COMP760: Harmonic Analysis of Boolean functions

Hamed Hatami

School of Computer Science, McGill University, Montréal, Canada
E-mail address: hatami@cs.mcgill.ca

Contents

CHAPTER 1

Basic Functional Analysis

The aim of this lecture is to introduce the necessary definitions, notations, and basic results from measure theory, and functional analysis for this course.

1.1. Some basic inequalities

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

THEOREM 1.1.1. *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*

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

In 1906 Jensen founded the theory of convex functions. This enabled him to prove a considerable extension of the AM-GM inequality. Recall that a subset D of a real vector space is called *convex* if every convex linear combination of a pair of points of D is in 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, a function $f : D \to \mathbb{R}$ is called *convex* if for every $t \leq [0,1]$,

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

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

Trivially f is convex if and only if $\{(x,y) \in D \times \mathbb{R} : y \geq f(x)\}$ is convex. Also note that $f : D \to \mathbb{R}$ is convex if and only if $f_{xy} : [x,y] \to \mathbb{R}$ defined as $f_{xy} : tx + (1-t)y \mapsto tf(x) + (1-t)f(y)$ is convex. By Rolle's theorem if f_{xy} is twice differentiable, then this is equivalent to $f''_{xy} \geq 0$.

A function $f : D \to \mathbb{R}$ is *concave* if $-f$ is convex. The following important inequality is often called Jensen's inequality.

THEOREM 1.1.2. *If $f : D \to \mathbb{R}$ is a concave function, then for every $x_1, \ldots, x_n \in D$ and $t_1, \ldots, t_n \geq 0$ with $\sum_{i=1}^n t_i = 1$ we have*

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

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

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

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

THEOREM 1.1.4 (Hölder's inequality). *Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be complex numbers, and $p, q > 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

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

The numbers p and q appearing in Theorem 1.1.4 are called *conjugate exponents*. In fact 1 and ∞ are also called conjugate exponents, and Hölder's inequality in this case becomes:

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

The next theorem is called Minkowski's inequality.

THEOREM 1.1.5 (Minkowski's inequality). *If $p > 1$ is a real number, and $x_1, \ldots, x_n$ are complex numbers, then*

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

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

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

1.2. Measure spaces

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

- It includes $\emptyset$. That is, we have $\emptyset \in \mathcal{F}$.
- It is closed under complementation. That is, if $A \in \mathcal{F}$, then the complement of A also belongs to $\mathcal{F}$.
- It is closed under countable unions of its members. That is, if $A_1, A_2, \ldots$ belong to $\mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

EXAMPLE 1.2.1. Let Ω be an arbitrary set. Then the family consisting only of the empty set and the set Ω 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. ■

There is a natural partial order between σ-algebras 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 $\mathcal{F}_1$ is *finer* than $\mathcal{F}_2$, or that $\mathcal{F}_2$ is *coarser* than $\mathcal{F}_1$. Note that the trivial σ-algebra is the coarsest σ-algebra over Ω, whilst the maximal σ-algebra is the finest σ-algebra over Ω.

DEFINITION 1.2.2. *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 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 called a probability measure.*

EXAMPLE 1.2.3. The counting measure on Ω is defined in the following way. The measure of a subset is taken to be the number of elements in the subset, if the subset is finite, and ∞ if the subset is infinite. ■

A measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$ is called *σ-finite*, if Ω is the countable union of measurable sets of finite measure.

Every measure space in this 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 one specifies an "algebra" of elements of Ω which generates $\mathcal{F}$.

DEFINITION 1.2.4. *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 minimal σ-algebra containing $\mathcal{A}$ is called the σ-algebra generated by $\mathcal{A}$.

EXAMPLE 1.2.5. Let $\mathcal{A}$ be the set of all finite unions of (open, closed, or half-open) intervals in $\mathbb{R}$. Then $\mathcal{A}$ is an algebra over $\mathbb{R}$. ■

Before proceeding, let us mention that $\mu : \mathcal{A} \to [0, \infty) \cup \{+\infty\}$ is called a measure over $\mathcal{A}$ if for every finite set of $E_1, \ldots, F_m \in \mathcal{A}$, we have

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

The following theorem, due to Carathéodory, is one of the basic theorems in measure theory. It says that if the measure μ is defined on the algebra, then we can automatically extend it to the σ-algebra generated by $\mathcal{A}$.

THEOREM 1.2.6 (Carathéodory's extension theorem). *Let $\mathcal{A}$ be an algebra of subsets of a given set Ω. One can always extend every σ-finite measure defined on $\mathcal{A}$ to the σ-algebra generated by $\mathcal{A}$; moreover, the extension is unique.*

EXAMPLE 1.2.7. Let $\mathcal{A}$ be the algebra on $\mathbb{R}$, defined in Example 1.2.5. Let μ be the measure on $\mathcal{A}$, defined by setting the measure of an interval to its length. By Carathéodory's extension theorem, μ extends uniquely to the σ-algebra generated by $\mathcal{A}$. The resulting measure is called the *Borel* measure on $\mathbb{R}$. ■

Consider two measure spaces $\mathcal{M} := (\Omega, \mathcal{F}, \mu)$ and $\mathcal{N} := (\Sigma, \mathcal{G}, \nu)$. The *product measure* $\mu \times \nu$ on $\Omega \times \Sigma$ is defined in the following way: For $F \in \mathcal{F}$ and $G \in \mathcal{G}$, define $\mu \times \nu(F \times G) = \mu(F) \times \nu(G)$. So far we defined the measure $\mu \times \nu$ on $A := \{F \times G : F \in \mathcal{F}, G \in \mathcal{G}\}$. Note that A is an algebra in that $\emptyset \in A$, and A is closed under complementation and *finite* unions of its members. However, A is not necessarily a σ-algebra, as it is possible that A is not closed under countable unions of its members. Let $\mathcal{F} \times \mathcal{G}$ be the σ-algebra generated by A, i.e. it is obtained by closing A under complementation and countable unions. 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 the cartesian product of $\mathcal{F}$ and $\mathcal{G}$. Theorem 1.2.6 shows that $\mu \times \nu$ extends uniquely from A to a measure over all of $\mathcal{F} \times \mathcal{G}$. We denote the corresponding measure space by $\mathcal{M} \times \mathcal{N}$ which is called the *product measure* of $\mathcal{M}$ and $\mathcal{N}$.

Consider two measure spaces $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$ and $\mathcal{N} = (\Sigma, \mathcal{G}, \nu)$. A function $f : \Omega \to \Sigma$ is called *measurable* if the preimage of every set in $\mathcal{G}$ belongs to $\mathcal{F}$.

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

THEOREM 1.2.8 (Borel-Cantelli). *Let (E_n) be a sequence of events in some probability space. If the sum of the probabilities of the E_n is finite, then the probability that infinitely many of them occur is 0, that is,*

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

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: $d(x, z) \leq d(x, y) + d(y, z)$, 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 $\epsilon > 0$, there exist an integer N_ϵ, such that for every $m, n \geq N_\epsilon$, we have $d(x_m, x_n) \leq \epsilon$. A metric space (M, d) is called *complete* if every Cauchy sequence has a limit in M. A metric space is *compact* if and only if every sequence in the space has a convergent subsequence.

Now that we have defined the measure spaces in Section 1.2, let us state the Hoölder's and Minkowski's inequalities in a more general form.

THEOREM 1.3.1 (Hölder's inequality). *Consider a measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$, and two reals $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. IF the two measurable functions $f, g : \Omega \to \mathbb{C}$ are such that both $|f|^p$ and $|g|^q$ are integrable, then*

$$\left| \int f(x)\overline{g(x)} d\mu(x) \right| \leq \left(\int |f(x)|^p d\mu(x) \right)^{1/p} \left(\int |g(x)|^q d\mu(x) \right)^{1/q}.$$

THEOREM 1.3.2 (Minkowski's inequality). *Consider a measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$, a real $p \geq 1$, and two measurable functions $f, g : \Omega \to \mathbb{C}$ such that $|f|^p$ and $|g|^p$ are both integrable. Then*

$$\left(\int |f(x) + g(x)|^p d\mu(x) \right)^{1/p} \leq \left(\int |f(x)|^p d\mu(x) \right)^{1/p} + \left(\int |g(x)|^p d\mu(x) \right)^{1/p}.$$

Next we define concept of a normed space which is central to function analysis.

DEFINITION 1.3.3. *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 $x, y \in V$, $\|x + y\| \leq \|x\| + \|y\|$.*

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.*

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

$$\|T\| := \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X} < \infty. \tag{1}$$

The set of all bounded operators from X to Y is denoted by $B(X, Y)$. Note that the *operator norm* defined in (1) makes $B(X, Y)$ a normed space.

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

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

The set of all bounded 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 $\mathbf{B}_X := \{x : \|x\| \leq 1\}$ 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, in the sense that $\|\lambda x\| = \|x\|$ for every scalar λ with $|\lambda| = 1$. The non-degeneracy condition implies that $\mathbf{B}_X$ has non-empty interior.

Every compact symmetric convex subset of $\mathbb{R}^n$ with 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

$$K^\circ = \{x \in \mathbb{R}^n : \sum x_i y_i \leq 1, \ \forall y \in K\}. \tag{2}$$

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 functional on X, and furthermore 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 $\|\cdot\|^*$.

1.3.1. Hilbert Spaces. Consider a vector space V over $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Recall that 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 $\langle 0, 0 \rangle = 0$.

A vector space together with an inner product is called an *inner product space*.

EXAMPLE 1.3.4. 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

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

It is not difficult to verify that the above mentioned function is indeed an inner product. ■

An inner product can be used to define a norm on V. For a vector $x \in V$, define $\|x\| = \sqrt{\langle x, x \rangle}$.

LEMMA 1.3.5. *For an inner product space V, the function $\|\cdot\| : x \mapsto \sqrt{\langle x, x\rangle}$ is a norm.*

PROOF. The non-degeneracy and homogeneity conditions are trivially satisfied. It remains to verify the triangle inequality. Consider two vectors $x, y \in V$ and note that by the axioms of an inner product:

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

Now taking $\lambda := \sqrt{\frac{\langle x,x\rangle}{\langle y,y\rangle}} \times \frac{\langle x,y\rangle}{|\langle x,y\rangle|}$ will show that

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

which leads to the triangle inequality. □

A complete inner-product space is called a *Hilbert space*.

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

$$\|f\|_p := \left(\int |f(x)|^p d\mu(x)\right)^{1/p} < \infty.$$

Strictly speaking the elements of $L_p(\mathcal{M})$ are equivalent classes. Two functions f_1 and f_2 are equivalent and are considered identical, if they agree almost everywhere or equivalently $\|f_1 - f_2\|_p = 0$.

PROPOSITION 1.3.6. *For every measure space $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$, $L_p(\mathcal{M})$ 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:

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

which simplifies to the triangle inequality □

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

THEOREM 1.3.7. *Let $\mathcal{M} = (\Omega, \mathcal{F}, \mu)$ be a probability space, $1 \leq p \leq q \leq \infty$ be real numbers, and $f \in L_q(\mathcal{M})$. Then*

$$\|f\|_p \leq \|f\|_q.$$

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

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

□

Note that Theorem 1.3.7 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. We shall use the notation $\ell_p := L_p(\mathbb{N})$. In this case the ℓ_p norms are actually decreasing.

1.4. Basic Probabilistic Inequalities

Markov's inequality gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant. The application's of Markov's inequality sometimes referred to as the first moment method.

THEOREM 1.4.1 (Markov's inequality). *If X is a complex valued random variable and $a > 0$, then*

$$\Pr[|X| > a] \leq \frac{\mathbb{E}[|X|]}{a}.$$

PROOF. It is trivial. It follows from the definition of the expected value that

$$\mathbb{E}[|X|] \geq a\Pr[|X| > a].$$

□

In the second moment method, Chebyshev's inequality is applied to bound the probability that a random variable deviates far from the mean by its variance. Recall that the variance of a random variable is defined as

$$\mathrm{Var}[X] = \mathbb{E}\left[|X - \mathbb{E}[X]|^2\right] = \mathbb{E}[|X|^2] - |\mathbb{E}[X]|^2.$$

THEOREM 1.4.2 (Chebyshev's inequality). *If X is a complex valued random variable and $a > 0$, then*

$$\Pr[|X - \mathbb{E}[X]| > a] \leq \frac{\mathrm{Var}[X]}{a^2}.$$

PROOF. The theorem follows from Markov's inequality applied to the random variable $|X - \mathbb{E}[X]|^2$. □

It is possible to use Chebyshev's inequality to show that sums of independent random variables are concentrated around their expected value.

LEMMA 1.4.3. *Let $X_1, \ldots, X_n$ be independent complex valued random variables satisfying $|X_i| \leq 1$ for all $i = 1, \ldots, N$. Then*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\right| > t\right] \leq \frac{n}{t^2}.$$

PROOF. Denote $A = X_1 + \ldots + X_n$. Then by independence of X_i's we have

$$\mathrm{Var}[A] = \sum_{i,j=1}^{n} \mathbb{E}[X_i\overline{X_j}] - \mathbb{E}[X_i]\mathbb{E}[\overline{X_j}] = \sum_{i=1}^{n} \mathbb{E}[|X_i|^2] - |\mathbb{E}[X_i]|^2 = \sum_{i=1}^{n} \mathrm{Var}[X_i] \leq n.$$

Then Chebyshev's inequality implies the result. □

However in these situations, there are different inequalities that provide much stronger bounds compared to Chebyshev's inequality. We state two of them:

LEMMA 1.4.4 (Chernoff Bound). *Suppose that $X_1, \ldots, X_n$ are independent Bernoulli variables each occurring with probability p. Then for any $0 < t \leq np$,*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - np\right| > t\right] < 2e^{\frac{-t^2}{3np}}.$$

LEMMA 1.4.5 (Hoeffding's Inequality). *Suppose that $X_1, \ldots, X_n$ are independent random variables with $|X_i| \leq 1$ for each $1 \leq i \leq n$. Then for any $t > 0$,*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\ \right| > t\right] < 2e^{\frac{-t^2}{2n}}.$$

Exercises

EXERCISE 1.4.6. *Let $x = \langle x_1, \ldots, x_n\rangle$ and $y = \langle y_1, \ldots, y_n\rangle$ be complex vectors. By studying the derivative of $\langle x + ty, y\rangle$ with respect to t, prove Theorem 1.1.3.*

EXERCISE 1.4.7. *Deduce Theorem 1.1.5 from Hölder's inequality.*

EXERCISE 1.4.8. *Let $1 \leq p \leq q \leq \infty$. Show that for every $f \in \ell_p$, we have $\|f\|_q \leq \|f\|_p$.*

EXERCISE 1.4.9. *Recall that by Hölder's inequality, if $p, q \geq 1$ are conjugate exponents and $a_1, \ldots, a_n, b_1, \ldots, b_n$ are complex numbers, then*

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

Deduce from this, that if $p_1, \ldots, p_n$ are non-negative numbers with $\sum_{i=1}^{n} p_i = 1$, then

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

EXERCISE 1.4.10. *Let X be a probability space, and $p, q \geq 1$ be conjugate exponents. Show that for everyf $\in L_p(X)$, we have*

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

EXERCISE 1.4.11. *Suppose that (X, μ) is a measure space and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, for $p, q, r \geq 1$. Show that if $f \in L_p(X)$, $g \in L_q(X)$, and $h \in L_r(X)$, then*

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

EXERCISE 1.4.12. *Suppose that X is a measure space and $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$, for $p, q, r \geq 1$. Show that if $f \in L_p(X)$ and $g \in L_q(X)$, then*

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

EXERCISE 1.4.13. *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 $p, q \geq 1$, 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

In this chapter we develop 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 product is defined as $a + b := a + b \pmod N$. Finite Abelian groups can be characterized as the products of cyclic groups:

THEOREM 2.0.14. *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 be mostly interested in the group $\mathbb{Z}_2^n$ as it can be identified with the set $\{0,1\}^n$. Hence boolean functions $f : \{0,1\}^n \to \{0,1\}$ can be identified with functions $f : \mathbb{Z}_2^n \to \{0,1\}$, and this shall allow us to use the Fourier analysis of $\mathbb{Z}_2^n$ to study boolean functions.

2.1. Basic Fourier Theory

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 is a group homomorphism. That is, $\chi(a+b) = \chi(a)\chi(b)$ for all $a, b \in G$, and $\chi(0) = 1$, where 0 is the identity of G. Note that the constant function 1 is always a character and it 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, a is of some finite order n (that is $na = 0$ where na refers to adding a to itself n times). Hence $1 = \chi(0) = \chi(na) = \chi(a)^n$ which shows that $\chi(a)$ is an n-th root of unity. In particular, every character χ of G satisfies

$$\chi : G \to \mathbb{T}, \tag{3}$$

where $\mathbb{T}$ is the unit complex circle.

THEOREM 2.1.1. *If G is a finite Abelian group, then the characters of G together with the usual point-wise product of complex valued functions form a group $\widehat{G}$.*

PROOF. The principal character 1 is the identity of $\widehat{G}$. Note that if χ and ξ are characters of G, then $\chi\xi$ is also a character. Indeed $\chi(ab)\xi(ab) = \chi(a)\xi(a)\chi(b)\xi(b)$, 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. □

The group $\widehat{G}$ is called the *Pontryagin dual* of G. Fourier analysis is based on expressing functions $f : G \to \mathbb{C}$ as linear combinations of characters. It will be convenient to treat the set of these functions as a Hilbert space: Let $L_2(G)$ denote the set of functions $f : G \to \mathbb{C}$, where here G is endowed with the uniform probability measure. Recall (see Section 1.3.1 and 1.3.2) that $L_2(G)$ is a Hilbert space with the inner product

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

In the sequel, we will often consider G as a probability space, and $\mathbb{E}_{x\in G}$ shall always mean that x is a random variable that takes values in G uniformly at random. To simplify the notation we usually abbreviate $\mathbb{E}_{x\in G}$ to simply $\mathbb{E}$. For a function $f : G \to \mathbb{C}$, the notation $\mathbb{E}[f]$ means $\mathbb{E}_{x\in G}[f(x)]$ (which is equal to $\frac{1}{|G|}\sum_{x\in G} f(x)$).

EXAMPLE 2.1.2. Consider the group $\mathbb{Z}_2^n$. For every $a = (a_1, \ldots, a_n) \in \mathbb{Z}_2^n$, one can construct a corresponding character χ_a that maps x to $\prod_{i:a_i=1}(-1)^{x_i} = (-1)^{\sum_{i:a_i=1} x_i}$. 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 indeed all the characters of $\mathbb{Z}_2^n$. Note that in this case the characters are actually real-valued (they only take values 1 and -1), but as we shall see below for all other Abelian groups there are characters that 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

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

and as later when we take a probabilistic approach to decomposing functions, this notation extends to general product spaces (where there is no group structure). ■

Our next goal will be to prove that the characters form an orthonormal basis for $L_2(G)$. First let us prove a simple lemma.

LEMMA 2.1.3. *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 y was arbitrary, we conclude that χ must be the principal character which is a contradiction. □

Now we can prove the orthogonality of the characters.

LEMMA 2.1.4. *The characters of a finite Abelian group G are orthonormal functions in $L_2(G)$.*

PROOF. It follows from (3) that every $\chi \in \widehat{G}$ satisfies

$$\|\chi\|_2^2 = \mathbb{E}\left[|\chi(x)|^2\right] = \mathbb{E}[1] = 1.$$

So characters are unit vectors in $L_2(G)$. It remains to verify the orthogonality. Let $\chi \neq \xi$ be two different characters. Then $\chi\overline{\xi} = \chi\xi^{-1}$ is a non-principal character of G (why?). Hence by Lemma 2.1.3, we have

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

□

So far we have discussed the Pontryagin dual of G in an abstract manner. Since finite Abelian groups have simple structures (Theorem 2.0.14), it is quite easy to describe the characters of G. We start with the basic case of $G = \mathbb{Z}_N$. For every $a \in \mathbb{Z}_N$, define $\chi_a \in L_2(G)$ as

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

Let us verify that χ_a is actually a character. Indeed $\chi_a(0) = e^{\frac{2\pi i}{N}0} = e^0 = 1$, and since $e^{2\pi i} = 1$, we have

$$\chi_a(x)\chi_a(y) = e^{\frac{2\pi i}{N}ax} e^{\frac{2\pi i}{N}ay} = e^{\frac{2\pi i}{N}a(x+y \pmod N)} = \chi_a(x+y).$$

Note that $L_2(G)$ is $|G|$-dimensional, and hence by Lemma 2.1.4, G has at most $|G|$ characters. It follows that $\{\chi_a : a \in G\}$ are all the characters of G. The principal character is $\chi_0 \equiv 1$. Also $\chi_a\chi_b = \chi_{a+b}$ which shows that the dual group $\widehat{G}$ is isomorphic to G. As we shall see below this is in general true for all finite Abelian groups.

Now let us consider the general case of $G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$ for some positive integers $N_1, \ldots, N_k$. For every $a = (a_1, \ldots, a_k) \in G$, define $\chi_a \in L_2(G)$ as the product of the characters $\chi_{a_1}, \ldots, \chi_{a_k}$ of the groups $\mathbb{Z}_{N_1}, \ldots, \mathbb{Z}_{N_k}$ applied to the coordinates of $x \in G$ respectively. More precisely

$$\chi_a : x \mapsto \prod_{j=1}^{k} e^{\frac{2\pi i}{N_j}a_j x_j}.$$

As in the case of $\mathbb{Z}_N$, it is straightforward to verify that χ_a is a character by showing that $\chi_a(0) = 1$, and $\chi_a(x+y) = \chi_a(x)\chi_a(y)$. Again Lemma 2.1.4 shows that $\{\chi_a : a \in G\}$ are all the characters of G. We also have the identify $\chi_a\chi_b = \chi_{a+b}$ which implies the following theorem.

THEOREM 2.1.5. *If G is a finite Abelian group, then the characters of G form an orthonormal basis for $L_2(G)$. Furthermore we have $\widehat{G} \cong G$.*

Theorem 2.1.5 shows that G is isomorphic to its dual $\widehat{G}$, and so it shall be convenient to identify the two groups in the sequel, and denote the characters by χ_a where $a \in G$. Since the characters form an orthonormal basis for $L_2(G)$, every function $f : G \to \mathbb{C}$ can be expressed in a unique way as a linear combination of the characters $f = \sum_{a \in G} \widehat{f}(a)\chi_a$. The corresponding coefficients $\widehat{f}(a) \in G$ are referred to as the Fourier coefficients. This leads to the definition of the Fourier transform.

DEFINITION 2.1.6. *The Fourier transform of a function $f : G \to \mathbb{C}$ is the unique function $\widehat{f} : \widehat{G} \to \mathbb{C}$ defined as*

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

We will often use the notation $\widehat{f}(a)$ to denote $\widehat{f}(\chi_a)$.

Let us state a simple example of the Fourier transform of a function on $\mathbb{Z}_2^n$.

EXAMPLE 2.1.7. Let $f : \mathbb{Z}_2^n \to \mathbb{C}$ be the parity function $f : x \mapsto \sum_{i=1}^n x_i (\text{mod } 2)$. Then

$$\widehat{f}(0) = \mathbb{E} f(x)\chi_0 = \mathbb{E} f(x) = \frac{1}{2}.$$

We also have

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

as $f(x) = 1$ if and only if $\sum_{j=1}^n x_j = 1 (\text{mod } 2)$. Next consider $a \in \mathbb{Z}_2^n$ with $a \neq (1, \ldots, 1)$ and $a \neq 0$. Let j_0, j_1 be such that $a_{j_0} = 0$ and $a_{j_1} = 1$. We have (why?)

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

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 $\chi_a(x) = -\chi_a(x + e_{j_0} + e_{j_1})$. We conclude that $\widehat{f}(a) = 0$ for every $a \in \mathbb{Z}_2^n$ satisfying $a \neq (1, \ldots, 1)$ and $a \neq 0$. ∎

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

LEMMA 2.1.8. *The Fourier transform considered as an operator from $L_1(G)$ to $L_\infty(\widehat{G})$ is norm decreasing:*

$$\|\widehat{f}\|_\infty \leq \|f\|_1.$$

PROOF. By (3) for every $a \in G$, we have

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

□

The 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 $\widehat{\mathbf{1}_A}(0) = \frac{|A|}{|G|}$ corresponds to the density of A.

It follows from the fact that the characters from an orthonormal basis for $L_2(G)$ that

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

and that this expansion of f as a linear combination of characters is unique. This formula is called the *Fourier inversion formula* as it shows how the functions f can be reconstructed from its Fourier transform.

If $A \subseteq G$, then the orthogonal complement of A is defined as

$$A^\perp = \{a \in G : \chi_a(x) = 1 \ \forall x \in A\}.$$

It follows from the identities $\chi_0 = 1$ and $\chi_a \chi_b = \chi_{a+b}$ that $S^\perp$ is a subgroup of G. The Fourier transform of the indicator function of a subgroup of G has a simple form:

LEMMA 2.1.9. *If H is a subgroup of G, then for every $a \in G$, we have*

$$\widehat{\mathbf{1}_H}(a) = \begin{cases} |H|/|G| & a \in H^\perp \\ 0 & a \notin H^\perp \end{cases}$$

PROOF. If $a \in H^\perp$, then

$$\widehat{\mathbf{1}_H}(a) = \langle \mathbf{1}_H, \chi_a \rangle = \mathbb{E}\mathbf{1}_H(x)\overline{\chi_a(x)} = \mathbb{E}\mathbf{1}_H(x) = |H|/|G|.$$

On the other hand if $a \notin H^\perp$, then there exists $y \in H$ such that $\chi_a(y) \neq 1$. Then

$$\sum_{z \in H} \overline{\chi_a(z)} = \chi_a(y) \sum_{z \in H} \overline{\chi_a(z-y)} = \chi_a(y) \sum_{z \in H} \overline{\chi_a(z)},$$

which shows that $\sum_{z \in H} \overline{\chi_a(z)} = 0$. Hence

$$\widehat{\mathbf{1}_H}(a) = \mathbb{E}\mathbf{1}_H(x)\overline{\chi_a(x)} = \frac{1}{|G|} \sum_{z \in H} \mathbb{E}\mathbf{1}_H(z) = 0.$$

□

REMARK 2.1.10. It follows from Lemma 2.1.9 that if $A = y + H$ is a coset of H in G (i.e. H is a subgroup of G and $y \in G$), then for every $a \in G$,

$$\begin{aligned}\widehat{\mathbf{1}_A}(a) &= \mathbb{E}\mathbf{1}_A(x)\overline{\chi_a(x)} = \mathbb{E}\mathbf{1}_H(x-y)\overline{\chi_a(x)} = \mathbb{E}\mathbf{1}_H(x)\overline{\chi_a(x+y)} = \overline{\chi(y)}\widehat{\mathbf{1}_H}(a) \\ &= \begin{cases} \overline{\chi(y)}|H|/|G| & a \in H^\perp \\ 0 & a \notin H^\perp \end{cases}\end{aligned}$$

■

EXAMPLE 2.1.11. Let us revisit Example 2.1.7 in light of Remark 2.1.10. Note that $H = \{x \in \mathbb{Z}_2^n : \sum_{i=1}^n x_i = 0 \pmod 2\}$ is a subgroup of $\mathbb{Z}_2^n$. Now the function f defined in Example 2.1.7 is the indicator function of $A = e_1 + H$. Note that

$$H^\perp = \{a : (-1)^{\sum_{i=1}^n x_i a_i} = 1 \ \forall x \in H\} = \{(0,\ldots,0),(1,\ldots,1)\}.$$

Hence

$$\widehat{f}(a) = \widehat{\mathbf{1}_A}(a) = \begin{cases} \overline{\chi_a(e_1)}|H|/|G| & a \in H^\perp \\ 0 & a \notin H^\perp \end{cases}$$

We conclude that $\widehat{f}(0) = 1/2$ and $\widehat{f}(1,\ldots,1) = -1/2$, and $\widehat{f}(a) = 0$ for every $a \in \mathbb{Z}_2^n$ satisfying $a \neq (1,\ldots,1)$ and $a \neq 0$. ■

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

THEOREM 2.1.12 (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 identify now follows from orthonormality of characters:

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

□

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

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

As the first example of an application of the Parseval identity, let us show that for every subgroup H of G, we have

$$|H||H^\perp| = |G|. \tag{4}$$

Indeed by Lemma 2.1.9, we have

$$\frac{|H|}{|G|} = \mathbb{E}\mathbf{1}_H = \mathbb{E}\mathbf{1}_H^2 = \langle \mathbf{1}_H, \mathbf{1}_H\rangle = \|\mathbf{1}_H\|_2^2 = \sum_{a\in G} |\widehat{\mathbf{1}_H}(a)|^2 = \sum_{a\in H^\perp} (|H|/|G|)^2 = \frac{|H|^2|H^\perp|}{|G|^2}$$

which simplifies to (4).

