

Fourier-Analytic Methods in Discrete Mathematics

Tibor Szabó Uli Wagner

Institute of Theoretical Computer Science

ETH Zürich

8092 Zürich, Switzerland

{szabo|uli}@inf.ethz.ch

February 7, 2007

Contents

0	Overview	2
1	Preliminaries	4
1.1	Discrete Measures, Integrals, Expectations	4
1.2	Tensor Products and Product Measures	5
1.3	Norms	6
2	Characters and the Fourier Transform	10
2.1	Characters and the Dual Group	10
2.2	The Fourier Transform	13
2.3	Convolutions	15
3	Boolean Functions	16
3.1	Examples and Various Viewpoints	16
3.1.1	Fourier Transforms of Boolean Functions	19
3.2	Linearity Testing	21
4	Influences	23
4.1	Influences and the Edge-Isoperimetric Inequality	24
4.2	Influences and the Fourier Transform	26
5	The Noise Operator	32
5.1	The Definition of T_ξ	32
5.2	The Hypercontractive Inequality	34
6	Influences for General Product Measures	37
7	Thresholds	38
8	Bounds for Error-Correcting Codes	39
8.1	Linear Codes.	41
8.2	Asymptotically Good Codes	45

Chapter 0

Overview

These are lecture notes for our joint course “Fourier-Analytic Methods in Discrete Mathematics” taught at ETH Zürich in fall, 2006. These are very rough notes, written mostly for ourselves in preparation for the class, and they come without any guarantee for correctness or completeness.

Partial syllabus for the course

1. Some examples of problems and questions to which Fourier-analytic methods have been applied: Threshold phenomena for monotone properties of random graphs; linearity testing; voting schemes, Boolean functions, and influences.
2. Basics notions of discrete Fourier Analysis: The dual group \hat{G} of characters of a finite group G ; orthogonality relations; the Fourier transform and Fourier inversion; Plancherel formula and Parseval’s identity; convolutions; the special case of the cube \mathbf{Z}_2^n .
3. Boolean functions and voting schemes. Computation of the Fourier transform for some examples. Linearity testing. The Noise Operator. “Dictatorship testing”.
4. Influences; isoperimetric inequalities; examples; KKL Theorem: statement and proof modulo the Hypercontractive inequality.
5. Proof of the Hypercontractive Inequality for the Noise Operator.
6. KKL2; nonuniform measures; BKKKL; monotoneization.
7. Monotone properties and thresholds. Sharp thresholds and functions without thresholds.
8. Applications to error-correcting codes; Packing Bound; Gilbert-Varshamov Bound; Macwilliams’ Identity for linear codes; Delsarte’s Linear Program and the resulting asymptotic bound.
9. Equations over finite fields.

Acknowledgements

The choice of material as well as our presentation are heavily based on the following sources (however, as usual, we are to blame for all errors and omissions):

1. A 5-week intensive graduate course “Harmonic Analysis in Computer Science and Combinatorics” at Charles University, Prague, from February 2 to March 3, 2006, organized by Jiří Matoušek and Jaroslav Nešetřil and jointly taught by Gil Kalai and Nati Linial, with guest lectures by Oded Regev, Alex Samorodnisky, Assaf Naor, Alex Iosevich, and the first author.
2. The scribe notes for the course “Analytical Methods in Combinatorics and CS” taught by Irit Dinur and Ehud Friedgut at the Hebrew University of Jerusalem during the Winter Semester 2005/06.
3. The scribe notes for the course “Harmonic Analysis and Combinatorial Applications” taught by Nati Linial at the University of Washington, Seattle, during the Winter Term 2005.
4. The lecture notes “The Fourier Transform and Equations over Finite Abelian Groups” by László Babai.
5. The scribe notes for the course “Algorithmic Introduction to Coding Theory” taught by Madhu Sudan at MIT in Fall 2002.

Chapter 1

Preliminaries

1.1 Discrete Measures, Integrals, Expectations

Let X be a finite set, and let μ be a (nonnegative) *measure* on X . This means that μ assigns a nonnegative real number $\mu(A)$ to every subset $A \subseteq X$ and that this assignment is *additive*, i.e., $\mu(A) = \sum_{x \in A} \mu(x)$, where $\mu(x)$ is a shorthand for $\mu(\{x\})$.¹ The pair (X, μ) is called a *finite measure space*.

We will use the following notations interchangeably²: For $f : X \rightarrow \mathbf{C}$,

$$\int_X f(x) d\mu(x) := \sum_{x \in X} f(x) \mu(x).$$

If μ is understood from the context, we also use the simplified notation $\int_X f$, or even $\int f$.

If $\mu(X) = 1$, then we call μ a *probability measure* on X . In this case, we also often write integrals as expectations,

$$\mathbf{E}_{x \sim \mu}[f(x)] := \int_X f(x) d\mu(x).$$

Again, we often simplify this to $\mathbf{E}_x[f(x)]$ or $\mathbf{E}[f]$ if μ is clear from the context (for example, if μ is the uniform probability measure on X , i.e., $\mu(x) = 1/|X|$ for all $x \in X$).

The *variance* of f is defined as

$$\text{Var}[f] := \mathbf{E}[|f - \mathbf{E}[f]|^2] = \mathbf{E}[|f|^2] - |\mathbf{E}[f]|^2.$$

¹Thus, in our case, μ is completely determined by the values $\mu(x)$ on the $x \in X$, so we could simply view μ as a function on X . In more general measure spaces, such as the interval $[0, 1]$ with the 1-dimensional volume, there might be uncountably infinite subsets whose measure cannot be defined as the sum of measures of 1-element subsets, since one cannot assign a finite value to a sum of uncountably many nonzero terms. In such a situation, one needs to delve into the technicalities of σ -algebras, measurable sets, etc., which we want to avoid here. Nonetheless, we would like to keep the distinction between a function and a measure.

²If we were considering more general infinite measure spaces, we would have to worry about summability or integrability.

1.2 Tensor Products and Product Measures

For sets X, Y , we denote by Y^X the set of all functions $X \rightarrow Y$. If Y is a field, e.g. $Y = \mathbf{C}$, then \mathbf{C}^X is a vector space under componentwise addition and scalar multiplication.

Tensor products. For sets X, Y and functions $f : X \rightarrow \mathbf{C}$ and $g : Y \rightarrow \mathbf{C}$, we define the *tensor product*³ $f \otimes g : X \times Y \rightarrow \mathbf{C}$ by $(f \otimes g)(x, y) := f(x) \cdot g(y)$. Observe that the tensor product is *bilinear*, i.e., $a(f \otimes g) = (af) \otimes g = f \otimes (ag)$ for all $a \in \mathbf{C}$ and $(f_1 + f_2) \otimes g = f_1 \otimes g + f_2 \otimes g$ and $f \otimes (g_1 + g_2) = f \otimes g_1 + f \otimes g_2$ for all $f_1, f_2 : X \rightarrow \mathbf{C}$ and $g_1, g_2 : Y \rightarrow \mathbf{C}$.

Remark 1.1. If X and Y are finite, then every function $h : X \times Y \rightarrow \mathbf{C}$ is a *finite* linear combination of such tensor products; in fact, given bases $\{f_a\}$ and $\{g_b\}$ of \mathbf{C}^X and \mathbf{C}^Y , respectively, the set $\{f_a \otimes g_b\}$ forms a basis for $\mathbf{C}^{X \times Y}$.

For example, if $\{e_a\}_{a \in X}$ and $\{e_b\}_{b \in Y}$ are the standard bases of \mathbf{C}^X and \mathbf{C}^Y , respectively, where $e_a(b) = 1$ if $a = b$ and $e_a(b) = 0$ otherwise, then $e_a \otimes e_b = e_{(a,b)}$, so we obtain again the standard basis of $X \times Y$.

Tensor products of linear operators. Given linear maps⁴ $S : \mathbf{C}^X \rightarrow \mathbf{C}^W$ and $T : \mathbf{C}^Y \rightarrow \mathbf{C}^Z$, we define their tensor product $S \otimes T : \mathbf{C}^{X \times Y} \rightarrow \mathbf{C}^{W \times Z}$ by setting

$$(S \otimes T)(f \otimes g) := (Sf) \otimes (Tg)$$

for generators $f \otimes g \in \mathbf{C}^{X \times Y}$ and extending by linearity.⁵

We note that everything we said about tensor products carries over almost verbatim to any finite number of factors.

Product measures. If $(X_1, \mu_1), \dots, (X_n, \mu_n)$ are finite measure spaces⁶ then the tensor product $\mu_1 \otimes \dots \otimes \mu_n$ given by $(\mu_1 \otimes \dots \otimes \mu_n)(x_1, \dots, x_n) = \prod_i \mu_i(x_i)$ is called the *product measure* on $X_1 \times \dots \times X_n$.

Example 1.2. Let $0 \leq p \leq 1$ and define $\mu_p = \mu_p^{(1)}$ on $\mathbf{Z}_2 = \{0, 1\}$ by $\mu_p(1) = p$ and $\mu_p(0) = 1 - p$. If $(X_i, \mu_i) = (\mathbf{Z}_2, \mu_p)$ for all i , then the product measure $\mu_p = \mu_p^{(n)} = \mu_p^{(1)} \otimes \dots \otimes \mu_p^{(1)}$ on \mathbf{Z}_2^n is given by $\mu_p(x) = p^{|x|}(1 - p)^{n - |x|}$, where $|x| = |\{i : x_i = 1\}|$. In particular, $\mu_{1/2}$ is the uniform measure on \mathbf{Z}_2^n .

³If you are familiar with abstract tensor products (which can be defined as appropriate quotient spaces, or via universal properties), this defines an embedding of the abstract tensor product $\mathbf{C}^X \otimes_{\mathbf{C}} \mathbf{C}^Y$ into $\mathbf{C}^{X \times Y}$. In the finite case, embedding is also surjective. In general, one needs to take the closure of the image.

⁴When we think of the elements of \mathbf{C}^X as functions, we will also speak of *linear operators*.

⁵This is well-defined, i.e., does not depend on how we decompose a function $h : X \times Y$ as a linear combination of generators, because we could first define $S \otimes T$ on a subset of generators that form a basis.

⁶In general measure spaces, the definition of the product measure is more complicated, first assigning a measure to product sets $A_1 \times \dots \times A_n$ (or even so-called cylinders, in the case of infinitely many factors) and then using a limiting process.

In our finite setting, integrals are simply finite sums, integrals with respect to a product measure can be evaluated by iterated univariate integrals, and we can freely exchange the order of integration (summation).⁷ For convenience, we just state this in the bivariate case.

Lemma 1.3. *For finite measure spaces (X, μ) and (Y, ν) and $h : X \times Y \rightarrow \mathbf{C}$,*

$$\int_{X \times Y} h(x, y) d(\mu \otimes \nu)(x, y) = \int_X \int_Y h(x, y) d\nu(y) d\mu(x) = \int_Y \int_X h(x, y) d\mu(x) d\nu(y)$$

In the special case that the function on the product space that we want to integrate has itself a product structure, the integration becomes particularly simple:

Corollary 1.4. *If $f_i : X_i \rightarrow \mathbf{C}$, $1 \leq i \leq n$, then*

$$\int_{X_1 \times \dots \times X_n} (f_1 \otimes \dots \otimes f_n) d(\mu_1 \otimes \dots \otimes \mu_n) = \prod_{i=1}^n \int_{X_i} f_i d\mu_i.$$

If all the measures are probability distributions, then the product measure captures mutual independence of the distributions. In particular, in this case the corollary states that the expectation of a product of independent random variables equals the product of the individual expectations: If $y = (y_1, \dots, y_n) \sim \mu := \mu_1 \otimes \dots \otimes \mu_n$ then

$$\mathbf{E}_{y \sim \mu} \left[\prod_{i=1}^n f_i(y_i) \right] = \prod_{i=1}^n \mathbf{E}_{y_i \sim \mu_i} [f_i(y_i)].$$

1.3 Norms

Let V be a real or complex vector space. A *norm* on V is a map $\|\cdot\| : V \rightarrow \mathbf{R}$ such that the following properties hold for all $u, v \in V$ and all scalars λ :

1. (*nonnegativity and strictness*) $\|v\| \geq 0$, and $\|v\| = 0$ iff $v = 0$;
2. (*homogeneity*) $\|\lambda v\| = |\lambda| \|v\|$;
3. (*triangle inequality*) $\|u + v\| \leq \|u\| + \|v\|$.

p -Norms and L^p -spaces. Let (X, μ) be a finite measure space. For $1 \leq p < \infty$, the *p -norm* of a function $f : X \rightarrow \mathbf{C}$ is defined by

$$\|f\|_p := \left(\int_X |f|^p d\mu \right)^{\frac{1}{p}};$$

⁷In general measure spaces, one needs suitable assumptions on the function f : either that f is nonnegative (allowing the possibility that $\int f = \infty$) or *absolute integrability*, i.e., $\int |f| < \infty$. This is called the *Tonelli-Fubini Theorem*.

for $p = \infty$, the ∞ -norm is defined by

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

The space \mathbf{C}^X together with the norm⁸ is denoted by $L^p(X, \mu)$. For $1 < p < \infty$, the triangle inequality $\|f+g\|_p \leq \|f\|_p + \|g\|_p$ is nontrivial to verify and known as *Minkowski's Inequality*. It is a consequence of *Hölder's Inequality*, which we state without a proof:

Fact 1.5 (Hölder's Inequality). *Let $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$ (with the convention that $\frac{1}{\infty} = 0$). Such p and q are called conjugate exponents. Then, for all $f, g : X \rightarrow \mathbf{C}$,*

$$\|fg\|_1 = \int_X |f(x)g(x)| d\mu(x) \leq \|f\|_p \|g\|_q.$$

Later on, we will need a generalization of Minkowski's Inequality from sums to integrals. Let (X, μ) and (Y, ν) be finite measure spaces. If $f : X \rightarrow \mathbf{C}$ and $g : Y \rightarrow \mathbf{C}$ then by Corollary 1.4,

$$\|f \otimes g\|_p = \|f\|_p \cdot \|g\|_p.$$

Consider now a general function $h : X \times Y \rightarrow \mathbf{C}$, and let $1 \leq p, q \leq \infty$. If we fix the second coordinate, say to $y \in Y$, then $h(\cdot, y)$ is a function $X \rightarrow \mathbf{C}$, and we can consider the p -norm of that function:

$$\|h\|_{X:p}(y) := \left(\int_X |h(x, y)|^p d\mu(x) \right)^{1/p}.$$

This defines a function $\|h\|_{X:p} : Y \rightarrow \mathbf{R}_+$, for which we can consider the q -norm,

$$\| \|h\|_{X:p} \|_{Y:q} = \left(\int_Y \|h\|_{X:p}^q d\nu(y) \right)^{1/q}.$$

Alternatively, we could first take the q -norm in the Y -direction and then take the p -norm in the X -direction. If $p = q$, the order does not matter:

$$\begin{aligned} \| \|h\|_{X:p} \|_{Y:p} &= \left(\int_Y \|h\|_{X:p}^p d\nu(y) \right)^{1/p} \\ &= \left(\int_Y \left(\left(\int_X |h(x, y)|^p d\mu(x) \right)^{1/p} \right)^p d\nu(y) \right)^{1/p} \\ &= \left(\int_Y \int_X |h(x, y)|^p d\mu(x) d\nu(y) \right)^{1/p} \\ &= \left(\int_{X \times Y} |h(x, y)|^p d(\mu \otimes \nu)(x, y) \right)^{1/p} = \|h\|_p, \end{aligned}$$

⁸In general measure spaces, $\|f\|_p$ might be infinite, and one defines $L^p(X, \mu)$ as the space of functions with $\|f\|_p < \infty$. Moreover, there might be nonempty sets of measure zero in X , and in order to guarantee strictness of $\|\cdot\|_p$, one needs to identify functions that differ only on a set of measure zero. Furthermore, in general the maximum in the definition of $\|\cdot\|_\infty$ has to be replaced by the so-called "essential supremum".

and symmetrically $\| \|h\|_{Y:p} \|_{X:p} = \|h\|_p$ if we exchange the roles of the first and second factor.

