\ell(\cdot)$ is a universal function that does not depend on the choice of the underlying group G. They first proved the special case of this theorem for the groups $\mathbb{Z}_2^n$ and afterwards generalized it to all locally compact Abelian groups in [GS08c]. The bounds obtained in these two papers were later improved by Sanders [San19, San20]. We will state their theorem for the group $\mathbb{Z}_2^n$.

Theorem 23.10 (Green and Sanders [GS08a])**.** *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be a boolean function, and suppose that the Fourier algebra norm $\|f\|_A$ is at most M. Then There exists affine subspaces $V_1, \dots, V_L$ of $\mathbb{Z}_2^n$ for some $L \leqslant \ell(M)$ such that*

$$f = \sum_{j=1}^{L} \pm \mathbf{1}_{V_j} \tag{23.3}$$

Remark 23.11. The original paper of Green and Sanders [GS08b] proves a bound of $L \leqslant 2^{2^{O(M^4)}}$. Sanders later improve this bound to $L \leqslant 2^{O(M^3 \operatorname{polylog}(M))}$. Recently, Gowers, Green, Manners, and Tao [GGMT23] proved Morton's conjecture (aka polynomial Freiman–Ruzsa conjecture). Substituting this result in Sanders proof for [San19, Proposition 2] shows that in the case of $\mathbb{Z}_2^n$, one may take $\ell(M) = 2^{O(M \operatorname{polylog}(M))}$.

We will not prove Theorem 23.10 in this course as its proof is based on various results from additive combinatorics. However, we will discuss the extreme case of $\|f\|_A \leqslant 1$. The following proposition shows that the only non-zero Boolean matrices $f : G \to \{0,1\}$ with $\|f\|_A \leqslant 1$ are the indicator functions of the cosets.

Theorem 23.12. *A Boolean function $f : G \to \{0,1\}$ satisfies $\|f\|_A < \sqrt{\frac{4}{3}}$ if and only if f is the indicator function of a coset, in which case $\|f\|_A = 1$.*

Proof. Let $S \subseteq G$ denote the support of f. If $f \neq 0$, then $\|f\|_\infty \geqslant 1$, which immediately implies $\|f\|_A \geqslant 1$.

If S is not the coset of a subgroup of G, then by Lemma 23.13 below, there are $a, b, c \in S$ such that $a + b - c \notin S$. Consider the 2×2 submatrix M of L_f induced by the rows $\{c, b\}$ and columns $\{0, c - a\}$. Note

$$M = \begin{bmatrix} f(c) & f(a) \\ f(b) & f(a+b-c) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix},$$

and therefore, by Theorem 23.8, we have $\|f\|_A \geqslant \|M\|_{\gamma_2}$. On the other hand, by Lemma 23.14 below, $\|M\|_{\gamma_2} \geqslant \sqrt{4/3}$. □

Lemma 23.13. *Let G be an Abelian group. A set $S \subseteq G$ is a coset of a subgroup of G iff for every $a, b, c \in S$, we have $a + b - c \in S$.*

Proof. One direction is obvious. If S is a coset, then for every $a, b, c \in S$, we have $a + b - c \in S$.

For the converse direction, suppose for every $a, b, c \in S$, we have $a+b-c \in S$. Let $H = S - S$. It is straightforward to verify that H is a subgroup: $0 \in H$; if $x \in H$, then $-x \in H$; finally, if $x = s_1 - s_2 \in S - S = H$ and $x' = s_1' - s_2' \in S - S = H$, then by our assumption, $x + x' = (s_1 + s_1' - s_2) - s_2' \in S - S = H$.

Take any $s \in S$. We claim $S = s + H$. Obviously, $S \subseteq S - S + s = H + s$. On the other hand, every element x of $H + s$ is of the form $x = s_1 - s_2 + s$ with all the terms in S. Hence, $x \in S$ by our assumption. □

Lemma 23.14 (Livshits [Liv95])**.** *We have*

$$\left\| \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \right\|_{\gamma_2} = \frac{2}{\sqrt{3}} > 1.$$

23.5 Fourier folding and Shpilika-Tal-Volk

Theorem 23.15 (Shpilka, Tal, and Volk [SlV13])**.** *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfy $\|\widehat{f}\|_1 \leqslant M$. There exists a co-set V of co-dimension at most M^2 such that f is constant on V.*

The proof relies on the simple equation $f^2 = 1$. By expanding the Fourier representation of both sides, we obtain that for every $b \neq 0$,

$$\sum_{a \in \mathbb{Z}_2^n} \widehat{f}(a)\widehat{f}(a+b) = 0.$$

This identity could be interpreted as saying that the Fourier mass of pairs whose product is positive is the same as the mass of pairs whose product is negative. In particular, if we consider the two heaviest elements in the Fourier spectrum, say, $|\widehat{f}(\alpha)|$ and $|\widehat{f}(\beta)|$, and let $\delta = \alpha + \beta$, then by restricting f to one of the subspaces $\chi_\delta = 1$ or $\chi_\delta = -1$ we obtain a substantial decrease in the Fourier algebra norm. This decrease occurs since there is a significant L_1 mass on some pairs $\widehat{f}(\lambda)$ and $\widehat{f}(\lambda + \delta)$ that have different signs.

Before starting the proof of Theorem 23.15, let us discuss the effect of restricting a function to a coset on the Fourier spectrum. Consider $f : \mathbb{Z}_2^n \to \mathbb{R}$, and let $a \in \mathbb{Z}_2^n$ be a *non-zero* element. Consider the $(n-1)$-dimensional subspace $V = \{a\}^\perp = \{x : \chi_a(x) = 1\}$ and its coset $W = \{x : \chi_a(x) = -1\}$. Since V is a subspace over $\mathbb{Z}_2$, it can be identified with $\mathbb{Z}_2^{n-1}$, and hence it is meaningful to discuss the Fourier transform of $f|_V$. For every $b \in V$, the coefficients $\widehat{f}(b)$ and $\widehat{f}(b+a)$ collapse to a single coefficient:

$$\widehat{f|_V}(b) \coloneqq \widehat{f}(b) + \widehat{f}(a+b). \tag{23.4}$$

Similarly for every $b \in W$,

$$\widehat{f|_W}(b) \coloneqq \widehat{f}(b) - \widehat{f}(a+b). \tag{23.5}$$

This phenomenon is sometimes called *Fourier folding*.

The following Lemma 23.16 is the key element of the proof of Theorem 23.15

Lemma 23.16. *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be a Boolean function such that $\|f\|_A = M > 1$. Then there exists $\gamma \in \mathbb{Z}_2^n$ and $b \in \{0,1\}$ such that $\|\widehat{f|_{\chi_\gamma = b}}\|_1 \leqslant M - \frac{1}{M}$.*

Proof. Let $\widehat{f}(\alpha)$ be the maximal Fourier coefficient of f in absolute value, and $\widehat{f}(\beta)$ be the second largest. It follows from $\sum |\widehat{f}(a)| = M$ and the Parseval identity $\sum |\widehat{f}(a)|^2 = 1$ that $|\widehat{f}(\alpha)| \geqslant \frac{1}{M}$. We can assume that $\widehat{\beta} \neq 0$, as otherwise the function f must be of the form $\pm\chi_\alpha$, and that corresponds to an $(n-1)$-dimensional coset.

Without loss of generality, assume that $\widehat{f}(\alpha)\widehat{f}(\beta) > 0$, i.e., these two Fourier coefficients have the same sign; the other case is completely analogous. By taking the Fourier transform of both sides of $f^2 = 1$, we have

$$\sum_{\gamma \in \mathbb{Z}_2^n} \widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma) = \widehat{1}(\alpha+\beta) = 0. \tag{23.6}$$

Let $N_{\alpha+\beta} \subseteq \mathbb{Z}_2^n$ be the set of $\gamma \in \mathbb{Z}_2^n$ with $\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma) < 0$. By our assumption $\widehat{f}(\alpha)\widehat{f}(\beta) > 0$, we have $\alpha, \beta \notin N_{\alpha+\beta}$. Switching sides in Equation (23.6), we obtain

$$2\left|\widehat{f}(\alpha)\widehat{f}(\beta)\right| = \sum_{\gamma \in N_{\alpha+\beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right| - \sum_{\substack{\gamma \notin N_{\alpha+\beta} \\ \gamma \neq \alpha,\beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right|.$$

In particular,

$$|\widehat{f}(\alpha)||\widehat{f}(\beta)| \leqslant \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right|. \tag{23.7}$$

We now use the fact that $\widehat{f}(\beta)$ is the second largest in absolute value, and $\widehat{f}(\alpha)$ does not appear in the sum, to bound the right-hand side:

$$\sum_{\gamma \in N_{\alpha+\beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right| \leqslant |\widehat{f}(\beta)| \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{|\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)|\right\}. \tag{23.8}$$

Then (23.7) and (23.8) together with the assumption $|\widehat{f}(\beta)| \neq 0$ imply

$$|\widehat{f}(\alpha)| \leqslant \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{|\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)|\right\}. \tag{23.9}$$

Let $f' = f|_{\chi_{\alpha+\beta}=1}$. Then by (23.4), for every γ, the coefficients $\widehat{f}(\gamma)$ and $\widehat{f}(\alpha+\beta+\gamma)$ collapse to a single coefficient whose absolute value is $|\widehat{f}(\gamma) + \widehat{f}(\alpha+\beta+\gamma)|$. For $\gamma \in N_{\alpha+\beta}$,

$$|\widehat{f}(\gamma) + \widehat{f}(\alpha+\beta+\gamma)| = \left||\widehat{f}(\gamma)| - |\widehat{f}(\alpha+\beta+\gamma)|\right|$$

which reduces the L_1 norm of f' compared to that of f by at least $\min(|\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)|)$. In total, since both γ and $\alpha+\beta+\gamma$ belong to $N_{\alpha+\beta}$, we obtain

$$\|f'\|_A \leqslant \|f\|_A - \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{|\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)|\right\}.$$

Therefore by (23.9) we have

$$\|f'\|_A \leqslant \|f\|_A - |\widehat{f}(\alpha)| \leqslant M - \frac{1}{M}.$$

□

Proof of Theorem 23.15. Apply Lemma 23.16 iteratively on f. After less than M^2 steps, we are left with a function g, which is a restriction of f on a coset defined by the restrictions so far, such that $\|g\|_A \leqslant 1$. Then Theorem 23.12 finishes the proof. □

23.6 Concluding remarks and open problems

Tsang, Wong, Xie, and Zhang [TWXZ13] noticed that a slight twist in the proof of Theorem 23.15 improves the co-dimension to $O(M)$.

It is not difficult to see that in Lemma 23.16, the restriction $f|_{\chi_\gamma \neq b}$ also provides some decrease in the Fourier algebra norm. That is $\left\|\widehat{f|_{\chi_\gamma \neq b}}\right\|_A \leqslant \|f\|_A - |\widehat{f}(\beta)|$, where $\widehat{f}(\beta)$ is the second largest Fourier coefficient in absolute value. Using this observation, one can prove the following theorem.

Theorem 23.17 (Shpilka, Tal, and Volk [SlV13])**.** *Every $f : \{0,1\}^n \to \{0,1\}$ with $\|f\|_A \leqslant M$ is computable by a parity decision tree of* size *at most $2^{M^2} n^M$.*

An important class of Boolean functions with small Fourier algebra norms are the *Fourier sparse functions*.

Definition 23.18 (Fourier Sparsity)**.** Let G be a finite Abelian group. The *Fourier sparsity* of $f : G \to \mathbb{C}$, denoted by $\mathrm{sp}(f)$, is the number of non-zero Fourier coefficients of f.

Remark 23.19. We showed that the eigenvalues of L_f are the Fourier coefficients of f. Therefore, $\mathrm{sp}(f) = \mathbf{rk}(L_f)$.

Since the absolute values of the Fourier coefficients of a Boolean function $f : G \to \{0,1\}$ are at most 1, Boolean functions satisfy $\|f\|_A \leqslant \mathrm{sp}(f)$.

Conjecture 23.20 ([MO09, ZS10])**.** *Every $f : \{0,1\}^n \to \{0,1\}$ is computable by a parity decision tree of depth at most* $\mathrm{polylog}(\mathrm{sp}(f))$.

The same techniques used in the proof of Theorem 23.17 can prove the following theorem.

Theorem 23.21 (Shpilka, Tal, and Volk [SlV13])**.** *Every $f : \{0,1\}^n \to \{0,1\}$ with $\|f\|_A \leqslant M$ is computable by a parity decision tree of depth at most $M^2 \log(\mathrm{sp}(f))$.*

23.6.1 Boolean matrices with small γ_2-norm

Next, we discuss a conjecture about extending Green and Sanders' idempotent theorem to matrices. First, we will characterize the Boolean matrices whose γ_2-norm is at most 1.

Proposition 23.22. *For every m, the γ_2-norm of the $m \times m$ identity matrix $\mathbf{I}_m$ is 1,*

Proof. Since the standard basis $e_1, \ldots, e_m \in \mathbb{R}^m$ satisfies

$$\langle e_i, e_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases},$$

$\mathbf{I}_m$ satisfies $\|\mathbf{I}_m\|_{\gamma_2} \leqslant 1$. Moreover, it is clear from the definition of the γ_2-norm that for every matrix A, we have $\|A\|_{\gamma_2} \geqslant \|A\|_\infty := \max_{x,y} |A(x,y)|$, and therefore, $\|\mathbf{I}_m\|_{\gamma_2} \geqslant 1$. □

Proposition 23.22 shows that all identity matrices have γ_2-norm 1. Note that the proof of Proposition 23.22 generalizes to a larger class of matrices, which we call *blocky matrices*.

Definition 23.23 (Blocky matrices)**.** A Boolean matrix $M \in \{0,1\}^{\mathcal{X} \times \mathcal{Y}}$ is *blocky* if there exist disjoint sets $\mathcal{X}_i \subseteq \mathcal{X}$ and disjoint sets $\mathcal{Y}_i \subseteq \mathcal{Y}$ such that the support of M is exactly $\bigcup_i \mathcal{X}_i \times \mathcal{Y}_i$.

It turns out that blocky matrices are precisely the set of Boolean matrices with γ_2-norm at most 1.

Proposition 23.24 (Livshits [Liv95])**.** *A non-zero Boolean matrix M satisfies $\|F\|_{\gamma_2} = 1$ iff M is a blocky matrix.*

Proof. By Lemma 23.14, a Boolean matrix M with $\|M\|_{\gamma_2} \leqslant 1$ cannot have any 2×2 submatrices with exactly 3 ones. It is straightforward to verify that a Boolean matrix satisfying this property must be blocky. □

Since every Boolean matrix with γ_2-norm 1 is blocky, it is natural to ask whether Boolean matrices of bounded γ_2-norm can be characterized through blocky matrices. The following conjecture is an analogue of Theorem 23.10.

Conjecture 23.25 ([HHH23]). *Suppose that M is a Boolean matrix with $\|M\|_{\gamma_2} \leqslant c$. Then we may write*

$$M = \sum_{i=1}^{L} \pm B_i, \tag{23.10}$$

where B_i are blocky matrices and $L \leqslant \ell(c)$ for some integer $\ell(c)$ depending only on c.

Conjecture 23.25, inspired by Cohen's idempotent theorem, is known to be true for a large class of Boolean matrices.

Proposition 23.26. *Conjecture 23.25 is true for the matrices of $L_f \in \{0,1\}^{G\times G}$ where G is a finite Abelian group, $f : G \to \{0,1\}$, and $L_f(x,y) = f(x-y)$ for every (x,y).*

Proof. Observe if $g = \mathbf{1}_{H+a}$ where $H + a$ is a coset, then L_g is a Blocky matrix. The proposition follows from Theorem 23.10. □

Remark 23.27. Proposition 23.26 is true even when G is a non-Abelian group and $L_f(x,y) = f(xy^{-1})$. This follows from a theorem of Sanders [San11b] that establishes a quantitative Cohen's theorem (i.e., analogue of Theorem 23.10) for non-Abelian groups.

Chapter 24

Pseudorandom Generators

The content of this chapter is mostly taken from the survey [HH24]. A *pseudorandom generator* (PRG) uses a small amount of true randomness, called the *seed*, to generate a long sequence that appears to be random in certain aspects. PRGs have many applications in computational theory and practice. One motivation is that we think of randomness as a scarce computational resource akin to time or space, so whenever we get our hands on random bits, we want to stretch them as far as possible. Furthermore, when the seed length s is small, we can derandomize certain probabilistic algorithms by *exhaustively trying all possible seeds* of a PRG.

To model PRGs mathematically, we consider some *observer*, modelled as a function f. Let U_n denote the uniform random variable over $\{0,1\}^n$. We want to *fool* f into mistaking a random variable X with U_n

Definition 24.1 (Fooling)**.** Suppose $f : \{0,1\}^n \to \mathbb{R}$ is a function, X is a random variable taking values in $\{0,1\}^n$, and $\varepsilon > 0$. We say that X *fools* f *with error* ε, or ε*-fools* f, if

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(U_n)]| \leqslant \varepsilon.$$

If $\varepsilon = 0$, we say that X *perfectly* fools f.

Definition 24.1 says that although X might not be uniform, X and a truly uniform random variable are nevertheless *indistinguishable* by f. A PRG's job is to use a few truly random bits to sample a distribution that fools f.

Definition 24.2 (PRGs)**.** Suppose $f : \{0,1\}^n \to \mathbb{R}$ and $G : \{0,1\}^s \to \{0,1\}^n$ are functions and $\varepsilon > 0$. We say that G is an ε*-PRG for* f if $G(U_s)$ fools f with error ε.

One of the great ideas in the theory of computing is to try to design a PRG that fools *all computationally efficient* observers. Given such a PRG and a truly random seed, we could execute any randomized algorithm worth executing. After all, there's no point running a program if one won't even survive long enough to see the output! Such a PRG could also be used in cryptographic settings because we can safely assume that eavesdroppers and hackers only have so much computational power.

For example, given a random seed $X_0 \in \{1, 2, \ldots, M-1\}$, the Blum-Blum-Shub (BBS) generator [BBS86] outputs the sequence $(X_1 \bmod 2, X_2 \bmod 2, X_3 \bmod 2, \ldots)$ where

$$X_{i+1} = X_i^2 \bmod M.$$

For a suitably chosen modulus M, the BBS generator is believed to fool all polynomial-time algorithms!

Fooling all efficient observers is a well-defined and well-motivated *goal*. Unfortunately, nobody can prove that some efficiently computable PRG *satisfies* this marvellous property.

To be clear, a substantial body of evidence indicates that such PRGs exist. For example, Blum, Blum, and Shub showed that their generator fools all polynomial-time observers under the plausible but unproven assumption that no good algorithm exists for the *quadratic residuosity problem* [BBS86]. There are many other examples of PRGs that fool all polynomial-time observers under reasonable cryptographic or complexity-theoretic assumptions.

The problem of designing PRGs that unconditionally fool all efficient observers is very challenging, with connections to deep topics such as the famous **P** vs. **NP** problem. Therefore, much of the research on PRGs focuses on interesting

and well-defined *restricted model of computation*. Then, we design PRGs that fool the chosen model of computation (unconditionally – with no unproven assumptions) and try to optimize the seed length of the PRG.

A toy example might clarify the idea. Let us design a PRG $G : \{0,1\}^2 \to \{0,1\}^3$ that fools every observer f that only looks at two of the three output bits. This problem is not completely trivial because we do not know which two bits f will observe. Nevertheless, the problem can be solved by defining

$$G(u_1, u_2) = (u_1, u_2, u_1 \oplus u_2).$$

When u_1 and u_2 are chosen uniformly at random, the three output bits are correlated, but any two of the bits are independent and uniformly random.

We will be especially interested in fooling computation models with a complexity theory flavour, i.e., we want the output of the PRG to appear random to any observer that is sufficiently efficient in some sense. Arguably, the two most important models in this field are *constant-depth circuits* and *read-once branching programs*.

24.1 The generic probabilistic existence proof

For many classes $\mathcal{F}$, including classes defined by standard computational models such as decision trees, circuits, and branching programs, there is a generic argument showing that there exist PRGs that fool $\mathcal{F}$ with a small seed length.

Proposition 24.3 (Nonexplicit PRGs)**.** *Let $\mathcal{F}$ be a class of functions $f : \{0,1\}^n \to \{0,1\}$. For every $\varepsilon > 0$, there exists an ε-PRG for $\mathcal{F}$ with seed length* $\log\log|\mathcal{F}| + 2\log(1/\varepsilon) + O(1)$.

Proof. Pick a function $\boldsymbol{G}\colon \{0,1\}^s \to \{0,1\}^n$ uniformly at random. Consider any arbitrary $f \in \mathcal{F}$. For each seed y, the value $f(\boldsymbol{G}(y))$ is a random bit satisfying

$$\mathbb{E}_{\boldsymbol{G}}[f(\boldsymbol{G}(y))] = \mathbb{E}[f].$$

Furthermore, as y ranges over all 2^s possible seeds, these random variables $f(\boldsymbol{G}(y))$ are independent. Therefore, by Hoeffding's inequality,

$$\Pr_{\boldsymbol{G}}\left[\left|\mathbb{E}[f] - 2^{-s}\sum_{y\in\{0,1\}^s} f(\boldsymbol{G}(y))\right| > \varepsilon\right] \leqslant 2e^{-2\varepsilon^2 2^s}.$$

By the union bound, the probability that $\boldsymbol{G}$ fails to ε-fool $\mathcal{F}$ is bounded by $2|\mathcal{F}|e^{-2\varepsilon^2 2^s}$. For $s = \log\log|\mathcal{F}| + 2\log(1/\varepsilon) + O(1)$, this probability is less than 1, i.e., there exists a G that does ε-fool $\mathcal{F}$. □

In a typical case – e.g., if $\mathcal{F}$ is the set of all circuits of size at most $n^{O(1)}$ – each function $f \in \mathcal{F}$ can be described using $\mathrm{poly}(n)$ bits, i.e., $|\mathcal{F}| \leqslant 2^{\mathrm{poly}(n)}$. In this case, the PRG guaranteed by Proposition 24.3 has seed length $O(\log(n/\varepsilon))$.