Next we introduce the important notion of *convolution*.

DEFINITION 2.1.13. *Let G be a finite Abelian group. For two functions $f, g : G \to \mathbb{C}$, we define their convolution $f * g : G \to \mathbb{C}$ as*

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

Note that $f * g(x)$ is the average of $f(a)f(b)$ over all pairs a, b with $a + b = x$. This gives a combinatorial nature to convolution which makes it very useful in dealing with certain discrete problems. Consider a set $A \subseteq G$. Then $f * \mathbf{1}_A(x)$ is the average of f over the set $x - A := \{x - y : y \in A\}$. 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. These types of averaging operators usually "smooth" f, and makes it more similar to a constant functions. This smoothing property of the convolution is one of the main tools in harmonic analysis and this course.

Next let us list some basic facts about the convolution. We define the *support* of $f : G \to \mathbb{C}$, denoted by $\mathrm{Supp}(f)$, to be the set of the points $x \in G$ with $f(x) \neq 0$.

LEMMA 2.1.14. *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*

$$Supp(f * g) \subseteq Supp(f) + Supp(g).$$

(e) *We have*

$$\|f * g\|_\infty \leq \|f\|_1 \|g\|_\infty.$$

(f) *More generally, if p and q are conjugate exponents, then*

$$\|f * g\|_\infty \leq \|f\|_p \|g\|_q.$$

(g) *We have*

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1.$$

PROOF. **(a)** For every $x \in G$, we have

$$f * g(x) = \mathbb{E}_y[f(x-y)g(y)] = \mathbb{E}_y[f(x-y)g(x-(x-y))] = \mathbb{E}_z[f(z)g(x-z)] = g * f(x).$$

(b) By Part **(a)**,

$$\begin{aligned}(f * g) * h(x) &= (g * f) * h(x) = \mathbb{E}_z \mathbb{E}_y[g(x-z-y)f(y)]h(z) = \\ &= \mathbb{E}_{y,z} g(x-z-y)f(y)h(z) = (h * g) * f(x) = f * (g * h)(x).\end{aligned}$$

(c) is trivial.

(d) follows from the fact that $f(x)$ is the average of $f(a)g(b)$ over all pairs of points $a, b \in G$ with $a + b = x$.

(e) is a special case of **(f)**.

(f) Note that for every $x \in G$, by Hölder's inequality we have

$$|f * g(x)| \leq \mathbb{E}_{y \in G}|f(x-y)||g(y)| \leq (\mathbb{E}|f(x-y)|^p)^{1/p} (\mathbb{E}|g(y)|^q)^{1/q} = (\mathbb{E}|f(y)|^p)^{1/p} \|g\|_q = \|f\|_p \|g\|_q.$$

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

(g) We have

$$\|f*g\|_1 = \mathbb{E}_x\,|f*g(x)| \le \mathbb{E}_{x,y}|f(x-y)||g(y)| = \mathbb{E}_{z,y}|f(z)||g(y)| = \mathbb{E}_z|f(z)|\mathbb{E}_y|g(y)| = \|f\|_1\|g\|_1.$$

□

The relevance of the Fourier transform to convolution lies in the following lemma.

LEMMA 2.1.15. *If $f, g : G \to \mathbb{C}$, then*

$$\widehat{f*g} = \widehat{f}\cdot\widehat{g}.$$

PROOF. We have

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

□

Note that Lemma 2.1.15 in particular shows that

$$\mathbb{E}f(x)\mathbb{E}g(x) = \widehat{f}(0)\widehat{g}(0) = \widehat{f*g}(0) = \mathbb{E}f*g(x).$$

We also have the dual version of Lemma 2.1.15,

$$\widehat{f\cdot g}(x) = \sum_{y\in G}\widehat{f}(x-y)\widehat{g}(y), \tag{5}$$

which converts point-wise product back to convolution.

For a function $h : G \to \mathbb{C}$, define $\tilde{h} : G \to \mathbb{C}$ as $h : x \mapsto \overline{h(-x)}$. Note that $\tilde{h} = \sum_{a\in G}\overline{\widehat{h}(a)}\chi_a$. Hence it follows from the Parseval identity and Lemma 2.1.15 that for $f, g, h : G \to \mathbb{C}$, we have

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

2.2. Fourier analysis of $\mathbb{Z}_2^n$ and polynomials

As we saw above the characters of the group $\mathbb{Z}_2^n$ are of the form $\chi_S(x) = (-1)^{\sum_{i\in S}x_i}$ for all $S \subseteq [n]$. As a result sometimes it is more intuitive to work with the domain $\{-1,1\}^n$ rather than $\mathbb{Z}_2^n \equiv \{0,1\}^n$ by changing the role of 0 to -1. Note that this change of the domain will convert the character χ_S with the function $w_S(x) := \prod_{i\in S}x_i$. Hence every function $f : \{-1,1\}^n \to \mathbb{C}$ can be expressed as

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

Note that the functions $w_S(x)$ are monomials of degree $|S|$ in which every variable appears with degree at most 1. Hence this representation is basically a representation of the function f as a polynomial of degree at most n. Note that the *Fourier degree* of f, often denoted by $\deg_{\mathcal{F}}(f)$, is the largest $|S|$ such that $\widehat{f}(S) \neq 0$.

Exercises

EXERCISE 2.2.1. *Prove the Identity (5).*

EXERCISE 2.2.2. *Suppose that for $f : \mathbb{Z}_2^n \to \{0,1\}$ we have $\widehat{f}(S) = 0$ for all $|S| \geq 2$ (that is $\deg_{\mathcal{F}}(f) \leq 1$). 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 2.2.3. *Let G be a finite Abelian group, and H be a subgroup of G. Prove that*

$$\left(H^{\perp}\right)^{\perp} = H.$$

EXERCISE 2.2.4. *Given two subsets $\mathcal{A}, \mathcal{B} \subseteq 2^{[n]}$. For $S \subseteq [n]$, let*

$$\text{PARITY}_S(\mathcal{A}) := \{A \in \mathcal{A} \mid |A \cap S| \equiv 0 \bmod 2\},$$

and

$$\text{PARITY}_S(\mathcal{B}) := \{B \in \mathcal{B} \mid |B \cap S| \equiv 0 \bmod 2\}.$$

- *Prove that if $\text{PARITY}_S(\mathcal{A}) = \text{PARITY}_S(\mathcal{A})$ for every S, then $\mathcal{A} = \mathcal{B}$.*
- *Suppose $|\text{PARITY}_S(\mathcal{A}) - \text{PARITY}_S(\mathcal{A})| \leq \delta$ for every S. Bound $|\,(\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A})\,|$ in terms of δ.*

EXERCISE 2.2.5. *Let G be a finite Abelian group, and H be a subgroup of G. Prove that for every $f : G \to \mathbb{C}$, we have*

$$\mathbb{E}_{x \in H} f(x) = \sum_{a \in H^{\perp}} \widehat{f}(\chi_a).$$

EXERCISE 2.2.6. *Let G be a finite Abelian group and $f, g : G \to \mathbb{C}$. Show that for every positive integer m,*

$$\|f * g\|_m \leq \|f\|_1 \|g\|_m.$$

EXERCISE 2.2.7. *Consider a function $f : \mathbb{Z}_2^n \to \{0,1\}$ and it's Fourier expansion $f = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S$. Define the discrete derivative of f in direction $i \in \{1, \ldots, n\}$ as $\Delta_i f : x \mapsto f(x + e_i) - f(x)$. Write the Fourier expansion of $\Delta_i f$ in terms of the Fourier coefficients of f.*

EXERCISE 2.2.8. *Suppose that for $f : \mathbb{Z}_2^n \to \{0,1\}$ we have $\widehat{f}(S) = 0$ for all $|S| > k$ (that is $\deg_{\mathcal{F}}(f) \leq k$). Show that all the Fourier coefficients of f are of the form $\frac{r}{2^k}$ where $r \in \{0, \pm 1, \ldots, \pm 2^k\}$. Conclude that every such function depends on at most $k2^{2k}$ coordinates.*

CHAPTER 3

Applications to Computer Science: Property 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, and accordingly probabilistically distinguish between the case that f is a linear function and the case that f has to be modified on at least $\epsilon > 0$ fraction of points to become a linear function. Inspired by this observation, Rubinfeld and Sudan [**RS93**] defined the concept of property testing which is now a major 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 there has been a significant amount of research on classifying the testable properties in combinatorial and algebraic settings.

3.1. Linearity test

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

DEFINITION 3.1.1. *A function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ is called* linear *if $f(x+y) = f(x) + f(y)$ for all $x, y \in \mathbb{Z}_2^n$.*

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

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 few points, and with some significant probability distinguish correctly between the following two cases

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

$$\Pr[f(x) \neq \ell(x)] \geq \epsilon.$$

More precisely for every $\epsilon > 0$, 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

(1) always accept f if it is linear.
(2) reject f with probability at least $\delta > 0$ if it is ϵ-far from every linear function.

This is a *one-sided-error* test as it always accepts f if it satisfies the property. Also note that one can easily boost the probability of the success of the test by applying the test several times. More precisely, one can run the test N times, and accept f if all the N executions accept f, and reject it otherwise. 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

$N = 1000\delta^{-1}$. However, note that in this case we are making $3N$ queries to f. Now let us finally state the 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 bulk of the analysis lies in proving that if f is ϵ-far from every linear function, then f is rejected with probability at least $\delta > 0$.

In order to apply Fourier analysis, we need the range of f to be $\mathbb{C}$ rather than $\mathbb{Z}_2$. The most trivial way to achieve this is to identify $\mathbb{Z}_2$ with $\{0, 1\}$ and consider f as a function from $\mathbb{Z}_2^n$ to $\mathbb{Z}_2$. However for this problem, it is more natural to achieve this by replacing f with $(-1)^f : \mathbb{Z}_2^n \to \{-1, 1\}$. Note that f is linear if and only if $(-1)^f$ is multiplicative (i.e. $(-1)^{f(x+y)} = (-1)^{f(x)}(-1)^{f(y)}$) which is equivalent to being a character of $\mathbb{Z}_2^n$. So the linearity test is can be reformulated as a "character test":

Blum, Luby, and Rubinfeld's [**BLR90**] *linearity test:*
- Given a function $f : \mathbb{Z}_2^n \to \{-1, 1\}$.
- Pick two random points $x, y \in \mathbb{Z}_2^n$.
- If $f(x)f(y) \neq f(x + y)$, then Reject, otherwise Accept.

And our goal is to show that for every $\epsilon > 0$, there exists a $\delta > 0$ such that the BLR test

(1) accept f if if is a character of $\mathbb{Z}_2^n$.
(2) reject f with probability at least $\delta > 0$ if it is ϵ-far from every character.

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

Consider a character χ_a, and note that

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

So if f is ϵ-far from every character, then

$$\epsilon \leq \frac{1}{2} - \frac{1}{2}\max_a \widehat{f}(a),$$

or equivalently

$$\max_a \widehat{f}(a) \leq 1 - 2\epsilon. \tag{6}$$

Now let us analyze the probability that f is not rejected by the BLR algorithm. Note that

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

Replacing f with its Fourier expansion, we get

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

Note that $a+c=0$ if and only if $a=c$, and thus (see Lemma 2.1.3),

$$\mathbb{E}[\chi_{a+c}(x)] = \begin{cases} 1 & a=c \\ 0 & a\neq c \end{cases}$$

Similarly

$$\mathbb{E}[\chi_{b+c}(y)] = \begin{cases} 1 & b=c \\ 0 & a\neq c \end{cases}$$

Hence

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

which shows that

$$\Pr_{x,y}[f(x)f(y)=f(x+y)] = \frac{1}{2}+\frac{1}{2}\sum_a \widehat{f}(a)^3 \leq \frac{1}{2}+\frac{1}{2}\left(\max_a \widehat{f}(a)\right)\sum_a \widehat{f}(a)^2. \tag{7}$$

By the Parseval identity

$$\sum_{a\in G}\widehat{f}(a)^2 = \|f\|_2^2 = 1.$$

So

$$\Pr_{x,y}[f(x)f(y)=f(x+y)] \leq \frac{1}{2}+\frac{1}{2}\max_a \widehat{f}(a). \tag{8}$$

Now to finish the proof note that by (6) and (8) if f is ϵ-far from every character, then

$$\Pr_{x,y}[f(x)f(y)=f(x+y)] \leq 1-\epsilon,$$

and the probability of rejection is at least $\delta := \epsilon > 0$.

3.1.2. Testing in general. In general testability with one-sided error of a property of functions is defined as the following:

DEFINITION 3.1.2 (Testability with one-sided error). *A property $\mathcal{P}$ is said to be* testable with one-sided error *if there are functions $q:(0,1)\to\mathbb{Z}_{>0}$, $\delta:(0,1)\to(0,1)$, and an algorithm T that, given as input a parameter $\epsilon>0$ and oracle access to a function f, makes at most $q(\epsilon)$ queries to the oracle for f,*

- *always accepts if $f\in\mathcal{P}$.*
- *rejects with probability at least $\delta(\epsilon)$ if f is ϵ-far from $\mathcal{P}$.*

If, furthermore, q is a constant function, then $\mathcal{P}$ is said to be proximity-obliviously testable (PO testable).

The BLR test of Section 3.1 shows that the property of being linear for functions $f:\mathbb{Z}_2^n\to\mathbb{Z}_2$ is proximity-obliviously testable with one-sided error using only 3 queries.

The properties of functions $f:\mathbb{Z}_2^n\to\mathbb{Z}_2$ that are proximity-obliviously testable with one-sided error are characterized in [**BFH**$^+$**13**] by Bhattacharyya, Fischer, H. Hatami, P. Hatami, and Lovett. Note that if we do not impose any conditions on the property, then the algebraic structure of $\mathbb{Z}_2^n$

is ignored, and we are treating $\mathbb{Z}_2^n$ as a generic set of size 2^n. Hence in order to take the algebraic structure of $\mathbb{Z}_2^n$ into account, one assumes that $\mathcal{P}$ to be affine-invariant:

DEFINITION 3.1.3 (Affine invariance). *A property $\mathcal{P}$ of functions $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ is* affine-invariant *if the following holds. For every positive integer n, if $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfies the property, then for every positive integer m, and every affine transformation $A : \mathbb{Z}_2^m \to \mathbb{Z}_2^n$, the function $f \circ A : \mathbb{Z}_2^m \to \{0,1\}$ also satisfies the property. (An affine transformation A is of the form $L + c$ where L is a linear transformation and c is a constant).*

It is not difficult to see that if an affine-invariant property is proximity-obliviously testable with one-sided error, then it must be locally characterized in the following sense.

DEFINITION 3.1.4. *For a positive integer k, a property $\mathcal{P}$ of functions $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ is called* k-locally characterized *if there the following holds: f satisfies $\mathcal{P}$ if and only if $f \circ A : \mathbb{Z}_2^k \to \mathbb{Z}_2$ satisfies $\mathcal{P}$ for all affine transformations $A : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$. We say $\mathcal{P}$ is* locally characterized *if it is k-locally characterized for some constant k.*

On the other hand, if a property $\mathcal{P}$ is k-locally characterized, then there is a natural candidate for a test that makes 2^k queries to the function:

- Given a function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$.
- Pick a random affine transformation $A : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$.
- If $f \circ A \in \mathcal{P}$, then Accept, otherwise Reject.

However proving that this test indeed rejects every f that is ϵ-far from $\mathcal{P}$ with probability $\delta(\epsilon, \mathcal{P}) > 0$ is not straightforward, and [**BFH+13**] uses several tools from an area of mathematics called higher-order Fourier analysis to establish this fact:

THEOREM 3.1.5 ([**BFH+13**]). *An affine-invariant property $\mathcal{P}$ is proximity-obliviously testable with one-sided error if and only if it is locally characterized.*

An important example of a k-locally characterized property is the property of having degree at most $k-1$ as a polynomial from $\mathbb{Z}_2^n$ to $\mathbb{Z}_2$. Note that this notion of degree is different from the Fourier degree discussed in Section 2.2. For example the function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ defined as $f : x \mapsto x_1 + \ldots + x_n$ is of of degree 1 in this sense, while by looking at its Fourier transform $f = \frac{1}{2} - \frac{1}{2}\chi_{\vec{1}}(x)$ we see that it has Fourier degree n.

In [**BFH+13**] it is shown that a wide class of properties, referred to as degree-structural properties are locally characterized. We formally defined them below in Definition 3.1.6, but first let us list some examples of degree-structural properties. Let d be a fixed positive integer. Each of the following properties defines a degree-structural property. Consider a prime number p.

- **Degree $\leq d$:** The degree of the function $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ as a polynomial is at most d;
- **Splitting:** A function $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ *splits* if it can be written as a product of at most d linear functions;
- **Factorization:** A function $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ *factors* if $f = gh$ for polynomials $g, h : \mathbb{Z}_p^n \to \mathbb{Z}_p$ such that $\deg(g) \leq d-1$ and $\deg(h) \leq d-1$;
- **Sum of two products:** A function $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ is a *sum of two products* if there are polynomials $g_1, g_2, g_3, g_4 : \mathbb{Z}_p^n \to \mathbb{Z}_p$ such that $f = g_1 g_2 + g_3 g_4$ and $\deg(g_i) \leq d-1$ for $i \in \{1,2,3,4\}$;
- **Having a square root:** A function $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ *has a square root* if $f = g^2$ for a polynomial $g : \mathbb{Z}_p^n \to \mathbb{Z}_p$ with $\deg(g) \leq d/2$;

In fact, roughly speaking, any property that can be described as the property of decomposing into a known structure of low-degree polynomials is degree-structural.

DEFINITION 3.1.6 (Degree-structural property). *Given an integer $c > 0$, a vector of non-negative integers $\mathbf{d} = (d_1, \ldots, d_c) \in \mathbb{Z}^c_{\geq 0}$, and a function $\Gamma : \mathbb{Z}_p^c \to \mathbb{Z}_p$, define the* $(c, \mathbf{d}, \Gamma)$-structured property *to be the collection of functions $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ for which there exist polynomials $P_1, \ldots, P_c : \mathbb{Z}_p^n \to \mathbb{Z}_p$ satisfying $f(x) = \Gamma(P_1(x), \ldots, P_c(x))$ for all $x \in \mathbb{Z}_p^n$ and $\deg(P_i) \leq d_i$ for all $i \in [c]$.*

We say a property $\mathcal{P}$ is degree-structural *if there exist integers $\sigma, \Delta > 0$ and a set of tuples $S \subset \{(c, \mathbf{d}, \Gamma) \mid c \in [\sigma], \mathbf{d} \in [0, \Delta]^c, \Gamma : \mathbb{Z}_p^c \to \mathbb{Z}_p\}$, such that a function $f : \mathbb{Z}_p^n \to \mathbb{Z}_p$ satisfies $\mathcal{P}$ if and only if f is $(c, \mathbf{d}, \Gamma)$-structured for some $(c, \mathbf{d}, \Gamma) \in S$. We call σ the* scope *and Δ the* max-degree *of the degree-structural property $\mathcal{P}$.*

THEOREM 3.1.7 ([**BFH+13**]). *Every degree-structural property is k-locally characterized for some k.*

The proof presented in [**BFH+13**] uses higher-order Fourier analysis and does not provide any explicit bounds on k. It seems that there should be a different proof for this theorem which would provide reasonable bounds on k.

PROBLEM 3.1.8 (Open Problem). *Find a reasonable upper-bound on k in terms of the max-degree and the size of the scope of degree-structural properties such that every degree-structural property with those parameters is k-locally characterized.*

This problem is interesting even for the simple and explicit properties that we listed above.

PROBLEM 3.1.9 (Open Problem). *For each one of the properties listed above (i.e. Factorization, Sum of two products, Having a square root) find the smallest k such that the property is k-locally characterized.*

Exercises

EXERCISE 3.1.10. *Prove that the property of having polynomial degree at most d for functions $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ is $(d+1)$-locally characterized.*

EXERCISE 3.1.11. *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 having degree at most 1. Prove that the test works correctly.*

EXERCISE 3.1.12. *By looking at $f * \ldots * f$, the k-fold convolution of f by itself, construct a linearity test with one-sided error that makes $2k$ queries to f. W Probability is the probability that this test makes an error and accepts a function that is ϵ-far from being linear? Compare this to applying the BLR test N times.*

CHAPTER 4

Applications to Computer Science: Bounded Depth Circuits

In 1949 Shannon proposed the size of Boolean circuits as a measure of computation difficulty of a function. Circuits are closely related in computational power to Turing machines, and thus they provide a nice framework for understanding the time complexity. On the other hand their especially simple definition makes them amenable to various combinatorial, algebraic, and analytic methods.

A burst of activity in circuit complexity exploded about 30 years ago with first exponential lower bounds for some circuit models, like bounded depth circuits, monotone circuits, restricted branching programs, etc. There has been quick progress made for about two decades, but soon various barriers are discovered.

A *Boolean circuit* is a directed acyclic graph. The vertices of indegree 0 are called *inputs*, and are labeled with a variable x_i or with a constant 0 or 1. The vertices of indegree $k > 0$ are called *gates* and are labeled with a Boolean function on k inputs. The indegree of a vertex is called its *fanin* and its outdegree is called its *fanout*. The most standard circuits are restricted to have gates $\wedge$, $\vee$, $\neg$. One of the nodes is designated the *output* node, and then the circuit represents a Boolean function in a natural way. The size of a circuit is its number of gates.

A simple counting argument establishes the following strong lower-bound. Roughly speaking, there are too many Boolean functions $f : \{0,1\}^n \to \{0,1\}$ (there are 2^{2^n} of those functions) compared to the number of small circuits.

THEOREM 4.0.13 (Muller 1956). *Almost every Boolean function $f : \{0,1\}^n \to \{0,1\}$ requires fanin 2 circuits of size $\Omega(2^n/n)$. On the other hand every function $f : \{0,1\}^n \to \{0,1\}$ can be computed by a fanin 2 circuit of size $O(2^n/n)$*

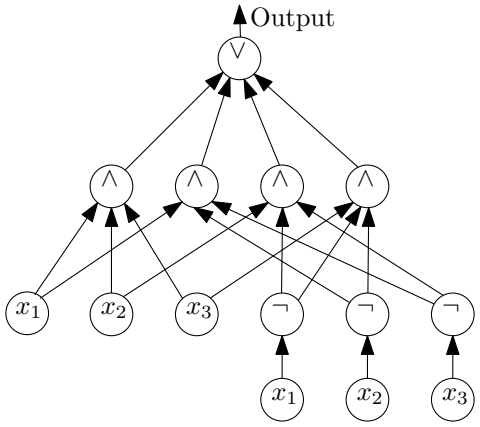

FIGURE 1. A circuit that computes the function $f(x_1, x_2, x_3) = x_1 + x_2 + x_3 \bmod 2$.

Theorem 4.0.13 has a major shortcoming. It does not provide any *explicit* example of a function which requires a large circuit. Also unfortunately it does not provide any example of a function in NP that requires circuits of superpolynomial size. Despite the importance of lower bounds on the circuit complexity, the best explicit known construction due to Blum 1984 provides a function which requires finin 2 circuits of size $3n - o(n)$.

4.1. Bounded depth circuits

Considering our inability in proving lower bounds on the circuit complexity of explicit Boolean functions, we need to impose strong restrictions on the circuits in order to be able to prove meaningful lower bounds. We will restrict to bounded depth circuits. 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 [**Has86**] further strengthened and simplified this argument, and obtained near optimal bounds.

Let us start by defining our constant depth circuits. As we mentioned earlier we are 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 fanin so that the circuit may 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. Let AC[d] denote the set of all alternating circuits of depth at most d.

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.

Definition 4.1.1 (Conjunctive Normal Form, $\wedge$ of $\vee$). *A formula is in conjunctive normal form, abbreviated to CNF, if it is a conjunction (i.e. $\vee$) of clauses, where a clause is a disjunction (i.e $\vee$) of literals (i.e. x_i or $\neg x_i$), where a literal and its negation cannot appear in the same clause*

For example $(x_1 \vee x_2) \wedge (\neg x_1 \vee x_2 \vee x_3)$ is a formula in conjunctive normal form. It corresponds to an alternating circuit of depth 2 with 3 gates.

Definition 4.1.2 (Disjunctive Normal Form, $\vee$ of $\wedge$). *A formula is in disjunctive normal form, abbreviated to DNF, if it is a disjunction (i.e. $\vee$) of conjunctive clauses (i.e $\vee$ of literals).*

Consider a fixed point $y = (y_1, \ldots, y_n) \in \{0,1\}^n$, and $T = \{i : y_i = 1\}$. Note that the only assignment that satisfies the clause

$$\left(\bigwedge_{i \in T} x_i\right) \wedge \left(\bigwedge_{i \notin T} \neg x_i\right)$$

is the assignment $x := y$. Hence given a Boolean function $f : \{0,1\}^n \to \{0,1\}$, for every point y with $f(y) = 1$ we can create a clause which is satisfied only if $x = y$. By taking the $\vee$ of these clauses we create a DNF formula that represents the function f.

EXAMPLE 4.1.3. Consider the function $f : \{0,1\}^2 \to \{0,1\}$ such that $f(0,0) = f(0,1) = f(1,1) = 1$ and $f(1,0) = 0$. Then the DNF

$$(\neg x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2) \vee (x_1 \wedge x_2)$$

represents f. ■

By changing the role of 0's and 1's and $\wedge$ and $\vee$, we can represent f in CNF. We conclude the following observation which says that the depth 2 alternating circuits are powerful enough to compute any Boolean function.

