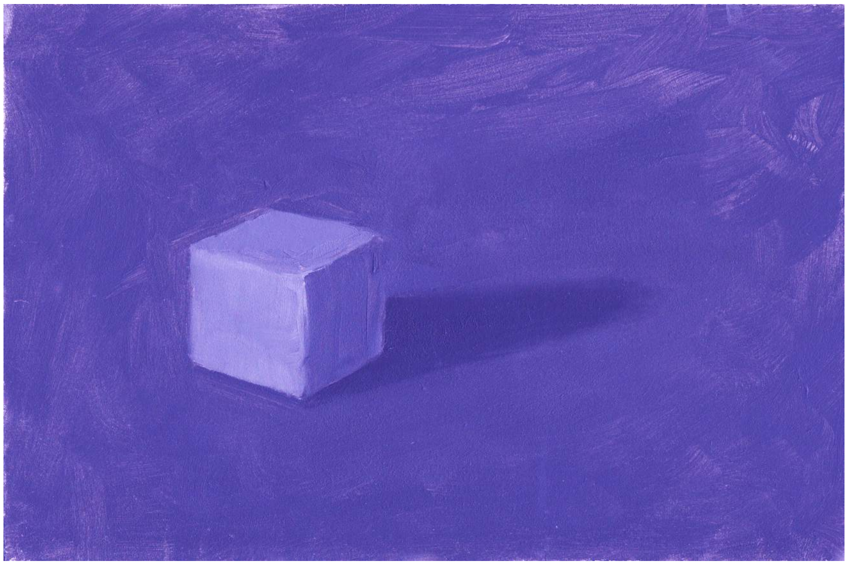
ANALYSIS OF BOOLEAN FUNCTIONS

Ryan O'Donnell



Published 2014 by Cambridge University Press.
Cover illustration by A. A. Williams.

To Zeynep,

for her unending support and encouragement.

Contents

Preface

The subject of this textbook is the *analysis of Boolean functions*. Roughly speaking, this refers to studying Boolean functions $f : \{0,1\}^n \to \{0,1\}$ via their Fourier expansion and other analytic means. Boolean functions are perhaps the most basic object of study in theoretical computer science, and Fourier analysis has become an indispensable tool in the field. The topic has also played a key role in several other areas of mathematics, from combinatorics, random graph theory, and statistical physics, to Gaussian geometry, metric/Banach spaces, and social choice theory.

The intent of this book is both to develop the foundations of the field and to give a wide (though far from exhaustive) overview of its applications. Each chapter ends with a "highlight" showing the power of analysis of Boolean functions in different subject areas: property testing, social choice, cryptography, circuit complexity, learning theory, pseudorandomness, hardness of approximation, concrete complexity, and random graph theory.

The book can be used as a reference for working researchers or as the basis of a one-semester graduate-level course. The author has twice taught such a course at Carnegie Mellon University, attended mainly by graduate students in computer science and mathematics but also by advanced undergraduates, postdocs, and researchers in adjacent fields. In both years most of Chapters 1–5 and 7 were covered, along with parts of Chapters 6, 8, 9, and 11, and some additional material on additive combinatorics. Nearly 500 exercises are provided at the ends of the book's chapters.

Additional material related to the book can be found at its website:

http://analysisofbooleanfunctions.org

This includes complete lecture notes from the author's 2007 course, complete lecture videos from the author's 2012 course, blog updates related to analysis of Boolean functions, an electronic draft of the book, and errata. The author would like to encourage readers to post any typos, bugs, clarification requests, and suggestions to this website.

Acknowledgments

My foremost acknowledgment is to all of the people who have taught me analysis of Boolean functions, especially Guy Kindler and Elchanan Mossel. I also learned a tremendous amount from my advisor Madhu Sudan, and my coauthors and colleagues Per Austrin, Eric Blais, Nader Bshouty, Ilias Diakonikolas, Irit Dinur, Uri Feige, Ehud Friedgut, Parikshit Gopalan, Venkat Guruswami, Johan Håstad, Gil Kalai, Daniel Kane, Subhash Khot, Adam Klivans, James Lee, Assaf Naor, Joe Neeman, Krzysztof Oleszkiewicz, Yuval Peres, Oded Regev, Mike Saks, Oded Schramm, Rocco Servedio, Amir Shpilka, Jeff Steif, Benny Sudakov, Li-Yang Tan, Avi Wigderson, Karl Wimmer, John Wright, Yi Wu, Yuan Zhou, and many others. Ideas from all of them have strongly informed this book.

Many thanks to my PhD students who suffered from my inattention during the completion of this book: Eric Blais, Yuan Zhou, John Wright, and David Witmer. I'd also like to thank the students who took my 2007 and 2012 courses on analysis of Boolean functions; special thanks to Deepak Bal, Carol Wang, and Patrick Xia for their very helpful course writing projects.

Thanks to my editor Lauren Cowles for her patience and encouragement, to the copyediting team of David Anderson and Rishi Gupta, and to Cambridge University Press for welcoming the free online publication of this book. Thanks also to Amanda Williams for the use of the cover image on the book's website.

I'm very grateful to all of the readers of the blog serialization who suggested improvements and pointed out mistakes in the original draft of this work: Amirali Abdullah, Stefan Alders, anon, Arda Antikacıoğlu, Albert Atserias, Per Austrin, Deepak Bal, Paul Beame, Tim Black, Ravi Boppana, Clément Canonne, Sankardeep Chakraborty, Bireswar Das, Andrew Drucker, Kirill Elagin, John Engbers, Diodato Ferraioli, Magnus Find, Michael Forbes, Matt Franklin, David Gajser, David García Soriano, Dmitry Gavinsky, Daniele Gewurz, Mrinalkanti Ghosh, Sivakanth Gopi, Tom Gur, Zachary Hamaker,

Prahladh Harsha, Justin Hilyard, Dmitry Itsykson, Hamidreza Jahanjou, Mitchell Johnston, Gautam Kamath, Shiva Kaul, Brian Kell, Pravesh Kothari, Chin Ho Lee, Euiwoong Lee, Holden Lee, Jerry Li, Noam Lifshitz, Tengyu Ma, Mladen Mikša, Aleksandar Nikolov, David Pritchard, Swagato Sanyal, Pranav Senthilnathan, Igor Shinkar, Lior Silberman, Marla Slusky, Dmitry Sokolov, Aravind Srinivasan, Avishay Tal, Li-Yang Tan, Roei Tell, Suresh Venkatasubramanian, Marc Vinyals, Emanuele Viola, Poorvi Vora, Amos Waterland, Karl Wimmer, Chung Hoi Wong, Xi Wu, Yi Wu, Mingji Xia, Yuichi Yoshida, Shengyu Zhang, and Yu Zhao. Special thanks in this group to Matt Franklin and Li-Yang Tan; extra-special thanks in this group to Noam Lifshitz.

I'm grateful to Denis Thérien for inviting me to lecture at the Barbados Complexity Workshop, to Cynthia Dwork and the STOC 2008 PC for inviting me to give a tutorial, and to the Simons Foundation who arranged for me to co-organize a symposium together with Elchanan Mossel and Krzysztof Oleskiewicz, all on the topic of analysis of Boolean functions. These opportunities greatly helped me to crystallize my thoughts on the topic.

I worked on this book while visiting the Institute for Advanced Study in 2010–2011 (supported by the Von Neumann Fellowship and in part by NSF grants DMS-0835373 and CCF-0832797); I'm very grateful to them for having me and for the wonderful working environment they provided. The remainder of the work on this book was done at Carnegie Mellon; I'm of course very thankful to my colleagues there and to the Department of Computer Science. "Reasonable" random variables were named after the department's "Reasonable Person Principle". I was also supported in this book-writing endeavor by the National Science Foundation, specifically grants CCF-0747250 and CCF-1116594. As usual: "This material is based upon work supported by the National Science Foundation under grant numbers listed above. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF)."

Finally, I'd like to thank all of my colleagues, friends, and relatives who encouraged me to write and to finish the book, Zeynep most of all.

– Ryan O'Donnell
Pittsburgh
October 2013

List of Notation

$\circ$	entry-wise multiplication of vectors
∇	the gradient: $\nabla f(x) = (\mathrm{D}_1 f(x), \dots, \mathrm{D}_n f(x))$
$\neg$	logical NOT
$\ni$	$S \ni i$ is equivalent to $i \in S$
$\oplus$	logical XOR (exclusive-or)
$\hat{\lVert} f \hat{\rVert}_p$	$(\sum_{\gamma \in \widehat{\mathbb{F}_2^n}} \lvert\widehat{f}(\gamma)\rvert^p)^{1/p}$
$\triangle$	symmetric difference of sets; i.e., $S \triangle T = \{i : i \text{ is in exactly one of } S, T\}$
$\vee$	logical OR
$\wedge$	logical AND
$*$	the convolution operator
$[z^k]F(z)$	coefficient on z^k in the power series $F(z)$
1_A	0-1 indicator function for A
$\mathbf{1}_B$	0-1 indicator random variable for event B
2^A	the set of all subsets of A
$\#\alpha$	if α is a multi-index, denotes the number of nonzero components of α
$\lvert\alpha\rvert$	if α is a multi-index, denotes $\sum_i \alpha_i$
AND_n	the logical AND function on n bits: False unless all inputs are True
$A^\perp$	$\{\gamma : \gamma \cdot x = 0 \text{ for all } x \in A\}$
$\mathrm{Aut}(f)$	the group of automorphisms of Boolean function f

$\mathrm{BitsToGaussians}_M^d$	on input the bit matrix $x \in \{-1,1\}^{d \times M}$, has output $z \in \mathbb{R}^d$ equal to $\frac{1}{\sqrt{M}}$ times the column-wise sum of x; if d is omitted it's taken to be 1
$\mathbb{C}$	the complex numbers
$\chi(b)$	when $b \in \mathbb{F}_2^n$, denotes $(-1)^b \in \mathbb{R}$
$\chi_S(x)$	when $x \in \mathbb{R}^n$, denotes $\prod_{i \in S} x_i$, where $S \subseteq [n]$; when $x \in \mathbb{F}_2^n$, denotes $(-1)^{\sum_{i \in S} x_i}$
$\operatorname{codim} H$	for a subspace $H \leq \mathbb{F}^n$, denotes $n - \dim H$
$\mathbf{Cov}[f,g]$	the covariance of f and g, $\mathbf{Cov}[f] = \mathbf{E}[fg] - \mathbf{E}[f]\mathbf{E}[g]$
D_i	the ith discrete derivative: $\mathrm{D}_i f(x) = \frac{f(x^{(i \mapsto 1)}) - f(x^{(i \mapsto -1)})}{2}$
$d_{\chi^2}(\varphi, 1)$	chi-squared distance of the distribution with density φ from the uniform distribution
$\deg(f)$	the degree of f; the least k such that f is a real linear combination of k-juntas
$\deg_{\mathbb{F}_2}(f)$	for Boolean-valued f, the degree of its $\mathbb{F}_2$-polynomial representation
$\Delta(x,y)$	the Hamming distance, $\#\{i : x_i \neq y_i\}$
$\Delta^{(\pi)}(f)$	the expected number of queries made by the best decision tree computing f when the input bits are chosen from the distribution π
$\delta^{(\pi)}(f)$	the revealment of f; i.e., $\min\{\max_i \delta_i^{(\pi)}(\mathcal{T}) : \mathcal{T}$ computes $f\}$
$\Delta^{(\pi)}(\mathcal{T})$	the expected number of queries made by randomized decision tree $\mathcal{T}$ when the input bits are chosen from the distribution π
$\delta_i^{(\pi)}(\mathcal{T})$	the probability randomized decision tree $\mathcal{T}$ queries coordinate i when the input bits are chosen from the distribution π
$\Delta_y f$	for $f : \mathbb{F}_2^n \to \mathbb{F}_2$, the function $\mathbb{F}_2^n \to \mathbb{F}_2$ defined by $\Delta_y f(x) = f(x+y) - f(x)$
$\operatorname{dist}(g,h)$	the relative Hamming distance; i.e., the fraction of inputs on which g and h disagree
$\mathrm{DNF}_{\mathrm{size}}(f)$	least possible size of a DNF formula computing f
$\mathrm{DNF}_{\mathrm{width}}(f)$	least possible width of a DNF formula computing f
$\mathrm{DT}(f)$	least possible depth of a decision tree computing f
$\mathrm{DT}_{\mathrm{size}}(f)$	least possible size of a decision tree computing f
$d_{\mathrm{TV}}(\varphi, \psi)$	total variation distance between the distributions with densities φ, ψ

E_i	the ith expectation operator: $\mathrm{E}_i f(x) = \mathbf{E}_{\boldsymbol{x}_i}[f(x_1,\dots,x_{i-1},\boldsymbol{x}_i,x_{i+1},\dots,x_n))]$
E_I	the expectation over coordinates I operator
$\mathbf{Ent}[f]$	for a nonnegative function on a probability space, denotes $\mathbf{E}[f\ln f] - \mathbf{E}[f]\ln\mathbf{E}[f]$
$\mathbf{E}_{\pi_p}[\cdot]$	an abbreviation for $\mathbf{E}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[\cdot]$
$f \oplus g$	if $f:\{-1,1\}^m \to \{-1,1\}$ and $g:\{-1,1\}^n \to \{-1,1\}$, denotes the function $h:\{-1,1\}^{m+n}\to\{-1,1\}$ defined by $h(x,y) = f(x)g(y)$
$f \otimes g$	if $f:\{-1,1\}^m \to \{-1,1\}$ and $g:\{-1,1\}^n \to \{-1,1\}$, denotes the function $h:\{-1,1\}^{mn}\to\{-1,1\}$ defined by $h(x^{(1)},\dots,x^{(m)}) = f(g(x^{(1)}),\dots,g(x^{(m)}))$
$f^{\otimes d}$	if $f:\{-1,1\}^n\to\{-1,1\}$, then $f^{\otimes d}:\{-1,1\}^{n^d}\to\{-1,1\}$ is defined inductively by $f^{\otimes 1} = f$, $f^{\otimes(d+1)} = f\otimes f^{\otimes d}$
f^{*n}	the n-fold convolution, $f * f * \cdots * f$
$f^{\dagger}$	the Boolean dual defined by $f^{\dagger}(x) = -f(-x)$
f^{+z}	if $f:\mathbb{F}_2^n\to\mathbb{R}$, $z\in\mathbb{F}_2^n$, denotes the function $f^{+z}(x) = f(x+z)$
f_H^{+z}	denotes $(f^{+z})_H$
$\mathbb{F}_2$	the finite field of size 2
$\widehat{\mathbb{F}_2^n}$	the group (vector space) indexing the Fourier characters of functions $f:\mathbb{F}_2^n\to\mathbb{R}$
f^{even}	the even part of f, $(f(x)+f(-x))/2$
$\langle f,g\rangle$	$\mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})g(\boldsymbol{x})]$
f_H	if $f:\mathbb{F}_2^n\to\mathbb{R}$, $H\le\mathbb{F}_2^n$, denotes the restriction of f to H
$\widehat{f}(i)$	shorthand for $\widehat{f}(\{i\})$ when $i\in\mathbb{N}$
$f^{\subseteq J}$	the function (depending only on the J coordinates) defined by $f^{\subseteq J}(x) = \mathbf{E}_{\boldsymbol{x}'_{\overline{J}}}[f(x_J,\boldsymbol{x}'_{\overline{J}})]$; in particular, it's $\sum_{S\subseteq J}\widehat{f}(S)\chi_S$ when $f:\{-1,1\}^n\to\mathbb{R}$
$f_{\mid z}$	if $f:\Omega^n\to\mathbb{R}$, $J\subseteq[n]$, and $z\in\Omega^{\overline{J}}$, denotes the restriction of f given by fixing the coordinates in $\overline{J}$ to z
$f_{J\mid z}$	if $f:\Omega^n\to\mathbb{R}$, $J\subseteq[n]$, and $z\in\Omega^{\overline{J}}$, denotes the restriction of f given by fixing the coordinates in $\overline{J}$ to z
$f^{=k}$	$\sum_{\lvert S\rvert=k}\widehat{f}(S)\chi_S$
$f^{\le k}$	$\sum_{\lvert S\rvert\le k}\widehat{f}(S)\chi_S$
f^{odd}	the odd part of f, $(f(x)-f(-x))/2$
$\mathbb{F}_{p^\ell}$	for p prime and $\ell\in\mathbb{N}^+$, denotes the finite field of p^ℓ elements

$\widehat{f}(S)$	the Fourier coefficient of f on character χ_S
$\mathrm{F}_{S\mid\overline{J}}f(z)$	for $S \subseteq J \subseteq [n]$, denotes $\widehat{f_{J\mid z}}(S)$
$\widetilde{f}$	the randomization/symmetrization of f, defined by $\widetilde{f}(r,x) = \sum_S \boldsymbol{r}^S f^{=S}(\boldsymbol{x})$
$\gamma^+(\partial A)$	the Gaussian Minkowski content of ∂A
$\mathscr{G}(v,p)$	the Erdős–Rényi random graph distribution, $\pi_p^{\otimes\binom{v}{2}}$
h_j	the jth (normalized) Hermite polynomial, $h_j = \frac{1}{\sqrt{j!}}H_j$
h_α	for $\alpha \in \mathbb{N}^n$ a multi-index, the n-variate (normalized) Hermite polynomial $h_\alpha(z) = \prod_{j=1}^n h_{\alpha_j}(z_j)$
H_j	the jth probabilists' Hermite polynomial, defined by $\exp(tz - \frac{1}{2}t^2) = \sum_{j=0}^\infty \frac{1}{j!}H_j(z)t^j$
$\mathbf{Inf}_i[f]$	the influence of coordinate i on f
$\mathbf{Inf}_i^{(\rho)}[f]$	the ρ-stable influence, $\mathbf{Stab}_\rho[\mathrm{D}_i f]$
$\widetilde{\mathbf{Inf}}_J[f]$	the coalitional influence of $J \subseteq [n]$ on $f : \{-1,1\}^n \to \{-1,1\}$, namely $\mathbf{Pr}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[f_{J\mid\boldsymbol{z}}$ is not constant$]$
$\widetilde{\mathbf{Inf}}_J^b[f]$	for $b \in \{-1,1\}$, equals $\mathbf{Pr}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[f_{J\mid\boldsymbol{z}} \not\equiv -b] - \mathbf{Pr}[f = b]$
$\overline{J}$	if $J \subseteq [n]$, denotes $[n] \setminus J$
$L^2(\{-1,1\}^n)$	denotes $L^2(\{-1,1\}^n, \pi_{1/2}^{\otimes n})$
$L^2(G^n)$	if G is a finite abelian group, denotes the complex inner product space of functions $G^n \to \mathbb{R}$ with inner product $\langle f,g\rangle = \mathbf{E}_{\boldsymbol{x}\sim G^n}[f(\boldsymbol{x})\overline{g(\boldsymbol{x})}]$
$L^2(\Omega,\pi)$	the inner product space of (square-integrable) functions $\Omega \to \mathbb{R}$ with inner product $\langle f,g\rangle = \mathbf{E}_{\boldsymbol{x}\sim\pi}[f(\boldsymbol{x})g(\boldsymbol{x})]$
$\Lambda_\rho(\alpha,\beta)$	$\mathbf{Pr}[\boldsymbol{z}_1 \leq t, \boldsymbol{z}_2 \leq t']$, where $\boldsymbol{z}_1,\boldsymbol{z}_2$ are standard Gaussians with correlation $\mathbf{E}[\boldsymbol{z}_1\boldsymbol{z}_2] = \rho$, and $t = \Phi^{-1}(\alpha)$, $t' = \Phi^{-1}(\beta)$
$\Lambda_\rho(\alpha)$	denotes $\Lambda_\rho(\alpha,\alpha)$
$\mathrm{L}f$	the Laplacian operator applied to the Boolean function f, defined by $\mathrm{L}f = \sum_{i=1}^n \mathrm{L}_i f$ (or, the Ornstein–Uhlenbeck operator if f is a function on Gaussian space)
L_i	the ith coordinate Laplacian operator: $\mathrm{L}_i f = f - \mathrm{E}_i f$
$\ln x$	$\log_e x$
$\log x$	$\log_2 x$
Maj_n	the majority function on n bits
$\mathbf{MaxInf}[f]$	$\max_i\{\mathbf{Inf}_i[f]\}$
$[n]$	$\{1,2,3,\ldots,n\}$