Proposition 24.3 has a major weakness: it does not guarantee that the PRG is *efficiently computable* since its proof is in some sense "nonconstructive."

Definition 24.4 (Explicitness)**.** A PRG $G\colon \{0,1\}^s \to \{0,1\}^n$ is *explicit* if it can be computed in time $\mathrm{poly}(n)$.

The default conjecture: Explicit PRGs exist! For each "reasonable" class $\mathcal{F}$, the standard conjecture is that there exists an *explicit* PRG with essentially the same seed length as the generic nonexplicit bound (Proposition 24.3). Often, this conjecture can be supported with evidence in the form of conditional constructions. For example, consider the class $\mathcal{F}$ of all CNF formulas of size at most n. The nonexplicit PRG has seed length $O(\log(n/\varepsilon))$. Under plausible complexity-theoretic assumptions, there is indeed an explicit PRG for *all* size-n Boolean circuits (whether CNF formulas or not) with seed length $O(\log(n/\varepsilon))$ [IW97].

24.2 Fourier uniformity and PRGs

Definition 24.5. A random variable X taking values in $\{0,1\}^n$ is called *ε-Fourier uniform* if for every non-principal character χ_S, we have

$$|\mathbb{E}[\chi_S(X)]| \leqslant \varepsilon.$$

Since the expected value of every non-principal character is 0, X is ε-Fourier uniform iff it fools all the non-principal characters.

If X is uniform over a set $A \subseteq \{0,1\}^n$ of density α, then ε-Fourier uniformity of X means $|\widehat{A}(S)| \leqslant \varepsilon\alpha$ for all non-principal characters χ_S. Note that this definition is closely related to Definition 5.2 but with a different normalization of the uniformity parameter ε.

Construction of Fourier Uniform PRGs: Naor and Naor [NN93] and independently Peralta [Per90] gave explicit constructions of ε-Fourier uniform PRGs with seed length $O(\log(n/\varepsilon))$; we'll present a simpler construction due to Alon, Goldreich, Håstad, and Peralta [AGHP92].

Theorem 24.6 (Fourier uniform PRGs [NN93, Per90])**.** *For every $n \in \mathbb{N}, \varepsilon > 0$, there is an explicit ε-Fourier uniform generator with output length n and seed length $O(\log(n/\varepsilon))$.*

Proof of Theorem 24.6. Let $q = n/\varepsilon$, and assume without loss of generality that $q = 2^k$ for an integer k. As vector spaces over $\mathbb{F}_2$, identify the field $\mathbb{F}_{2^k}$ with $\mathbb{F}_2^k$. Our pseudo-random generator $G\colon \mathbb{F}_2^k \times \mathbb{F}_{2^k} \to \{0,1\}^n$ is defined by

$$G(y,z) \coloneqq \left(\langle y, z^1\rangle_{\mathbb{F}_2}, \langle y, z^2\rangle_{\mathbb{F}_2}, \ldots, \langle y, z^n\rangle_{\mathbb{F}_2}\right) \in \mathbb{F}_2^n$$

where z^i refers to the i-th power of z in $\mathbb{F}_{2^k}$, and the inner product is defined by identifying $z^i \in \mathbb{F}_{2^k}$ with its corresponding elements in $\mathbb{F}_2^k$.

To prove that G fools all non-principal characters, let $f\colon \{0,1\}^n \to \{0,1\}$ be a nonzero parity function, say $f(x) = \bigoplus_{i\in S} x_i$. Then doing arithmetic in $\mathbb{F}_2$,

$$f(G(y,z)) = \sum_{i\in S} \langle y, z^i\rangle_{\mathbb{F}_2} = \left\langle y, \sum_{i\in S} z^i \right\rangle_{\mathbb{F}_2}.$$

Define $g(z) = \sum_{i\in S} z^i$. Then g is a nonzero polynomial in $\mathbb{F}_q[z]$ of degree at most n, and therefore, it has at most n roots. Since $f(G(y,z)) = \langle y, g(z)\rangle_{\mathbb{F}_2}$, when z is a root of g_S, we have $f(G(y,z)) = 0$. On the other hand, when z is not a root of g, if we sample $Y \in \mathbb{F}_q$ uniformly at random, $f(G(Y,z))$ is a uniform random bit. Therefore, when we sample $Y, Z \in \mathbb{F}_q$ independently and uniformly at random,

$$\mathop{\mathbf{E}}_{Y,Z}[f(G(Y,Z))] = \frac{1}{2} \cdot \mathop{\mathbf{Pr}}_{Z}[g(Z) \neq 0] \in \left[\frac{1}{2} - \frac{n}{2q}, \frac{1}{2}\right].$$

Since $\mathbb{E}[f] = \frac{1}{2}$, our PRG G fools parity functions with error $\frac{n}{2q} = \frac{\varepsilon}{2}$, and hence it fools character functions with error at most ε. □

Fourier uniformity fools small Fourier L_1: Since Fourier uniform random variables fool all the non-principal characters, it easily follows that they fool every function with a small Fourier algebra norm.

Proposition 24.7. *Let X be an ε-Fourier uniform random variable taking values in $\{0,1\}^n$. Every $f : \{0,1\}^n \to \mathbb{R}$ satisfies*

$$|\mathbb{E}f(X) - \mathbb{E}f(U_n)| \leqslant \varepsilon \|f\|_A.$$

Proof.

$$|\mathbb{E}f(X) - \mathbb{E}f(U_n)| = \left|\widehat{f}(\emptyset) + \sum_{S\neq\emptyset} \widehat{f}(S)\mathbb{E}[\chi_S(X)] - \widehat{f}(\emptyset)\right| \leqslant \varepsilon \sum_{S\neq\emptyset} |\widehat{f}(S)| = \varepsilon\|f\|_A.$$

□

Recall by Proposition 23.3, we have $\|f\|_A \leqslant 2^{\mathrm{pdt}(f)}$ where $\mathrm{pdt}(f)$ denotes the parity decision tree complexity of f. Therefore, by Theorem 24.6 and Proposition 24.7 explicit PRGs with seed length $\mathrm{polylog}(n/\varepsilon)$ that ε-fool functions $f : \{0,1\}^n \to \{0,1\}$ with $\mathrm{pdt}(f) = \mathrm{polylog}(n)$.

24.3 k-wise uniformity

Let $X = (X_1, \ldots, X_n) \in \{0,1\}^n$ be a random variable. For $k \in [n]$, we say that X is *k-wise independent*, if for every $S \subseteq [n]$ with $|S| = k$, the bits $(X_i : i \in S)$ are independent.

We say that X if for every $S \subseteq [n]$ with $|S| = k$, X_S is uniform over $\{0,1\}^S$. Note that a k-wise uniform X is k-wise independent but not vice versa.[1]

Since a k-junta inspects at most k variables, it cannot distinguish between a k-wise uniform random variable and the uniform random variable U_n.

Proposition 24.8. *Every k-wise uniform X taking values in $\{0,1\}^n$ perfectly fools all k-juntas $f : \{0,1\}^n \to \mathbb{R}$.*

Constructing k-wise uniform PRGs: We start with the case of pairwise uniform, which has a simple construction.

Proposition 24.9. *For every $n \in \mathbb{N}$, there is an explicit pseudo-random generator $G : \{0,1\}^s \to \{0,1\}^n$ with seed length $s = \lfloor \log n \rfloor + 1$ that is pairwise uniform. In particular, it perfectly fools 2 juntas on n bits.*

Proof. If Y is a uniform random variable taking values in $\{0,1\}^s$, then

$$(\oplus_{i \in S} y_i : \ S \subseteq [s], S \neq \emptyset)$$

is a collection of $n = 2^s - 1$ pairwise uniform random bits. □

For the more general case, the key idea is that if we pick a random polynomial p of degree $d > k$, then for any fixed distinct $z_1, \ldots, z_k$, the random variables $p(z_1), \ldots, p(z_k)$ are mutually independent.

Theorem 24.10 (k-wise uniform bits)**.** *For every $n, k \in \mathbb{N}$, there is an explicit pseudo-random generator $G : \{0,1\}^s \to \{0,1\}^n$ with seed length $s = O(k \log n)$ that is k-wise uniform. In particular, it perfectly fools k-juntas on n bits.*

Proof. Let $\mathbb{F}_q$ be a finite field with at least n elements, and let $\mathcal{P}$ be the set of univariate polynomials over $\mathbb{F}_q$ of degrees less than k. Let $z_1, \ldots, z_k \in \mathbb{F}_q$ be distinct. In preparation for defining the PRG, define $H \colon \mathcal{P} \to \mathbb{F}_q^k$ by

$$H(p) = (p(z_1), \ldots, p(z_k)).$$

Since two polynomials with degrees less than k can be equal on at most k points, The function H is injective. Furthermore, $|\mathcal{P}| = |\mathbb{F}_q^k| = q^k$, since a polynomial $p \in \mathcal{P}$ can be specified by k coefficients from $\mathbb{F}_q$. Therefore, H is bijective, and hence if $P \in \mathcal{P}$ is sampled uniformly at random, $H(P)$ is a uniform random vector.

Now let $z_1, \ldots, z_n \in \mathbb{F}_q$ be distinct, and define $G \colon \mathcal{P} \to \mathbb{F}_q^n$ by

$$G(p) = (p(z_1), \ldots, p(z_n)).$$

By the above analysis, when $P \in \mathcal{P}$ is sampled uniformly at random, any k coordinates of $G(P)$ are independent and uniform random.

All that remains is to bridge the gap between field elements and bits. Let q be a power of two, so that field elements can be naturally encoded as bitstrings. The seed of our PRG describes a polynomial $p \in \mathcal{P}$ by giving the encodings of its k coefficients; this requires $k \log q = k \cdot \lceil \log n \rceil$ bits if we pick q to be the smallest power of two that is at least n. The output of our PRG is the sequence of first bits of the encodings of the coordinates of $G(p)$. □

Fooling low-depth decision trees: Every leaf of a decision tree of depth at most k corresponds to a k-junta. Therefore, k-wise uniform random variables perfectly fool decision trees of depth at most k.

Proposition 24.11 (Perfect PRGs for low-depth decision trees)**.** *Let $n, k \in \mathbb{N}$ and let X be a k-wise uniform distribution over $\{0,1\}^n$. Then X perfectly fools depth-k decision trees.*

[1] Unfortunately, in the literature, people often use the term k-wise independence to refer to k-wise uniform. This practice is a little sloppy because it does not clarify the marginal distributions of the individual coordinates of X.

Proof. Let f be a depth-k decision tree. Let $\mathcal{L}$ be the set of accepting leaves of f, i.e., leaves labelled 1. For each leaf $u \in \mathcal{L}$, define $f_u \colon \{0,1\}^n \to \{0,1\}$ by letting $f_u(x) = 1$ iff f arrives at u when it reads x. Note that f_u is a k-junta. Furthermore, we can express f as

$$f(x) = \sum_{u \in \mathcal{L}} f_u(x).$$

Therefore, by linearity of expectation,

$$\mathbb{E}[f(X)] = \mathbb{E}\left[\sum_{u \in A} f_u(X)\right] = \sum_{u \in \mathcal{L}} \mathbb{E}[f_u(X)] = \sum_{u \in \mathcal{L}} \mathbb{E}[f_u] = \mathbb{E}\left[\sum_{u \in \mathcal{L}} f_u\right] = \mathbb{E}[f]. \qquad \square$$

24.4 Almost k-wise uniformity

Let $X \in \{0,1\}^n$ be a random variable and recall the following notions.

- X is ε-Fourier uniform if $|\mathbb{E}[\chi_S(X)]| \leqslant \varepsilon$ for all nonempty $S \subseteq [n]$.
- X is k-wise uniform if $\mathbb{E}[\chi_S(X)] = 0$ for all nonempty $S \subseteq [n]$.

In other words, ε-Fourier uniform PRGs ε-fool all Fourier characters, and k-wise uniform PRGs *perfectly* fool all characters up to level k.

What if we try to ε-fool all characters up to level k? Can we obtain a smaller seed lengths in this case?

Definition 24.12. We say that a random variable $X \in \{0,1\}^n$ is *ε-almost k-wise uniform* if for every nonempty set $S \subseteq [n]$ with $|S| \leqslant k$, we have $|\mathbb{E}[\chi_S(X)]| \leqslant \varepsilon$.

Theorem 24.6 provides an explicit ε-Fourier uniform PRG with seed length $O(\log(n/\varepsilon))$, and Theorem 24.10 provides an explicit k-wise uniform PRG with seed length $O(k \log(n))$, and both of these bounds are optimal.

Theorem 24.13 (almost k-wise uniform generators [NN93])**.** *For every $n, k \in \mathbb{N}$ and every $\varepsilon > 0$, there is an explicit ε-almost k-wise generator with output length n and seed length $O(\log(k/\varepsilon) + \log\log n)$.*

Proof. Let $G \colon \{0,1\}^s \to \{0,1\}^n$ be a k-wise uniform generator that is also a linear transformation when we think of it as a map between vector spaces, $G \colon \mathbb{F}_2^s \to \mathbb{F}_2^n$. One can verify that the k-wise uniform generator that we constructed to prove Theorem 24.10 is indeed a linear transformation. Let Y be an ε-Fourier uniform distribution over $\{0,1\}^s$. We will show that $G(Y)$ fools parities of at most k bits. Indeed, let $f(x) = \sum_{i \in S} x_i$, where $x \in \mathbb{F}_2^n$ and $|S| \leqslant k$. Let $M \in \mathbb{F}_2^{n \times s}$ be the matrix representation of G, with rows $M_1, \ldots, M_n \in \mathbb{F}_2^s$. Then for any $y \in \mathbb{F}_2^s$,

$$f(G(y)) = \sum_{i \in S} \langle M_i, y \rangle = \sum_{i \in S} \sum_{j=1}^{s} M_{ij} y_j = \sum_{j=1}^{s} \left(\sum_{i \in S} M_{ij} \right) y_j.$$

This is a parity function of the variables $y_1, \ldots, y_s$. Therefore, since Y is ε-Fourier uniform,

$$|\,\mathbf{E}[f(G(Y))] - \mathbf{E}[f(G(U))]| \leqslant \varepsilon/2.$$

Furthermore, since G is k-wise uniform and f is a k-junta, $\mathbf{E}[f(G(U))] = \mathbf{E}[f]$. Therefore, $G(Y)$ is k-wise ε-biased. To achieve the promised seed length, we can plug in the constructions of Theorems 24.6 and 24.10 for G and Y, respectively. $\square$

Once again, the seed length of Theorem 24.13 is optimal up to constant factors.

24.5 The sandwiching lemma

Suppose we wish to show that a random variable X fools some class $\mathcal{F}$. A common approach has two steps:

1. Prove that X fools some simpler class $\mathcal{F}'$.

2. Prove a transfer theorem, saying that every distribution that fools $\mathcal{F}'$ also fools $\mathcal{F}$ (possibly with some loss in the error parameter).

Suppose X is a distribution that fools $\mathcal{F}'$, and $\mathcal{F}'$ *approximately* simulates $\mathcal{F}$ in some way. For instance, for every $f \in \mathcal{F}$, there might exist an $f' \in \mathcal{F}'$ such that $\mathbb{E}[|f - f'|]$ is small. However, this alone does not guarantee that X fools $\mathcal{F}$. While f and f' behave similarly under the uniform distribution, it is not obvious whether they behave similarly under the pseudorandom distribution X. A common technique to address this challenge requires a stronger notion of approximation called *sandwiching*.

Definition 24.14 (Sandwiching)**.** Let $f, f_\ell, f_u \colon \{0,1\}^n \to \mathbb{R}$. We say that f is *δ-sandwiched* between f_ℓ and f_u if $f_\ell \leqslant f \leqslant f_u$ and $\mathbb{E}[f_u - f_\ell] \leqslant \delta$.

Lemma 24.15 (Sandwiching Lemma)**.** *Suppose f is δ-sandwiched between f_ℓ and f_u, and suppose X fools f_ℓ and f_u with error ε. Then X fools f with error $\varepsilon + \delta$.*

Proof.

$$\mathbb{E}[f(X)] \leqslant \mathbb{E}[f_u(X)] \leqslant \mathbb{E}[f_u] + \varepsilon \leqslant \mathbb{E}[f] + \varepsilon + \delta$$
$$\mathbb{E}[f(X)] \geqslant \mathbb{E}[f_\ell(X)] \geqslant \mathbb{E}[f_\ell] - \varepsilon \geqslant \mathbb{E}[f] - \varepsilon - \delta.$$

□

To illustrate the sandwiching technique, let us return to the decision tree model. Recall that we showed that k-wise uniform generators fool depth-k decision trees (Proposition 24.11). We now show that k-wise uniform generators also fool bounded-*size* decision trees.

Proposition 24.16 (Limited independence fools bounded-size decision trees)**.** *If X is a k-wise uniform distribution, then X fools size-m decision trees with error $m \cdot 2^{-k}$.*

Proof. Let f be a size-m decision tree. Define a depth-k decision tree f_ℓ by starting with f and replacing each internal node at depth exactly k with a leaf labelled 0 and deleting all of its descendants. Similarly, define f_u by replacing each internal node at depth k with a leaf labelled 1. Let us show that f is δ-sandwiched between f_ℓ and f_u, for $\delta = m \cdot 2^{-k}$.

Clearly $f_\ell \leqslant f \leqslant f_u$. For each "new" leaf u of f_ℓ or f_u (i.e., u was not a leaf in f), the probability of reaching u on a uniform random input is precisely 2^{-k}. The number of new leaves is the number of internal nodes of f at depth k, which is at most m. Therefore, by the union bound, $\mathbb{E}[f_u - f_\ell] \leqslant m \cdot 2^{-k}$.

The Sandwiching Lemma completes the proof because X fools f_ℓ and f_u with error 0. □

Remark 24.17. Proposition 24.16 implies that using k-wise uniform generators, we can ε-fool size-m decision trees using a seed of length $O(\log(m/\varepsilon) \cdot \log n)$. This seed length is *inferior* to the seed length that we can obtain from Proposition 24.7 using the fact that $\|f\|_A \leqslant m$. However, sometimes, it is useful to understand the effect of specific classes of distributions, such as k-wise uniform distributions, on a given model of computation.

24.6 Braverman's theorem: Poly-logarithmic independence fools AC^0

In this section, we will show that every k-wise uniform generator fools constant-depth polynomial-size AC circuits for a suitable $k = \mathrm{polylog}(n)$. This result was first conjectured by Linial and Nisan [LN90]. Two decades later, Bazzi [Baz09] proved it true for depth-2 circuits, with a simpler proof subsequently provided by Razborov [Raz09]. Finally, Braverman [Bra10] proved that k-wise independence for polylogarithmic k fools AC^0 circuits. Further improvements to the parameters were made by Tal [Tal17] and Harsha and Srinivasan [HS19].

The proof of Braverman's theorem hinges on the two classical approximation of AC^0 circuits by low degree polynomials, Theorem 22.2 of Linial, Mansour, and Nisan [LMN93] and Theorem 22.4 of Razborov and Smolensky [Raz87, Smo87]. The bounds in both theorems have been improved recently, and we will need these stronger bounds to obtain the promised bound of $(\log m)^{O(d)} \cdot \log(1/\varepsilon)$ in Theorem 24.20.

Theorem 24.18 (Improved Razborov-Smolensky [HS19])**.** *Let μ be a probability distribution over $\{0,1\}^n$, and let $f : \{0,1\}^n \to \{0,1\}$ be computed by an AC circuit of depth d and size M. For every $\delta > 0$, there is a polynomial $g : \{0,1\}^n \to \mathbb{R}$ with*

1. *(error bound)*
$$\Pr_{\mathbf{x}\sim\mu}[f(\mathbf{x}) \neq g(\mathbf{x})] \leqslant \delta.$$

2. *(degree bound)*
$$\deg(g) = (\log M)^{O(d)} \log(1/\delta).$$

3. *(boundedness)*
$$\|g\|_\infty \leqslant 2^{(\log M)^{O(d)} \log(1/\delta)}.$$

4. *(Error detection) Moreover, there exists an* AC *circuit* E *with depth* $d + O(1)$ *and size* $M^{O(1)}$ *such that*
$$f(x) \neq g(x) \iff E(x) = 1.$$

Perhaps the most mysterious part of Theorem 24.18 is the fourth assertion regarding the existence of a small circuit that can inform us whether the approximation g matches the value of $f(x)$. To verify this, recall the proof of Theorem 22.4, where we recursively replaced each $\wedge$ and $\vee$ gate with a simple random polynomial. The errors introduced in these replacements are localized, and a small depth 2 circuit can check whether the polynomial approximation accurately computes the desired value at such a gate.