The situation becomes more interesting if $p \neq q$. Minkowski's Integral Inequality, which we state without a proof, tells us that we get the larger value if we take the norm with the larger exponent first.

Fact 1.6 (Minkowski's Integral Inequality). *If $1 \leq p \leq q \leq \infty$ then*

$$\| \|h\|_{X:p} \|_{Y:q} \leq \| \|h\|_{Y:q} \|_{X:p}$$

For example, if $X = \{0, 1\}$ is a two-element set with the counting measure $\mu(0) = \mu(1) = 1$, then setting $p = 1$, we retrieve the triangle inequality for the q -norm for the two functions $h_x(y) := h(x, y)$, $x \in \{0, 1\}$, defined on Y : We have $\|h\|_{X:1}(y) = |h_0(y)| + |h_1(y)|$ and $\|h\|_{Y:q}(x) = \|h_x\|_q$, and so

$$\|h_0 + h_1\|_q \leq \| |h_0| + |h_1| \|_{Y:q} \leq \|h_0\|_q + \|h_1\|_q.$$

Operator norms. If V and W are normed vector spaces (where we use the same symbol $\|\cdot\|$ to denote the norm on either of these spaces) and if $T : V \rightarrow W$ is a linear map (or "linear operator", if we think of V, W as spaces of functions) then the *operator norm* of T is defined as

$$\|T\| := \sup_{v \neq 0} \frac{\|Tv\|}{\|v\|} = \sup_{\|v\|=1} \|Tv\|.$$

In other words, $\|T\|$ is the smallest number α (more precisely, the infimum) that satisfies $\|Tv\| \leq \alpha\|v\|$ for all $v \in V$.

We will need the following lemma in the proof of the so-called *Hypercontractive Inequality* for the *Noise Operator* in Section 5.2, which in turn will be the main technical tool in the proof of the KKL Theorem 4.17 on influences of boolean functions.

Lemma 1.7. *Let X_1, X_2, Y_1, Y_2 be finite measure spaces, let $1 \leq p \leq q \leq \infty$, and let $T_1 : L^p(X_1) \rightarrow L^q(Y_1)$ and $T_2 : L^p(X_2) \rightarrow L^q(Y_2)$ be linear operators. Then the norm of the operator $T_1 \otimes T_2 : L^p(X_1 \times X_2) \rightarrow L^q(Y_1 \times Y_2)$ satisfies*

$$\|T_1 \otimes T_2\| \leq \|T_1\| \|T_2\|.$$

Observe that this estimate is trivial for generators $f_1 \otimes f_2 \in L^p(X_1 \times X_2)$, because $\|(T_1 \otimes T_2)(f_1 \otimes f_2)\|_q = \|(T_1 f_1) \otimes (T_2 f_2)\|_q = \|T_1 f_1\|_q \|T_2 f_2\|_q \leq \|T_1\| \|T_2\| \|f_1\|_p \|f_2\|_p = \|T_1\| \|T_2\| \|f_1 \otimes f_2\|_p$. For general functions, we first prove an auxiliary lemma:

Lemma 1.8. *Let X, Y, Z be finite measure spaces, let $1 \leq p, q \leq \infty$ (here, we do not assume $p \leq q$), and let $T : L^p(X) \rightarrow L^q(Y)$ be a linear operator. Then, for any function $k : X \times Z \rightarrow \mathbf{C}$ and any $z \in Z$, we have*

$$\|(T \otimes \text{id}_Z)k\|_{Y:q}(z) \leq \|T\| \|k\|_{X:p}(z)$$

and a symmetric statement holds for the operator $\text{id}_Z \otimes T : L^p(Z \times X) \rightarrow L^q(Z \times Y)$.

Proof. Let $k : X \times Z \rightarrow \mathbf{C}$. We can write $k = \sum_{z \in Z} k_z \otimes e_z$, where $k_z = k(\cdot, z) : X \rightarrow \mathbf{C}$ for each $z \in Z$. Thus, $(T \otimes \text{id}_Z)k = \sum_{z \in Z} T k_z \otimes e_z$. In other words, for each fixed $z \in Z$, $k(x, z) = k_z(x)$ and $((T \otimes \text{id}_Z)k)(y, z) = (T k_z)(y)$. Therefore,

$$\begin{aligned} \|(T \otimes \text{id}_Z)k\|_{Y:q}(z) &= \left(\int_Y |((T \otimes \text{id}_Z)k)(y, z)|^q dy \right)^{1/q} \\ &= \|T k_z\|_q \leq \|T\| \|k_z\|_p = \|T\| \|k\|_{X:p}(z). \end{aligned}$$

□

Proof of Lemma 1.7. Let $h : X_1 \times X_2 \rightarrow \mathbf{C}$. We can write $T_1 \otimes T_2 = (T_1 \otimes \text{id}_{Y_2})(\text{id}_{X_1} \otimes T_2)$, where the product of functions means their composition. Therefore,

$$\begin{aligned} \|(T_1 \otimes T_2)h\|_{Y_1 \times Y_2:q} &= \left\| \underbrace{\|(T_1 \otimes \text{id}_{Y_2:q})(\text{id}_{X_1} \otimes T_2)h\|_{Y_1:q}}_{\leq \|T_1\| \|(\text{id}_{X_1} \otimes T_2)h\|_{X_1:p}} \right\|_{Y_2:q} \\ &\leq \|T_1\| \left\| \|(\text{id}_{X_1} \otimes T_2)h\|_{X_1:p} \right\|_{Y_2:q} \\ &\leq \|T_1\| \left\| \underbrace{\|(\text{id}_{X_1} \otimes T_2)h\|_{Y_2:q}}_{\leq \|T_2\| \|h\|_{X_2:p}} \right\|_{X_1:p} \leq \|T_1\| \|T_2\| \|h\|_{X_1 \times X_2:p}, \end{aligned}$$

as desired, where we first apply Lemma 1.8 with $k = (\text{id}_{X_1} \otimes T_2)h$, then Minkowski's Integral Inequality, and then once again Lemma 1.8 with $k = h$. □

Chapter 2

Characters and the Fourier Transform

2.1 Characters and the Dual Group

Let G be a finite abelian group, written additively. A *character* of G is a homomorphism $\chi : G \rightarrow \mathbf{C}^\times$ from G to the multiplicative group of nonzero complex numbers, i.e.,

$$\chi(a + b) = \chi(a)\chi(b)$$

for all $a, b \in G$.

Observe that for $n := |G|$, we have

$$n \cdot a := \underbrace{a + a + \dots + a}_{n \text{ times}} = 0$$

for all $a \in G$, hence $\chi(a)^n = 1$, i.e., all values of χ are n^{th} roots of unity. Thus,

$$\chi(-a) = \frac{1}{\chi(a)} = \overline{\chi(a)}$$

for all $a \in G$, where the bar denotes complex conjugation.

The *principal character* is defined by

$$\chi_0(a) = 1$$

for all $a \in G$.

Lemma & Definition 2.1 (Dual Group \widehat{G}). Let G be a finite abelian group, and let \widehat{G} be the set of characters of G . Consider the operation of pointwise multiplication of \widehat{G} , i.e., the product $\chi\psi$ of two characters $\chi, \psi \in \widehat{G}$ is defined by $(\chi\psi)(a) := \chi(a)\psi(a)$. This pointwise product defines again a character, and \widehat{G} with this product forms an abelian group, called the *dual group* of G , with unit element χ_0 and with group inverse $\chi^{-1} = 1/\chi = \bar{\chi}$.

The proof is an easy exercise. We recall that every finite abelian group can be written as a direct sum of finite cyclic groups. We use the notation $\mathbf{Z}_r = \{0, \dots, r-1\}$ (with addition modulo r) for the finite cyclic group of order r .

Fact 2.2 (Fundamental Theorem of Finite Abelian Groups). *Let G be a finite abelian group. Then there are nonnegative integers n_1, \dots, n_k such that*

$$G \cong \mathbf{Z}_{n_1} \oplus \dots \oplus \mathbf{Z}_{n_k}.$$

(In fact, one can require $n_1 | n_2 | \dots | n_k$, where $n|m$ means that n is a divisor of m , and this makes the tuple (n_1, \dots, n_k) unique.)

The complex n^{th} roots of unity form a cyclic group of order n (with multiplication). The following lemma says that any choice of a generator of this group determines an isomorphism $\widehat{\mathbf{Z}}_n \cong \mathbf{Z}_n$.

Lemma 2.3. *Let ω be a principal complex n^{th} root of unity. Then, for $j \in \mathbf{Z}$, the map $\chi_j : \mathbf{Z}_n \rightarrow \mathbf{C}^\times$ defined by $\chi_j(a) := \omega^{ja}$ is a character of \mathbf{Z}_n . Moreover, the map $j \mapsto \chi_j$ is a homomorphism, i.e., $\chi_{j+k} = \chi_j \chi_k$ (the addition in the subscript is modulo n), which is injective, i.e., $\chi_j = \chi_k$ if and only if $j \equiv k \pmod{n}$, and surjective, i.e., $\{\chi_0, \chi_1, \dots, \chi_{n-1}\} = \widehat{\mathbf{Z}}_n$. Thus, $\widehat{\mathbf{Z}}_n \cong \mathbf{Z}_n$.*

The proof is left as an exercise. The next lemma tells us that also a direct sum decomposition carries over to the dual group:

Lemma 2.4. *Suppose $G = A \oplus B$, and let $\chi \in \widehat{A}$ and $\psi \in \widehat{B}$. Then $\chi \otimes \psi \in \widehat{G}$, where*

$$(\chi \otimes \psi)(a, b) = \chi(a)\psi(b).$$

Moreover, this map $(\chi, \psi) \mapsto \chi \otimes \psi$ defines an isomorphism $\widehat{A} \oplus \widehat{B} \cong \widehat{A \oplus B} = \widehat{G}$.

Proof. Given $\phi \in \widehat{A \oplus B}$, we can define $\phi|_A(a) := \phi(a, 0)$ and $\phi|_B(b) := \phi(0, b)$ for $a \in A$ and $b \in B$. It is easy to check that the maps $(\chi, \psi) \mapsto \chi \otimes \psi$ and $\phi \mapsto (\phi|_A, \phi|_B)$ are homomorphisms between $\widehat{A \oplus B}$ and $\widehat{A} \oplus \widehat{B}$ and inverses of each other. \square

Corollary 2.5. $G \cong \widehat{G}$.

Note that there is no natural isomorphism between G and \widehat{G} , because the direct sum decomposition is generally not unique, and even for cyclic groups the isomorphism that we constructed depends on the choice of the generator ω . A noteworthy exception is the group $G = \mathbf{Z}_2^n$, provided we consider the direct sum decomposition as given¹, because $\omega = -1$ is the only principal square root of unity. For future reference, we explicitly note the characters in this case:

¹To see that the direct sum decomposition is not unique, consider, for example, the group $G = \mathbf{Z}_2^2$ and the elements $a = (1, 0)$, $b = (0, 1)$ and $c = (1, 1)$. Then G is the direct sum of any two of the three cyclic subgroups generated by a , b , and c , respectively.

Corollary 2.6 (Characters of the Discrete Cube \mathbf{Z}_2^n). For $a, x \in \mathbf{Z}_2$, define $\chi_a(x) = (-1)^{ax}$. Then $\widehat{\mathbf{Z}}_2 = \{\chi_0, \chi_1\}$. For $x = (x_1, \dots, x_n)$ and $a = (a_1, \dots, a_n)$ in \mathbf{Z}_2^n , let

$$\chi_a(x) = (\chi_{a_1} \otimes \cdots \otimes \chi_{a_n})(x_1, \dots, x_n) = (-1)^{\langle a, x \rangle},$$

where $\langle a, x \rangle = \sum_i a_i x_i$ is the standard scalar product. Observe that the value of $(-1)^{\langle a, x \rangle}$ remains the same whether we consider $\langle a, x \rangle$ as an integer or whether we evaluate the sum modulo 2. Then $a \mapsto \chi_a$ defines an isomorphism $\mathbf{Z}_2^n \cong \widehat{\mathbf{Z}}_2^n$. (Note that in the case of the zero vector $0 \in \mathbf{Z}_2^n$, this is consistent with our earlier notation for the principal character $\chi_0 \equiv 1$.) In particular,

$$\chi_a \chi_b = \chi_{a \oplus b}$$

for all $a, b \in \mathbf{Z}_2^n$, where “ \oplus ” denotes addition modulo 2 (componentwise for vectors).

Note that there is a natural correspondence between vectors $x \in \mathbf{Z}_2^n$ and subsets $X \subseteq [n] = \{1, \dots, n\}$,

$$x \longleftrightarrow X = \{i : x_i = 1\}.$$

It is quite common, and sometimes convenient, to label the characters of \mathbf{Z}_2^n by such subsets, i.e.,

$$\chi_X(y) = \prod_{i \in X} (-1)^{y_i} = (-1)^{|X \cap Y|}.$$

Remark 2.7. While there is no natural isomorphism between G and \widehat{G} , there is a natural isomorphism

$$\begin{aligned} G &\cong \widehat{\widehat{G}} \\ a &\mapsto \tilde{a} : \widehat{G} \rightarrow \mathbf{C}^\times \\ &\tilde{a}(\chi) := \chi(a) \end{aligned}$$

(this is analogous to the situation for duals of finite dimensional vector spaces). The nontrivial part is to show injectivity of this mapping (surjectivity then follows since $|G| = |\widehat{G}| = |\widehat{\widehat{G}}|$, by Corollary 2.5), which boils down to showing the following “separation property”: For $a \neq 0$ in G , there exists a character $\chi \in \widehat{G}$ with $\chi(a) \neq 1$. This follows from the Second Orthogonality Relation for Characters (Corollary 2.10) below.

An inner product on \mathbf{C}^G and orthogonality relations for characters. We denote by \mathbf{C}^G the space of all functions $G \rightarrow \mathbf{C}$. For two such functions f, g , we define their *inner product*

$$\langle f, g \rangle := \mathbf{E}[f\bar{g}] = \int_G f\bar{g} = \frac{1}{|G|} \sum_{x \in G} f(x)\overline{g(x)}.$$

(Expectation and integral notation are with respect to the uniform measure on G .)

Lemma 2.8 (First Orthogonality relation for characters). *For characters χ, ψ of G ,*

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

In particular, for every nonprincipal character $\chi \neq \chi_0$ of G , we have $\sum_{a \in G} \chi(a) = 0$.

Proof. If $\chi = \psi$ then $\chi\bar{\psi} = \chi\chi^{-1} = \chi_0 \equiv 1$, i.e., $\mathbf{E}[\chi\bar{\psi}] = \mathbf{E}[1] = 1$. Otherwise, there exists $b \in G$ with $\chi(b) \neq \psi(b)$. Then, by linearity of expectation,

$$\underbrace{\chi(b)\bar{\psi(b)}}_{\neq 1} \langle \chi, \psi \rangle = \mathbf{E}_{a \in G} \left[\underbrace{\chi(a)\chi(b)}_{\chi(a+b)} \underbrace{\bar{\psi(a)}\bar{\psi(b)}}_{\bar{\psi(a+b)}} \right] \stackrel{c=a+b}{=} \mathbf{E}_{c \in G} [\chi(c)\bar{\psi(c)}] = \langle \chi, \psi \rangle,$$

hence $\langle \chi, \psi \rangle = 0$. □

Corollary 2.9. *Since $|G| = |\widehat{G}|$, it follows from the preceding lemma that the characters form an orthonormal basis of the space \mathbf{C}^G .*

In other words, if we write down a matrix $U \in \mathbf{C}^{G \times \widehat{G}}$, with rows indexed by the elements of G , with columns indexed by the characters in \widehat{G} , and with entries $u_{g,\chi} = \chi(g)/\sqrt{|G|}$, then U is a unitary matrix: the columns of U are pairwise orthogonal vectors in $\mathbf{C}^{|G|}$ (with the standard inner product). It follows that the rows of U are also pairwise orthogonal. Moreover, note that if we index the first row and column of U by the zero element of G and the principal character, respectively, then both of them are equal to the constant vector $(1/\sqrt{|G|}, \dots, 1/\sqrt{|G|})$.