$\mathbb{N}$	$\{0,1,2,3,\dots\}$
$\mathbb{N}^+$	$\{1,2,3,\dots\}$
$\mathbb{N}_{<m}$	$\{0,1,\dots,m-1\}$
$N_\rho(x)$	when $x \in \{-1,1\}^n$, denotes the probability distribution generating a string ρ-correlated to x
$N_\rho(z)$	when $z \in \mathbb{R}^n$, denotes the probability distribution of $\rho z + \sqrt{1-\rho^2}\boldsymbol{g}$ where $\boldsymbol{g} \sim \mathrm{N}(0,1)^n$
$\mathbf{NS}_\delta[f]$	the noise sensitivity of f at δ; i.e., $\frac{1}{2} - \frac{1}{2}\mathbf{Stab}_{1-2\delta}[f]$
$\mathrm{N}(0,1)$	the standard Gaussian distribution
$\mathrm{N}(0,1)^d$	the distribution of d independent standard Gaussians; i.e., $\mathrm{N}(0,I_{d\times d})$
$\mathrm{N}(\mu,\Sigma)$	for $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ positive semidefinite, the d-variate Gaussian distribution with mean μ and covariance matrix Σ
OR_n	the logical OR function on n bits: True unless all inputs are False
ϕ	the standard Gaussian pdf, $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$
Φ	the standard Gaussian cdf, $\Phi(t) = \int_{-\infty}^{t}\phi(z)\,dz$
$\overline{\Phi}$	the standard Gaussian complementary cdf, $\overline{\Phi}(t) = \int_{t}^{\infty}\phi(z)\,dz$
φ_A	the density function for the uniform probability distribution on A; i.e., $1_A/\mathbf{E}[1_A]$
ϕ_α	given functions $\phi_0,\dots,\phi_{m-1}$ and a multi-index α, denotes $\prod_{i=1}^{n}\phi_{\alpha_i}$
$\pi^{\otimes n}$	if π is a probability distribution on Ω, denotes the associated product probability distribution on Ω^n
$\pi_{1/2}$	the uniform distribution on $\{-1,1\}$
π_p	the "p-biased" distribution on bits: $\pi_p(-1) = p$, $\pi_p(1) = 1-p$
$\mathbf{Pr}_{\pi_p}[\cdot]$	an abbreviation for $\mathbf{Pr}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[\cdot]$
$\mathbb{R}$	the real numbers
$\mathbb{R}^{\geq 0}$	the nonnegative real numbers
$\mathrm{RDT}(f)$	the zero-error randomized decision tree complexity of f
$\mathbf{RS}_A(\delta)$	the rotation sensitivity of A at δ; i.e., $\mathbf{Pr}[1_A(\boldsymbol{z}) \neq 1_A(\boldsymbol{z}')]$ for a $\cos\delta$-correlated pair $(\boldsymbol{z},\boldsymbol{z}')$
$\mathrm{sens}_f(x)$	the number of pivotal coordinates for f at x
$\mathrm{sgn}(t)$	$+1$ if $t \geq 0$, -1 if $t < 0$
S_n	the symmetric group on $[n]$

sparsity(f)	$\mathbf{Pr}_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq 0]$
sparsity$(\widehat{f})$	$\lvert\mathrm{supp}(\widehat{f})\rvert$
$\mathbf{Stab}_\rho[f]$	the noise stability of f at ρ: $\mathbf{E}[f(\boldsymbol{x})f(\boldsymbol{y})]$ where $\boldsymbol{x}, \boldsymbol{y}$ are a ρ-correlated pair
supp(α)	if α is a multi-index, denotes $\{i : \alpha_i \neq 0\}$
supp(f)	if f is a function, denotes the set of inputs where f is nonzero
T_ρ	the noise operator: $\mathrm{T}_\rho f(x) = \mathbf{E}_{\boldsymbol{y} \sim N_\rho(x)}[f(\boldsymbol{y})]$
T^i_ρ	the operator defined by $\mathrm{T}^i_\rho f(x) = \rho f + (1-\rho)\mathrm{E}_i f$
T_r	for $r \in \mathbb{R}^n$, denotes the operator defined by $\mathrm{T}^1_{r_1}\mathrm{T}^2_{r_2}\cdots\mathrm{T}^n_{r_n}$
$\mathscr{U}$	the Gaussian isoperimetric function, $\mathscr{U} = \phi \circ \Phi^{-1}$
U_ρ	the Gaussian noise operator: $\mathrm{U}_\rho f(z) = \mathbf{E}_{\boldsymbol{z}' \sim N_\rho(z)}[f(\boldsymbol{z}')]$
$\mathbf{Var}[f]$	the variance of f, $\mathbf{Var}[f] = \mathbf{E}[f^2] - \mathbf{E}[f]^2$
Var_i	the operator defined by $\mathrm{Var}_i f(x) = \mathbf{Var}_{\boldsymbol{x}_i}[f(x_1, \dots, x_{i-1}, \boldsymbol{x}_i, x_{i+1}, \dots, x_n))]$
$\mathrm{vol}_\gamma(A)$	$\mathbf{Pr}_{\boldsymbol{z} \sim \mathrm{N}(0,1)^n}[\boldsymbol{z} \in A]$, the Gaussian volume of A
$\mathbf{W}^k[f]$	the Fourier weight of f at degree k
$\mathbf{W}^{>k}[f]$	the Fourier weight of f at degrees above k
$x^{(i \mapsto b)}$	the string $(x_1, \dots, x_{i-1}, b, x_{i+1}, \dots, x_n)$
$x^{\oplus i}$	$(x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)$
$\boldsymbol{x} \sim \varphi$	the random variable $\boldsymbol{x}$ is chosen from the probability distribution with density φ
x^S	$\prod_{i \in S} x_i$, with the convention $x^\emptyset = 1$
$\boldsymbol{x} \sim A$	the random variable $\boldsymbol{x}$ is chosen uniformly from the set A
$\boldsymbol{x} \sim \{-1,1\}^n$	the random variable $\boldsymbol{x}$ is chosen uniformly from $\{-1,1\}^n$
(y,z)	if $J \subseteq [n]$, $y \in \{-1,1\}^J$, $z \in \{-1,1\}^{\overline{J}}$, denotes the natural composite string in $\{-1,1\}^n$
$\mathbb{Z}$	the additive group of integers modulo m
$\widehat{\mathbb{Z}_m^n}$	the group indexing the Fourier characters of functions $f : \mathbb{Z}_m^n \to \mathbb{C}$

Chapter 1

Boolean functions and the Fourier expansion

In this chapter we describe the basics of analysis of Boolean functions. We emphasize viewing the Fourier expansion of a Boolean function as its representation as a real multilinear polynomial. The viewpoint based on harmonic analysis over $\mathbb{F}_2^n$ is mostly deferred to Chapter 3. We illustrate the use of basic Fourier formulas through the analysis of the Blum–Luby–Rubinfeld linearity test.

1.1. On analysis of Boolean functions

This is a book about Boolean functions,

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

Here f maps each length-n binary vector, or *string*, into a single binary value, or *bit*. Boolean functions arise in many areas of computer science and mathematics. Here are some examples:

- In circuit design, a Boolean function may represent the desired behavior of a circuit with n inputs and one output.
- In graph theory, one can identify v-vertex graphs G with length-$\binom{v}{2}$ strings indicating which edges are present. Then f may represent a property of such graphs; e.g., $f(G) = 1$ if and only if G is connected.
- In extremal combinatorics, a Boolean function f can be identified with a "set system" $\mathscr{F}$ on $[n] = \{1,2,\dots,n\}$, where sets $X \subseteq [n]$ are identified with their 0-1 indicators and $X \in \mathscr{F}$ if and only if $f(X) = 1$.

- In coding theory, a Boolean function might be the indicator function for the set of messages in a binary error-correcting code of length n.
- In learning theory, a Boolean function may represent a "concept" with n binary attributes.
- In social choice theory, a Boolean function can be identified with a "voting rule" for an election with two candidates named 0 and 1.

We will be quite flexible about how bits are represented. Sometimes we will use True and False; sometimes we will use -1 and 1, thought of as real numbers. Other times we will use 0 and 1, and these might be thought of as real numbers, as elements of the field $\mathbb{F}_2$ of size 2, or just as symbols. Most frequently we will use -1 and 1, so a Boolean function will look like

$$f : \{-1,1\}^n \to \{-1,1\}.$$

But we won't be dogmatic about the issue.

We refer to the domain of a Boolean function, $\{-1,1\}^n$, as the *Hamming cube* (or hypercube, n-cube, Boolean cube, or discrete cube). The name "Hamming cube" emphasizes that we are often interested in the *Hamming distance* between strings $x, y \in \{-1,1\}^n$, defined by

$$\Delta(x,y) = \#\{i : x_i \neq y_i\}.$$

Here we've used notation that will arise constantly: x denotes a bit string, and x_i denotes its ith coordinate.

Suppose you have a problem involving Boolean functions with the following two characteristics:

- the Hamming distance is relevant;
- you are *counting* strings, or the uniform probability distribution on $\{-1,1\}^n$ is involved.

These are the hallmarks of a problem for which *analysis of Boolean functions* may help. Roughly speaking, this means deriving information about Boolean functions by analyzing their *Fourier expansion*.

1.2. The "Fourier expansion": functions as multilinear polynomials

The *Fourier expansion* of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ is simply its representation as a real, multilinear polynomial. (*Multilinear* means that no variable x_i appears squared, cubed, etc.) For example, suppose $n = 2$ and

$f = \mathrm{max}_2$, the "maximum" function on 2 bits:

$$\begin{aligned}\mathrm{max}_2(+1,+1) &= +1,\\ \mathrm{max}_2(-1,+1) &= +1,\\ \mathrm{max}_2(+1,-1) &= +1,\\ \mathrm{max}_2(-1,-1) &= -1.\end{aligned}$$

Then max_2 can be expressed as a multilinear polynomial,

$$\mathrm{max}_2(x_1,x_2) = \tfrac{1}{2} + \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 - \tfrac{1}{2}x_1x_2; \tag{1.1}$$

this is the "Fourier expansion" of max_2. As another example, consider the *majority function* on 3 bits, $\mathrm{Maj}_3 : \{-1,1\}^3 \to \{-1,1\}$, which outputs the ±1 bit occurring more frequently in its input. Then it's easy to verify the Fourier expansion

$$\mathrm{Maj}_3(x_1,x_2,x_3) = \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{2}x_3 - \tfrac{1}{2}x_1x_2x_3. \tag{1.2}$$

The functions max_2 and Maj_3 will serve as running examples in this chapter.

Let's see how to obtain such multilinear polynomial representations in general. Given an arbitrary Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ there is a familiar method for finding a polynomial that interpolates the 2^n values that f assigns to the points $\{-1,1\}^n \subset \mathbb{R}^n$. For each point $a = (a_1,\dots,a_n) \in \{-1,1\}^n$ the *indicator polynomial*

$$1_{\{a\}}(x) = \left(\tfrac{1+a_1x_1}{2}\right)\left(\tfrac{1+a_2x_2}{2}\right)\cdots\left(\tfrac{1+a_nx_n}{2}\right)$$

takes value 1 when $x = a$ and value 0 when $x \in \{-1,1\}^n \setminus \{a\}$. Thus f has the polynomial representation

$$f(x) = \sum_{a\in\{-1,1\}^n} f(a)1_{\{a\}}(x).$$

Illustrating with the $f = \mathrm{max}_2$ example again, we have

$$\begin{aligned}\mathrm{max}_2(x) \quad &= \quad (+1)\left(\tfrac{1+x_1}{2}\right)\left(\tfrac{1+x_2}{2}\right)\\ &+ \quad (+1)\left(\tfrac{1-x_1}{2}\right)\left(\tfrac{1+x_2}{2}\right)\\ &+ \quad (+1)\left(\tfrac{1+x_1}{2}\right)\left(\tfrac{1-x_2}{2}\right)\\ &+ \quad (-1)\left(\tfrac{1-x_1}{2}\right)\left(\tfrac{1-x_2}{2}\right) \quad = \quad \tfrac{1}{2} + \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 - \tfrac{1}{2}x_1x_2.\end{aligned} \tag{1.3}$$

Let us make two remarks about this interpolation procedure. First, it works equally well in the more general case of *real-valued Boolean functions*, $f : \{-1,1\}^n \to \mathbb{R}$. Second, since the indicator polynomials are multilinear when expanded out, the interpolation always produces a multilinear polynomial. Indeed, it makes sense that we can represent functions $f : \{-1,1\}^n \to \mathbb{R}$ with multilinear polynomials: since we only care about inputs x where $x_i = \pm 1$, any factor of x_i^2 can be replaced by 1.

We have illustrated that every $f : \{-1,1\}^n \to \mathbb{R}$ can be represented by a real multilinear polynomial; as we will see in Section 1.3, this representation is unique. The multilinear polynomial for f may have up to 2^n terms, corresponding to the subsets $S \subseteq [n]$. We write the monomial corresponding to S as

$$x^S = \prod_{i \in S} x_i \qquad \text{(with } x^\emptyset = 1 \text{ by convention)},$$

and we use the following notation for its coefficient:

$$\widehat{f}(S) = \text{coefficient on monomial } x^S \text{ in the multilinear representation of } f.$$

This discussion is summarized by the *Fourier expansion theorem*:

Theorem 1.1. *Every function $f : \{-1,1\}^n \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial,*

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) x^S. \tag{1.4}$$

This expression is called the Fourier expansion *of f, and the real number $\widehat{f}(S)$ is called the* Fourier coefficient of f on S. *Collectively, the coefficients are called the* Fourier spectrum *of f.*

As examples, from (1.1) and (1.2) we obtain:

$$\widehat{\max_2}(\emptyset) = \tfrac{1}{2}, \quad \widehat{\max_2}(\{1\}) = \tfrac{1}{2}, \quad \widehat{\max_2}(\{2\}) = \tfrac{1}{2}, \quad \widehat{\max_2}(\{1,2\}) = -\tfrac{1}{2};$$

$$\widehat{\mathrm{Maj}_3}(\{1\}),\ \widehat{\mathrm{Maj}_3}(\{2\}),\ \widehat{\mathrm{Maj}_3}(\{3\}) = \tfrac{1}{2}, \quad \widehat{\mathrm{Maj}_3}(\{1,2,3\}) = -\tfrac{1}{2}, \quad \widehat{\mathrm{Maj}_3}(S) = 0 \text{ else}.$$

We finish this section with some notation. It is convenient to think of the monomial x^S as a function on $x = (x_1, \dots, x_n) \in \mathbb{R}^n$; we write it as

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

Thus we sometimes write the Fourier expansion of $f : \{-1,1\}^n \to \mathbb{R}$ as

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

So far our notation makes sense only when representing the Hamming cube by $\{-1,1\}^n \subseteq \mathbb{R}^n$. The other frequent representation we will use for the cube is $\mathbb{F}_2^n$. We can define the Fourier expansion for functions $f : \mathbb{F}_2^n \to \mathbb{R}$ by "encoding" input bits $0, 1 \in \mathbb{F}_2$ by the real numbers $-1, 1 \in \mathbb{R}$. We choose the encoding $\chi : \mathbb{F}_2 \to \mathbb{R}$ defined by

$$\chi(0_{\mathbb{F}_2}) = +1, \quad \chi(1_{\mathbb{F}_2}) = -1.$$

This encoding is not so natural from the perspective of Boolean logic; e.g., it means the function $\max_2$ we have discussed represents logical AND. But it's mathematically natural because for $b \in \mathbb{F}_2$ we have the formula $\chi(b) = (-1)^b$. We now extend the χ_S notation:

Definition 1.2. For $S \subseteq [n]$ we define $\chi_S : \mathbb{F}_2^n \to \mathbb{R}$ by

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

which satisfies

$$\chi_S(x + y) = \chi_S(x)\chi_S(y). \tag{1.5}$$

In this way, given any function $f : \mathbb{F}_2^n \to \mathbb{R}$ it makes sense to write its Fourier expansion as

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

In fact, if we are really thinking of $\mathbb{F}_2^n$ the n-dimensional vector space over $\mathbb{F}_2$, it makes sense to identify subsets $S \subseteq [n]$ with vectors $\gamma \in \mathbb{F}_2^n$. This will be discussed in Chapter 3.2.

1.3. The orthonormal basis of parity functions

For $x \in \{-1,1\}^n$, the number $\chi_S(x) = \prod_{i \in S} x_i$ is in $\{-1,1\}$. Thus $\chi_S : \{-1,1\}^n \to \{-1,1\}$ is a Boolean function; it computes the logical *parity*, or *exclusive-or* (XOR), of the bits $(x_i)_{i \in S}$. The parity functions play a special role in the analysis of Boolean functions: the Fourier expansion

$$f = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S \tag{1.6}$$

shows that any f can be represented as a linear combination of parity functions (over the reals).

It's useful to explore this idea further from the perspective of linear algebra. The set of all functions $f : \{-1,1\}^n \to \mathbb{R}$ forms a vector space V, since we can add two functions (pointwise) and we can multiply a function by a real scalar. The vector space V is 2^n-dimensional: if we like we can think of the functions in this vector space as vectors in $\mathbb{R}^{2^n}$, where we stack the 2^n values $f(x)$ into a tall column vector (in some fixed order). Here we illustrate the Fourier expansion (1.1) of the $\max_2$ function from this perspective:

$$\max_2 = \begin{bmatrix} +1 \\ +1 \\ +1 \\ -1 \end{bmatrix} = (1/2)\begin{bmatrix} +1 \\ +1 \\ +1 \\ +1 \end{bmatrix} + (1/2)\begin{bmatrix} +1 \\ -1 \\ +1 \\ -1 \end{bmatrix} + (1/2)\begin{bmatrix} +1 \\ +1 \\ -1 \\ -1 \end{bmatrix} + (-1/2)\begin{bmatrix} +1 \\ -1 \\ -1 \\ +1 \end{bmatrix}. \tag{1.7}$$

More generally, the Fourier expansion (1.6) shows that every function $f : \{-1,1\}^n \to \mathbb{R}$ in V is a linear combination of the parity functions; i.e., the parity functions are a *spanning set* for V. Since the number of parity functions is $2^n = \dim V$, we can deduce that they are in fact a *linearly independent basis* for V. In particular this justifies the uniqueness of the Fourier expansion stated in Theorem 1.1.

We can also introduce an inner product on pairs of function $f,g : \{-1,1\}^n \to \mathbb{R}$ in V. The usual inner product on $\mathbb{R}^{2^n}$ would correspond to $\sum_{x\in\{-1,1\}^n} f(x)g(x)$, but it's more convenient to scale this by a factor of 2^{-n}, making it an average rather than a sum. In this way, a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ will have $\langle f, f\rangle = 1$, i.e., be a "unit vector".

Definition 1.3. We define an inner product $\langle\cdot,\cdot\rangle$ on pairs of function $f,g : \{-1,1\}^n \to \mathbb{R}$ by

$$\langle f,g\rangle = 2^{-n}\sum_{x\in\{-1,1\}^n} f(x)g(x) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[f(\boldsymbol{x})g(\boldsymbol{x})]. \tag{1.8}$$

We also use the notation $\|f\|_2 = \sqrt{\langle f,f\rangle}$, and more generally,

$$\|f\|_p = \mathbf{E}[|f(\boldsymbol{x})|^p]^{1/p}.$$

Here we have introduced probabilistic notation that will be used heavily throughout the book:

Notation 1.4. We write $\boldsymbol{x} \sim \{-1,1\}^n$ to denote that $\boldsymbol{x}$ is a uniformly chosen random string from $\{-1,1\}^n$. Equivalently, the n coordinates $\boldsymbol{x}_i$ are independently chosen to be $+1$ with probability $1/2$ and -1 with probability $1/2$. We always write random variables in **boldface**. Probabilities **Pr** and expectations **E** will always be with respect to a uniformly random $\boldsymbol{x} \sim \{-1,1\}^n$ unless otherwise specified. Thus we might write the expectation in (1.8) as $\mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})g(\boldsymbol{x})]$ or $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{x})]$ or even $\mathbf{E}[fg]$.

Returning to the basis of parity functions for V, the crucial fact underlying all analysis of Boolean functions is that this is an *orthonormal basis*.

Theorem 1.5. *The 2^n parity functions $\chi_S : \{-1,1\}^n \to \{-1,1\}$ form an orthonormal basis for the vector space V of functions $\{-1,1\}^n \to \mathbb{R}$; i.e.,*

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

Recalling the definition $\langle \chi_S, \chi_T\rangle = \mathbf{E}[\chi_S(\boldsymbol{x})\chi_T(\boldsymbol{x})]$, Theorem 1.5 follows immediately from two facts:

Fact 1.6. *For $x \in \{-1,1\}^n$ it holds that $\chi_S(x)\chi_T(x) = \chi_{S\triangle T}(x)$, where $S\triangle T$ denotes symmetric difference.*

Proof. $\chi_S(x)\chi_T(x) = \prod_{i\in S} x_i \prod_{i\in T} x_i = \prod_{i\in S\triangle T} x_i \prod_{i\in S\cap T} x_i^2 = \prod_{i\in S\triangle T} x_i = \chi_{S\triangle T}(x).$ □

Fact 1.7. $\mathbf{E}[\chi_S(\boldsymbol{x})] = \mathbf{E}\Big[\prod_{i\in S}\boldsymbol{x}_i\Big] = \begin{cases} 1 & \textit{if } S = \emptyset, \\ 0 & \textit{if } S \neq \emptyset. \end{cases}$

Proof. If $S = \emptyset$ then $\mathbf{E}[\chi_S(\boldsymbol{x})] = \mathbf{E}[1] = 1$. Otherwise,

$$\mathbf{E}\Big[\prod_{i \in S} \boldsymbol{x}_i\Big] = \prod_{i \in S} \mathbf{E}[\boldsymbol{x}_i]$$

because the random bits $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are independent. But each of the factors $\mathbf{E}[\boldsymbol{x}_i]$ in the above (nonempty) product is $(1/2)(+1) + (1/2)(-1) = 0$. □

1.4. Basic Fourier formulas

As we have seen, the Fourier expansion of $f : \{-1,1\}^n \to \mathbb{R}$ can be thought of as the representation of f over the orthonormal basis of parity functions $(\chi_S)_{S \subseteq [n]}$. In this basis, f has 2^n "coordinates", and these are precisely the Fourier coefficients of f. The "coordinate" of f in the χ_S "direction" is $\langle f, \chi_S \rangle$; i.e., we have the following formula for Fourier coefficients:

Proposition 1.8. *For $f : \{-1,1\}^n \to \mathbb{R}$ and $S \subseteq [n]$, the Fourier coefficient of f on S is given by*

$$\widehat{f}(S) = \langle f, \chi_S \rangle = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})].$$

We can verify this formula explicitly:

$$\langle f, \chi_S \rangle = \left\langle \sum_{T \subseteq [n]} \widehat{f}(T)\chi_T, \chi_S \right\rangle = \sum_{T \subseteq [n]} \widehat{f}(T)\langle \chi_T, \chi_S \rangle = \widehat{f}(S), \tag{1.9}$$

where we used the Fourier expansion of f, the linearity of $\langle \cdot, \cdot \rangle$, and finally Theorem 1.5. This formula is the simplest way to calculate the Fourier coefficients of a given function; it can also be viewed as a streamlined version of the interpolation method illustrated in (1.3). Alternatively, this formula can be taken as the *definition* of Fourier coefficients.

The orthonormal basis of parities also lets us measure the squared "length" (2-norm) of $f : \{-1,1\}^n \to \mathbb{R}$ efficiently: it's just the sum of the squares of f's "coordinates" – i.e., Fourier coefficients. This simple but crucial fact is called *Parseval's Theorem*.

Parseval's Theorem. *For any $f : \{-1,1\}^n \to \mathbb{R}$,*

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

In particular, if $f : \{-1,1\}^n \to \{-1,1\}$ is Boolean-valued then

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

As examples we can recall the Fourier expansions of $\max_2$ and Maj_3:

$$\max_2(x) = \tfrac{1}{2} + \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 - \tfrac{1}{2}x_1x_2, \qquad \mathrm{Maj}_3(x) = \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{2}x_3 - \tfrac{1}{2}x_1x_2x_3.$$

In both cases the sum of squares of Fourier coefficients is $4 \times (1/4) = 1$.

More generally, given two functions $f,g : \{-1,1\}^n \to \mathbb{R}$, we can compute $\langle f,g \rangle$ by taking the "dot product" of their coordinates in the orthonormal basis of parities. The resulting formula is called *Plancherel's Theorem*.