OBSERVATION 4.1.4. *Every function $f : \{0,1\}^n \to \{0,1\}$ can be represented in both DNF and CNF formulas using at most 2^n clauses.*

4.2. Håstad's switching lemma

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 fanin, then 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.

As we mentioned earlier, Håstad further explored these ideas. The core of his proof is an important lemma known as switching lemma. It is a key tool for proving lower bounds on the size of the constant-depth Boolean circuits.

DEFINITION 4.2.1. *Let $X = \{x_1, \ldots, x_n\}$ be the input variables to a circuit C computing a function f. A* restriction *ρ is an element in $\{0,1,*\}^X$.*

A restriction ρ is interpreted as setting the variables assigned 0's and 1's and leaving the variables that are assigned $*$'s. 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.

For a Boolean function $f : \{0,1\}^n \to \{0,1\}$, let $\mathcal{D}(f)$ denote the smallest $s \geq 0$ such that f can be expressed as a DNF formula that satisfies the following two properties:

- Each clause has size at most s;
- The clauses all accept disjoint sets of points. *I.e.* there is no $x \in \{0,1\}^n$ that satisfies more than one clause.

Note that the construction following Definition 4.1.2 shows that always $\mathcal{D}(f) \leq n$.

LEMMA 4.2.2 (Håstad's switching lemma). *Let f be given by a CNF formula where each clause has size at most t. Choose a random restriction ρ by setting every variable independently to $*$ with probability p, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\Pr[\mathcal{D}(f_\rho) > s] \leq (5pt)^s.$$

We are going to prove this lemma by induction. But for the induction to work one needs to strengthen the statement:

LEMMA 4.2.3 (Håstad's switching lemma, stronger version). *Let f be given by a CNF formula where each clause has size at most t. Choose a random restriction ρ by setting every variable independently to $*$ with probability p, and to 0 and 1 each with probability $\frac{1-p}{2}$. For every function $F : \{0,1\}^n \to \{0,1\}$, we have*

$$\Pr[\mathcal{D}(f_\rho) > s | F_\rho \equiv 1] \leq (5pt)^s. \tag{9}$$

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 without loss of generality, we can assume that there are no negated literals in C_1, and hence

$$C_1 = \bigvee_{i \in T} x_i,$$

for a subset $T \subseteq \{1, \ldots, n\}$. To prove (9) it suffices to prove both

$$\Pr[\mathcal{D}(f_\rho) > s | F_\rho \equiv 1,\ \rho_T \notin \{0,*\}^T] \leq \alpha^s, \tag{10}$$

and

$$\Pr[\mathcal{D}(f_\rho) > s | F_\rho \equiv 1,\ \rho_T \in \{0,*\}^T] \leq \alpha^s. \tag{11}$$

To prove (10) note that

$$\text{L.H.S of (10)} \quad = \quad \Pr[\mathcal{D}(f_\rho) > s \mid (F \wedge C_1)_\rho \equiv 1] = \Pr[\mathcal{D}((\wedge_{i=2}^m C_i)_\rho) > s \mid (F \wedge C_1)_\rho \equiv 1] \leq \alpha^s,$$

where in the last inequality we used the induction hypothesis. It remains to prove (11). Note that if $\rho_T = \vec{0}$, then $f_\rho \equiv 0$ and thus $\mathcal{D}(f_\rho) = 0$. Hence

$$\begin{aligned}
\text{L.H.S of (11)} \quad &= \quad \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} \Pr[\mathcal{D}(f_\rho) > s,\ \rho_Y = \vec{*},\ \rho_{T-Y} = \vec{0} \mid F_\rho \equiv 1,\ \rho_T \in \{0,*\}^T] \\
&\leq \quad \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} \Pr[\rho_Y = \vec{*}, \rho_{T-Y} = \vec{0} \mid F_\rho \equiv 1,\ \rho_T \in \{0,*\}^T] \times \\
&\qquad\qquad \Pr[\mathcal{D}(f_\rho) > s \mid F_\rho \equiv 1,\ \rho_Y = \vec{*},\ \rho_{T-Y} = \vec{0},\ \rho_T \in \{0,*\}^T] \\
&\leq \quad \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} \Pr\left[\rho_Y = \vec{*} \middle| F_\rho \equiv 1, \rho_T \in \{0,*\}^T\right] \times \Pr\left[\mathcal{D}(f_\rho) > s \middle| F_\rho \equiv 1, \rho_Y = \vec{*}, \rho_{T-Y} = \vec{0}\right]. \qquad (12)
\end{aligned}$$

Observation 1: Since setting variables in Y to $*$ cannot increase the probability that $F_\rho \equiv 1$, we have

$$\Pr[F_\rho \equiv 1 \mid \rho_Y = \vec{*},\ \rho_T \in \{0,*\}^T] \leq \Pr[F_\rho \equiv 1 \mid \rho_T \in \{0,*\}^T],$$

which by a straightforward argument (See Exercise 4.6.4) using the formula $\Pr[A|B]\Pr[B] = \Pr[A \wedge B]$ leads to

$$\Pr[\rho_Y = \vec{*} \mid F_\rho \equiv 1,\ \rho_T \in \{0,*\}^T] \leq \Pr[\rho_Y = \vec{*} \mid \rho_T \in \{0,*\}^T] = \left(\frac{2p}{1+p}\right)^{|Y|} \leq (2p)^{|Y|}.$$

Observation 2: Define $G : \{0,1\}^n \to \{0,1\}$ as

$$G : x \mapsto \begin{cases} 0 & x_{T\setminus Y} \neq \vec{0} \\ F(x) & x_{T\setminus Y} = \vec{0}. \end{cases}$$

Consider $\rho \in \{0,*\}^n$ with $\rho_Y = \vec{*}$. For $\sigma \in \{0,1\}^Y$, let ρ_σ denote the restriction $\rho_\sigma = (\sigma, \rho_{[n]\setminus Y})$. That is ρ_σ is the same as ρ except that to the variables in Y, the restriction ρ_σ assigns the values of σ instead of $*$'s. Note that $\mathcal{D}(f_\rho) \leq |Y| + \max_\sigma \mathcal{D}(f_{\rho_\sigma})$ (See Exercise 4.6.5). Thus by induction hypothesis,

$$\begin{aligned} &\Pr[\mathcal{D}(f_\rho) > s \mid F_\rho \equiv 1, \rho_Y = \vec{*}, \rho_{T\setminus Y} = \vec{0}] \\ &\leq \Pr\left[\exists \sigma \in \{0,1\}^Y,\ \mathcal{D}(f_{\rho_\sigma}) > s - |Y| \ \middle|\ F_\rho \equiv 1, \rho_{T\setminus Y} = \vec{0}\right] \\ &\leq \sum_{\sigma \in \{0,1\}^Y} \Pr[\mathcal{D}(f_{\rho_\sigma}) > s - |Y| \mid G_\rho \equiv 1] \\ &\leq \sum_{\sigma \in \{0,1\}^Y} \alpha^{s-|Y|} \leq (2^{|Y|} - 1)\alpha^{s-|Y|}. \end{aligned}$$

Combining the two observations with (12), we finish the proof:

$$\text{L.H.S of (11)} \leq \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} (2^{|Y|} - 1)\alpha^{s-|Y|}(2p)^{|Y|} = \alpha^s \sum_{\substack{Y \subseteq T \\ Y \neq \emptyset}} \left(\left(\frac{4p}{\alpha}\right)^{|Y|} - \left(\frac{2p}{\alpha}\right)^{|Y|}\right) \leq \alpha^s.$$

□

REMARK 4.2.4. Since the negation of a CNF is a DNF and vice versa, the switching lemma can be used to convert a DNF formula with clauses of size at most t to a CNF with clauses of size at most s in the same way as Lemma 4.2.3. However the statement that "the (conjunctive) clauses in the obtained DNF accept different points" now becomes that "the (disjunctive) clauses in the obtained CNF reject different points". ■

COROLLARY 4.2.5. *Let f be a Boolean function computed by an* AC *circuit of size M and depth d whose output gate is $\wedge$. Choose a random restriction ρ by setting every variable independently to $*$ with probability $p = \frac{1}{10^d s^{d-1}}$, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\Pr[\mathcal{D}(f_\rho) > s] \leq M 2^{-s}.$$

PROOF. We view the restriction ρ as obtained by first having a random restriction ρ_0 with $\Pr[*] = 1/10$, and then $d-1$ consecutive restrictions $\rho_1, \ldots, \rho_{d-1}$ each with $\Pr[*] = \frac{1}{10s}$. With high probability, after the restriction ρ_0, at the bottom level of the circuit all fanins are at most s. To see this, consider two cases for each gate at the bottom level of the original circuit:

(1) The original fanin is at least $2s$. In this case, the probability that the gate was not eliminated by ρ_0, that is, that no input to this gate got assigned a 1 (assuming without loss of generality that the bottom level is an $\vee$ level) is at most $(0.55)^{2s} < 2^{-s}$.

(2) The original fanin is at most $2s$. In this case, the probability that at least s inputs got assigned a $*$ by ρ_0 is at most $\binom{2s}{s}(1/10)^s \leq 2^{-s}$.

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 $\Pr[*] = \frac{1}{10s}$. After each of these, we use Håstad's switching lemma (see Remark 4.2.4) to convert the lower two levels from CNF to DNF (or vice versa), and 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 $\mathcal{D}(g_{\rho_i}) > s$, is bounded by $(5\frac{1}{10s}s)^s \leq 2^{-s}$. 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 fanin at most s. We now apply the last restriction and by switching lemma we get a function f_ρ with $\mathcal{D}(f_\rho) \geq s$. 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 contributed 2^{-s} probability of failure exactly once. □

Note that if in the above proof we stop before applying the last restriction ρ_{d-1}, then we obtain the following corollary which uses a larger value for p.

COROLLARY 4.2.6. *Let f be a Boolean function computed by an* AC *circuit of size M and depth $d \geq 2$ whose output gate is $\wedge$. Choose a random restriction ρ by setting every variable independently to $*$ with probability $p = \frac{1}{10^{d-1}s^{d-2}}$, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\Pr[f_\rho \text{ does not have a CNF with fanin} \leq s] \leq M2^{-s}.$$

Similarly if the output gate of the original circuit is $\vee$, then the probability that f_ρ does not have a DNF with fanin $\leq s$ is bounded by $M2^{-s}$.

4.3. Influences in bounded depth circuits

Let us now introduce an important notion in the study of Boolean functions $f : \{0,1\}^n \to \{0,1\}$.

DEFINITION 4.3.1 (Influence). *Let $f : \mathbb{Z}_2^n \to \{0,1\}$. The* influence *of the ith variable on f is the probability that changing the ith coordinate changes the the value of f. That is,*

$$I_i(f) = \Pr[f(x) \neq f(x + e_i)],$$

where $x \in \{0,1\}^n$ is sampled uniformly and e_i is the i-th standard vector. The total influence of f is defined as

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

Note that always $0 \leq I_i(f) \leq 1$ and $0 \leq I_f \leq n$. The parity function $f : x \mapsto x_1 + \ldots + x_n \pmod 2$ has total influence n, and a constant function has total influence 0.

The *sensitivity* of a point x with respect to f, denoted by $s_f(x)$, is the number of coordinates i for which $f(x) \neq f(x + e_i)$. Since

$$I_f = \mathbb{E}[s_f(x)], \tag{13}$$

sometimes I_f is called the average sensitivity of f.

Our next goal is to show that the total influence of small circuits of small depth cannot be large. First we consider the CNF and the DNF circuits with small clauses.

LEMMA 4.3.2. *Let f be a CNF or a DNF formula where all the clauses are of size at most s. Then $I_f \leq 2s$.*

PROOF. We prove the lemma for the DNF case, and the CNF case follows by replacing f with $1-f$. For $x \in \mathbb{Z}_2^n$, let $s_{1\to 0}(x)$ denote the number of $i \in [n]$ such that $f(x) = 1$ and $f(x+e_i) = 0$. Trivially if $f(x) = 0$, then $s_{1\to 0}(x)$. Since f is represented by a DNF with clauses of size at most s, for every x, we have $s_{1\to 0}(x) \leq s$. Hence

$$I_i(f) = \sum_{i=1}^{n} \Pr[f(x) \neq f(x+e_i)] = \sum_{i=1}^{n} 2\Pr[(f(x)=1) \wedge (f(x+e_i)=0)] = 2\mathbb{E}[s_{1\to 0}(x)] \leq 2s.$$

□

THEOREM 4.3.3 (Boppana [**Bop97**]). *Let f be a Boolean function computed by an AC circuit of depth d and size M, then*

$$I_f \leq 3(20\log M)^{d-1}.$$

PROOF. Applying Corollary 4.2.6 with $s = 2\log M$ and $p = \frac{1}{10^{d-1}s^{d-2}}$, and combining it with Lemma 4.3.2 we conclude that

$$\Pr[I_{f_\rho} \geq 2s] \leq M2^{-s} \leq \frac{1}{M} \leq \frac{1}{n}.$$

Here we are assuming $n \leq M$ by counting the input gates in the size of the circuit. Hence

$$\mathbb{E}_\rho[I_{f_\rho}] \leq \Pr[I_{f_\rho} > 2s]n + 2s \leq \frac{1}{n}n + 2s \leq 2s+1 \leq 3s.$$

On the other hand trivially

$$\mathbb{E}_\rho[I_{f_\rho}] = \mathbb{E}_{\rho,x}[s_{f_\rho}(x)] = p\mathbb{E}[s_f(x)] = pI_f.$$

Hence

$$I_f \leq \frac{3s}{p} \leq 3(10s)^{d-1} = 3(20\log M)^{d-1}.$$

□

4.4. The Fourier tail of functions with small bounded depth circuits

Recall that the characters of $\mathbb{Z}_2^n$ are $\chi_S : x \mapsto (-1)^{\sum_{i\in S} x_i}$ for $S \subseteq [n]$. So in this notation, the Fourier expansion of $f : \mathbb{Z}_2^n \to \mathbb{C}$ is $f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S$. We think of $|S|$ as the "frequency" of the character χ_S. This corresponds to the fact that when $|S|$ is small, χ_S is more stable under local changes (e.g. change of one random coordinate).

As we defined in Section 2.2, the Fourier degree of a function $f : \mathbb{Z}_2^n \to \mathbb{C}$, denoted by $\deg_{\mathcal{F}}(f)$, is the size of the largest S such that $\widehat{f}(S) \neq 0$. For a positive integer k, and a function $f : \mathbb{Z}_2^n \to \mathbb{C}$, we define

$$f^{=k} := \sum_{S:|S|\leq k} \widehat{f}(S)\chi_S,$$

and $f^{\leq k}$, $f^{\geq k}$, $f^{<k}$ and $f^{>k}$ are defined similarly. Note that by the Parseval identity,

$$\|f\|_2^2 = \sum_{k=0}^{n} \|f^{=k}\|_2^2.$$

We leave the proof of the following easy lemma as an exercise to the reader (See Exercise 4.6.6).

LEMMA 4.4.1. *Let $f : \{0,1\}^n \to \{0,1\}$ be computed by a $\wedge$-clause of size s without repeated variables. Then $\deg_{\mathcal{F}}(f) = s$.*

REMARK 4.4.2. Identifying $\{0,1\}^n$ with the n-dimensional hypercube, note that f in Lemma 4.4.1 is the indicator function of a sub-cube of codimension s. So Lemma 4.4.1 is equivalent to the fact that the degree of the indicator function of a subcube of the hypercube $\{0,1\}^n$ is its codimension. ■

COROLLARY 4.4.3. *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 $*$ with probability $p = \frac{1}{10^d s^{d-1}}$, and to 0 and 1 each with probability $\frac{1-p}{2}$. Then*

$$\Pr[\deg_{\mathcal{F}}(f_\rho) > s] \leq M2^{-s}.$$

PROOF. Since $\deg_{\mathcal{F}}(g) = \deg_{\mathcal{F}}(1-g)$ for every function g, we can assume that the output gate of the circuit computing f is $\vee$ (otherwise we replace f with $1-f$ and negate the circuit). Now by Corollary 4.2.5 with probability at least $1 - M2^{-s}$, we have $f_\rho = \vee_{i=1}^m C_i$ for $\wedge$ clauses $C_1, \ldots, C_m$, each of size at most s, such that the clauses all accept disjoint sets of points (*i.e.* no $x \in \{0,1\}^n$ satisfies more than one clause). By the latter property we can write $f_\rho = \sum_{i=1}^m C_i$, where here we are identifying clauses with the functions represented by them. By Lemma 4.4.1, we know $\deg_{\mathcal{F}}(C_i) \leq s$ for all $1 \leq i \leq m$. Hence the degree of their sum is also at most s. We conclude

$$\Pr[\deg_{\mathcal{F}}(f_\rho) > s] \leq M2^{-s}.$$

□

Now we are at the point to prove the main theorem of this section.

THEOREM 4.4.4 (Linial, Mansour, Nisan [**LMN93**]). *Let f be a Boolean function computed by an AC circuit of depth d and size M, and let t be any integer. Then*

$$\sum_{|S|>t} |\widehat{f}(S)|^2 \leq 2M2^{-t^{1/d}/20}.$$

PROOF. Consider a random restriction $\rho \in \{0,1,*\}^n$ with $\Pr[*] = p \leq \frac{1}{10^d k^{d-1}}$ for values of k and p to be determined later. We sample ρ in two steps. First we pick $T \subseteq [n]$ corresponding to the positions that are not assigned a $*$. Then we pick $x_T \in \{0,1\}^T$ uniformly at random, and ρ is defined as $\rho := (x_T, \vec{*})$. Set $f_{x_T} := f_\rho = f(x_T, \cdot)$. Since $\chi_S(x) = \prod_{i \in S}(-1)^{x_i}$, we can decompose it as

$$\chi_S(x) = \chi_{S \cap T}(x_T)\chi_{S \setminus T}(x_{\overline{T}}).$$

Note that $f_{x_T} : \{0,1\}^{\overline{T}} \to \{0,1\}$ and since

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S(x) = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_{S \cap T}(x_T)\chi_{S \setminus T}(x_{\overline{T}}) = \sum_{A \subseteq \overline{T}} \left(\sum_{B \subseteq T} \widehat{f}(A \cup B)\chi_B(x_T) \right) \chi_A(x_{\overline{T}}),$$

we have

$$\widehat{f_{x_T}}(A) = \sum_{B \subseteq T} \widehat{f}(A \cup B)\chi_B(x_T),$$

for every $A \subseteq \overline{T}$. Hence by the Parseval identity

$$\mathbb{E}_{x_T} \left| \widehat{f_{x_T}}(A) \right|^2 = \sum_{B \subseteq T} |\widehat{f}(A \cup B)|^2,$$

which shows that

$$\mathbb{E}_{x_T}\left\|f_{x_T}^{>k}\right\|_2^2 = \mathbb{E}_{x_T}\sum_{\substack{A\subseteq\overline{T}\\|A|>k}}\left|\widehat{f_{x_T}}(A)\right|^2 = \sum_{\substack{A\subseteq\overline{T}\\|A|>k}}\sum_{B\subseteq T}|\widehat{f}(A\cup B)|^2 = \sum_{S:|S\cap\overline{T}|>k}|\widehat{f}(S)|^2.$$

Now we use the randomness in T. Since $f_{x_T}^{>k}=0$ if $\deg_{\mathcal{F}}(f_\rho)\le k$, and that always $\|f_{x_T}^{>k}\|_2^2 \le \|f_{x_T}\|_2^2\le 1$, we have

$$\mathbb{E}_T\left[\sum_{S:|S\cap\overline{T}|>k}|\widehat{f}(S)|^2\right] = \mathbb{E}_T\mathbb{E}_{x_T}\left\|f_{x_T}^{>k}\right\|_2^2 = \mathbb{E}_\rho\left\|f_\rho^{>k}\right\|_2^2 \le \Pr[\deg_{\mathcal{F}}(f_\rho)>k]\le M2^{-k}, \tag{14}$$

where the last inequality follows from Corollary 4.4.3 as we chose $\Pr[*]=p\le\frac{1}{10^d k^{d-1}}$. Also we can bound the left-hand size of (14) from below:

$$\text{L.H.S. of (14)} = \sum_{S\subseteq[n]}\Pr[|S\cap\overline{T}|>k]|\widehat{f}(S)|^2 \ge \sum_{|S|>t}\Pr[|S\cap\overline{T}|>k]|\widehat{f}(S)|^2.$$

Taking $p=\frac{1}{10t^{(d-1)/d}}$ and $k=t^{1/d}/20$, satisfies $p\le\frac{1}{10^d k^{d-1}}$, and by the Chernoff bound (Lemma 1.4.4) for $|S|>t$, the probability of $|S\cap\overline{T}|>k=pt/2$ is at least $1-2e^{\frac{-pt}{12}}\ge\frac{1}{2}$. Hence by (14), we have

$$\sum_{S:|S|>t}\frac{1}{2}|\widehat{f}(S)|^2 \le M2^{-t^{1/d}/20}.$$

□

4.5. The Razborov-Smolensky Theorem

Taking $g=f^{\le t}$, Theorem 4.4.4 shows that $\|f-g\|_2^2\le 2M2^{-t^{1/d}/20}$. In other words circuits of low depth and small size can be approximated by functions of low Fourier degree in the L_2 norm. The next theorem shows a different type of approximating such functions with low degree functions.

THEOREM 4.5.1 ([**Raz87**], [**Smo87**]). *Let $f:\{0,1\}^n\to\{0,1\}$ be computed by a circuit of depth d and size M. For any s, there is a function g with degree $r\le(s\log M)^d$ such that*

$$\Pr[f(x)\ne g(x)]\le\left(1-\frac{1}{2e}\right)^s M.$$

PROOF. The function g is constructed in an inductive way. We will show how to make a step with an $\wedge$ gate. Since the whole construction is symmetric with respect to 0 and 1, the step also holds for an $\vee$ gate. Let

$$f=\wedge_{i=1}^k f_i$$

where $k<M$. For convenience, let us assume that $k=2^\ell$ is a power of 2. For every $p=2^{-1},2^{-2},\dots,2^{-\ell}=1/k$ we pick s random subsets of $\{1,\dots,k\}$ by including every element in the subset independently with probability p. We obtain a collection of sets $S_1,\dots,S_t$ with $t=s\ell\le s\log M$. Let $g_1,\dots,g_k$ be the approximating functions for $f_1,\dots,f_k$ provided by the previous inductive step. We set

$$g:=\prod_{i=1}^t(1-|S_j|+\sum_{j\in S_i}g_j).$$

By the induction assumption, the degrees of g_j are $\leq (s \log m)^{d-1}$, hence the degree of f is bounded by $t(s \log m)^{d-1} \leq (s \log m)^d$. Next we bound the probability of $f(x) \neq g(x)$ conditioned on the event that all of the inputs $f_1, \ldots, f_k$ are calculated correctly. We have

$$\Pr[f(x) \neq g(x) | g_j = f_j \text{ for all } j] = \Pr\left[\prod_{i=1}^{t} \left(1 - |S_i| + \sum_{j \in S_i} f_j\right) \neq \prod_{j=1}^{k} f_j\right].$$

To bound this we fix a vector of specific values $f_1(x), \ldots, f_k(x)$ and calculate the probability of an error over the possible choices of the random sets S_i. Note that 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). Let $1 \leq z \leq k$ be the number of zeros among $f_1(x), \ldots, f_k(x)$, and $\alpha \in \mathbb{Z}$ be such that $2^\alpha \leq z < 2^{\alpha+1}$. Let S be a random set with parameter $p = 2^{-\alpha-1}$. Our approximation will be correct if S hits exactly one 0 among the z zeros of $f_1(x), \ldots, f_k(x)$. The probability of this event is exactly

$$zp(1-p)^{z-1} \geq \frac{1}{2}(1-p)^{1/p-1} > \frac{1}{2e}.$$

Hence the probability of being wrong after s such sets are being chosen is bounded by $(1 - \frac{1}{2e})^s$ and

$$\Pr\left[\prod_{i=1}^{t}(1 - |S_i| + \sum_{j \in S_i} f_j) \neq \prod_{j=1}^{k} f_j\right] < \left(1 - \frac{1}{2e}\right)^s.$$

By making the same probabilistic argument at every node, by the union bound we conclude that the probability that an error happens is at most $M\left(1 - \frac{1}{2e}\right)^s$. □

4.6. Conclusion and open problems

As we shall see later in (**??**), we have $I_f = 4 \sum_{S \subseteq [n]} |S| |\widehat{f}(S)|^2$. Hence for every $\epsilon > 0$,

$$\sum_{S : |S| \geq I_f/\epsilon} |\widehat{f}(S)|^2 \leq \epsilon.$$

Note that by Theorem 4.3.3, an AC circuit of polynomial size and constant depth d satisfies $I_f = O((\log n)^{d-1})$. Thus most of the L_2 Fourier mass of such functions is concentrated on the first $O((\log n)^{d-1})$ levels. In the case of $d = 2$, where f is a CNF or a DNF of polynomial size, we conclude that $I_f = O(\log n)$ and the L_2 Fourier mass is concentrated on the first $O(\log n)$ levels. There are $n^{O(\log n)}$ Fourier coefficients in those levels. Mansour[**Man95**] conjectured that these Fourier coefficients are concentrated only on polynomially many values of S.

CONJECTURE 4.6.1 (Mansour[**Man95**]). *Consider a Boolean function $f : \{0,1\}^n \to \{0,1\}$ described by a CNF or a DNF of polynomial size, and a constant $\epsilon > 0$. There is a set $\mathcal{S}$ of polynomial size (depending on ϵ) such that*

$$\sum_{S \notin \mathcal{S}} \widehat{f}(S)^2 \leq \epsilon.$$

For monotone functions an even stronger statement is conjectured in [**BKS99**]. No counter-example is known even for the non-monotone case.

CONJECTURE 4.6.2 ([**BKS99**, Conjecture 7.2]). *Let $\epsilon > 0$ be a fixed real number and $d \geq 1$ be a fixed integer. Let $f : \{0,1\}^n \to \{0,1\}$ be a monotone function expressed by a depth-d circuit of*

size M. There is a set S of polynomial size in M (where the polynomial depends on d and ϵ) so that

$$\sum_{S \notin \mathcal{S}} \widehat{f}(S)^2 \leq \epsilon.$$

It would be really interesting to prove an inverse for Boppana's theorem, Theorem 4.3.3. In fact 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 $O(\log n)$.

Focusing only on the $O(\log n)$ case. I do believe both a negative or a positive answer to the following problem would be very interesting.