Corollary 2.10 (Second Orthogonality Relation for Characters). *For $a, b \in G$,*

$$\sum_{\chi \in \widehat{G}} \chi(a)\bar{\chi(b)} = \begin{cases} 0 & , \text{ if } a \neq b \\ |G| & , \text{ if } a = b. \end{cases}$$

In particular, for $a \neq 0$, $\sum_{\chi \in \widehat{G}} \chi(a) = 0$.

2.2 The Fourier Transform

Let G be a finite abelian group. Since the characters in \widehat{G} form a basis for the space \mathbf{C}^G , every function $f : G \rightarrow \mathbf{C}$ has a unique representation as a linear combination of characters,

$$f = \sum_{\chi \in \widehat{G}} c_\chi \chi. \tag{2.1}$$

Since the values of the characters are n^{th} roots of unity, such a linear combination is often called a *trigonometric polynomial*. The coefficients c_χ are called the *Fourier coefficients* of f and denoted by $\widehat{f}(\chi)$. Thus, since the characters are pairwise orthogonal, we have

$$\widehat{f}(\chi) = \langle f, \chi \rangle = \int_G f \bar{\chi} = \frac{1}{|G|} \sum_{a \in G} f(a) \bar{\chi(a)}. \tag{2.2}$$

The function $\widehat{f} : \widehat{G} \rightarrow \mathbf{C}$ defined in this way is called the *Fourier transform* of f . Note that the map $f \mapsto \widehat{f}$ is linear, and that it commutes with complex conjugation: $\widehat{\overline{f}} = \overline{\widehat{f}}$. In the finite setting in which we work, we do not have to deal with issues of convergence that arise when studying Fourier transforms in a continuous setting, and the inverse of the Fourier transform is simply given by Equation (2.1). With our notation for the Fourier coefficients,

$$f = \sum_{\chi \in \widehat{G}} \widehat{f}(\chi) \chi. \quad (2.3)$$

This is called the *Fourier expansion* of the function f .

Consider the standard basis $\{e_a\}$ for the space \mathbf{C}^G , where $e_a(b) = 1$ if $a = b$ and $e_a(b) = 0$ otherwise. Note that $\langle e_a, e_b \rangle = 0$ for $a \neq b$ and that $\langle e_a, e_a \rangle = 1/n$, where $n := |G|$. Thus, $\{\sqrt{n}e_a : a \in G\}$ is another orthonormal basis of \mathbf{C}^G , and up to multiplication by \sqrt{n} , the Fourier expansion is simply a change of coordinates from one orthonormal basis to another. Alternatively, one can consider the Fourier transform as a linear transformation $\mathbf{C}^G \rightarrow \mathbf{C}^{\widehat{G}}$, $f \mapsto \widehat{f}$. The matrix for the linear transformation is the unitary matrix U described above. In particular, scalar products are preserved. This is the content of the following

Theorem 2.11 (Plancherel Formula). *For $f, g : G \rightarrow \mathbf{C}$,*

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

Proof. Consider the Fourier expansions $f = \sum_{\chi} \widehat{f}(\chi) \chi$ and $g = \sum_{\psi} \widehat{g}(\psi) \psi$. By bilinearity of the inner product,

$$\langle f, g \rangle = \langle \sum_{\chi} \widehat{f}(\chi) \chi, \sum_{\psi} \widehat{g}(\psi) \psi \rangle = \sum_{\chi, \psi} \widehat{f}(\chi) \overline{\widehat{g}(\psi)} \langle \chi, \psi \rangle = \sum_{\chi} \widehat{f}(\chi) \overline{\widehat{g}(\chi)}.$$

□

Corollary 2.12 (Parseval's Identity). *For $f : G \rightarrow \mathbf{C}$,*

$$\|f\|_2^2 = \mathbf{E}[|f|^2] = \sum_{\chi} |\widehat{f}(\chi)|^2$$

We remark that we will mostly be working with real-valued functions, for which we can forget about complex conjugation and also simply write f^2 instead of $|f|^2$. Note that in our finite setting, Parseval's Identity is just an instance of Pythagoras' Theorem: the squared length of a vector equals the sum of squares of its coordinates with respect to an orthonormal basis.

2.3 Convolutions

Definition 2.13 (Convolution of functions). Given $f, g : G \rightarrow \mathbf{C}$, we define their *convolution* $f * g : G \rightarrow \mathbf{C}$ by

$$(f * g)(x) := \mathbf{E}_y[f(x - y)g(y)] = \int_y f(x - y)g(y) = \frac{1}{|G|} \sum_{y \in G} f(x - y)g(y).$$

Translation. Given a function $f : G \rightarrow \mathbf{C}$ and a fixed $y \in G$, we define

$$f_y(x) := f(x + y),$$

the translation of f by y .

Lemma 2.14. Let $f, g, h : G \rightarrow \mathbf{C}$, $\chi \in \widehat{G}$, and $z \in G$.

1. *Commutativity of convolution:* $f * g = g * f$.
2. *Associativity of convolution:* $(f * g) * h = f * (g * h)$
3. *Convolution and translation commute:* $(f * g)_z = (f_z) * g = f * (g_z)$.
4. *Translation Fourier transforms to multiplication by “phase”:* $\widehat{f_z}(\chi) = \chi(z)\widehat{f}(\chi)$
5. *Convolution Fourier transforms to multiplication:* $\widehat{f * g} = \widehat{f}\widehat{g}$.

Proof. We prove the last statement and leave the other ones as exercises. For $\chi \in \widehat{G}$, we have

$$\begin{aligned} \widehat{f * g}(\chi) &= \langle f * g, \chi \rangle = \mathbf{E}_x \left[\underbrace{(f * g)(x)}_{\mathbf{E}_y[f(x-y)g(y)]} \cdot \underbrace{\overline{\chi}(x)}_{\overline{\chi}(x-y)\overline{\chi}(y)} \right] \\ &= \mathbf{E}_y \left[\underbrace{\mathbf{E}_x[f(x-y)\overline{\chi}(x-y)]}_{\stackrel{z=x-y}{=} \mathbf{E}_z[f(z)\overline{\chi}(z)] = \widehat{f}(\chi)} g(y)\overline{\chi}(y) \right] \quad \left(\begin{array}{l} \text{exchange order of} \\ \text{summation/integration} \end{array} \right) \\ &= \widehat{f}(\chi)\mathbf{E}_y[g(y)\overline{\chi}(y)] = \widehat{f}(\chi)\widehat{g}(\chi). \end{aligned}$$

□

Chapter 3

Boolean Functions

A *boolean function* is a function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, for some natural number n which is called the *dimension* or *number of variables*. The set $\{0, 1\}^n$ is called the *discrete cube of dimension n* . If we consider the set $\{0, 1\}$ with addition modulo 2, which we denote by ‘ \oplus ’, we obtain the cyclic group \mathbf{Z}_2 of order 2; if on top of that, we also consider \mathbf{Z}_2 to be equipped with (the usual) multiplication, we obtain the prime field \mathbf{F}_2 of order 2.

When applying Fourier-analytic methods to the study of boolean functions, we view the domain of definition $\{0, 1\}^n$ as the abelian group \mathbf{Z}_2^n and the range $\{0, 1\}$ as a subset of the real numbers \mathbf{R} (or the complex numbers \mathbf{C}).

3.1 Examples and Various Viewpoints

There are several equivalent ways of viewing or interpreting boolean functions:

1. **Subsets of $\{0, 1\}^n$.** Given a set M , we can interpret any function $f : M \rightarrow \{0, 1\}$ as the indicator function of a subset of X , which yields the correspondence

$$f \longleftrightarrow \{x \in M : f(x) = 1\}.$$

If the ground set M is clear from the context, we denote the indicator function of a subset $A \subseteq M$ by $\mathbf{1}_A$. Thus, $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ if $x \notin A$. In particular, boolean functions are in one-to-one correspondence with subsets of the discrete cube.

2. **Boolean Formulas.** A boolean formula is built from a finite set V of variables using the basic binary operators \wedge (‘and’, *conjunction*), \vee (‘or’, *disjunction*) and the unitary operator \neg (‘not’, *negation*),¹ e.g.,

$$(\neg x_1 \vee x_2) \wedge (\neg x_2 \vee x_1). \tag{3.1}$$

By interpreting 0 as ‘false’ and 1 as ‘true’ and numbering the variables as x_1, \dots, x_n , a boolean formula corresponds to a boolean function. For instance, the formula

¹One of the two binary operators would be enough, since either one of them can be expressed through the other and negation, e.g., $x \vee y \equiv \neg(\neg x \wedge \neg y)$.

(3.1) corresponds to the function $f : \{0, 1\}^2 \rightarrow \{0, 1\}$ given by $f(0, 0) = f(1, 1) = 1$ and $f(0, 1) = f(1, 0) = 0$. To see this systematically, we can *arithmetize* a formula by replacing each conjunction $x \wedge y$ by a multiplication xy , and each negation $\neg x$ by $1 - x = 1 \oplus x$.² If we do this for the formula (3.1), we see that we obtain the function

$$\begin{aligned} f(x_1, x_2) &= (1 \oplus (1 \oplus (1 \oplus x_1)) \cdot (1 \oplus x_2)) \cdot (1 \oplus (1 \oplus x_1) \cdot (1 \oplus (1 \oplus x_2))) \\ &= (1 \oplus x_1(1 \oplus x_2)) \cdot (1 \oplus (1 \oplus x_1)x_2) \\ &= 1 \oplus x_1(1 \oplus x_2) \oplus (1 \oplus x_1)x_2 + \underbrace{x_1x_2(1 \oplus x_1)(1 \oplus x_2)}_{\equiv 0 \text{ since } x_1, x_2 \in \{0, 1\}} \\ &= 1 \oplus x_1 \oplus x_2, \end{aligned}$$

which is exactly the function described before. Note that the operation ‘ \oplus ’ of addition modulo 2 in the arithmetization corresponds to the binary boolean operator ‘xor’ (exclusive disjunction).

Conversely, any boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ can be represented by a boolean formula, e.g. by

$$\bigvee_{\substack{a \in \{0, 1\}^n \\ f(a) = 1}} \left(\bigwedge_{\substack{i \in [n] \\ a_i = 1}} x_i \wedge \bigwedge_{\substack{i \in [n] \\ a_i = 0}} \neg x_i \right).$$

3. Voting Schemes. Consider a population of n individuals deciding about an issue or public office for which there are two choices or candidates (for all practical purposes, the presidential elections in the USA can serve as an example, because there are usually only two relevant candidates).³ Any method of arriving at a decision for the whole population based on the individual votes is called a *voting scheme*. If we label the two possible choices by 0 and 1, then voting schemes are nothing else but boolean functions.

For a voting scheme it should arguably play no role which of the two outcomes we label ‘0’ and which ‘1’. Thus, the corresponding boolean function should be *antipodal*, i.e., $f(\mathbb{1} \oplus x) = 1 \oplus f(x)$ for all $x \in \{0, 1\}^n$, where $\mathbb{1} = (1, \dots, 1)$.

²As remarked above, this takes also care of disjunctions.

³If there are more than two possible choices, the study of voting mechanisms becomes even more interesting. One natural approach is to use the notion of “rankings” or “preference lists” for each voter, which is simply a linear ordering (according to decreasing preference, say) of the possible choices. In this setting, the following paradoxical situation can arise: Suppose that there are three possible choices, A , B , and C , and three voters with the preference lists (A, B, C) , (B, C, A) , and (C, A, B) , respectively. In this situation, whichever candidate X we chose, there will be a 2/3-majority of the population that would have preferred some other candidate Y . This is called *Condorcet’s Paradox*, discovered by the Marquis de Condorcet in the 18th century. Moreover, this is no accident: There is a result called *Arrow’s Theorem* which asserts that as soon as there are three or more candidates, any voting system based on ranked preferences and satisfying some fairly reasonable and weak assumptions will produce similar paradoxes, except for a dictatorship in which one individual alone decides.

Another property, which is implied by antipodality but strictly weaker, is that of being balanced: a boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is called *balanced* if $\Pr[f = 1] = \Pr[f = 0]$ (with respect to the uniform probability distribution on the discrete cube), i.e., if $\mathbf{E}[f] = \Pr[f = 1] = 1/2$.

Examples 3.1. The names of the first few examples are motivated by the voting scheme viewpoint.

1. “*Dictatorship*”. For any fixed k , $1 \leq k \leq n$, the *projection* onto the k^{th} variable, $x = (x_1, \dots, x_n) \mapsto x_k$, is also often called a *dictatorship*. A dictatorship function is antipodal and hence balanced.

2. “*Majority*”. The function

$$\text{maj}(x_1, \dots, x_n) := \begin{cases} 1 & , \text{ if } |\{i : x_i = 1\}| > n/2 \\ 0 & , \text{ otherwise} \end{cases}$$

is called the *majority function* (more precisely, one could speak of “*strict majority*”). The majority function is perfectly antipodal if n is odd. If n is even, the “middle level” $\{x \in \{0, 1\}^n : |x| = n/2\}$ violates this antipodality. However, by Stirling’s Formula, $\Pr_x[|x| = n/2] \sim \sqrt{2/(\pi n)}$, so the effect of the middle level becomes negligible if n is large.

3. “*Majority of Majorities*”. If we fix a partition of the variables, e.g., $\{x_1, \dots, x_n\} = \{x_1, \dots, x_{n_1}\} \cup \{x_{n_1+1}, \dots, x_{n_2}\} \cup \dots \cup \{x_{n_{r-1}+1}, \dots, x_n\}$, the function

$$\text{maj}(\text{maj}(x_1, \dots, x_{n_1}), \text{maj}(x_{n_1+1}, \dots, x_{n_2}), \dots, \text{maj}(x_{n_{r-1}+1}, \dots, x_n))$$

is called a *majority of majorities* function.

4. “*Juntas*”. We are often interested not in a single boolean function, but in a whole infinite family $\{f_n : \{0, 1\}^n \rightarrow \{0, 1\}, n \in \mathbf{N}\}$ of boolean functions, usually in their “asymptotic behaviour” (in some suitable sense) as $n \rightarrow \infty$. If each function f_n depends only on a bounded number of variables, we call it a *junta*. For example, $f(x_1, \dots, x_n) := 1 \oplus (1 \oplus x_1)x_2 \oplus x_3$ is a junta.

5. “*Tribes*”. This function will play a very important role in what follows, in particular in the study of *influences* (Chapter 4). Let $b|n$ be a parameter, which we will determine later. Consider a partition $[n] = B_1 \cup \dots \cup B_{n/b}$ into n/b disjoint parts, called “*tribes*”, of size b each. Using the notation for boolean formulas, we define the tribes function as

$$\text{tribes}(x_1, \dots, x_n) = \bigvee_{j=1}^{n/b} \bigwedge_{i \in B_j} x_i.$$

In other words, $\text{tribes}(x) = 1$ iff there exists some tribe B_j such that all “members” of that tribe unanimously “vote” 1.

Now we set the parameter b so as to make the tribes function balanced. The probability that at least one member of a given tribe votes 0 is $1 - 2^{-b}$, so

$$\Pr[\text{tribes} = 0] = (1 - 2^{-b})^{\frac{n}{b}}.$$

We define b as the number that makes this probability equal to $1/2$ (we ignore the small error incurred by the fact that the exact b may not be an integer or not a divisor of n). Using the estimate $(1 - 2^{-b})^{\frac{n}{b}} \sim e^{-2^{-b}n/b}$ and solving for b , we see that $b + \log_2 b = \log_2 n - \log_2(\ln 2 + o(1))$, hence $b = \log_2 n - \log_2 \log_2 n + O(1)$.