Plancherel's Theorem. *For any* $f,g : \{-1,1\}^n \to \mathbb{R}$,

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

We can verify this formula explicitly as we did in (1.9):

$$\langle f,g \rangle = \Big\langle \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S, \sum_{T\subseteq[n]} \widehat{g}(T)\chi_T \Big\rangle = \sum_{S,T\subseteq[n]} \widehat{f}(S)\widehat{g}(T)\langle \chi_S, \chi_T \rangle = \sum_{S\subseteq[n]} \widehat{f}(S)\widehat{g}(S).$$

Now is a good time to remark that for Boolean-valued functions $f,g : \{-1,1\}^n \to \{-1,1\}$, the inner product $\langle f,g \rangle$ can be interpreted as a kind of "correlation" between f and g, measuring how similar they are. Since $f(x)g(x) = 1$ if $f(x) = g(x)$ and $f(x)g(x) = -1$ if $f(x) \neq g(x)$, we have:

Proposition 1.9. *If* $f,g : \{-1,1\}^n \to \{-1,1\}$,

$$\langle f,g \rangle = \mathbf{Pr}[f(\boldsymbol{x}) = g(\boldsymbol{x})] - \mathbf{Pr}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] = 1 - 2\mathrm{dist}(f,g).$$

Here we are using the following definition:

Definition 1.10. Given $f,g : \{-1,1\}^n \to \{-1,1\}$, we define their *(relative Hamming) distance* to be

$$\mathrm{dist}(f,g) = \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})],$$

the fraction of inputs on which they disagree.

With a number of Fourier formulas now in hand we can begin to illustrate a basic theme in the analysis of Boolean functions: interesting combinatorial properties of a Boolean function f can be "read off" from its Fourier coefficients. Let's start by looking at one way to measure the "bias" of f:

Definition 1.11. The *mean* of $f : \{-1,1\}^n \to \mathbb{R}$ is $\mathbf{E}[f]$. When f has mean 0 we say that it is *unbiased*, or *balanced*. In the particular case that $f : \{-1,1\}^n \to \{-1,1\}$ is Boolean-valued, its mean is

$$\mathbf{E}[f] = \mathbf{Pr}[f = 1] - \mathbf{Pr}[f = -1];$$

thus f is unbiased if and only if it takes value 1 on exactly half of the points of the Hamming cube.

Fact 1.12. *If* $f : \{-1,1\}^n \to \mathbb{R}$ *then* $\mathbf{E}[f] = \widehat{f}(\emptyset)$.

This formula holds simply because $\mathbf{E}[f] = \langle f,1 \rangle = \widehat{f}(\emptyset)$ (taking $S = \emptyset$ in Proposition 1.8). In particular, a Boolean function is unbiased if and only if its empty-set Fourier coefficient is 0.

Next we obtain a formula for the *variance* of a real-valued Boolean function (thinking of $f(\boldsymbol{x})$ as a real-valued random variable):

Proposition 1.13. *The* variance *of* $f : \{-1,1\}^n \to \mathbb{R}$ *is*

$$\mathbf{Var}[f] = \langle f - \mathbf{E}[f], f - \mathbf{E}[f] \rangle = \mathbf{E}[f^2] - \mathbf{E}[f]^2 = \sum_{S \neq \emptyset} \widehat{f}(S)^2.$$

This Fourier formula follows immediately from Parseval's Theorem and Fact 1.12.

Fact 1.14. *For* $f : \{-1,1\}^n \to \{-1,1\}$*,*

$$\mathbf{Var}[f] = 1 - \mathbf{E}[f]^2 = 4\,\mathbf{Pr}[f(\boldsymbol{x}) = 1]\,\mathbf{Pr}[f(\boldsymbol{x}) = -1] \in [0,1].$$

In particular, a Boolean-valued function f has variance 1 if it's unbiased and variance 0 if it's constant. More generally, the variance of a Boolean-valued function is proportional to its "distance from being constant".

Proposition 1.15. *Let* $f : \{-1,1\}^n \to \{-1,1\}$*. Then* $2\epsilon \le \mathbf{Var}[f] \le 4\epsilon$*, where*

$$\epsilon = \min\{\mathrm{dist}(f, 1), \mathrm{dist}(f, -1)\}.$$

The proof of Proposition 1.15 is an exercise. See also Exercise 1.17.

By using Plancherel in place of Parseval, we get a generalization of Proposition 1.13 for *covariance*:

Proposition 1.16. *The* covariance *of* $f, g : \{-1,1\}^n \to \mathbb{R}$ *is*

$$\mathbf{Cov}[f,g] = \langle f - \mathbf{E}[f], g - \mathbf{E}[g] \rangle = \mathbf{E}[fg] - \mathbf{E}[f]\mathbf{E}[g] = \sum_{S \neq \emptyset} \widehat{f}(S)\widehat{g}(S).$$

We end this section by discussing the *Fourier weight distribution* of Boolean functions.

Definition 1.17. The *(Fourier) weight* of $f : \{-1,1\}^n \to \mathbb{R}$ on set S is defined to be the squared Fourier coefficient, $\widehat{f}(S)^2$.

Although we lose some information about the Fourier coefficients when we square them, many Fourier formulas only depend on the weights of f. For example, Proposition 1.13 says that the variance of f equals its Fourier weight on nonempty sets. Studying Fourier weights is particularly pleasant for Boolean-valued functions $f : \{-1,1\}^n \to \{-1,1\}$ since Parseval's Theorem says that they always have total weight 1. In particular, they define a *probability distribution* on subsets of $[n]$.

Definition 1.18. Given $f : \{-1,1\}^n \to \{-1,1\}$, the *spectral sample* for f, denoted $\mathscr{S}_f$, is the probability distribution on subsets of $[n]$ in which the set S has probability $\widehat{f}(S)^2$. We write $\boldsymbol{S} \sim \mathscr{S}_f$ for a draw from this distribution.

For example, the spectral sample for the $\max_2$ function is the uniform distribution on all four subsets of $[2]$; the spectral sample for Maj_3 is the uniform distribution on the four subsets of $[3]$ with odd cardinality.

Given a Boolean function it can be helpful to try to keep a mental picture of its weight distribution on the subsets of $[n]$, partially ordered by inclusion. Figure 1.1 is an example for the Maj_3 function, with the white circles indicating weight 0 and the shaded circles indicating weight 1/4.

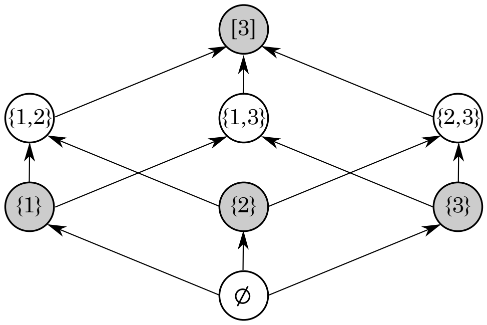


Figure 1.1. Fourier weight distribution of the Maj_3 function

Finally, as suggested by the diagram we often stratify the subsets $S \subseteq [n]$ according to their cardinality (also called "height" or "level"). Equivalently, this is the *degree* of the associated monomial x^S.

Definition 1.19. For $f : \{-1,1\}^n \to \mathbb{R}$ and $0 \le k \le n$, the *(Fourier) weight of f at degree k* is

$$\mathbf{W}^k[f] = \sum_{\substack{S\subseteq[n]\\|S|=k}} \widehat{f}(S)^2.$$

If $f : \{-1,1\}^n \to \{-1,1\}$ is Boolean-valued, an equivalent definition is

$$\mathbf{W}^k[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{S}\sim\mathcal{S}_f}[|\boldsymbol{S}| = k].$$

By Parseval's Theorem, $\mathbf{W}^k[f] = \|f^{=k}\|_2^2$ where

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

is called the *degree k part of f*. We will also sometimes use notation like $\mathbf{W}^{>k}[f] = \sum_{|S|>k} \widehat{f}(S)^2$ and $f^{\le k} = \sum_{|S|\le k} \widehat{f}(S)\,\chi_S$.

1.5. Probability densities and convolution

For variety's sake, in this section we write the Hamming cube as $\mathbb{F}_2^n$ rather than $\{-1,1\}^n$. In developing the Fourier expansion, we have generalized from *Boolean-valued Boolean functions* $f : \mathbb{F}_2^n \to \{-1,1\}$ to *real-valued Boolean functions* $f : \mathbb{F}_2^n \to \mathbb{R}$. Boolean-valued functions arise more often in combinatorial problems, but there are important classes of real-valued Boolean functions. One example is *probability densities*.

Definition 1.20. A *(probability) density* function on the Hamming cube $\mathbb{F}_2^n$ is any nonnegative function $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ satisfying

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[\varphi(\boldsymbol{x})] = 1.$$

We write $\boldsymbol{y} \sim \varphi$ to denote that $\boldsymbol{y}$ is a random string drawn from the associated probability distribution, defined by

$$\mathop{\mathbf{Pr}}_{\boldsymbol{y} \sim \varphi}[\boldsymbol{y} = y] = \varphi(y)\frac{1}{2^n} \quad \forall y \in \mathbb{F}_2^n.$$

Here you should think of $\varphi(y)$ as being the *relative* density of y with respect to the uniform distribution on $\mathbb{F}_2^n$. For example, we have:

Fact 1.21. *If φ is a density function and $g : \mathbb{F}_2^n \to \mathbb{R}$, then*

$$\mathop{\mathbf{E}}_{\boldsymbol{y} \sim \varphi}[g(\boldsymbol{y})] = \langle \varphi, g \rangle = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[\varphi(\boldsymbol{x})g(\boldsymbol{x})].$$

The simplest example of a probability density is just the constant function 1, which corresponds to the uniform probability distribution on $\mathbb{F}_2^n$. The most common case arises from the uniform distribution over some subset $A \subseteq \mathbb{F}_2^n$.

Definition 1.22. If $A \subseteq \mathbb{F}_2^n$ we write $1_A : \mathbb{F}_2^n \to \{0,1\}$ for the 0-1 *indicator function* of A; i.e.,

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Assuming $A \neq \emptyset$ we write φ_A for the density function associated to the uniform distribution on A; i.e.,

$$\varphi_A = \tfrac{1}{\mathbf{E}[1_A]} 1_A.$$

We typically write $\boldsymbol{y} \sim A$ rather than $\boldsymbol{y} \sim \varphi_A$.

A simple but useful example is when A is the singleton set $A = \{0\}$. (Here 0 is denoting the vector $(0,0,\dots,0) \in \mathbb{F}_2^n$.) In this case the function $\varphi_{\{0\}}$ takes value 2^n on input $0 \in \mathbb{F}_2^n$ and is zero elsewhere on $\mathbb{F}_2^n$. In Exercise 1.1 you will verify the Fourier expansion of $\varphi_{\{0\}}$:

Fact 1.23. *Every Fourier coefficient of $\varphi_{\{0\}}$ is 1; i.e., its Fourier expansion is*

$$\varphi_{\{0\}}(y) = \sum_{S \subseteq [n]} \chi_S(y).$$

We now introduce an operation on functions that interacts particularly nicely with density functions, namely, *convolution*.

Definition 1.24. Let $f, g : \mathbb{F}_2^n \to \mathbb{R}$. Their *convolution* is the function $f * g : \mathbb{F}_2^n \to \mathbb{R}$ defined by

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

Since subtraction is equivalent to addition in $\mathbb{F}_2^n$ we may also write

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

If we were representing the Hamming cube by $\{-1, 1\}^n$ rather than $\mathbb{F}_2^n$ we would replace $x + \boldsymbol{y}$ with $x \circ \boldsymbol{y}$, where $\circ$ denotes entry-wise multiplication.

Exercise 1.25 asks you to verify that convolution is associative and commutative:

$$f * (g * h) = (f * g) * h, \qquad f * g = g * f.$$

Using Fact 1.21 we can deduce the following two simple results:

Proposition 1.25. *If φ is a density function on $\mathbb{F}_2^n$ and $g : \mathbb{F}_2^n \to \mathbb{R}$ then*

$$\varphi * g(x) = \mathop{\mathbf{E}}_{\boldsymbol{y} \sim \varphi}[g(x - \boldsymbol{y})] = \mathop{\mathbf{E}}_{\boldsymbol{y} \sim \varphi}[g(x + \boldsymbol{y})].$$

In particular, $\mathbf{E}_{\boldsymbol{y} \sim \varphi}[g(\boldsymbol{y})] = \varphi * g(0)$.

Proposition 1.26. *If $g = \psi$ is itself a probability density function then so is $\varphi * \psi$; it represents the distribution on $\boldsymbol{x} \in \mathbb{F}_2^n$ given by choosing $\boldsymbol{y} \sim \varphi$ and $\boldsymbol{z} \sim \psi$ independently and setting $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$.*

The most important theorem about convolution is that it corresponds to multiplication of Fourier coefficients:

Theorem 1.27. *Let $f, g : \mathbb{F}_2^n \to \mathbb{R}$. Then for all $S \subseteq [n]$,*

$$\widehat{f * g}(S) = \widehat{f}(S)\widehat{g}(S).$$

Proof. We have

$$\begin{aligned}
\widehat{f * g}(S) &= \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[(f * g)(\boldsymbol{x})\chi_S(\boldsymbol{x})] && \text{(the Fourier formula)} \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}\left[\mathop{\mathbf{E}}_{\boldsymbol{y} \sim \mathbb{F}_2^n}[f(\boldsymbol{y})g(\boldsymbol{x} - \boldsymbol{y})]\chi_S(\boldsymbol{x})\right] && \text{(by definition)} \\
&= \mathop{\mathbf{E}}_{\substack{\boldsymbol{y}, \boldsymbol{z} \sim \mathbb{F}_2^n \\ \text{independently}}}[f(\boldsymbol{y})g(\boldsymbol{z})\chi_S(\boldsymbol{y} + \boldsymbol{z})] && \text{(as } x - \boldsymbol{y} \text{ is uniform on } \mathbb{F}_2^n \ \forall x) \\
&= \mathop{\mathbf{E}}_{\boldsymbol{y}, \boldsymbol{z} \sim \mathbb{F}_2^n}[f(\boldsymbol{y})\chi_S(\boldsymbol{y})g(\boldsymbol{z})\chi_S(\boldsymbol{z})] && \text{(by identity (1.5))} \\
&= \widehat{f}(S)\widehat{g}(S) && \text{(Fourier formula, independence),}
\end{aligned}$$

as claimed. □

1.6. Highlight: Almost linear functions and the BLR Test

In linear algebra there are two equivalent definitions of what it means for a function to be linear:

Definition 1.28. A function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is *linear* if either of the following equivalent conditions hold:

(1) $f(x+y) = f(x) + f(y)$ for all $x, y \in \mathbb{F}_2^n$;

(2) $f(x) = a \cdot x$ for some $a \in \mathbb{F}_2^n$; i.e., $f(x) = \sum_{i \in S} x_i$ for some $S \subseteq [n]$.

Exercise 1.26 asks you to verify that the conditions are indeed equivalent. If we encode the output of f by $\pm 1 \in \mathbb{R}$ in the usual way then the "linear" functions $f : \mathbb{F}_2^n \to \{-1, 1\}$ are precisely the 2^n parity functions $(\chi_S)_{S \subseteq [n]}$.

Let's think of what it might mean for a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ to be *approximately* linear. Definition 1.28 suggests two possibilities:

(1′) $f(x+y) = f(x) + f(y)$ for *almost all* pairs $x, y \in \mathbb{F}_2^n$;

(2′) there is some $S \subseteq [n]$ such that $f(x) = \sum_{i \in S} x_i$ for *almost all* $x \in \mathbb{F}_2^n$.

Are these equivalent? The proof of (2) $\implies$ (1) in Definition 1.28 is "robust": it easily extends to show (2′) $\implies$ (1′) (see Exercise 1.26). But the natural proof of (1) $\implies$ (2) in Definition 1.28 does not have this robustness property. The goal of this section is to show that (1′) $\implies$ (2′) nevertheless holds.

Motivation for this problem comes from an area of theoretical computer science called *property testing*, which we will discuss in more detail in Chapter 7. Imagine that you have "black-box" access to a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$, meaning that the function f is unknown to you but you can "query" its value on inputs $x \in \mathbb{F}_2^n$ of your choosing. The function f is "supposed" to be a linear function, and you would like to try to verify this.

The only way you can be certain f is indeed a linear function is to query its value on all 2^n inputs; unfortunately, this is very expensive. The idea behind "property testing" is to try to verify that f has a certain property – in this case, linearity – by querying its value on just a few random inputs. In exchange for efficiency, we need to be willing to only approximately verify the property.

Definition 1.29. If f and g are Boolean-valued functions we say they are *ϵ-close* if $\mathrm{dist}(f, g) \le \epsilon$; otherwise we say they are *ϵ-far*. If $\mathscr{P}$ is a (nonempty) property of n-bit Boolean functions we define $\mathrm{dist}(f, \mathscr{P}) = \min_{g \in \mathscr{P}}\{\mathrm{dist}(f, g)\}$. We say that f is ϵ-close to $\mathscr{P}$ if $\mathrm{dist}(f, \mathscr{P}) \le \epsilon$; i.e., f is ϵ-close to some g satisfying $\mathscr{P}$.

In particular, in property testing we take property (2′) above to be the notion of "approximately linear": we say f is ϵ-close to being linear if $\mathrm{dist}(f,g) \le \epsilon$ for some truly linear $g(x) = \sum_{i\in S} x_i$.

In 1990 Blum, Luby, and Rubinfeld [**BLR90**] showed that indeed (1′) $\Longrightarrow$ (2′) holds, giving the following "test" for the property of linearity that makes just 3 queries:

BLR Test. *Given query access to $f : \mathbb{F}_2^n \to \mathbb{F}_2$:*

- *Choose $\boldsymbol{x} \sim \mathbb{F}_2^n$ and $\boldsymbol{y} \sim \mathbb{F}_2^n$ independently.*
- *Query f at $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{x}+\boldsymbol{y}$.*
- *"Accept" if $f(\boldsymbol{x}) + f(\boldsymbol{y}) = f(\boldsymbol{x}+\boldsymbol{y})$.*

We now show that if the BLR Test accepts f with high probability then f is close to being linear. The proof works by directly relating the acceptance probability to the quantity $\sum_S \widehat{f}(S)^3$; see equation (1.10) below.

Theorem 1.30. *Suppose the BLR Test accepts $f : \mathbb{F}_2^n \to \mathbb{F}_2$ with probability $1-\epsilon$. Then f is ϵ-close to being linear.*

Proof. In order to use the Fourier transform we encode f's output by $\pm 1 \in \mathbb{R}$; thus the acceptance condition of the BLR Test becomes $f(\boldsymbol{x})f(\boldsymbol{y}) = f(\boldsymbol{x}+\boldsymbol{y})$. Since

$$\tfrac{1}{2} + \tfrac{1}{2} f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{x}+\boldsymbol{y}) = \begin{cases} 1 & \text{if } f(\boldsymbol{x})f(\boldsymbol{y}) = f(\boldsymbol{x}+\boldsymbol{y}), \\ 0 & \text{if } f(\boldsymbol{x})f(\boldsymbol{y}) \neq f(\boldsymbol{x}+\boldsymbol{y}), \end{cases}$$

we conclude

$$\begin{aligned} 1-\epsilon = \mathbf{Pr}[\text{BLR accepts } f] &= \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}}[\tfrac{1}{2} + \tfrac{1}{2} f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{x}+\boldsymbol{y})] \\ &= \tfrac{1}{2} + \tfrac{1}{2} \mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \cdot \mathop{\mathbf{E}}_{\boldsymbol{y}}[f(\boldsymbol{y})f(\boldsymbol{x}+\boldsymbol{y})]] \\ &= \tfrac{1}{2} + \tfrac{1}{2} \mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \cdot (f * f)(\boldsymbol{x})] && \text{(by definition)} \\ &= \tfrac{1}{2} + \tfrac{1}{2} \sum_{S\subseteq[n]} \widehat{f}(S)\widehat{f*f}(S) && \text{(Plancherel)} \\ &= \tfrac{1}{2} + \tfrac{1}{2} \sum_{S\subseteq[n]} \widehat{f}(S)^3 && \text{(Theorem 1.27).} \end{aligned}$$

We rearrange this equality and then continue:

$$\begin{aligned} 1-2\epsilon &= \sum_{S\subseteq[n]} \widehat{f}(S)^3 && (1.10) \\ &\le \max_{S\subseteq[n]}\{\widehat{f}(S)\} \cdot \sum_{S\subseteq[n]} \widehat{f}(S)^2 \\ &= \max_{S\subseteq[n]}\{\widehat{f}(S)\} && \text{(Parseval).} \end{aligned}$$

But $\widehat{f}(S) = \langle f, \chi_S \rangle = 1 - 2\mathrm{dist}(f, \chi_S)$ (Proposition 1.9). Hence there exists some $S^* \subseteq [n]$ such that $1-2\epsilon \le 1-2\mathrm{dist}(f, \chi_{S^*})$; i.e., f is ϵ-close to the linear function χ_{S^*}. □

In fact, for small ϵ one can show that f is more like $(\epsilon/3)$-close to linear, and this is sharp. See Exercise 1.28.

The BLR Test shows that given black-box access to $f : \mathbb{F}_2^n \to \{-1,1\}$, we can "test" whether f is close to some linear function χ_S using just 3 queries. The test does not reveal *which* linear function χ_S is close to (indeed, determining this takes at least n queries; see Exercise 1.27). Nevertheless, we can still determine the value of $\chi_S(x)$ with high probability for *every* $x \in \mathbb{F}_2^n$ of our choosing using just 2 queries. This property is called *local correctability* of linear functions.

Proposition 1.31. *Suppose $f : \mathbb{F}_2^n \to \{-1,1\}$ is ϵ-close to the linear function χ_S. Then for every $x \in \mathbb{F}_2^n$, the following algorithm outputs $\chi_S(x)$ with probability at least $1-2\epsilon$:*

- *Choose $\boldsymbol{y} \sim \mathbb{F}_2^n$.*
- *Query f at $\boldsymbol{y}$ and $x+\boldsymbol{y}$.*
- *Output $f(\boldsymbol{y})f(x+\boldsymbol{y})$.*

We emphasize the order of quantifiers here: if we just output $f(x)$ then this will equal $\chi_S(x)$ for *most* x; however, the above "local correcting" algorithm determines $\chi_S(x)$ (with high probability) for *every* x.