Theorem 24.19 (Improved LMN bound [Tal17])**.** *Let* $f\colon \{0,1\}^n \to \{0,1\}$ *be computable by an* AC *circuit of depth* d *and size* M*, and let* $\gamma > 0$*. There exists* $g\colon \{0,1\}^n \to \mathbb{R}$ *such that:*

1. *(L_2 approximation)*
$$\|f - g\|_2 \leqslant \gamma.$$

2. *(degree bound)*
$$\deg(g) \leqslant O(\log M)^{d-1} \cdot \log(1/\gamma).$$

With Theorem 24.18 and Theorem 24.19 in hand, we can prove Braverman's theorem.

Theorem 24.20 (Braverman's theorem [Bra10, Tal17, HS19])**.** *For every* $n, M, d \in \mathbb{N}$ *and* $\varepsilon > 0$*, there is a value*
$$k = (\log M)^{O(d)} \cdot \log(1/\varepsilon)$$
such that if X *is a random variable with a* k*-wise uniform distribution over* $\{0,1\}^n$*, then it* ε*-fools any* AC *of size* $M \geqslant n$ *and depth* d*. Consequently, there is an explicit* ε*-PRG for* AC *circuits of size* $M \geqslant n$ *and depth* d *that has seed length*
$$(\log M)^{O(d)} \log(1/\varepsilon).$$

Proof. Let $f : \{0,1\}^n \to \{0,1\}$ be the function computed by an AC circuit of size M and depth d, and let X be a random variable with a k-wise uniform distribution μ_X over $\{0,1\}^n$ with the specified k. Let ν be the uniform distribution on $\{0,1\}^n$. We need to show that $|\mathbb{E}_{\mu_X}[f] - \mathbb{E}_\nu[f]| \leqslant \varepsilon$. Since one can replace f with $\neg f$ in the statement of the theorem, without loss of generality, it suffices to prove
$$\mathbb{E}_\nu[f] - \mathbb{E}_{\mu_X}[f] \leqslant \varepsilon. \tag{24.1}$$

Define $\mu \coloneqq \frac{1}{2}(\mu + \nu)$, and let $\widetilde{f}$ be the corresponding polynomial approximation for f from Theorem 24.18, and let $E(x) \coloneqq \mathbf{1}_{[f(x)\neq\widetilde{f}(x)]}$ be the corresponding error function satisfying $\mathbb{E}_\mu[E] \leqslant \delta \coloneqq \frac{\varepsilon}{8}$. We have
$$\deg(\widetilde{f}) \leqslant O(\log M)^{d-1} \cdot \log(1/\delta) \quad \text{and} \quad \left\|\widetilde{f}\right\|_\infty \leqslant 2^{(\log M)^{O(d)} \log(1/\varepsilon)},$$
and by the definition of μ,
$$\mathbb{E}_{\mu_X}[E] \leqslant 2\delta = \frac{\varepsilon}{4} \quad \text{and} \quad \mathbb{E}_\nu[E] \leqslant \frac{\varepsilon}{4}.$$
Define $F \coloneqq f \vee E$. Since $E(x) = 0$ implies $f(x) = F(x)$, we have
$$|\mathbb{E}_{\mu_X}(F) - \mathbb{E}_{\mu_X}[f]| \leqslant \mathbb{E}_{\mu_X}(E) \leqslant \frac{\varepsilon}{4} \text{ and } |\mathbb{E}_\nu(F) - \mathbb{E}_\nu[f]| \leqslant \mathbb{E}_\nu(E) \leqslant \frac{\varepsilon}{4}.$$

Thus, to prove Equation (24.1), it suffices to prove

$$\mathbb{E}_\nu[F] - \mathbb{E}_{\mu_X}[F] \leqslant \frac{\varepsilon}{3}. \tag{24.2}$$

We will prove (24.2) by providing an appropriate lower sandwiching polynomial p_ℓ.

Since the error function E is computable by an AC circuit of size $M^{O(1)}$ and depth $d + O(1)$, we may apply Theorem 24.19 to obtain a polynomial approximation $\widetilde{E}$ that satisfies $\left\|E - \widetilde{E}\right\|_2 \leqslant \gamma$ for an error parameter γ that will be specified later. Define

$$q \coloneqq \widetilde{f}(1 - \widetilde{E}) \quad \text{and} \quad p_\ell \coloneqq 1 - (1 - q)^2.$$

See Figure 24.1 for an illustration of these functions.

Claim 24.21. *We have $p_\ell \leqslant F$. Furthermore, for sufficiently small $\gamma = 2^{-(\log M)^{O(d)} \log(1/\varepsilon)}$, we have $\|F - p_\ell\|_1 \leqslant \frac{\varepsilon}{3}$.*

Proof. First, we show that for every x, we have $p_\ell(x) \leqslant F(x)$. If $F(x) = 1$, then $p_\ell(x) \leqslant 1 \leqslant F(x)$. If $F(x) = 0$, then $f(x) = E(X) = 0$, which implies $0 = \widetilde{f}(x) = p_\ell(x) = q(x)$. The latter also shows

$$\|F - p_\ell\|_1 = \mathbb{E}_\nu|F - p_\ell| = \mathbb{E}_\nu|F - p_\ell|\mathbf{1}_{[F=1]} = \mathbb{E}_\nu|1 - q|^2\mathbf{1}_{[F=1]} \leqslant \mathbb{E}_\nu|F - q|^2 = \|F - q\|_2^2.$$

On the other hand, by the triangle inequality, we have

$$\begin{aligned}\|F - q\|_2 &\leqslant \left\|F - \widetilde{f}(1 - E)\right\|_2 + \left\|\widetilde{f}(1 - E) - \widetilde{f}(1 - \widetilde{E})\right\|_2 \leqslant \sqrt{\mathop{\mathbf{Pr}}_\nu[E = 1]} + \left\|\widetilde{f}\right\|_\infty \left\|E - \widetilde{E}\right\|_2 \\ &\leqslant \sqrt{\frac{\varepsilon}{4}} + \left\|\widetilde{f}\right\|_\infty \gamma \leqslant \sqrt{\frac{\varepsilon}{4}} + 2^{(\log M)^{O(d)} \log(1/\varepsilon)}\gamma \leqslant \sqrt{\frac{\varepsilon}{3}},\end{aligned}$$

provided that $\gamma = 2^{-(\log M)^{O(d)} \log(1/\varepsilon)}$ is sufficiently small. □

Our choice of γ yields

$$\deg(p_\ell) \leqslant 2\deg(\widetilde{f})\deg(\widetilde{E}) \leqslant (\log M)^{O(d)} \log(1/\varepsilon).$$

Taking $k = \deg(p_\ell) + 1$, we have $\mathbb{E}_{\mu_X} p_\ell = \mathbb{E}_\nu p_\ell$. Therefore, by the Claim 24.21,

$$\mathbb{E}_\nu[F] - \mathbb{E}_{\mu_X}[F] \leqslant \mathbb{E}_\nu[F] - \mathbb{E}_{\mu_X}[p_\ell] \leqslant \frac{\varepsilon}{3} + \mathbb{E}_\nu[p_\ell] - \mathbb{E}_{\mu_X}[p_\ell] = \frac{\varepsilon}{3},$$

which verifies Equation (24.2), and completes the proof. □

Figure 24.1: TO DO: Braverman's functions.

Braverman's theorem represents neither the first nor the strongest known unconditional PRG for AC^0. Instead, the advantage of Braverman's theorem is that k-wise uniformity is a particularly *simple and general* PRG construction.

It is an open problem to improve the parameters of Braverman's theorem even further and find the optimal k such that every k-wise uniform random variable ε-fools AC circuits of size M and depth d. There are counterexamples [LV96] showing that $k = \Omega((\log m)^{d-1} \log(1/\varepsilon))$, but that still leaves a significant gap between the lower and upper bounds.

Conjecture 24.22 (Improved Braverman's theorem)**.** *For every $m, d \in \mathbb{N}$ and $\varepsilon > 0$, there exists a value*

$$k = (\log M)^{d+O(1)} \log(1/\varepsilon)$$

such that every k-wise uniform distribution ε-fools AC circuits of size M and depth d.

Explicit PRGs for AC^0 circuits with seed length $(\log M)^{d+O(1)} \log(1/\varepsilon)$ are already known [TX13, Tal17, ST19, Kel21, Lyu22]; the question is whether a generic k-wise uniform generator suffices.

Chapter 25

PRGs from polarizing random walks

In chapter chapter, we will discuss a new method due to [CHHL19b] for constructing pseudorandom generators that uses the novel idea of polarizing random walks. The applications of these ideas go beyond the construction of pseudorandom generators; for instance, Raz and Tal [RT22] utilized these techniques to establish their groundbreaking result on the oracle separation of the quantum complexity class BQP and the polynomial hierarchy PH.

Similar to Chapter 24, most of the content of this chapter is taken from the survey [HH24].

25.1 Fractional PRGs

For the purposes of this chapter, it will be convenient to work with functions $f\colon \{-1,1\}^n \to \mathbb{R}$ instead of the domain $\{0,1\}^n$. The strategy of [CHHL19b] is to first design a relaxation of a PRG, called a *fractional PRG*, that takes values in the continuous cube $[-1,1]^n$. Then, we will use the polarized random walk to round the fractional PRG into a genuine PRG in $\{-1,1\}^n$.

The Fourier expansion of f naturally extends f to a function $f : [-1,1]^n \to \mathbb{R}$ as follows.

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \prod_{i \in S} x_i \qquad \text{for } x \in [-1,1]^n.$$

We can naturally define the fractional PRGs using this extension.

Definition 25.1 (Fractional PRGs)**.** Let $f\colon \{-1,1\}^n \to \mathbb{R}$, and extend f to the domain $[-1,1]^n$ via the Fourier expansion. A random variable $X \in [-1,1]^n$ is said to *ε-fool f* if

$$|\mathbb{E}[f(X)] - \mathbb{E}[f]| \leqslant \varepsilon.$$

A *fractional ε-PRG* with seed length s for a family $\mathcal{F}$ of Boolean functions is a function $G\colon \{-1,1\}^s \to [-1,1]^n$ such that $G(U_s)$ fools every $f \in \mathcal{F}$ with error ε, where U_s is the uniform random variable over $\{-1,1\}^s$.

Since $\mathbb{E}[f] = f(\vec{0})$, the trivial random variable that always takes value $\vec{0} \in [-1,1]^n$ perfectly fools every function f. Obviously, this construction is too trivial and useless for constructing genuine PRGs. To address this issue, we focus on random variables $X \in [-1,1]^n$ where each X_i has a large variance. We will show how to transform such random variables into $\{-1,1\}^n$-valued random variables with comparable fooling properties.

Definition 25.2 (Noticeability)**.** We say that a random variable $X \in [-1,1]^n$ is *q-noticeable* for a parameter $q \geqslant 0$, if for every $i \in [n]$, we have $\mathbb{E}[X_i^2] \geqslant q$.

25.2 From fractional PRGs to PRGs

In this section, we will show that it is possible to transform fractional PRGs into PRGs. We will require the fractional PRG to be symmetric.

Definition 25.3 (Symmetric random variables)**.** Let X be a random variable distributed over $[-1,1]^n$. We say that X is *symmetric* if for every $x \in [-1,1]^n$, we have $\mathbf{Pr}[X = x] = \mathbf{Pr}[X = -x]$.

In addition to the symmetry condition, it will also be essential that the PRG not only fools $f \in \mathcal{F}$, but it also fools all the restrictions of f. Consequently, in the following theorem, we require that the class $\mathcal{F}$ is closed under restrictions.

Theorem 25.4 (Fractional PRG to PRG [CHHL19b])**.** *Suppose that $\mathcal{F}$ is a family of functions $f\colon \{-1,1\}^n \to \{0,1\}$ that is closed under restrictions. Assume there exists an explicit q-noticeable symmetric fractional PRG for $\mathcal{F}$ with error ε and seed length s. Then there exists an explicit PRG for $\mathcal{F}$ with seed length $O(s \cdot q^{-1} \cdot \log(n/\varepsilon))$ and error $O(\varepsilon \cdot q^{-1} \cdot \log(n/\varepsilon))$.*

To prove Theorem 25.4, we will take a *random walk* through the solid hypercube $[-1,1]^n$. The construction of the random walk is quite natural. We will take $Y^{(0)} = 0^n$ as the starting point, since $\mathbb{E}[f] = f(0^n)$. We wish to have a random walk that converges quickly to the Boolean cube $\{-1,1\}^n$, while each step does not incur much error. The fractional PRG provides us with the first step of the random walk. If we set $Y^{(1)} = Y^{(0)} + X = X$, then $\mathbb{E}[f(X)] \approx \mathbb{E}[f]$ since X fools f. For the next step, naturally, we wish to take another step using the fractional PRG and set $Y^{(2)} = Y^{(1)} + X'$ where X' is an i.i.d. copy of X. However, since $Y^{(1)} + X'$ might fall outside of the cube $[-1,1]^n$, we have to scale the coordinates of X' to guarantee that $Y^{(1)} + X'$ remains in the cube $[-1,1]^n$. See Figure 25.1.

Figure 25.1: TO DO: polarized random walk.

For two vectors $x, x' \in [-1,1]^n$, define $x \odot x' \in [-1,1]^n$ to be their coordinate-wise product. Moreover, for every vector $y \in [-1,1]^n$ define $d_y \in [0,1]^n$ to be the vector with i-th coordinate $(d_y)_i = 1 - |y_i|$, i.e., $(d_y)_i$ is the distance from $y_i \in [-1,1]$ to the Boolean endpoints $\{-1,1\}$. The vector d_y defines the dimensions of the largest subcube inside $[-1,1]^n$ centred at y. Using this notation, we can now define the random walk. let $X^{(1)}, \ldots, X^{(t)}$ be t independent samples of X where t is to be determined later.

- $Y^{(0)} = 0^n$, and
- For $j > 0$, let $Y^{(j)} = Y^{(j-1)} + d_{Y^{(j-1)}} \odot X^{(j)}$.

We will show that this random walk quickly approaches $\{-1,1\}^n$. Still, there is a chance that the coordinates of $Y^{(t)}$ are never *exactly* integers. The final construction takes care of this by outputting the coordinate-wise signs of $Y^{(t)}$. To this end, for $x \in \mathbb{R}^n$ define $\mathrm{sgn}(x) \in \{-1,1\}^n$ to be the vector with i-th coordinate $\mathrm{sgn}(x)_i = 1 \iff x_i > 0$.

The Generator G:

1. Let $X_1, \ldots, X_t$ be independent copies of X for a suitable value $t = O(q^{-1} \cdot \log(n/\varepsilon))$
2. Let $Y^{(0)} = 0^n$, and for $j > 0$ define $Y^{(j)} = Y^{(j-1)} + d_{Y^{(j-1)}} \odot X^{(j)}$
3. Output $\mathrm{sgn}(Y^{(t)})$

25.2.1 Analysis of the random walk

To prove the correctness of the generator G, we will prove that the random walk has three properties:

(a) Each step introduces little error: For every $f \in \mathcal{F}$ and $j \in [t]$,

$$\left| \mathbb{E}\left[f(Y^{(j)}) \right] - \mathbb{E}\left[f(Y^{(j+1)}) \right] \right| \leqslant \varepsilon.$$

(b) The walk *polarizes* with high probability:

$$\mathbf{Pr}[\|d_{Y^{(t)}}\|_\infty \leqslant \varepsilon/n] \geqslant 1 - \varepsilon.$$

(c) The final rounding operation introduces little error: For every $f \in \mathcal{F}$, conditioned on polarization,

$$|f(Y^{(t)}) - f(\operatorname{sgn}(Y^{(t)}))| \leqslant \varepsilon.$$

We prove these properties in the next three lemmas.

Lemma 25.5 (Steps incur small error)**.** *Let $\mathcal{F}$ be a family of functions $f\colon \{-1,1\}^n \to \mathbb{R}$ that is closed under restrictions, and suppose $X \in [-1,1]^n$ fools $\mathcal{F}$ with error ε. Then for every $f \in \mathcal{F}$ and $y \in [-1,1]^n$,*

$$|f(y) - \mathbb{E}[f(y + d_y \odot X)]| \leqslant \varepsilon.$$

In particular, for every $j \in [t]$,

$$\left|\mathbb{E}\left[f(Y^{(j)})\right] - \mathbb{E}\left[f(Y^{(j+1)})\right]\right| \leqslant \varepsilon.$$

Proof. Let $y \in [-1,1]^n$ be fixed. Sample a random restriction $\boldsymbol{\rho} \in \{-1,1,\star\}^n$, independent of X, where the coordinates of $\boldsymbol{\rho}$ are independent and distributed as follows:

$$\boldsymbol{\rho}_i = \begin{cases} \operatorname{sgn}(y_i) & \text{with probability } |y_i| \\ \star & \text{with probability } 1 - |y_i|. \end{cases}$$

for $x \in [-1,1]^n$, let $\boldsymbol{\rho} \circ x$ be defined by filling in the $\star$ coordinates of $\boldsymbol{\rho}$ using x. That way, for each coordinate $i \in [n]$, we have

$$\mathbb{E}_{\boldsymbol{\rho}}[(\boldsymbol{\rho} \circ x)_i] = |y_i| \cdot \operatorname{sgn}(y_i) + (1 - |y_i|) \cdot x_i = y_i + (1 - |y_i|) \cdot x_i,$$

and hence overall,

$$\mathbb{E}_{\boldsymbol{\rho}}[\boldsymbol{\rho} \circ x] = y + d_y \odot x.$$

It follows that

$$\mathbb{E}_{\boldsymbol{\rho}}[f_{\boldsymbol{\rho}}(x)] = \mathbb{E}_{\boldsymbol{\rho}}[f(\boldsymbol{\rho} \circ x)] = f\left(\mathbb{E}_{\boldsymbol{\rho}}[\boldsymbol{\rho} \circ x]\right) = f(y + d_y \odot x).$$

Consequently,

$$|f(y) - \mathbb{E}_X[f(y + d_y \odot X)]| = |\mathbb{E}_{\boldsymbol{\rho}}[f_{\boldsymbol{\rho}}(0^n)] - \mathbb{E}_{\boldsymbol{\rho},X}[f_{\boldsymbol{\rho}}(X)]| \leqslant \mathbb{E}_{\boldsymbol{\rho}}\left[|f_{\boldsymbol{\rho}}(0^n) - \mathbb{E}_X[f_{\boldsymbol{\rho}}(X)]|\right] \leqslant \varepsilon,$$

where the last step uses that $\mathcal{F}$ is closed under restriction, and hence X fools $f_{\boldsymbol{\rho}}$ with error ε. □

Next, we will show that the random walk quickly converges to $\{-1,1\}^n$. For this argument, we need the assumption that X is q-noticeable for a large enough $q > 0$ and symmetric. The symmetry assumption is helpful because of the following lemma concerning the case $n = 1$.

Lemma 25.6. *Let $X \in [-1,1]$ be a symmetric q-noticeable random variable. Then*

$$\mathbb{E}\left[\sqrt{1 - X}\right] \leqslant 1 - q/8.$$

Proof. Let $Y = |X|$, and sample $Z \in \{\pm 1\}$ independently of X. Then the product YZ is distributed identically to X. Furthermore, for each fixed value $y \in [0,1]$, we have

$$\left(\mathbb{E}\left[\sqrt{1 - yZ}\right]\right)^2 = \left(\frac{\sqrt{1-y} + \sqrt{1+y}}{2}\right)^2 = \frac{1 + \sqrt{1-y^2}}{2} \leqslant 1 - \frac{y^2}{4}.$$

Therefore,

$$\mathbb{E}\left[\sqrt{1 - X}\right] = \mathbb{E}_Y\left[\mathbb{E}_Z\left[\sqrt{1 - YZ}\right]\right] \leqslant \mathbb{E}_Y\left[\sqrt{1 - Y^2/4}\right] \leqslant \mathbb{E}_Y[1 - Y^2/8] \leqslant 1 - q/8.$$ □

Next, let us use Lemma 25.6 to show that coordinate-wise polarization happens with high probability. Indeed, looking ahead, the probability will be high enough to allow a union bound over all coordinates.

Lemma 25.7 (Key observation: Polarization)**.** *Let $A^{(1)}, \dots, A^{(t)} \in [-1,1]$ be independent symmetric q-noticeable random variables. Define $B^{(0)} \coloneqq 0$, and for $j > 0$, define*

$$B^{(j)} \coloneqq B^{(j-1)} + (1 - |B^{(j-1)}|) \cdot A^{(j)}. \tag{25.1}$$

Then $\mathbf{Pr}[1 - |B^{(t)}| \geqslant e^{-tq/8}] \leqslant e^{-tq/16}$.