PROBLEM 4.6.3 (Open problem). *Is it true that for every $\epsilon, C > 0$ there exist constants $d, k \in \mathbb{N}$ and $\delta > 0$ such that for every $f : \{0,1\}^n \to \{0,1\}$ with total influence $I_f \leq C \log n$, there exists an AC circuit g of depth d and size n^k satisfying*

$$\Pr[f(x) \neq g(x)] \leq \epsilon.$$

Exercises

EXERCISE 4.6.4. *In the proof of Lemma 4.2.3, fill the gaps in the argument of Observation 1.*

EXERCISE 4.6.5. *For $\rho \in \{0,*\}^n$ with $\rho_Y = \vec{*}$ show that $\mathcal{D}(f_\rho) \leq |Y| + \max_\sigma \mathcal{D}(f_{\rho_\sigma})$ where the maximum is over all $\sigma \in \{0,1\}^Y$.*

EXERCISE 4.6.6. *Prove Lemma 4.4.1.*

EXERCISE 4.6.7. *Let $f : \{0,1\}^n \to \{0,1\}$ be computed by a decision tree (nodes are labeled with variables and at every node we branch according to the value of that variables) of depth d. Prove that $\mathcal{D}(f) \leq d$ and $\mathcal{D}(1-f) \leq d$.*

CHAPTER 5

Applications to Computer Science: Machine Learning

In this chapter we overview some of the applications of the Fourier analysis to an important area of computer science called machine learning. Machine learning concerns the construction and study of systems that can learn from data. Theoretical results in machine learning mainly deal with a type of i learning called supervised learning. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples that are not in the training data. In other words the algorithm observes the value of a function on some points (training data), and then tries to predict the value of the function on the rest of the points.

One of the most popular computational learning models is Valiant's *Probably approximately correct* (PAC) model of learning from random examples []. In this framework, the learner receives samples from an known function f belonging to a certain class of possible functions (called concept class), and must produce a generalization function (called the *hypothesis*) that is a low-error approximation of f.

In this course, our concept classes $\mathcal{C}$ consist only of Boolean function. That is

$$\mathcal{C} \subseteq \{f : \{0,1\}^n \to \{0,1\} \ : \ n \in \mathbb{N}\}.$$

For example our concept class might be the class of all dictator functions, where a dictator function maps $x = (x_1, \ldots, x_n)$ to x_i for some fixed i.

The goal of a learning algorithm A for a concept class $\mathcal{C}$ is to identify an unknown function $f \in \mathcal{C}$ by using random examples from this function only. In particular, the probabilistic algorithm A takes as input an accuracy parameter $\epsilon > 0$ and a confidence parameter $\delta > 0$. It has access to an example oracle $\mathrm{EX}(f, \mu)$ where μ is a distribution on $\{0,1\}^n$. When queried, the example oracle provides the learning algorithm with an example $[x, f(x)]$, where x is drawn from the distribution μ. The output of A is a hypothesis h, which is a boolean function. The hypothesis h is said to be ϵ-close to f if

$$\mathrm{Pr}_{x\sim\mu}[f(x) \neq h(x)] \leq \epsilon.$$

We say that A is a learning algorithm for $\mathcal{C}$ if for all $f \in \mathcal{C}$ and distributions μ, when A is run with example oracle $\mathrm{EX}(f, \mu)$, with probability at least $1 - \delta$ it outputs a hypothesis which is ϵ-close to f.

Since PAC learning in its full generality seems to be very difficult for many natural concept classes, often some of its requirements are relaxed. One of the most frequently studied relaxations is uniform-distribution PAC learning in which the algorithm need only work when the distribution μ is the uniform distribution over $\{0,1\}^n$.

Another relaxation of the PAC model is to allow the learning algorithm to make membership queries, that is the learner is allowed to ask for the value of the target function f on points of its choosing. This model gives the learner considerably more power than usual and is thus a significant weakening.

5.1. Uniform-distribution PAC learning

Consider a target function $f : \{0,1\}^n \to \{0,1\}$ and suppose that we have an example oracle that provides us with samples $[x, f(x)]$ where x is drawn from the uniform distribution. Suppose for the moment that our task is to learn one of the Fourier coefficients of the function f, say $\widehat{f}(S)$. Note that $\widehat{f}(S) = \mathbb{E}[f(x)\chi_S(x)]$, and we can try to estimate this average by $a_S := \frac{1}{N}\sum_{i=1}^{N} f(x_i)\chi_S(x_i)$ where $x_1, \ldots, x_N \in \{0,1\}^n$ are provided by the example oracle and hence are independent, each is drawn from the uniform distribution. Using the Chernoff bound (Lemma **??**)

$$\Pr[|a_S - \widehat{f}(S)| \geq \lambda] \leq 2e^{-\lambda^2 N/2}. \tag{15}$$

Hence by taking N to be sufficiently large, with high probability, we can obtain a very accurate estimate of $\widehat{f}(S)$. Now let us see how one can use this to devise a general approach to uniform-distribution PAC learning.

Suppose that there exists a collection $\mathcal{B}$ of subsets of $[n]$ such that for every function f in the concept class $\mathcal{C}$, we have

$$\|f - \sum_{S\in\mathcal{B}} \widehat{f}(S)\chi_S\|^2 = \sum_{S\notin\mathcal{B}} |\widehat{f}(S)|^2 \leq \epsilon.$$

Then we can try to estimate $\widehat{f}(S)$ for all $S \in \mathcal{B}$ as above, and then approximate f with $g := \sum_{S\in\mathcal{B}} a_S\chi_S$. Note that

$$\|f - g\|_2^2 = \sum_{S\notin\mathcal{B}} |\widehat{f}(S)|^2 + \sum_{S\in\mathcal{B}} |\widehat{f}(S) - a_S|^2 \leq \epsilon + \sum_{S\in\mathcal{B}} |\widehat{f}(S) - a_S|^2,$$

and by taking $\lambda = \sqrt{\frac{\epsilon}{|\mathcal{B}|}}$ and $N = 2|\mathcal{B}|\epsilon^{-1}\log(2|\mathcal{B}|/\delta)$ in (15), we obtain that with probability $1 - 2e^{-\lambda^2 N/2}|\mathcal{B}| = 1 - \delta$, for all $S \in \mathcal{B}$, we have $|a_S - \widehat{f}(S)| < \lambda$ and consequently

$$\|f - g\|_2^2 \leq \epsilon + \lambda^2|\mathcal{B}| \leq 2\epsilon. \tag{16}$$

Hence with probability at least $1 - \delta$, the function g is a good approximation of f as $\|f - g\|_2^2 \leq 2\epsilon$. However we can not output g as the hypothesis function as it is not necessarily a Boolean function. This can be easily remedied by rounding g to a Boolean function. Namely, we define

$$h(x) := \begin{cases} 1 & g(x) \geq \frac{1}{2} \\ 0 & g(x) < \frac{1}{2} \end{cases}$$

Since f is a Boolean function itself, we have

$$\Pr[f(x) \neq h(x)] = \mathbb{E}[|f(x) - h(x)|^2] \leq 4\mathbb{E}[|f(x) - g(x)|^2] = 4\|f - g\|_2^2,$$

which is at most 8ϵ if (16) holds. Thus we proved the following theorem.

THEOREM 5.1.1 (Linial, Mansour, Nisan [**LMN93**]). *Suppose that for a concept class $\mathcal{C}$, there exists a collection $\mathcal{B}$ of size m of subsets of $[n]$ such that for every function $f \in \mathcal{C}$ we have*

$$\sum_{S\notin\mathcal{B}} |\widehat{f}(S)|^2 \leq \epsilon.$$

Then there is an algorithm that given access to $\mathcal{B}$ and the oracle example for f with uniform distribution, runs in time $O(m\epsilon^{-1}\log(m/\delta))$ and with probability at least $1 - \delta$ produces a function $h : \{0,1\}^n \to \{0,1\}$ such that $\Pr[f(x) - h(x)] = O(\epsilon)$.

As we discussed in Section 4.6 it follows from Boppana's theorem (Theorem 4.3.3) that if f is computable by a circuit of polynomial size and depth d, then for some $K = O_\epsilon((\log n)^{d-1})$, we have

$$\sum_{|S|>K} |\widehat{f}(S)|^2 \leq \epsilon.$$

Note that there are $\sum_{i=0}^{K} \binom{n}{i} = n^K$ subsets of $[n]$ of size at most K. Hence we conclude that the concept class of functions computable by circuits of depth d and of size bounded by a given polynomial are PAC learn-able under the uniform distribution in time $n^{O((\log n)^{d-1})} = e^{C(\log n)^d}$ for some $C = O(1)$.

5.2. PAC learning under the query model

In the previous section we showed how to learn concept classes for which the Fourier spectrum of the functions in the class concentrated on a small set $\mathcal{B}$, by approximating their Fourier coefficients on sets that belong to $\mathcal{B}$. However, it might be the case that a function can be approximated by a small number of coefficients, but these coefficients do not come from a fixed set $\mathcal{B}$ for all the functions in the concept class. In this section we describe a learning algorithm that learns the target function without necessarily having access to the set $\mathcal{B}$. However, there is a great disadvantage, that is the algorithm uses the query model.

THEOREM 5.2.1. *Suppose that for every function $f : \{0,1\}^n \to \{0,1\}$ in a concept class $\mathcal{C}$, there exists a collection $\mathcal{B}$ of size m of subsets of $[n]$ such that*

$$\sum_{S \notin \mathcal{B}} |\widehat{f}(S)|^2 \leq \epsilon.$$

There is an algorithm that queries the value of f on $O(m\epsilon^{-1}\log(m/\delta))$ points and produces a function $h : \{0,1\}^n \to \{0,1\}$ such that $\Pr[f(x) - h(x)] = O(\epsilon)$ with probability at least $1 - \delta$.

Consider the Fourier expansion of f,

$$f = \sum_{z \in \mathbb{Z}_2^n} \widehat{f}(z)\chi_z.$$

The learning algorithm in Theorem 5.2.1 is based on detecting the large Fourier coefficients of f. Suppose that we want to detect all the Fourier coefficients that are larger than $\theta := \sqrt{\frac{\epsilon}{|\mathcal{B}|}} > 0$ in absolute value. By Parseval the number of such coefficients is at most $1/\theta^2$. Defining $g = \sum_{z:|\widehat{f}(z)|\geq\theta} \widehat{f}(z)\chi_z$, we have

$$\|f - g\|_2^2 = \sum_{S:|\widehat{f}(z)|<\theta} |\widehat{f}(z)|^2 \leq \epsilon + \sum_{\substack{z:|\widehat{f}(z)|<\theta \\ z\in\mathcal{B}}} |\widehat{f}(z)|^2 \leq \epsilon + \theta^2|\mathcal{B}| \leq 2\epsilon.$$

If we can obtain a good approximation of g, then we can round it to a Boolean function as in the proof of Theorem 5.1.1 by defining

$$h(x) := \begin{cases} 1 & g(x) \geq \frac{1}{2} \\ 0 & g(x) < \frac{1}{2} \end{cases}$$

We have $\Pr[f(x) \neq h(x)] \leq 4\|f - g\|_2^2 \leq 8\epsilon$.

So from this point on, we will focus on finding an approximation of the function g. To this end it suffices to find all z for which $|\widehat{f}(z)| \geq \theta$, as once we find those z we can approximate $\widehat{f}(z)$ empirically as in (15).

The algorithm partitions coefficients according to their prefix. For $\alpha \in \mathbb{Z}_2^k$, let $f_\alpha : \mathbb{Z}_2^{n-k} \to \mathbb{R}$ be defined as

$$f_\alpha(x) = \mathbb{E}_{y \in \mathbb{Z}_2^k} f(y,x)\chi_\alpha(y).$$

We can simplify this as

$$f_\alpha(x) = \sum_{(z_1,z_2)\in\mathbb{Z}_2^n} \widehat{f}(z_1,z_2)\chi_{z_2}(x)\mathbb{E}_y\left[\chi_{z_1+\alpha}(y)\right] = \sum_{(\alpha,z_2)\in\mathbb{Z}_2^n} \widehat{f}(\alpha,z_2)\chi_{z_2}(x),$$

where in the first sum $z_1 \in \mathbb{Z}_2^k$ and $z_2 \in \mathbb{Z}_2^{n-k}$. By Parseval

$$\mathbb{E}_x[f_\alpha(x)^2] = \sum_{(\alpha,z_2)\in\mathbb{Z}_2^n} |\widehat{f}(\alpha,z_2)|^2.$$

Note that

$$\mathbb{E}_x[f_\alpha(x)^2] = \mathbb{E}_{x\in\mathbb{Z}_2^{n-k}}\left(\mathbb{E}_{y\in\mathbb{Z}_2^k} f(y,x)\chi_\alpha(y)\right)^2 = \mathbb{E}_{x\in\mathbb{Z}_2^{n-k}}\mathbb{E}_{y_1,y_2\in\mathbb{Z}_2^k}\left[f(y_1,x)\chi_\alpha(y_1)f(y_2,x)\chi_\alpha(y_2)\right]$$

Since $|f(y_1,x)\chi_\alpha(y_1)f(y_2,x)\chi_\alpha(y_2)| \le 1$, Chernoff bound implies that if we take sufficiently many random points x, y_1, y_2, we can obtain an accurate approximation for $\mathbb{E}_x[f_\alpha(x)^2]$. More precisely, similar to (15), we can average over random triples x, y_1, y_2, and obtain an approximation a_α such that

$$\Pr\left[\left|a_\alpha - \mathbb{E}_x[f_\alpha(x)^2]\right| \ge \lambda\right] \le 2e^{-\lambda^2 N/2} \tag{17}$$

By setting λ and N properly we can assume that with high probability we can obtain sufficiently precise approximation of $a_\alpha \approx \mathbb{E}_x[f_\alpha(x)^2]$ for a given value of α. Hence to simplify the presentation we assume from now on that we can learn the value of $\mathbb{E}_x[f_\alpha(x)^2]$ correctly. Note that if $\mathbb{E}_x[f_\alpha(x)^2] < \theta^2$, then we know that for all $z = (\alpha, z_2)$ we have $|\widehat{f}(z)| < \theta$. Hence running the following subroutine for $\alpha = \emptyset$ outputs all the Fourier coefficients that are at least θ in absolute value

Subroutine SA(α)

- if $\mathbb{E}_x[f_\alpha(x)^2] \ge \theta^2$ then
- if $|\alpha| = n$, then OUTPUT α
- else run SA$(\alpha, 0)$ and SA$(\alpha, 1)$

Now as mentioned earlier, knowing all the values of z for which $|\widehat{f}(z)| > \theta$, we can find a good approximation h for f.

5.2.1. Functions with bounded spectral norm. The L_1 norm of the Fourier spectrum of a function $\|\widehat{f}\|_1 := \sum |\widehat{f}(a)|$ is sometimes referred to as the spectral norm of f. The concept classes consisting of Boolean functions with small spectrum norms are an important case where Theorem 5.2.1 can be applied.

LEMMA 5.2.2. *Let $\epsilon > 0$, and let $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfy $\|\widehat{f}\|_1 \le M$, then for $\theta = \epsilon/M$, we have*

$$\sum_{S\notin\mathcal{B}_\theta} |\widehat{f}(S)|^2 \le \epsilon,$$

where

$$\mathcal{B}_\theta = \{S : |\widehat{f}(S)| \ge \theta\}.$$

Furthermore $|\mathcal{B}_\theta| \le 1/\theta^2 = M^2/\epsilon^2$.

PROOF. We have

$$\sum_{S \notin \mathcal{B}_\theta} |\widehat{f}(S)|^2 \leq \left(\max_{S \notin \mathcal{B}_\theta} |\widehat{f}(S)|\right) \left(\sum |\widehat{f}(S)|\right) \leq \theta M \leq \epsilon.$$

Furthermore by Parseval

$$1 \geq \sum_{S \in \mathcal{B}_\theta} |\widehat{f}(S)|^2 \geq |\mathcal{B}_\theta| \theta^2,$$

which leads to the desired conclusion. □

Let us now study Boolean functions with small spectral norm. First we recall some facts from Lemma 2.1.9 and Remark 2.1.10. Let $V \subseteq \mathbb{Z}_2^n$ be a linear subspace of co-dimension d. Then there exists linearly independent $a_1, \ldots, a_d \in \mathbb{Z}_2^n$ such that

$$V = \{x : \forall i \ \langle a_i, x \rangle = 0 (\text{mod } 2)\},$$

or equivalently

$$V = \{x : \forall i \ \chi_{a_i}(x) = 1\}.$$

Then $V^\perp = \text{span}\{a_1, \ldots, a_d\}$, and note that

$$\mathbf{1}_V = \sum_{v \in V^\perp} \frac{1}{2^d} \chi_v.$$

It follows that

$$\|\widehat{\mathbf{1}_V}\|_1 = 2^d \frac{1}{2^d} = 1.$$

More generally consider a co-set of V, say $W = y + V$ for some $y \in \mathbb{Z}_2^n$. Then

$$\mathbf{1}_W = \sum_{v \in V^\perp} \frac{\chi_v(y)}{2^d} \chi_v.$$

and thus $\|\widehat{\mathbf{1}_W}\|_1 = 1$. So we established that every co-set of $\mathbb{Z}_2^n$ has spectral norm 1. Exercise 5.3.4 shows that these are the only sets with spectral norm 1. We will use these facts to show that small decision trees, or more generally small parity decision trees provide examples of Boolean functions with small spectral norm.

DEFINITION 5.2.3 (Decision tree). *A decision tree is a labeled binary tree. Each internal node of the tree is labeled with a variable x_i, 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's a leaf, and output its label. Otherwise, query its label x_i. If $x_i = 0$, then recursively evaluate the left subtree, and if $x_i = 1$, evaluate the right subtree.*

For a leaf ℓ of a decision tree, let $L_\ell \subseteq \mathbb{Z}_2^n$ denote the set of all $x \in \mathbb{Z}_2^n$ whose computational path ends in ℓ. Let f be the function computed by the decision tree. We have

$$f = \sum_{\ell : \text{label}(\ell) = 1} \mathbf{1}_{L_\ell}.$$

Note that L_ℓ is a co-set of the subspace

$$\{x : x_i = 0 \text{ for all } i \text{ on the root to } \ell \text{ path}\},$$

and hence $\|\widehat{\mathbf{1}_{L_\ell}}\|_1 \leq 1$. Consequently

$$\|\widehat{f}\|_1 \leq \sum_{\ell : \text{label}(\ell) = 1} \|\widehat{\mathbf{1}_{L_\ell}}\|_1 \leq |\text{Leaves}|.$$

These facts can be generalized to the so called parity decision trees.

DEFINITION 5.2.4 (Parity Decision tree). *A parity decision tree (also denoted as $\oplus$-decision tree) is a labeled binary tree. Each internal node of the tree is labeled with a character χ_a, 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's a leaf, and output its label. Otherwise, query its label $\chi_a(x)$. If $\chi_a(x) = 1$, then recursively evaluate the left subtree, and if $\chi_a(x) = -1$, evaluate the right subtree.*

Note that the value of $\chi_a(x)$ is determined by $\sum_{i \in S} x_i \pmod 2$ where $S \subseteq [n]$ is the corresponding subset of a. Hence one can equivalently assume that the internal nodes are labeled with parity functions $\oplus_S(x) := \sum_{i \in S} x_i \pmod 2$. This justifies the term parity decision trees.

Consider a leaf ℓ of a $\oplus$-decision tree, and let $a_1, \ldots, a_d, \ell$ be the nodes on the path from the root to the leaf ℓ. Then L_ℓ is the set of all x such that the value of $\chi_{a_i}(x)$ is consistent with the path for all $i = 1, \ldots, d$. Thus L_ℓ is a co-set of the subspace

$$\{a_1, \ldots, a_d\}^\perp := \{x : \chi_{a_i} = 1 \ \forall 1 \leq i \leq d\}.$$

So similar to the case of the decision tree we conclude that $\|\widehat{\mathbf{1}_{L_\ell}}\|_1 \leq 1$, and consequently $\|\widehat{f}\|_1$ is bounded by the number of the leaves of the tree. Let us state this as a lemma for the future reference.

LEMMA 5.2.5. *Let f be a Boolean function computed by a $\oplus$-decision tree. Then $\|\widehat{f}\|_1$ is bounded by the number of leaves of the tree.*

Combining this with Lemma 5.2.2 and Theorem 5.2.1 we conclude that the class of $\oplus$-decision trees of size m is learn-able in the query model in time $O\left(m^2\epsilon^{-3}\log\left(\frac{m^2}{\epsilon^2\delta}\right)\right)$.

5.2.2. Inverse Theorems for Small Spectral norm. Lemma 5.2.5 provides a way to construct Boolean functions with small spectral norms. Now let us investigate the structure of all the functions with small spectral norm.

THEOREM 5.2.6 (Sanders 2014+). *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be a boolean function, and suppose that the spectral norm $\|\widehat{f}\|_1$ is at most M. Then There exists subspaces $V_1, \ldots, V_L$ of $\mathbb{Z}_2^n$ for some $L \leq e^{M^{O(1)}}$ such that*

$$f = \sum_{j=1}^{L} \pm \mathbf{1}_{V_j} \tag{18}$$

REMARK 5.2.7. Note that every function f satisfying (18) has spectral norm bounded by L. Theorem 5.2.6 is an improvement over an earlier bout of $L \leq 2^{2^{O(M^4)}}$ due to Sanders and Ben Green [**GS08**]. ∎

We will not prove Theorem 5.2.6 in this course as its proof is based on various results from additive combinatorics. Instead we will prove a recent different inverse theorem for such functions by Shpilka, Tal, and Volk [**SlV13**].

THEOREM 5.2.8 (Shpilka, Tal, and Volk [**SlV13**]). *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfy $\|\widehat{f}\|_1 \leq M$. There exists a co-set V of co-dimension at most M^2 such that f is constant on V.*

To prove this theorem it is more convenient to work with functions $f : \mathbb{Z}_2^n \to \{-1,1\}$. As we have mentioned earlier a function $f : \mathbb{Z}_2^n \to \{0,1\}$ can be easily converted to such a function via the

affine transformation $f \mapsto 2f - 1$, and this will not have a significant effect on the Fourier spectrum of the function.

Consider a function $f : \mathbb{Z}_2^n \to \{-1, 1\}$. 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 mass on pairs whose product is positive is the same as the mass on pairs whose product is negative. In particular, if we consider the two heaviest elements in the Fourier spectrum, say, $f(\alpha)$ and $f(\beta)$, and let $\delta = \alpha + \beta$, then by restricting f to one of the subspaces $\chi_\delta = 1$ or $\chi_\delta = -1$ we get a substantial saving in the spectral norm. This happens since there is a significant L_1 mass on pairs $f(\lambda)$ and $f(\lambda + \delta)$ that have different signs.

Before stating the proof of Theorem 5.2.8, let us discuss the effect that restricting a function to a coset has 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 $f|_V$. Then for every $b \in V$, the coefficients $\widehat{f}(b)$ and $\widehat{f}(b+a)$ collapse to a single coefficient:

$$\widehat{f|_V}(b) := \widehat{f}(b) + \widehat{f}(a+b). \tag{19}$$

Similarly for every $b \in W$,

$$\widehat{f|_V}(b) := \widehat{f}(b) - \widehat{f}(a+b). \tag{20}$$

The following Lemma 5.2.9 is the key part of the proof of Theorem 5.2.8

LEMMA 5.2.9. *Let $f : \mathbb{Z}_2^n \to \{-1, 1\}$ be a Boolean function such that $\|\widehat{f}\|_1 = M > 1$. Then there exists $\gamma \in \mathbb{Z}_2^n$ and $b \in \{-1, 1\}$ such that $\|\widehat{f|_{\chi_\gamma = b}}\|_1 \leq M - 1/M$.*

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 $|\widehat{f}(\alpha)| \geq \frac{1}{M}$. We can assume that $\widehat{\beta} \neq 0$, as otherwise the function f must be of the form $\pm\chi_\alpha$, and that correspond to an $(n-1)$-dimensional coset.

Without loss of generality assume that $\widehat{f}(\alpha)\widehat{f}(\beta) > 0$, i.e. they have the same sign, the other case is completely analogous. By taking the Fourier transform of both sides of $f^2 = 1$, we get that

$$\sum_{\gamma \in \mathbb{Z}_2^n} \widehat{f}(\gamma)\widehat{f}(\alpha + \beta + \gamma) = \widehat{1}(\alpha + \beta) = 0. \tag{21}$$

Let $N_{\alpha+\beta} \subseteq \mathbb{Z}_2^n$ be the set of vectors γ such that $\widehat{f}(\gamma)\widehat{f}(\alpha + \beta + \gamma) < 0$. Note that by assumption, $\alpha, \beta \notin N_{\alpha+\beta}$. Switching sides in (21), we get

$$2\left|\widehat{f}(\alpha)\widehat{f}(\beta)\right| = \sum_{\gamma \in N_{\alpha+\beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right| - \sum_{\substack{\gamma \notin N_{\alpha+\beta} \\ \gamma \neq \alpha, \beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right|.$$

In particular,

$$|\widehat{f}(\alpha)||\widehat{f}(\beta)| \leq \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \left|\widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma)\right|. \tag{22}$$

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:

$$\sum_{\gamma \in N_{\alpha+\beta}} \left| \widehat{f}(\gamma)\widehat{f}(\alpha+\beta+\gamma) \right| \leq |\widehat{f}(\beta)| \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{ |\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)| \right\}. \tag{23}$$

Then (22) and (23) (as well as the assumption $|\widehat{f}(\beta)| \neq 0$) together imply

$$|\widehat{f}(\alpha)| \leq \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{ |\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)| \right\}. \tag{24}$$

Let $f' = f|_{\chi_{\alpha+\beta}=1}$. Then for every γ the coefficients $\widehat{f}(\gamma)$ and $\widehat{f}(\alpha+\beta+\gamma)$ collapse to a single coefficient whose absolute value is $|\widehat{f}(\gamma) + \widehat{f}(\alpha+\beta+\gamma)|$ (recall Equation (19)). 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 get:

$$\|\widehat{f'}\|_1 \leq \|\widehat{f}\|_1 - \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{ |\widehat{f}(\gamma)|, |\widehat{f}(\alpha+\beta+\gamma)| \right\}.$$

Therefore by (24) we have

$$\|\widehat{f'}\|_1 \leq \|\widehat{f}\|_1 - |\widehat{f}(\alpha)| \leq M - \frac{1}{M}.$$

□

Proof of Theorem 5.2.8. Apply Lemma 5.2.9 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 $\|g\|_1 = 1$. Then Exercise 5.3.4 finishes the proof. □

5.3. Concluding remarks and open problems

Recently Tsang, Wong, Xie, and Zhang [**TWXZ13**] noticed that a slight twist in the proof of Theorem 5.2.8 improves the co-dimension in to $O(M)$.

It is not difficult to see that in Lemma 5.2.9 the restriction $f|_{\chi_\gamma \neq b}$ also provides some decrease in the spectral norm. That is $\|\widehat{f|_{\chi_\gamma \neq b}}\|_1 \leq \|f\|_1 - |\widehat{f}(\beta)|$, where $\widehat{f}(\beta)$ is the second largest Fourier coefficient in absolute value. Using this one can prove the following theorem which unfortunately provides a bound that depends on n.