Proof. Since $\boldsymbol{y}$ and $x+\boldsymbol{y}$ are both uniformly distributed on $\mathbb{F}_2^n$ (though not independently) we have $\mathbf{Pr}[f(\boldsymbol{y}) \neq \chi_S(\boldsymbol{y})] \le \epsilon$ and $\mathbf{Pr}[f(x+\boldsymbol{y}) \neq \chi_S(x+\boldsymbol{y})] \le \epsilon$ by assumption. By the union bound, the probability of either event occurring is at most 2ϵ; when neither occurs,

$$f(\boldsymbol{y})f(x+\boldsymbol{y}) = \chi_S(\boldsymbol{y})\chi_S(x+\boldsymbol{y}) = \chi_S(x)$$

as desired. □

1.7. Exercises and notes

1.1 Compute the Fourier expansions of the following functions:

(*a*) $\mathrm{min}_2 : \{-1,1\}^2 \to \{-1,1\}$, the minimum function on 2 bits (also known as the logical OR function);

(*b*) $\mathrm{min}_3 : \{-1,1\}^3 \to \{-1,1\}$ and $\mathrm{max}_3 : \{-1,1\}^3 \to \{-1,1\}$;

(*c*) the indicator function $1_{\{a\}} : \mathbb{F}_2^n \to \{0,1\}$, where $a \in \mathbb{F}_2^n$;

(*d*) the density function $\varphi_{\{a\}} : \mathbb{F}_2^n \to \mathbb{R}^{\ge 0}$, where $a \in \mathbb{F}_2^n$;

(*e*) the density function $\varphi_{\{a,a+e_i\}} : \mathbb{F}_2^n \to \mathbb{R}^{\ge 0}$, where $a \in \mathbb{F}_2^n$ and $e_i = (0,\dots,0,1,0,\dots,0)$ with the 1 in the ith coordinate;

(f) the density function corresponding to the product probability distribution on $\{-1,1\}^n$ in which each coordinate has mean $\rho \in [-1,1]$;

(g) the *inner product mod 2 function*, $\mathrm{IP}_{2n} : \mathbb{F}_2^{2n} \to \{-1,1\}$ defined by $\mathrm{IP}_{2n}(x_1,\dots,x_n,y_1,\dots,y_n) = (-1)^{x\cdot y}$;

(h) the *equality function* $\mathrm{Equ}_n : \{-1,1\}^n \to \{0,1\}$, defined by $\mathrm{Equ}_n(x) = 1$ if and only if $x_1 = x_2 = \cdots = x_n$;

(i) the *not-all-equal function* $\mathrm{NAE}_n : \{-1,1\}^n \to \{0,1\}$, defined by $\mathrm{NAE}_n(x) = 1$ if and only if the bits $x_1,\dots,x_n$ are not all equal;

(j) the *selection function*, $\mathrm{Sel} : \{-1,1\}^3 \to \{-1,1\}$, which outputs x_2 if $x_1 = -1$ and outputs x_3 if $x_1 = 1$;

(k) $\mathrm{mod}_3 : \mathbb{F}_2^3 \to \{0,1\}$, which is 1 if and only if the number of 1's in the input is divisible by 3;

(l) $\mathrm{OXR} : \mathbb{F}_2^3 \to \{0,1\}$ defined by $\mathrm{OXR}(x_1,x_2,x_3) = x_1 \vee (x_2 \oplus x_3)$. Here $\vee$ denotes logical OR, $\oplus$ denotes logical XOR;

(m) the *sortedness function* $\mathrm{Sort}_4 : \{-1,1\}^4 \to \{-1,1\}$, defined by $\mathrm{Sort}_4(x) = -1$ if and only if $x_1 \le x_2 \le x_3 \le x_4$ or $x_1 \ge x_2 \ge x_3 \ge x_4$;

(n) the *hemi-icosahedron function* $\mathrm{HI} : \{-1,1\}^6 \to \{-1,1\}$ (also known as the *Kushilevitz function*), defined to be the number of facets labeled $(+1,+1,+1)$ in Figure 1.2, minus the number of facets labeled $(-1,-1,-1)$, modulo 3.

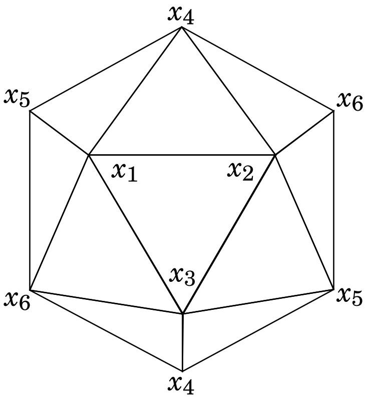


Figure 1.2. The hemi-icosahedron

(Hint: First compute the real multilinear interpolation of the analogue $\mathrm{HI} : \{0,1\}^6 \to \{0,1\}$.)

(o) the majority functions $\mathrm{Maj}_5 : \{-1,1\}^5 \to \{-1,1\}$ and $\mathrm{Maj}_7 : \{-1,1\}^7 \to \{-1,1\}$;

(p) the *complete quadratic function* $\mathrm{CQ}_n : \mathbb{F}_2^n \to \{-1,1\}$ defined by $\mathrm{CQ}_n(x) = \chi(\sum_{1\le i<j\le n} x_i x_j)$. (Hint: Determine $\mathrm{CQ}_n(x)$ as a function of the number of 1's in the input modulo 4. You'll want to distinguish whether n is even or odd.)

1.2 How many Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ have exactly 1 nonzero Fourier coefficient?

1.3 Let $f : \mathbb{F}_2^n \to \{0,1\}$ and suppose $\#\{x : f(x) = 1\}$ is odd. Prove that all of f's Fourier coefficients are nonzero.

1.4 Let $f : \{-1,1\}^n \to \mathbb{R}$ have Fourier expansion $f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) x^S$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be the extension of f which is also defined by $F(x) = \sum_{S \subseteq [n]} \widehat{f}(S) x^S$. Show that if $\mu = (\mu_1, \dots, \mu_n) \in [-1,1]^n$ then

$$F(\mu) = \mathop{\mathbf{E}}_{\boldsymbol{y}}[f(\boldsymbol{y})],$$

where $\boldsymbol{y}$ is the random string in $\{-1,1\}^n$ defined by having $\mathbf{E}[\boldsymbol{y}_i] = \mu_i$ independently for all $i \in [n]$.

1.5 Prove that any $f : \{-1,1\}^n \to \{-1,1\}$ has at most one Fourier coefficient with magnitude exceeding 1/2. Is this also true for any $f : \{-1,1\}^n \to \mathbb{R}$ with $\|f\|_2 = 1$?

1.6 Use Parseval's Theorem to prove uniqueness of the Fourier expansion.

1.7 Let $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ be a random function (i.e., each $\boldsymbol{f}(x)$ is ± 1 with probability 1/2, independently for all $x \in \{-1,1\}^n$). Show that for each $S \subseteq [n]$, the random variable $\widehat{\boldsymbol{f}}(S)$ has mean 0 and variance 2^{-n}. (Hint: Parseval.)

1.8 The *(Boolean) dual* of $f : \{-1,1\}^n \to \mathbb{R}$ is the function $f^\dagger$ defined by $f^\dagger(x) = -f(-x)$. The function f is said to be *odd* if it equals its dual; equivalently, if $f(-x) = -f(x)$ for all x. The function f is said to be *even* if $f(-x) = f(x)$ for all x. Given any function $f : \{-1,1\}^n \to \mathbb{R}$, its *odd part* is the function $f^{\text{odd}} : \{-1,1\}^n \to \mathbb{R}$ defined by $f^{\text{odd}}(x) = (f(x) - f(-x))/2$, and its *even part* is the function $f^{\text{even}} : \{-1,1\}^n \to \mathbb{R}$ defined by $f^{\text{even}}(x) = (f(x) + f(-x))/2$.

(a) Express $\widehat{f^\dagger}(S)$ in terms of $\widehat{f}(S)$.

(b) Verify that $f = f^{\text{odd}} + f^{\text{even}}$ and that f is odd (respectively, even) if and only if $f = f^{\text{odd}}$ (respectively, $f = f^{\text{even}}$).

(c) Show that

$$f^{\text{odd}} = \sum_{\substack{S \subseteq [n] \\ |S| \text{ odd}}} \widehat{f}(S) \chi_S, \qquad f^{\text{even}} = \sum_{\substack{S \subseteq [n] \\ |S| \text{ even}}} \widehat{f}(S) \chi_S.$$

1.9 In this problem we consider representing $\mathsf{False}, \mathsf{True}$ as $0, 1 \in \mathbb{R}$.

(a) Using the interpolation method from Section 1.2, show that every $f : \{\mathsf{False}, \mathsf{True}\}^n \to \{\mathsf{False}, \mathsf{True}\}$ can be represented as a real multilinear polynomial

$$q(x) = \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i, \tag{1.11}$$

"over $\{0,1\}$", meaning mapping $\{0,1\}^n \to \{0,1\}$.

(*b*) Show that this representation is unique. (Hint: If q as in (1.11) has at least one nonzero coefficient, consider $q(a)$ where $a \in \{0,1\}^n$ is the indicator vector of a minimal S with $c_S \neq 0$.)

(*c*) Show that all coefficients c_S in the representation (1.11) will be integers in the range $[-2^n, 2^n]$.

(*d*) Let $f : \{\mathsf{False}, \mathsf{True}\}^n \to \{\mathsf{False}, \mathsf{True}\}$. Let $p(x)$ be f's multilinear representation when $\mathsf{False}, \mathsf{True}$ are $1, -1 \in \mathbb{R}$ (i.e., p is the Fourier expansion of f) and let $q(x)$ be f's multilinear representation when $\mathsf{False}, \mathsf{True}$ are $0, 1 \in \mathbb{R}$. Show that $q(x) = \frac{1}{2} - \frac{1}{2}p(1-2x_1, \dots, 1-2x_n)$.

1.10 Let $f : \{-1,1\}^n \to \mathbb{R}$ be not identically 0. The *(real) degree* of f, denoted $\deg(f)$, is defined to be the degree of its multilinear (Fourier) expansion; i.e., $\max\{|S| : \widehat{f}(S) \neq 0\}$.

(*a*) Show that $\deg(f) = \deg(a + bf)$ for any $a, b \in \mathbb{R}$ (assuming $b \neq 0$, $a + bf \neq 0$).

(*b*) Show that $\deg(f) \leq k$ if and only if f is a real linear combination of functions $g_1, \dots, g_s$, each of which depends on at most k input coordinates.

(*c*) Which functions in Exercise 1.1 have "nontrivial" degree? (Here $f : \{-1,1\}^n \to \mathbb{R}$ has "nontrivial" degree if $\deg(f) < n$.)

1.11 Suppose that $f : \{-1,1\}^n \to \{-1,1\}$ has $\deg(f) = k \geq 1$.

(*a*) Show that f's real multilinear representation over $\{0,1\}$ (see Exercise 1.9), call it $q(x)$, also has $\deg(q) = k$.

(*b*) Using Exercise 1.9(*c*),(*d*), deduce that f's Fourier spectrum is "2^{1-k}-granular", meaning each $\widehat{f}(S)$ is an integer multiple of 2^{1-k}.

(*c*) Show that $\sum_{S \subseteq [n]} |\widehat{f}(S)| \leq 2^{k-1}$.

1.12 A *Hadamard Matrix* is any $N \times N$ real matrix with ± 1 entries and orthogonal rows. Particular examples are the *Walsh–Hadamard Matrices* H_N, inductively defined for $N = 2^n$ as follows: $H_1 = \begin{bmatrix} 1 \end{bmatrix}$, $H_{2^{n+1}} = \begin{bmatrix} H_{2^n} & H_{2^n} \\ H_{2^n} & -H_{2^n} \end{bmatrix}$.

(*a*) Let's index the rows and columns of H_{2^n} by the integers $\{0, 1, 2, \dots, 2^n - 1\}$ rather than $[2^n]$. Further, let's identify such an integer i with its binary expansion $(i_0, i_1, \dots, i_{n-1}) \in \mathbb{F}_2^n$, where i_0 is the least significant bit and i_{n-1} the most. For example, if $n = 3$, we identify the index $i = 6$ with $(0,1,1)$. Now show that the (γ, x) entry of H_{2^n} is $(-1)^{\gamma \cdot x}$.

(*b*) Show that if $f : \mathbb{F}_2^n \to \mathbb{R}$ is represented as a column vector in $\mathbb{R}^{2^n}$ (according to the indexing scheme from part (*a*)) then $2^{-n} H_{2^n} f = \widehat{f}$. Here we think of $\widehat{f}$ as also being a function $\mathbb{F}_2^n \to \mathbb{R}$, identifying subsets $S \subseteq \{0, 1, \dots, n-1\}$ with their indicator vectors.

(*c*) Show how to compute $H_{2^n} f$ using just $n 2^n$ additions and subtractions (rather than 2^{2n} additions and subtractions as the usual matrix-vector multiplication algorithm would require). This computation is called the *Fast Walsh–Hadamard Transform* and is the method of choice for computing the Fourier expansion of a generic function $f : \mathbb{F}_2^n \to \mathbb{R}$ when n is large.

(*d*) Show that taking the Fourier transform is essentially an "involution": $\widehat{\widehat{f}} = 2^{-n} f$ (using the notations from part (*b*)).

1.13 Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $0 < p \le q < \infty$. Show that $\|f\|_p \le \|f\|_q$. (Hint: Use Jensen's inequality with the convex function $t \mapsto t^{q/p}$.) Extend the inequality to the case $q = \infty$, where $\|f\|_\infty$ is defined to be $\max_{x \in \{-1,1\}^n} \{|f(x)|\}$.

1.14 Compute the mean and variance of each function from Exercise 1.1.

1.15 Let $f : \{-1,1\}^n \to \mathbb{R}$. Let $K \subseteq [n]$ and let $z \in \{-1,1\}^K$. Suppose $g : \{-1,1\}^{[n]\setminus K} \to \mathbb{R}$ is the subfunction of f gotten by restricting the K-coordinates to be z. Show that $\mathbf{E}[g] = \sum_{T \subseteq K} \widehat{f}(T) z^T$.

1.16 If $f : \{-1,1\}^n \to \{-1,1\}$, show that $\mathbf{Var}[f] = 4 \cdot \mathrm{dist}(f,1) \cdot \mathrm{dist}(f,-1)$. Deduce Proposition 1.15.

1.17 Extend Fact 1.14 by proving the following: If $\boldsymbol{F}$ is a $\{-1,1\}$-valued random variable with mean μ then

$$\mathbf{Var}[\boldsymbol{F}] = \mathbf{E}[(\boldsymbol{F}-\mu)^2] = \tfrac{1}{2}\mathbf{E}[(\boldsymbol{F}-\boldsymbol{F}')^2] = 2\,\mathbf{Pr}[\boldsymbol{F} \neq \boldsymbol{F}'] = \mathbf{E}[|\boldsymbol{F}-\mu|],$$

where $\boldsymbol{F}'$ is an independent copy of $\boldsymbol{F}$.

1.18 For any $f : \{-1,1\}^n \to \mathbb{R}$, show that

$$\langle f^{=k}, f^{=\ell} \rangle = \begin{cases} \mathbf{W}^k[f] & \text{if } k = \ell, \\ 0 & \text{if } k \neq \ell. \end{cases}$$

1.19 Let $f : \{-1,1\}^n \to \{-1,1\}$.

(*a*) Suppose $\mathbf{W}^1[f] = 1$. Show that $f(x) = \pm\chi_S$ for some $|S| = 1$.

(*b*) Suppose $\mathbf{W}^{\le 1}[f] = 1$. Show that f depends on at most 1 input coordinate.

(*c*) Suppose $\mathbf{W}^{\le 2}[f] = 1$. Must f depend on at most 2 input coordinates? At most 3 input coordinates? What if we assume $\mathbf{W}^2[f] = 1$?

1.20 Let $f : \{-1,1\}^n \to \mathbb{R}$ satisfy $f = f^{=1}$. Show that $\mathbf{Var}[f^2] = 2\sum_{i \neq j} \widehat{f}(i)^2 \widehat{f}(j)^2$.

1.21 Prove that there are no functions $f : \{-1,1\}^n \to \{-1,1\}$ with exactly 2 nonzero Fourier coefficients. What about exactly 3 nonzero Fourier coefficients?

1.22 Verify Propositions 1.25 and 1.26.

1.23 In this exercise you will prove some basic facts about "distances" between probability distributions. Let φ and ψ be probability densities on $\mathbb{F}_2^n$.

(*a*) Show that the *total variation distance* between φ and ψ, defined by

$$d_{\mathrm{TV}}(\varphi,\psi) = \max_{A \subseteq \mathbb{F}_2^n} \left\{ \left| \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim\varphi}[\boldsymbol{y} \in A] - \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim\psi}[\boldsymbol{y} \in A] \right| \right\},$$

is equal to $\frac{1}{2}\|\varphi - \psi\|_1$.

(*b*) Show that the *collision probability* of φ, defined to be

$$\mathop{\mathbf{Pr}}_{\substack{\boldsymbol{y},\boldsymbol{y}' \sim \varphi \\ \text{independently}}}[\boldsymbol{y} = \boldsymbol{y}'],$$

is equal to $\|\varphi\|_2^2/2^n$.

(*c*) The χ^2*-distance of* φ *from* ψ is defined by

$$d_{\chi^2}(\varphi,\psi) = \mathop{\mathbf{E}}_{\boldsymbol{y}\sim\psi}\left[\left(\frac{\varphi(\boldsymbol{y})}{\psi(\boldsymbol{y})} - 1\right)^2\right],$$

assuming ψ has full support. Show that the χ^2-distance of φ from uniform is equal to $\mathbf{Var}[\varphi]$.

(*d*) Show that the total variation distance of φ from uniform is at most $\frac{1}{2}\sqrt{\mathbf{Var}[\varphi]}$.

1.24 Let $A \subseteq \{-1,1\}^n$ have "volume" δ, meaning $\mathbf{E}[1_A] = \delta$. Suppose φ is a probability density *supported* on A, meaning $\varphi(x) = 0$ when $x \notin A$. Show that $\|\varphi\|_2^2 \geq 1/\delta$ with equality if $\varphi = \varphi_A$, the uniform density on A.

1.25 Show directly from the definition that the convolution operator is associative and commutative.

1.26 Verify that (1) $\iff$ (2) in Definition 1.28.

1.27 Suppose an algorithm is given query access to a linear function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ and its task is to determine *which* linear function f is. Show that querying f on n inputs is necessary and sufficient.

1.28 (*a*) Generalize Exercise 1.5 as follows: Let $f : \mathbb{F}_2^n \to \{-1,1\}$ and suppose that $\text{dist}(f, \chi_{S^*}) = \delta$. Show that $|\widehat{f}(S)| \leq 2\delta$ for all $S \neq S^*$. (Hint: Use the union bound.)

(*b*) Deduce that the BLR Test rejects f with probability at least $3\delta - 10\delta^2 + 8\delta^3$.

(*c*) Show that this lower bound cannot be improved to $c\delta - O(\delta^2)$ for any $c > 3$.

1.29 (*a*) We call $f : \mathbb{F}_2^n \to \mathbb{F}_2$ an *affine* function if $f(x) = a \cdot x + b$ for some $a \in \mathbb{F}_2^n$, $b \in \mathbb{F}_2$. Show that f is affine if and only if $f(x)+f(y)+f(z) = f(x+y+z)$ for all $x,y,z, \in \mathbb{F}_2^n$

(*b*) Let $f : \mathbb{F}_2^n \to \mathbb{R}$. Suppose we choose $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \sim \mathbb{F}_2^n$ independently and uniformly. Show that $\mathbf{E}[f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z})f(\boldsymbol{x}+\boldsymbol{y}+\boldsymbol{z})] = \sum_S \widehat{f}(S)^4$.

(*c*) Give a 4-query test for a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ with the following property: if the test accepts with probability $1-\epsilon$ then f is ϵ-close to being affine. All four query inputs should have the uniform distribution on $\mathbb{F}_2^n$ (but of course need not be independent).

(*d*) Give an alternate 4-query test for being affine in which three of the query inputs are uniformly distributed and the fourth is not random.

(Hint: Show that f is affine if and only if $f(x)+f(y)+f(0)=f(x+y)$ for all $x,y\in\mathbb{F}_2^n$.)

1.30 Permutations $\pi\in S_n$ act on strings $x\in\{-1,1\}^n$ in the natural way: $(x^\pi)_i = x_{\pi(i)}$. They also act on functions $f:\{-1,1\}^n\to\mathbb{R}$ via $f^\pi(x)=f(x^\pi)$ for all $x\in\{-1,1\}^n$. We say that functions $g,\ h:\{-1,1\}^n\to\{-1,1\}$ are *(permutation-)isomorphic* if $g=h^\pi$ for some $\pi\in S_n$. We call $\mathrm{Aut}(f)=\{\pi\in S_n : f^\pi=f\}$ the *(permutation-)automorphism group* of f.

(*a*) Show that $\widehat{f^\pi}(S)=\widehat{f}(\pi^{-1}(S))$ for all $S\subseteq[n]$.

For future reference, when we write $(\widehat{f}(S))_{|S|=k}$, we mean the sequence of degree-k Fourier coefficients of f, listed in lexicographic order of the k-sets S.

Given complete truth tables of some g and h we might wish to determine whether they are isomorphic. One way to do this would be to define a *canonical form* $\mathrm{can}(f):\{-1,1\}^n\to\{-1,1\}$ for each $f:\{-1,1\}^n\to\{-1,1\}$, meaning that: (i) $\mathrm{can}(f)$ is isomorphic to f; (ii) if g is isomorphic to h then $\mathrm{can}(g)=\mathrm{can}(h)$. Then we can determine whether g is isomorphic to h by checking whether $\mathrm{can}(g)=\mathrm{can}(h)$. Here is one possible way to define a canonical form for f:

```
1. Set $P_0 = S_n$.
2. For each $k = 1,2,3,\dots,n$,
3.     Define $P_k$ to be the set of all $\pi \in P_{k-1}$ that make the sequence
       $(\widehat{f^\pi}(S))_{|S|=k}$ maximal in lexicographic order on $\mathbb{R}^{\binom{n}{k}}$.
4. Let $\mathrm{can}(f) = f^\pi$ for (any) $\pi \in P_n$.
```

(*b*) Show that this is well-defined, meaning that $\mathrm{can}(f)$ is the same function for any choice of $\pi\in P_n$.

(*c*) Show that $\mathrm{can}(f)$ is indeed a canonical form; i.e., it satisfies (i) and (ii) above.

(*d*) Show that if $\widehat{f}(\{1\}),\dots,\widehat{f}(\{n\})$ are distinct numbers then $\mathrm{can}(f)$ can be computed in $\widetilde{O}(2^n)$ time.

(*e*) We could more generally consider $g,h:\{-1,1\}^n\to\{-1,1\}$ to be isomorphic if $g(x)=h(\pm x_{\pi(1)},\dots,\pm x_{\pi(n)})$ for some permutation π on $[n]$ and some choice of signs. Extend the results of this exercise to handle this definition.