Proof. Let us investigate the change in the distance $1 - |B^{(\cdot)}|$ when we apply the update rule (25.1). If $\operatorname{sgn}(A^{(j)}) = \operatorname{sgn}(B^{(j-1)})$ (the "good case"), the distance decreases by a factor of $1 - |A^{(j)}|$. If $\operatorname{sgn}(A^{(j)}) \neq \operatorname{sgn}(B^{(j-1)})$ (the "bad case"), the distance might increase, but at most, it increases by a factor of $1 + |A^{(j)}|$. Either way, for $j > 0$, we have

$$1 - |B^{(j)}| \leqslant (1 - |B^{(j-1)}|) \cdot (1 - A^{(j)} \cdot \operatorname{sgn}(B^{(j-1)})).$$

We have assumed that $A^{(1)}, \dots, A^{(j-1)}$ are symmetric. It follows that $B^{(j-1)}$ is also symmetric. Therefore, $|B^{(j-1)}|$ and $A^{(j)} \cdot \operatorname{sgn}(B^{(j-1)})$ are independent. As a consequence,

$$\mathbb{E}\left[\sqrt{1 - |B^{(j)}|}\right] \leqslant \mathbb{E}\left[\sqrt{1 - |B^{(j-1)}|}\right] \cdot \mathbb{E}\left[\sqrt{1 - A^{(j)} \cdot \operatorname{sgn}(B^{(j-1)})}\right].$$

The random variable $A^{(j)} \cdot \operatorname{sgn}(B^{(j-1)})$ is symmetric and q-noticeable, so we may apply Lemma 25.6, giving us

$$\mathbb{E}\left[\sqrt{1 - |B^{(j)}|}\right] \leqslant \mathbb{E}\left[\sqrt{1 - |B^{(j-1)}|}\right] \cdot (1 - q/8).$$

By induction, this implies

$$\mathbb{E}\left[\sqrt{1 - |B^{(t)}|}\right] \leqslant (1 - q/8)^t \leqslant e^{-qt/8}.$$

The lemma follows from Markov's inequality. □

We show that the final rounding step does not introduce too much error.

Lemma 25.8 (Rounding Error)**.** *Let $f\colon \{-1,1\}^n \to \{0,1\}$ be a function, and extend it to the domain $[-1,1]^n$ via the Fourier expansion. For every $y \in [-1,1]^n$,*

$$|f(y) - f(\operatorname{sgn}(y))| \leqslant \sum_{i=1}^{n} (1 - |y_i|) \leqslant n \cdot \|d_y\|_\infty.$$

Proof. For $y \in [-1,1]^n$, let $\Pi_y \in \{-1,1\}^n$ be the random variable with independent coordinates satisfying $\mathbb{E}[\Pi_y] = y$. We have

$$|f(y) - f(\operatorname{sgn}(y))| = |\mathbb{E}[f(\Pi_y)] - f(\operatorname{sgn}(y))| \leqslant \mathbf{Pr}[\Pi_y \neq \operatorname{sgn}(y)] \leqslant \sum_{i=1}^{n} \frac{1 - |y_i|}{2},$$

where the final inequality follows from the union bound. □

We can now analyze G and complete the proof of Theorem 25.4. The output of the generator G is $\operatorname{sgn}(Y^{(t)})$ for $t = 16 \log(n/\varepsilon)/q$. The seed for G is determined by t independent samples from the fractional generator, and hence has seed-length $ts = O(s \log(n/\varepsilon)/q)$. Let E denote the event that $\|d_{Y^{(t)}}\|_\infty \leqslant e^{-tq/8} \leqslant \varepsilon/n$. Then, we can bound the error of the generator $\operatorname{sgn}(Y^{(t)})$ as follows:

$$\begin{aligned}
|\mathbb{E}[f] - \mathbb{E}[f(\operatorname{sgn}(Y^{(t)}))]| &\leqslant |\mathbb{E}[f(\operatorname{sgn}(Y^{(t)}))] - \mathbb{E}[f(Y^{(t)})]| + \sum_{j=1}^{t} |\mathbb{E}[f(Y^{(j)})] - \mathbb{E}[f(Y^{(j-1)})]| \\
&\leqslant |\mathbb{E}[f(\operatorname{sgn}(Y^{(t)})) - f(Y^{(t)}) \mid \mathrm{E}]| + 2\,\mathbf{Pr}[\mathrm{E}] + \varepsilon t && \text{(by Lemma 25.5)} \\
&\leqslant n\frac{\varepsilon}{n} + 2n \cdot e^{-tq/16} + \varepsilon t \leqslant O(\varepsilon \log(n/\varepsilon)/q). && \text{(by Lemmas 25.7 and 25.8)}
\end{aligned}$$

25.3 Constructing fractional PRGs

Now that we know how to convert a fractional PRG with noticeable coordinates to a standard PRG let's discuss the construction of fractional PRGs.

25.3.1 Fractional PRGs from random restrictions

Let $\mathcal{F}$ be a class of functions $f\colon \{-1,1\}^n \to \{0,1\}$ that we wish to fool. Suppose we have shown that functions in our class $\mathcal{F}$ simplify under random restrictions. Let $\boldsymbol{\rho}_p$ denote the random restriction that sets every variable independently to $\star$ with probability p, and to 0 and 1 each with probability $\frac{1-p}{2}$. Suppose we have identified a class $\mathcal{F}_{\text{simp}}$ of simpler functions and values $p, \delta > 0$ such that for each $f \in \mathcal{F}$, we have

$$\mathbf{Pr}[f_{\boldsymbol{\rho}_p} \in \mathcal{F}_{\text{simp}}] \geqslant 1 - \delta.$$

The following lemma shows that if random restrictions simplify a class $\mathcal{F}$ to another class $\mathcal{F}_{\text{simp}}$ for which we have good PRGs, then we obtain a good *fractional* PRG for the original class $\mathcal{F}$.

Lemma 25.9 (Simplification implies fractional PRGs)**.** *Let $\mathcal{F}$ and $\mathcal{F}_{\text{simp}}$ be classes of functions $f\colon \{-1,1\}^n \to \{0,1\}$. Let $p, \delta > 0$, and suppose that for each $f \in \mathcal{F}$, we have*

$$\mathbf{Pr}[f_{\boldsymbol{\rho}_p} \in \mathcal{F}_{\text{simp}}] \geqslant 1 - \delta.$$

Let X be a distribution over $\{-1,1\}^n$ that ε-fools $\mathcal{F}_{\text{simp}}$. Then pX is p^2-noticeable and fools $\mathcal{F}$ with error $\varepsilon + 2\delta$.

Proof. Since the coordinates of pX take $\pm p$-values, it is trivially p^2-noticeable. For each fixed string $x \in \{-1,1\}^n$, the composition $\boldsymbol{\rho}_p \circ x$ has a product distribution over $\{-1,1\}^n$, where

$$\mathbb{E}[(\boldsymbol{\rho}_p \circ x)_i] = (1-p)0 + px_i = p \cdot x_i.$$

Therefore, $\mathbb{E}[f(\boldsymbol{\rho}_p \circ x)] = f(px) = T_p f(x)$ where T_p is the noise operator. Consequently,

$$\mathbb{E}_X[f(pX)] = \mathbb{E}_{\boldsymbol{\rho}_p, X}[f(\boldsymbol{\rho}_p \circ X)].$$

Since X fools the functions in $\mathcal{F}_{\text{simp}}$ with error ε, we have

$$\left|\mathbb{E}_{\boldsymbol{\rho}_p, X}[f(\boldsymbol{\rho}_p \circ X)] - \mathbb{E}[f]\right| = \mathbf{Pr}[f_{\boldsymbol{\rho}_p} \in \mathcal{F}_{\text{simp}}]\varepsilon + \mathbf{Pr}[f_{\boldsymbol{\rho}_p} \notin \mathcal{F}_{\text{simp}}] \leqslant (1-\delta)\varepsilon + \delta \leqslant \varepsilon + 2\delta.$$

□

We saw in Corollary 21.11, that as a consequence of Håstad's switching lemma, if $f\colon \{0,1\}^n \to \{0,1\}$ is computable by an AC circuit of depth d and size M, then for $p = 1/\Theta_\delta(\log M)^{d-1}$, we have

$$\mathbf{Pr}[\mathrm{dt}(f_{\boldsymbol{\rho}_p}) \leqslant \log(M/\delta)] \geqslant 1 - \delta.$$

Combining with the PRGs for low-depth decision trees, we obtain the following fractional PRG for AC^0.

Corollary 25.10 (Fractional PRGs for AC^0)**.** *For every $n, m, d \in \mathbb{N}$ and every $\varepsilon > 0$, there is $q = 1/\Theta_q(\log M)^{2d-2}$ and an explicit q-noticeable fractional PRG with seed length $O(\log(1/\varepsilon) + \log\log n)$ that ε-fools AC circuits of depth d and size M.*

For example, by combining Theorem 25.4 and Corollary 25.10, we get the following PRG for AC^0.

Corollary 25.11 (PRG for AC^0 based on simplification under truly random restrictions)**.** *For every $n, m, d \in \mathbb{N}$ and $\varepsilon > 0$, there is an explicit ε-PRG for depth-d size-m AC^0 circuits on n input bits that has seed length*

$$\tilde{O}(\log m)^{2d-2} \cdot \tilde{O}(\log(n/\varepsilon) \cdot \log(1/\varepsilon)).$$

25.3.2 Fractional PRGs from Fourier growth

Recall from Chapter 23 that the Fourier algebra norm $\|f\|_A := \left\|\widehat{f}\right\|_1$ is the L_1-sum of the absolute values of the Fourier coefficients. Given a function $f : \{-1,1\}^n \to \mathbb{R}$ and $t \in [n]$, define

$$L_{1,t}(f) = \left\|f^{=t}\right\|_A = \sum_{\substack{S \subseteq [n] \\ |S|=t}} \left|\widehat{f}(S)\right|.$$

Tail bounds on $L_{1,d}(f)$ are quite useful in complexity theory and theory of learning and they have been established for several classes of functions. For example, [LPV22] recently showed that the so-called *regular read-once branching programs* of width w satisfy $L_{1,t}(f) \leqslant (w-1)^t$. Similarly, Tal [Tal17] proved the following tail bound for AC^0 circuits.

Theorem 25.12 (Tal [Tal17])**.** *Let $f : \{-1,1\}^n \to \{0,1\}$ be a Boolean function computed by an AC circuit of depth d and size M, and let t be any integer. Then*

$$L_{1,t}(f) \leqslant O(\log(M)^{d-1})^t.$$

The following proposition shows that tail bounds on $L_{1,t}(f)$ imply the existence of explicit noticeable fractional PRGs against f.

Proposition 25.13 (Fourier growth and fractional PRGs)**.** *Let $a, b > 0$ and $\varepsilon > 0$ be constants and suppose $f : \{-1,1\}^n \to \mathbb{R}$ satisfies the tail bound*

$$L_{1,t}(f) \leqslant a \cdot b^t \tag{25.2}$$

for every $t \in [n]$. Let $X \in \{-1,1\}^n$ be a δ-almost k-wise uniform random variable where

$$\delta := \frac{\varepsilon}{2a} \quad \text{and} \quad k := \left\lceil \log\left(\frac{2a}{\varepsilon}\right)\right\rceil.$$

Then for $p := \frac{1}{2b}$, the random variable pX is p^2-noticeable and it ε-fools f.

Proof. Note $\mathbb{E}[f(pX)] = T_p f(X)$ and

$$|\mathbb{E}[T_p f(X)] - \mathbb{E}[f]| = \left|\sum_{\substack{S \subseteq [n] \\ S \neq \emptyset}} p^{|S|} \widehat{f}(S) \mathbb{E}\chi_S(X)\right| \leqslant \sum_{\substack{S \subseteq [n] \\ S \neq \emptyset}} p^{|S|} \left|\widehat{f}(S)\right| \cdot |\mathbb{E}\chi_S(X)|.$$

When $|S| \leqslant k$, we have $|\mathbb{E}\chi_S(X)| \leqslant \delta$. When $|S| > k$, we use the trivial bound $|\mathbb{E}\chi_S(X)| \leqslant 1$. Plugging these bounds into the above inequality, we get

$$|\mathbb{E}[T_p f(X)] - \mathbb{E}[T_p f]| \leqslant \delta \cdot a \cdot \sum_{t=1}^{k} (pb)^t + a \cdot \sum_{t=k+1}^{n} (pb)^t \leqslant \delta a + 2^{-k} a \leqslant \varepsilon.$$

□

We can use the polarized random walk to convert the fractional PRG of Proposition 25.13 into a genuine PRG.

Corollary 25.14. *Let $a, b > 0$ and $\varepsilon > 0$ be constants, and let $\mathcal{F}$ be a class of Boolean functions $f : \{-1,1\}^n \to \{0,1\}$. Suppose $\mathcal{F}$ is closed under restrictions, and every $f \in \mathcal{F}$ satisfies $L_{1,t}(f) \leqslant a \cdot b^t$ for every $t \in [n]$. There is an explicit ε-PRG for f with seed length*

$$O(b^2 \cdot \log n \cdot (\log(ab/\varepsilon) + \log\log n)).$$

Proof. Apply Proposition 25.13 with a sufficiently small error parameter $\varepsilon_1 = O(\frac{\varepsilon}{b^2 \log(n) \log(b/\varepsilon)})$. Let $p = \frac{1}{2b}$ and

$$\delta = \varepsilon_1/(2a) \quad \text{and} \quad k = \lceil \log(2a/\varepsilon_1) \rceil$$

as in the statement of Proposition 25.13. By Theorem 24.13, one can generate the fractional PRG pX using a seed of length $s = O(\log(k/\delta) + \log\log n)$. Applying the polarized random walk of Theorem 25.4 converts the fractional PRG pX to a standard PRG for $\mathcal{F}$ with seed length

$$O(s \cdot p^{-2} \cdot \log(n/\varepsilon_1)) = O(b^2 \cdot \log n \cdot (\log(ab/\varepsilon) + \log\log n)).$$

and error

$$O(\varepsilon_1 \cdot p^{-2} \cdot \log(n/\varepsilon_1)) \leqslant \varepsilon.$$

□

25.4 Concluding remarks

More recently, it was shown in [CHLT18] that if instead of using pX for a δ-almost k-wise uniform PRGs X, one can use a more elaborate construction of fractional PRGs, then we obtain a fractional PRG using only *second-level* Fourier bounds. Combined with the discussed polarized random walk, their construction implies the following theorem.

Theorem 25.15 ([CHLT18])**.** *Let $\mathcal{F}$ be a class of Boolean functions closed under restrictions. Let $\varepsilon > 0$ be constants and suppose every $f \in \mathcal{F}$ satisfies*

$$L_{1,2}(f) = \left\| f^{=2} \right\|_A \leqslant \alpha. \tag{25.3}$$

Then there is an explicit ε-PRG for f with seed length $O\left((\alpha/\varepsilon)^{2+o(1)} \operatorname{polylog}(n)\right)$.

Subsequently, [CGL+21] showed that better bounds could be achieved if bounds on higher Fourier levels are available, and interestingly, that fractional PRGs can be achieved even from bounds on $|\sum_{S:|S|=d} \hat{f}(S)|$ where one can have cancellations, as opposed to L_1 bounds.

These works show that certain improved bounds on the Fourier tails of $\mathbb{F}_2$-polynomials will lead to new PRGs. For isntance, the resolution of the following conjecture would be a major breakthrough in complexity theory, as currently no explicit PRGs are known for polynomials of degree $\Omega(\log(n))$.

Conjecture 25.16 ([CHLT18])**.** *If $p : \mathbb{F}_2^n \to \mathbb{F}_2$ is a polynomial of degree d and $f : \{0,1\}^n \to \{0,1\}$ is the corresponding Boolean function, then*

$$L_{1,2}(f) = O(d^2).$$

Chapter 26

Draft: The Semigroup method

In Chapter 10, we introduced the noise operator T_ρ and studied some of its key properties. In this chapter, we take a broader perspective by examining the noise operator as a specific instance of a more general class of operators.

The general class of operators we consider arises from random walks, and they are parameterized by time $t \in [0, \infty)$. Given a (continuous time) random walk, the corresponding operator Q_t maps f to the function

$$Q_t f : a \mapsto \mathbb{E}[f(X^a(t))]$$

where $X^a(t)$ is the position of the random walk at time t starting at point a.

We begin this chapter by analyzing the simplest case: the discrete random walk on the Boolean hypercube.

26.1 Poisson random walk on Hypercube

Consider the n-dimensional hypercube with the vertex set $\{0,1\}^n$, where two vertices are neighbours if and only if they differ in one coordinate. Let us consider the standard discrete random walk on this graph, starting from a vertex $a \in \{0,1\}^n$. The random walk is defined as follows:

- **Initialization:** $Y^a(0) = a$ is the starting point.
- **Transition rule:** At each discrete time $t \in \mathbb{N}$, we choose the next vertex $Y^a(t)$ uniformly at random from the n neighbours of the current vertex $Y^a(t-1)$.

Let f_t be the distribution of Y_t, meaning $f_t(x) = \mathbf{Pr}[Y^a(t) = x]$. Initially, this distribution is given by

$$f_0(x) = \mathbf{1}_a(x) = 2^{-n} \sum_{S \subseteq [n]} \chi_S(a)\chi_S(x).$$

Define the operator $K : L_2(\{0,1\}^n) \to L_2(\{0,1\}^n)$ as

$$Kf(x) = \frac{1}{n} \sum_{i=1}^{n} f(x \oplus e_i), \tag{26.1}$$

so that $f_t = K^t f_0$ for every integer $t > 1$. For every character χ_S, we have

$$K\chi_S = \frac{1}{n}\left((n - |S|)\chi_S - |S|\chi_S\right) = \left(1 - \frac{2|S|}{n}\right)\chi_S.$$

Thus, χ_S are eigenvectors of the operator K with corresponding eigenvalues $(1 - 2|S|/n)$. It follows that

$$f_t = 2^{-n} \sum_{S \subseteq [n]} \left(1 - \frac{2|S|}{n}\right)^t \chi_S(a)\chi_S.$$

Since

$$\left|1 - \frac{2|S|}{n}\right| < 1 - \frac{2}{n} \quad \text{for} \quad S \notin \{\emptyset, [n]\},$$

we have

$$\left\| f_t - 2^{-n} \left(\chi_\emptyset + (-1)^t \chi_{[n]}\right) \right\|_\infty \leqslant (1 - 2/n)^t \leqslant e^{-2t/n}.$$

Therefore, as $t \to \infty$, we observe the following convergences:

- For even times:
$$f_{2t} \to 2^{-n} \left(\chi_\emptyset + \chi_{[n]}\right) = 2^{1-n} \mathbf{1}_{[\sum x_i \equiv 0 \mod 2]},$$
- For odd times:
$$f_{2t+1} \to 2^{-n} \left(\chi_\emptyset - \chi_{[n]}\right) = 2^{1-n} \mathbf{1}_{[\sum x_i \equiv 1 \mod 2]}.$$

In other words, on even time steps, this random walk quickly converges to the uniform measure on points with even parity. On odd time steps, it converges to the uniform distribution on the points with odd parity. This phenomenon corresponds to the bipartite structure of the hypercube, and it prevents the random walk from being fully ergodic (i.e., converging to the uniform distribution over all vertices).

The lazy random walk: To make the random walk ergodic, we modify the standard walk by introducing a *lazy* transition rule. Given a parameter $\lambda \in (0, \frac{1}{2})$, define the random walk $Z(t) \coloneqq Z^{a,\lambda}(t)$ as

- **Initialization:** $Z(0) = a$ is the starting point.
- **Transition rule:** At time $t \in \mathbb{N}$
  - With probability $1 - \lambda$, stay at the current location: $Z(t) = Z(t-1)$;
  - With probability λ, move to a uniformly random neighbour of $Z(t-1)$.

Let f_t denote the distribution of the random walk at time t. The evolution of f_t follows $f_t = K_\lambda f_{t-1}$, where $K_\lambda = (1-\lambda)\,\mathrm{Id} + \lambda K$ where K is defined in (26.1). Every character χ_S satisfies

$$K_\lambda \chi_S = (1-\lambda)\chi_S + \lambda \left(1 - \frac{2|S|}{n}\right) \chi_S = \left(1 - \frac{2\lambda|S|}{n}\right) \chi_S,$$

and consequently

$$f_t = K_\lambda^t f_0 = 2^{-n} \sum_{S \subseteq [n]} \left(1 - \frac{2|S|\lambda}{n}\right)^t \chi_S(a) \chi_S.$$

Since $1 - 2\lambda|S|/n < 1$ unless $S = \emptyset$, the distribution f_t of the lazy random walk will converge to the uniform measure on the cube as t tends to infinity. We obtained a fully ergodic random walk by using laziness to eliminate the periodicity of the standard random walk.

Continuous Limit of the Lazy Random Walk: When t is large, by the law of large numbers, the random walk moves at roughly λ fraction of time steps. Hence, it is more natural to rescale time by considering $f_{\lfloor nt/\lambda \rfloor}$. In other words, consider n epochs, each consisting of $1/\lambda$ step so that, on average, we expect to make one move in every epoch. By taking the limit $\lambda \to 0$, we obtain a *continuous* version of the walk. This leads to the formula

$$\lim_{\lambda \to 0} f_{\lfloor nt/\lambda \rfloor} = 2^{-n} \sum_{S \subseteq [n]} \chi_S(a) e^{-2t|S|} \chi_S.$$

We can rescale time by another factor of 2 to obtain the nicer formula:

$$\lim_{\lambda \to 0} f_{\lfloor nt/2\lambda \rfloor} = 2^{-n} \sum_{S \subseteq [n]} \chi_S(a) e^{-t|S|} \chi_S.$$

The continuous random walk $(X^a(t))_{t\in[0,\infty)}$ that is obtained as the limit in this way has the property that

$$X^a(t) \sim 2^{-n} \sum_{S\subseteq[n]} \chi_S(a) e^{-t|S|} \chi_S. \tag{26.2}$$

Note further that if instead of f_0, we start with an arbitrary distribution μ and pick the starting point a randomly according to μ, then the distribution at time t will be

$$\lim_{\lambda\to 0} \mu_{\lfloor nt/2\lambda \rfloor} = \sum_{S\subseteq[n]} e^{-t|S|} \widehat{\mu}(S) \chi_S. \tag{26.3}$$

This is equal to $T_{e^{-t|S|}}\mu$. Now for the moment we depart from analyzing this random walk as the limit of the discrete random walks, and consider a different and more direct perspective.