6. “Parity”. We define

$$\text{parity}(x_1, \dots, x_n) = x_1 \oplus \dots \oplus x_n.$$

The parity function is always perfectly balanced. If n is odd, it is even antipodal.

3.1.1 Fourier Transforms of Boolean Functions

Recall that by Corollary 2.6, the characters of the discrete cube are the functions given by

$$\chi_a(x) = (-1)^{\langle a, x \rangle}$$

for $a, x \in \{0, 1\}^n$. We will often simply write $\widehat{f}(a)$ instead of $\widehat{f}(\chi_a)$. Sometimes it will also be convenient to use the correspondence $a \leftrightarrow A = \{i : a_i = 1\}$ and to denote the Fourier coefficients by $\widehat{f}(A)$, $A \subseteq [n]$. Moreover, for all $a, b \in \{0, 1\}^n$,

$$\chi_a \chi_b = \chi_{a \oplus b}.$$

The Fourier expansion of a function $f : \{0, 1\}^n \rightarrow \mathbf{C}$ is given by

$$\widehat{f}(a) = \int_x f(x) \chi_a(x) dx = \frac{1}{2^n} \sum_{x \in \{0, 1\}^n} f(x) (-1)^{\langle a, x \rangle}.$$

Next, we will compute the Fourier transforms of some examples of boolean functions mentioned above. The characters of the discrete cube are exactly the homomorphisms from \mathbf{Z}_2^n (with componentwise addition modulo 2) into the multiplicative group $\{+1, -1\}$, and the latter is isomorphic to the additive group $\mathbf{Z}_2 = \{0, 1\}$ via

$$\begin{aligned} \{0, 1\} &\leftrightarrow \{+1, -1\} \\ x &\mapsto (-1)^x = 1 - 2x \\ \frac{1-z}{2} &\leftrightarrow z \end{aligned}$$

Using this correspondence, it is sometimes easier to recognize functions directly as simple linear combinations of characters, rather than computing their Fourier transform using the above formula.

Examples 3.2. 1. *Dictatorships.* Consider the dictatorship $\pi_k(x_1, \dots, x_n) := x_k$. Moreover, let $e_k = (0, \dots, 0, 1, 0, \dots, 0)$ be the k^{th} standard basis vector of $\{0, 1\}^n$. Then $\chi_{e_k}(x) = (-1)^{\langle x, e_k \rangle} = (-1)^{x_k} = 1 - 2x_k$. Thus, $\pi_k = \frac{1}{2}(\underbrace{1}_{=\chi_0} - \chi_{e_k})$, in other words

$$\widehat{\pi}_k(a) = \begin{cases} 1/2 & , \text{ if } a = 0 \\ -1/2 & , \text{ if } a = e_k \\ 0 & , \text{ otherwise.} \end{cases}$$

2. *Parity.* Observe that $\chi_{\mathbb{1}} = (-1)^{\text{parity}} = 1 - 2 \cdot \text{parity}$, hence $\text{parity} = \frac{1}{2}(\chi_0 - \chi_{\mathbb{1}})$.
3. *Global 'and'.* Consider the function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, $f(x) = \prod_{i=1}^n x_i$. We know that we can express $x_i = \frac{1}{2}(\chi_0(x) - \chi_{e_i}(x))$. Thus,

$$f = \prod_{i=1}^n \frac{1}{2}(\chi_0 - \chi_{e_i}) = \sum_{a \in \{0, 1\}^n} \underbrace{\frac{(-1)^{|a|}}{2^n}}_{=\widehat{f}(a)} \chi_a,$$

where we use the fact that $\chi_a \chi_b = \chi_{a \oplus b}$.

4. *Majority and its relatives.* Fix a parameter k , $0 \leq k \leq n$, and consider the indicator function $f := \mathbf{1}_{\{|x| \leq k\}}$ given by $f(x) = 1$ iff $|x| \leq k$ and $f(x) = 0$ otherwise. (In particular, if $k = \lceil n/2 \rceil - 1$, then $\text{maj}(x) = 1 - f(x)$.) First, observe that f is symmetric under permuting the variables. Thus, $\widehat{f}(a)$ depends only on $|a|$ (check this!). We have $\widehat{f}(\chi_0) = \mathbf{E}[f] = \frac{1}{2^n} \sum_{j=0}^k \binom{n}{j}$. Next, $\widehat{f}(a) = \widehat{f}(e_1)$ whenever $|a| = 1$, and

$$2^n \widehat{f}(e_1) = \sum_{|x| \leq k} (-1)^{\langle e_1, x \rangle} = \sum_{\substack{|x| \leq k \\ x_1=0}} 1 - \sum_{\substack{|x| \leq k \\ x_1=1}} 1 = \sum_{j=0}^k \binom{n-1}{j} - \binom{n-1}{j-1} = \binom{n-1}{k}.$$

Similarly, $\widehat{f}(a) = \widehat{f}(e_1 \oplus e_2)$ if $|a| = 2$, and

$$2^n \widehat{f}(e_1 \oplus e_2) = \underbrace{\sum_{j=0}^k \binom{n-2}{j} - \binom{n-2}{j-1}}_{\binom{n-2}{k}} - \underbrace{\sum_{j=0}^k \binom{n-2}{j-1} - \binom{n-2}{j-2}}_{\binom{n-2}{k-1}} = \binom{n-2}{k} - \binom{n-2}{k-1}.$$

More generally, by the same type of argument, one can show that for $1 \leq m \leq n$,

$$2^n \widehat{f}(e_1 \oplus \dots \oplus e_m) = \sum_{i=0}^{m-1} (-1)^i \binom{m-1}{i} \binom{n-m}{k-i}.$$

3.2 Linearity Testing

In this section, we describe a first algorithmic application. A function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is called *linear* if it is linear over the field \mathbf{F}_2 . Since the only scalars are 0 and 1, this boils down to f being a homomorphism of groups, i.e., $f(x \oplus y) = f(x) \oplus f(y)$ for all $x, y \in \{0, 1\}^n$ (where addition is modulo 2, and componentwise for vectors). If we switch from additive to multiplicative notation, i.e., if we replace f by $h = (-1)^f = 1 - 2f : \{0, 1\}^n \rightarrow \{-1, +1\}$, $h(x) = (-1)^{f(x)} = 1 - 2f(x)$, then linearity of f corresponds to $h(x + y) = h(x)h(y)$ for all $x, y \in \mathbf{Z}_2^n$, i.e., f is linear iff $(-1)^f$ is a character.

We define \mathcal{L}_n (or simply \mathcal{L} , when n is understood from the context) as the set of linear functions $\{0, 1\}^n \rightarrow \{0, 1\}$. (Note that \mathcal{L}_n is just the dual vector space $(\mathbf{F}_2^n)^*$.) The “dictatorship” functions $\pi_k(x_1, \dots, x_n) = x_k$, $1 \leq k \leq n$, form a basis (over \mathbf{F}_2) for \mathcal{L}_n . (This is the dual basis to the standard basis $\{e_k : 1 \leq k \leq n\}$ of \mathbf{F}_2^n .) Thus, the linear functions are precisely those that can be obtained as a \mathbf{F}_2 -linear combination of dictatorships: $f \in \mathcal{L}$ iff there exists $A \subseteq [n]$ with $f = \bigoplus_{k \in A} \pi_k$, i.e., $f(x_1, \dots, x_n) = \bigoplus_{k \in A} x_k$. Equivalently this is the case iff $(-1)^f = \chi_a$, where $A = \{i : a_i = 1\}$.

For functions f, g defined on $\{0, 1\}^n$ (without specifying where exactly they take their values), we define

$$\text{dist}(f, g) := \Pr_x[f(x) \neq g(x)],$$

where $x \in \{0, 1\}^n$ uniformly at random, and $\text{dist}(f, \mathcal{L}) := \min_{g \in \mathcal{L}} \text{dist}(f, g)$. Note that for functions with values in $\{0, 1\}$, we have $\text{dist}(f, g) = \|f - g\|_1$. Moreover, with the notation used in the preceding paragraph, $\text{dist}(f, \bigoplus_{k \in A} \pi_k) = \text{dist}((-1)^f, \chi_a)$.

Theorem 3.3. *For all $\varepsilon, \delta > 0$, there exists a randomized algorithm \mathcal{A} with the following specifications. Given a function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ (where we assume that the function is given implicitly, by an “oracle”), the algorithm requires $O(\frac{1}{\delta} \log(\frac{1}{\varepsilon}))$ many evaluations of f and afterwards accepts or rejects f , and we are guaranteed that*

1. $f \in \mathcal{L} \Rightarrow \Pr[\mathcal{A} \text{ accepts } f] = 1$,
2. $\text{dist}(f, \mathcal{L}) \geq \delta \Rightarrow \Pr[\mathcal{A} \text{ rejects } f] \geq 1 - \varepsilon$.

The algorithm is very simple to describe: Repeat the following test $\frac{1}{\delta} \log(\frac{1}{\varepsilon})$ times (with independent random choices): Choose $x, y \in \{0, 1\}^n$ independently and uniformly at random. Test if $f(x) \oplus f(y) = f(x \oplus y)$. In the end, accept f if all the tests succeed, otherwise reject. A linear function clearly passes all tests. On the other hand, if $\text{dist}(f, \mathcal{L}) \geq \delta$, then by the following lemma, the probability that at least one test fails is at least $1 - (1 - \delta)^{\frac{1}{\delta} \log(\frac{1}{\varepsilon})} \geq 1 - e^{-\delta^{\frac{1}{\delta}} \log(\frac{1}{\varepsilon})} = 1 - \varepsilon$, which implies Theorem 3.3. Note that we analyzed the number of calls to the “oracle”, not the running time of the algorithm, which will depend on n (for instance, computing $x \oplus y$ requires time $O(n)$).

Lemma 3.4. *Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$. Then the probability that a single test as described above fails is at least $\text{dist}(f, \mathcal{L})$.*

Proof. We pass from additive to multiplicative notation, i.e., we consider $h = (-1)^f$. Consider a character χ_a , $a \in \{0, 1\}^n$. We have

$$\begin{aligned} \text{dist}(h, \chi_a) &= \Pr_x[h(x) \neq \chi_a(x)] \\ &= \mathbf{E}_x\left[\frac{1 - h(x)\chi_a(x)}{2}\right] \\ &= \frac{1}{2}(1 - \mathbf{E}_x[h(x)\chi_a(x)]) = \frac{1}{2}(1 - \hat{h}(a)). \end{aligned}$$

Thus, $\text{dist}(f, \mathcal{L}) = \min_a \text{dist}(h, \chi_a) = \frac{1}{2}(1 - \max_a \hat{h}(a))$. On the other hand, a single test involving x and y succeeds iff $f(x) \oplus f(y) \oplus f(x \oplus y) = 0$, i.e., iff $h(x)h(y)h(x \oplus y) = 1$, and it fails iff $h(x)h(y)h(x \oplus y) = -1$. Thus, the probability of a failed test equals

$$\mathbf{E}_{x,y}\left[\frac{1 - h(x)h(y)h(x \oplus y)}{2}\right] = \frac{1}{2}(1 - \mathbf{E}_{x,y}[h(x)h(y)h(x \oplus y)]),$$

and

$$\begin{aligned} \mathbf{E}_{x,y}[h(x)h(y)h(x \oplus y)] &= \mathbf{E}_x[h(x) \underbrace{\mathbf{E}_y[h(y)h(x \oplus y)]}_{(h * h)(x)}] \\ &= \langle h, h * h \rangle = \sum_{a \in \{0,1\}^n} \hat{h}(a) \cdot \widehat{h * h}(a) = \sum_a \hat{h}(a) \cdot (\hat{h}(a))^2 \\ &\leq \max_a \hat{h}(a) \cdot \sum_a (\hat{h}(a))^2 = \max_a \hat{h}(a) \cdot \|h\|_2^2 = \max_a \hat{h}(a). \end{aligned}$$

□

Chapter 4

Influences

Definition 4.1 (Influences). Let μ be a probability measure on the discrete cube $\{0, 1\}^n$, and let $f : \{0, 1\}^n \rightarrow \mathbf{C}$. For $1 \leq k \leq n$, the *influence of the k -th variable on f* is defined as the probability that the value of f at a random point $x \sim \mu$ changes if we flip the k^{th} coordinate,

$$I_k^\mu(f) := \Pr_{x \sim \mu} [f(x) \neq f(x \oplus e_k)],$$

where $e_k = (0, \dots, 0, 1, 0, \dots, 0)$ is the k^{th} standard basis vector (the 1 is in the k^{th} position). The *total influence* is defined as the sum

$$I^\mu(f) := \sum_{k=1}^n I_k^\mu(f).$$

In this chapter, we will focus on the case that $\mu = \mu_{\frac{1}{2}}$ is the uniform probability measure on the discrete cube and to simplify notation, we will drop the superscript and simply write $I_k(f)$ and $I(f)$. Later on (in Chapter 6), we will also consider more general product measures, in particular the measures μ_p , $0 \leq p \leq 1$.

Examples 4.2. 1. *Dictatorships.* Consider the “dictatorship” $\pi_1(x_1, \dots, x_n) = x_1$. Then

$$I_k(\pi_1) = \begin{cases} 1 & , \text{ if } k = 1, \\ 0 & , \text{ if } k \neq 1. \end{cases}$$

Consequently, $I(\pi_1) = 1$.

2. *Majority.* Note that for any boolean function f ,

$$I_k(f) = 2 \Pr[f(x) = 0 \text{ and } f(x \oplus e_k) = 1].$$

In the case of the majority function,

$$\text{maj}(x) = 0 \text{ and } \text{maj}(x \oplus e_k) = 1 \iff |x| = \lfloor n/2 \rfloor \text{ and } x_k = 0.$$

The probability for this to happen equals $\binom{n-1}{\lfloor n/2 \rfloor} / 2^n \sim \sqrt{2/(\pi n)}$. Thus,

$$I_k(\text{maj}) = \Theta(1/\sqrt{n}) \quad \text{and} \quad I(\text{maj}) = \Theta(\sqrt{n}).$$

3. *Parity*. For the parity function, $I_k(\text{parity}) = 1$ for all k , and so $I(\text{parity}) = n$.
4. *Tribes*. Recall that the tribes function is defined as follows. We partition the set of variables into n/b “tribes” of size b each,

$$[n] = B_1 \cup \dots \cup B_{n/b}, \quad |B_j| = b \text{ for all } j,$$

where b is a parameter. Then the tribes function is given by

$$f(x_1, \dots, x_n) = \bigvee_{j=1}^{n/b} \bigwedge_{i \in B_j} x_i.$$

The parameter b is chosen such that $\mathbf{E}[f] = \Pr[f = 1] = \Pr[f = 0] = 1/2$, i.e., $1/2 = (1 - 2^{-b})^{n/b} \sim e^{-2^{-b} \frac{n}{b}}$. Thus,

$$b = \log_2 n - \log_2 \log_2 n + O(1).$$

To determine the influences $I_k(f)$, consider the tribe that x_k belongs to, say $k \in B_1$. Then flipping the variable x_k affects the value of $f(x)$ iff for each “tribe” B_j with $j \neq 1$, $x_i = 0$ for at least one $i \in B_j$, and $x_i = 1$ for all $i \in B_1 \setminus \{k\}$. Thus,

$$I_k(f) = (1 - 2^{-b})^{n/b-1} \cdot 2^{-b+1} = \frac{\frac{1}{2} \cdot 2 \cdot 2^{-b}}{1 - 2^{-b}} = \frac{1}{2^b - 1} = \Theta\left(\frac{\log n}{n}\right)$$

for $1 \leq k \leq n$. As we will see below (Corollary 4.17), this is the minimum for any balanced boolean function: any such function has at least one variable with influence $\Omega(\log(n)/n)$.