Notes. The Fourier expansion for real-valued Boolean functions dates back to Walsh [**Wal23**] who introduced a complete orthonormal basis for $L^2([0,1])$ consisting of ±1-valued functions, constant on dyadic intervals. Using the ordering introduced by Paley [**Pal32**], the nth Walsh basis function $w_n:[0,1]\to\{-1,1\}$ is defined by $w_n(x)=\prod_{i=0}^\infty r_i(x)^{n_i}$, where $n=\sum_{i=0}^\infty n_i 2^i$ and $r_i(x)$ (the "ith Rademacher function at x") is defined to be $(-1)^{x_i}$, with $x=\sum_{i=0}^\infty x_i 2^{-(i+1)}$ for non-dyadic $x\in[0,1]$. Walsh's interest was in comparing and contrasting the properties of this basis with the usual basis of trigonometric polynomials and also Haar's basis [Haa10].

The first major study of the Walsh functions came in the remarkable paper of Paley [Pal32], which included strong results on the L^p-norms of truncations of Walsh series. Sadly, Paley died in an avalanche one year later (at age 26) while skiing near Banff. The next major development in the study of Walsh series was conceptual, with Vilenkin [Vil47] and Fine [Fin49] independently suggesting the more natural viewpoint of the Walsh functions as characters of the discrete group $\mathbb{Z}_2^n$. There was significant subsequent work in the 1950s and 1960s, but it's somewhat unnatural from our point of view because it relies fundamentally on ordering the Rademacher and Walsh functions according to binary expansions. Bonami [Bon68] and Kiener [Kie69] seem to have been the first authors to take our viewpoint, treating bits $x_1, x_2, x_3, \dots$ symmetrically and ordering Fourier characters χ_S according to $|S|$ rather than $\max(S)$. Bonami also obtained the first *hypercontractivity* result for the Boolean cube. This proved to be a crucial tool for analysis of Boolean functions; see Chapter 9. For an early survey on Walsh series, see Balashov and Rubinshtein [BR73].

Turning to Boolean functions and computer science, the idea of using Boolean logic to study "switching functions" (as engineers originally called Boolean functions) dates to the late 1930s and is usually credited to Nakashima [Nak35], Shannon [Sha37], and Shestakov [She38]. Muller [Mul54b] seems to be the first to have used Fourier coefficients in the study of Boolean functions; he mentions computing them while classifying all functions $f : \{0,1\}^4 \to \{0,1\}$ up to certain equivalences. The first publication devoted to Boolean Fourier coefficients was by Ninomiya [Nin58], who expanded on Muller's use of Fourier coefficients for the classification of Boolean functions up to various isomorphisms. Golomb [Gol59] independently pursued the same project (his work is the content of Exercise 1.30); he was also the first to recognize the connection to Walsh series. The use of "Fourier–Walsh analysis" in the study of Boolean functions quickly became well known in the early 1960s. Several symposia on applications of Walsh functions took place in the early 1970s, with Lechner's 1971 monograph [Lec71] and Karpovsky's 1976 book [Kar76] becoming the standard references. However, the use of Boolean analysis in theoretical computer science seemed to wane until 1988, when the outstanding work of Kahn, Kalai, and Linial [KKL88] ushered in a new area of sophistication.

The original analysis by Blum, Luby, and Rubinfeld [BLR90] for their linearity test was combinatorial; our proof of Theorem 1.30 is the elegant analytic one due to by Bellare, Coppersmith, Håstad, Kiwi, and Sudan [BCH$^+$96]. In fact, the essence of this analysis appears already in the 1953 work of

Roth [**Rot53**] (in the context of the cyclic group $\mathbb{Z}_N$ rather than $\mathbb{F}_2^n$). The work of Bellare et al. also gives additional analysis improving the results of Theorem 1.30 and Exercise 1.28. See also the work of Kaufman, Litsyn, and Xie [**KLX10**] for further slight improvement.

In Exercise 1.1, the sortedness function was introduced by Ambainis [**Amb03, LLS06**]; the hemi-icosahedron function was introduced by Kushilevitz [**NW95**]. The fast algorithm for computing the Fourier transform mentioned in Exercise 1.12 is due to Lechner [**Lec63**].

Chapter 2

Basic concepts and social choice

In this chapter we introduce a number of important basic concepts including influences and noise stability. Many of these concepts are nicely motivated using the language of *social choice*. The chapter is concluded with Kalai's Fourier-based proof of Arrow's Theorem.

2.1. Social choice functions

In this section we describe some rudiments of the mathematics of *social choice*, a topic studied by economists, political scientists, mathematicians, and computer scientists. The fundamental question in this area is how best to *aggregate* the opinions of many agents. Examples where this problem arises include citizens voting in an election, committees deciding on alternatives, and independent computational agents making collective decisions. Social choice theory also provides very appealing interpretations for a number of important functions and concepts in the analysis of Boolean functions.

A Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ can be thought of as a *voting rule* or *social choice function* for an election with 2 candidates and n voters; it maps the votes of the voters to the winner of the election. Perhaps the most familiar voting rule is the majority function:

Definition 2.1. For n odd, *the majority* function $\mathrm{Maj}_n : \{-1,1\}^n \to \{-1,1\}$ is defined by $\mathrm{Maj}_n(x) = \mathrm{sgn}(x_1 + x_2 + \cdots + x_n)$. (Occasionally, for n even we say that f is *a* majority function if $f(x)$ equals the sign of $x_1 + \cdots + x_n$ whenever this number is nonzero.)

The Boolean AND and OR functions correspond to voting rules in which a certain candidate is always elected unless all voters are unanimously opposed. Recalling our somewhat nonintuitive convention that -1 represents True and $+1$ represents False:

Definition 2.2. The function $\mathrm{AND}_n : \{-1,1\}^n \to \{-1,1\}$ is defined by $\mathrm{AND}_n(x) = +1$ unless $x = (-1,-1,\dots,-1)$. The function $\mathrm{OR}_n : \{-1,1\}^n \to \{-1,1\}$ is defined by $\mathrm{OR}_n(x) = -1$ unless $x = (+1,+1,\dots,+1)$.

Another voting rule commonly encountered in practice:

Definition 2.3. The ith *dictator* function $\chi_i : \{-1,1\}^n \to \{-1,1\}$ is defined by $\chi_i(x) = x_i$.

Here we are simplifying notation for the singleton monomial from $\chi_{\{i\}}$ to χ_i. Even though they are extremely simple functions, the dictators play a very important role in analysis of Boolean functions; to highlight this we prefer the colorful terminology "dictator functions" to the more mathematically staid "projection functions". Generalizing:

Definition 2.4. A function $f : \{-1,1\}^n \to \{-1,1\}$ is called a *k-junta* for $k \in \mathbb{N}$ if it depends on at most k of its input coordinates; i.e., $f(x) = g(x_{i_1},\dots,x_{i_k})$ for some $g : \{-1,1\}^k \to \{-1,1\}$ and $i_1,\dots,i_k \in [n]$. Informally, we say that f is a "junta" if it depends on only a "constant" number of coordinates.

For example, the number of functions $f : \{-1,1\}^n \to \{-1,1\}$ which are 1-juntas is precisely $2n+2$: the n dictators, the n negated-dictators, and the 2 constant functions ± 1.

The European Union's Council of Ministers adopts decisions based on a weighted majority voting rule:

Definition 2.5. A function $f : \{-1,1\}^n \to \{-1,1\}$ is called a *weighted majority* or *(linear) threshold function* if it is expressible as $f(x) = \mathrm{sgn}(a_0 + a_1x_1 + \cdots + a_nx_n)$ for some $a_0, a_1, \dots, a_n \in \mathbb{R}$.

Exercise 2.2 has you verify that majority, AND, OR, dictators, and constants are all linear threshold functions.

The leader of the United States (and many other countries) is elected via a kind of "two-level majority". We make a natural definition along these lines:

Definition 2.6. The *depth-d recursive majority of n* function, denoted $\mathrm{Maj}_n^{\otimes d}$, is the Boolean function of n^d bits defined inductively as follows: $\mathrm{Maj}_n^{\otimes 1} = \mathrm{Maj}_n$, and $\mathrm{Maj}_n^{\otimes(d+1)}(x^{(1)},\dots,x^{(n)}) = \mathrm{Maj}_n(\mathrm{Maj}_n^{\otimes d}(x^{(1)}),\dots,\mathrm{Maj}_n^{\otimes d}(x^{(n)}))$ for $x^{(i)} \in \{-1,1\}^{n^d}$.

In our last example of a 2-candidate voting rule, the voters are divided into "tribes" of equal size and the outcome is True if and only if at least one tribe is unanimously in favor of True. This rule is only somewhat plausible in practice, but it plays a very important role in the analysis of Boolean functions:

Definition 2.7. The *tribes* function of width w and size s, $\mathrm{Tribes}_{w,s} : \{-1,1\}^{sw} \to \{-1,1\}$, is defined by $\mathrm{Tribes}_{w,s}(x^{(1)},\dots,x^{(s)}) = \mathrm{OR}_s(\mathrm{AND}_w(x^{(1)}),\dots,\mathrm{AND}_w(x^{(s)}))$, where $x^{(i)} \in \{-1,1\}^w$.

Here are some natural properties of 2-candidate social choice functions which may be considered desirable:

Definition 2.8. We say that a function $f : \{-1,1\}^n \to \{-1,1\}$ is:

- *monotone* if $f(x) \leq f(y)$ whenever $x \leq y$ coordinate-wise;
- *odd* if $f(-x) = -f(x)$;
- *unanimous* if $f(1,\dots,1) = 1$ and $f(-1,\dots,-1) = -1$;
- *symmetric* if $f(x^\pi) = f(x)$ for all permutations $\pi \in S_n$ (using the notation from Exercise 1.30); i.e., $f(x)$ only depends on the number of 1's in x.

The definitions of monotone, odd, and symmetric are also natural for $f : \{-1,1\}^n \to \mathbb{R}$.

Example 2.9. The majority function (for n odd) has all four properties in Definition 2.8; indeed, *May's Theorem* (Exercise 2.3) states that it is the only monotone, odd, symmetric function. The dictator functions have the first three properties above, as do recursive majority functions. The AND and OR functions are monotone, unanimous, and symmetric, but not odd. The tribes functions are monotone and unanimous; although they are not symmetric they have an important weaker property:

Definition 2.10. A function $f : \{-1,1\}^n \to \{-1,1\}$ is *transitive-symmetric* if for all $i, i' \in [n]$ there exists a permutation $\pi \in S_n$ taking i to i' such $f(x^\pi) = f(x)$ for all $x \in \{-1,1\}^n$.

Intuitively, a function is transitive-symmetric if any two coordinates $i, j \in [n]$ are "equivalent".

One more natural desirable property of a 2-candidate voting rule is that it be *unbiased* as defined in Chapter 1.4, i.e., "equally likely" to elect ±1. Of course, this presupposes the uniform probability distribution on votes.

Definition 2.11. The *impartial culture assumption* is that the n voters' preferences are independent and uniformly random.

Although this assumption might seem somewhat unrealistic, it gives a good basis for comparing voting rules in the absence of other information.

One might also consider it as a model for the votes of just the "undecided" or "party-independent" voters.

2.2. Influences and derivatives

Given a voting rule $f : \{-1,1\}^n \to \{-1,1\}$ it's natural to try to measure the "influence" or "power" of the ith voter. One can define this to be the "probability that the ith vote affects the outcome".

Definition 2.12. We say that coordinate $i \in [n]$ is *pivotal* for $f : \{-1,1\}^n \to \{-1,1\}$ on input x if $f(x) \neq f(x^{\oplus i})$. Here we have used the notation $x^{\oplus i}$ for the string $(x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)$.

Definition 2.13. The *influence* of coordinate i on $f : \{-1,1\}^n \to \{-1,1\}$ is defined to be the probability that i is pivotal for a random input:

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

Influences can be equivalently defined in terms of "geometry" of the Hamming cube:

Fact 2.14. *For $f : \{-1,1\}^n \to \{-1,1\}$, the influence $\mathbf{Inf}_i[f]$ equals the fraction of* dimension-i edges *in the Hamming cube which are* boundary *edges. Here (x, y) is a dimension-i edge if $y = x^{\oplus i}$; it is a boundary edge if $f(x) \neq f(y)$.*

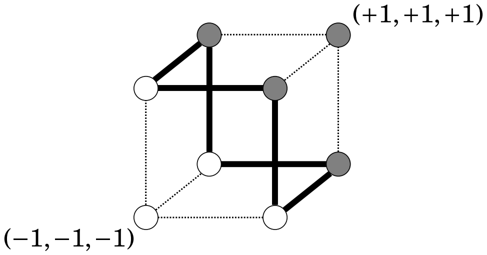


Figure 2.1. Boundary edges of the Maj_3 function

Example 2.15. For the ith dictator function χ_i we have that coordinate i is pivotal for every input x; hence $\mathbf{Inf}_i[\chi_i] = 1$. On the other hand, if $j \neq i$ then coordinate j is never pivotal; hence $\mathbf{Inf}_j[\chi_i] = 0$ for $j \neq i$. Note that the same two statements are true about the negated-dictator functions. For the constant functions ± 1, all influences are 0. For the OR_n function, coordinate 1 is pivotal for exactly two inputs, $(-1,1,1,\dots,1)$ and $(1,1,1,\dots,1)$; hence $\mathbf{Inf}_1[\mathrm{OR}_n] = 2^{1-n}$. Similarly, $\mathbf{Inf}_i[\mathrm{OR}_n] = \mathbf{Inf}_i[\mathrm{AND}_n] = 2^{1-n}$ for all $i \in [n]$. The Maj_3 is depicted in Figure 2.1; the points where it's $+1$ are colored gray and the points where it's -1 are colored white. Its boundary edges are highlighted in black; there are 2 of them in each of the 3 dimensions. Since there are 4 total edges in each dimension, we conclude $\mathbf{Inf}_i[\mathrm{Maj}_3] = 2/4 = 1/2$ for all $i \in [3]$. For majority in higher dimensions, $\mathbf{Inf}_i[\mathrm{Maj}_n]$ equals the probability that among $n-1$ random bits, exactly half of them are 1. This is roughly $\frac{\sqrt{2/\pi}}{\sqrt{n}}$ for large n; see Exercise 2.22 or Chapter 5.2.

Influences can also be defined more "analytically" by introducing the *derivative* operators.

Definition 2.16. The *ith (discrete) derivative operator* D_i maps the function $f : \{-1,1\}^n \to \mathbb{R}$ to the function $\mathrm{D}_i f : \{-1,1\}^n \to \mathbb{R}$ defined by

$$\mathrm{D}_i f(x) = \frac{f(x^{(i\mapsto 1)}) - f(x^{(i\mapsto -1)})}{2}.$$

Here we have used the notation $x^{(i\mapsto b)} = (x_1,\dots,x_{i-1},b,x_{i+1},\dots,x_n)$. Notice that $\mathrm{D}_i f(x)$ does not actually depend on x_i. The operator D_i is a linear operator: i.e., $\mathrm{D}_i(f+g) = \mathrm{D}_i f + \mathrm{D}_i g$.

If $f : \{-1,1\}^n \to \{-1,1\}$ is Boolean-valued then

$$\mathrm{D}_i f(x) = \begin{cases} 0 & \text{if coordinate } i \text{ is not pivotal for } x, \\ \pm 1 & \text{if coordinate } i \text{ is pivotal for } x. \end{cases} \tag{2.1}$$

Thus $\mathrm{D}_i f(x)^2$ is the 0-1 indicator for whether i is pivotal for x and we conclude that $\mathbf{Inf}_i[f] = \mathbf{E}[\mathrm{D}_i f(\boldsymbol{x})^2]$. We take this formula as a *definition* for the influences of real-valued Boolean functions.

Definition 2.17. We generalize Definition 2.13 to functions $f : \{-1,1\}^n \to \mathbb{R}$ by defining the influence of coordinate i on f to be

$$\mathbf{Inf}_i[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\mathrm{D}_i f(\boldsymbol{x})^2] = \|\mathrm{D}_i f\|_2^2.$$

Definition 2.18. We say that coordinate $i \in [n]$ is *relevant* for $f : \{-1,1\}^n \to \mathbb{R}$ if and only if $\mathbf{Inf}_i[f] > 0$; i.e., $f(x^{(i\mapsto 1)}) \neq f(x^{(i\mapsto -1)})$ for at least one $x \in \{-1,1\}^n$.

The discrete derivative operators are quite analogous to the usual partial derivatives. For example, $f : \{-1,1\}^n \to \mathbb{R}$ is monotone if and only if $\mathrm{D}_i f(x) \geq 0$ for all i and x. Further, D_i acts like formal differentiation on Fourier expansions:

Proposition 2.19. *Let $f : \{-1,1\}^n \to \mathbb{R}$ have the multilinear expansion $f(x) = \sum_{S\subseteq[n]} \widehat{f}(S)x^S$. Then*

$$\mathrm{D}_i f(x) = \sum_{\substack{S\subseteq[n]\\ S\ni i}} \widehat{f}(S)\, x^{S\setminus\{i\}}. \tag{2.2}$$

Proof. Since D_i is a linear operator, the claim follows immediately from the observation that

$$\mathrm{D}_i x^S = \begin{cases} x^{S\setminus\{i\}} & \text{if } i \in S, \\ 0 & \text{if } i \not\in S. \end{cases}$$ □

By applying Parseval's Theorem to the Fourier expansion (2.2), we obtain a Fourier formula for influences:

Theorem 2.20. *For $f : \{-1,1\}^n \to \mathbb{R}$ and $i \in [n]$,*

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

In other words, the influence of coordinate i on f equals the sum of f's Fourier weights on sets containing i. This is another good example of being able to "read off" an interesting combinatorial property of a Boolean function from its Fourier expansion. In the special case that $f : \{-1,1\}^n \to \{-1,1\}$ is monotone there is a much simpler way to read off its influences: they are the degree-1 Fourier coefficients. In what follows, we write $\widehat{f}(i)$ in place of $\widehat{f}(\{i\})$.

Proposition 2.21. *If $f : \{-1,1\}^n \to \{-1,1\}$ is monotone, then $\mathbf{Inf}_i[f] = \widehat{f}(i)$.*

Proof. By monotonicity, the ±1 in (2.1) is always 1; i.e., $\mathrm{D}_i f(x)$ is the 0-1 indicator that i is pivotal for x. Hence $\mathbf{Inf}_i[f] = \mathbf{E}[\mathrm{D}_i f] = \widehat{\mathrm{D}_i f}(\emptyset) = \widehat{f}(i)$, where the third equality used Proposition 2.19. □

This formula allows us a neat proof that for any 2-candidate voting rule that is monotone and transitive-symmetric, all of the voters have *small influence*:

Proposition 2.22. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be transitive-symmetric and monotone. Then $\mathbf{Inf}_i[f] \leq 1/\sqrt{n}$ for all $i \in [n]$.*

Proof. Transitive-symmetry of f implies that $\widehat{f}(i) = \widehat{f}(i')$ for all $i, i' \in [n]$ (using Exercise 1.30(*a*)); thus by monotonicity, $\mathbf{Inf}_i[f] = \widehat{f}(i) = \widehat{f}(1)$ for all $i \in [n]$. But by Parseval, $1 = \sum_S \widehat{f}(S)^2 \geq \sum_{i=1}^n \widehat{f}(i)^2 = n\widehat{f}(1)^2$; hence $\widehat{f}(1) \leq 1/\sqrt{n}$. □

This bound is slightly improved in Proposition 2.58 and Exercise 2.24.

The derivative operators are very convenient for functions defined on $\{-1,1\}^n$ but they are less natural if we think of the Hamming cube as $\{\mathsf{True}, \mathsf{False}\}^n$; for the more general domains we'll look at in Chapter 8 they don't even make sense. We end this section by introducing some useful definitions that will generalize better later.

Definition 2.23. The *ith expectation operator* E_i is the linear operator on functions $f : \{-1,1\}^n \to \mathbb{R}$ defined by

$$\mathrm{E}_i f(x) = \mathop{\mathbf{E}}_{\boldsymbol{x}_i}[f(x_1, \dots, x_{i-1}, \boldsymbol{x}_i, x_{i+1}, \dots, x_n)].$$

Whereas $\mathrm{D}_i f$ isolates the part of f depending on the ith coordinate, $\mathrm{E}_i f$ isolates the part *not* depending on the ith coordinate. Exercise 2.15 asks you to verify the following:

Proposition 2.24. *For $f : \{-1,1\}^n \to \mathbb{R}$,*

- $\mathrm{E}_i f(x) = \dfrac{f(x^{(i\mapsto 1)}) + f(x^{(i\mapsto -1)})}{2}$,
- $\mathrm{E}_i f(x) = \sum_{S \not\ni i} \widehat{f}(S) x^S$,
- $f(x) = x_i \mathrm{D}_i f(x) + \mathrm{E}_i f(x)$.

Note that in the decomposition $f = x_i \mathrm{D}_i f + \mathrm{E}_i f$, neither $\mathrm{D}_i f$ nor $\mathrm{E}_i f$ depends on x_i. This decomposition is very useful for proving facts about Boolean functions by induction on n.

Finally, we will also define an operator very similar to D_i called the *ith Laplacian*:

Definition 2.25. The *ith coordinate Laplacian operator* L_i is defined by

$$\mathrm{L}_i f = f - \mathrm{E}_i f.$$

Notational warning: Elsewhere you might see the negated definition, $\mathrm{E}_i f - f$.

Exercise 2.16 asks you to verify the following:

Proposition 2.26. *For $f : \{-1,1\}^n \to \mathbb{R}$,*

- $\mathrm{L}_i f(x) = \dfrac{f(x) - f(x^{\oplus i})}{2}$,
- $\mathrm{L}_i f(x) = x_i \mathrm{D}_i f(x) = \sum_{S \ni i} \widehat{f}(S) x^S$,
- $\langle f, \mathrm{L}_i f\rangle = \langle \mathrm{L}_i f, \mathrm{L}_i f\rangle = \mathbf{Inf}_i[f]$.

2.3. Total influence

A very important quantity in the analysis of a Boolean function is the sum of its influences.