Theorem 5.3.1 (Shpilka, Tal, and Volk [**SlV13**]). *Every $f : \mathbb{Z}_2^n \to \{0,1\}$ with $\|\widehat{f}\|_1 \leq M$ can be computed by a $\oplus$-decision tree of size at most $2^{M^2} n^M$.*

An interesting class of Boolean functions with small spectral norm are the Fourier sparse functions. Consider a Boolean function $f : \mathbb{Z}_2^n \to \{0,1\}$ and denote $\mathrm{Support}(\widehat{f}) := \{a : \widehat{f}(a) \neq 0\}$. Note that the spectral norm is bounded by the size of the Fourier support: $\|\widehat{f}\|_1 \leq |\mathrm{Support}(\widehat{f})|$.

Conjecture 5.3.2 ([**MO09, ZS10**]). *Every $f : \mathbb{Z}_2^n \to \{0,1\}$ can be computed by a $\oplus$-decision tree of depth at most $O(\mathrm{poly}(\log s))$ where $s = |\mathrm{Support}(\widehat{f})|$.*

The same techniques that are used in the proof of Theorem 5.3.1 can be applied to prove the following theorem.

THEOREM 5.3.3 (Shpilka, Tal, and Volk [**SlV13**]). *Every $f : \mathbb{Z}_2^n \to \{0,1\}$ with $\|\widehat{f}\|_1 \leq M$ can be computed by a $\oplus$-decision tree of depth at most $M^2 \log s$ where $s = |\mathrm{Support}(\widehat{f})|$.*

Exercises

EXERCISE 5.3.4. *Show that if $f : \mathbb{Z}_2^n \to \{-1,1\}$ satisfies $\|\widehat{f}\|_1 = 1$, then $f = \pm\chi_\alpha$ for some $\alpha \in \mathbb{Z}_2^n$.*

EXERCISE 5.3.5. *Show that if $f : \mathbb{Z}_2^n \to \{0,1\}$ can be computed by a $\oplus$-decision tree of depth d then $|\mathrm{Support}(\widehat{f})| \leq 4^d$.*

CHAPTER 6

Hypercontractivity, Friedgut's Theorem, KKL inequality

We start the study of an important tool in harmonic analysis. Namely, the *hypercontractivity* of the noise operator. The ideas and results were developed by different people (Bonami, Beckner, Ornstein-Uhlenbeck, Gross, Nelson) in different contexts.

6.1. The noise operator

We begin by formally introducing the noise operator in dimension one, i.e. on the space of functions $f : \mathbb{Z}_2 \to \mathbb{C}$. Let μ_p denote the Bernoulli distribution with success probability p (that is $\mu_p(\{1\}) = p$ and $\mu_p(\{0\}) = 1 - p$).

DEFINITION 6.1.1 (The 1-dimensional noise operator). *Let $0 \leq \rho \leq 1$ and set $p = \frac{1}{2}(1-\rho)$. For a function $f : \mathbb{Z}_2 \to \mathbb{C}$, define $T_\rho f : \mathbb{Z}_2 \to \mathbb{C}$ by*

$$T_\rho f(x) = \mathbb{E}_{y \sim \mu_p} f(x + y).$$

Note that $T_\rho f(x) = \mathbb{E}_z [f(z)]$, where z is a noisy copy of x (it is flipped with probability p):

$$z = \begin{cases} x & \text{with probability } 1-p, \\ 1-x & \text{with probability } p. \end{cases} \tag{25}$$

The value of p is chosen so that $\mathbb{E}[(-1)^x(-1)^z] = \rho$. From this one can deduce that for every $f : \mathbb{Z}_2 \to \mathbb{C}$ we have $T_\rho[f] = \widehat{f}(\emptyset) + \rho\widehat{f}(\{1\})\chi_{\{1\}}$ while $f = \widehat{f}(\emptyset) + \widehat{f}(\{1\})\chi_{\{1\}}$.

The operator T_ρ is linear

$$T_\rho(f + \lambda g) = T_\rho f + \lambda T_\rho g,$$

and its corresponding matrix is

$$\begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}. \tag{26}$$

Since T_ρ is an averaging operator it is contractive:

THEOREM 6.1.2 (Contractivity in dimension 1). *For $1 \leq p \leq \infty$, the operator T_ρ is a contractive operator from L_p to L_p. That is,*

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

PROOF. A simple application of Minkowski's Inequality (Theorem 1.3.2) gives the result.

$$\begin{aligned} \|T_\rho f\|_p &= \left(\mathbb{E}_x \left|\mathbb{E}_{y \sim \mu_p} f(x+y)\right|^p\right)^{1/p} \\ &\leq \mathbb{E}_{y \sim \mu_p} \left(\mathbb{E}_x |f(x+y)|^p\right)^{1/p} \\ &= \|f\|_p. \end{aligned}$$

□

The operator T_ρ satisfies a stronger property. Namely it is *hypercontractive*.

THEOREM 6.1.3 (Hypercontractivity - The one-dimensional case). *Let $1 < p \leq q < \infty$. Then for $0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$,*

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

PROOF. Consider $f : \mathbb{Z}_2 \to \mathbb{C}$ and set $\alpha = \frac{1}{2}(1-\rho)$. Then

$$\begin{aligned}
\|T_\rho f\|_q &= \left(\mathbb{E}_x \left|\mathbb{E}_{y\sim\mu_\alpha} f(x+y)\right|^q\right)^{1/q} \\
&= \left(\frac{1}{2}\big((1-\alpha)|f(0)| + \alpha|f(1)|\big)^q + \frac{1}{2}\big(\alpha|f(0)| + (1-\alpha)|f(1)|\big)^q\right)^{1/q} \\
&\leq \left(\frac{1}{2}|f(0)|^p + \frac{1}{2}|f(1)|^p\right)^{1/p} \\
&= \|f\|_p.
\end{aligned}$$

Above, the inequality can be derived using standard methods from calculus. □

Next we will consider the noise operator in the general case. Let μ_p^n denote the product probability measure on $\{0,1\}^n$ corresponding to the Bernoulli measure μ_p. In other words for $y \in \{0,1\}^n$, we have $\mu_p^n(y) = p^{\sum y_i}(1-p)^{n-\sum y_i}$. The noise operator in general is defined as the tensor of the 1-dimensional noise operators: $T_\rho \otimes \ldots \otimes T_\rho$:

DEFINITION 6.1.4 (Noise operator in general). *Let $0 \leq \rho \leq 1$ and set $p = \frac{1}{2}(1-\rho)$. For a function $f : \mathbb{Z}_2^n \to \mathbb{C}$, define $T_\rho f : \mathbb{Z}_2^n \to \mathbb{C}$ by*

$$T_\rho f(x) = \mathbb{E}_{y\sim\mu_p^n} f(x+y).$$

There are several equivalent ways to define the noise operator. First observe that for every $x \in \mathbb{Z}_2^n$, we have

$$\mathbb{E}_{y\sim\mu_p^n}\left[f(x+y)\right] = 2^n \mathbb{E}_y\left[f(x+y)\mu_p^n(y)\right] = 2^n f * \mu_p^n(x),$$

where in the second expected value $y \in \mathbb{Z}_2^n$ is chosen according to the uniform distribution. We can also write $T_\rho f(x) = \mathbb{E}_z\left[f(z)\right]$, where

$$z_i = \begin{cases} x_i & \text{with probability } 1-p, \\ 1-x_i & \text{with probability } p, \end{cases} \tag{27}$$

independently for each i. In other words z is a noisy copy of x (each coordinate is flipped with probability p). Again T_ρ is a linear operator

$$T_\rho(f + \lambda g) = T_\rho f + \lambda T_\rho g,$$

and its corresponding matrix can be obtained by taking the n-th tensor power of the matrix in (26).

Note that T_ρ has a smoothing property. When $\rho = 1$, we have $T_\rho f = f$, but as one decreases ρ, the function $T_\rho f$ "converges" to the constant $\mathbb{E}[f]$ and indeed, for $\rho = 0$, we have $T_\rho f = \mathbb{E}[f]$. Note $T_\rho f(x)$ takes the average of f evaluated at points sampled according to z. When $\rho = 1$, the random variable z is concentrated on point x, and thus the average is just over x so we obtain the original function f. As ρ decreases, the variable z becomes more spread out. Finally $\rho = 0$, we lose the information about x and z is distributed uniformly over all points in $\mathbb{Z}_2^n$. Therefore in this case we get the constant function $\mathbb{E}[f]$. Recall from Lecture 3, when introducing the concept of convolution, we saw that if S is the Hamming ball of radius r around 0 in $\mathbb{Z}_2^n$, then $f * \mathbf{1}_S(x)$ is the average of f over the Hamming ball of radius r around x. The noise operator, which is basically

a convolution itself, has a smoother definition and the Hamming ball is replaced by a distribution centered at x.

Let us now see the effect of the noise operator on the Fourier spectrum.

LEMMA 6.1.5. *If $f : \mathbb{Z}_2^n \to \mathbb{C}$, then*

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

PROOF. Since T_ρ is linear it suffices to show that for every $S \subseteq [n]$, we have

$$T_\rho \chi_S = \rho^{|S|} \chi_S.$$

Indeed we have

$$\begin{aligned} T_\rho \chi_S(x) &= \mathbb{E}_{y \sim \mu_p^n} \chi_S(x+y) = \chi_S(x) \mathbb{E}_{y \sim \mu_p^n} \chi_S(y) = \chi_S(x) \mathbb{E}_{y_i \sim \mu_p} \prod_{i \in S} (-1)^{y_i} \\ &= \chi_S(x) \prod_{i \in S} \mathbb{E}_{y_i \sim \mu_p} (-1)^{y_i} = \chi_S(x) \rho^{|S|}. \end{aligned}$$

□

In other words, the noise operator dampens the high frequency Fourier coefficients, and the dampening effect increases exponentially with the frequency.

In the above proof, we utilized the fact that the noise operator acts on each coordinate independently. In fact, in many results regarding the noise operator we can employ the same trick: analyze the effect of the noise in one coordinate and then use the direct product structure to obtain the desired result. It was for this reason that we first treated the 1-dimensional case separately.

Note that the proof of Theorem 6.1.2 remains valid for the general case.

THEOREM 6.1.6 (Contractivity). *For $1 \leq p \leq \infty$, the operator T_ρ (acting on the space of functions $\mathbb{Z}_2^n \to \mathbb{C}$) is a contractive operator from L_p to L_p. That is,*

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

Next we will show that T_ρ in general is hypercontractive. Before stating this theorem and presenting its proof, we introduce some notation.

As stated before, the direct product structure of $\mathbb{Z}_2^n$ is very useful and is often exploited in proofs. For this reason we introduce some notation for product probability spaces. For a distribution μ over X and a distribution ν over Y, consider the product probability distribution $\mu \times \nu$. Consider $f : (X \times Y, \mu \times \nu) \to \mathbb{C}$. We define $\|f\|_{L_p(\nu)}$ to be the function $x \mapsto \|f_x\|_{L_p(\nu)}$, where $f_x = f(x, \cdot)$. Similarly, define $\|f\|_{L_p(\mu)}$ to be 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 $f : \mathbb{Z}_2^n \to \mathbb{C}$ as a function $f : \mathbb{Z}_2^S \times \mathbb{Z}_2^{\bar{S}} \to \mathbb{C}$. Then it is straightforward to verify,

$$\|f\|_q = \left\| \|f\|_{L_q(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})}. \tag{28}$$

It can be instructive to see how $\left\| \|f\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})}$ expands out:

$$\left\| \|f\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} = \left(\mathbb{E}_{y \in \mathbb{Z}_2^{\bar{S}}} \left| \|f_y\|_{L_p(\mathbb{Z}_2^S)} \right|^q \right)^{1/q} = \left(\mathbb{E}_{y \in \mathbb{Z}_2^{\bar{S}}} \left| \left(\mathbb{E}_{x \in \mathbb{Z}_2^S} |f_y(x)|^p \right)^{1/p} \right|^q \right)^{1/q}. \tag{29}$$

Equation (28) follows immediately as $f_y(x) = f(x, y)$.

As in the proof of Theorem 6.1.2, a simple application of Minkowski's Inequality gives

$$\left\|\|f\|_{L_1(\mu)}\right\|_{L_p(\nu)} \leq \left\|\|f\|_{L_p(\nu)}\right\|_{L_1(\mu)}.$$

Indeed,

$$\left\|\|f\|_{L_1(\mu)}\right\|_{L_p(\nu)} = \|\mathbb{E}_{x\sim\mu}|f(x,\cdot)|\|_{L_p(\nu)} \leq \mathbb{E}_{x\sim\mu}\| \, |f(x,\cdot)| \, \|_{L_p(\nu)} = \left\|\|f\|_{L_p(\nu)}\right\|_{L_1(\mu)}.$$

This is in fact a special case of a more general inequality:

THEOREM 6.1.7 (Generalized Minkowski's Inequality). *For $1 \leq p \leq q \leq \infty$, we have*

$$\left\|\|f\|_{L_p(\nu)}\right\|_{L_q(\mu)} \leq \left\|\|f\|_{L_q(\mu)}\right\|_{L_p(\nu)}.$$

Now we have all the tools we need to prove the Bonami-Beckner inequality. Recall that L_p norms are increasing on probability space. That is for $1 \leq p \leq q \leq \infty$ we have $\|f\|_p \leq \|f\|_q$. The Bonami-Beckner inequality says that if we sufficiently smooth f by applying the operator T_ρ, we can reverse the direction of this inequality.

THEOREM 6.1.8 (Hypercontractivity - Bonami 1970, Beckner 1975, Nelson 1973, Gross 1975). *Let $1 < p \leq q < \infty$. Then for $0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$,*

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

PROOF. The proof is by induction on n. We have already verified the inequality for $n = 1$ in Theorem 6.1.3. Next we exploit the direct product structure to prove it for all n.

Consider $f : \mathbb{Z}_2^n \to \mathbb{C}$. For $S \subseteq [n]$, let T_ρ^S denote the noise operator applied to the coordinates in S. That is, it 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 $S = \{1\}$. In light of Equation (29), we have

$$\begin{aligned}
\|T_\rho f\|_q &= \|T_\rho^S T_\rho^{\bar{S}} f\|_q \\
&= \left\| \|T_\rho^S T_\rho^{\bar{S}} f\|_{L_q(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} && \text{(Equation (28))} \\
&\leq \left\| \|T_\rho^{\bar{S}} f\|_{L_p(\mathbb{Z}_2^S)} \right\|_{L_q(\mathbb{Z}_2^{\bar{S}})} && \text{(Induction Hypothesis)} \\
&\leq \left\| \|T_\rho^{\bar{S}} f\|_{L_q(\mathbb{Z}_2^{\bar{S}})} \right\|_{L_p(\mathbb{Z}_2^S)} && \text{(Generalized Minkowski)} \\
&\leq \left\| \|f\|_{L_p(\mathbb{Z}_2^{\bar{S}})} \right\|_{L_p(\mathbb{Z}_2^S)} && \text{(Induction Hypothesis)} \\
&= \|f\|_p && \text{(Equation (28)).}
\end{aligned}$$

□

Next we state a very useful corollary of the Bonami-Beckner inequality.

COROLLARY 6.1.9. *Let $f : \mathbb{Z}_2^n \to \mathbb{C}$ be a function and $k > 0$ be an integer. Then for $1 < p \leq 2$*

$$\|f^{\leq k}\|_2 \leq \left(\frac{1}{\sqrt{p-1}}\right)^k \|f\|_p,$$

and for $2 \leq q < \infty$,

$$\|f^{\leq k}\|_q \leq \left(\sqrt{q-1}\right)^k \|f\|_2.$$

PROOF. In the case of $1 < p \leq 2$, we can apply the Bonami-Beckner inequality with $\rho = \sqrt{p-1}$ and get

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

Observe that

$$\|T_\rho f\|_2^2 = \sum_S \rho^{2|S|} |\widehat{f}(S)|^2 \geq \rho^{2k} \sum_{S:|S|\leq k} |\widehat{f}(S)|^2 = \rho^{2k} \|f^{\leq k}\|_2^2.$$

Therefore

$$\|f^{\leq k}\|_2 \leq \frac{1}{\rho^k} \|f\|_p = \left(\frac{1}{\sqrt{p-1}}\right)^k \|f\|_p.$$

Case $q \geq 2$ follows by duality. Let p satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Note that $1 < p \leq 2$ so we can apply the first part using the L_p norm. Since L_p and L_q are dual norms we have (see question 2 of assignment 1):

$$\|f^{\leq k}\|_q = \sup_{g \neq 0} \frac{\langle f^{\leq k}, g \rangle}{\|g\|_p} \leq \sup_{g \neq 0} \frac{\langle f^{\leq k}, g \rangle}{(\sqrt{p-1})^k \|g^{\leq k}\|_2} = (\sqrt{q-1})^k \sup_{g \neq 0} \frac{\langle f^{\leq k}, g^{\leq k} \rangle}{\|g^{\leq k}\|_2}.$$

Since the dual of the L_2 norm is the L_2 norm itself,

$$(\sqrt{q-1})^k \sup_{g \neq 0} \frac{\langle f^{\leq k}, g^{\leq k} \rangle}{\|g^{\leq k}\|_2} = (\sqrt{q-1})^k \|f^{\leq k}\|_2 \leq (\sqrt{q-1})^k \|f\|_2.$$

□

EXERCISE 6.1.10. *Prove the $q \geq 2$ case of Corollary 6.1.9 by applying the Bonami-Beckner inequality to $g = \sum_S \rho^{-|S|} \widehat{f}(S) \chi_S$.*

Recall that the L_p norms are increasing, that is, $\|f\|_p \leq \|f\|_q$ when $1 \leq p \leq q \leq \infty$. An immediate consequence of Corollary 6.1.9 is that if $\deg(f) \leq k$, then for $1 < p \leq 2$,

$$\|f\|_p \leq \|f\|_2 \leq \left(\frac{1}{\sqrt{p-1}}\right)^k \|f\|_p,$$

and for $2 \leq q < \infty$,

$$\|f\|_2 \leq \|f\|_q \leq \left(\sqrt{q-1}\right)^k \|f\|_2.$$

REMARK 6.1.11. The above inequalities show that the function is "flat". If there are large fluctuations in f, then we cannot hope to have such strong equivalences between the different norms. In this sense, one can think of the Bonami-Beckner inequality as a concentration inequality. Indeed, viewing f as a random variable, by bounding the q-norms in terms of the 2-norm, we are essentially bounding the moments of f in terms of the standard deviation of f. ■

REMARK 6.1.12. The case of $\deg(f) = 1$ is known as Khintchine inequality: for $a_1, a_2, \ldots, a_n \in \mathbb{C}$,

$$\sqrt{p-1} \left(\sum_i |a_i|^2\right)^{1/2} \leq \left(\mathbb{E}\left|\sum_i \epsilon_i a_i\right|^p\right)^{1/p} \leq \left(\sum_i |a_i|^2\right)^{1/2},$$

where the expectation is over $\{\epsilon_i\}$ which are ± 1 valued i.i.d. random variables with $\Pr[\epsilon_i = 1] = 1/2$. By setting $f = \sum_i a_i \chi_{\{i\}}$, we see the correspondence immediately. ■

6.2. Influence and Friedgut's Theorem

The Bonami-Beckner inequality is a powerful tool in the analysis of Boolean functions. Recall that in Definition 4.3.1, we defined the *influence* of the ith variable on f is the probability that changing the ith coordinate changes the the value of f. That is,

$$I_i(f) = \Pr[f(x) \neq f(x + e_i)],$$

where $x \in \{0,1\}^n$ is sampled uniformly and e_i is the i-th standard vector, and the total influence of f is defined as

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

REMARK 6.2.1. Considering the support of f, $\mathrm{Supp}(f) = \{x : f(x) \neq 0\}$, as a subset of the hypercube $\mathcal{Q}_n$, I_f corresponds to the *edge boundary* of $\mathrm{Supp}(f)$. For a subset S of the hypercube, the edge boundary of S, denoted ∂S, is the set of edges of $\mathcal{Q}_n$ with one end point in S and the other endpoint outside of S. It follows by definition that

$$I_f = \frac{2|\partial \mathrm{Supp}(f)|}{2^n}.$$

■

When studying influences, it is natural to consider $f_{(i)} : \mathbb{Z}_2^n \to \{-1, 0, 1\}$ (sometimes referred to as the ith derivative of f), which is defined as

$$f_{(i)}(x) = f(x) - f(x + e_i).$$

Indeed, since $|f_{(i)}(x)| = |f_{(i)}(x)|^2$, we have

$$I_i(f) = \mathbb{E}_x |f_{(i)}(x)| = \mathbb{E}_x |f_{(i)}(x)|^2 = \|f_{(i)}\|_2^2.$$

The Fourier expansion of $f_{(i)}$ is

$$f_{(i)}(x) = \sum_S \widehat{f}(S)\chi_S(x + e_i) - \widehat{f}(S)\chi_S(x) = 2 \sum_{S: i \in S} \widehat{f}(S)\chi_S(x),$$

and therefore

$$I_i(f) = 4 \sum_{S: i \in S} |\widehat{f}(S)|^2.$$

We can also get a nice expression for the total influence of f in terms of its Fourier coefficients:

$$I_f = \sum_i 4 \sum_{S: i \in S} |\widehat{f}(S)|^2 = 4 \sum_S |S| |\widehat{f}(S)|^2.$$

With this, we can get a simple lower bound for the total influence in terms of the variance of f. Note that $\mathrm{Var}(f) = \mathbb{E}[f^2] - (\mathbb{E}[f])^2 = \sum_{S: S \neq \emptyset} |\widehat{f}(S)|^2$ and therefore

$$I_f \geq 4\mathrm{Var}(f). \tag{30}$$

This bound in general can be quite weak. A stronger bound can be obtained by a discrete isoperimetric inequality.

THEOREM 6.2.2 (Edge Isoperimetric Inequality). *For S a subset of the vertices of the hypercube $\mathcal{Q}_n$ we have*

$$|\partial S| \geq -|S| \log_2 \frac{|S|}{2^n}.$$

Equality is achieved when S is a subcube.

PROOF. The proof is quite straightforward using induction on n. The base case, $n = 1$, is easily verified so we directly move to the induction step. Partition $\mathcal{Q}_n$ into two disjoint subcubes $\mathcal{Q}^1_{n-1}$ and $\mathcal{Q}^2_{n-1}$ of dimension $n-1$ each. Similarly partition S into two sets $S_1 = S \cap V(\mathcal{Q}^1_{n-1})$ and $S_2 = S \cap V(\mathcal{Q}^2_{n-1})$. Without loss of generality assume $|S_1| = |S_2| + t$. Now the boundary of S will have edges from the boundary of S_1 in $\mathcal{Q}^1_{n-1}$, edges from the boundary of S_2 in $\mathcal{Q}^2_{n-1}$, and also at least t edges that must go between the two subcubes. Using the induction hypothesis, we have

$$\begin{aligned}|\partial S| &\geq |S_1|(n-1-\log|S_1|) + |S_2|(n-1-\log|S_2|) + t \\ &= |S_1|n - |S_1| - |S_1|\log|S_1| + |S_2|n - |S_2| - |S_2|\log|S_2| + t \\ &= |S_1|n - |S_1|\log|S_1| + |S_2|n - |S_2|\log|S_2| - 2|S_2|.\end{aligned}$$

Note that $|S|(n - \log|S|) = |S_1|n + |S_2|n - (|S_1| + |S_2|)\log(|S_1| + |S_2|)$, so we are done provided

$$|S_1|\log|S_1| + |S_2|\log|S_2| + 2|S_2| \leq (|S_1| + |S_2|)\log(|S_1| + |S_2|).$$

This inequality is easily derived using simple manipulations. □

Defining f such that $\mathrm{Supp}(f) = S$, we can rewrite the Edge Isoperimetric Inequality in terms of the total influence:

$$I_f \geq -\mathbb{E}[f]\log_2 \mathbb{E}[f]. \tag{31}$$

Consider a balanced function $f : \mathbb{Z}_2^n \to \{0,1\}$, i.e. $\mathbb{E}[f] = 1/2$. Both lower bounds (30) and (31) on I_f imply that $I_f \geq 1/2$, which shows

$$\max_i I_i(f) \geq \frac{1}{2n}.$$

Note that the lower bound $I_f \geq 1/2$ is tight for half-cubes, i.e. for $f(x) = x_i$ or $f(x) = -x_i$ for some i. Two questions naturally arise:

(1) (Ben-Or Linial) How small can $\max_i I_i(f)$ be for balanced functions?
(2) What are the functions with small total influence?

We first give an answer to the second question. For this we need to define a *junta*.

DEFINITION 6.2.3. *A Boolean function $f : \{0,1\}^n \to \{0,1\}$ is called a k-junta if there exists $J \subseteq [n]$ of size at most k and $g : \{0,1\}^J \to \{0,1\}$ such that $f(x) = g(x_J)$. In other words, f is a k-junta if its output only depends on at most k input coordinates.*

Observe that if f is a k-junta then $I_f \leq k/2$. This is because every variable that f does not depend on has influence 0, and every other variable has influence at most $1/2$. Friedgut's Theorem gives a partial converse to this observation and states that a Boolean function with small total influence is well approximated by a k-junta with a small k.

THEOREM 6.2.4 (Friedgut). *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. Then there exists a $2^{O(I_f/\epsilon)}$-junta $g : \{0,1\}^n \to \{0,1\}$ such that*

$$\Pr[f(x) \neq g(x)] \leq \epsilon.$$

PROOF. First note that the probabilistic approximation can be interpreted in terms of the L_2 difference:

$$\|f - g\|_2^2 = \Pr[f(x) \neq g(x)].$$

Let J be the set of most influential variables of f, that is, $J = \{i \in [n] \mid I_i(f) \geq \delta\}$ for some parameter δ to be determined later. It is natural to try to find a g that depends only on the variables in J. Define h to be

$$h = \sum_{S \subseteq J} \widehat{f}(S)\chi_S.$$

Clearly h depends only on the variables in J, but it is not necessarily a Boolean function. Nevertheless we can round h to make it Boolean. Let $g(x) = 1$ if $h(x) > 1/2$ and let $g(x) = 0$ if $h(x) \leq 1/2$. By rounding we haven't lost much in the following sense. If $\|f - h\|_2^2 \leq \epsilon$, then $\|f - g\|_2^2 \leq 4\epsilon$. This is easy to see since for any x with $f(x) \neq g(x)$, $|f(x) - h(x)|^2 \geq 1/4$.

Thus our task reduces to showing that

$$\|f - h\|_2^2 \leq \frac{\epsilon}{4}.$$

By Parseval we have

$$\|f - h\|_2^2 = \sum_S (\widehat{f}(S) - \widehat{h}(S))^2 = \sum_{S \not\subseteq J} \widehat{f}(S)^2.$$

So we want to upper bound the ℓ_2 mass of $\widehat{f}$ on sets S with $S \not\subseteq J$. To do this we will divide the above sum into two parts, the low degree part and the high degree part, and deal with them separately.