4.1 Influences and the Edge-Isoperimetric Inequality

The *Hamming cube* is the graph with vertex set $V = \{0, 1\}^n$, and two vertices are joined by an edge if they have *Hamming distance* 1, i.e., if they differ in exactly one component. Formally,

$$\text{dist}_H(v, w) := |\{i : v_i \neq w_i\}| = \underbrace{\|v - w\|_1}_{=\sum_{i=1}^n |v_i - w_i|}$$

and

$$E := \{\{v, w\} \in \binom{V}{2} : \text{dist}_H(v, w) = 1\}.$$

In other words, the neighbors of v are exactly the vertices $v \oplus e_k$, $k = 1, \dots, n$.

For a subset $A \subseteq V$, its *edge boundary* is defined as the set of edges

$$\partial_E A := E[A, V \setminus A],$$

where generally, for disjoint sets A and B , $E[A, B]$ is the set of edges that have one endpoint in A and the other one in B .

Observation 4.3. *If $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is the indicator function of A , i.e., $A = \{x \in \{0, 1\}^n : f(x) = 1\}$, then the total influence of f and the size of the edge boundary of A are the same up to a factor that depends only on n :*

$$I(f) = \sum_k I_k(f) = \frac{2}{2^n} |\partial_E A|.$$

(To see this, note that $I_k(f) = 2 \Pr_x[x \in A \text{ and } x \oplus e_k \notin A]$.)

Next, we note that the degree of every vertex in the whole Hamming graph equals n . Thus,

$$n|A| = |\partial_E A| + 2|E[A]|, \tag{4.1}$$

where $E[A]$ is the set of edges both of whose endpoints lie in A . An ‘‘isoperimetric inequality’’ gives a lower bound for the size of the boundary of a set in terms of the size of the set.¹ By the previous equation, for the edge boundary in the Hamming cube, this is equivalent to bounding the size of $|E[A]|$ from above.

Theorem 4.4 (Edge-Isoperimetric Inequality for the Hamming Cube). *Let $A \subseteq V$. Then*

$$|E[A]| \leq \frac{1}{2} |A| \log_2 |A|$$

(with the understanding that $0 \log 0 = 0$). Equivalently,

$$|\partial_E A| \geq |A|(n - \log_2 |A|).$$

Equality is attained iff A is the vertex set of a subcube (which means that there is a partition $[n] = X \dot{\cup} Y \dot{\cup} Z$ of the index set such that $A = \{0\}^X \times \{1\}^Y \times \{0, 1\}^Z$).

Proof. By induction on n . The base case $n = 0$ is clear. For $n > 0$, we split A into two subsets according to the last coordinate: For $i \in \{0, 1\}$, define $A_i := \{v \in A : v_n = i\}$. Then

$$E[A] = E[A_0] \dot{\cup} E[A_1] \dot{\cup} E[A_0, A_1].$$

We have $E[A_0, A_1] \leq \min\{|A_0|, |A_1|\}$. Moreover, the sets A_i , $i = 0, 1$, live in cubes of dimension $n - 1$, so by induction

$$|E[A_i]| \leq \frac{1}{2} |A_i| \log_2 |A_i|.$$

Thus,

$$|E[A]| \leq \frac{1}{2} (|A_0| \log_2 |A_0| + |A_1| \log_2 |A_1|) + \min\{|A_0|, |A_1|\}. \tag{4.2}$$

We note that without the factor of $\frac{1}{2}$, the estimate would immediately follow by induction.² However, we are aiming for the exact constant and claim that the right-hand side

¹For instance, in the classical isoperimetric inequality in the plane, the size of a set is its area, and the size of the boundary is the circumference.

²If $|A_0| \leq |A_1|$, say, then $|A_0| \log_2 |A_0| \leq |A_0|(\log_2 |A| - 1)$, which absorbs the additional $|A_0|$ term.

of (4.2) is bounded from above by $\frac{1}{2}|A|\log_2|A|$. This claim essentially follows from the fact that the entropy function $H(x) := x\log x + (1-x)\log(1-x)$ is convex (where x represents the fraction $\min\{|A_0|, |A_1|\}/|A| \leq 1/2$).³ Therefore, the right-hand-side of (4.2) attains its maximum at $x = 0$ or $x = 1/2$, and in both cases, it equals the desired bound.

It is easy to verify by induction that equality is attained in all our estimates iff A is a subcube. \square

By the correspondence between boolean functions and subsets of $\{0, 1\}^n$, the following is another equivalent reformulation of the edge-isoperimetric inequality:

Corollary 4.5. *Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$, and let $p = \mathbf{E}[f] = \Pr[f = 1]$. Then*

$$I(f) \geq 2p \log_2(1/p).$$

For instance, if f is balanced, i.e., $\mathbf{E}[f] = 1/2$, we get $I(f) \geq 1$. Below, we will see an alternative way of arriving at this conclusion using Fourier Analysis.

Remark 4.6. There is also a notion of a *vertex boundary* and a corresponding isoperimetric inequality: For $A \subseteq V = \{0, 1\}^n$, define the vertex boundary as the set of vertices at Hamming distance exactly 1,

$$\partial_V(A) := \{v \in V \setminus A : \text{dist}_H(v, A) = 1\},$$

where $\text{dist}_H(v, A) := \min\{\text{dist}_H(v, w) : w \in A\}$. Then the *Vertex-Isoperimetric Inequality* for the Hamming cube states that for all sets A of a given size, $|\partial_V(A)|$ is minimized if A is (approximately) a *Hamming ball*, i.e., if there exist $c \in V$ and a radius $0 \leq r \leq n$ such that $B_r(c) \subseteq A \subseteq B_{r+1}(c)$, where $B_r(c) := \{v \in V : \text{dist}(c, v) \leq r\}$ is the Hamming ball of radius r centered at c (for a proof, see for instance, Chapter 16 of Béla Bollobás, *Combinatorics*, Cambridge University Press, 1986).

4.2 Influences and the Fourier Transform

All probabilities (expectations, variances, influences) in this section will be with respect to the uniform measure $\mu_{\frac{1}{2}}$ on the discrete cube.

Our goal is to show that for any boolean function, there always exists a variable with “large influence”, where “large” means of the order $\text{Var}[f] \cdot \log(n)/n$. This is known

³Here are the details: Setting $a := |A|$ and $x := \min\{|A_0|, |A_1|\}/|A| \in [0, 1/2]$, we can rewrite the right-hand-side of (4.2) as

$$\frac{1}{2}(xa \log_2(xa) + (1-x)a \log_2((1-x)a)) + xa = a \left(\frac{\log_2(a)}{2} + \underbrace{\frac{1}{2}[x \log_2(x) + (1-x) \log_2(1-x)] + x}_{(*)} \right)$$

The nonconstant part $(*)$ is a convex function of x on the interval $[0, 1/2]$: The first derivative is $\frac{1}{2\ln 2}(\ln x - \ln(1-x)) + 1$, and the second derivative is $\frac{1}{2\ln 2}(1/x + 1/(1-x)) > 0$.

as the *KKL Theorem* (after the authors' last names: Kahn, Kalai, and Linial). As the “tribes” example shows, this bound is best possible.

We can relate influences to the Fourier transform by means of the following auxiliary functions: For $f : \{0, 1\}^n \rightarrow \{0, 1\}$ and $1 \leq k \leq n$, we define

$$\partial_k f := f - \tau_{e_k} f : \{0, 1\}^n \rightarrow \{-1, 0, 1\},$$

in other words,

$$\partial_k f(x) := f(x) - f(x \oplus e_k) = \begin{cases} -1 & , \text{ if } f(x) = 0 \text{ and } f(x \oplus e_k) = 1 \\ 0 & , \text{ if } f(x) = f(x \oplus e_k) \\ 1 & , \text{ if } f(x) = 1 \text{ and } f(x \oplus e_k) = 0. \end{cases} \quad (4.3)$$

Observation 4.7.

$$I_k(f) = \Pr[\partial_k f \neq 0] = \mathbf{E}[(\partial_k f)^2] = \|\partial_k f\|_2^2.$$

By Parseval's Identity that we can express the square norm of a function in terms of its Fourier coefficients. By the definition of $\partial_k f = f - \tau_{e_k} f$, Part 4 of Lemma 2.14 on the Fourier transform and translations immediately implies $\widehat{\partial_k f}(a) = \widehat{f}(a)(1 - \chi_a(e_k))$. Therefore:

Lemma 4.8. For $a \in \{0, 1\}^n$,

$$\widehat{\partial_k f}(a) = \begin{cases} 0 & , \text{ if } a_k = 0, \\ 2\widehat{f}(a) & , \text{ if } a_k = 1. \end{cases}$$

In other words,

$$\partial_k f = 2 \sum_{\substack{a \in \{0, 1\}^n \\ a_k = 1}} \widehat{f}(a) \chi_a.$$

Corollary 4.9. For $1 \leq k \leq n$,

$$I_k(f) = 4 \sum_{a: a_k=1} \widehat{f}(a)^2,$$

and therefore

$$I(f) = 4 \sum_{a \in \{0, 1\}^n} \widehat{f}(a)^2 |a|.$$

We also need some general facts about the expectation and variance of functions on the discrete cube:

Observation 4.10. Let $f : \{0, 1\}^n \rightarrow \mathbf{R}$.

1. $\mathbf{E}[f] = \widehat{f}(0)$.

2. The variance of f is defined by

$$\text{Var}[f] := \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

By Parseval's Identity and Part 1,

$$\text{Var}[f] = \sum_{\substack{a \in \{0,1\}^n \\ a \neq 0}} \widehat{f}(a)^2.$$

3. If f is boolean, i.e., takes only values in $\{0, 1\}$ then $f^2 = f$. Thus,

$$\text{Var}[f] = \mathbf{E}[f](1 - \mathbf{E}[f]).$$

For the rest of this section, let $f : \{0, 1\}^n \rightarrow \{0, 1\}$ be a boolean function, let $I := I(f)$ be the total influence of f , and let $v := \text{Var}[f]$ be the variance of f . A simple averaging argument implies that the sum of the “large” squared Fourier coefficients of f is “small” compared to the total influence:

Lemma 4.11. For any $t \geq 0$, $\sum_{|a| > t} \widehat{f}(a)^2 \leq \frac{I}{4t}$.

Proof. Let S be the sum we want to bound. Then $4tS \leq 4 \sum_{|a| > t} \widehat{f}(a)^2 |a| \leq I$. \square

As an immediate corollary (setting $t = 0$), we get the following bound (which we previously derived from the edge-isoperimetric inequality):

Corollary 4.12. For a boolean function f with $\mathbf{E}[f] = 1/2$, the total influence $I(f)$ is at least 1.

Corollary 4.13. Setting $t = \frac{I}{2v}$, we obtain $\sum_{|a| > \frac{I}{2v}} \widehat{f}(a)^2 < v/2$. Since $v = \sum_{a \neq 0} \widehat{f}(a)^2$,

$$\sum_{0 < |a| \leq \frac{I}{2v}} \widehat{f}^2(a) \geq \frac{v}{2}.$$

Up to one technical ingredient, which we will introduce below when it is needed and prove in the next chapter, we are now ready to prove the main technical lemma:

Lemma 4.14. For a suitable parameter $0 < \xi < 1$,

$$\sum_{k=1}^n I_k(f)^{\frac{2}{1+\xi^2}} \geq I \cdot \xi^{I/v}.$$

We will prove this lemma with $\xi = 1/\sqrt{e}$. However, we postpone the proof of this lemma to the end of this section and first derive some consequences. All we will use is that the exponent $\frac{2}{1+\xi^2}$ is strictly larger than 1. For this, any $\xi < 1$ will do. In fact, for any $0 < \xi < 1$, the lemma remains true provided that I/v is sufficiently large (see Remark 4.18).

Let us first see how this lemma can be used to derive a lower bound on individual influences. For future purposes, it will be convenient to derive the following technical theorem first, from which the actual lower bound on the maximum single influences follows as a corollary.

Theorem 4.15 (“Parametric” KKL Theorem). *There is a universal constant $c > 0$ such that the following holds: For any boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ with variance $v := \text{Var}[f]$ and any parameter $0 < \delta < 1$,*

$$\max_{1 \leq k \leq n} I_k(f) \leq \delta \implies I \geq c \cdot v \cdot \ln(1/\delta).$$

Proof. Let ξ be as in Lemma 4.14. For $\alpha := \frac{2}{1+\xi^2} > 1$, the function $x \mapsto x^\alpha$ is convex. Consequently, a finite sum $\sum_k x_k^\alpha$ with nonnegative real numbers x_k subject to the constraints $\max_k x_k \leq \delta$ and $\sum_k x_k = B$ is maximized if B/δ of the x_k ’s are as large as possible and the remaining ones are zero, i.e., $\sum_k x_k^\alpha \leq (B/\delta) \cdot \delta^\alpha = B \cdot \delta^{\alpha-1}$. (In this somewhat imprecise reasoning, we ignore the fact that B/δ need not be an integer. To argue more formally, invoke Jensen’s inequality.) Applying this with $x_k = I_k(f)$ and $B = I$, it follows from Lemma 4.14 and the assumption $\max_k I_k \leq \delta$ that

$$I \cdot \xi^{I/v} \leq \sum_k I_k^\alpha \leq I \cdot \delta^{\alpha-1},$$

i.e., $\xi^{I/v} \leq \delta^{\alpha-1}$. Taking logarithms and multiplying the inequality by $\frac{v}{\ln \xi} = -\frac{v}{\ln(1/\xi)}$, the theorem follows with $c = \frac{\alpha-1}{\ln(1/\xi)} > 0$. \square

Remark 4.16. We remark that (using L’Hôpital’s Rule, for instance) one can show that

$$c = \frac{\frac{2}{1+\xi^2} - 1}{\ln(1/\xi)} \rightarrow 1$$

as $\xi \rightarrow 1$, so we can make c arbitrarily close to one if we can choose ξ arbitrarily close to 1. This is possible using a bootstrapping argument, see Remark 4.18 below. The same applies to the constant c' in the following corollary: There is a constant that works for all n . Moreover, as we will see, we have $c' = c + o(1)$, so for any $\varepsilon > 0$, there is some N such that the corollary holds with $c' = 1 - \varepsilon$ for all $n \geq N$. Note, furthermore, that even our choice $\xi = 1/\sqrt{e}$ already gives a pretty good constant $c = 2\left(\frac{e-1}{e+1}\right) \approx 0.92$.