Definition 2.27. The *total influence* of $f : \{-1,1\}^n \to \mathbb{R}$ is defined to be

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

For Boolean-valued functions $f : \{-1,1\}^n \to \{-1,1\}$ the total influence has several additional interpretations. First, it is often referred to as the *average sensitivity* of f because of the following proposition:

Proposition 2.28. *For $f : \{-1,1\}^n \to \{-1,1\}$*

$$\mathbf{I}[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}}[\mathrm{sens}_f(\boldsymbol{x})],$$

where $\mathrm{sens}_f(x)$ *is the* sensitivity *of f at x, defined to be the number of pivotal coordinates for f on input x.*

Proof.

$$\begin{aligned}\mathbf{I}[f] = \sum_{i=1}^{n} \mathbf{Inf}_i[f] &= \sum_{i=1}^{n} \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})] \\ &= \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{x}}[\mathbf{1}_{f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})}] = \mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{i=1}^{n} \mathbf{1}_{f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})}\right] = \mathop{\mathbf{E}}_{\boldsymbol{x}}[\mathrm{sens}_f(\boldsymbol{x})]. \quad \square\end{aligned}$$

The total influence of $f : \{-1,1\}^n \to \{-1,1\}$ is also closely related to the size of its *edge boundary*; from Fact 2.14 we deduce:

Fact 2.29. *The fraction of edges in the Hamming cube $\{-1,1\}^n$ which are boundary edges for $f : \{-1,1\}^n \to \{-1,1\}$ is equal to $\frac{1}{n}\mathbf{I}[f]$.*

Example 2.30. (Recall Example 2.15.) For Boolean-valued functions $f : \{-1,1\}^n \to \{-1,1\}$ the total influence ranges between 0 and n. It is minimized by the constant functions ±1 which have total influence 0. It is maximized by the parity function $\chi_{[n]}$ and its negation which have total influence n; every coordinate is pivotal on every input for these functions. The dictator functions (and their negations) have total influence 1. The total influence of OR_n and AND_n is very small: $n2^{1-n}$. On the other hand, the total influence of Maj_n is fairly large: roughly $\sqrt{2/\pi}\sqrt{n}$ for large n.

By virtue of Proposition 2.21 we have another interpretation for the total influence of *monotone* functions:

Proposition 2.31. *If $f : \{-1,1\}^n \to \{-1,1\}$ is monotone, then*

$$\mathbf{I}[f] = \sum_{i=1}^{n} \widehat{f}(i).$$

This sum of the degree-1 Fourier coefficients has a natural interpretation in social choice:

Proposition 2.32. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a voting rule for a 2-candidate election. Given votes $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, let $\boldsymbol{w}$ be the number of votes that agree with the outcome of the election, $f(\boldsymbol{x})$. Then*

$$\mathbf{E}[\boldsymbol{w}] = \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \widehat{f}(i).$$

Proof. By the formula for Fourier coefficients,

$$\sum_{i=1}^{n} \widehat{f}(i) = \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x})\boldsymbol{x}_i] = \mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x})(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_n)]. \tag{2.3}$$

Now $\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$ equals the difference between the number of votes for candidate 1 and the number of votes for candidate -1. Hence $f(\boldsymbol{x})(\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n)$ equals the difference between the number of votes for the winner and the number of votes for the loser; i.e., $\boldsymbol{w} - (n - \boldsymbol{w}) = 2\boldsymbol{w} - n$. The result follows. □

Rousseau [Rou62] suggested that the ideal voting rule is one which maximizes the number of votes that agree with the outcome. Here we show that the majority rule has this property (at least when n is odd):

Theorem 2.33. *The unique maximizers of* $\sum_{i=1}^{n} \widehat{f}(i)$ *among all* $f : \{-1,1\}^n \to \{-1,1\}$ *are the majority functions. In particular,* $\mathbf{I}[f] \leq \mathbf{I}[\mathrm{Maj}_n] = \sqrt{2/\pi}\sqrt{n} + O(n^{-1/2})$ *for all monotone* f.

Proof. From (2.3),

$$\sum_{i=1}^{n} \widehat{f}(i) = \mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x})(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_n)] \leq \mathop{\mathbf{E}}_{\boldsymbol{x}}[|\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_n|],$$

since $f(\boldsymbol{x}) \in \{-1,1\}$ always. Equality holds if and only if $f(x) = \mathrm{sgn}(x_1 + \cdots + x_n)$ whenever $x_1 + \cdots + x_n \neq 0$. The second statement of the theorem follows from Proposition 2.31 and Exercise 2.22. □

Let's now take a look at more analytic expressions for the total influence. By definition, if $f : \{-1,1\}^n \to \mathbb{R}$, then

$$\mathbf{I}[f] = \sum_{i=1}^{n} \mathbf{Inf}_i[f] = \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{x}}[\mathrm{D}_i f(\boldsymbol{x})^2] = \mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{i=1}^{n} \mathrm{D}_i f(\boldsymbol{x})^2\right]. \tag{2.4}$$

This motivates the following definition:

Definition 2.34. The *(discrete) gradient operator* ∇ maps the function $f : \{-1,1\}^n \to \mathbb{R}$ to the function $\nabla f : \{-1,1\}^n \to \mathbb{R}^n$ defined by

$$\nabla f(x) = (\mathrm{D}_1 f(x), \mathrm{D}_2 f(x), \ldots, \mathrm{D}_n f(x)).$$

Note that for $f : \{-1,1\}^n \to \{-1,1\}$ we have $\|\nabla f(x)\|_2^2 = \mathrm{sens}_f(x)$, where $\|\cdot\|_2$ is the usual Euclidean norm in $\mathbb{R}^n$. In general, from (2.4) we deduce:

Proposition 2.35. *For* $f : \{-1,1\}^n \to \mathbb{R}$*,*

$$\mathbf{I}[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}}[\|\nabla f(\boldsymbol{x})\|_2^2].$$

An alternative analytic definition involves introducing the *Laplacian*:

Definition 2.36. The *Laplacian operator* L is the linear operator on functions $f : \{-1,1\}^n \to \mathbb{R}$ defined by $\mathrm{L} = \sum_{i=1}^{n} \mathrm{L}_i$.

Exercise 2.17 asks you to verify the following:

Proposition 2.37. *For* $f : \{-1,1\}^n \to \mathbb{R}$*,*

- $\mathrm{L}f(x) = (n/2)\big(f(x) - \mathop{\mathrm{avg}}_{i\in[n]}\{f(x^{\oplus i})\}\big)$,
- $\mathrm{L}f(x) = f(x) \cdot \mathrm{sens}_f(x)$ *if* $f : \{-1,1\}^n \to \{-1,1\}$*,*
- $\mathrm{L}f = \sum_{S\subseteq[n]} |S|\widehat{f}(S)\chi_S$,

- $\langle f, \mathrm{L}f \rangle = \mathbf{I}[f]$.

We can obtain a Fourier formula for the total influence of a function using Theorem 2.20; when we sum that theorem over all $i \in [n]$ the Fourier weight $\widehat{f}(S)^2$ is counted exactly $|S|$ times. Hence:

Theorem 2.38. *For $f : \{-1,1\}^n \to \mathbb{R}$,*

$$\mathbf{I}[f] = \sum_{S \subseteq [n]} |S| \widehat{f}(S)^2 = \sum_{k=0}^{n} k \cdot \mathbf{W}^k[f]. \tag{2.5}$$

For $f : \{-1,1\}^n \to \{-1,1\}$ we can express this using the spectral sample:

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

Thus the total influence of $f : \{-1,1\}^n \to \{-1,1\}$ also measures the average "height" or degree of its Fourier weights.

Finally, from Proposition 1.13 we have $\mathbf{Var}[f] = \sum_{k>0} \mathbf{W}^k[f]$; comparing this with (2.5) we immediately deduce a simple but important fact called the *Poincaré Inequality*.

Poincaré Inequality. *For any $f : \{-1,1\}^n \to \mathbb{R}$, $\mathbf{Var}[f] \leq \mathbf{I}[f]$.*

Equality holds in the Poincaré Inequality if and only if all of f's Fourier weight is at degrees 0 and 1; i.e., $\mathbf{W}^{\leq 1}[f] = \mathbf{E}[f^2]$. For Boolean-valued $f : \{-1,1\}^n \to \{-1,1\}$, Exercise 1.19 tells us this can only occur if $f = \pm 1$ or $f = \pm \chi_i$ for some i.

For Boolean-valued $f : \{-1,1\}^n \to \mathbb{R}$, the Poincaré Inequality can be viewed as an (edge-)isoperimetric inequality, or *(edge-)expansion bound*, for the Hamming cube. If we think of f as the indicator function for a set $A \subseteq \{-1,1\}^n$ of "measure" $\alpha = |A|/2^n$, then $\mathbf{Var}[f] = 4\alpha(1-\alpha)$ (Fact 1.14) whereas $\mathbf{I}[f]$ is n times the (fractional) size of A's edge boundary. In particular, the Poincaré Inequality says that subsets $A \subseteq \{-1,1\}^n$ of measure $\alpha = 1/2$ must have edge boundary at least as large as those of the dictator sets.

For $\alpha \notin \{0, 1/2, 1\}$ the Poincaré Inequality is not sharp as an edge-isoperimetric inequality for the Hamming cube; for small α even the asymptotic dependence is not optimal. Precisely optimal edge-isoperimetric results (and also vertex-isoperimetric results) are known for the Hamming cube. The following simplified theorem is optimal for α of the form 2^{-i}:

Theorem 2.39. *For $f : \{-1,1\}^n \to \{-1,1\}$ with $\alpha = \min\{\mathbf{Pr}[f = 1], \mathbf{Pr}[f = -1]\}$,*

$$2\alpha \log(1/\alpha) \leq \mathbf{I}[f].$$

This result illustrates an important recurring concept in the analysis of Boolean functions: The Hamming cube is a "small-set expander". Roughly speaking, this is the idea that "small" subsets $A \subseteq \{-1,1\}^n$ have unusually large "boundary size".

2.4. Noise stability

Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is a voting rule for a 2-candidate election. Making the impartial culture assumption, the n voters independently and uniformly randomly choose their votes $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. Now imagine that when each voter goes to the ballot box there is some chance that their vote is *misrecorded*. Specifically, say that each vote is correctly recorded with probability $\rho \in [0,1]$ and is garbled – i.e., changed to a random bit – with probability $1-\rho$. Writing $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ for the votes that are finally recorded, we may ask about the probability that $f(\boldsymbol{x}) = f(\boldsymbol{y})$, i.e., whether the misrecorded votes affected the outcome of the election. This has to do with the *noise stability* of f.

Definition 2.40. Let $\rho \in [0,1]$. For fixed $x \in \{-1,1\}^n$ we write $\boldsymbol{y} \sim N_\rho(x)$ to denote that the random string $\boldsymbol{y}$ is drawn as follows: for each $i \in [n]$ independently,

$$\boldsymbol{y}_i = \begin{cases} x_i & \text{with probability } \rho, \\ \text{uniformly random} & \text{with probability } 1-\rho. \end{cases}$$

We extend the notation to all $\rho \in [-1,1]$ as follows:

$$\boldsymbol{y}_i = \begin{cases} x_i & \text{with probability } \frac{1}{2} + \frac{1}{2}\rho, \\ -x_i & \text{with probability } \frac{1}{2} - \frac{1}{2}\rho. \end{cases}$$

We say that $\boldsymbol{y}$ is *ρ-correlated* to x.

Definition 2.41. If $\boldsymbol{x} \sim \{-1,1\}^n$ is drawn uniformly at random and then $\boldsymbol{y} \sim N_\rho(\boldsymbol{x})$, we say that $(\boldsymbol{x}, \boldsymbol{y})$ is a *ρ-correlated pair* of random strings. This definition is symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$; it is equivalent to saying that independently for each $i \in [n]$, the pair of random bits $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ satisfies $\mathbf{E}[\boldsymbol{x}_i] = \mathbf{E}[\boldsymbol{y}_i] = 0$ and $\mathbf{E}[\boldsymbol{x}_i \boldsymbol{y}_i] = \rho$.

With these definitions in hand we can now define the important concept of noise stability, which measures the correlation between $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$ when $(\boldsymbol{x}, \boldsymbol{y})$ is a ρ-correlated pair.

Definition 2.42. For $f : \{-1,1\}^n \to \mathbb{R}$ and $\rho \in [-1,1]$, the *noise stability of f at ρ* is

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \\ \rho\text{-correlated}}}[f(\boldsymbol{x})f(\boldsymbol{y})].$$

If $f : \{-1,1\}^n \to \{-1,1\}$ we have

$$\begin{aligned} \mathbf{Stab}_\rho[f] &= \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \\ \rho\text{-correlated}}}[f(\boldsymbol{x}) = f(\boldsymbol{y})] - \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \\ \rho\text{-correlated}}}[f(\boldsymbol{x}) \neq f(\boldsymbol{y})] \\ &= 2 \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \\ \rho\text{-correlated}}}[f(\boldsymbol{x}) = f(\boldsymbol{y})] - 1. \end{aligned}$$

In the voting scenario described above, the probability that the misrecording of votes doesn't affect the election outcome is $\frac{1}{2} + \frac{1}{2}\mathbf{Stab}_\rho[f]$.

When ρ is close to 1 (i.e., the "noise" is small) it's sometimes more natural to ask about the probability that reversing a small fraction of the votes reverses the outcome of the election.

Definition 2.43. For $f : \{-1,1\}^n \to \{-1,1\}$ and $\delta \in [0,1]$ we write $\mathbf{NS}_\delta[f]$ for *noise sensitivity of f at δ*, defined to be the probability that $f(\boldsymbol{x}) \neq f(\boldsymbol{y})$ when $\boldsymbol{x} \sim \{-1,1\}^n$ is uniformly random and $\boldsymbol{y}$ is formed from $\boldsymbol{x}$ by reversing each bit independently with probability δ. In other words,

$$\mathbf{NS}_\delta[f] = \frac{1}{2} - \frac{1}{2}\mathbf{Stab}_{1-2\delta}[f].$$

Example 2.44. The constant functions ± 1 have noise stability 1 for every ρ. The dictator functions χ_i satisfy $\mathbf{Stab}_\rho[\chi_i] = \rho$ for all ρ (equivalently, $\mathbf{NS}_\delta[\chi_i] = \delta$ for all δ). More generally,

$$\mathbf{Stab}_\rho[\chi_S] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y}) \\ \rho\text{-correlated}}}[\boldsymbol{x}^S \boldsymbol{y}^S] = \mathbf{E}\left[\prod_{i \in S}(\boldsymbol{x}_i \boldsymbol{y}_i)\right] = \prod_{i \in S} \mathbf{E}[\boldsymbol{x}_i \boldsymbol{y}_i] = \prod_{i \in S} \rho = \rho^{|S|},$$

where we used the fact that the bit pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are independent across i to convert the expectation of a product to a product of an expectation.

There is no convenient expression for the noise stability of the majority function $\mathbf{Stab}_\rho[\mathrm{Maj}_n]$. However, for a fixed noise rate, the noise stability/sensitivity tends to a nice limit as $n \to \infty$:

Theorem 2.45. *For any $\rho \in [-1,1]$,*

$$\lim_{\substack{n \to \infty \\ n \text{ odd}}} \mathbf{Stab}_\rho[\mathrm{Maj}_n] = \tfrac{2}{\pi}\arcsin\rho = 1 - \tfrac{2}{\pi}\arccos\rho.$$

Equivalently, for $\delta \in [0,1]$,

$$\lim_{\substack{n \to \infty \\ n \text{ odd}}} \mathbf{NS}_\delta[\mathrm{Maj}_n] = \tfrac{1}{\pi}\arccos(1-2\delta).$$

Using $\cos(z) = 1 - \frac{1}{2}z^2 + O(z^4)$*, hence* $\arccos(1-2\delta) = 2\sqrt{\delta} + O(\delta^{3/2})$*, we deduce*

$$\lim_{\substack{n \to \infty \\ n \text{ odd}}} \mathbf{NS}_\delta[\mathrm{Maj}_n] = \tfrac{2}{\pi}\sqrt{\delta} + O(\delta^{3/2}).$$

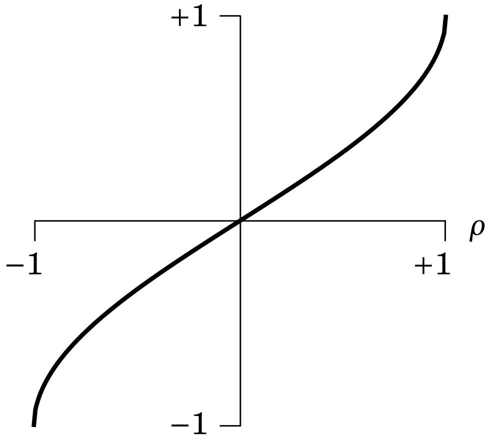


Figure 2.2. Plot of $\frac{2}{\pi}\arcsin\rho$ as a function of ρ

We prove Theorem 2.45 in Chapter 5.2.

There is a simple Fourier formula for the noise stability of a Boolean function; it's one of the most powerful links between the combinatorics of Boolean functions and their Fourier spectra. To determine it, we begin by introducing the most important operator in analysis of Boolean functions: the *noise operator*, denoted T_ρ for historical reasons.

Definition 2.46. For $\rho \in [-1,1]$, the *noise operator with parameter* ρ is the linear operator T_ρ on functions $f : \{-1,1\}^n \to \mathbb{R}$ defined by

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

Proposition 2.47. *For $f : \{-1,1\}^n \to \mathbb{R}$, the Fourier expansion of $\mathrm{T}_\rho f$ is given by*

$$\mathrm{T}_\rho f = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}(S)\chi_S = \sum_{k=0}^{n} \rho^k f^{=k}.$$

Proof. Since T_ρ is a linear operator, it suffices to verify that $\mathrm{T}_\rho\chi_S = \rho^{|S|}\chi_S$:

$$\mathrm{T}_\rho\chi_S(x) = \mathop{\mathbf{E}}_{\boldsymbol{y}\sim N_\rho(x)}[\boldsymbol{y}^S] = \prod_{i\in S}\mathop{\mathbf{E}}_{\boldsymbol{y}\sim N_\rho(x)}[\boldsymbol{y}_i] = \prod_{i\in S}(\rho x_i) = \rho^{|S|}\chi_S(x).$$

Here we used the fact that for $\boldsymbol{y} \sim N_\rho(x)$ the bits $\boldsymbol{y}_i$ are independent and satisfy $\mathbf{E}[\boldsymbol{y}_i] = \rho x_i$. □

Exercise 2.25 gives an alternate way of looking at this proof. Yet another proof using probability densities and convolution is outlined in Exercise 2.30.

The connection between T_ρ and noise stability is that

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

hence:

Fact 2.48. $\mathbf{Stab}_\rho[f] = \langle f, \mathrm{T}_\rho f\rangle$.

From Plancherel's Theorem and Proposition 2.47 we deduce the Fourier formula for noise stability:

Theorem 2.49. *For $f : \{-1,1\}^n \to \mathbb{R}$,*

$$\mathbf{Stab}_\rho[f] = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}(S)^2 = \sum_{k=0}^{n} \rho^k \cdot \mathbf{W}^k[f].$$

Hence for $f : \{-1,1\}^n \to \{-1,1\}$ we have

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\boldsymbol{S}\sim\mathcal{S}_f}[\rho^{|\boldsymbol{S}|}], \tag{2.6}$$

$$\mathbf{NS}_\delta[f] = \tfrac{1}{2}\sum_{k=0}^{n}(1-(1-2\delta)^k)\cdot\mathbf{W}^k[f]. \tag{2.7}$$

Thus the noise stability of f at ρ is equal to the sum of its Fourier weights, attenuated by a factor which decreases exponentially with degree. A simple but important corollary is that dictators (and their negations) maximize noise stability:

Proposition 2.50. *Let $\rho \in (0,1)$. If $f : \{-1,1\}^n \to \{-1,1\}$ is unbiased, then $\mathbf{Stab}_\rho[f] \leq \rho$, with equality if and only if $f = \pm\chi_i$ for some $i \in [n]$.*

Proof. For unbiased f we have $\mathbf{W}^0[f] = 0$ and hence $\mathbf{Stab}_\rho[f] = \sum_{k\geq 1}\rho^k\mathbf{W}^k[f]$. Since $\rho^k < \rho$ for all $k > 1$, noise stability is maximized if all of f's Fourier weight is on degree 1. This occurs if and only if $f = \pm\chi_i$, by Exercise 1.19(a). □

For a fixed function f, it's often interesting to see how $\mathbf{Stab}_\rho[f]$ varies as a function of ρ. From Theorem 2.49 we see that $\mathbf{Stab}_\rho[f]$ is a (univariate) *polynomial* with nonnegative coefficients; in particular, it's an increasing function of ρ on $[0,1]$. The derivatives of this polynomial at 0 and 1 have nice interpretations, as can be immediately deduced from Theorem 2.49:

Proposition 2.51. *For $f : \{-1,1\}^n \to \mathbb{R}$,*

$$\frac{d}{d\rho}\mathbf{Stab}_\rho[f]\Big|_{\rho=0} = \mathbf{W}^1[f],$$
$$\frac{d}{d\rho}\mathbf{Stab}_\rho[f]\Big|_{\rho=1} = \mathbf{I}[f].$$

For $f : \{-1,1\}^n \to \{-1,1\}$ we have that $\mathbf{NS}_\delta[f]$ is an increasing function of δ on $[0,1/2]$, and the second identity is equivalent to

$$\frac{d}{d\delta}\mathbf{NS}_\delta[f]\Big|_{\delta=0} = \mathbf{I}[f].$$

We conclude this section by introducing a version of influences that also incorporates noise.

Definition 2.52. For $f : \{-1,1\}^n \to \mathbb{R}$, $\rho \in [0,1]$ and $i \in [n]$, the *ρ-stable influence* of i on f is

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

with 0^0 interpreted as 1. We also define $\mathbf{I}^{(\rho)}[f] = \sum_{i=1}^n \mathbf{Inf}_i^{(\rho)}[f]$.

Exercise 2.40 asks you to verify the following:

Fact 2.53. $\mathbf{I}^{(\rho)}[f] = \frac{d}{d\rho}\mathbf{Stab}_\rho[f] = \sum_{k=1}^n k\rho^{k-1} \cdot \mathbf{W}^k[f]$.