Intuitively, a function with small total influence should not have large ℓ_2 mass on high degree characters as high degree characters, viewed as 0/1 valued functions, have large total influence. This intuition is easy to formalize. Set $k = 2I_f/\epsilon$. Then

$$I_f = 4\sum_S |S||\widehat{f}(S)|^2 \geq 4k \sum_{|S| \geq k} |\widehat{f}(S)|^2,$$

which implies

$$\sum_{S:|S| \geq k} |\widehat{f}(S)|^2 \leq \frac{I_f}{4k} \leq \frac{\epsilon}{2}.$$

Thus,

$$\|f - h\|_2^2 \leq \frac{\epsilon}{8} + \sum_{\substack{S:|S|<k \\ S \not\subseteq J}} \widehat{f}(S)^2.$$

Now to bound the low degree part we will use Bonami-Beckner inequality (the form given in Corollary 6.1.9). First observe that

$$\sum_{\substack{S:|S|<k \\ S \not\subseteq J}} \widehat{f}(S)^2 = \sum_{i \notin J} \sum_{\substack{S:|S|<k \\ i \in S}} \widehat{f}(S)^2. \tag{32}$$

We want to bound the inside sum on the RHS above. Recall that

$$f_{(i)}(x) = f(x) - f(x + e_i) = 2\sum_{i \in S} \widehat{f}(S)\chi_S(x),$$

So the quantity we want to bound is $\|f_{(i)}^{<k}\|_2^2$. We apply Corollary 6.1.9 with $p = 4/3$ to get

$$\|f_{(i)}^{<k}\|_2 \leq 3^{k/2}\|f_{(i)}\|_{4/3},$$

and so

$$\|f_{(i)}^{<k}\|_2^2 \leq 3^k \left(\mathbb{E}_x |f_{(i)}(x)|^{4/3}\right)^{3/2} = 3^k \left(\mathbb{E}_x |f_{(i)}(x)|^2\right)^{3/2}.$$

Recall that $I_i(f) = \|f_{(i)}\|_2^2$. Also, since $i \notin J$, $I_i(f) < \delta$. Thus,

$$\|f_{(i)}^{<k}\|_2^2 \leq 3^k I_i(f)^{3/2} \leq 3^k \delta^{1/2} I_i(f).$$

Equivalently,

$$4 \sum_{\substack{S:|S|<k \\ i \in S}} \widehat{f}(S)^2 \leq 3^k \delta^{1/2} I_i(f).$$

Going back to (32), we have

$$\sum_{i \notin J} \sum_{\substack{S:|S|<k \\ i \in S}} \widehat{f}(S)^2 \leq \sum_{i \notin J} 3^k \delta^{1/2} I_i(f) \leq 3^k \delta^{1/2} I_f.$$

Putting things together

$$\|f - h\|_2^2 \leq \frac{\epsilon}{8} + 3^k \delta^{1/2} I_f \leq \frac{\epsilon}{8} + \frac{\epsilon}{8} = \frac{\epsilon}{4},$$

when δ is set to be sufficiently small. Recall that we set $k = 2I_f/\epsilon$, i.e. $I_f = k\epsilon/2$. Now a simple calculation shows that we can set $\delta = 1/3^{3k+2}$. With this δ, we have

$$|J| \leq \frac{I_f}{\delta} \leq 2^{O(I_f/\epsilon)},$$

as required.

□

6.3. Kahn-Kalai-Linial Theorem

In this lecture we are going to prove the Kahn-Kalai-Linial (KKL) Theorem that says that every balanced function has an influential variable, that is, there is some $i \in [n]$ such that $I_i(f) = \Omega(\frac{\log n}{n})$. The proof is essentially the same as Friedgut's Theorem[1]. We separate the Fourier spectrum of f into high degree and low degree parts. The high degree part is easy to handle and for the low degree part we apply the Bonami-Beckner inequality. The reason why Bonami-Beckner inequality is effective can be seen as follows. For $1 \leq p < 2$, when g is a Boolean function, we have $\mathbb{E}[|g|] = \mathbb{E}[|g|^p] = \mathbb{E}[|g|^2]$, which implies that $\|g\|_p = \|g\|_2^{2/p}$. Now if $\|g\|_2 =: \delta$ is small, then $\|g\|_p = \delta \cdot \delta^{(2/p-1)}$ is very small. So applying Corollary 6.1.9 to g, we get a good bound on $\|g^{<k}\|_2$ and gain a factor of $\delta^{2/p-1}$.

THEOREM 6.3.1 (Kahn-Kalai-Linial). *Let $f : \mathbb{Z}_2^n \to \{0,1\}$ be such that $\mathbb{E}[f] = \alpha$. If $\delta = \max_i I_i(f)$, then*

$$I_f \geq \Omega\left(\alpha(1-\alpha)\log 1/\delta\right)$$

In particular

$$\delta \geq \Omega\left(\alpha(1-\alpha)\frac{\log n}{n}\right).$$

PROOF. Recall that $\mathrm{Var}(f) = \mathbb{E}[f^2] - (\mathbb{E}[f])^2 = \alpha - \alpha^2 = \alpha(1-\alpha)$. Also since $\mathbb{E}[f]^2 = \widehat{f}(\emptyset)^2$,

$$\mathrm{Var}(f) = \sum_{S:|S|\geq 1} |\widehat{f}(S)|^2.$$

[1]Historically the KKL Theorem came before Friedgut's Theorem.

In (30) we observed that $\mathrm{Var}(f) \leq \frac{1}{4} I_f$ and that this leads to the bound $\delta \geq \frac{1}{n}$ for balanced functions. Our goal now is to obtain a better upper bound on the variance, which will lead to a better lower bound on δ. In particular we are aiming for the upper bound

$$\mathrm{Var}(f) = \sum_{S:|S|\geq 1} |\widehat{f}(S)|^2 \lesssim \frac{I_f}{\log 1/\delta}.$$

Our strategy will be as in the proof of Friedgut's Theorem. We divide the sum into the low degree and high degree parts, and upper bound each part separately.

Recall that

$$I_f = 4 \sum_S |S| |\widehat{f}(S)|^2 \geq 4k \sum_{|S|>k} |\widehat{f}(S)|^2.$$

This implies

$$\sum_{|S|>k} |\widehat{f}(S)|^2 \leq \frac{I_f}{4k}.$$

Setting $k \approx \log 1/\delta$, the upper bound above is what we want for $\mathrm{Var}(f)$. So with this choice of k, we would like to show an upper bound on the low degree part that is negligible compared to $I_f/4k$.

To handle the low degree part, we will apply Bonami-Beckner inequality to $\|f_{(i)}\|_2$ with $p = 3/2$:

$$\sum_{1\leq|S|\leq k} |\widehat{f}(S)|^2 \leq \sum_{i=1}^n \sum_{\substack{i\in S \\ |S|\leq k}} |\widehat{f}(S)|^2 = \frac{1}{4} \sum_{i=1}^n \|f_{(i)}^{\leq k}\|_2^2 \leq \frac{1}{4} \sum_{i=1}^n 2^k \|f_{(i)}\|_{3/2}^2.$$

Using the fact that $|f_{(i)}(x)| \in \{0, 1\}$, we have

$$\frac{1}{4} \sum_{i=1}^n 2^k \|f_{(i)}\|_{3/2}^2 = \frac{1}{4} 2^k \sum_{i=1}^n \|f_{(i)}\|_2^{8/3} = \frac{1}{4} 2^k \sum_{i=1}^n I_i(f)^{4/3} \leq 2^k \delta^{1/3} \sum_{i=1}^n I_i(f) = 2^k \delta^{1/3} I_f.$$

Putting things together we get

$$\alpha(1-\alpha) = \sum_{S:|S|\geq 1} |\widehat{f}(S)|^2 \leq \frac{I_f}{2k} + 2^k \delta^{1/3} I_f.$$

Setting $k = \frac{1}{10} \log 1/\delta$ shows

$$\frac{1}{10} \alpha(1-\alpha) \log 1/\delta \leq I_f.$$

We also know that $I_f \leq \delta n$. These upper and lower bounds on I_f imply by a straightforward calculation that

$$\delta \geq \Omega\left(\alpha(1-\alpha) \frac{\log n}{n}\right).$$

□

The KKL Theorem is tight, which can be seen by considering the *tribes* function. Let

$$f(x) = \bigvee_{i=1}^m \bigwedge_{j=1}^k x_{ij},$$

where $k = \log n - \log\ln n$ and $m = n/k$. Without loss of generality consider the first variable. For x_1 to be able to change the output, all other variables in the first clause must be set to 1, and all other clauses must be evaluating to 0. Thus,

$$\begin{aligned} I_1(f) &= \Pr[f(x) \neq f(x+e_i)] = (1-2^{-k})^{m-1} \cdot 2^{-k+1} \\ &= 2^{1-k}(1-2^{-k})^{m-1} = \frac{2\ln n}{n}(1-\frac{\ln n}{n})^{m-1} = \frac{2\ln n}{n}(1-o(1)). \end{aligned}$$

Now we will see some corollaries to KKL Theorem and some related conjectures.

COROLLARY 6.3.2. *If a balanced function $f : \mathbb{Z}_2^n \to \{0,1\}$ satisfies $I_1(f) = I_2(f) = \cdots = I_n(f)$ (e.g. f is invariant under certain symmetries), then $I_f \gtrsim \log n$.*

Bourgain and Kalai show that under strong symmetry assumptions, the above bound can be improved significantly. For instance if f is a symmetric function, i.e. f's output only depends on the Hamming weight of the input, then $I_f \gtrsim \sqrt{n}$.

A Boolean function $f : \{0,1\}^n \to \{0,1\}$ is called *increasing* (or *monotone*) if $f(x) \leq f(y)$ whenever $x_i \leq y_i$ for all i.

COROLLARY 6.3.3. *Let $f : \{0,1\}^n \to \{0,1\}$ increasing balanced function. Then there is a set $J \subseteq [n]$ of size $O_\epsilon(\frac{n}{\log n})$ such that*

$$\mathbb{E}\left[f(x)|x_J = \vec{1}\right] \geq 1-\epsilon,$$

and

$$\mathbb{E}\left[f(x)|x_J = \vec{0}\right] \leq \epsilon,$$

PROOF SKETCH. Let $i \in [n]$ have the highest influence. Then setting $x_i = 1$ will increase the average of f by at least $\Omega(\frac{\log n}{n})$. Repeat with the new function to obtain a set J_1 of size $O_\epsilon(\frac{n}{\log n})$ with

$$\mathbb{E}\left[f(x)|x_{J_1} = \vec{1}\right] \geq 1-\epsilon.$$

Repeating the same process but setting the variables to 0 leads to another set J_2 of size $O_\epsilon(\frac{n}{\log n})$ with

$$\mathbb{E}\left[f(x)|x_{J_2} = \vec{0}\right] \leq \epsilon.$$

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 6.3.3 cannot be improved significantly in any direction.

THEOREM 6.3.4 ([**AL93**]). *There exists a balanced and increasing 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*

$$\mathbb{E}\left[f(x)|x_J = \vec{1}\right] - \mathbb{E}[f] = o(1),$$

and

$$\mathbb{E}\left[f(x)|x_J = \vec{0}\right] - \mathbb{E}[f] = o(1).$$

PROBLEM 6.3.5 (Open Problem). *It is believed that in Theorem 6.3.4, the bound $o(\frac{n}{\log^2(n)})$ can be improved to close to $o(\frac{n}{\log(n)})$ matching the bound in Corollary 6.3.3*

CONJECTURE 6.3.6 (Freidgut). *Let $f : [0,1]^n \to \{0,1\}$ be an increasing function. Then there exists a subset $J \subseteq [n]$ with $|J| = o_\epsilon(n)$ such that*

$$\mathbb{E}\left[f(x)|x_J = \vec{0}\right] \leq \epsilon \quad \text{or} \quad \mathbb{E}\left[f(x)|x_J = \vec{1}\right] \geq 1 - \epsilon.$$

CONJECTURE 6.3.7 (Freidgut). *Suppose $f : \{0,1\}^n \to \{0,1\}$ is increasing and $\max_i I_i(f) \leq \frac{c \log n}{n}$, for some constant c. Then there is $J \subseteq [n]$ of size $O_{\epsilon,c}(\log n)$ such that*

$$\mathbb{E}\left[f(x)|x_J = \vec{0}\right] \leq \epsilon \quad \text{or} \quad \mathbb{E}\left[f(x)|x_J = \vec{1}\right] \geq 1 - \epsilon.$$

We can make a similar conjecture for non-monotone functions.

CONJECTURE 6.3.8. *Suppose $f : \{0,1\}^n \to \{0,1\}$ satisfies $\max_i I_i(f) \leq \frac{c \log n}{n}$, for some constant c. Then there is $J \subseteq [n]$ of size $O_{\epsilon,c}(\log n)$ and $y \in \{0,1\}^J$ such that*

$$\mathbb{E}\left[f(x)|x_J = y\right] \leq \epsilon \quad \text{or} \quad \mathbb{E}\left[f(x)|x_J = y\right] \geq 1 - \epsilon.$$

The influences of increasing Boolean functions have a very special and useful characterization in terms of f's Fourier coefficients. It is not hard to verify that

$$I_i(f) = \Pr[f(x) \neq f(x + e_i)] = -\mathbb{E} f(x)\chi_{\{i\}}(x) = -\widehat{f}(\{i\}).$$

Using this and the Cauchy-Schwarz inequality, it is easy to get an upper bound on the total influence of increasing functions:

$$I_f = \sum_i |\widehat{f}(\{i\})| \leq \sqrt{n} \sum_i \left(|\widehat{f}(\{i\})|^2\right)^{1/2} \leq \sqrt{n}.$$

Note that for non-monotone functions we can have $I_f = n$ (e.g. $f = \text{PARITY}$). The above bound is tight since $I_{\text{MAJ}} = \Theta(\sqrt{n})$, where MAJ denotes the majority function:

$$\text{MAJ}(x) := \begin{cases} 1 & \text{if } \sum_i x_i \geq n/2, \\ 0 & \text{otherwise.} \end{cases}$$

CHAPTER 7

The Semigroup method

In Chapter 6 we introduced the noise operator T_ρ, and studied some of its useful properties. In this chapter we take a more general approach, and study the noise operator as an instance of a big class of operators. This point of view will also shed some light on the definition of the noise operator.

These general classes of operators are defined through random walks. They are parametrized by time $t \in [0, \infty)$, and defined in the following way. 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 if it is started at point a.

We start this chapter by studying the simple discrete random walk on the cube.

7.1. The Poisson random walk on the cube

Consider the n-dimensional hypercube with vertex set $\mathbb{Z}_2^n$, where two vertices are neighbours if and only if they differ in one coordinate. Let us examine the standard discrete random walk on this graph started at a vertex $a \in \mathbb{Z}_2^n$:

- $Y^a(0) = a$ is the starting point.
- At time $t \in \mathbb{N}$, we choose $Y^a(t)$ from the n neighbors of $Y^a(t-1)$ uniformly at random.

Let f_t be the distribution of Y_t, i.e. $f_t(x) = \Pr[Y^a(t) = x]$. Note that

$$f_0(x) = \mathbf{1}_a(x) = 2^{-n} \prod_{i=1}^{n} (1 - (-1)^{a_i}(-1)^{x_i}) = 2^{-n} \sum_{S \subseteq [n]} \chi_S(a)\chi_S(x).$$

Define the operator $K : L_2(\mathbb{Z}_2^n) \to L_2(\mathbb{Z}_2^n)$ as

$$Kf(x) = \frac{1}{n} \sum_{i=1}^{n} f(x + e_i), \tag{33}$$

so that $f_t = Kf_{t-1} = K^2 f_{t-2} = \ldots = K^t f_0$ for every integer $t > 1$. Note that for every character χ_S, we have

$$K\chi_S = \frac{1}{n}\left((n - |S|)\chi_S - |S|\chi_S\right) = \left(1 - \frac{2|S|}{n}\right)\chi_S.$$

This shows that χ_S are eigenvectors of the operator K with corresponding eigenvalues $(1 - 2|S|/n)$. It follows that

$$f_t = 2^{-n} \sum_{S \subseteq [n]} \left(1 - \frac{2|S|}{n}\right)^t \chi_S(a)\chi_S,$$

and since $|1 - 2|S|/n| < 1 - 2/n$ for $S \neq \emptyset, [n]$, we have

$$\left\| f_t - 2^{-n}\left(\chi_\emptyset + (-1)^t \chi_{[n]}\right)\right\|_\infty \leq (1 - 2/n)^t \leq e^{-2t/n}.$$

So as $t \to \infty$, we obtain an exponentially fast convergence of the form

$$f_{2t} \to 2^{-n}\left(\chi_\emptyset + \chi_{[n]}\right) = 2^{1-n}\mathbf{1}_{[\sum x_i \equiv_2 0]},$$

and similarly

$$f_{2t+1} \to 2^{-n}\left(\chi_\emptyset - \chi_{[n]}\right) = 2^{1-n}\mathbf{1}_{[\sum x_i \equiv_2 1]}.$$

In other words, on even times, this random walk quickly converges to the uniform measure on points with even parity, and on odd times it converges to the uniform distribution on the points with odd parity. This means that the random walk is not fully ergodic (i.e. it does not converge to the uniform measure), as the bipartite structure of the cube prevents it from being so.

To obtain an ergodic random walk, we modify the transitions slightly by making them "lazy". We choose a parameter $\lambda(0, \frac{1}{2})$ and define the random walk $Z(t) := Z^{a,\lambda}(t)$ as

- $Z(0) = a$ is the starting point.
- At time $t \in \mathbb{N}$, with probability $1 - \lambda$ we set $Z(t) = Z(t-1)$, and with probability λ, we choose $Z(t)$ from the n neighbors of $Z(t-1)$ uniformly at random.

Now if we again denote the distribution at time t by f_t, then $f_t = K_\lambda f_{t-1}$, where now $K_\lambda = (1-\lambda)\mathrm{Id} + \lambda K$ with K is defined in (33). Hence for a character χ_S,

$$K_\lambda \chi_S = (1-\lambda)\chi_S + \lambda\left(1 - \frac{2|S|}{n}\right)\chi_S = \left(1 - \frac{2\lambda|S|}{n}\right)\chi_S,$$

and consequently

$$f_t = K_\lambda^t f_0 = 2^{-n}\sum_{S\subseteq[n]}\left(1 - \frac{2|S|\lambda}{n}\right)^t \chi_S(a)\chi_S.$$

Note that $1 - 2\lambda|S|/n < 1$ unless $S = \emptyset$, and this time f_t will converge to the uniform measure on the cube as t tends to infinity. Here we obtained a fully ergodic random walk by using the laziness to destroy the periodicity of the original random walk.

By the law of large numbers, when t is large, we make a move at roughly λ fraction of time steps. Hence it is more natural to consider a different rescaling of time and study $f_{\lfloor nt/\lambda \rfloor}$. That is now we are considering n epochs, each consisting of $1/\lambda$ steps, and on average we expect to make one move in every epoch. By tending λ to 0 we obtain a "continuous" version of the walk in the limit. 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

$$X^a(t) \sim 2^{-n}\sum_{S\subseteq[n]} \chi_S(a)e^{-t|S|}\chi_S. \tag{34}$$

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

$$\lim_{\lambda\to 0} \mu_{\lfloor nt/2\lambda \rfloor} = \sum_{S\subseteq[n]} e^{-t|S|}\widehat{\mu}(S)\chi_S. \tag{35}$$

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)

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

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 $\Pr(E \leq s+t|E > s) = \Pr(E \leq t)$. 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 $\text{Pois}(\lambda)$, is the probability distribution on $\{0,1,2,\ldots\}$ defined by

$$\Pr[\{k\}] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

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 7.1.1. *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 $0 \leq s \leq t$, we have $N(t) - N(s) \sim N(t-s) \sim \text{Pois}(t-s)$.*

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 $\mathbb{Z}_2$. 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 7.1.2. *Show that the process $(M(t))_{t\in[0,\infty)}$ defined above is a Markov process.*

EXERCISE 7.1.3. *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 7.1.4. *For $0 \leq s < t$, we have*

$$\Pr[M(t) = 0|M(s) = 0] = \Pr[M(t) = 1|M(s) = 1] = \frac{(1 + e^{-2(t-s)})}{2},$$

and

$$\Pr[M(t) = 0|M(s) = 1] = \Pr[M(t) = 1|M(s) = 0] = \frac{(1 - e^{-2(t-s)})}{2}.$$

PROOF. We have

$$\begin{aligned}\Pr[M(t) = 0|M(s) = 0] &= \Pr[N(t) \equiv_2 0|N(s) \equiv_2 0] = \Pr[N(t) - N(s) \equiv_2 0|N(s) \equiv_2 0] \\ &= \Pr[N(t-s) \equiv_2 0] = \mathrm{Pois}_{t-s}(\{0, 2, 4, \ldots\}) = \frac{(1 + e^{-2(t-s)})}{2}.\end{aligned}$$

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

$$\Pr[X(t) = 0|X(s) = 0] = \Pr[X(t) = 1|X(s) = 1] = \frac{(1 + e^{-(t-s)})}{2}, \tag{36}$$

and

$$\Pr[X(t) = 0|X(s) = 1] = \Pr[X(t) = 1|X(s) = 0] = \frac{(1 - e^{-(t-s)})}{2}. \tag{37}$$

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 (36) and (37) 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 $\Pr[X(t) = y|X(s) = x]$ according to (36) and (37), we need to verify that for $0 \leq s < t < u$, and $z, x \in \{0, 1\}$, we have

$$p_{s,u}(x, z) = \sum_{y\in\{0,1\}} p_{s,t}(x, y)p_{t,u}(y, z). \tag{38}$$

Then Kolmogorov's extension theorem guarantees that there is a Markov process $(X(t))_{t\in[0,\infty)}$ satisfying the transition inequalities (36) and (37).

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 $\mathbb{Z}_2$. 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 $1 \leq i \leq n$), and so on.

Denoting by f_t the distribution of $X^a(t)$, by (36) and (37), 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 (34) as the limit of the discrete lazy walks with proper rescaling of time. As we will formally see in Section 7.2, 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 7.1.5. [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

$$\Pr[\text{Nothing is produced}] = (1-p\lambda)(1-(1-p)\lambda) = 1-\lambda+p(1-p)\lambda^2,$$

and

$$\Pr[\text{A Head is produced}] = p\lambda(1-(1-p)\lambda) = p\lambda - p(1-p)\lambda^2,$$

and

$$\Pr[\text{A Tail is produced}] = (1-p\lambda)(1-p)\lambda = (1-p)\lambda - p(1-p)\lambda^2,$$

and

$$\Pr[\text{A Head and a Tail are produced}] = p(1-p)\lambda^2.$$

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.

7.2. Semigroups

Consider the Poisson random walk $(X^a(t))_{t\in[0,\infty)}$ constructed in Section 7.1. 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 $t \geq 0$ 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 (36) and (37), 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 $f : \mathbb{Z}_2^n \to \mathbb{C}$, we have

$$P_t f = \sum_{S\subseteq[n]} e^{-t|S|}\widehat{f}(S)\chi_S. \tag{39}$$

Hence, not surprisingly at this point, similar to (35), we have $P_t f = T_{e^{-t}} f$.

The operators P_t are clearly linear operators from $L_2(\mathbb{Z}_2^n)$ to $L_2(\mathbb{Z}_2^n)$. The next lemma shows that they form a semigroup.

LEMMA 7.2.1. *We have* $P_0 = \mathrm{Id}$*, and* $P_t \circ P_s = P_{t+s}$ *for* $s, t \geq 0$*.*

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 (38) 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 $f \geq 0$, then $P_t f \geq 0$.
- *Preserves Order:* If $f \geq g$, then $P_t f \geq P_t g$.

These observations motivate the following definition.

DEFINITION 7.2.2. *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: $Q_t f \geq 0$ *almost everywhere if* $f \geq 0$ *almost everywhere,*
(3) Preserves Order: *If* $f \geq g$ *almost everywhere, then* $P_t f \geq P_t g$ *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

$$q_t(a, b) := (Q_t \mathbf{1}_b)(a), \tag{40}$$

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 $s \geq 0$, to have

$$\Pr[Y_{s+t} = b | Y_s = a] = q_t(a, b) := (Q_t \mathbf{1}_b)(a).$$

Since Q_t preserves positivity, we have $q_t(a, b) \geq 0$, 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 (38) 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 7.2.3. *If $(Q_t)_{t\in[0,\infty)}$ is Markovian semigroup, and $q_t(a,b)$ is defined as in (40). Show that $q_t(a,b)$ satisfies the Chapman-Kolmogorov equation (38).*

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 $(Y^a_t)_{t\in[0,\infty)}$ 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 7.2.4. *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 $t>0$,*

$$\sum_{x\in\Omega} \mu(\{x\})q_t(x,y) = \mu(y). \tag{41}$$

This means that the total "immigration" to y balances "emigration" from y. Note that (41) 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

$$\mathbb{E}[Q_t f] = \langle Q_t f, 1\rangle = \langle f, Q_t 1\rangle = \langle f, 1\rangle = \mathbb{E}[f]. \tag{42}$$

7.2.1. 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 7.2.5. *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 $t \geq 0$, we always have that $(-\frac{d}{dt}Q_t\big|_{t=0^+})1 = 0$.

REMARK 7.2.6. 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 $\{Q_\epsilon g : \epsilon > 0, g \in L_2\}$. ■

Let us go back to the semi-group $(P_t)_{t\in(0,\infty]}$ that we constructed from the parity Poisson process.

REMARK 7.2.7. 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 $|\rho| \leq 1$ 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 7.2.8. 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 $\frac{d}{dt}Q_t = -LQ_t$ 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 (33).

In Theorem 6.1.6 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 7.2.9. *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 $t \geq 0$ and every function $f \in L_2$ we have*

$$\mathbb{E}[\Phi(Q_t f)] \leq \mathbb{E}[\Phi(f)].$$

In particular taking $\Phi = |\cdot|^p$ for $p \geq 1$ we obtain $\|Q_t f\|_p \leq \|f\|_p$.

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 $\Phi(f) \geq a_\alpha f + b_\alpha$ which using the order-preserving property of Markovian semigroups reduces to the pointwise inequality

$$Q_t(\Phi(f)) \geq Q_t(a_\alpha f + b_\alpha) = a_\alpha(Q_t f) + b_\alpha.$$

Taking the supremum we obtain $Q_t(\Phi(f)) \geq \Phi(Q_t f)$. Taking the expectation and using the fact that symmetric Markovian semigroups preserve expectation (See 42), we obtain

$$\mathbb{E}[\Phi(f)] = \mathbb{E}[Q_t(\Phi(f))] \geq \mathbb{E}[\Phi(Q_t f)].$$

□

7.3. Some Examples

We close this chapter by mentioning some examples of Markovian semigroups.