Corollary 4.17 (KKL Theorem). *There is a universal constant c' such that the following holds: For $f : \{0, 1\}^n \rightarrow \{0, 1\}$ and $v := \text{Var}[f] = \mathbf{E}[f](1 - \mathbf{E}[f])$,*

$$\max_k I_k(f) \geq c' \cdot v \cdot \frac{\ln n}{n}.$$

Proof. Note that for a boolean function f , we have $v = \text{Var}[f] \leq 1/4$ (we only need that it is bounded). Thus, if $I_k(f) \geq \ln(n)/n$ for some k , we are done, and therefore we may assume that $\max_k I_k(f) \leq \delta := \ln(n)/n$. Then Theorem 4.15 implies $I(f) \geq c \cdot v \cdot (\ln n - \ln \ln n)$. Thus, there must be at least one k with $I_k \geq c \cdot v \cdot (\ln n - \ln \ln n)/n \geq c' \cdot v \cdot \ln(n)/n$, where $c' = c + o(1)$. \square

Proof of Lemma 4.14. By Corollary 4.13, we know that the squares of the “low” Fourier coefficients $\widehat{f}(a)$, i.e., of those with $0 < |a| < \frac{I}{2v}$ make up at least half of the variance. On the other hand, by Corollary 4.9, we know that

$$I(f) = 4 \sum_a \widehat{f}(a)^2 |a|.$$

To relate these two things, we would also like to restrict the latter sum to the low Fourier coefficients. Instead, we will introduce exponentially decaying weights. Let $0 < \xi < 1$ be a parameter which we will determine later. For a function $g : \{0, 1\}^n \rightarrow \mathbf{C}$, define $T_\xi g : \{0, 1\}^n \rightarrow \mathbf{C}$ by its Fourier expansion,

$$\widehat{T_\xi g}(a) := \widehat{g}(a) \cdot \xi^{|a|}.$$

Since we shrink all the Fourier coefficients, it follows immediately from Parseval’s Identity that $\|T_\xi g\|_2 \leq \|g\|_2$. In the next chapter, we will prove the stronger inequality

$$\|T_\xi g\|_2 \leq \|g\|_{1+\xi^2}. \quad (5.1)$$

This inequality is stronger since for functions on a probability space, $\|g\|_p \leq \|g\|_q$ for $1 \leq p \leq q \leq \infty$. If g takes values only in $\{-1, 0, +1\}$, like our functions $\partial_k f$, then (5.1) can be rewritten as

$$\|T_\xi g\|_2 \leq \|g\|_2^{\frac{2}{1+\xi^2}}. \quad (4.4)$$

We now apply this to the functions $\partial_k f$. By Observation 4.7, $I_k(f) = \|\partial_k f\|_2^2$, and so combining (4.7), Parseval’s Identity, Lemma 4.8, and the definition of T_ξ , we obtain

$$I_k^{\frac{2}{1+\xi^2}} = \left(\|\partial_k f\|_2^{\frac{2}{1+\xi^2}} \right)^2 \geq \|T_\xi(\partial_k f)\|_2^2 = \sum_a T_\xi(\widehat{\partial_k f})(a)^2 = 4 \sum_{a:a_k=1} \widehat{f}(a)^2 \xi^{2|a|}.$$

Summing up over all k , we arrive at

$$\sum_{k=1}^n I_k^{\frac{2}{1+\xi^2}} \geq 4 \sum_a \widehat{f}(a)^2 \xi^{2|a|} |a| \geq 4 \sum_{0 < |a| \leq \frac{I}{2v}} \widehat{f}(a)^2 \xi^{2|a|} |a|.$$

We claim that in the restricted sum on the right-hand side the weight $\xi^{2|a|} |a|$ is minimized for the highest terms, i.e., $\xi^{2|a|} |a| \geq \xi^{\frac{I}{v}} \frac{I}{2v}$ for $0 < |a| \leq \frac{I}{2v}$. Assuming for the moment that this claim is true, Corollary 4.13 implies

$$\sum_{k=1}^n I_k^{\frac{2}{1+\xi^2}} \geq 4 \underbrace{\left(\sum_{0 < |a| \leq \frac{I}{2v}} \widehat{f}(a)^2 \right)}_{\geq v/2} \xi^{I/v} \frac{I}{2v} = I \xi^{I/v},$$

as desired. So it remains to verify the claim. For this, we note that the function $h_\xi(s) = \xi^{2s} \cdot s$ is monotonically increasing for $0 < s < \frac{1}{2\ln(1/\xi)}$ and afterwards monotonically decreasing—consider the derivative $h'_\xi(s) = \xi^{2s}(2\ln \xi + 1)$. Therefore, $h_\xi(s) \geq h_\xi(\frac{I}{2v})$ for all *integers* $0 < s \leq \frac{I}{2v}$ provided $\xi^2 = h_\xi(1) \geq h_\xi(\frac{I}{2v})$. Thus, if we chose $\xi = 1/\sqrt{e}$, we are on the safe side since then $\frac{1}{2\ln(1/\xi)} = 1$ and so $h_\xi(s)$ attains its global maximum at $s = 1$. \square

Remark 4.18. We have proved Lemma 4.14 for a fixed parameter $\xi = 1/\sqrt{e}$. As we have seen, it holds for any $0 < \xi < 1$, provided that I/v is sufficiently large (so that $h_\xi(\frac{I}{2v}) \leq h_\xi(1)$). Based on this, we can use the following bootstrapping argument. Let $\varepsilon > 0$. As remarked after the proof of Theorem 4.15, if ξ is sufficiently close to 1, then $c = \frac{2}{\log(1/\xi)} > 1 - \varepsilon$. Moreover, the $\frac{1}{\sqrt{e}}$ -version of Lemma 4.14 implies that if $\max_k I_k \leq \delta$ then $I/v \geq 2\frac{e-1}{e+1} \cdot \log(1/\delta)$, which is sufficiently large for any given $\xi < 1$ if δ is sufficiently small. Moreover, we have already seen that the constant c' in Corollary 4.17 satisfies $c' = 1 + o(1)$, and $\delta = \log(n)/n$ is as small as we like if n is sufficiently large.

Chapter 5

The Noise Operator

Nonuniform product measures on the discrete cube. Recall the definition of the p -measure on the discrete cube: For a real number $0 \leq p \leq 1$, the p -measure μ_p on the 1 -dimensional cube $\{0, 1\}$ is defined by $\mu_p(0) = 1 - p$ and $\mu_p(1) = p$. If $\mathbb{p} = (p_1, \dots, p_n)$ is an n -tuple of real numbers in $[0, 1]$, then the product measure $\mu_{\mathbb{p}} := \mu_{p_1} \otimes \dots \otimes \mu_{p_n}$ on the n -dimensional cube $\{0, 1\}^n$ is given by

$$\mu_{\mathbb{p}}(x) = \prod_{i=1}^n \mu_{p_i}(x_i).$$

This corresponds to tossing n independent 0/1-coins and writing down the sequence of outcomes, where the probability for a 1 to appear in toss i is p_i . In the special case that all $p_i = p$ are equal, we denote the n -dimensional product measure again by μ_p ,

$$\mu_p(x) = p^{|x|}(1 - p)^{n - |x|}.$$

In particular, $\mu_{\frac{1}{2}}$ is the uniform measure on the cube.

5.1 The Definition of T_{ξ}

Let $0 \leq \xi \leq 1$ and let $p := (1 - \xi)/2$; thus, $0 \leq p \leq 1/2$ and $\xi = 1 - 2p$. We define an operator

$$T_{\xi} : \mathbf{C}^{\{0,1\}^n} \rightarrow \mathbf{C}^{\{0,1\}^n},$$

called the *noise operator*. This operator “eats” a function $f : \{0, 1\}^n \rightarrow \mathbf{C}$ and “spits out” another function $T_{\xi}f : \{0, 1\}^n \rightarrow \mathbf{C}$ according the following rule:

$$T_{\xi}f(x) := \mathbf{E}_{y \sim \mu_p}[f(x \oplus y)] = \sum_{y \in \{0,1\}^n} f(x \oplus y) p^{|y|} (1 - p)^{n - |y|}.$$

That is, to compute the value of $T_{\xi}f$ at a point x , we first apply some random noise by flipping each bit of x independently with probability p (this corresponds to adding $y \sim \mu_p$), then evaluate the original function f at the perturbed point, and finally average

over all perturbations. An equivalent way of describing the noise in terms of the parameter ξ is by saying that we do the following independently for each bit of x : with probability ξ , leave the bit unchanged; otherwise, assign it a value chosen uniformly from $\{0, 1\}$.

Example 5.1. In the extreme cases $\xi = 1$ ($p = 0$) and $\xi = 0$ ($p = 1/2$), respectively, $T_1 f = f$ and $T_0 \equiv \mathbf{E}[f]$.

Observation 5.2 (Linearity of T_ξ). *By linearity of expectation, T_ξ is a linear operator, i.e., $T_\xi(f + g) = T_\xi f + T_\xi g$ and $T_\xi(a \cdot f) = a \cdot T_\xi f$ for all $f, g : \{0, 1\}^n$ and all $a \in \mathbf{C}$.*

Next, we observe that noise operator has a product structure. In order to stress the dimension n of the underlying cube, we denote the n -dimensional noise operator by $T_\xi^{(n)}$.

Lemma 5.3. *The n -dimensional operator is the n -fold tensor product of the corresponding 1-dimensional operators,*

$$T_\xi^{(n)} = T_\xi^{(1)} \otimes \cdots \otimes T_\xi^{(1)}.$$

Recall from Section 1.2 that this means the following: If $f : \{0, 1\}^n \rightarrow \mathbf{C}$ happens to be a tensor product $f = f_1 \otimes \cdots \otimes f_n$ of 1-dimensional functions $f_i : \{0, 1\} \rightarrow \mathbf{C}$, i.e.,

$$f(x) = \prod_{i=1}^n f_i(x_i)$$

then

$$T_\xi^{(n)} f(x) = \prod_{i=1}^n (T_\xi^{(1)} f_i)(x_i).$$

Proof. This follows immediately from the fact that the expectation of a product of independent random variables is the product of the expectations (Corollary 1.4): If $f = f_1 \otimes \cdots \otimes f_n$ then

$$T_\xi^{(n)} f(x) = \mathbf{E}_{y \sim \mu_p} \left[\underbrace{f(x \oplus y)}_{\prod_i f_i(x_i \oplus y_i)} \right] = \prod_{i=1}^n \mathbf{E}_{y_i \sim \mu_p} [f_i(x_i \oplus y_i)] = \prod_{i=1}^n (T_\xi^{(1)} f_i)(x_i)$$

for all $x \in \{0, 1\}^n$. □

In particular, when we apply this to the characters $\chi_a = \chi_{a_1} \otimes \cdots \otimes \chi_{a_n}$, $a \in \{0, 1\}^n$ of the discrete cubes, we immediately get all the eigenvectors and -values of the linear operator T_ξ :

Corollary 5.4. *For $a \in \{0, 1\}^n$,*

$$T_\xi \chi_a = \xi^{|a|} \chi_a.$$

In other words, each χ_a is an eigenvector with eigenvalue $\xi^{|a|}$ of the linear operator T_ξ . Since the characters form a basis of the space $\mathbf{C}^{\{0,1\}^n}$, this is a complete list of eigenvectors. An equivalent way of restating this (by the linearity of T_ξ) is in terms of the Fourier expansion:

Corollary 5.5 (Fourier Expansion for the Noise Operator). *For $a \in \{0, 1\}^n$ and $f : \{0, 1\}^n \rightarrow \mathbf{C}$,*

$$\widehat{T_\xi f}(a) = \xi^{|a|} \widehat{f}(a),$$

hence

$$T_\xi f = \sum_S \widehat{f}(S) \xi^{|S|} \chi_S.$$

Since the characters form a basis of the space $\mathbf{C}^{\{0,1\}^n}$, we could also define the noise operator T_ξ in terms of its Fourier expansion, as we did during the proof of Lemma 4.14. For completeness, we state the following lemma, which we will not need in what follows. The proof is left as an exercise for the reader.

Lemma 5.6. *A linear operator R is a convolution (i.e., $Rf = f * g$ for some fixed g) iff its eigenvectors are precisely the characters.*

Lemma 5.7. *For $0 \leq \xi_1, \xi_2 \leq 1$,*

$$T_{\xi_1} T_{\xi_2} = T_{\xi_1 \xi_2}.$$

Proof. By Corollary 5.5, $\widehat{T_\xi f}(S) = \xi^{|S|} \widehat{f}(S)$ for all S and ξ . Applying this twice, we conclude that $(\widehat{T_{\xi_1} T_{\xi_2} f})(S) = \xi_1^{|S|} \widehat{T_{\xi_2} f}(S) = \xi_1^{|S|} \xi_2^{|S|} \widehat{f}(S) = \widehat{T_{\xi_1 \xi_2} f}(S)$ for all $S \subseteq [n]$ and all functions f . By Fourier inversion, this proves the lemma. \square

Corollary 5.8.

$$\mathbf{E}[T_\xi f] = T_0 T_\xi f = T_0 f = \mathbf{E}[f].$$

5.2 The Hypercontractive Inequality

As before, we fix a parameter $0 \leq \xi \leq 1$. In this section, we prove the inequality that we used in the proof of Lemma 4.14.

As remarked above, it follows immediately from Corollary 5.5 that

$$\|T_\xi f\|_2 \leq \|f\|_2$$

for all $f : \{0, 1\}^n$, because of Parseval's Identity and because we shrink all Fourier coefficients. Thus, as a linear operator from the normed space $L^2(\{0, 1\}^n) = (\mathbf{C}^{\{0,1\}^n}, \|\cdot\|_2)$ into itself, T_ξ is *contracting*. The goal is to prove the following strengthening:

Theorem 5.9 (Hypercontractive Inequality). *For any function $f : \{0, 1\}^n \rightarrow \mathbf{C}$,*

$$\|T_\xi f\|_2 \leq \|f\|_{1+\xi^2}. \tag{5.1}$$

In other words, the linear operator between normed spaces

$$T_\xi : L^{1+\xi^2}(\{0, 1\}^n) \rightarrow L^2(\{0, 1\}^n)$$

has operator norm $\|T\| \leq 1$ (as defined in Section 1.3).

We note that this estimate is sharp, because $T_\xi \chi_0 = \chi_0$ for the constant 1 function χ_0 , and the p -norm $\|\chi_0\|_p = 1$ for every $1 \leq p \leq \infty$. We also remark the hypercontractive inequality (5.1) is indeed a strengthening of the first estimate for the 2-norms:

Lemma 5.10. *Let (X, μ) be a probability space and let $1 \leq p \leq q \leq \infty$. Then*

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

for any function $f : X \rightarrow \mathbf{C}$.

Proof. Since $\mu(X) = 1$, the statement is clearly true for $q = \infty$. For $q < \infty$, let $r = q/p$, let r' be the conjugate exponent to r defined by $1/r + 1/r' = 1$, and let $\mathbf{1}$ be the constant 1 function on X . Then Hölder's Inequality implies

$$\|f\|_p = \left(\int_X |f|^p d\mu \right)^{1/p} = \| |f|^p \mathbf{1} \|_1^{1/p} \leq \| |f|^p \|_r \cdot \underbrace{\| \mathbf{1} \|_{r'}}_{=\mu(X)=1} = \left(\int_X |f|^{p \frac{q}{p}} \right)^{1/q},$$

and taking p^{th} roots gives the desired result. \square

Proof of Theorem 5.9. As we observed in Lemma 5.3, the n -dimensional noise operator is a tensor product of 1-dimensional ones. Therefore, by Lemma 1.7, we have the following estimate for the operator norms of the linear operators $T_\xi^{(n)} : L^{1+\xi^2}(\{0, 1\}^n) \rightarrow L^2(\{0, 1\}^n)$ and $T_\xi^{(1)} : L^{1+\xi^2}(\{0, 1\}) \rightarrow L^2(\{0, 1\})$:

$$\|T_\xi^{(n)}\| \leq \left(\|T_\xi^{(1)}\| \right)^n$$

Thus, it suffices to show $\|T_\xi^{(1)}\| \leq 1$, i.e., the Hypercontractive Inequality in dimension 1, i.e., for functions $f : \{0, 1\} \rightarrow \mathbf{C}$. First of all, note that by the definition of T_ξ as an expectation, we have $|T_\xi f| \leq T_\xi |f|$ pointwise. Therefore, it suffices to prove the inequality in the case that f takes nonnegative real values. Then we can write

$$f(x) = a\chi_0(x) + b\chi_1(x) = \begin{cases} a + b & , \text{ if } x = 0 \\ a - b & , \text{ if } x = 1 \end{cases}$$

for some real number a, b , and the nonnegativity of f implies $|b| \leq a$. Hence, either $f \equiv 0$, in which case the inequality is trivial, or $a > 0$ and we may assume that $a = 1$ (because the inequality is homogeneous and we can multiply both sides by $1/a$). Then

$$T_\xi f(x) = \begin{cases} 1 + \xi b & , \text{ if } x = 0 \\ 1 - \xi b & , \text{ if } x = 1 \end{cases}$$

and we want to show

$$\left(\frac{(1 + \xi b)^2 + (1 - \xi b)^2}{2} \right)^{\frac{1+\xi^2}{2}} \leq \frac{(1 + b)^{1+\xi^2} + (1 - b)^{1+\xi^2}}{2}. \quad (5.2)$$

First, consider the right-hand side of the equation. For real numbers $s \geq 1$ and $0 \leq x \leq 1$, we have ¹

$$(1+x)^s = \sum_{k=0}^{\infty} \binom{s}{k} x^k,$$

where $\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!}$. Note that for $1 < s < 2$, the sign of $\binom{s}{k}$ is $(-1)^k$ for $k > 1$. Consequently,

$$\frac{(1+x)^s + (1-x)^s}{2} = \sum_{2|k} \binom{s}{k} x^k \geq 1 + \binom{s}{2} x^2$$

for $-1 \leq x \leq 1$ and $1 \leq s \leq 2$. Applying this with $x = b$ and $s = 1 + \xi^2$, we see that the right-hand side of (5.2) is bounded from below by

$$1 + \frac{\xi^2(1 + \xi^2)}{2} b^2 \tag{5.3}$$

Now we turn to the left-hand side. First, we note that the term inside the big parentheses equals $1 + (\xi b)^2$. Next, consider again the function $h_s(x) = (1+x)^s$, $0 \leq x \leq 1$, this time with $0 < s \leq 1$. If we wanted to apply the Binomial Theorem, we might have a small convergence problem for $s < 1$ and $x = 1$. However, a much easier argument is available: Differentiating, we see that $h'_s(x) = s(1+x)^{s-1} \geq 0$, i.e., the function $h_s(x)$ is concave. Therefore, $h_s(x) \leq 1 + h'_s(0) \cdot x = 1 + sx$. Applying this with $x = \xi^2 b^2$ and $s = \frac{1+\xi^2}{2}$, we see that the left-hand side of (5.2) is bounded from above by (5.3), which completes the proof. \square

¹For $x < 1$ (in fact, for $-1 < x < 1$), the binomial series actually converges for any real (or complex) number s ; this is the Binomial Theorem. Also, for integer s , convergence is never an issue because the series is a finite sum. Finally, for $s > 1$ and $x = 1$, we need to observe that $|s - j| \leq \max s, k - 1$. Thus, $|\binom{s}{k}| \leq \max \left\{ \frac{s^k}{k!}, \frac{s}{k(k-1)} \right\}$, and hence $\sum_{k \geq 0} \binom{s}{k}$ converges.