Recall that the exponential distribution with parameter λ is defined through its probability density function (pdf)

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0. \end{cases}$$

An exponential distribution is supported on the interval $[0,\infty)$. Here $\lambda > 0$ is the parameter of the distribution, and if $\lambda = 1$ the distribution is called the standard exponential distribution. Exponential distribution is the continuous analogue of the geometric distribution, and can be interpreted as the time that it takes for an event to happen if it has the occurrence rate of λ per unit of time (say a customer showing up in a store). It has the key property of being memoryless, that is if E is exponentially distributed, then $\mathbf{Pr}(E \leqslant s+t | E > s) = \mathbf{Pr}(E \leqslant t)$. This means that as you continue to wait, the chance of something happening "soon" neither increases nor decreases.

The *Poisson distribution* with parameter λ, denoted by $\mathrm{Pois}(\lambda)$, is the probability distribution on $\{0,1,2,\ldots\}$ defined by

$$\mathbf{Pr}[\{k\}] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

The Poisson distribution can be obtained using exponential random variables as time increments. Let $E_1, E_2, \ldots$ be i.i.d. exponential random variables with parameter λ, and suppose the first event happens at time E_1, the second at time E_1+E_2, the third at time $E_1+E_2+E_3$, etc. Then $\max_k(\sum_{i=1}^k E_i < t)$, which is the number events that happen until time t, has Poisson distribution with parameter λt. Now our goal is to define a continuous random walk on the cube. First we need to define the standard Poisson process.

Definition 26.1. The standard Poisson process $(N(t))_{t\in[0,\infty)}$ is an increasing integer-valued Markov process with independent Poisson increments:

- $N(0) = 0$;
- For $0 \leqslant s \leqslant t$, we have $N(t) - N(s) \sim N(t-s) \sim \mathrm{Pois}(t-s)$.

It follows from the above discussion that a Poisson process can be generated using increments with exponential distributions: If $E_1, E_2, \ldots$ are i.i.d. random variables with standard exponential distribution, then defining

$$N(t) = \max_k(\sum_{i=1}^k E_i < t)$$

we obtain the standard Poisson process. The Poisson process has the Markov property which is defined as being memoryless in the sense that the conditional probability distribution of future states of the process conditional on both past and present values depends only upon the present state, not on the sequence of events that preceded it.

Since we are only concerned with the cube $\{0,1\}^n$, we project the Poisson process into $\{0,1\}$. Define the $\{0,1\}$-valued process $(N(t))_{t\in[0,\infty)}$ with $M(t) = N(t) \bmod 2$. Note that in general it is not true that the image of a Markov process is always a Markov process, but in this case it is easy to see that $(M(t))_{t\in(0,\infty]}$ is a Markov process.

Exercise 26.1. Show that the process $(M(t))_{t\in[0,\infty)}$ defined above is a Markov process.

Exercise 26.2. Construct a Markov process $(X(t))_{t\in[0,\infty)}$ and a function f such that $(f(X(t)))_{t\in[0,\infty)}$ is not a Markov process.

Let us now calculate the transition probabilities of this process.

Claim 26.2. *For $0 \leqslant s < t$, we have*

$$\mathbf{Pr}[M(t) = 0 | M(s) = 0] = \mathbf{Pr}[M(t) = 1 | M(s) = 1] = \frac{(1 + e^{-2(t-s)})}{2},$$

and

$$\mathbf{Pr}[M(t) = 0 | M(s) = 1] = \mathbf{Pr}[M(t) = 1 | M(s) = 0] = \frac{(1 - e^{-2(t-s)})}{2}.$$

Proof. We have

$$\begin{aligned}\mathbf{Pr}[M(t) = 0 | M(s) = 0] &= \mathbf{Pr}[N(t) \equiv_2 0 | N(s) \equiv_2 0] = \mathbf{Pr}[N(t) - N(s) \equiv_2 0 | N(s) \equiv_2 0] \\ &= \mathbf{Pr}[N(t-s) \equiv_2 0] = \mathrm{Pois}_{t-s}(\{0, 2, 4, \ldots\}) = \frac{(1 + e^{-2(t-s)})}{2}.\end{aligned}$$

The other cases are similar. □

We rescale time and define the process $(X(t))_{t \in [0,\infty)}$ as $X(t) = M(t/2)$ so that

$$\mathbf{Pr}[X(t) = 0 | X(s) = 0] = \mathbf{Pr}[X(t) = 1 | X(s) = 1] = \frac{(1 + e^{-(t-s)})}{2}, \tag{26.4}$$

and

$$\mathbf{Pr}[X(t) = 0 | X(s) = 1] = \mathbf{Pr}[X(t) = 1 | X(s) = 0] = \frac{(1 - e^{-(t-s)})}{2}. \tag{26.5}$$

This process is homogeneous in both time and space. It is time homogeneous as the distribution of $X(t) - X(s)$ depends only on $t - s$, and it is space homogeneous as it is symmetric with respect to 0 and 1.

Let us also remark that we could construct the process $(X(t))_{t \in [0,\infty)}$ directly from the transition equations (26.4) and (26.5) without starting from the Poisson process. Indeed one only needs to verify that the Chapman-Kolmogorov equations are satisfied. That is, setting $p_{s,t}(x, y)$ to the value of $\mathbf{Pr}[X(t) = y | X(s) = x]$ according to (26.4) and (26.5), we need to verify that for $0 \leqslant s < t < u$, and $z, x \in \{0, 1\}$, we have

$$p_{s,u}(x, z) = \sum_{y \in \{0,1\}} p_{s,t}(x, y) p_{t,u}(y, z). \tag{26.6}$$

Then Kolmogorov's extension theorem guarantees that there is a Markov process $(X(t))_{t \in [0,\infty)}$ satisfying the transition inequalities (26.4) and (26.5).

Now that we have constructed the Markov process $(X(t))_{t \in [0,\infty)}$ on $\{0, 1\}$, we will use it to define a continuous random walk on the cube $\{0, 1\}^n$. Let $a \in \{0, 1\}^n$ be the starting point, and let $(X_1(t), \ldots, X_n(t))_{t \in [0,\infty)}$ be i.i.d. copies of $(X(t))_{t \in [0,\infty)}$. Define the process $(X^a(t))_{t \in [0,\infty)}$ as

$$X^a(t) = (a_1 + X_1(t), \ldots, a_n + X_n(t)),$$

where the additions are in $\{0, 1\}$. Note that the process starts at $X^a(0) = a$, and then when the first change occurs in $(X_1(t), \ldots, X_n(t))$, it jumps to the corresponding neighbors of a in the cube (i.e. to $a + e_i$ for some $1 \leqslant i \leqslant n$), and so on.

Denoting by f_t the distribution of $X^a(t)$, by (26.4) and (26.5), we have

$$f_t(x) = \prod_{i=1}^{n} \left(\frac{1 + (-1)^{a_i + x_i} e^{-t}}{2} \right) = 2^{-n} \sum_{S \subseteq [n]} \chi_S(a) e^{-t|S|} \chi_S.$$

This is the same distribution that we obtained in (26.2) as the limit of the discrete lazy walks with proper rescaling of time. As we will formally see in Section 26.2, this means that the two random walks coincide. This is a curious fact. In the lazy random walk, there is no coordinate-wise independence, as at every move we change exactly one of the coordinates. However in the Poisson random walk, coordinates behave totally independently. So it might seem

mysterious that in the limit, the lazy random walk converges to the Poisson random walk and the coordinate-wise dependencies disappear. Indeed this is part of a more general phenomenon that is called Poissonization. Let us explain this using a simple example.

Example 26.3 (Poissonization)**.** Consider a biased coin that comes up Head with probability p, and Tail with probability $1-p$. We flip the coin infinitely many times, and let H_n and T_n respectively denote the number of heads and tails until time n. Obviously, these two random variables are totally dependent as $H_n = n - T_n$.

Now consider the following different process. Let $E_1, E_2, \ldots$ be an i.i.d. sequence of standard exponential random variables. We wait until time E_1 and toss the coin for the first time, then we wait for another E_2 units of time and toss the coin again, etc. For $t \in [0,\infty)$, let H'_t and T'_t respectively denote the number of heads and tails until time t. Then H'_t has distribution $\mathrm{Pois}(pt)$ and T'_t has distribution $\mathrm{Pois}((1-p)t)$, and it is not hard to see that they are (rather miraculously) independent.

The reason for this independence becomes apparent when we examine how the second process can be obtained as the limit of the first one. Let N be a large number and set $\lambda = \frac{1}{N}$, and consider the lazy version of the first process, where now at time t, we do nothing with probability $1-\lambda$, and with probability λ we flip our biased coin.

Let us compare this to two independent processes, one responsible for producing heads, and the other one for producing tails. In the first one, at each time step, with probability λp we produce a Head, and we do nothing we probability $1-\lambda p$. In the second process, at every time step, with probability $\lambda(1-p)$ we produce a Tail, and we do nothing we probability $1-\lambda(1-p)$. Observing these two processes simultaneously at a single time step, we see that

$$\mathbf{Pr}[\text{Nothing is produced}] = (1-p\lambda)(1-(1-p)\lambda) = 1-\lambda+p(1-p)\lambda^2,$$

and

$$\mathbf{Pr}[\text{A Head is produced}] = p\lambda(1-(1-p)\lambda) = p\lambda - p(1-p)\lambda^2,$$

and

$$\mathbf{Pr}[\text{A Tail is produced}] = (1-p\lambda)(1-p)\lambda = (1-p)\lambda - p(1-p)\lambda^2,$$

and

$$\mathbf{Pr}[\text{A Head and a Tail are produced}] = p(1-p)\lambda^2.$$

Now if we let λ tend to 0, the quadratic terms in λ become negligible, and the process becomes indistinguishable form the lazy biased coin process that we described above. That is in the limit, after proper rescaling of the time, the lazy biased coin process, and the independent production of heads and tails converge to the same limit, the Poisson process $(H'_t, T'_t)_{t\in[0,\infty)}$ that we described above. This in particular verifies the independence for $(H'_t, T'_t)_{t\in[0,\infty)}$.

The independence achieved by Poissonization of the discrete lazy random walk on the cube is highly desirable, and it is one of the main motivations behind considering the random Poisson processes rather than the more elementary object of the discrete random walk on the cube.

26.2 Semigroups

Consider the Poisson random walk $(X^a(t))_{t\in[0,\infty)}$ constructed in Section 26.1. This random walk can be used to define a class of operators. For $f : \{0,1\}^n \to \mathbb{R}$, $a \in \{0,1\}^n$ and $t \geqslant 0$ define

$$P_t f(a) = \mathbb{E}[f(X^a(t))].$$

In other words, to evaluate $P_t f$ at a point a, we start our random walk at a, and look at the expected value of f on the point $X^a(t)$ obtained by running the random walk until time t. Note that by (26.4) and (26.5), we have

$$P_t\chi_S(a) = \mathbb{E}[\chi_S(X^a(t))] = \mathbb{E}\prod_{i\in S}(-1)^{a_i+X_i(t)} = \chi_S(a)\prod_{i\in S}\mathbb{E}[(-1)^{X_i(t)}] = \chi_S(a)\prod_{i\in S}e^{-t} = e^{-t|S|}\chi_S(a).$$

Thus $P_t\chi_S = \chi_S$ and consequently for every function $f : \{0,1\}^n \to \mathbb{C}$, we have

$$P_t f = \sum_{S\subseteq[n]} e^{-t|S|}\widehat{f}(S)\chi_S. \tag{26.7}$$

Hence, not surprisingly at this point, similar to (26.3), we have $P_t f = T_{e^{-t}} f$.

The operators P_t are clearly linear operators from $L_2(\{0,1\}^n)$ to $L_2(\{0,1\}^n)$. The next lemma shows that they form a semigroup.

Lemma 26.4. *We have $P_0 = \mathrm{Id}$, and $P_t \circ P_s = P_{t+s}$ for $s, t \geqslant 0$.*

Proof. The fact that P_0 is trivial. The identity $P_t \circ P_s = P_{t+s}$ can be verified using the definition $P_t f(a) = \mathbb{E}[f(X^a(t))]$ through Chapman-Kolmogorov equation (26.6) for the random walk. We leave the details as an exercise to the reader. □

Trivially P_t satisfies the following basic properties

- *Preserves Identity:* $P_t 1 = 1$.
- *Preserves Positivity:* If $f \geqslant 0$, then $P_t f \geqslant 0$.
- *Preserves Order:* If $f \geqslant g$, then $P_t f \geqslant P_t g$.

These observations motivate the following definition.

Definition 26.5. A set of linear operators $(Q_t)_{t\in[0,1)}$ is called a *semigroup* if $Q_0 = \mathrm{Id}$, and $Q_t \circ Q_s = Q_{t+s}$ for $t, s \in [0, \infty)$. If it furthermore satisfies

1. *Preserves Identity:* $Q_t 1 = 1$,
2. *Preserves Positivity:* $Q_t f \geqslant 0$ almost everywhere if $f \geqslant 0$ almost everywhere,
3. *Preserves Order:* If $f \geqslant g$ almost everywhere, then $P_t f \geqslant P_t g$ almost everywhere.

then it is called a *Markovian semigroup*.

Note that preserving order follows from preserving positivity, and can be omitted from the definition. Obviously the semigroup $(P_t)_{t\in[0,\infty)}$ constructed above is Markovian. Next we will show that in fact every Markovian semigroup can be constructed through a Makrov process. Consider a Markovian semigroup $(Q_t)_{t\in[0,\infty)}$ and define the transition probabilities of a time homogenuous random walk as

$$q_t(a, b) := (Q_t \mathbf{1}_b)(a), \tag{26.8}$$

where $\mathbf{1}_b$ is the indicator function of the point $\{b\}$. That is in the corresponding Markov process $(Y_t)_{t\in[0,\infty)}$, we would like for every $s \geqslant 0$, to have

$$\mathbf{Pr}[Y_{s+t} = b | Y_s = a] = q_t(a, b) := (Q_t \mathbf{1}_b)(a).$$

Since Q_t preserves positivity, we have $q_t(a, b) \geqslant 0$, and since $Q_t 1 = 1$ we have that

$$\sum_b q_t(a, b) = \sum_b (Q_t \mathbf{1}_b)(a) = (Q_t 1)(a) = 1.$$

The Chapman-Kolmogorov equation (26.6) can also be verified using the semigroup property $Q_t \circ Q_s = Q_{s+t}$ which we leave to the reader as an exercise.

Exercise 26.3. If $(Q_t)_{t\in[0,\infty)}$ is Markovian semigroup, and $q_t(a, b)$ is defined as in (26.8). Show that $q_t(a, b)$ satisfies the Chapman-Kolmogorov equation (26.6).

Hence by Kolmogorov's extension theorem, there exists a corresponding Markov process $(Y^a(t))_{t\in[0,\infty)}$ with transition probabilities $q_t(a, b)$. Now note that

$$\mathbb{E}[f(Y^a(t))] = \sum_b q_t(a, b) f(b) = \sum_b (Q_t \mathbf{1}_b)(a) f(b) = Q_t(\sum_b \mathbf{1}_b f(b))(a) = (Q_t f)(a).$$

Hence the semigroup $(Q_t)_{t\in[0,\infty)}$ could be recovered as $Q_t f(a) = \mathbb{E}[f(Y^a(t))]$.

To summarize we showed that Markov processes $(Y_t^a)_{t\in[0,\infty)}$ are in one to one correspondence with Markovian semigroup $(Q_t)_{t\in[0,\infty)}$ via the formulas $Q_t f(a) = \mathbb{E}[f(X^a(t))]$ and $q_t(a,b) = (Q_t \mathbf{1}_b)(a)$.

Now that we established this equivalence, we can mention an important property of Markovian semigroups, namely that they preserve expectation with respect to the so called *invariant measure*.

Definition 26.6. A probability measure μ on a finite set Ω is an *invariant measure* for a Markovian semigroup $(Q_t)_{t\in[0,\infty)}$, or a *stationary distribution* for the corresponding Markov process q_t, if for every $y \in \Omega$ and $t > 0$,

$$\sum_{x\in\Omega} \mu(\{x\}) q_t(x,y) = \mu(y). \tag{26.9}$$

This means that the total "immigration" to y balances "emigration" from y. Note that (26.9) is equivalent to $\mathbb{E}_\mu[Q_t \mathbf{1}_y] = \mathbf{1}_y$. Since $\{\mathbf{1}_y : y \in \Omega\}$ spans the set of all functions on Ω, we see that μ is invariant for the semigroup if and only if $\mathbb{E}_\mu[Q_t f] = f$ for every $f : \Omega \to \mathbb{R}$. Hence A Markovian semigroup preserve expectation with respect to invariant measure. When we work with a semigroup or a Markov process, invariant measures are the "right" measures to consider on the space. From this point on when we talk about a semigroup or a Markov process we always assume that the underlying measure space is an invariant measure for the semigroup, and that expectations are taken with respect to that measure.

The operators P_t is a symmetric (a.k.a. Hermitian) operator, and in fact self-adjoint as it is defined everywhere. Indeed by Plancherel,

$$\langle P_t f, g\rangle = \sum_{S\subseteq[n]} e^{-t|S|} \widehat{f}\widehat{g} = \langle f, P_t g\rangle.$$

In the more general case of the Markovian semi-groups when the invariant measure μ is nonuniform, the symmetry of the operator Q_t does not mean that the transition matrix $q_t(x,y)$ is symmetric. For example, in the finite case, it means that

$$\mu(\{x\}) q_t(x,y) = \langle Q_t \mathbf{1}_x, \mathbf{1}_y\rangle = \langle \mathbf{1}_x, Q_t \mathbf{1}_y\rangle = \mu(\{y\}) q_t(y,x),$$

In general the semigroup $(Q_t)_{t\in[0,\infty)}$ is symmetric if and only if the corresponding Markov process is time reversible. A symmetric Markovian semigroup preserve expectation. Indeed

$$\mathbb{E}[Q_t f] = \langle Q_t f, 1\rangle = \langle f, Q_t 1\rangle = \langle f, 1\rangle = \mathbb{E}[f]. \tag{26.10}$$

26.2.1 Generator of a semigroup

To define the *generator* of a semigroup we would like to differentiate Q_t in t, but unfortunately a Markovian semigroup need not even be continuous with respect to the parameter t: As an example one may consider $Q_0 f := f$ and $Q_t[f] := \mathbb{E}[f]$ for $t > 0$, which is not continuous in time unless f is constant almost surely. However, in many cases Markov semigroups are not only continuous but also differentiable with respect to time.

Definition 26.7. The linear operator $-\frac{d}{dt} Q_t\big|_{t=0^+}$ is called a generator of the semigroup $(Q_t)_{t\in[0,\infty)}$.

Note that for a Markovian operator, since $Q_t 1 = 1$ for every $t \geqslant 0$, we always have that $(-\frac{d}{dt} Q_t\big|_{t=0^+})1 = 0$.

Remark 26.8. For non-discrete spaces usually the generator cannot be defined on the whole L_2 function space but only a dense linear subspace. There are many technical problems and extensive literature concerning relations between a Markov semigroup and its generator. The assumption that is usually used is that Q_t is strongly continuous, i.e. it is continuous in t in the strong operator topology. Then it is not difficult to see that $\frac{d}{dt} Q_t\big|_{t=0^+}$ is well-defined on the dense set of all "smoothed" functions $\{Q_\varepsilon g : \varepsilon > 0, g \in L_2\}$.

Let us go back to the semi-group $(P_t)_{t\in(0,\infty]}$ that we constructed from the parity Poisson process.

Remark 26.9. We have shown that $T_{e^{-t}} \equiv P_t$. The notation P_t is preferred by probability theorists. Harmonic analysts however prefer the notation T_ρ as for example it allows considering complex values of ρ with $|\rho| \leqslant 1$ which leads to the definition of the so called holomorphic semigroups. Computer scientists also adopted the notation T_ρ as it is simpler, however there is a price to this, as the intuition that t corresponds to time, and that this operator is defined through a Markov process becomes less apparent.

Note that taking the derivative of

$$P_t f = \sum_{S \subseteq [n]} e^{-t|S|} \widehat{f}(S) \chi_S.$$

we see that the generator $L := -\frac{d}{dt} P_t \big|_{t=0^+}$ of this semigroup is defined as

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

Our semigroup P_t can be easily recovered from its generator:

$$P_t := e^{-tL} = \mathrm{Id} + \sum_{k=1}^{\infty} \frac{(-t)^k L^k}{k!}.$$

Indeed for a character, we have

$$P_t \chi_S = \left(1 + \sum_{k=1}^{\infty} \frac{(-t)^k |S|^k}{k!}\right) \chi_S = e^{-t|S|} \chi_S.$$

Remark 26.10. For this approach to work, it is necessary that the generator is a bounded operator (as it is the case for L, the generator of P_t). However in the more general settings of Markovian semigroups, the generator is not always defined on all of the space. Nevertheless, the notation e^{-tL} is still used, and it usually means the solution to the differential equation $\frac{d}{dt} Q_t = -L Q_t$ with the boundary condition $Q_0 = \mathrm{Id}$.