EXAMPLE 7.3.1. 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

$$P_t = e^{-tL} = \mathrm{Id} + \sum_{k=1}^{\infty} \frac{(-t)^k}{k!} L^k, \tag{43}$$

we obtain

$$P_t f(x) = f(x) + t f'(x) + \frac{t^2}{2!} f''(x) + \dots$$

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 $f(x+t) \neq f(x) + t f'(x) + \frac{t^2}{2!} f''(x) + \dots$ 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 (43). ■

EXAMPLE 7.3.2. [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 $0 \le s_1 \le t_1 \le s_2 \le t_2$.
- $B_t - B_s \sim N(0, t-s)$ 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 $\lambda \ge 0$ 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

$$\begin{aligned}\frac{d}{dt}P_tf(x) &= \frac{d}{dt}\mathbb{E}[f(x+\sqrt{t}G)] = \frac{1}{2\sqrt{t}}\mathbb{E}[f'(x+\sqrt{t}G)G] = \frac{1}{2\sqrt{t}}\frac{1}{\sqrt{2\pi}}\int f'(x+\sqrt{t}y)e^{-y^2/2}ydy \\ &= \frac{1}{2\sqrt{2t\pi}}\int f'(x+\sqrt{t}y)\frac{d}{dy}(-e^{-y^2/2})dy = \frac{1}{2\sqrt{2t\pi}}\int \sqrt{t}f''(x+\sqrt{t}y)e^{-y^2/2}dy \\ &= \frac{1}{2}\mathbb{E}f''(x+\sqrt{t}G), \qquad (44)\end{aligned}$$

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 (44) shows that

$$\frac{d}{dt}P_tf(x) = \frac{1}{2}\Delta_x(P_tf(x)),$$

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 (44), 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

$$\frac{d}{dt}P_tf(x) = \frac{1}{2}\Delta(P_tf(x)),$$

holds. ∎

Example 7.3.3. [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 $\mathbb{Z}_2^n$. 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

$$U_tf(x) = \mathbb{E}\left[f(e^{-t}x + \sqrt{1-e^{-2t}}G)\right].$$

To describe the connection to the semigroup $(P_t)_{t\in[0,\infty)}$ on the cube $\mathbb{Z}_2^n$, consider a function $f : (\mathbb{R}, \gamma) \to \mathbb{R}$, and define $g_n : \mathbb{Z}_2^n \to \mathbb{R}$ 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 7.3.4. *Show that the generator of the n-dimensional Ornstein-Uhlenbeck semigroup is*

$$(Lf)(x) = \langle x, \nabla f(x)\rangle - (\Delta f)(x).$$

CHAPTER 8

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.

8.0.1. 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 $\mathcal{E}(f) := \mathcal{E}(f,f) = \sum |S||\widehat{f}(S)|^2 \geq 0$.

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 $t \geq 0$ 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

$$\Psi(t) \leq \|f\|_2^2 = \|P_0 f\|_2^2 = \Psi(0),$$

so that $\Psi'(0^+) \leq 0$. Thus $\mathcal{E}(f) := \mathcal{E}(f,f) \geq 0$, 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

$$\mathcal{E}(f) = \mathcal{E}(f,f) = 2^{-n} \sum_{x \sim y} \left(\frac{f(x) - f(y)}{2} \right)^2. \tag{45}$$

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

$$\mathcal{E}(f) = \frac{1}{2} \int |\nabla f|^2 dx dy, \tag{46}$$

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 8.0.5 (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 (46). 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. $|\Psi(a) - \Psi(b)| \leq C|a - b|$ 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

$$\mathcal{E}(\Psi(f)) \leq C^2 \mathcal{E}(f).$$

In particular, $\mathcal{E}(|f|) \leq \mathcal{E}(f)$. This can be generalized to other symmetric Markovian sermigroups under some mild technical conditions.

8.1. 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

$$\int_D f^2 \leq C_D \int_D |\nabla f|^2.$$

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

$$\mathrm{Var}_\nu(f) := \int f^2 d\nu - \left(\int f d\nu \right)^2 \leq C \int |\nabla f|^2 d\nu.$$

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

$$\mathbb{E}[f^2] - \mathbb{E}[f]^2 \leq \mathcal{E}(f) := \mathbb{E}\left[|\nabla f|^2 \right] := \mathbb{E}[fLf].$$

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.

8.2. 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 8.2.1. *For $p > 1$ and $a, b \geq 0$ we have*

$$(p-2)^2(a^p + b^p) - p^2(a^{p-1}b + ab^{p-1}) + 8(p-1)a^{p/2}b^{p/2} \geq 0.$$

PROOF. Because of the homogeneity, it suffices to prove that for $t \geq 1$

$$u(t) = (p-2)^2 t^p - p^2 t^{p-1} + 8(p-1)t^{p/2} - p^2 t + (p-2)^2 \geq 0.$$

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 $p \geq 2$,

$$\frac{p-2}{p} + \frac{2}{p}t^{-p/2} = \frac{p-2}{p} \cdot 1 + \frac{2}{p}t^{-p/2} \cdot t^{-p/2} \geq 1^{\frac{p-2}{p}} \left(t^{-p/2} \right)^{2/p} = t^{-1},$$

while for $p \in (1,2]$,

$$\frac{2-p}{2} + \frac{p}{2}t^{-1} = \frac{2-p}{2} \cdot 1 + \frac{p}{2}t^{-1} \geq 1^{\frac{p-2}{2}} \left(t^{-1}\right)^{p/2} = t^{-p/2},$$

we conclude $u''(t) \geq 0$ and the proof is finished. □

Now we will deduce the Stroock-Varopoulos inequality from Lemma 8.2.1. 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 8.2.2 (Stroock-Varopoulos). *For any $f : \mathbb{Z}_2^n \to [0,\infty)$, and every $p > 1$, we have*

$$\mathcal{E}(f^{p/2}) := \mathbb{E}\left[f^{p/2} L(f^{p/2})\right] \leq \frac{p^2}{4(p-1)} \mathbb{E}[f^{p-1} Lf].$$

PROOF. By Lemma 8.2.1, for any $a \geq 0$, we have the ponitwise inequality

$$(p-2)^2(a^p + f^p) - p^2(a^{p-1}f + af^{p-1}) + 8(p-1)a^{p/2}f^{p/2} \geq 0.$$

Since P_t is linear and order preserving for any $t \geq 0$, it holds pointwise that

$$(p-2)^2(a^p + P_t(f^p)) - p^2(a^{p-1}P_t f + aP_t(f^{p-1})) + 8(p-1)a^{p/2}P_t(f^{p/2}) \geq 0.$$

Hence setting $a = f$ we have

$$(p-2)^2(f^p + P_t(f^p)) - p^2(f^{p-1}P_t f + fP_t(f^{p-1})) + 8(p-1)f^{p/2}P_t(f^{p/2}) \geq 0.$$

We can take the expected value and arrive at

$$(p-2)^2(\mathbb{E}[f^p] + \mathbb{E}[P_t(f^p)]) - p^2(\mathbb{E}[f^{p-1}P_t f] + \mathbb{E}[fP_t(f^{p-1})]) + 8(p-1)\mathbb{E}[f^{p/2}P_t(f^{p/2})] \geq 0.$$

Since P_t is symmetric, it preserves expectation, and the above reduces to

$$\beta(t) = 2(p-2)^2\mathbb{E}[f^p] - 2p^2\mathbb{E}[f^{p-1}P_t f] + 8(p-1)\mathbb{E}[f^{p/2}P_t(f^{p/2})] \geq 0. \tag{47}$$

Now as $P_0 = \mathrm{Id}$, we have

$$\beta(0) = (2(p-2)^2 - 2p^2 + 8(p-1))\mathbb{E}[f^p] \geq 0,$$

and thus (47) implies that $\beta'(0^+) \geq 0$. But as $L = -\frac{d}{dt}P_t f\big|_{0^+}$, we have

$$0 \leq \beta'(0^+) = 2p^2\mathbb{E}[f^{p-1}Lf] - 8(p-1)\mathbb{E}[f^{p/2}L(f^{p/2})],$$

which completes the proof. □

REMARK 8.2.3. Note that in Theorem 8.2.2 we have equality when $p = 2$. ■

REMARK 8.2.4. 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 8.2.2 for any $p > 1$. ■

8.3. Entropy and Logarithmic Sobolev inequalities

We start by defining the notion of entropy.

DEFINITION 8.3.1. *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 $\mathrm{Ent}[g] \geq 0$, 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:

$$\mathrm{Ent}[f^2] \leq C\mathcal{E}(f).$$

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 8.3.2. *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*

$$\mathbb{E}_\mu[f^2 \ln(f^2)] - \mathbb{E}_\mu[f^2] \ln \mathbb{E}_\mu[f^2] \leq C\mathbb{E}_\mu[fLf].$$

It turns out that logarithmic Sobolev inequalities are equivalent to hyper-contractive inequalities. Recall that in Theorem 6.1.8 we showed that for $1 < p \leq q < \infty$, and $0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$, we always have

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

Using our semigroup notation, we can rewrite this as $\|P_t f\|_q \leq \|f\|_p$ for $0 \leq t \leq \frac{1}{2}(\ln(p-1) - \ln(q-1))$. 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 $0 \leq t \leq t(p,q)$ we have

$$\|Q_t f\|_q \leq \|f\|_p.$$

THEOREM 8.3.3 (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 8.3.3 combined with the hypercontractive estimates that we obtained in Theorem 6.1.8 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

$$\mathbb{E}[f^2 \ln(f^2)] - \mathbb{E}[f^2] \ln \mathbb{E}[f^2] \leq 2\mathbb{E}[fLf].$$

In order to prove Theorem 8.3.3 we first need the following lemma whose proof is based on the Stroock-Varopoulos theorem.

LEMMA 8.3.4. *The following statements are equivalent:*

(a): *For every* $f : \mathbb{Z}_2^n \to \mathbb{R}$,

$$\mathbb{E}[f^2 \ln(f^2)] - \mathbb{E}[f^2] \ln \mathbb{E}[f^2] \leq C\mathbb{E}[fLf].$$

(b): *For every nonnegative* $f : \mathbb{Z}_2^n \to \mathbb{R}$,

$$\mathbb{E}[f^2 \ln(f^2)] - \mathbb{E}[f^2] \ln \mathbb{E}[f^2] \leq C\mathbb{E}[fLf].$$

(c): *For every nonnegative* $f : \mathbb{Z}_2^n \to \mathbb{R}$, *and every* $p > 1$,

$$\mathbb{E}[f^p \ln(f^p)] - \mathbb{E}[f^p] \ln \mathbb{E}[f^p] \leq \frac{Cp^2}{4(p-1)} \mathbb{E}[f^{p-1} Lf].$$

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) ⇒ (a): This follows from $\mathcal{E}(|f|) \leq \mathcal{E}(f)$ which we proved in Section 8.0.1.

(b) ⇒ (c): This follows immediately from applying (b) to $f^{p/2}$ and then using the Stroock-Varopoulos inequality (Theorem 8.2.2). □

PROOF OF THEOREM 8.3.3. For $p \geq q > 1$, 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 $\phi_q(p) \leq \phi_q(q)$ for $p \geq q$. For $p \geq q$ denote

$$f_p := P_{t_q(p)} f \geq 0.$$

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

$$\frac{d}{dp}\phi_q(p) \leq 0 \iff \mathrm{Ent}(f_p^p) \leq \frac{Cp^2}{4(p-1)} \mathbb{E}[f_p^{p-1} L(f_p)]$$

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 $\frac{d}{dp}\phi_q(p)\Big|_{p=q} \leq 0$. Since $f_q = f$, this gives

$$\mathrm{Ent}(f^q) \leq \frac{Cq^2}{4(q-1)} \mathbb{E}[f^{q-1} L(f)],$$

which verifies the logarithmic Sobolev inequality by setting $q = 2$. □

EXERCISE 8.3.5. *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 $2C$.*

8.3.1. Tensorization of logarithmic Sobolev inequality. Recall that in Chapter 6 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 8.3.3 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 8.3.6 (Subadditivity of Entropy). *For $f : \mathbb{Z}_2^n \to [0,\infty)$, we have*

$$\mathrm{Ent}(f) \le \sum_{i=1}^{n} \mathrm{Ent}_i(f).$$

EXERCISE 8.3.7. *Prove the variational formulation of entropy:*

$$\mathrm{Ent}(f) = \sup\{\langle f, g\rangle : \mathbb{E}[e^g] \le 1,\ g : \mathbb{Z}_2^n \to \mathbb{R}\},$$

for every $f : \mathbb{Z}_2^n \to [0,\infty)$.

EXERCISE 8.3.8. *Prove Lemma 8.3.6 using the variational formulation of entropy.*

EXERCISE 8.3.9. *Use 8.3.6 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 8.3.10. *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 $\mathbb{Z}_2^n$.*

8.4. Reverse Hypercontractivity

In this lecture we are going to address a question related to expansion. We choose an element in a product space and change each coordinate with a small probability. How large is the probability that starting in a given small set A, the new point lands in another small set B? We are going to prove a lower bound on this probability that depends on the relative densities of such sets. To this end we introduce new tools concerning $\|\cdot\|p$ with $p < 1$.

Recall that for $0 \le \rho \le 1$, the noise operator is defined as

$$T_\rho f(x) = \mathbb{E}_y f(y),$$

where y is obtained by changing each coordinate $x \in \{0,1\}_2^n$ is obtained by changing each coordinate independently with probability $\frac{1-\rho}{2}$. Here y is called a ρ-correlated copy of x. Equivalently, we can

take $y = X^x(t_0)$ when $e^{-t_0} = \rho$ and $(X^x(t))_{t\in[0,\infty)}$ is the continuous random walk that we defined in Section 7.1 started at x.

Consider $A, B \subseteq \{0,1\}^n$ with relative densities α, β. That is

$$\frac{|A|}{2^n} = \alpha, \qquad \frac{|B|}{2^n} = \beta.$$

Note that α, β are small but may not be constant. We are interested in the following question: Pick a random $x \in \{0,1\}^n$ and ρ-correlated copy y of x; How small can the following probability can be?

$$\Pr[x \in A, y \in B] = \alpha \Pr[y \in B | x \in A].$$

Minimizing and and maximizing this quantity are both highly nontrivial. We will focus on minimization here, and defer the maximization problem to a later chapter where we discuss noise-stability. If we want to minimize this probability, intuitively, we would like to choose two opposite corners of the cube. The probability in this case can be upper-bounded using the following lemma whose proof we omit.

LEMMA 8.4.1. *Fix $a, b > 0$ and let $A, B \subseteq \{0,1\}^n$ be*

$$A = \left\{x | \sum x_i \leq \frac{n}{2} - a\sqrt{n}\right\},$$
$$B = \left\{x | \sum x_i \geq \frac{n}{2} + b\sqrt{n}\right\}.$$

Let $x \in \{0,1\}^n$ be uniform and y be a correlated copy of x. Then we have the following upper bound

$$\lim_{n\to\infty} \Pr[x \in A, y \in B] \leq \frac{\sqrt{1-\rho^2}}{2\pi a(\rho a + b)} e\left\{-\frac{1}{2} \cdot \frac{a^2 + b^2 + 2\rho ab}{1-\rho^2}\right\}$$

The main term in the bound above is the exponential one and it involves the relative densities of A and B as

$$\lim_{n\to\infty} \frac{|A|}{2^n} = \frac{1}{\sqrt{2\pi a}} e^{-a^2/2},$$
$$\lim_{n\to\infty} \frac{|B|}{2^n} = \frac{1}{\sqrt{2\pi b}} e^{-b^2/2}.$$

We are going to establish a lower-bound in Theorem 8.4.6 that almost matches the upper-bound of Lemma 8.4.1.

Let us first try the straight forward Fourier analytic approach that correspond to a spectral gap method as the eigenvectors of T_ρ are the characters because

$$T_\rho \chi_S = \rho^{|S|} \chi_S.$$

Therefore, the eigenvalues of T_ρ are $\rho^{|S|}$, the largest is 1, corresponding to the principal character ($S = \emptyset$), and the second largest is ρ, recall $\rho < 1$. To compute the Fourier expansions, fix x and average over y,

$$\Pr[x \in A, y \in B] = \mathbb{E} 1_A(x) 1_B(y) = \mathbb{E} 1_A(x) T_\rho 1_B(x).$$

We can use an spectral gap method, that is, to remove the first coefficient and bound the other ones by the second largest, finally we use Cauchy-Schwarz to derive the following,

$$\begin{aligned}\sum \widehat{1_A}(S)\widehat{1_B}(S)\rho^{|S|} &\geq \widehat{1_A}(\emptyset)\widehat{1_B}(\emptyset) - \rho\sum_{S\neq\emptyset}\left|\widehat{1_A}(S)\widehat{1_B}(S)\right| \\ &\geq \alpha\beta - \rho\left(\sum_{S\neq\emptyset}\left|\widehat{1_A}(S)\right|^2\right)^{1/2}\left(\sum_{S\neq\emptyset}\left|\widehat{1_A}(S)\right|^2\right)^{1/2} = \alpha\beta - \rho\sqrt{\alpha-\alpha^2}\sqrt{\beta-\beta^2}.\end{aligned}$$

When α and β are small, the second term in the right hand side is larger than the first term and the bound is negative (and useless) unless ρ is very small. So we need a deeper approach.

8.4.1. Reverse hypercontractivity and its applications. We are going to use "L_p-norms" for $p<1$, and obtain a reverse hypercontractivity inequality. It is similar to the hypercontractivity theorem but the direction of the inequality is reversed and applies to $-\infty < q \leq p < 1$. Also unlike the original hypercontractivity inequality, it only applies to non-negative functions. In fact all of the next four theorems and lemmas require the functions to be non-negative.

THEOREM 8.4.2 (Inverse Hölder Inequality). *If $f,g\geq 0$ are measurable functions with respect to a measure space then*

$$\langle f,g\rangle \geq \|f\|p\|g\|q,$$

where $-\infty<q,p<1$ and $\frac{1}{p}+\frac{1}{q}=1$.

REMARK 8.4.3. When $p<1$, although the function $\|\cdot\|p$ is not a norm, we still use this notation to denote

$$\left(\int |f|^p\right)^{1/p}.$$

In fact if $f,g\geq 0$ then the triangle inequality is reversed for $-\infty<p<1$,

$$\|f+g\|p \geq \|f\|p + \|g\|p.$$

To see this note that by Inverse Hölder Inequality

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

which simplifies to the desired inequality. ∎

THEOREM 8.4.4 (Reverse Hypercontractivity inequality). *Let $f:\{0,1\}^n\to[0,\infty)$, then*

$$\|T_\rho f\|q \geq \|f\|p,$$

for $0\leq\rho\leq\sqrt{\frac{p-1}{q-1}}$ and $-\infty<q\leq p<1$.

The proof is similar to the Hypercontractivity inequality. First one proves it for the 1-dimensional case and then induction establishes the general case.

COROLLARY 8.4.5. *Let $f,g:\{0,1\}^n\to[0,\infty)$ and $x\in\{0,1\}^n$ uniform and a ρ-correlated y copy of x, then*

$$\mathbb{E}f(x)g(y) \geq \|f\|p\|g\|q,$$

where $0<\rho\leq\sqrt{(1-p)(1-q)}\leq 1$ and $-\infty<q,p<1$.

PROOF. Let $p' = \frac{p}{p-1}$, so that p, p' are conjugate exponents. We use the reverse Hölder's inequality and then apply the inverse hypercontractivity inequality,

$$\begin{aligned}\mathbb{E}f(x)g(y) &= \mathbb{E}f(x)T_\rho g(x)\\ &\geq \|f\|p\|T_\rho g\|p'\\ &\geq \|f\|p\|g\|q,\end{aligned}$$

where the last inequality requires $0 < \rho \leq \sqrt{\frac{1-q}{1-p'}} = \sqrt{(1-p)(1-q)}$. □

Now we can prove the main theorem of the lecture, regarding the lower bound on the probability that a ρ-correlated copy of a uniform element that is in A lands in B.

THEOREM 8.4.6. *Let $A, B \subseteq \{0,1\}^n$ have relative densities*

$$\frac{|A|}{2^n} = e^{-a^2/2} \qquad\qquad \frac{|B|}{2^n} = e^{-b^2/2},$$

and let $x \in \{0,1\}^n$ be uniform and y be a ρ-correlated copy of x. Then

$$\Pr[x \in A, y \in B] \geq e\left\{-\frac{1}{2} \cdot \frac{a^2 + b^2 + 2\rho ab}{1-\rho^2}\right\}.$$

PROOF. Let p, q be such that $\rho^2 = (1-p)(1-q)$, by corollary 8.4.5 we have that

$$\Pr[x \in A, y \in B] = \mathbb{E}1_A(x)1_B(y) \geq \|1_A\|p\|1_B\|q.$$

Now our task is to optimize p so that the the R.H.S. is maximized. Note that the L_p norm can be expressed in term of the relative density because we are dealing with an indicator function

$$\|1_A\|p = e^{-a^2/2p} \qquad\qquad \|1_B\|p = e^{-b^2/2q}.$$

To simplify computations, write $p = 1 - \rho r$ and $q = 1 - \frac{\rho}{r}$ with $r > 0$, where

$$r = \frac{1-p}{\rho} = \frac{\rho}{1-q}.$$

Then the optimal solution is achieved when

$$r = \frac{\frac{b}{a} + \rho}{1 + \rho\frac{b}{a}}.$$

This gives the claimed lower bound as for the optimal value of r,

$$\frac{a^2}{p} + \frac{b^2}{q} = \frac{a^2 + b^2 + 2\rho ab}{1-\rho^2}.$$

□

We obtain the following corollary.

COROLLARY 8.4.7. *Let $A, B \subseteq \{0,1\}^n$ with relative densities $\alpha > 0$ and $\alpha^\sigma > 0$ respectively, where $\sigma > 0$. Let $x \in \{0,1\}^n$ be uniform and y be a ρ-correlated copy of x. Then*

$$\Pr[x \in A, y \in B] \geq \alpha\alpha^{(\sqrt{\sigma}+\rho)^2/(1-\rho^2)}.$$

In particular, if $|A| = |B|$, the this probability is at least $\alpha^{(1+\rho)/(1-\rho)}$.

Another interesting corollary of the inverse hypercontractivity inequality is that we can measure how T_ρ "smooths" the "peaks" of the function f. That is, we can bound $\Pr[T_\rho f(x) > 1 - \delta]$.

THEOREM 8.4.8. *Let $f : \{0,1\}^n \to [0,1]$ with $\mathbb{E}f = \alpha$. Then for any $0 < \rho < 1$ and $0 \le \epsilon \le 1-\alpha$ we have*

$$\Pr[T_\rho f > 1 - \delta] < \epsilon$$

provided that $0 \le \delta < \epsilon^{\rho^2/(1-\rho^2)+O(\kappa)}$, where $\kappa = \sqrt{\frac{\alpha \log(e/(1-\alpha))}{1-\rho}}$.

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 $\epsilon' := \mathbb{E}g \le \epsilon$. By the first moment method,

$$\alpha = \mathbb{E}f \ge (1 - \mathbb{E}h)b,$$

then

$$\mathbb{E}h > 1 - \frac{\alpha}{b} = \frac{1-\alpha}{1+\alpha},$$

and therefore support of h is not very small. Now, when $g(x) = 1$ we have $T_\rho(1 - f(x)) \le \delta$, so

$$T_\rho[(1-b)h(x)] < \delta$$

and so

$$T_\rho[h(x)] \le \frac{\delta}{1-b}.$$

Thus

$$\mathbb{E}[gT_\rho h] \le \frac{\delta\epsilon'}{1-b} = \frac{2\delta\epsilon'}{1-\alpha}. \tag{48}$$

Meanwhile, by Corollary 8.4.7

$$\mathbb{E}[gT_\rho h] \ge \epsilon' \cdot {\epsilon'}^{\frac{(\sqrt{\beta}+\rho)^2}{1-\rho^2}} \tag{49}$$

where $\beta = \frac{\log \mathbb{E}h}{\log \epsilon'}$. Now (48) and (49) together with our assumption on δ leads to the desired bound $\epsilon' \le \epsilon$. □

CHAPTER 9

Noise Stability

In Section **??** we studied lower bounds for

$$\Pr[x \in A, y \in B] = \alpha \Pr[y \in B | x \in A],$$

when the densities $\alpha = 2^{-n}|A|$ and $\beta = 2^{-n}|B|$ are fixed, and y is a ρ-correlated copy of x. We have

$$\Pr[x \in A, y \in B] = \mathbb{E}[1_A(x)1_B(y)] = \mathbb{E}[1_A(x)T_\rho(x)].$$

In this chapter we focus on maximizing this probability. A particular case of interest is when $A = B$. This leads to the notion of noise stability, which measures the proportion of elements x that remains in the support of a function f when we add some noise to them.

DEFINITION 9.0.9. *For $0 \leq \rho \leq 1$, the noise stability of $f : \{0,1\}^n \to \mathbb{R}$ is*

$$\mathbb{S}_\rho(f) := \mathbb{E} f(x) f(y),$$

where y is a ρ-correlated copy of x and $x \in \{0,1\}^n$ is uniform.

We are interested in finding the Boolean functions that have large noise stability. Note that for fixed x, $\mathbb{E}f(y) = T_\rho f(x)$, so we can look at $\mathbb{S}_\rho f$ as the correlation between f and $T_\rho f$, then

$$\mathbb{S}_\rho(f) = \mathbb{E} f(x) T_\rho f(x) = \sum_{S \subseteq [n]} \rho^{|S|} |\hat{f}(S)|^2.$$

Now, for a function $f : \{0,1\}^n \to \{0,1\}$ with $\mathbb{E}[f] = 1/2$ what is the largest possible value of $\mathbb{S}_\rho f(x)$? A first approach is to use the spectral gap. That is to separate the principal coefficient and upper bound $\rho^{|S|}$ by ρ to get

$$\begin{aligned} \mathbb{S}_\rho(f) &= \sum_{S \subseteq [n]} \rho^{|S|} |\hat{f}(S)|^2 \\ &\leq |\mathbb{E}f|^2 + \rho \sum_{S \neq \emptyset} |\hat{f}(S)|^2 = \frac{1}{4} + \frac{\rho}{4}, \end{aligned}$$

where in the last equality we used the assumption that f is balanced and boolean. On the other hand, half-cubes achieve this upper bound, for example, if $f(x) = x_1$, then the Fourier expansion is

$$f = \frac{1}{2} + \frac{1}{2}\chi_{\{i\}},$$

and so

$$\mathbb{S}_\rho(f) = \frac{1}{4} + \frac{\rho}{4}.$$

In general if the value of the function f depends only on 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.

In contrast with the half-cubes where the Fourier coefficients are concentrated in the first level, the next theorem states that even when all the influences are small, if one looks at the levels above d, the sum of the squared Fourier coefficients is still large.

THEOREM 9.0.10 (Bourgain 2000). *If $f : \{0,1\}^n \to \{0,1\}$ is balanced and $I_i \leq 10^{-d}$ for all $i \in [n]$, then*

$$\|f^{\geq d}\|2^2 = \sum_{|S|\geq d} |\hat{f}(S)|^2 \geq d^{-1/2-o\left(\sqrt{\frac{\ln\ln d}{\ln d}}\right)},$$

which is $d^{-1/2-o(1)}$.