Chapter 6

Influences for General Product Measures

Chapter 7

Thresholds

Chapter 8

Bounds for Error-Correcting Codes

Suppose we want to store data (an image, a typed text, music), given as a finite binary string $x \in \{0, 1\}^*$, on a digital storage device, for instance on a CD or DVD. The storage devices are generally not 100% flawless, and some bits will not be stored correctly: there is a certain probability that a 0 gets stored as a 1, or vice versa. The same phenomenon occurs when sending the data (from a satellite to the Earth, or within a cellular phone network, etc.), because of noise in the transmission. (In the case of storage, errors may also be caused by the gradual wear and tear on the device over time.)

Thus, when we read out the data from the storage device or receive it, we will not get the original string x , but a string $y = x \oplus e$, where e is the error, hopefully containing few 1's. (Of course, some bits could also become entirely unreadable or get lost.) Coding theory deals with the question of how to encode data so as to make it more resilient to these inevitable errors.¹

In order to achieve this goal, we have to introduce some redundancy into the data.² A very simple way of doing this is the following encoding: simply repeat each symbol of the original string three times, say. In this way, we can reconstruct the original string even in the presence of errors, provided that there is at most one error per block (simply take the majority vote among the three bits in each block of length 3 of the encoding). We say that this encoding can *correct* 1 error. We pay for this by a threefold increase in the storage requirement (or the time needed to transmit the message).

More generally, we can break the original data/message into contiguous blocks of length k , say, and then somehow encode each of the 2^k possible blocks of length k by a longer block of length $n \geq k$. Another very basic example of this is the introduction of a *parity bit*. For example, in the ASCII standard, there are 128 basic characters. Each of these can be represented by a 7-bit string $x_1x_2 \dots x_7$, and an eighth *parity bit*

¹Thus, the goal is different from that of cryptography, where we want to *encrypt* the data in such a way that an unauthorized eavesdropper will not be able to understand it (or at least not within a reasonable span of time). Nonetheless, the theory of error-correcting codes also has applications in cryptography.

²A phenomenon we are very familiar with from our everyday lives: natural languages contain plenty of redundancy, and it usually takes quite a few misprints or mispronounced letters to seriously impede understanding.

$x_8 := x_1 \oplus x_2 \oplus \dots \oplus x_7$ is appended, so that the resulting 8-bit ASCII code word $x = x_1x_2 \dots x_8$ has even parity. Thus, if a single bit get flips, we can detect this because of the wrong parity, even though we will not be able to reconstruct what the original data was before the error. We say that the ASCII code can *detect* one error. Moreover, the storage requirement goes up by a modest factor of 8/7.

Definition 8.1. A *code*³ of *block length* n is a simply a subset of $\{0, 1\}^n$. The elements of C are the *code words*. The *distance* or *minimum distance* is defined as the minimum Hamming distance between any two elements of the code,

$$\text{mindist}(C) := \min_{\substack{v, w \in C \\ v \neq w}} \text{dist}_H(v, w).$$

The *rate* or *efficiency* of the code is defined as

$$\text{rate}(C) := \frac{\log_2 |C|}{n}.$$

If $\lfloor \log_2 |S| \rfloor = k$, then we can use $|C|$ to encode blocks of length k of the original data, and so the encoding increases the storage requirement by a factor of n/k , which is essentially the reciprocal of the rate of C .

Furthermore, if the minimum distance of $|C|$ is d , then the code *corrects* $\lfloor \frac{d-1}{2} \rfloor$ (or fewer) errors, in the following sense: If $x \in C$ and if $|e| \leq \lfloor \frac{d-1}{2} \rfloor$, then x is the code word closest in Hamming distance to $y := x \oplus e$, so the *nearest neighbor decoding* correctly reconstructs the original code word. If d is even and $|e| = d/2$, then there may be two different code words at equal distance $d/2$ to y , so we can no longer reconstruct the original one with absolute certainty, but we can still detect that an error occurred which is beyond the tolerance limit for our code.

The fundamental problem in the theory of error-correcting is to find codes that achieve the conflicting goals of having both a large size (rate) *and* large minimum distance. On a theoretical level, the first question is simply whether a code with given rate and minimum distance *exists*. We define

$$A(n, d) := \max\{|C| : C \subseteq \{0, 1\}^n : \text{mindist}(C) \geq d\}.$$

For a code to be practical, one must also be able first of all to *explicitly construct* the code (we will not try to make the notion of explicit construction precise), and secondly to *efficiently* encode and decode the data. We will largely ignore these algorithmic issues, however.

Here is a first, almost trivial, upper bound on $A(n, d)$. For $0 \leq r \leq n$ and $v \in \{0, 1\}^n$, consider the Hamming ball

$$B_r(v) := \{w \in \{0, 1\}^n : \text{dist}_H(v, w) \leq r\}.$$

³More precisely, a *binary block code*.

We have $|B_r(v)| = \Phi_r(n)$, where

$$\Phi_r(n) := \sum_{i=0}^r \binom{n}{i}$$

for $0 \leq r \leq n$. Suppose that $C \subseteq \{0, 1\}^n$ is a code with $\text{mindist}(C) \geq d$. Then the Hamming balls $B_{\lfloor \frac{d-1}{2} \rfloor}(w)$, $w \in C$, are *pairwise disjoint* (this is precisely the observation that allows us to correct $\lfloor \frac{d-1}{2} \rfloor$ errors). Consequently:

Proposition 8.2 (*Volume or Sphere-Packing Bound*).

$$A(n, d) \leq \left\lfloor \frac{2^n}{\Phi_{\lfloor \frac{d-1}{2} \rfloor}(n)} \right\rfloor.$$

This bound is tight whenever there is a partition of the discrete cube into Hamming balls of a given radius r : then the centers of the balls form a code of minimum distance $d = 2r + 1$ such that the volume bound is attained; such a code is called a *perfect code*. Below, we will see one family of perfect codes, *Hamming codes*, but generally, perfect codes are rare.

8.1 Linear Codes.

A very important special kind of codes are *linear codes*. Consider the discrete cube $\{0, 1\}^n$ as the n -dimensional vector space \mathbf{F}_2^n over the 2-element field \mathbf{F}_2 . A code C is *linear* if it is a linear subspace of this vector space. Thus, since the only scalars are 0 and 1, C is linear iff $v \oplus w \in C$ whenever $v, w \in C$.

A linear code $C \subseteq \mathbf{F}_2^n$ always has size 2^k for some integer k . Moreover, it allows two compact⁴ descriptions. One is by means of a *generator matrix* G : pick a basis $\{v_1, \dots, v_k\}$ of C and let G be the $k \times n$ -matrix with rows v_i^T (where \cdot^T denotes the transpose). Then G defines a linear mapping $\mathbf{F}_2^k \rightarrow \mathbf{F}_2^n$ and C is the image of that mapping. Note that this provides an efficient encoding algorithm. The second way of defining a linear code C is by means of a *parity-check* matrix, which we shall define presently.

Definition 8.3 (Dual Code). Let $C \subseteq \mathbf{F}_2^n$ be a linear code.⁵ Then the *dual code* is given by

$$C^\perp := \{w \in \mathbf{F}_2^n : \langle v, w \rangle \equiv 0 \pmod{2} \text{ for all } v \in C\},$$

where $\langle v, w \rangle = \sum_i v_i w_i$ is the standard inner product. We can consider this inner product sometimes as an integer and sometimes modulo 2, and both options are useful, so we will specify which one we mean.⁶

⁴In the colloquial, non-topological sense of “compact”.

⁵The definition would also make sense for nonlinear codes, but the dual code is always linear.

⁶Note when evaluating the inner product modulo 2, things may happen that we are not used to from our euclidean geometric intuition: C and C^\perp may have nontrivial intersection or even coincide; this is the case, for instance, for $C = \{(0, 0), (1, 1)\} \subseteq \mathbf{F}_2^2$.

Note that the rate of the dual code is $n - k$. Suppose we pick a basis $\{w_1, \dots, w_{n-k}\}$ for C^\perp . If H is the $n \times (n - k)$ -matrix with columns w_i^T then

$$v \in C \iff Hv = 0,$$

i.e., C is the kernel of the linear map $\mathbf{F}_2^n \rightarrow \mathbf{F}_2^{n-k}$ given by H . Such a matrix H is called a *parity-check matrix* for the code C . Conversely, any $n \times (n - k)$ -matrix over \mathbf{F}_2 of full rank $n - k$ defines a linear code $C = \ker H$ with parity-check matrix H .

One example is the ASCII code $C \subseteq \mathbf{F}_2^8$ that we saw above: Here, the parity check matrix has only one row, namely the all-one vector $\mathbb{1}$. Another example are the aforementioned Hamming codes.

Example 8.4 (Hamming codes). Let $s \geq 1$ be integer, and set $n := 2^s - 1$. Let H be the $s \times n$ -matrix whose columns are all nonzero vectors in \mathbf{F}_2^s . The Hamming code of block length n and rate $n - s$ is defined as the code with H as a parity-check matrix.

Observation 8.5. *The Hamming distance is invariant under translation by a fixed vector, i.e., for every $u \in \mathbf{F}_2^n$, it holds that*

$$\text{dist}_H(v, w) = \text{dist}_H(u \oplus v, u \oplus w)$$

for all $v, w \in \mathbf{F}_2^n$. Consequently, if C is a linear code, then the minimum distance of C equals the minimum weight of any nonzero code word,

$$\text{mindist}(C) = \min_{v \in C \setminus \{0\}} |v|.$$

The following is just a simple, but very useful, reformulation of the definition of a parity-check matrix:

Observation 8.6. *Let $C \subseteq \mathbf{F}_2^n$ be a linear code of rate $k = \dim C$, and let $H \in \mathbf{F}_2^{n \times (n-k)}$ be a parity-check matrix for C with columns $a_1 \dots a_n \in \mathbf{F}_2^{n-k}$. Then the code words in C correspond to the linear dependencies (over \mathbf{F}_2) of the columns a_i ,*

$$w = (w_1, \dots, w_n) \in C \iff w_1 a_1 \oplus \dots \oplus w_n a_n = 0.$$

Hence, $\text{mindist}(C) \geq d$ iff any $d - 1$ or fewer of the a_i 's are linearly independent.

Proposition 8.7. *The Hamming code is a perfect code of minimum distance 3.*

Proof. Exercise. □

MacWilliams's identity for the weight enumerator. Sometimes it is useful to have more information about the distribution of pairwise distances in a code rather than just the minimum distance. By Observation 8.5, for linear codes, it is enough to know the distribution of weights of code words. More precisely, for a linear code $C \subseteq \mathbf{F}_2^n$ and $0 \leq i \leq n$, define

$$A_i = A_i(C) := \{w \in C : |w| = i\}.$$

A standard way of representing such a sequence of numbers is by means of a generating function, i.e., by considering the polynomial $\sum_{i=0}^n A_i x^i \in \mathbf{Z}[x]$, where x is a formal variable. In our case, it turns out to be useful to make this polynomial homogeneous by introducing an additional variable. Thus, we arrive at the following notion: The *weight enumerator* of C is defined as the homogeneous bivariate polynomial

$$P_C(x, y) := \sum_{i=0}^n A_i x^i y^{n-i} = \sum_{w \in C} x^{|w|} y^{n-|w|} \in \mathbf{Z}[x, y]$$

It turns out that the weight enumerators of a linear code and its dual, respectively, determine each other by a simple change of variables:

Theorem 8.8 (MacWilliams' Identity). *Let $C \subseteq \mathbf{F}_2^n$ be a linear code. Then*

$$P_C(x, y) = \frac{|C|}{2^n} P_{C^\perp}(y - x, y + x).$$

As a consequence, the weight distributions of C and C^\perp , respectively, determine each other linearly.

Corollary 8.9. *Let $C \subseteq \mathbf{F}_2^n$ be a linear code. For $0 \leq s \leq n$,*

$$A_s(C) = \frac{|C|}{2^n} \sum_{i=0}^n K_s(i; n) \cdot A_i(C^\perp),$$

where

$$K_s(i; n) := \sum_{j=0}^n (-1)^j \binom{i}{j} \binom{n-i}{s-j}.$$

These coefficients are called Kravchuk polynomials.⁷

Proof of the corollary. Let $A'_i := A_i(C^\perp)$. Then

$$\begin{aligned} \frac{2^n}{|C|} P_C(x, 1) &= P_{C^\perp}(1 - x, 1 + x) \\ &= \sum_i A'_i (1 - x)^i (1 + x)^{n-i} \\ &= \sum_i \sum_j \sum_k A'_i (-1)^j \binom{i}{j} \binom{n-i}{k} x^{j+k}. \end{aligned}$$

We are interested in A_s , i.e., in the coefficient of x^s in $P_C(x, 1)$. Thus, $k = s - j$ in the above sum, and we get

$$A_s = \frac{|C|}{2^n} \sum_i A'_i \underbrace{\sum_j (-1)^j \binom{i}{j} \binom{n-i}{s-j}}_{=K_s(i;n)},$$

as desired. □

⁷More precisely, n is usually considered to be fixed, s is considered as a parameter, and the definition is extended from integer i to a real variable x (using $\binom{x}{k} := \frac{x(x-1)\cdots(x-k+1)}{k!}$ for integers $k > 0$, $\binom{x}{0} = 1$, and $\binom{x}{k} = 0$ for integers $k < 0$). The resulting polynomial $K_s(x; n)$ is called the s^{th} Kravchuk polynomial.