For the semigroup $(P_t)_{t \in [0,\infty)}$, there is a more direct way to define the generator L. We have

$$Lf = \frac{1}{2} \sum_{i=1}^{n} f(x) - f(x + e_i),$$

as it can be easily verified using the Fourier transform. Hence $L = \frac{n}{2}(\mathrm{Id} - K)$, where K is defined in (26.1).

In Theorem 10.11 we saw that the operator T_ρ is a contractive operator from L_p to L_p. This phenomenon holds for general symmetric Markovian semigroups.

Theorem 26.11. *Let $(Q_t)_{t \in [0,\infty)}$ be a* symmetric *Markovian semigroup, and $\Phi : \mathbb{R} \to \mathbb{R}$ be a convex function. For every $t \geqslant 0$ and every function $f \in L_2$ we have*

$$\mathbb{E}[\Phi(Q_t f)] \leqslant \mathbb{E}[\Phi(f)].$$

In particular taking $\Phi = |\cdot|^p$ for $p \geqslant 1$ we obtain $\|Q_t f\|_p \leqslant \|f\|_p$.

Proof. Since Φ is convex we have $\Phi(x) = \sup_{\alpha \in \mathcal{I}} a_\alpha x + b_\alpha$ for some family $\mathcal{I}$ of affine functions $a_\alpha x + b_\alpha$. Then for every $\alpha \in \mathcal{I}$, we have the pointwise inequality $\Phi(f) \geqslant a_\alpha f + b_\alpha$ which using the order-preserving property of Markovian semigroups reduces to the pointwise inequality

$$Q_t(\Phi(f)) \geqslant Q_t(a_\alpha f + b_\alpha) = a_\alpha (Q_t f) + b_\alpha.$$

Taking the supremum we obtain $Q_t(\Phi(f)) \geqslant \Phi(Q_t f)$. Taking the expectation and using the fact that symmetric Markovian semigroups preserve expectation (See 26.10), we obtain

$$\mathbb{E}[\Phi(f)] = \mathbb{E}[Q_t(\Phi(f))] \geqslant \mathbb{E}[\Phi(Q_t f)].$$

□

26.3 Some Examples

We close this chapter by mentioning some examples of Markovian semigroups.

Example 26.12. Consider the space $(\mathbb{R}, \lambda)$ where λ is the Lesbagues measure. Define the semigroup $(P_t)_{t\in[0,\infty)}$ as $P_t : f(\cdot) \to f(\cdot + t)$. It can be easily seen that this is a Markovian semigroup. Note that the generator L of this semigroup is equal to $-D$ where D is the differentiation:

$$(Lf)(x) = -\frac{d}{d_t} P_t f\Big|_{0^+} = -f'(x).$$

Then if try to recover P_t from the generator using the formula

$$P_t = e^{-tL} = \mathrm{Id} + \sum_{k=1}^{\infty} \frac{(-t)^k}{k!} L^k, \tag{26.11}$$

we obtain

$$P_t f(x) = f(x) + tf'(x) + \frac{t^2}{2!} f''(x) + \dots$$

This is the Taylor expansion for $f(x + t)$, and is equal to $f(x + t)$ when f is analytic. Note that there are smooth functions that are not analytic. For example, it is well-known that the function

$$f(x) = \begin{cases} e^{-1/x^2} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is smooth (i.e. it has derivatives of all orders), but it is easy to see that $f^{(k)}(0) = 0$ for all k, and thus $f(x + t) \neq f(x) + tf'(x) + \frac{t^2}{2!} f''(x) + \dots$ for $x = 0$. Note that even if we replace our original space $(\mathbb{R}, \lambda)$ with the compact space $(\mathbb{R}/\mathbb{Z}, \lambda)$, this example still shows that it is not always possible to recover the semigroup from its generator using (26.11).

Example 26.13 (Heat semigroup)**.** Joseph Fourier initiated the investigation of Fourier series and their applications to problems of heat transfer and vibrations. He discovered the law of heat conduction, also known as Fourier's law, which states that the time rate of heat transfer through a material is proportional to the negative gradient in the temperature and to the area, at right angles to that gradient, through which the heat flows. Fourier's law combined with conservation of energy implies the so called heat equation. Suppose one has a function $f(x)$ that describes the temperature at a given location of a metal bar. This function will change over time as heat spreads throughout space. The heat equation can be used to determine the change in the function f over time. It says that if $P_t f$ denotes the distribution of the temperature at time t, then

$$\frac{d}{dt}(P_t f)(x) = \alpha \frac{\partial^2}{\partial^2 x} P_t f(x),$$

where $\alpha > 0$ is a constant depending on the material and is called the thermal diffusivity. If instead of a bar, we consider a 3-dimensional object, and denote the temperature at point $x = (x_1, x_2, x_3)$ with $f(x_1, x_2, x_3)$, then the heat equation becomes

$$\frac{d}{dt}(P_t f) = \alpha \Delta (P_t f)(x),$$

where Δ denotes the Laplacian $\Delta := \frac{\partial^2}{\partial^2 x_1} + \frac{\partial^2}{\partial^2 x_2} + \frac{\partial^2}{\partial^2 x_3}$.

The heat equation is used in probability and describes random walks. It is also applied in financial mathematics for this reason.vIt is also important in Riemannian geometry and thus topology: it was adapted by Richard Hamilton when he defined the Ricci flow that was later used by Grigori Perelman to solve the topological Poincaré conjecture.

The heat equation can be understood through the heat semigroup. First we need to introduce the Brownian motion, an important notion that occurs frequently in pure and applied mathematics, economics and physics. The (1-dimensional) Brownian motion (a.k.a. Wiener process) is a continuous-time stochastic process $(B_t)_{t\in[0,\infty)}$ that is characterized by four facts:

- $B_0 = 0$.
- B_t is almost surely continuous.
- B_t has independent increments (i.e. $B_{t_1} - B_{s_1}$ is independent of $B_{t_2} - B_{s_2}$ for $0 \leqslant s_1 \leqslant t_1 \leqslant s_2 \leqslant t_2$.

- $B_t - B_s \sim N(0, t-s)$ for $t > s$, where $N(0, t-s)$ denotes the normal distribution with expected value 0 and variance $t - s$.

The Brownian motion can be obtained as the limit of the following discrete random walks. Let $\lambda > 0$ be the time increment. The random walk starts at the origin $X_0 = 0$, and at time $(t+1)\lambda$ its value $X_{(t+1)\lambda}$ is set with equal probability to either $X_{t\lambda} + \sqrt{\lambda}$ or $X_{t\lambda} - \sqrt{\lambda}$ (we make a left or right jump of magnitude $\sqrt{\lambda}$ with equal probability). Now as the time increment $\lambda \geqslant 0$ goes to 0, this random walk converges to the Brownian motion.

Now that we have a process $(B_t)_{t\in[0,\infty)}$, we can consider the corresponding semigroup $(P_t)_{t\in[0,\infty)}$. It maps every function $f : \mathbb{R} \to \mathbb{R}$, that satisfies certain integrability conditions, to $P_t f(x) := \mathbb{E}[f(B_t^x)]$, where $(B_t^x)_{t\in[0,\infty)}$ is the Brownian motion started at point x. Note that B_t^x has the same distribution as $x + B_t$ as the Brownian motion is space homogeneous, and hence

$$P_t f(x) = \mathbb{E}[f(B_t^x)] = \mathbb{E}[f(x + B_t)] = \mathbb{E}[f(x + \sqrt{t}G)],$$

where $G \sim N(0,1)$ is the standard Gaussian random variable, so that $\sqrt{t}G \sim N(0,t)$.

To find the generator, differentiating the operator and using the formula for the density of the normal distribution, we get

$$\begin{aligned}
\frac{d}{dt}P_t f(x) &= \frac{d}{dt}\mathbb{E}[f(x+\sqrt{t}G)] = \frac{1}{2\sqrt{t}}\mathbb{E}[f'(x+\sqrt{t}G)G] = \frac{1}{2\sqrt{t}}\frac{1}{\sqrt{2\pi}}\int f'(x+\sqrt{t}y)e^{-y^2/2}y\,dy \\
&= \frac{1}{2\sqrt{2t\pi}}\int f'(x+\sqrt{t}y)\frac{d}{dy}(-e^{-y^2/2})dy = \frac{1}{2\sqrt{2t\pi}}\int \sqrt{t}f''(x+\sqrt{t}y)e^{-y^2/2}dy \\
&= \frac{1}{2}\mathbb{E}f''(x+\sqrt{t}G), \qquad (26.12)
\end{aligned}$$

where in the integration by part we assumed that f vanishes at $\pm\infty$. Taking the limit $t \to 0$ we obtain that the generator is $Lf = \frac{-1}{2}f''$, or in other words $L = \frac{-1}{2}\Delta$ where Δ is the (one-dimensional) Laplace operator. Note that (26.12) shows that

$$\frac{d}{dt}P_t f(x) = \frac{1}{2}\Delta_x(P_t f(x)),$$

where Δ_x denotes the Laplacian with respect to x. This is the famous heat equation discussed above which roughly means that the flow of heat can be approximated as the movement of many small particles, where each particle moves according to a Brownian motion.

The heat semigroup can be defined on the n-dimensional space. Let $B_1(t), \ldots, B_n(t)$ be independent 1-dimensional Brownian motions as defined above. The n-dimensional Brownian motion $(B_t)_{t\in[0,1)}$ is defined as

$$B_t = \left(\frac{B_1(t)}{\sqrt{n}}, \ldots, \frac{B_n(t)}{\sqrt{n}}\right)_{t\in[0,\infty)}.$$

The normalization factor $\frac{1}{\sqrt{n}}$ is chosen so that $B_t \sim N_n(0,1)$ is an n-dimensional Gaussian random variable and thus has density

$$\Phi_n(x) := \frac{1}{(2\pi)^{n/2}}e^{-\|x\|_2^2/2} = \frac{1}{(2\pi)^{n/2}}e^{-(\sum_{i=1}^n x_i^2)/2}.$$

Repeating the calculation in (26.12), we see that the generator of the heat semigroup defined via this process is $\frac{-1}{2}\Delta$ where $\Delta = \frac{\partial^2}{\partial^2 x_1} + \ldots + \frac{\partial^2}{\partial^2 x_n}$ is the Laplace operator, and the heat equation

$$\frac{d}{dt}P_t f(x) = \frac{1}{2}\Delta(P_t f(x)),$$

holds.

Example 26.14 (The Ornstein-Uhlenbeck semigroup)**.** This semigroup is defined on $(\mathbb{R}, \gamma)$ where γ is the Gaussian measure. In some aspects it is closely related to the semigroup $(P_t)_{t\in[0,\infty)}$ that we defined on the cube $\{0,1\}^n$. We will define a process similar to the Brownian motion. Consider a time increment $\lambda > 0$, and define the process $(X_{t\lambda})_{t\in\mathbb{Z}_+}$ in the following way. To make a move from a point a, we first dilate a by multiply it by $e^{-\lambda}$ and then we make a jump of magnitude $\sqrt{\lambda}$ either to the left or right with equal probability. That is $X_{(t+1)\lambda}$ is set to one of $e^{-\lambda}X_{t\lambda} \pm \sqrt{\lambda}$

with equal probability. If we take the limit as $\lambda \to 0$, we obtain a Gaussian process $(X_t)_{t\in[0,\infty)}$. Now, because of the dilation, unlike the Brownian motion, X_t does not escape to infinity as t grows, and in fact X_t converges to $N(0,1)$ in distribution. It is not difficult to see that if we start the process at a point a, then $X_t^a \sim e^{-t}a + \sqrt{1-e^{-2t}}G$, where $G \sim N(0,1)$ is a standard Gaussian. Hence the corresponding semigroup U_t satisfies

$$U_t f(x) = \mathbb{E}\left[f(e^{-t}x + \sqrt{1-e^{-2t}}G)\right].$$

To describe the connection to the semigroup $(P_t)_{t\in[0,\infty)}$ on the cube $\{0,1\}^n$, consider a function $f : (\mathbb{R},\gamma) \to \mathbb{R}$, and define $g_n : \{0,1\}^n \to \mathbb{R}$ as $g_n(x_1,\ldots,x_n) = f(\frac{2(\sum x_i)-n}{\sqrt{n}})$. Note that $P_t g_n$ is a symmetric function and thus $P_t g_n(x_1,\ldots,x_n) = f_n(\frac{2(\sum x_i)-n}{\sqrt{n}})$ for a function $f_n : \mathbb{R} \to \mathbb{R}$. It is not difficult to see that

$$\lim_{n\to\infty} f_n = U_t f,$$

which can be interpreted as

$$\lim_{n\to\infty} P_t g_n = U_t f.$$

The same trick of approximating a gaussian with $\frac{2(\sum x_i)-n}{\sqrt{n}}$ allows one to deduce many geometric results in the Gaussian space from results on the cube. Going in the opposite direction is usually much harder, but there are some tools like the invariance principle that we will see later in Chapter **??** that allow it under some conditions.

the Ornstein-Uhlenbeck semigroup, similar to the heat-semigroup, can be defined in the n-dimensional space endowed with the Gaussian measure. For $f : (\mathbb{R}^n,\gamma_n) \to \mathbb{R}$ we have

$$U_t f(x) = \mathbb{E}\left[f(e^{-t}x + \sqrt{1-e^{-2t}}G)\right],$$

where G is the standard n-dimensional Gaussian random variable.

We leave to the reader to verify that the generator of the Ornstein-Uhlenbeck semigroup in general is

$$(Lf)(x) = \langle x, \nabla f(x)\rangle - (\Delta f)(x).$$

Exercise 26.4. Show that the generator of the n-dimensional Ornstein-Uhlenbeck semigroup is

$$(Lf)(x) = \langle x, \nabla f(x)\rangle - (\Delta f)(x).$$

Chapter 27

Isoperimetric Type Inequalities

Consider the hypercube with vertex set $\mathbb{Z}_2^n$, and let $S \subseteq \mathbb{Z}_2^n$ be a subset of the vertices. As we have discussed earlier the total influence of the indicator function of S corresponds to the size of the *edge boundary* of S. In other words for $f := \mathbf{1}_S$, we have

$$I_f = \mathbb{E}\left[\sum_{i=1}^{n} |f(x) - f(x+e_i)|\right] = \frac{2|\partial f|}{2^n},$$

where the edge boundary of S, denoted ∂S, is the set of edges of the cube with one endpoint in S and the other endpoint outside of S. In this chapter we study concepts related to edge-boundries.

27.0.1 Energy functions

Consider the semigroup $(P_t)_{t\in[0,\infty)}$ that we constructed from the Poisson random walk on the cube. Define the bi-linear form $\mathcal{E}(\cdot,\cdot)$ via the generator L of the semigroup $(P_t)_{t\in[0,\infty)}$ as

$$\mathcal{E}(f,g) := \langle f, Lg\rangle = \langle Lf, g\rangle.$$

This is a positive semi-definite form as $\mathcal{E}(f) := \mathcal{E}(f,f) = \sum |S||\widehat{f}(S)|^2 \geqslant 0$.

The positive semi-definiteness of the $\mathcal{E}$ can also be verified directly, without appealing to Fourier expansion, from contractivity of the semi-group. Indeed, for $t \geqslant 0$ and $f \in L_2$, set $\Psi(t) = \|P_t f\|_2^2 = \mathbb{E}[(P_t f)^2]$. Then taking the derivative with respect to t, we obtain

$$\Psi'(t) = 2\mathbb{E}\left[(P_t f)\frac{d}{dt}P_t f\right] = 2\mathbb{E}[-(P_t f)\cdot L(P_t f)],$$

and thus $\Psi'(0^+) := \lim_{t\to 0}\Psi'(t) = -2\mathbb{E}[f\cdot Lf] = -2\mathcal{E}(f,f)$. On the other hand, because of the contractivity of $(P_t)_{t\in[0,\infty)}$ we have

$$\Psi(t) \leqslant \|f\|_2^2 = \|P_0 f\|_2^2 = \Psi(0),$$

so that $\Psi'(0^+) \leqslant 0$. Thus $\mathcal{E}(f) := \mathcal{E}(f,f) \geqslant 0$, and $\mathcal{E}$ is positive semidefinite.

Let us now find a combinatorial way of describing $\mathcal{E}$. Using the formula

$$Lf(x) = \frac{1}{2}\sum_{i=1}^{n} f(x) - f(x+e_i),$$

we obtain

$$\mathcal{E}(f,g) = \langle f, Lg\rangle = 2^{-n-1}\sum_{x\sim y} f(x)g(x) - f(x)g(y),$$

where $x \sim y$ means that x and y are neighbours in the cube (i.e. $y = x + e_i$ for some $i \in [n]$). Using

$$f(x)g(x) - f(x)g(y) + f(y)g(y) - f(y)g(x) = (f(x) - f(y))(g(x) - g(y)),$$

we can simplify this to

$$\mathcal{E}(f,g) = 2^{-n-2}\sum_{x\sim y}(f(x)-f(y))(g(x)-g(y)),$$

which in particular shows

$$\mathcal{E}(f) = \mathcal{E}(f,f) = 2^{-n}\sum_{x\sim y}\left(\frac{f(x)-f(y)}{2}\right)^2. \tag{27.1}$$

The last expression is a discrete counterpart of the averaged $|\nabla f|^2$. The similarity to the physical kinetic energy notion explains the name given to this quadratic form. Quadratic forms of this type (under some additional conditions) are called Dirichlet forms and play important role in the theory of Markov semigroups. Given an open set $\Omega \subseteq \mathbb{R}^n$ and function $f : \Omega \to \mathbb{R}$, the Dirichlet's energy of the function f is the real number

$$\mathcal{E}(f) = \frac{1}{2}\int |\nabla f|^2 dxdy, \tag{27.2}$$

where $\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)$ is the gradient of the function f.

Definition 27.1 (Discrete gradient)**.** For a function $f : \mathbb{Z}_2^n \to \mathbb{R}$, define the discrete gradient of f at point x as

$$\nabla f(x) = \left(\frac{f(x)-f(x+e_1)}{2}, \ldots, \frac{f(x)-f(x+e_n)}{2}\right).$$

With this notation we have for every function $f : \mathbb{Z}_2^n \to \mathbb{R}$

$$\mathcal{E}(f) = 2^{-n}\sum_{x\sim y}\left(\frac{f(x)-f(y)}{2}\right)^2 = \mathbb{E}|\nabla f(x)|^2,$$

which reminisces the Dirichlet energy formula (27.2). There is an extensive literature that investigates the conditions under which the generator of a semigroup can be constructed from a Dirichlet form. In the case of the finite spaces the following are indeed equivalent.

$$\text{Markov processes} \sim \text{semigroups} \sim \text{generators} \sim \text{Dirichlet forms}$$

Let us finish this section by mentioning that the energy function behaves nicely when composed with Lipschitz maps. Let $\Psi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz map with constant C, i.e. $|\Psi(a) - \Psi(b)| \leqslant C|a-b|$ for all $a, b \in \mathbb{R}$. Then the formula

$$\mathcal{E}(\Psi(f)) = 2^{-n}\sum_{x\sim y}\left(\frac{\Psi(f(x))-\Psi(f(x))}{2}\right)^2$$

shows that for every $f \in \mathbb{Z}_2^n \to \mathbb{R}$, we have

$$\mathcal{E}(\Psi(f)) \leqslant C^2\mathcal{E}(f).$$

In particular, $\mathcal{E}(|f|) \leqslant \mathcal{E}(f)$. This can be generalized to other symmetric Markovian sermigroups under some mild technical conditions.

27.1 Poincaré inequalities

The classical Poincaré inequality comes from partial differential equations. It says that given a bounded connected open subset $D \subseteq \mathbb{R}^n$ with a sufficiently "regular" boundary, there exists a constant C_D such that for every function $f \in \mathcal{C}^1(D)$ (that is f differentiable and it's derivative is continuous) satisfying $\int_D f = 0$, we have

$$\int_D f^2 \leqslant C_D \int_D |\nabla f|^2.$$

The probabilistic analog of this is more relevant to us. A probability Borel measure ν on $\mathbb{R}^n$ is said to satisfy the Poincaré inequality with constant C if for every $\mathcal{C}^1$ function $f : \mathbb{R}^n \to \mathbb{R}$ with $\int f d\nu < \infty$, we have

$$\operatorname*{Var}_{\nu}(f) := \int f^2 d\nu - \left(\int f d\nu \right)^2 \leqslant C \int |\nabla f|^2 d\nu.$$

On the discrete group $\mathbb{Z}_2^n$, using the discrete gradient, the Energy function will take the place of $\int |\nabla f|^2 d\nu$, and we will obtain the following Poincaré inequality

$$\mathbb{E}[f^2] - \mathbb{E}[f]^2 \leqslant \mathcal{E}(f) := \mathbb{E}\left[|\nabla f|^2\right] := \mathbb{E}[fLf].$$

This follows by noticing that the left hand side is equal to $\sum_{S \neq \emptyset} |\widehat{f}(S)|^2$ while the right hand side is equal to

$$\langle f, Lf \rangle = \sum_{S \subseteq [n]} |S| |\widehat{f}(S)|^2.$$

The above variance-energy inequality is also called an spectral gap inequality. It holds because there is a gap in the spectrum $\sigma(L)$ between the eigenvalue 0, associated to the constant function 1 (principal character), and the second smallest eigenvalue in absolute value (which is 1 and it associated to the characters χ_S for $|S| = 1$).