Theorem 9.0.10 whose proof is highly nontrivial provides an upperbound on noise stability

COROLLARY 9.0.11. *If $f : \{0,1\}^n \to \{0,1\}$ is balanced and $I_i(f) = 2^{-O(1/\epsilon)}$ for all $i \in [n]$, then*

$$\mathcal{S}_{1-\epsilon}(f) \leq \frac{1}{2} - \epsilon^{1/2+o(1)}$$

EXERCISE 9.0.12. *Prove Corollary 9.0.11.*

Note that Corollary 9.0.11 is a great improvement compared with the bound obtained through the spectral gap, namely $\frac{1}{2} - \frac{\epsilon}{4}$. However, this is not sharp as the majority function has an even larger noise stability and it is conjecture that it achieves essentially the maximum noise stability among balanced functions. The majority function $\mathrm{Maj}_n : \{0,1\}^n \to \{0,1\}$ is defined as

$$\mathrm{Maj}_n : x \to \begin{cases} 1 & \sum x_i \geq n/2 \\ 0 & \sum x_i < n/2. \end{cases}$$

THEOREM 9.0.13. *The limit as n tends to ∞ of the noise stability of the majority function is*

$$\lim_{n\to\infty} \mathbb{S}_\rho(\mathrm{Maj}_n) = \frac{1}{4} + \frac{\arcsin\rho}{2\pi}.$$

It was conjectured by Subash Khot that under the condition of low influences, the majority function is the stablest Boolean function. The analogous statement in the Gaussian setting was proved in 1983 by Borell. Recently Mossel, O'Donnell, Oleszkiewicz found a method to deduce the discrete case from Borell's result.

THEOREM 9.0.14 (Majority is stablest). *For $0 < \rho < 1$, if $f : \{0,1\}^n \to \{0,1\}$ is balanced and $I_i(f) \leq \epsilon$ for all $i \in [n]$, then*

$$\mathbb{S}_\rho(f) \leq \frac{1}{4} + \frac{\arcsin\rho}{2\pi} + O\left(\frac{\log\log 1/\epsilon}{\log 1/\epsilon}\right).$$

Note that $O\left(\frac{\log\log 1/\epsilon}{\log 1/\epsilon}\right) = o(\epsilon)$. This theorem together with the so called "unique games conjecture" imply strong results about hardness of approximation.

For the proof of Theorem 9.0.14, we actually have to use geometry. The rest of the lecture we will define gaussian random variables in $\mathbb{R}^n$, state some of their basics properties and settle an analogous setup for the noise operators. In the next lecture we will prove the analogue of Theorem 9.0.13 in $\mathbb{R}^n$ and then translate it back to the discrete case.

DEFINITION 9.0.15. *The normal distribution on $\mathbb{R}$ is the probability distribution γ_1 on $\mathbb{R}$ with density function*

$$\frac{e^{-x^2/2}}{\sqrt{2\pi}},$$

that is

$$\gamma_1([a,b]) = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

A random variable g with distribution γ_1 is called a gaussian, these random variables have the property that $\mathbb{E}g = 0$ and $\mathbb{E}g^2 = 1$.

DEFINITION 9.0.16. *Let γ_n denote the corresponding product probability distribution on $\mathbb{R}^n$. In other words,*

$$\gamma_n(\{x \in \mathbb{R}^n | a_i \leq x_i \leq b_i\}) = \prod_{i=1}^n \gamma_1([a_i, b_i]). \tag{50}$$

The density function of γ_n is

$$\Phi_n(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{\frac{-\|x\|^2}{2}},$$

that is, if $A \in \mathbb{R}^n$, then $\gamma_n(A) = \int_A \Phi_n(x)dx$.

REMARK 9.0.17. The way that we defined the gaussian measure on $\mathbb{R}^n$ as the product space in (50) it is natural to expect that gaussians depend on the coordinates but they do not. The measure γ_n is uniformly distributed in spheres centered at the origin; when one fixes a sphere, the function Φ_n becomes constant and therefore independent of the coordinates.

In particular if $g_1, \ldots, g_n$ are i.i.d. gaussians and $\alpha, \beta \in \mathbb{R}^n$ with $\|\alpha\|_2 = \|\beta\|_2$, then the random variables

$$\alpha_1 g_1, + \cdots + \alpha_n g_n \qquad \text{and} \qquad \beta_1 g_1 + \cdots + \beta_n g_n,$$

have the same distribution. In particular, $\sum \alpha_i g_i$ has the same distribution as $\|\alpha\|_2 g$, where g is a one dimensional gaussian. ■

Now we consider the characters of $\mathbb{Z}_2^n$ in the gaussian space. Let $S \subseteq [n]$, then we define

$$\omega_S : x \mapsto \prod_{i \in S} x_i.$$

LEMMA 9.0.18. *The functions $\omega_S : (\mathbb{R}^n, \gamma_n) \to \mathbb{R}$ are orthonormal.*

PROOF. The inner product of any two function ω_S, ω_T can be express as the expected value of its product with respect to γ_n, so

$$\langle \omega_S, \omega_T \rangle \quad = \int \omega_S(x)\omega_T(x) d\gamma_n(x) = \mathbb{E}\omega_S(g_1, \ldots, g_n)\omega_S(g_1, \ldots, g_n),$$

where g_i are i.i.d. gaussians, so by independence

$$\mathbb{E}\omega_S(g_1, \ldots, g_n)\omega_S(g_1, \ldots, g_n) = \mathbb{E}\prod_{i \in S} g_i \prod_{i \in T} g_i = \left(\prod_{i \in S \cap T} \mathbb{E}g_i^2\right)\left(\prod_{i \in S \triangle T} \mathbb{E}g_i\right) = \begin{cases} 0 & S \triangle T \neq \emptyset \\ 1 & \text{otherwise} \end{cases}$$

Therefore, the inner product of any two of those functions is zero unless they are the same functions, and the norm of all of them is 1. □

REMARK 9.0.19. $\{\omega_S\}_{S \subseteq [n]}$ do not generate all of $L_2(\mathbb{R}^n, \gamma_n)$ but one can extend them using the so called Hermité polynomials to a basis for $L_2(\mathbb{R}^n, \gamma_n)$. ■

To define the noise stability we have to define what a ρ-correlated of a gaussian is:

DEFINITION 9.0.20. *Let $0 \leq \rho \leq 1$, two gaussians g, h are ρ-correlated if*

$$g = \rho h + \sqrt{1-\rho^2} g',$$

where g' is a gaussian independent of g, h.

Here g and h have ρ correlation. To see this use the definition of g to get

$$\mathbb{E}(g(x)h(x)) = \mathbb{E}[\rho h^2 + \sqrt{1-\rho^2} h g'] = \mathbb{E}\rho = \rho,$$

we again used that the expected value of a gaussian is zero and the second moment is 1. On the other hand, the coefficients are chosen so that g is a gaussian, note that $g = \rho h + \sqrt{1-\rho^2} g'$ has the same distribution as

$$(\rho^2 + \sqrt{1-\rho^2}) g'' = g'',$$

where g'' is a gaussian. Now, we define the analogue of the "noise operator" for the functions on the gaussian space.

DEFINITION 9.0.21. *Let $0 \leq \rho \leq 1$, then the Ornestein-Uhlenbeck operator acting on $L_2(\mathbb{R}^n, \gamma_n)$ is defined as*

$$U_\rho f(x) = \mathbb{E} f(y),$$

where $y = \rho x + \sqrt{1-\rho^2} g$ is a ρ-correlated copy of x.

LEMMA 9.0.22. *We have $U_\rho \omega_S = \rho^{|S|} \omega_S$.*

PROOF. For fixed x_i's we have

$$U_\rho \omega_S(x) = \mathbb{E} \prod_{i \in S} (\rho x_i + \sqrt{1-\rho^2} g_i) = \rho^{|S|} \prod_{i \in S} x_i = \rho^{|S|} \omega_S(x).$$

□

9.1. Noise Stability in Gaussian Space

Finally, we get to define the noise stability for function on the guassian space.

DEFINITION 9.1.1. *The noise stability of $f : (\mathbb{R}^n, \gamma_n) \to \mathbb{R}$ is defined as*

$$\mathbb{S}_\rho(f) := \mathbb{E}(f U_\rho f) = \mathbb{E} f(x) f(y),$$

where x has distribution γ_n and y is a ρ-correlated copy of x.

The following is the analogue of Theorem 9.0.13 in the gaussian space.

THEOREM 9.1.2 (Noise Stability of Majority Function). *Let* $\mathrm{Maj}_n : (R^n, \gamma_n) \to \{0, 1\}$

$$\mathrm{Maj}_n : x \to \begin{cases} 1 & \sum x_i \geq 0 \\ 0 & \sum x_i < 0 \end{cases}$$

then we have

$$\mathbb{S}_\rho(\mathrm{Maj}_n) = \frac{1}{4} + \frac{\arcsin(\rho)}{2\pi}$$

PROOF.

$$S_\rho(\mathrm{Maj}_n) = \mathbb{E}1_{[\sum x_i \geq 0]} 1_{[\sum y_i \geq 0]} = \mathbb{E}1_{[\rho \sum y_i + \sqrt{1-\rho^2} \sum g_i \geq 0]} 1_{[\sum y_i \geq 0]}$$

where y_i's and g_i's are i.i.d. Gaussians. $\sum y_i$ has the same distribution as $\sqrt{n}h$, where h is a Gaussian in $\mathbb{R}$. Similarly, $\sum g_i$ has distribution the same as $\sqrt{n}h'$. Therefore, the expected value is equal to:

$$\mathbb{E}1_{[\rho h + \sqrt{1-\rho^2}h' \geq 0]} 1_{[h \geq 0]} = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}$$

□

DEFINITION 9.1.3 (Gaussian Rearrangement). *Given $A \subset \mathbb{R}^n$ its Gaussian Rearrangement A^* is defined to be the interval (t, ∞) with $\gamma_1(t, \infty) = \gamma_n(A)$.*

Recall that γ_i is the Gaussian measure on $\mathbb{R}^*$. The following is the analogue of Theorem 9.0.14 in the gaussian space.

THEOREM 9.1.4 (Borrell 83). *Let $A, B \subseteq \mathbb{R}^n$. Then for any $0 \leq \rho \leq 1$ and $q \geq 1$ we have:*

$$\mathbb{E}(U_\rho A)^q B \leq \mathbb{E}(U_\rho A^*)^q B^*$$

In particular,

$$\mathbb{S}_\rho(A) = \mathbb{E}AU_\rho A \leq \mathbb{E}A^* U_\rho A^* = \mathbb{S}_\rho(A^*)$$

Hence, $\gamma_n(A) = \frac{1}{2}$ then $\mathbb{S}_\rho(A) \leq \mathbb{S}_\rho(\mathrm{Maj}_n) = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}$.

Note that unlike in Theorem 9.0.14, the previous theorem does not require any conditions on the influences.

9.2. Invariance Principle

THEOREM 9.2.1 (Invariance Principal I). *Let $Q(x_1, \ldots, x_n) = \sum_{S \leq [n]} \alpha_S \prod_{i \in S} x_i$ satisfies:*

$$\deg(Q) \leq d \tag{51}$$

$$\sum_{|S|>0} \alpha_S^2 = 1 \tag{52}$$

$$I_i := \sum_{S: i \in S} \alpha_S^2 \leq \tau \ \ \forall i : 1, \ldots, n \tag{53}$$

Then:

$$\sup_t |prob[Q(\varepsilon_1, \ldots, \varepsilon_n) \leq t] - \Pr[Q(g_1, \ldots, g_n) \leq t]| \leq O(d\tau^{\frac{1}{8d}})$$

Where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. ± 1 uniform random variables and $g_1, \ldots, g_n$ are i.i.d. Gaussians.

DEFINITION 9.2.2 (Rademacher Random Variable). *A uniform ± 1 random variable is called a rademacher random variable.*

THEOREM 9.2.3 (Invariance Principal II).

$$|\mathbb{E}[\Psi(Q(\varepsilon_1, \ldots, \varepsilon_n))] - \mathbb{E}[\Psi(Q(g_1, \ldots, g_n))]| \leq O(d9^d B\tau)$$

for all $\Psi : \mathbb{R} \to \mathbb{R}$ in C^4 (four times differentiable) with $|\Psi^{(4)}(t)| < B$ for all t.

Remark that if we could take $\Psi : x \to \begin{cases} |x| & x \leq t \\ 0 & otherwise \end{cases}$ then theorem II would imply theorem I. One instead has to approximate functions with bounded fourth derivatives.

PROOF. Let $Z_i = Q(g_1, \ldots, g_i, \varepsilon_{i+1}, \ldots, \varepsilon_n)$. We claim that $|\mathbb{E}\Psi(Z_{i-1} - \mathbb{E}\Psi(Z_i)| \leq O(B9^d I_i^2)$. First we show that the theorem can be extracted from this claim. Indeed,

$$|\mathbb{E}\Psi(Z_0) - \mathbb{E}\Psi(Z_n)| \leq \sum_{i=1}^{n} |\mathbb{E}\Psi((Z_{i-1}) - \mathbb{E}\Psi(Z_i)|$$

$$\leq O(B9^d) \sum_{i=1}^{n} I_i^2 = O(B9^d)$$

$$(\max I_i) \sum I_i \leq O(B9^d \tau) \sum I_i = O(B9^d \tau) \sum_{|S|>0} |S| \alpha_S^2 \leq O(dB9^d \tau) \sum_{|S|>0} \alpha_S^2 = O(\tau B9^d d)$$

To prove the claim $Q(x_1, \ldots, x_n) = \sum_{S: i \notin S} \alpha_S \prod_{j \in s} x_j \ + \ x_i \sum_{S: i \in S} \alpha_S \prod_{j \in s \setminus \{i\}} x_j = r(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) + x_i s(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, let $R = r(g_1, \ldots, g_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n)$ and $S = s(g_1, \ldots, g_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n)$. We have $Z_{i-1} = R + \varepsilon_i S$ and $Z_i = R + g_i S$. Now using Taylor's theorem:

$$\begin{aligned} |\mathbb{E}\Psi(Z_{i-1}) - \mathbb{E}\Psi(Z_i)| \ \leq \ & |\mathbb{E}\Psi(R) + \varepsilon_i S \Psi'(R) + \frac{(\varepsilon_i S)^2}{2} \Psi''(R) + \frac{(\varepsilon_i S)^3 \Psi^{(3)}(R)}{6} + E_1 \\ & - \mathbb{E}\Psi(R) - g_i S \Psi'(R) - \frac{(g_i S)^2}{2} \Psi''(R) - \frac{(g_i S)^3 \Psi^{(3)}(R)}{6} - E_2| \end{aligned}$$

Where $|E_1| \leq \frac{|\Psi^{(4)}(\xi)|(\varepsilon_i S)^4}{24} \leq \frac{B(\varepsilon_i S)^4}{24}$ for some ξ between R and $R + \varepsilon_i S$. Similarly, $|E_2| \leq \frac{B(g_i S)^4}{24}$. All terms get canceled except the error terms E_1 and E_2. So the expression is bounded by:

$$\mathbb{E}|\frac{B(\varepsilon_i S)^4}{24}| + \mathbb{E}|\frac{B(g_i S)^4}{24}| \leq \frac{B}{24}\mathbb{E}S^4 + \frac{3B}{24}\mathbb{E}S^4 \leq \frac{B}{6}\mathbb{E}S^4$$

by *Hypercontractivity*

$$\leq \frac{B9^d}{6}(\mathbb{E}S^2)^2 = \frac{B9^d}{6} \sum_{i \in S} \alpha_S^2 = \frac{B9^d}{6} I_i^2$$

□

Above we claimed that *Majority Function* is the stablest in Gaussian space. Now, we are going to prove this fact using the properties of *Threshold Function*.

DEFINITION 9.2.4. $T_\rho Q = \sum \rho^{|S|} \alpha_S \prod_{i \in S} x_i$

DEFINITION 9.2.5 (Threshold Function). *For any $\mu \in [-1, 1]$, the function $Thr^{(\mu)} : (\mathbb{R}, \gamma_1) \to \{-1, 1\}$ is defined as:*

$$Thr^{(\mu)} : x \to \begin{cases} 1 & x \geq t_0 \\ -1 & x < t_0 \end{cases}$$

with $\mathbb{E}Thr^{(\mu)} = \mu$.

THEOREM 9.2.6 (Majority is stablest in Gaussian space). *Let $f : (\mathbb{R}^n, \gamma_n) \to [1, -1]$ with $\mathbb{E}f = \mu$. Then:*

$$\mathbb{S}_\rho(f) \leq \mathbb{S}_\rho(Thr^{(\mu)})$$

THEOREM 9.2.7 (Majority is stablest in discrete setting). *Let $f : \{0, 1\}^n \to [-1, 1]$ and $I_i(f) = \sum_{S \ni i} |\widehat{f(s)}|^2 \leq \tau$ for every i. Then for every $0 \leq \rho \leq 1$, $\mathbb{S}_\rho(f) \leq \mathbb{S}_\rho(Thr^{(\mu)}) + O_\rho(\frac{loglog\frac{1}{\tau}}{log\frac{1}{\tau}})$ where $\mu = \mathbb{E}f$*

PROOF. Express $f = \sum \widehat{f(S)}\chi_S$. Let $Q(x_1, \ldots, x_n) = \sum_{s\subseteq[[n]} \widehat{f(S)} \prod_{i\in S} x_i$. Therefore, $f(x_1, \ldots, x_n) = Q(\varepsilon_1, \ldots, \varepsilon_n)$ where $\varepsilon_i = (-1)^{x_i}$. Let $(g_1, \ldots, g_n)$ be an i.i.d. Gaussian. We have

$$\mathbb{S}_\rho(f) = \sum \rho^{|S|}|\widehat{f(S)}|^2 = \mathbb{S}_\rho(Q(g_1, \ldots, g_n))$$

We would like to apply invariance principal to replace rademachers with Gaussians. However, since the degree of Q can be as large as n, we cannot apply invariance directly to Q. Instead, we apply a *smoothed* version of the theorem, which can be applied on $T_\beta Q$ for $\beta < 1$. Let $\rho = \rho'\beta^2$ where $\beta < 1$ is very close to 1. $(0 < 1 - \beta << 1 - \rho)$ to be determined later.

$$\mathbb{S}_\rho(f) = \sum \rho^{|S|}|\widehat{f(S)}|^2 = \sum(\rho'\beta^2)^{|S|}|\widehat{f(S)}|^2 = \mathbb{S}_{\rho'}(T_\beta Q(g_1, \ldots, g_n)).$$

Now using the smoothed invariance, $T_\beta Q(g_1, \ldots, g_n)$ is close in distribution to $T_\beta Q(\varepsilon_1, \ldots, \varepsilon_n)$ and hence it cannot be far from being in $[-1, 1]$. To make this precise we define function ξ as follows:

$$\xi : t \to \begin{cases} 0 & |t| \le 1 \\ (|t| - 1)^2 & |t| > 1 \end{cases}$$

Note that ξ measures the L_2-distance of t from its truncated value in [-1, 1]. By invariance principle of random variables $R = T_\beta Q(\varepsilon_1, \ldots, \varepsilon_n)$ and $S = T_\beta Q(g_1, \ldots, g_n)$ satisfy $|\mathbb{E}\xi(R) - \mathbb{E}\xi(S)| \le \tau^{\Omega(1-\beta)}$. Let S' be the truncation of S to the interval [-1, 1]:

$$S' = \begin{cases} S & |S| \le 1 \\ 1 & S > 1 \\ -1 & S < -1 \end{cases}$$

By assumption, $Q(\varepsilon_1, \ldots, \varepsilon_n) \in [-1, 1]$ and since T_β is an averaging operator, $T_\beta Q(\varepsilon_1, \ldots, \varepsilon_n) \in [-1, 1]$ and hence $\xi(R) = 0$.
Thus,

$$\mathbb{E}|\xi(S)| = \mathbb{E}(S - S')^2 \le \tau^{\Omega(1-\beta)}$$

$$\Rightarrow |\mathbb{S}_{\rho'}(S) - \mathbb{S}_{\rho'}(S')| = |\mathbb{E}SU_{\rho'}S - \mathbb{E}S'U_{\rho'}S'|$$

$$\le |\mathbb{E}SU_{\rho'}S - \mathbb{E}S'U_{\rho'}S| + |\mathbb{E}S'U_{\rho'}S - \mathbb{E}S'U_{\rho'}S'|$$

$$\le \|S - S'\|_2\|U_{\rho'}S\|_2 + \|S'\|_2\|U_{\rho'}(S - S')\|_2$$

$$\le \|S - S'\|_2\|S\|_2 + \|S'\|_2\|S - S'\|_2 \le \tau^(\Omega(1 - \beta)).$$

By Borrell's theorem, $\mathbb{S}_{\rho'}(S') \le \mathbb{S}_{\rho'}(Thr^{\mu'})$ where $\mu' = \mathbb{E}S'$. Now, we just have to show that $\mu = \mu'$:

$$|\mu - \mu'| = |\mathbb{E}(S - S')| \le \|S - S'\|_2 \le \tau^(\Omega(1 - \beta))$$

$$\Rightarrow |\mathbb{S}_{\rho'}(Thr^{\mu}) - \mathbb{S}_{\rho'}(Thr^{\mu'})| \le O(\frac{1-\beta}{1-\rho})$$

$$\Rightarrow \mathbb{S}_\rho(f) = \mathbb{S}_\rho(Thr^{(\mu)}) + O(\tau^{(\Omega(1-\beta)}) + \frac{1-\beta}{1-\rho}$$

and by optimizing the last expression over β the result yields to the theorem claim. □

9.3. Applications of "Majority is Stablest" Theorem

DEFINITION 9.3.1 (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 his preferences. The first bit indicates whether he prefers A to B; The second one shows his preference between B and C and the third one shows the same fact over 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* $(f(x), f(y), f(z))$.

DEFINITION 9.3.2 (Condorcet Paradox). *If f is the Majority function we 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 9.3.3. *A triple* $(a, b, c) \in \{-1, 1\}^3$ *is called rational, if it corresponds to a non-cyclic ordering.*

THEOREM 9.3.4 (Ken 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 $x, y, z \in \{-1, 1\}^n$ be the corresponding random string. Note that:

$$1_{[a_1=a_2=a3]} = \frac{1}{4} + \frac{1}{4}a_1a_2 + \frac{1}{4}a_1a_3 + \frac{1}{4}a_2a_3$$

$$\Rightarrow \Pr[(f(x), f(y), f(z))] = 1 - \mathbb{E}1_{[f(x)=f(y)=f(z)]} = \frac{3}{4} - \frac{1}{4}\mathbb{E}f(x)f(y) - \frac{1}{4}\mathbb{E}f(x)f(z) - \frac{1}{4}\mathbb{E}f(y)f(z)$$

$$= \frac{3}{4} - \frac{3}{4}\mathbb{E}f(x)f(y) = \frac{3}{4} - \frac{3}{4}\sum \widehat{f(S)}\widehat{f(T)}\mathbb{E}\chi_S(x)\chi_T(y)$$

Now we know that,

$$\mathbb{E}\chi_S(x)\chi_T(y) = (\prod_{i\in S\cap T} \mathbb{E}x_iy_i)(\prod_{i\in S\setminus T} \mathbb{E}x_i)(\prod_{i\in T\setminus S} \mathbb{E}y_i)$$

Since $\mathbb{E}y_i = \mathbb{E}x_i = 0$ and $\mathbb{E}x_iy_i = \frac{2}{6} - \frac{4}{6} = -\frac{1}{3}$, so $\mathbb{E}\chi_S(x)\chi_T(y) = \begin{cases} 0 & S \neq T \\ (\frac{-1}{3})^{|S|} & S = T \end{cases}$. Hence,

$$\Pr[(f(x), f(y), f(z)) \text{ is rational}] = \frac{3}{4} + \frac{3}{4}\sum(\frac{-1}{3})^{|S|}|\widehat{f(S)}|^2 \leq \frac{3}{4} + \frac{3}{4}\mathbb{S}_{\frac{1}{3}}(f)$$

COROLLARY 9.3.5. *If f satisfies* $I_i(f) = o_n(1)$ *and* $\mathbb{E}f = 0$, *then rationality of* $f \leq \frac{3}{4} + \frac{3}{4}\arcsin\frac{1}{3} + o_n(1) \leq 0.9123 + o_n(1)$.

Bibliography

[Ajt83] M. Ajtai. Σ_1^1-formulae on finite structures. *Ann. Pure Appl. Logic*, 24(1):1–48, 1983.

[AL93] Miklós Ajtai and Nathal Linial. The influence of large coalitions. *Combinatorica*, 13(2):129–145, 1993.

[BFH+13] Arnab Bhattacharyya, Eldar Fischer, Hamed Hatami, Pooya Hatami, and Shachar Lovett. Every locally characterized affine-invariant property is testable. In *Proceedings of the 45th Annual ACM Symposium on Symposium on Theory of Computing*, STOC '13, pages 429–436, New York, NY, USA, 2013. ACM.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, (90):5–43 (2001), 1999.

[BLR90] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. In *STOC '90: Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 73–83, New York, NY, USA, 1990. ACM.

[Bop97] Ravi B. Boppana. The average sensitivity of bounded-depth circuits. *Inform. Process. Lett.*, 63(5):257–261, 1997.

[FSS84] Merrick Furst, James B. Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Math. Systems Theory*, 17(1):13–27, 1984.

[GGR98] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.

[GS08] Ben Green and Tom Sanders. Boolean functions with small spectral norm. *Geom. Funct. Anal.*, 18(1):144–162, 2008.

[Has86] J Hastad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC '86, pages 6–20, New York, NY, USA, 1986. ACM.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. Assoc. Comput. Mach.*, 40(3):607–620, 1993.

[Man95] Yishay Mansour. An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *J. Comput. System Sci.*, 50(3, part 3):543–550, 1995. Fifth Annual Workshop on Computational Learning Theory (COLT) (Pittsburgh, PA, 1992).

[MO09] Ashley Montanaro and Tobias Osborne. On the communication complexity of xor functions. *CoRR*, abs/0909.3392, 2009.

[OW07] Ryan O'Donnell and Karl Wimmer. Approximation by DNF: examples and counterexamples. In *Automata, languages and programming*, volume 4596 of *Lecture Notes in Comput. Sci.*, pages 195–206. Springer, Berlin, 2007.

[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(4):333–338, 1987.

[RS93] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials and their applications to program testing. Technical report, Ithaca, NY, USA, 1993.

[SlV13] Amir Shpilka and Ben lee Volk. On the structure of boolean functions with small spectral norm. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:49, 2013.

[Smo87] Roman Smolensky. Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *STOC*, pages 77–82, 1987.

[TWXZ13] Hing Yin Tsang, Chung Hoi Wong, Ning Xie, and Shengyu Zhang. Fourier sparsity, spectral norm, and the log-rank conjecture. In *FOCS*, pages 658–667. IEEE Computer Society, 2013.

[Yao85] Andrew Chi-Chih Yao. Separating the polynomial-time hierarchy by oracles. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science*, SFCS '85, pages 1–10, Washington, DC, USA, 1985. IEEE Computer Society.

[ZS10] Zhiqiang Zhang and Yaoyun Shi. On the parity complexity measures of boolean functions. *Theor. Comput. Sci.*, 411(26-28):2612–2618, 2010.