Theorem 8.8 is proved using Fourier analysis. We will need the following auxiliary lemma. Let $C \subseteq \mathbf{F}_2^n$ be a linear code, let $u \in \mathbf{F}_2^n$. For $i = 0, 1$, define

$$C_i = C_i(u) := \{v \in C : \langle u, v \rangle \equiv i \pmod{2}\}.$$

Lemma 8.10. *If $u \notin C^\perp$ then*

$$|C_0| = |C_1|.$$

Proof. By definition, $u \notin C^\perp$ means that $C_1 \neq \emptyset$. Choose and fix an arbitrary $w \in C_1$. Translation by w determines a bijective map $C_0 \rightarrow C_1$, $v \mapsto v \oplus w$. \square

Lemma 8.11. *Let $C \subseteq \mathbf{F}_2^n$ be a linear code. Then*

$$\widehat{\mathbf{1}}_C = \frac{|C|}{2^n} \mathbf{1}_{C^\perp}.$$

Proof. By Lemma 8.10, we have

$$\widehat{\mathbf{1}}_C(u) = \frac{1}{2^n} \sum_{v \in C} (-1)^{\langle u, v \rangle} = \frac{1}{2^n} (|C_0| - |C_1|) = \begin{cases} \frac{|C|}{2^n} & , \text{if } u \in C^\perp, \\ 0 & , \text{otherwise.} \end{cases}$$

\square

Proof of Theorem 8.8. We want to show that two bivariate polynomials in $\mathbf{Z}[x, y]$ are equal, i.e., that all their coefficients agree. Each of these polynomials determines a function $\mathbf{R}^2 \rightarrow \mathbf{R}$, and a sufficient condition for the equality of the polynomial is that the two corresponding functions agree on some open set. We take the set $O = \{(x, y) \in \mathbf{R}^2 : y \neq 0\}$. Fix $(x, y) \in O$. This yields a function $f = f_{(x,y)} : \{0, 1\}^n \rightarrow \mathbf{R}$, $f(w) := x^{|w|} y^{n-|w|}$. With this notation,

$$P_C(x, y) = 2^n \langle \mathbf{1}_C, f \rangle = 2^n \sum_u \widehat{\mathbf{1}}_C(u) \widehat{f}(u),$$

by Plancherel, where the left-hand side denotes the value of the polynomial function at $(x, y) \in O$. Thus, it remains to compute the Fourier transform of f . By definition,

$$2^n \widehat{f}(u) = \sum_v x^{|v|} y^{n-|v|}.$$

We group the terms in this sum according to $s := \langle u, v \rangle$ (considered as an integer, not modulo 2) and $t := |v| - s$. Thus,

$$\begin{aligned} 2^n \widehat{f}(u) &= \sum_{s,t=0}^n \binom{|u|}{s} \binom{n-|u|}{t} (-1)^s x^{s+t} y^{n-s-t} \\ &= y^n \sum_{s=0}^n \underbrace{\binom{|u|}{s} \left(-\frac{x}{y}\right)^s}_{(1-x/y)^{|u|}} \sum_{t=0}^n \underbrace{\binom{n-|u|}{t} \left(\frac{x}{y}\right)^t}_{(1+x/y)^{n-|u|}} \\ &= (y-x)^{|u|} (y+x)^{n-|u|}. \end{aligned}$$

Combining this with Lemma 8.11, we obtain

$$P_C(x, y) = \frac{|C|}{2^n} \sum_u \mathbf{1}_{C^\perp}(u) \cdot (y - x)^{|u|} (y + x)^{n-|u|} = P_{C^\perp}(y - x, y + x),$$

as desired. \square

8.2 Asymptotically Good Codes

One can turn the simple volume argument of Proposition 8.2 around to construct codes with a given minimum distance in a recursive fashion: Fix n and d with $d \leq n$. Pick an arbitrary first codeword $w_1 \in \{0, 1\}^n$. Consider the Hamming ball $B_{d-1}(w_1)$ centered at w_1 . If $2^n > \Phi_{d-1}(n)$ then $\{0, 1\}^n \setminus B_{d-1}(w_1) \neq \emptyset$; pick an arbitrary element w_2 in this complement. By construction, we have $\text{dist}_H(w_1, w_2) \geq d$. Assume now that we have constructed $C_{j-1} = \{w_1, \dots, w_{j-1}\}$ with minimum pairwise distance at least d . If the union $\bigcup_{i=1}^{j-1} B_{d-1}(w_i)$ does not cover the whole cube then we can choose an arbitrary element w_j in the complement and $C_j := C_{j-1} \cup \{w_j\}$ will still have minimum distance at least d . A sufficient condition to guarantee this is $(j-1) \cdot \Phi_{d-1}(n) < 2^n$. Thus, this construction shows that $A(n, d) \geq 2^n / \Phi_{d-1}(n)$. With some more care, we can extend this idea to construct linear codes of the almost the same size:

Theorem 8.12 (Gilbert-Varshamov Lower Bound). *Let $1 \leq k, d \leq n$ be integers. If*

$$\Phi_{d-2}(n-1) < 2^{n-k}$$

then there exists a linear code $C \subseteq \mathbf{F}_2^n$ of rate at least k and minimum distance at least d .

Proof. We define C by means of a parity check matrix $H \in \mathbf{F}_2^{(n-k) \times n}$, with columns $a_1, \dots, a_n \in \mathbf{F}_2^{n-k}$, which we construct column by column. By Observation 8.6, we have to guarantee that

$$\text{any } d-1 \text{ or fewer of the columns } a_i \text{ are linearly independent} \quad (*)$$

For the first column, choose an arbitrary *nonzero* vector $a_1 \in \mathbf{F}_2^{n-k}$. Suppose now that we have constructed $a_1 \dots a_{j-1}$ such that $(*)$ is satisfied for $i < j$. When we choose the next column a_j , we may not take any vector that is a linear combination of $d-2$ or fewer of the smaller a_i 's; this is the only requirement. Since the only scalars are 0 and 1, there are exactly $\Phi_{d-2}(j-1)$ forbidden vectors. Thus, if $\Phi_{d-2}(j-1) < 2^{n-k}$, we may choose the next column; this proves the theorem, since $\Phi_r(m)$ is monotone increasing in m . \square

Remark 8.13. The preceding proof yields an exponential-time deterministic algorithm to construct the parity-check matrix H and thus the linear code. Alternatively, there is an extremely simple polynomial-time randomized construction (of Monte Carlo type, i.e., it may fail with small probability) that yields essentially the same bound: Pick a matrix $G \in \mathbf{F}_2^{n \times k}$ uniformly at random (i.e., pick the entries independently uniformly at random), and set $C := G\mathbf{F}_2^k = \{Gx : x \in \mathbf{F}_2^k\}$. One can show that with high probability, C has 2^k distinct elements and that the minimum distance is at least d provided $2^k - 1 < 2^n / \Phi_{d-1}(n)$.

Asymptotics. There is a great number of special constructions that produce better codes (with a better rate for a given minimum distance, or vice versa) for particular values of the parameters, but the Gilbert-Varshamov argument gives the best results in general, in the following sense. An infinite family \mathcal{C} of codes is called *asymptotically good* if both the minimum distance and the rate are $\Omega(n)$, i.e., if there exist constants $\delta, R > 0$ such that $\text{rate}(C) \geq Rn$ and $\text{mindist}(C) \geq \delta n$ for every $c \in \mathcal{C}$, where n is the block length of C (the codes in \mathcal{C} need not be defined for every n , only for infinitely many values of n). For $0 < \delta < 1$, we define $R(\delta)$, the *asymptotic rate limit*, as the largest R such that such an asymptotically good family of codes exist:

$$R(\delta) := \limsup_{n \rightarrow \infty} \max_{\substack{C \subseteq \{0,1\}^n \\ \text{mindist}(C) \geq \delta n}} \frac{\log_2 |C|}{n}.$$

Note that $R(\delta)$ is a monotonically decreasing function of δ , and that $R(\delta) = 0$ for $\delta \geq 1/2$.

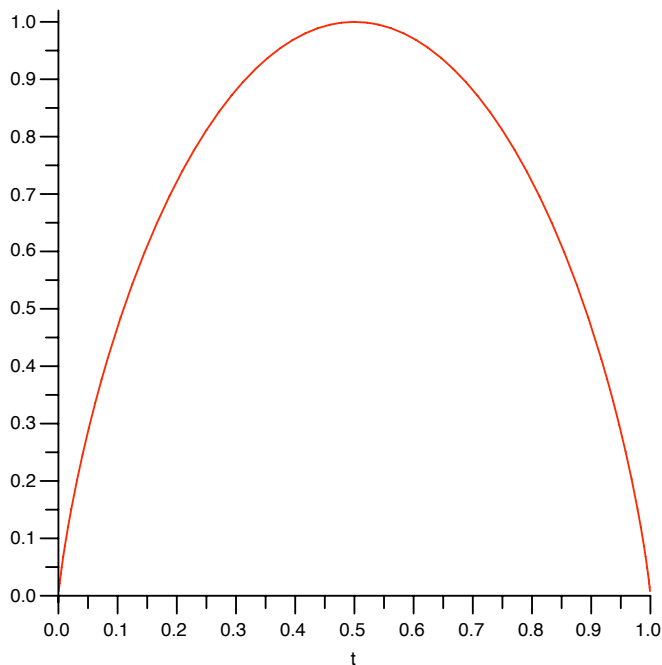


Figure 8.1: The binary entropy function.

In order to state the bounds obtained so far in their asymptotic form, we need the

notion of the (binary) *entropy function*. For $0 \leq t \leq 1$, let

$$H(t) := t \log_2 \left(\frac{1}{t} \right) + (1-t) \log_2 \left(\frac{1}{1-t} \right),$$

see Figure 8.1 (with the understanding that $H(0) = H(1) = 0$). It follows from Stirling's Formula that for any constant $0 < \alpha < 1$,

$$\binom{n}{\alpha n} = 2^{nH(\alpha) + O(\log n)}$$

and therefore also

$$\Phi_{\alpha n}(n) = 2^{nH(\alpha) + O(\log n)}.$$

Applying this to the bounds in Proposition 8.2 and Theorem 8.12, we conclude:

Corollary 8.14. *For $0 < \delta < 1/2$,*

1. *Asymptotic Packing Bound:* $R(\delta) \leq 1 - H(\delta/2)$.
2. *Asymptotic Gilbert-Varshamov Bound:* $R(\delta) \geq 1 - H(\delta)$.

Note that the asymptotic packing bound does not even tend to zero as $\delta \rightarrow 1/2$.

The Delsarte linear programming bound. We now derive an upper bound for the numbers $A(n, d)$ in terms of a linear program, due to Delsarte, which leads to the currently best asymptotic upper bounds for binary codes (for various small values of the parameters, there are further extensions and improvements, for instance based on semidefinite programming).

Let us first consider a *linear code* $C \subseteq \{0, 1\}^n$, and let $A_i := |\{w \in C : |w| = i\}|$, $0 \leq i \leq n$, as before. Clearly, $A_i \geq 0$ for all i , $A_0 = 1$, and

$$|C| = A_0 + \dots + A_n.$$

Moreover, if $\text{mindist}(C) = d$ then

$$A_i = 0, \quad 1 \leq i \leq d-1.$$

MacWilliams' Identity (in form of Corollary 8.9) gives us another set of linear inequalities for the numbers A_i :

$$\frac{|C^\perp|}{2^n} \sum_{i=0}^n K_s(i; n) \cdot A_i = A_s(C^\perp) \geq 0, \quad 1 \leq s \leq n.$$

(We remark that $|C^\perp|/2^n = 1/|C|$ and $K_0(i, n) = 1$ for all i . Therefore, MacWilliams's identity for $s = 0$ just yields the trivial equality $1 = 1$.)

Thus, the maximum size of a *linear* code $C \subseteq \{0, 1\}^n$ is bounded from above by the following linear program:

$$\begin{aligned}
& \text{Maximize} && x_0 + x_1 + \dots + x_n \\
& \text{subject to} && x_0 = 1 \\
& && x_i = 0 && 1 \leq i \leq d-1 \\
& && \sum_{i=0}^n K_s(i; n) \cdot x_i \geq 0 && 1 \leq s \leq n \\
& && x_i \geq 0 && 0 \leq i \leq n.
\end{aligned} \tag{D}$$

Somewhat surprisingly, this remains true for general codes.

Theorem 8.15. *For all integers $1 \leq d \leq n$, the optimal value of the linear program (D) is an upper bound for the maximum size $A(n, d)$ of any code $C \subseteq \{0, 1\}^n$ with minimum distance d .*

More precisely, consider the distance distribution of the code,

$$a_i(C) := \frac{1}{|C|} |\{(v, w) \in C \times C : \text{dist}_H(v, w) = i\}|$$

for $0 \leq i \leq n$; note that for linear codes, $a_i(C) = A_i(C)$. Then $|C| = a_0 + a_1 + \dots + a_n$ and the vector (a_0, a_1, \dots, a_n) forms a feasible solution for the linear program (D).

Proof. This is obvious, except for the constraints involving the Kravchuk polynomials. Let $C \subseteq \{0, 1\}^n$, $\text{mindist}(C) = d$. Consider the function $f = \frac{2^n}{|C|} \mathbf{1}_C * \mathbf{1}_C$, where $\mathbf{1}_C$ is the indicator function of C . Then

$$f(z) = \frac{2^n}{|C|} \frac{1}{2^n} \sum_{v \in \{0, 1\}^n} \mathbf{1}_C(v) \mathbf{1}_C(z \oplus v) = \frac{1}{|C|} |\{(v, w) \in C \times C : v \oplus w = z\}|. \tag{8.1}$$

for all $z \in \{0, 1\}^n$. In particular, $a_i = \sum_{|z|=i} f(z)$ for all i . Furthermore,

$$\widehat{f}_C(z) = \frac{2^n}{|C|} \underbrace{(\widehat{\mathbf{1}}_C(z))^2}_{\in \mathbf{R}} \geq 0 \tag{8.2}$$

for all $z \in \{0, 1\}^n$. For $0 \leq t \leq n$, let L_t be the indicator function of the t^{th} level, i.e., $L_t(x) = 1$ if $|x| = t$ and $L_t(x) = 0$ otherwise. Summing up (8.2) over all points of a given level, we see that

$$\sum_z L_t(z) \widehat{f}(z) \geq 0$$

for $0 \leq t \leq n$. Next, we observe that

$$\chi_y(x) = (-1)^{\langle x, y \rangle} = \chi_x(y)$$

for all $x, y \in \{0, 1\}^n$. Therefore, for any two functions $f, g : \{0, 1\}^n \rightarrow \mathbf{R}$,

$$\sum_y \widehat{f}(y) g(y) = \sum_y \frac{1}{2^n} \sum_x f(x) \chi_y(x) g(y) = \sum_x \frac{1}{2^n} \sum_y f(x) \chi_x(y) g(y) = \sum_x f(x) \widehat{g}(x).$$

Applying this to our particular f under consideration and to $g = L_t$, we conclude that

$$0 \leq \sum_z L_t(z) \widehat{f}(z) = \sum_y f(y) \widehat{L}_t(y)$$

for $0 \leq t \leq n$. Moreover,

$$\widehat{L}_t(y) = \sum_{x:|x|=t} (-1)^{\langle x,y \rangle} = K_t(|y|; n)$$

is essentially the t^{th} Kravchuk polynomial. Consequently,

$$0 \leq \sum_y f(y) \widehat{L}_t(y) = \sum_{i=0}^n K_t(i; n) \underbrace{\sum_{y:|y|=i} f(y)}_{=a_i}$$

for $0 \leq t \leq n$, as desired. \square

In order to derive an upper bound for the optimal solution of the Delsarte linear program, one considers the *dual linear program*:

$$\begin{aligned} &\text{Minimize} && \Lambda(0) \\ &\text{subject to} && \Lambda = \sum_{s=0}^n b_s K_s \\ & && b_0 = 1 \\ & && \Lambda(i) \leq 0 \quad d \leq i \leq n \\ & && b_s \geq 0 \quad 0 \leq s \leq n, \end{aligned} \tag{D*}$$

where we write K_s for $K_s(\cdot; n)$. Any feasible solution to this dual linear program yields an upper bound for the optimal solution of the primal program (actually, by the LP duality theorem, since the primal is both feasible and bounded, so is the dual, and the optimal values are equal, but we do not need this stronger statement).

In this way, one can show the following:

Theorem 8.16.

$$R(\delta) \leq H\left(\frac{1}{2} - \sqrt{\delta(1-\delta)}\right).$$

It is conjectured that this is the best asymptotic bound that can be derived from Delsarte's linear program for $\delta \geq 0.273\dots$

It is known that

$$R_{\text{LP}}(\delta) \geq \frac{1}{2}H(1 - 2\sqrt{\delta(1-\delta)})$$