The existence of the spectral gap for a symmetric Markov semigroup $(Q_t)_{t \in [0,\infty)}$ implies $Q_t f \to \mathbb{E}[f]$ as $t \to \infty$ and the size of the gap is responsible for the speed of convergence. This is of extreme importance in physics, and not surprisingly the Poincaré-type inequalities were considered in physics first, already in the middle of the nineteenth century.

27.2 Stroock-Varopoulos inequality

In this section we prove the Stroock-Varopoulos inequality which is an important inequality in the theory of semigroups. We start with an elementary inequality whose proof can be skipped by uninterested reader.

Lemma 27.2. *For $p > 1$ and $a, b \geqslant 0$ we have*

$$(p-2)^2(a^p + b^p) - p^2(a^{p-1}b + ab^{p-1}) + 8(p-1)a^{p/2}b^{p/2} \geqslant 0.$$

Proof. Because of the homogeneity, it suffices to prove that for $t \geqslant 1$

$$u(t) = (p-2)^2 t^p - p^2 t^{p-1} + 8(p-1)t^{p/2} - p^2 t + (p-2)^2 \geqslant 0.$$

Indeed, $u(1) = 2(p^2 - 4p + 4) - 2p^2 + 8p - 8 = 0$, and

$$u'(t) = p(p-2)^2 t^{p-1} - p^2(p-1)t^{p-2} + 4p(p-1)t^{p/2-1} - p^2,$$

so that $u'(1) = (p^3 - 4p^2 + 4p) - (p^3 - p^2) + (4p^2 - 4p) - p^2 = 0$. Now it suffices to note that

$$\begin{aligned} u''(t) &= p(p-1)(p-2)^2 t^{p-2} - p^2(p-1)(p-2)t^{p-3} + 2p(p-1)(p-2)t^{\frac{p}{2}-2} \\ &= p^2(p-1)(p-2)t^{p-2}\left(\frac{p-2}{p} + \frac{2}{p}t^{-p/2} - t^{-1}\right) \\ &= 2p(p-1)(p-2)t^{p-2}\left(\frac{2-p}{2} + \frac{p}{2}t^{-1} - t^{-p/2}\right). \end{aligned}$$

Since for $p \geqslant 2$,

$$\frac{p-2}{p} + \frac{2}{p}t^{-p/2} = \frac{p-2}{p} \cdot 1 + \frac{2}{p}t^{-p/2} \cdot t^{-p/2} \geqslant 1^{\frac{p-2}{p}} \left(t^{-p/2}\right)^{2/p} = t^{-1},$$

while for $p \in (1, 2]$,

$$\frac{2-p}{2} + \frac{p}{2}t^{-1} = \frac{2-p}{2} \cdot 1 + \frac{p}{2}t^{-1} \geqslant 1^{\frac{p-2}{2}} \left(t^{-1}\right)^{p/2} = t^{-p/2},$$

we conclude $u''(t) \geqslant 0$ and the proof is finished. $\square$

Now we will deduce the Stroock-Varopoulos inequality from Lemma 27.2. We state the proof for the semigroup P_t on the hupercube, but the same proof works for every symmetric Markov semigroup (under some additional assumptions about f).

Theorem 27.3 (Stroock-Varopoulos)**.** *For any $f : \mathbb{Z}_2^n \to [0, \infty)$, and every $p > 1$, we have*

$$\mathcal{E}(f^{p/2}) := \mathbb{E}\left[f^{p/2} L(f^{p/2})\right] \leqslant \frac{p^2}{4(p-1)} \mathbb{E}[f^{p-1} Lf].$$

Proof. By Lemma 27.2, for any $a \geqslant 0$, we have the ponitwise inequality

$$(p-2)^2(a^p + f^p) - p^2(a^{p-1}f + af^{p-1}) + 8(p-1)a^{p/2}f^{p/2} \geqslant 0.$$

Since P_t is linear and order preserving for any $t \geqslant 0$, it holds pointwise that

$$(p-2)^2(a^p + P_t(f^p)) - p^2(a^{p-1}P_t f + aP_t(f^{p-1})) + 8(p-1)a^{p/2}P_t(f^{p/2}) \geqslant 0.$$

Hence setting $a = f$ we have

$$(p-2)^2(f^p + P_t(f^p)) - p^2(f^{p-1}P_t f + fP_t(f^{p-1})) + 8(p-1)f^{p/2}P_t(f^{p/2}) \geqslant 0.$$

We can take the expected value and arrive at

$$(p-2)^2(\mathbb{E}[f^p] + \mathbb{E}[P_t(f^p)]) - p^2(\mathbb{E}[f^{p-1}P_t f] + \mathbb{E}[fP_t(f^{p-1})]) + 8(p-1)\mathbb{E}[f^{p/2}P_t(f^{p/2})] \geqslant 0.$$

Since P_t is symmetric, it preserves expectation, and the above reduces to

$$\beta(t) = 2(p-2)^2\mathbb{E}[f^p] - 2p^2\mathbb{E}[f^{p-1}P_t f] + 8(p-1)\mathbb{E}[f^{p/2}P_t(f^{p/2})] \geqslant 0. \tag{27.3}$$

Now as $P_0 = \mathrm{Id}$, we have

$$\beta(0) = (2(p-2)^2 - 2p^2 + 8(p-1))\mathbb{E}[f^p] \geqslant 0,$$

and thus (27.3) implies that $\beta'(0^+) \geqslant 0$. But as $L = -\frac{d}{dt}P_t f\big|_{0^+}$, we have

$$0 \leqslant \beta'(0^+) = 2p^2\mathbb{E}[f^{p-1}Lf] - 8(p-1)\mathbb{E}[f^{p/2}L(f^{p/2})],$$

which completes the proof. $\square$

Remark 27.4. Note that in Theorem 27.3 we have equality when $p = 2$.

Remark 27.5. Recall that for The Ornstein-Uhlenbeck semigroup on $(\mathbb{R}^n, (2\pi)^{-n/2}e^{-|x|/2}dx)$ the generator is given by

$$(Lf)(x) = \langle x, \nabla f(x)\rangle - (\Delta f)(x).$$

In this case for $f, g \in \mathcal{C}^\infty$, it is not difficult to see that

$$\mathbb{E}[f \cdot Lg] = (2\pi)^{-n/2} \int \langle \nabla f(x), \nabla g(x)\rangle f(x) e^{-|x|/2} dx = \mathbb{E}[\langle \nabla f(x), \nabla g(x)\rangle],$$

where the expectation is with respect to the Gaussian measure.

Note that in this case we will actually have equality in Theorem 27.3 for any $p > 1$.

27.3 Entropy and Logarithmic Sobolev inequalities

We start by defining the notion of entropy.

Definition 27.6. For an integrable non-negative function g on a probability space we define its entropy as

$$\mathrm{Ent}(g) = \mathbb{E}[g \ln g] - \mathbb{E}[g] \ln(\mathbb{E}[g]),$$

where we adopt a natural convention $0 \ln(0) = 0$.

Clearly, $\mathrm{Ent}[g] < \infty$ if and only if $g \ln g$ is integrable. Since $x \ln(x)$ is strictly convex, always $\mathrm{Ent}[g] \geqslant 0$, and $\mathrm{Ent}[g] = 0$ if and only if g is constant almost everywhere. Note also that

$$\mathrm{Ent}(\lambda g) = \lambda \, \mathrm{Ent}(g).$$

The logarithmic Sobolev inequality (called also entropy-energy inequality) was introduced by L. Gross. It resembles the Poincaré inequality - the variance functional on the left hand side is replaced by the entropy of the square of the function. The inequality has the form:

$$\mathrm{Ent}[f^2] \leqslant C\mathcal{E}(f).$$

Both sides of this inequality measure how far f is from being constant. Note that for a constant f, both $\mathrm{Ent}[f^2]$ and $\mathcal{E}(f)$ are 0.

Definition 27.7. A symmetric Markov semigroup $(Q_t)_{t\in[0,\infty)}$ on Ω, with an invariant measure μ and a self-adjoint (with respect to the $L_2(\Omega,\mu)$ structure) generator L, satisfies the *logarithmic Sobolev* inequality with constant $C > 0$ if for every function f belonging to the domain of L, we have

$$\mathbb{E}_\mu[f^2 \ln(f^2)] - \mathbb{E}_\mu[f^2] \ln \mathbb{E}_\mu[f^2] \leqslant C\mathbb{E}_\mu[fLf].$$

It turns out that logarithmic Sobolev inequalities are equivalent to hyper-contractive inequalities. Recall that in Theorem 10.13 we showed that for $1 < p \leqslant q < \infty$, and $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$, we always have

$$\|T_\rho f\|_q \leqslant \|f\|_p.$$

Using our semigroup notation, we can rewrite this as $\|P_t f\|_q \leqslant \|f\|_p$ for $0 \leqslant t \leqslant \frac{1}{2}(\ln(p-1) - \ln(q-1))$. A semigroup $(Q_t)_{t\in[0,\infty)}$ is (p,q)-hypercontractive with parameter $t(p,q)$ if for every f in the domain and every $0 \leqslant t \leqslant t(p,q)$ we have

$$\|Q_t f\|_q \leqslant \|f\|_p.$$

Theorem 27.8 (Gross)**.** *A symmetric generator L satisfies the logarithmic Sobolev inequality with constant C if and only if for all $p > q > 1$ the semigroup $(P_t)_{t\in[0,\infty}$ generated by L is (p,q)-hypercontractive with $t(p,q) = \frac{C}{4}(\ln(p-1) - \ln(q-1))$.*

Theorem 27.8 combined with the hypercontractive estimates that we obtained in Theorem 10.13 show that the semigroup $(P_t)_{t\in[0,\infty)}$ on the hypercube satisfies the logarithmic Sobolev inequality with constant 2, i.e. for every $f : \mathbb{Z}_2^n \to \mathbb{R}$, we have

$$\mathbb{E}[f^2 \ln(f^2)] - \mathbb{E}[f^2] \ln \mathbb{E}[f^2] \leqslant 2\mathbb{E}[fLf].$$

In order to prove Theorem 27.8 we first need the following lemma whose proof is based on the Stroock-Varopoulos theorem.

Lemma 27.9. *The following statements are equivalent:*

(a): *For every $f : \mathbb{Z}_2^n \to \mathbb{R}$,*

$$\mathbb{E}[f^2 \ln(f^2)] - \mathbb{E}[f^2] \ln \mathbb{E}[f^2] \leqslant C\mathbb{E}[fLf].$$

(b): *For every nonnegative $f : \mathbb{Z}_2^n \to \mathbb{R}$,*

$$\mathbb{E}[f^2 \ln(f^2)] - \mathbb{E}[f^2] \ln \mathbb{E}[f^2] \leqslant C\mathbb{E}[fLf].$$

(c): *For every nonnegative $f : \mathbb{Z}_2^n \to \mathbb{R}$, and every $p > 1$,*

$$\mathbb{E}[f^p \ln(f^p)] - \mathbb{E}[f^p] \ln \mathbb{E}[f^p] \leqslant \frac{Cp^2}{4(p-1)} \mathbb{E}[f^{p-1} Lf].$$

Proof. Obviously (a) implies (b), and also setting $P = 2$ in (c) we recover (b). So it suffices to show that (b) implies (a) and (c).

(b) $\Rightarrow$ (a): This follows from $\mathcal{E}(|f|) \leqslant \mathcal{E}(f)$ which we proved in Section 27.0.1.

(b) $\Rightarrow$ (c): This follows immediately from applying (b) to $f^{p/2}$ and then using the Stroock-Varopoulos inequality (Theorem 27.3). □

Proof of Theorem 27.8. For $p \geqslant q > 1$, define $t_q(p) = \frac{C}{4} \ln \frac{p-1}{q-1}$. Consider a nonnegative function $f \in L_2$, and set

$$\phi_q(p) = \ln \|P_{t_q(p)} f\|_q = \frac{1}{p} \mathbb{E}\left[\ln |P_{t_q(p)} f|^p\right].$$

Note that $t(q, q) = 0$ and thus $\phi_q(q) = \ln \|f\|_q$. Hence hypercontractivity is equivalent to $\phi_q(p) \leqslant \phi_q(q)$ for $p \geqslant q$. For $p \geqslant q$ denote

$$f_p := P_{t_q(p)} f \geqslant 0.$$

Using $\frac{d}{dt} P_t f = -L(p_t f)$, we obtain

$$\frac{d}{dp} f_p^p = \frac{1}{p} f_p^p \ln(f_p^p) - \frac{Cp}{4(p-1)} f_p^{p-1} L(f_p).$$

This shows

$$\begin{aligned}
\frac{d}{dp}\phi_q(p) &= \frac{1}{p} \frac{\mathbb{E}[\frac{d}{dp}(f_p^p)]}{\mathbb{E}[f_p^p]} - \frac{1}{p^2} \ln \mathbb{E}[f_p^p] \\
&= \frac{1}{p^2} \frac{\mathbb{E}[f_p^p \ln(f_p^p)]}{\mathbb{E}[f_p^p]} - \frac{C}{4(p-1)} \frac{\mathbb{E}[f_p^{p-1} L(f_p)]}{\mathbb{E}[f_p^p]} - \frac{1}{p^2} \ln \mathbb{E}[f_p^p] \\
&= \frac{1}{p^2 \mathbb{E}[f_p^p]} \left(\mathrm{Ent}(f_p^p) - \frac{Cp^2}{4(p-1)} \mathbb{E}[f_p^{p-1} L(f_p)] \right).
\end{aligned}$$

Hence

$$\frac{d}{dp}\phi_q(p) \leqslant 0 \iff \mathrm{Ent}(f_p^p) \leqslant \frac{Cp^2}{4(p-1)} \mathbb{E}[f_p^{p-1} L(f_p)]$$

Thus $\phi_q(p)$ is decreasing if the semigroup satisfies the logarithmic Sobolev inequality with constant C, and we obtain the desired hyper-contractive estimates.

To deduce the logarithmic Sobolev inequality from hyper-contractivity, it suffices to notice that if hypercontractivity holds, then $\frac{d}{dp}\phi_q(p)\Big|_{p=q} \leqslant 0$. Since $f_q = f$, this gives

$$\mathrm{Ent}(f^q) \leqslant \frac{Cq^2}{4(q-1)} \mathbb{E}[f^{q-1} L(f)],$$

which verifies the logarithmic Sobolev inequality by setting $q = 2$. □

Exercise 27.1. This exercises shows that the logarithmic Sobolev inequality is stronger than the Poincaré inequality (the converse is not true). Show that if a semigroup satisfies the logarithmic Sobolev inequality with constant C, then it satisfies the Poincaré inequality with constant $2C$.

27.3.1 Tensorization of logarithmic Sobolev inequality

Recall that in Chapter 10 to prove the hypercontractivity for the noise operator, first we proved it for dimension 1 and then used generalized Minkowski's inequality to show that the inequality tensorizes. Theorem 27.8 shows that hypercontractivy us equivalent to the logarithmic Sobolev inequality. This suggest that the logarithmic sobolev

inequlity must also tensorize. Indeed there is also a standard method of tensorizing both Poincaré and logarithmic Sobolev inequalities by using the subadditivity of the variance and entropy functionals.

Thus the logorithmic sobolev inequalty and hypercontractive estimates on the cube could also be obtained by proving the logarithmic Sobolev inequality on $0, 1$ and then deducing it on the general cube via subadditivity. For $f : \{0,1\}^n \to [0,\infty)$, and $i \in [n]$ define the coordinate-wise entropy as

$$\mathrm{Ent}_i(f) = \mathbb{E}_{x_{[n]\setminus\{i\}}}\left[\mathrm{Ent}\, f_{x_{[n]\setminus\{i\}}}(x_i)\right],$$

where $f_{x_{[n]\setminus\{i\}}} : x_i \mapsto f(x_1, \ldots, x_n)$.

Lemma 27.10 (Subadditivity of Entropy)**.** *For $f : \mathbb{Z}_2^n \to [0,\infty)$, we have*

$$\mathrm{Ent}(f) \leqslant \sum_{i=1}^{n} \mathrm{Ent}_i(f).$$

Exercise 27.2. Prove the variational formulation of entropy:

$$\mathrm{Ent}(f) = \sup\{\langle f, g\rangle : \mathbb{E}[e^g] \leqslant 1,\ g : \mathbb{Z}_2^n \to \mathbb{R}\},$$

for every $f : \mathbb{Z}_2^n \to [0,\infty)$.

Exercise 27.3. Prove Lemma 27.10 using the variational formulation of entropy.

Exercise 27.4. Use 27.10 to show that the logarithmic sobolev inequality tensorizes. That is if it holds with constant C for nonegative functions on $\mathbb{Z}_2$, then it holds with constant C for nonnegative functions on $\mathbb{Z}_2^n$.

Exercise 27.5. Use the subadditivity of variance to show that the Poincaré inequality tensorizes. That is if it holds with constant C for nonegative functions on $\mathbb{Z}_2$, then it holds with constant C for nonnegative functions on $\mathbb{Z}_2^n$.

Bibliography

[AA14] Scott Aaronson and Andris Ambainis, *The need for structure in quantum speedups*, Theory Comput. **10** (2014), 133–166. 77

[AF99] Dimitris Achlioptas and Ehud Friedgut, *A sharp threshold for k-colorability*, Random Structures Algorithms **14** (1999), no. 1, 63–70. 96

[AGHP92] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta, *Simple constructions of almost k-wise independent random variables*, Random Structures & Algorithms **3** (1992), no. 3, 289–304. 151

[Ajt83] M. Ajtai, Σ_1^1*-formulae on finite structures*, Ann. Pure Appl. Logic **24** (1983), no. 1, 1–48. MR 706289 (85b:03048) 127

[AL93] Miklós Ajtai and Nathal Linial, *The influence of large coalitions*, Combinatorica **13** (1993), no. 2, 129–145. MR 1237037 (94k:05018) 76

[Baz09] Louay M. J. Bazzi, *Polylogarithmic independence can fool DNF formulas*, SIAM J. Comput. **38** (2009), no. 6, 2220–2272. 154

[BBS86] L. Blum, M. Blum, and M. Shub, *A simple unpredictable pseudorandom number generator*, SIAM J. Comput. **15** (1986), no. 2, 364–383. MR 837589 149

[Bec75] William Beckner, *Inequalities in Fourier analysis*, Ann. of Math. (2) **102** (1975), no. 1, 159–182. MR 385456 59

[Beh46] F. A. Behrend, *On sets of integers which contain no three terms in arithmetical progression*, Proc. Nat. Acad. Sci. U.S.A. **32** (1946), 331–332. 25

[BK99] Jean Bourgain and Gil Kalai, *Threshold intervals under group symmetries*, Convex geometric analysis (Berkeley, CA, 1996), Math. Sci. Res. Inst. Publ., vol. 34, Cambridge Univ. Press, Cambridge, 1999, pp. 59–63. 75, 91

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm, *Noise sensitivity of Boolean functions and applications to percolation*, Inst. Hautes Études Sci. Publ. Math. (1999), no. 90, 5–43 (2001). MR 1813223 (2001m:60016) 132

[BLR90] M. Blum, M. Luby, and R. Rubinfeld, *Self-testing/correcting with applications to numerical problems*, STOC '90: Proceedings of the twenty-second annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 1990, pp. 73–83. 21, 22

[Bon70] Aline Bonami, *Étude des coefficients de fourier des fonctions de* $l^p(g)$, Annales de l'institut Fourier **20** (1970), no. 2, 335–402. 59

[Bop97] Ravi B. Boppana, *The average sensitivity of bounded-depth circuits*, Inform. Process. Lett. **63** (1997), no. 5, 257–261. MR 1475339 (98f:68093) 131, 132

[Bou80] J. Bourgain, *Walsh subspaces of* L^p*-product spaces*, Seminar on Functional Analysis, 1979–1980 (French), École Polytech., Palaiseau, 1980, pp. Exp. No. 4A, 9. MR 604387 85

[Bou99a] ———, *On triples in arithmetic progression*, Geom. Funct. Anal. **9** (1999), no. 5, 968–984. 25, 28

[Bou99b] Jean Bourgain, *An appendix to "Sharp thresholds of graph properties, and the k-sat problem" by E. Friedgut*, J. Amer. Math. Soc. **12** (1999), no. 4, 1017–1054. 92, 95, 96

[Bra10] Mark Braverman, *Polylogarithmic independence fools* AC^0 *circuits*, J. ACM **57** (2010), no. 5, 28:1–28:10. MR 2683586 154, 155

[BS21a] Nikhil Bansal and Makrand Sinha, *k-Forrelation optimally separates quantum and classical query complexity*, 53rd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2021, ACM, 2021, pp. 1303–1316. 139

[BS21b] Thomas F. Bloom and Olof Sisask, *Breaking the logarithmic barrier in roth's theorem on arithmetic progressions*, 2021, https://arxiv.org/abs/2007.03528. 25

[BS23] ———, *An improvement to the Kelley-Meka bounds on three-term arithmetic progressions*, 2023, https://arxiv.org/abs/2309.02353. 25