The ρ-stable influence $\mathbf{Inf}_i^{(\rho)}[f]$ increases from $\widehat{f}(i)^2$ up to $\mathbf{Inf}_i[f]$ as ρ increases from 0 to 1. For $0 < \rho < 1$ there isn't an especially natural combinatorial interpretation for $\mathbf{Inf}_i^{(\rho)}[f]$ beyond $\mathbf{Stab}_\rho[\mathrm{D}_i f]$; however, we will see later that the stable influences are technically very useful. One reason for this is that every function $f : \{-1,1\}^n \to \{-1,1\}$ has at most "constantly" many "stably-influential" coordinates:

Proposition 2.54. *Suppose $f : \{-1,1\}^n \to \mathbb{R}$ has $\mathbf{Var}[f] \le 1$. Given $0 < \delta, \epsilon \le 1$, let $J = \{i \in [n] : \mathbf{Inf}_i^{(1-\delta)}[f] \ge \epsilon\}$. Then $|J| \le \frac{1}{\delta\epsilon}$.*

Proof. Certainly $|J| \le \mathbf{I}^{(1-\delta)}[f]/\epsilon$ so it remains to verify $\mathbf{I}^{(1-\delta)}[f] \le 1/\delta$. Comparing Fact 2.53 with $\mathbf{Var}[f] = \sum_{k \neq 0} \mathbf{W}^k[f]$ term by term, it suffices to show that $(1-\delta)^{k-1}k \le 1/\delta$ for all $k \ge 1$. This is the easy Exercise 2.45. □

It's good to think of the set J in this proposition as the "notable" coordinates for function f. Had we used the usual influences in place of stable influences, we would not have been guaranteed a bounded number of "notable" coordinates (since, e.g., the parity function $\chi_{[n]}$ has all n of its influences equal to 1).

2.5. Highlight: Arrow's Theorem

When there are just 2 candidates, the majority function possesses all of the mathematical properties that seem desirable in a voting rule (e.g., May's Theorem and Theorem 2.33). Unfortunately, as soon as there are 3 (or more) candidates the problem of social choice becomes much more difficult. For example, suppose we have candidates a, b, and c, and each of n voters has a ranking of them. How should we aggregate these preferences to produce a winning candidate?

In his 1785 *Essay on the Application of Analysis to the Probability of Majority Decisions* [dC85], Condorcet suggested using the voters' preferences to conduct the three possible pairwise elections, a vs. b, b vs. c, and c vs. a. This calls for the use of a 2-candidate voting rule $f : \{-1,1\}^n \to \{-1,1\}$; Condorcet suggested $f = \mathrm{Maj}_n$ but we might consider any such rule. Thus a "3-candidate Condorcet election" using f is conducted as follows:

	Voters' Preferences					
	#1	#2	#3	$\cdots$		Societal Aggregation
a (+1) vs. b (−1)	$+1$	$+1$	-1	$\cdots$	$= x$	$f(x)$
b (+1) vs. c (−1)	$+1$	-1	$+1$	$\cdots$	$= y$	$f(y)$
c (+1) vs. a (−1)	-1	-1	$+1$	$\cdots$	$= z$	$f(z)$

In the above example, voter #1 ranked the candidates $a > b > c$, voter #2 ranked them $a > c > b$, voter #3 ranked them $b > c > a$, etc. Note that the ith voter has one of $3! = 6$ possible rankings, and these translate into a triple of bits (x_i, y_i, z_i) from the following set:

$$\Big\{(+1,+1,-1),(+1,-1,-1),(-1,+1,-1),(-1,+1,+1),(+1,-1,+1),(-1,-1,+1)\Big\}.$$

These are precisely the triples satisfying the *not-all-equal* predicate NAE_3 (see Exercise 1.1(*i*)).

In the example above, if $n = 3$ and $f = \mathrm{Maj}_3$ then the societal outcome would be $(+1,+1,-1)$, meaning that society elects a over b, b over c, and a over c. In this case it is only natural to declare a the overall winner.

Definition 2.55. In an election employing Condorcet's method with $f : \{-1,1\}^n \to \{-1,1\}$, we say that a candidate is a *Condorcet winner* if it wins all of the pairwise elections in which it participates.

Unfortunately, as Condorcet himself noted, there may not *be* a Condorcet winner. In the example above, if voter #2's ranking was instead $c > a > b$ (corresponding to $(+1,-1,+1)$), we would obtain the "paradoxical" outcome $(+1,+1,+1)$: society prefers a over b, b over c, and c over a! This lack of a Condorcet winner is termed *Condorcet's Paradox*; it occurs when the outcome $(f(x), f(y), f(z))$ is one of the two "all-equal" triples $\{(-1,-1,-1),(+1,+1,+1)\}$.

One might wonder if the Condorcet Paradox can be avoided by using a voting rule $f : \{-1,1\}^n \to \{-1,1\}$ other than majority. However, in 1950 Arrow [Arr50] famously showed that the only means of avoidance is an unappealing one:

Arrow's Theorem. *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is a unanimous voting rule used in a* 3*-candidate Condorcet election. If there is* always *a Condorcet winner, then f must be a dictatorship.*

(In fact, Arrow's Theorem is slightly stronger than this; see Exercise 2.51.)

In 2002 Kalai gave a new proof of Arrow's Theorem; it takes its cue from the title of Condorcet's work and computes the *probability* of a Condorcet winner. This is done under the "impartial culture assumption" for 3-candidate elections: each voter independently chooses one of the 6 possible rankings uniformly at random.

Theorem 2.56. *Consider a* 3*-candidate Condorcet election using* $f : \{-1,1\}^n \to \{-1,1\}$. *Under the impartial culture assumption, the probability of a Condorcet winner is precisely* $\frac{3}{4} - \frac{3}{4}\mathbf{Stab}_{-1/3}[f]$.

Proof. Let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{-1,1\}^n$ be the votes for the elections a vs. b, b vs. c, and c vs. a, respectively. Under impartial culture, the bit triples $(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ are independent and each is drawn uniformly from the 6 triples satisfying the not-all-equal predicate $\mathrm{NAE}_3 : \{-1,1\}^3 \to \{0,1\}$. There is a Condorcet winner if and only if $\mathrm{NAE}_3(f(\boldsymbol{x}), f(\boldsymbol{y}), f(\boldsymbol{z})) = 1$. Hence

$$\mathbf{Pr}[\exists \text{ Condorcet winner}] = \mathbf{E}[\mathrm{NAE}_3(f(\boldsymbol{x}), f(\boldsymbol{y}), f(\boldsymbol{z}))]. \tag{2.8}$$

The multilinear (Fourier) expansion of NAE_3 is

$$\mathrm{NAE}_3(w_1, w_2, w_3) = \tfrac{3}{4} - \tfrac{1}{4}w_1w_2 - \tfrac{1}{4}w_1w_3 - \tfrac{1}{4}w_2w_3;$$

thus

$$(2.8) = \tfrac{3}{4} - \tfrac{1}{4}\mathbf{E}[f(\boldsymbol{x})f(\boldsymbol{y})] - \tfrac{1}{4}\mathbf{E}[f(\boldsymbol{x})f(\boldsymbol{z})] - \tfrac{1}{4}\mathbf{E}[f(\boldsymbol{y})f(\boldsymbol{z})].$$

In the joint distribution of $\boldsymbol{x}, \boldsymbol{y}$ the n bit pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are independent. Further, by inspection we see that $\mathbf{E}[\boldsymbol{x}_i] = \mathbf{E}[\boldsymbol{y}_i] = 0$ and that $\mathbf{E}[\boldsymbol{x}_i\boldsymbol{y}_i] = (2/6)(+1) + (4/6)(-1) = -1/3$. Hence $\mathbf{E}[f(\boldsymbol{x})f(\boldsymbol{y})]$ is precisely $\mathbf{Stab}_{-1/3}[f]$. Similarly $\mathbf{E}[f(\boldsymbol{x})f(\boldsymbol{z})] = \mathbf{E}[f(\boldsymbol{y})f(\boldsymbol{z})] = \mathbf{Stab}_{-1/3}[f]$ and the proof is complete. □

Arrow's Theorem is now an easy corollary:

Proof of Arrow's Theorem. By assumption, the probability of a Condorcet winner is 1; hence

$$1 = \tfrac{3}{4} - \tfrac{3}{4}\mathbf{Stab}_{-1/3}[f] = \frac{3}{4} - \frac{3}{4}\sum_{k=0}^{n}(-1/3)^k\mathbf{W}^k[f].$$

Since $(-1/3)^k \geq -1/3$ for all k, the equality above can only occur if all of f's Fourier weight is on degree 1; i.e., $\mathbf{W}^1[f] = 1$. By Exercise 1.19(*a*) this implies that f is either a dictator or a negated-dictator. Since f is unanimous, it must in fact be a dictator. □

An advantage of Kalai's analytic proof of Arrow's Theorem is that we can deduce several more interesting results about the probability of a Condorcet winner. For example, combining Theorem 2.56 with Theorem 2.45 we get *Guilbaud's Formula*:

Guilbaud's Formula. *In a 3-candidate Condorcet election using* Maj_n*, the probability of a Condorcet winner tends to*

$$\tfrac{3}{2\pi}\arccos(-1/3) \approx 91.2\%.$$

as $n \to \infty$.

This is already a fairly high probability. Unfortunately, if we want to improve on it while still using a reasonably fair election scheme, we can only set our hopes higher by a sliver:

Theorem 2.57. *In a 3-candidate Condorcet election using an* $f : \{-1,1\}^n \to \{-1,1\}$ *with all* $\widehat{f}(i)$ *equal, the probability of a Condorcet winner is at most* $\frac{7}{9} + \frac{4}{9\pi} + o_n(1) \approx 91.9\%$.

The condition in Theorem 2.57 seems like it would be satisfied by most reasonably fair voting rules $f : \{-1,1\}^n \to \{-1,1\}$ (e.g., it is satisfied if f is transitive-symmetric or is monotone with all influences equal). In fact, we will show that Theorem 2.57's hypothesis can be relaxed in Chapter 5.4; we will further show in Chapter 11.7 that $\frac{7}{9} + \frac{4}{9\pi}$ can be improved to the tight value $\frac{3}{2\pi}\arccos(-1/3)$ of majority. To return to Theorem 2.57, it is an immediate consequence of the following two results, the first being Exercise 2.24 and the second being an easy corollary of Theorem 2.56.

Proposition 2.58. *Suppose* $f : \{-1,1\}^n \to \{-1,1\}$ *has all* $\widehat{f}(i)$ *equal. Then* $\mathbf{W}^1[f] \leq 2/\pi + o_n(1)$.

Corollary 2.59. *In a 3-candidate Condorcet election using* $f : \{-1,1\}^n \to \{-1,1\}$*, the probability of a Condorcet winner is at most* $\frac{7}{9} + \frac{2}{9}\mathbf{W}^1[f]$.

Proof. From Theorem 2.56, the probability is

$$\begin{aligned}
\tfrac{3}{4} - \tfrac{3}{4}\mathbf{Stab}_{-1/3}[f] &= \tfrac{3}{4} - \tfrac{3}{4}(\mathbf{W}^0[f] - \tfrac{1}{3}\mathbf{W}^1[f] + \tfrac{1}{9}\mathbf{W}^2[f] - \tfrac{1}{27}\mathbf{W}^3[f] + \cdots) \\
&\leq \tfrac{3}{4} + \tfrac{1}{4}\mathbf{W}^1[f] + \tfrac{1}{36}\mathbf{W}^3[f] + \tfrac{1}{324}\mathbf{W}^5[f] + \cdots \\
&\leq \tfrac{3}{4} + \tfrac{1}{4}\mathbf{W}^1[f] + \tfrac{1}{36}(\mathbf{W}^3[f] + \mathbf{W}^5[f] + \cdots) \\
&\leq \tfrac{3}{4} + \tfrac{1}{4}\mathbf{W}^1[f] + \tfrac{1}{36}(1 - \mathbf{W}^1[f]) \quad = \quad \tfrac{7}{9} + \tfrac{2}{9}\mathbf{W}^1[f].
\end{aligned}$$
□

Finally, using Corollary 2.59 we can prove a "robust" version of Arrow's Theorem, showing that a Condorcet election is *almost* paradox-free only if it is *almost* a dictatorship (possibly negated).

Corollary 2.60. *Suppose that in a 3-candidate Condorcet election using* $f : \{-1,1\}^n \to \{-1,1\}$*, the probability of a Condorcet winner is* $1-\epsilon$*. Then* f *is* $O(\epsilon)$*-close to* $\pm\chi_i$ *for some* $i \in [n]$.

Proof. From Corollary 2.59 we obtain that $\mathbf{W}^1[f] \geq 1 - \frac{9}{2}\epsilon$. The conclusion now follows from the FKN Theorem. □

Friedgut–Kalai–Naor (FKN) Theorem. *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{W}^1[f] \geq 1-\delta$. Then f is $O(\delta)$-close to $\pm\chi_i$ for some $i \in [n]$.*

We will see the proof of the FKN Theorem in Chapter 9.1. We'll also show in Chapter 5.4 that the $O(\delta)$ closeness can be improved to $\delta/4 + O(\delta^2 \log(2/\delta))$.

2.6. Exercises and notes

2.1 For each function in Exercise 1.1, determine if it is odd, transitive-symmetric, and/or symmetric.

2.2 Show that the n-bit functions majority, AND, OR, $\pm\chi_i$, and ± 1 are all linear threshold functions.

2.3 Prove *May's Theorem*:

(*a*) Show that $f : \{-1,1\}^n \to \{-1,1\}$ is symmetric and monotone if and only if it can be expressed as a weighted majority with $a_1 = a_2 = \cdots = a_n = 1$.

(*b*) Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is symmetric, monotone, and odd. Show that n must be odd, and that $f = \mathrm{Maj}_n$.

2.4 Subset $A \subseteq \{-1,1\}^n$ is called a *Hamming ball* if $A = \{x : \Delta(x,z) < r\}$ for some $z \in \{-1,1\}^n$ and real r. Show that $f : \{-1,1\}^n \to \{-1,1\}$ is the indicator of a Hamming ball if and only if it's expressible as a linear threshold function $f(x) = \mathrm{sgn}(a_0 + a_1x_1 + \cdots + a_nx_n)$ with $|a_1| = |a_2| = \cdots = |a_n|$.

2.5 Let $f : \{-1,1\}^n \to \{-1,1\}$ and $i \in [n]$. We say that f is *unate in the ith direction* if either $f(x^{(i\mapsto -1)}) \leq f(x^{(i\mapsto 1)})$ for all x (*monotone in the ith direction*) or $f(x^{(i\mapsto -1)}) \geq f(x^{(i\mapsto 1)})$ for all x (*antimonotone in the ith direction*). We say that f is *unate* if it is unate in all n directions.

(*a*) Show that $|\widehat{f}(i)| \leq \mathbf{Inf}_i[f]$ with equality if and only if f is unate in the ith direction.

(*b*) Show that the second statement of Theorem 2.33 holds even for all unate f.

2.6 Show that linear threshold functions are unate.

2.7 For each function f in Exercise 1.1, compute $\mathbf{Inf}_1[f]$.

2.8 Let $f : \{-1,1\}^n \to \{-1,1\}$. Show that $\mathbf{Inf}_i[f] \leq \mathbf{Var}[f]$ for each $i \in [n]$. (Hint: Show $\mathbf{Inf}_i[f] \leq 2\min\{\mathbf{Pr}[f = -1], \mathbf{Pr}[f = 1]\}$.)

2.9 Let $f : \{0,1\}^6 \to \{-1,1\}$ be given by the weighted majority $f(x) = \mathrm{sgn}(-58 + 31x_1 + 31x_2 + 28x_3 + 21x_4 + 2x_5 + 2x_6)$. Compute $\mathbf{Inf}_i[f]$ for all $i \in [6]$.

2.10 Say that coordinate i is *b-pivotal* for $f : \{-1,1\}^n \to \{-1,1\}$ on input x (for $b \in \{-1,1\}$) if $f(x) = b$ and $f(x^{\oplus i}) \neq b$. Show that $\mathbf{Pr}_{\boldsymbol{x}}[i \text{ is } b\text{-pivotal on } \boldsymbol{x}] = \frac{1}{2}\mathbf{Inf}_i[f]$. Deduce that $\mathbf{I}[f] = 2\,\mathbf{E}_{\boldsymbol{x}}[\#\ b\text{-pivotal coordinates on } \boldsymbol{x}]$.

2.11 Let $f : \{-1,1\}^n \to \{-1,1\}$ and suppose $\widehat{f}(S) \neq 0$. Show that each coordinate $i \in S$ is relevant for f.

2.12 Let $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ be a random function (as in Exercise 1.7). Compute $\mathbf{E}[\mathbf{Inf}_1[\boldsymbol{f}]]$ and $\mathbf{E}[\mathbf{I}[\boldsymbol{f}]]$.

2.13 Let $w \in \mathbb{N}$, $n = w2^w$, and write f for $\text{Tribes}_{w,2^w} : \{-1,1\}^n \to \{-1,1\}$.

(*a*) Compute $\mathbf{E}[f]$ and $\mathbf{Var}[f]$, and estimate them asymptotically in terms of n.

(*b*) Describe the function $\mathrm{D}_1 f$.

(*c*) Compute $\mathbf{Inf}_1[f]$ and $\mathbf{I}[f]$ and estimate them asymptotically.

2.14 Let $f : \{-1,1\}^n \to \mathbb{R}$. Show that $|\, \mathrm{D}_i|f|\, | \le |\mathrm{D}_i f|$ pointwise. Deduce that $\mathbf{Inf}_i[|f|] \le \mathbf{Inf}_i[f]$ and $\mathbf{I}[|f|] \le \mathbf{I}[f]$.

2.15 Prove Proposition 2.24.

2.16 Prove Proposition 2.26.

2.17 Prove Proposition 2.37.

2.18 Let $f : \{-1,1\}^n \to \mathbb{R}$. Show that

$$\mathrm{L}f(x) = \frac{d}{d\rho}\mathrm{T}_\rho f(x)\Big|_{\rho=1} = -\frac{d}{dt}\mathrm{T}_{e^{-t}} f(x)\Big|_{t=0}.$$

2.19 Suppose $f, g : \{-1,1\}^n \to \mathbb{R}$ have the property that f does not depend on the ith coordinate and g does not depend on the jth coordinate ($i \neq j$). Show that $\mathbf{E}[\boldsymbol{x}_i\boldsymbol{x}_j f(\boldsymbol{x})g(\boldsymbol{x})] = \mathbf{E}[\mathrm{D}_j f(\boldsymbol{x})\mathrm{D}_i g(\boldsymbol{x})]$.

2.20 For $f : \{-1,1\}^n \to \{-1,1\}$ we have that $\mathbf{E}[\text{sens}_f(\boldsymbol{x})] = \mathbf{E}_{\boldsymbol{S}\sim\mathscr{S}_f}[|\boldsymbol{S}|]$. Show that also $\mathbf{E}[\text{sens}_f(\boldsymbol{x})^2] = \mathbf{E}[|\boldsymbol{S}|^2]$. (Hint: Use Proposition 2.37.) Is it true that $\mathbf{E}[\text{sens}_f(\boldsymbol{x})^3] = \mathbf{E}[|\boldsymbol{S}|^3]$?

2.21 Let $f : \{-1,1\}^n \to \mathbb{R}$ and $i \in [n]$.

(*a*) Define $\mathrm{Var}_i f : \{-1,1\}^n \to \mathbb{R}$ by $\mathrm{Var}_i f(x) = \mathbf{Var}_{\boldsymbol{x}_i}[f(x_1,\dots,x_{i-1},\boldsymbol{x}_i,x_{i+1},\dots,x_n)]$. Show that $\mathbf{Inf}_i[f] = \mathbf{E}_{\boldsymbol{x}}[\mathrm{Var}_i f(\boldsymbol{x})]$.

(*b*) Show that

$$\mathbf{Inf}_i[f] = \tfrac{1}{2} \mathop{\mathbf{E}}_{\substack{\boldsymbol{x}_i,\boldsymbol{x}'_i\sim\{-1,1\}\\ \text{independent}}}\left[\left\| f_{|\boldsymbol{x}_i} - f_{|\boldsymbol{x}'_i}\right\|_2^2\right],$$

where $f_{|b}$ denotes the function of $n-1$ variables gotten by fixing the ith input of f to bit b.

2.22 (*a*) Show that $\mathbf{Inf}_i[\text{Maj}_n] = \binom{n-1}{\frac{n-1}{2}}2^{1-n}$ for all $i \in [n]$.

(*b*) Show that $\mathbf{Inf}_1[\text{Maj}_n]$ is a decreasing function of (odd) n.

(*c*) Use Stirling's Formula $m! = (m/e)^m(\sqrt{2\pi m} + O(m^{-1/2}))$ to deduce that $\mathbf{Inf}_1[\text{Maj}_n] = \frac{\sqrt{2/\pi}}{\sqrt{n}} + O(n^{-3/2})$.

(*d*) Deduce that $2/\pi \le \mathbf{W}^1[\text{Maj}_n] \le 2/\pi + O(n^{-1})$.

(*e*) Deduce that $\sqrt{2/\pi}\sqrt{n} \le \mathbf{I}[\text{Maj}_n] \le \sqrt{2/\pi}\sqrt{n} + O(n^{-1/2})$.

(*f*) Suppose n is even and $f : \{-1,1\}^n \to \{-1,1\}$ is a majority function. Show that $\mathbf{I}[f] = \mathbf{I}[\text{Maj}_{n-1}] = \sqrt{2/\pi}\sqrt{n} + O(n^{-1/2})$.

2.23 Using only Cauchy–Schwarz and Parseval, give a very simple proof of the following weakening of Theorem 2.33: If $f : \{-1,1\}^n \to \{-1,1\}$ is monotone then $\mathbf{I}[f] \le \sqrt{n}$. Extend also to the case of f unate (see Exercise 2.5).

2.24 Prove Proposition 2.58 with $O(n^{-1})$ in place of $o_n(1)$. (Hint: Show $\widehat{f}(i) \le \frac{\sqrt{2/\pi}}{\sqrt{n}} + O(n^{-3/2})$ using Theorem 2.33.)

2.25 Deduce $\mathrm{T}_\rho f(x) = \sum_S \rho^{|S|} \widehat{f}(S) x^S$ using Exercise 1.4.

2.26 For each function f in Exercise 1.1, compute $\mathbf{I}[f]$.

2.27 Which functions $f : \{-1,1\}^n \to \{-1,1\}$ with $\#\{x : f(x) = 1\} = 3$ maximize $\mathbf{I}[f]$?

2.28 Suppose $f : \{-1,1\}^n \to \mathbb{R}$ is an even function (recall Exercise 1.8). Show the improved Poincaré Inequality $\mathbf{Var}[f] \le \frac{1}{2}\mathbf{I}[f]$.

2.29 Let $f : \{-1,1\}^n \to \{-1,1\}$ be unbiased, $\mathbf{E}[f] = 0$, and let $\mathbf{MaxInf}[f]$ denote $\max_{i\in[n]}\{\mathbf{Inf}_i[f]\}$.

(a) Use the Poincaré Inequality to show $\mathbf{MaxInf}[f] \ge 1/n$.

(b) Prove that $\mathbf{I}[f] \ge 2 - n\mathbf{MaxInf}[f]^2$. (Hint: Prove $\mathbf{I}[f] \ge \mathbf{W}^1[f] + 2(1 - \mathbf{W}^1[f])$ and use Exercise 2.5.) Deduce that $\mathbf{MaxInf}[f] \ge \frac{2}{n} - \frac{4}{n^2}$.