[BT87] B. Bollobás and A. Thomason, *Threshold functions*, Combinatorica **7** (1987), no. 1, 35–38. 80

[CGL+21] Eshan Chattopadhyay, Jason Gaitonde, Chin Ho Lee, Shachar Lovett, and Abhishek Shetty, *Fractional Pseudorandom Generators from Any Fourier Level*, Proc. 36th Computational Complexity Conference (CCC), 2021, pp. 10:1–10:24. 163

[Cha02] Mei-Chu Chang, *A polynomial bound in Freiman's theorem*, Duke Math. J. **113** (2002), no. 3, 399–419. 67, 68

[CHHL19a] Eshan Chattopadhyay, Pooya Hatami, Kaave Hosseini, and Shachar Lovett, *Pseudorandom generators from polarizing random walks*, Theory of Computing **15** (2019), no. 10, 1–26. 139

[CHHL19b] ———, *Pseudorandom generators from polarizing random walks*, Theory Comput. **15** (2019), no. 1, 1–26. 157, 158

[CHLT18] Eshan Chattopadhyay, Pooya Hatami, Shachar Lovett, and Avishay Tal, *Pseudorandom Generators from the Second Fourier Level and Applications to AC0 with Parity Gates*, Proc. 10th Conference on Innovations in Theoretical Computer Science (ITCS), 2018, pp. 22:1–22:15. 163

[CLP17] Ernie Croot, Vsevolod F. Lev, and Péter Pál Pach, *Progression-free sets in* $\mathbb{Z}_4^n$ *are exponentially small*, Ann. of Math. (2) **185** (2017), no. 1, 331–337. 27

[Coh60] Paul J. Cohen, *On a conjecture of Littlewood and idempotent measures*, Amer. J. Math. **82** (1960), 191–212. 139, 144

[EG17] Jordan S. Ellenberg and Dion Gijswijt, *On large subsets of* $\mathbb{F}_q^n$ *with no three-term arithmetic progression*, Ann. of Math. (2) **185** (2017), no. 1, 339–343. 27

[EoR59] P. Erd″ os and A. Rényi, *On random graphs. I*, Publ. Math. Debrecen **6** (1959), 290–297. 79, 91

[ES81] B. Efron and C. Stein, *The jackknife estimate of variance*, Ann. Statist. **9** (1981), no. 3, 586–596. 57

[ET36] Paul Erdös and Paul Turán, *On Some Sequences of Integers*, J. London Math. Soc. **11** (1936), no. 4, 261–264. MR 1574918 25

[FF81] P. Frankl and Z. Füredi, *A short proof for a theorem of Harper about Hamming-spheres*, Discrete Math. **34** (1981), no. 3, 311–313. 55

[FGHK16] Magnus Gausdal Find, Alexander Golovnev, Edward A. Hirsch, and Alexander S. Kulikov, *A better-than-3n lower bound for the circuit complexity of an explicit function*, 57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016, IEEE Computer Soc., Los Alamitos, CA, 2016, pp. 89–98. MR 3630969 126

[FK96] Ehud Friedgut and Gil Kalai, *Every monotone graph property has a sharp threshold*, Proc. Amer. Math. Soc. **124** (1996), no. 10, 2993–3002. 83, 91, 136

[FK18] Michael A. Forbes and Zander Kelley, *Pseudorandom generators for read-once branching programs, in any order*, 59th IEEE Annual Symposium on Foundations of Computer Science, STOC 2018, IEEE Computer Society, 2018, pp. 946–955. 139

[FKN02] Ehud Friedgut, Gil Kalai, and Assaf Naor, *Boolean functions whose Fourier transform is concentrated on the first two levels*, Adv. in Appl. Math. **29** (2002), no. 3, 427–437. 67, 68, 69

[Fri98] Ehud Friedgut, *Boolean functions with low average sensitivity depend on few coordinates*, Combinatorica **18** (1998), no. 1, 27–35. 71, 72, 91

[Fri99] ———, *Sharp thresholds of graph properties, and the k-sat problem*, J. Amer. Math. Soc. **12** (1999), no. 4, 1017–1054, With an appendix by Jean Bourgain. 91, 92, 96, 97

[Fri04] ———, *Influences in product spaces: KKL and BKKKL revisited*, Combin. Probab. Comput. **13** (2004), no. 1, 17–29. 76

[Fri05] ———, *Hunting for sharp thresholds*, Random Structures Algorithms **26** (2005), no. 1-2, 37–51. 91, 96

[FSS84] Merrick Furst, James B. Saxe, and Michael Sipser, *Parity, circuits, and the polynomial-time hierarchy*, Math. Systems Theory **17** (1984), no. 1, 13–27. MR 738749 (86e:68048) 127

[GGMT23] W. T. Gowers, Ben Green, Freddie Manners, and Terence Tao, *On a conjecture of Marton*, 2023. 144

[GGR98] Oded Goldreich, Shari Goldwasser, and Dana Ron, *Property testing and its connection to learning and approximation*, J. ACM **45** (1998), no. 4, 653–750. 21

[GL89] O. Goldreich and L. A. Levin, *A hard-core predicate for all one-way functions*, Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '89, Association for Computing Machinery, 1989, p. 25–32. 121

[GL92] C. Gotsman and N. Linial, *The equivalence of two problems on the cube*, J. Combin. Theory Ser. A **61** (1992), no. 1, 142–146. 51

[Gow01] W. T. Gowers, *A new proof of Szemerédi's theorem*, Geom. Funct. Anal. **11** (2001), no. 3, 465–588. 25, 33, 34

[Gro75] Leonard Gross, *Logarithmic Sobolev inequalities*, Amer. J. Math. **97** (1975), no. 4, 1061–1083. 59

[GS08a] Ben Green and Tom Sanders, *Boolean functions with small spectral norm*, Geometric and Functional Analysis **18** (2008), no. 1, 144–162. 144

[GS08b] ———, *Boolean functions with small spectral norm*, Geom. Funct. Anal. **18** (2008), no. 1, 144–162. MR 2399099 (2009d:11039) 144

[GS08c] ———, *A quantitative version of the idempotent theorem in harmonic analysis*, Ann. of Math. (2) **168** (2008), no. 3, 1025–1054. 139, 144

[GT10] Ben Green and Terence Tao, *Linear equations in primes*, Ann. of Math. (2) (2010), 1753–1850. 34

[GTW21] Uma Girish, Avishay Tal, and Kewen Wu, *Fourier growth of parity decision trees*, 36th Computational Complexity Conference, LIPIcs, vol. 200, 2021, pp. 39:1–39:36. 139

[GW10] W. T. Gowers and J. Wolf, *The true complexity of a system of linear equations*, Proc. Lond. Math. Soc. (3) **100** (2010), no. 1, 155–176. 34

[GW11] ———, *Linear forms and higher-degree uniformity for functions on* $\mathbb{F}_p^n$, Geom. Funct. Anal. **21** (2011), no. 1, 36–69. 34

[Har66] L. H. Harper, *Optimal numberings and isoperimetric problems on graphs*, J. Combinatorial Theory **1** (1966), 385–393. 55

[Has86a] J Hastad, *Almost optimal lower bounds for small depth circuits*, Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '86, ACM, 1986, pp. 6–20. 127

[Has86b] Johan Hastad, *Almost Optimal Lower Bounds for Small Depth Circuits*, Proceedings of the 18th Annual ACM Symposium on Theory of Computing, ACM, 1986, pp. 6–20. 131

[Hat12] Hamed Hatami, *A structure theorem for Boolean functions with small total influences*, Ann. of Math. (2) **176** (2012), no. 1, 509–533. 92

[HB87] D. R. Heath-Brown, *Integer sets containing no arithmetic progressions*, J. London Math. Soc. (2) **35** (1987), no. 3, 385–394. 25

[Hel53] Henry Helson, *Note on harmonic functions*, Proc. Amer. Math. Soc. **4** (1953), no. 5, 686–691. 139

[HH24] Pooya Hatami and William Hoza, *Paradigms for unconditional pseudorandom generators*, Found. Trends Theor. Comput. Sci. **16** (2024), no. 1-2, 1–210. MR 4707623 149, 157

[HHH23] Lianna Hambardzumyan, Hamed Hatami, and Pooya Hatami, *Dimension-free bounds and structural results in communication complexity*, Israel Journal of Mathematics **253** (2023), no. 2, 555–616. 148

[Hoe48] Wassily Hoeffding, *A class of statistics with asymptotically normal distribution*, Ann. Math. Statistics **19** (1948), 293–325. 56

[Hos86] B. Host, *Le théorème des idempotents dans $B(G)$*, Bull. Soc. Math. France **114** (1986), no. 2, 215–223. 139

[HS19] Prahladh Harsha and Srikanth Srinivasan, *On polynomial approximations to AC^0*, Random Structures Algorithms **54** (2019), no. 2, 289–303. 154, 155

[Hua19] Hao Huang, *Induced subgraphs of hypercubes and a proof of the sensitivity conjecture*, Ann. of Math. (2) **190** (2019), no. 3, 949–955. 46, 49, 50

[IW97] Russell Impagliazzo and Avi Wigderson, *P = bpp if e requires exponential circuits: Derandomizing the xor lemma*, Proc. 29th Annual ACM Symposium on Theory of Computing (STOC), 1997, pp. 220–229. 150

[Kel21] Zander Kelley, *An improved derandomization of the switching lemma*, Proc. 53rd Annual ACM Symposium on Theory of Computing (STOC), 2021, p. 272–282. 156

[KI40] Yukiyosi Kawada and Kiyosi Itô, *On the probability distribution on a compact group. I*, Proc. Phys.-Math. Soc. Japan (3) **22** (1940), 977–998. MR 3462 139

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial, *The influence of variables on Boolean functions*, 29th Annual Symposium on Foundations of Computer Science, IEEE Comput. Soc. Press, Washington, DC, [1988] ©1988, pp. 68–80. 71, 74

[KKL+20] Esty Kelman, Guy Kindler, Noam Lifshitz, Dor Minzer, and Muli Safra, *Towards a proof of the Fourier-entropy conjecture?*, Geom. Funct. Anal. **30** (2020), no. 4, 1097–1138. MR 4153910 137

[KM93] Eyal Kushilevitz and Yishay Mansour, *Learning decision trees using the Fourier spectrum*, SIAM J. Comput. **22** (1993), no. 6, 1331–1348. 139

[KM23] Zander Kelley and Raghu Meka, *Strong bounds for 3-progressions*, 2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2023, pp. 933–973. 25, 68

[Lef72] Marcel Lefranc, *Sur certaines algèbres de fonctions sur un groupe*, C. R. Acad. Sci. Paris Sér. A-B **274** (1972), A1882–A1883. 139

[Liv95] Leo Livshits, *A note on* 0-1 *Schur multipliers*, Linear Algebra Appl. **222** (1995), 15–22. MR 1332920 145, 147

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan, *Constant depth circuits, Fourier transform, and learnability*, J. Assoc. Comput. Mach. **40** (1993), no. 3, 607–620. MR 1370363 (96h:68074) 120, 133, 154

[LN90] Nathan Linial and Noam Nisan, *Approximate inclusion-exclusion*, Combinatorica **10** (1990), no. 4, 349–365. 154

[Lov15] Shachar Lovett, *An exposition of sanders' quasi-polynomial freiman-ruzsa theorem*, no. 6, 1–14. 68

[LPV22] Chin Ho Lee, Edward Pyne, and Salil Vadhan, *Fourier Growth of Regular Branching Programs*, Proc. 26th International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM), 2022, pp. 2:1–2:21. 162

[LSS24] James Leng, Ashwin Sah, and Mehtaab Sawhney, *Improved bounds for Szemerédi's theorem*, 2024, https://arxiv.org/abs/2402.17995. 25

[Lup58] O. B. Lupanov, *The synthesis of contact circuits*, Dokl. Akad. Nauk SSSR (N.S.) **119** (1958), 23–26. MR 101176 126

[LV96] M. Luby and B. Veličković, *On deterministic approximation of DNF*, Algorithmica **16** (1996), no. 4/5, 415–433. 156

[Lyu22] Xin Lyu, *Improved Pseudorandom Generators for AC^0 Circuits*, Proc. 37th Computational Complexity Conference (CCC), 2022, pp. 34:1–34:25. 156

[Man95] Yishay Mansour, *An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution*, J. Comput. System Sci. **50** (1995), no. 3, part 3, 543–550, Fifth Annual Workshop on Computational Learning Theory (COLT) (Pittsburgh, PA, 1992). MR 1339562 (96h:68171) 136

[Mar74] G. A. Margulis, *Probabilistic characteristics of graphs with large connectivity*, Problemy Peredači Informacii **10** (1974), no. 2, 101–108. 79, 81

[Mes95] Roy Meshulam, *On subsets of finite abelian groups with no 3-term arithmetic progressions*, Journal of Combinatorial Theory, Series A **71** (1995), no. 1, 168–172. 25

[MO09] Ashley Montanaro and Tobias Osborne, *On the communication complexity of xor functions*, CoRR **abs/0909.3392** (2009). 147

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz, *Noise stability of functions with low influences: invariance and optimality*, Ann. of Math. (2) **171** (2010), no. 1, 295–341. 102, 105, 113

[MOR+06] Elchanan Mossel, Ryan O'Donnell, Oded Regev, Jeffrey E. Steif, and Benny Sudakov, *Non-interactive correlation distillation, inhomogeneous Markov chains, and the reverse Bonami-Beckner inequality*, Israel J. Math. **154** (2006), 299–336. MR 2254545 100, 101

[MOS03] Elchanan Mossel, Ryan O'Donnell, and Rocco P. Servedio, *Learning juntas*, Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing, ACM, New York, 2003, pp. 206–212. MR 2121045 119

[MP88] Marvin L. Minsky and Seymour A. Papert, *Perceptrons: expanded edition*, MIT Press, Cambridge, MA, USA, 1988. 43

[MR209] *24th Annual IEEE Conference on Computational Complexity*, IEEE Computer Society, Los Alamitos, CA, 2009. MR 2920336 134

[MRT19] Raghu Meka, Omer Reingold, and Avishay Tal, *Pseudorandom generators for width-3 branching programs*, Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, 2019, p. 13–23. 139

[Nis91] Noam Nisan, *CREW PRAMs and decision trees*, SIAM J. Comput. **20** (1991), no. 6, 999–1007. 45

[NN93] Joseph Naor and Moni Naor, *Small-bias probability spaces: efficient constructions and applications*, SIAM J. Comput. **22** (1993), no. 4, 838–856. MR 1227764 151, 153

[NS94] Noam Nisan and Márió Szegedy, *On the degree of Boolean functions as real polynomials*, vol. 4, 1994, Special issue on circuit complexity (Barbados, 1992), pp. 301–313. 44, 45, 46

[OW07] Ryan O'Donnell and Karl Wimmer, *Approximation by DNF: examples and counterexamples*, Automata, languages and programming, Lecture Notes in Comput. Sci., vol. 4596, Springer, Berlin, 2007, pp. 195–206. MR 2424683 (2009f:68076) 132

[Per90] Rene Peralta, *On the randomness complexity of algorithms*, University of Wisconsin, Milwaukee CS Research Report TR 90-1 (1990). 151

[Raz87] A.A. Razborov, *Lower bounds on the size of bounded depth circuits over a complete basis with logical addition*, Mathematical notes of the Academy of Sciences of the USSR **41** (1987), no. 4, 333–338 (English). 133, 134, 135, 154

[Raz09] Alexander Razborov, *A simple proof of Bazzi's theorem*, ACM Transactions on Computation Theory**1** (2009), no. 1. 154

[Rot53] Klaus F Roth, *On certain sets of integers*, J. London Math. Soc **28** (1953), no. 104-109, 3. 25

[RS93] Ronitt Rubinfeld and Madhu Sudan, *Robust characterizations of polynomials and their applications to program testing*, Tech. report, Ithaca, NY, USA, 1993. 21

[RSV13] Omer Reingold, Thomas Steinke, and Salil P. Vadhan, *Pseudorandomness for regular branching programs via fourier analysis*, APPROX-RANDOM 2013, Lecture Notes in Computer Science, vol. 8096, Springer, 2013, pp. 655–670. 139

[RT19] Ran Raz and Avishay Tal, *Oracle separation of BQP and PH*, Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, 2019, p. 13–23. 139

[RT22] ———, *Oracle separation of BQP and PH*, J. ACM **69** (2022), no. 4, Art. 30, 21. MR 4483114 157

[Rud59a] Walter Rudin, *Idempotent measures on Abelian groups*, Pacific J. Math. **9** (1959), 195–209. 139

[Rud59b] ———, *Measure algebras on abelian groups*, Bull. Amer. Math. Soc. **65** (1959), 227–247. 139

[Rud90] ———, *Fourier analysis on groups*, Wiley Classics Library, John Wiley & Sons, Inc., New York, 1990, Reprint of the 1962 original, A Wiley-Interscience Publication. MR 1038803 144

[Run07] Volker Runde, *Cohen-Host type idempotent theorems for representations on Banach spaces and applications to Figà-Talamanca-Herz algebras*, J. Math. Anal. Appl. **329** (2007), no. 1, 736–751. 139

[Rus81] Lucio Russo, *On the critical percolation probabilities*, Z. Wahrsch. Verw. Gebiete **56** (1981), no. 2, 229–237. 79, 81, 91

[Rus82] ———, *An approximate zero-one law*, Z. Wahrsch. Verw. Gebiete **61** (1982), no. 1, 129–139. 91

[San11a] Tom Sanders, *On Roth's theorem on progressions*, Ann. of Math. (2) **174** (2011), no. 1, 619–636. 25, 68

[San11b] ———, *A quantitative version of the non-abelian idempotent theorem*, Geom. Funct. Anal. **21** (2011), no. 1, 141–221. 139, 148

[San12] ———, *On the Bogolyubov-Ruzsa lemma*, Anal. PDE **5** (2012), no. 3, 627–655. 68

[San19] ———, *Boolean functions with small spectral norm, revisited*, Math. Proc. Cambridge Philos. Soc. **167** (2019), no. 2, 335–344. 144

[San20] ———, *Bounds in Cohen's idempotent theorem*, J. Fourier Anal. Appl. **26** (2020), no. 2, Paper No. 25, 64. 139, 144

[San21] ———, *Coset decision trees and the Fourier algebra*, J. Anal. Math. **144** (2021), no. 1, 227–259. 139

[Sha49] Claude E. Shannon, *The synthesis of two-terminal switching circuits*, Bell System Tech. J. **28** (1949), 59–98. 125, 126

[SlV13] Amir Shpilka and Ben lee Volk, *On the structure of boolean functions with small spectral norm.*, Electronic Colloquium on Computational Complexity (ECCC) **20** (2013), 49. 145, 147

[Smo87] Roman Smolensky, *Algebraic methods in the theory of lower bounds for boolean circuit complexity*, STOC, 1987, pp. 77–82. 133, 134, 135, 154

[ST19] Rocco A. Servedio and Li-Yang Tan, *Improved Pseudorandom Generators from Pseudorandom Multi-Switching Lemmas*, Proc. 28th International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM), 2019, pp. 45:1–45:23. 156

[STV17] Amir Shpilka, Avishay Tal, and Ben lee Volk, *On the structure of Boolean functions with small spectral norm*, Comput. Complexity **26** (2017), no. 1, 229–273. MR 3620782 139

[Sze75] Endre Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith **27** (1975), no. 199-245, 2. 25

[Sze90] E. Szemerédi, *Integer sets containing no arithmetic progressions*, Acta Math. Hungar. **56** (1990), no. 1-2, 155–158. 25, 28

[Tal94] Michel Talagrand, *On Russo's approximate zero-one law*, Ann. Probab. **22** (1994), no. 3, 1576–1587. 82, 91

[Tal17] Avishay Tal, *Tight Bounds on the Fourier Spectrum of AC0*, Proc. 32nd Computational Complexity Conference (CCC), 2017, pp. 15:1–15:31. 139, 154, 155, 156, 162

[Tal20] ———, *Towards optimal separations between quantum and randomized query complexities*, 61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, 2020, pp. 228–239. 139

[TWXZ13] Hing Yin Tsang, Chung Hoi Wong, Ning Xie, and Shengyu Zhang, *Fourier sparsity, spectral norm, and the Log-rank conjecture*, 54th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2013, 2013, pp. 658–667. 139, 147

[TX13] Luca Trevisan and Tongke Xue, *A derandomized switching lemma and an improved derandomization of AC0*, Proc. 28th Annual IEEE Conference on Computational Complexity (CCC), 2013, pp. 242–247. 156

[Val84] L. G. Valiant, *A theory of the learnable*, Commun. ACM **27** (1984), no. 11, 1134–1142. 119

[Wen54] J. G. Wendel, *Haar measure and the semigroup of measures on a compact group*, Proc. Amer. Math. Soc. **5** (1954), 923–929. 139

[Yao85] Andrew Chi-Chih Yao, *Separating the polynomial-time hierarchy by oracles*, Proceedings of the 26th Annual Symposium on Foundations of Computer Science (Washington, DC, USA), SFCS '85, IEEE Computer Society, 1985, pp. 1–10. 127

[ZS10] Zhiqiang Zhang and Yaoyun Shi, *On the parity complexity measures of boolean functions.*, Theor. Comput. Sci. **411** (2010), no. 26-28, 2612–2618. 147