2.30 Use Exercises 1.1(e),(f) to deduce the formulas $\mathrm{E}_i f = \sum_{S \not\ni i} \widehat{f}(S)\chi_S$ and $\mathrm{T}_\rho f = \sum_S \rho^{|S|}\widehat{f}(S)\chi_S$.

2.31 Show that T_ρ is *positivity-preserving* for $\rho \in [-1,1]$; i.e., $f \ge 0 \implies \mathrm{T}_\rho f \ge 0$. Show that T_ρ is *positivity-improving* for $\rho \in (-1,1)$; i.e., $f \ge 0, f \ne 0 \implies \mathrm{T}_\rho f > 0$.

2.32 Show that T_ρ satisfies the *semigroup property*: $\mathrm{T}_{\rho_1}\mathrm{T}_{\rho_2} = \mathrm{T}_{\rho_1\rho_2}$.

2.33 For $\rho \in [-1,1]$, show that T_ρ is a *contraction on* $L^p(\{-1,1\}^n)$ for all $p \ge 1$; i.e., $\|\mathrm{T}_\rho f\|_p \le \|f\|_p$ for all $f : \{-1,1\}^n \to \mathbb{R}$.

2.34 Show that $|\mathrm{T}_\rho f| \le T_\rho |f|$ pointwise for any $f : \{-1,1\}^n \to \mathbb{R}$. Further show that for $-1 < \rho < 1$, equality occurs if and only if f is everywhere nonnegative or everywhere nonpositive.

2.35 For $i \in [n]$ and $\rho \in \mathbb{R}$, let T^i_ρ be the operator on functions $f : \{-1,1\}^n \to \mathbb{R}$ defined by

$$\mathrm{T}^i_\rho f = \rho f + (1-\rho)\mathrm{E}_i f = \mathrm{E}_i f + \rho \mathrm{L}_i f.$$

(a) Show that for $\rho \in [-1,1]$ we have

$$\mathrm{T}^i_\rho f(x) = \mathop{\mathbf{E}}_{\boldsymbol{y}_i \sim N_\rho(x_i)}[f(x_1,\dots,x_{i-1},\boldsymbol{y}_i,x_{i+1},\dots,x_n)].$$

(b) Show that $\mathrm{T}^i_{\rho_1}\mathrm{T}^i_{\rho_2} = \mathrm{T}^i_{\rho_1\rho_2}$ (cf. Exercise 2.32) and that any two operators T^i_ρ and $\mathrm{T}^j_{\rho'}$ commute.

(c) For $(\rho_1,\dots,\rho_n) \in \mathbb{R}^n$ we define $\mathrm{T}_{(\rho_1,\dots,\rho_n)} = \mathrm{T}^1_{\rho_1}\mathrm{T}^2_{\rho_2}\cdots\mathrm{T}^n_{\rho_n}$. Show that $\mathrm{T}_{(\rho,\dots,\rho)}$ is simply T_ρ and that $\mathrm{T}_{(1,\dots,1,\rho,1,\dots,1)}$ (with the ρ in the ith position) is T^i_ρ.

(*d*) For $\rho_1, \dots, \rho_n \in [-1,1]$, show that $\mathrm{T}_{(\rho_1,\dots,\rho_n)}$ is a contraction on $L^p(\{-1,1\}^n)$ for all $p \geq 1$ (cf. Exercise 2.33).

2.36 Show that $\mathbf{Stab}_{-\rho}[f] = -\mathbf{Stab}_\rho[f]$ if f is odd and $\mathbf{Stab}_{-\rho}[f] = \mathbf{Stab}_\rho[f]$ if f is even.

2.37 For each function f in Exercise 1.1, compute $\mathbf{Stab}_\rho[f]$.

2.38 Compute $\mathbf{Stab}_\rho[\mathrm{Tribes}_{w,s}]$.

2.39 Suppose $f : \{-1,1\}^n \to \{-1,1\}$ has $\min(\mathbf{Pr}[f = 1], \mathbf{Pr}[f = -1]) = \alpha$. Show that $\mathbf{NS}_\delta[f] \leq 2\alpha$ for all $\delta \in [0,1]$.

2.40 Verify Fact 2.53.

2.41 Fix $f : \{-1,1\}^n \to \mathbb{R}$. Show that $\mathbf{Stab}_\rho[f]$ is a convex function of ρ on $[0,1]$.

2.42 Let $f : \{-1,1\}^n \to \{-1,1\}$. Show that $\mathbf{NS}_\delta[f] \leq \delta \mathbf{I}[f]$ for all $\delta \in [0,1]$.

2.43 (*a*) Define the *average influence* of $f : \{-1,1\}^n \to \mathbb{R}$ to be $\mathcal{E}[f] = \frac{1}{n}\mathbf{I}[f]$. Now for $f : \{-1,1\}^n \to \{-1,1\}$, show

$$\mathcal{E}[f] = \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim\{-1,1\}^n \\ \boldsymbol{i}\sim[n]}}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus \boldsymbol{i}})] \quad \text{and} \quad \tfrac{1-e^{-2}}{2}\mathcal{E}[f] \leq \mathbf{NS}_{1/n}[f] \leq \mathcal{E}[f].$$

(*b*) Given $f : \{-1,1\}^n \to \{-1,1\}$ and integer $k \geq 2$, define

$$A_k = \frac{1}{k}(\mathbf{W}^{\geq 1}[f] + \mathbf{W}^{\geq 2}[f] + \cdots + \mathbf{W}^{\geq k}[f]),$$

the "average of the first k tail weights". Generalizing the second statement in part (a), show that $\frac{1-e^{-2}}{2}A_k \leq \mathbf{NS}_{1/k}[f] \leq A_k$.

2.44 Suppose $f_1, \dots, f_s : \{-1,1\}^n \to \{-1,1\}$ satisfy $\mathbf{NS}_\delta[f_i] \leq \epsilon_i$. Let $g : \{-1,1\}^s \to \{-1,1\}$ and define $h : \{-1,1\}^n \to \{-1,1\}$ by $h = g(f_1, \dots, f_s)$. Show that $\mathbf{NS}_\delta[h] \leq \sum_{i=1}^s \epsilon_i$.

2.45 Complete the proof of Proposition 2.54 by showing that $(1-\delta)^{k-1}k \leq 1/\delta$ for all $0 < \delta \leq 1$ and $k \in \mathbb{N}^+$. (Hint: Compare both sides with $1 + (1-\delta) + (1-\delta)^2 + \cdots + (1-\delta)^{k-1}$.)

2.46 Fixing $f : \{-1,1\}^n \to \mathbb{R}$, show the following Lipschitz bound for $\mathbf{Stab}_\rho[f]$ when $0 \leq \rho - \epsilon \leq \rho < 1$:

$$\left|\mathbf{Stab}_\rho[f] - \mathbf{Stab}_{\rho-\epsilon}[f]\right| \leq \epsilon \cdot \frac{1}{1-\rho} \cdot \mathbf{Var}[f].$$

(Hint: Use the Mean Value Theorem and Exercise 2.45.)

2.47 Let $f : \{-1,1\}^n \to \{-1,1\}$ be a transitive-symmetric function; in the notation of Exercise 1.30, this means the group $\mathrm{Aut}(f)$ acts transitively on $[n]$. Show that $\mathbf{Pr}_{\boldsymbol{\pi}\sim\mathrm{Aut}(f)}[\boldsymbol{\pi}(i) = j] = 1/n$ for all $i, j \in [n]$.

2.48 Suppose that $\mathbf{F}$ is a functional on functions $f : \{-1,1\}^n \to \mathbb{R}$ expressible as $\mathbf{F}[f] = \sum_S c_S \widehat{f}(S)^2$ where $c_S \geq 0$ for all $S \subseteq [n]$. (Examples include $\mathbf{Var}$, $\mathbf{W}^k$, $\mathbf{Inf}_i$, $\mathbf{I}$, $\mathbf{Inf}_i^{(1-\delta)}$, and $\mathbf{Stab}_\rho$ for $\rho \geq 0$.) Show that $\mathbf{F}$ is convex, meaning $\mathbf{F}[\lambda f + (1-\lambda)g] \leq \lambda \mathbf{F}[f] + (1-\lambda)\mathbf{F}[g]$ for all f, g, and $\lambda \in [0,1]$.

2.49 Extend the FKN Theorem as follows: Suppose $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{W}^{\le 1}[f] \ge 1-\delta$. Show that f is $O(\delta)$-close to a 1-junta. (Hint: Consider $g(x_0,x) = x_0 f(x_0 x)$.)

2.50 Compute the precise probability of a Condorcet winner (under impartial culture) in a 3-candidate, 3-voter election using $f = \mathrm{Maj}_3$.

2.51 (*a*) Arrow's Theorem for 3 candidates is slightly more general than what we stated: it allows for three *different* unanimous functions $f,g,h : \{-1,1\}^n \to \{-1,1\}$ to be used in the three pairwise elections. But show that if using f, g, h always gives rise to a Condorcet winner then $f = g = h$. (Hint: First show $g(x) = -f(-x)$ for all x by using the fact that x, $y = -x$, and $z = (f(x),\dots,f(x))$ is always a valid possibility for the votes.)

(*b*) Extend Arrow's Theorem to the case of Condorcet elections with more than 3 candidates.

2.52 The *polarizations* of $f : \{-1,1\}^n \to \mathbb{R}$ (also known as compressions, down-shifts, or two-point rearrangements) are defined as follows. For $i \in [n]$, the i-polarization of f is the function $f^{\sigma_i} : \{-1,1\}^n \to \mathbb{R}$ defined by

$$f^{\sigma_i}(x) = \begin{cases} \max\{f(x^{(i\mapsto +1)}), f(x^{(i\mapsto -1)})\} & \text{if } x_i = +1, \\ \min\{f(x^{(i\mapsto +1)}), f(x^{(i\mapsto -1)})\} & \text{if } x_i = -1. \end{cases}$$

(*a*) Show that $\mathbf{E}[f^{\sigma_i}] = \mathbf{E}[f]$ and $\|f^{\sigma_i}\|_p = \|f\|_p$ for all p.

(*b*) Show that $\mathbf{Inf}_j[f^{\sigma_i}] \le \mathbf{Inf}_j[f]$ for all $j \in [n]$.

(*c*) Show that $\mathbf{Stab}_\rho[f^{\sigma_i}] \ge \mathbf{Stab}_\rho[f]$ for all $0 \le \rho \le 1$.

(*d*) Show that f^{σ_i} is monotone in the ith direction (recall Exercise 2.5). Further, show that if f is monotone in the jth direction for some $j \in [n]$ then f^{σ_i} is still monotone in the jth direction.

(*e*) Let $f^* = f^{\sigma_1\sigma_2\cdots\sigma_n}$. Show that f^* is monotone, $\mathbf{E}[f^*] = \mathbf{E}[f]$, $\mathbf{Inf}_j[f^*] \le \mathbf{Inf}_j[f]$ for all $j \in [n]$, and $\mathbf{Stab}_\rho[f^*] \ge \mathbf{Stab}_\rho[f]$ for all $0 \le \rho \le 1$.

2.53 The Hamming distance $\Delta(x,y) = \#\{i : x_i \neq y_i\}$ on the discrete cube $\{-1,1\}^n$ is an example of an ℓ_1 *metric space*. For $D \ge 1$, we say that the discrete cube can be *embedded into* ℓ_2 *with distortion* D if there is a mapping $F : \{-1,1\}^n \to \mathbb{R}^m$ for some $m \in \mathbb{N}$ such that:

$$\|F(x) - F(y)\|_2 \ge \Delta(x,y) \text{ for all } x,\ y; \qquad \text{(“no contraction”)}$$

$$\|F(x) - F(y)\|_2 \le D \cdot \Delta(x,y) \text{ for all } x,\ y. \qquad \text{(“expansion at most } D\text{”)}$$

In this exercise you will show that the least distortion possible is $D = \sqrt{n}$.

(*a*) Recalling the definition of f^{odd} from Exercise 1.8, show that for any $f : \{-1,1\}^n \to \mathbb{R}$ we have $\|f^{\mathrm{odd}}\|_2^2 \le \mathbf{I}[f]$ and hence

$$\mathop{\mathbf{E}}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - f(-\boldsymbol{x}))^2] \le \sum_{i=1}^n \mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\left(f(\boldsymbol{x}) - f(\boldsymbol{x}^{\oplus i})\right)^2\right].$$

(*b*) Suppose $F : \{-1,1\}^n \to \mathbb{R}^m$, and write $F(x) = (f_1(x), f_2(x), \dots, f_m(x))$ for functions $f_i : \{-1,1\}^n \to \mathbb{R}$. By summing the above inequality over $i \in [m]$, show that any F with no contraction must have expansion at least $\sqrt{n}$.

(*c*) Show that there is an embedding F achieving distortion $\sqrt{n}$.

2.54 Give a Fourier-free proof of the Poincaré Inequality by induction on n.

2.55 Let V be a vector space with norm $\|\cdot\|$ and fix $w_1, \dots, w_n \in V$. Define $g : \{-1,1\}^n \to \mathbb{R}$ by $g(x) = \|\sum_{i=1}^n x_i w_i\|$.

(*a*) Show that $\mathrm{L}g \le g$ pointwise. (Hint: Triangle inequality.)

(*b*) Deduce $2\,\mathbf{Var}[g] \le \mathbf{E}[g^2]$ and thus the following *Khintchine–Kahane Inequality*:

$$\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\left\|\sum_{i=1}^n \boldsymbol{x}_i w_i\right\|\right] \ge \frac{1}{\sqrt{2}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\left\|\sum_{i=1}^n \boldsymbol{x}_i w_i\right\|^2\right]^{1/2}.$$

(Hint: Exercise 2.28.)

(*c*) Show that the constant $\frac{1}{\sqrt{2}}$ above is optimal, even if $V = \mathbb{R}$.

2.56 In the *correlation distillation* problem, a *source* chooses $\boldsymbol{x} \sim \{-1,1\}^n$ uniformly at random and broadcasts it to q *parties*. We assume that the transmissions suffer from some kind of noise, and therefore the players receive imperfect copies $\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(q)}$ of $\boldsymbol{x}$. The parties are not allowed to communicate, and despite having imperfectly correlated information they wish to agree on a single random bit. In other words, the ith party will output a bit $f_i(\boldsymbol{y}^{(i)}) \in \{-1,1\}$, and the goal is to find functions $f_1, \dots, f_q$ that maximize the probability that $f_1(\boldsymbol{y}^{(1)}) = f_2(\boldsymbol{y}^{(2)}) = \cdots = f_q(\boldsymbol{y}^{(q)})$. To avoid trivial deterministic solutions, we insist that $\mathbf{E}[f_i(\boldsymbol{y}^{(j)})]$ be 0 for all $j \in [q]$.

(*a*) Suppose $q = 2$, $\rho \in (0,1)$, and $\boldsymbol{y}^{(j)} \sim N_\rho(\boldsymbol{x})$ independently for each j. Show that the optimal solution is $f_1 = f_2 = \pm\chi_i$ for some $i \in [n]$. (Hint: You'll need Cauchy–Schwarz.)

(*b*) Show the same result for $q = 3$.

(*c*) Let $q = 2$ and $\rho \in (\frac{1}{2}, 1)$. Suppose that $\boldsymbol{y}^{(1)} = \boldsymbol{x}$ exactly, but $\boldsymbol{y}^{(2)} \in \{-1,0,1\}^n$ has *erasures*: it's formed from $\boldsymbol{x}$ by setting $\boldsymbol{y}^{(2)}_i = \boldsymbol{x}_i$ with probability ρ and $\boldsymbol{y}^{(2)}_i = 0$ with probability $1-\rho$, independently for all $i \in [n]$. Show that the optimal success probability is $\frac{1}{2} + \frac{1}{2}\rho$ and there is an optimal solution in which $f_1 = \pm\chi_i$ for any $i \in [n]$. (Hint: Eliminate the source, and introduce a fictitious party $1'$…)

(*d*) Consider the previous scenario but with $\rho \in (0, \frac{1}{2})$. Show that if n is sufficiently large, then the optimal solution does *not* have $f_1 = \pm\chi_i$.

2.57 (*a*) Let $g : \{-1,1\}^n \to \mathbb{R}^{\ge 0}$ have $\mathbf{E}[g] = \delta$. Show that for any $\rho \in [0,1]$,

$$\rho \sum_{j=1}^n |\widehat{g}(j)| \le \delta + \sum_{k=2}^n \rho^k \|g^{=k}\|_\infty.$$

(Hint: Exercise 2.31.)

(*b*) Assume further that $g : \{-1,1\}^n \to \{0,1\}$. Show that $\|g^{=k}\|_\infty \leq \sqrt{\delta}\sqrt{\binom{n}{k}}$. (Hint: First bound $\|g^{=k}\|_2^2$.) Deduce $\rho \sum_{j=1}^n |\widehat{g}(j)| \leq \delta + 2\rho^2\sqrt{\delta}n$, assuming $\rho \leq \frac{1}{2\sqrt{n}}$.

(*c*) Show that $\sum_{j=1}^n |\widehat{g}(j)| \leq 2\sqrt{2}\delta^{3/4}\sqrt{n}$ (assuming $\delta \leq 1/4$). Deduce $\mathbf{W}^1[g] \leq 2\sqrt{2}\cdot\delta^{7/4}\sqrt{n}$. (Hint: show $|\widehat{g}(j)| \leq \delta$ for all j.)

(*d*) Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is monotone and $\mathbf{MaxInf}[f] \leq \delta$. Show $\mathbf{W}^2[f] \leq \sqrt{2}\cdot\delta^{3/4}\cdot\mathbf{I}[f]\cdot\sqrt{n}$.

(*e*) Suppose further that f is unbiased. Show that $\mathbf{MaxInf}[f] \leq o(n^{-2/3})$ implies $\mathbf{I}[f] \geq 3 - o(1)$; conclude $\mathbf{MaxInf}[f] \geq \frac{3}{n} - o(1/n)$. (Hint: Extend Exercise 2.29.) Use Exercise 2.52 to remove the assumption that f is monotone for these statements.

2.58 Let V be a vector space (over $\mathbb{R}$) with norm $\|\cdot\|_V$. If $f : \{-1,1\}^n \to V$ we can define its Fourier coefficients $\widehat{f}(S) \in V$ by the usual formula $\widehat{f}(S) = \mathbf{E}_{\boldsymbol{x}\in\{-1,1\}^n}[f(\boldsymbol{x})\boldsymbol{x}^S]$. We may also define $\|f\|_p = \mathbf{E}_{\boldsymbol{x}\in\{-1,1\}^n}[\|f(\boldsymbol{x})\|_V^p]^{1/p}$. Finally, if the norm $\|\cdot\|_V$ arises from an inner product $\langle\cdot,\cdot\rangle_V$ on V we can define an inner product on functions $f,g : \{-1,1\}^n \to V$ by $\langle f,g\rangle = \mathbf{E}_{\boldsymbol{x}\in\{-1,1\}^n}[\langle f(\boldsymbol{x}),g(\boldsymbol{x})\rangle_V]$. The material developed so far in this book has used $V = \mathbb{R}$ with $\langle\cdot,\cdot\rangle_V$ being multiplication. Explore the extent to which this material extends to the more general setting.

Notes. The mathematical study of social choice began in earnest in the late 1940s; see Riker [**Rik61**] for an early survey or the compilation [**BGR09**] for some modern results. Arrow's Theorem was the field's first major result; Arrow proved it in 1950 [**Arr50**] under the extra assumption of monotonicity (and with a minor error [**Bla57**]), with the refined version appearing in 1963 [**Arr63**]. He was awarded the Nobel Prize for this work in 1972. May's Theorem is from 1952 [**May52**]. Guilbaud's Formula is also from 1952 [**Gui52**], though Guilbaud only stated it in a footnote and wrote that it is computed "by the usual means in combinatorial analysis". The first published proof appears to be due to Garman and Kamien [**GK68**]; they also introduced the impartial culture assumption. The term "junta" appears to have been introduced by Parnas, Ron, and Samorodnitsky [**PRS01**].

The notion of influence $\mathbf{Inf}_i[f]$ was originally introduced by the geneticist Penrose [**Pen46**], who observed that $\mathbf{Inf}_i[\mathrm{Maj}_n] \sim \frac{\sqrt{2/\pi}}{\sqrt{n}}$. It was rediscovered by the lawyer Banzhaf in 1965 [**Ban65**]; he sued the Nassau County (NY) Board after proving that the voting system it used (the one in Exercise 2.9) gave some towns zero influence. Influence is sometimes referred to as the Banzhaf, Penrose–Banzhaf, or Banzhaf–Coleman index (Coleman being another rediscoverer [**Col71**]). Influences were first studied in the computer science literature by Ben-Or and Linial [**BL85**]; they introduced also introduced "tribes" as an example of a function with constant variance yet small influences. The Fourier formulas for influence may have first appeared in the work of Chor and Geréb-Graus [**CGG87**].

Total influence of Boolean functions has long been studied in combinatorics, since it is equivalent to edge-boundary size for subsets of the Hamming cube. For example, the edge-isoperimetric inequality was first proved by Harper in 1964 [**Har64**]. In the context of Boolean functions, Karpovsky [**Kar76**] proposed $\mathbf{I}[f]$ as a measure of the computational complexity of f, and Hurst, Miller, and Muzio [**HMM82**] gave the Fourier formula $\sum_S |S|\widehat{f}(S)^2$. The terminology "Poincaré Inequality" comes from the theory of functional inequalities and Markov chains; the inequality is equivalent to the *spectral gap* for the discrete cube graph.

The noise stability of Boolean functions was first studied explicitly by Benjamini, Kalai, and Schramm in 1999 [**BKS99**], though it plays an important role in the earlier work of Håstad [**Hås97**]. See O'Donnell [**O'D03**] for a survey. The noise operator was introduced by Bonami [**Bon70**] and independently by Beckner [**Bec75**], who used the notation T_ρ which was standardized by Kahn, Kalai, and Linial [**KKL88**]. For nonnegative noise rates it's often natural to use the alternate parameterization $\mathrm{T}_{e^{-t}}$ for $t \in [0,\infty]$.

The Fourier approach to Arrow's Theorem is due to Kalai [**Kal02**]; he also proved Theorem 2.57 and Corollary 2.60. The FKN Theorem is due to Friedgut, Kalai, and Naor [**FKN02**]; the observation from Exercise 2.49 is due to Kindler.

The polarizations from Exercise 2.52 originate in Kleitman [**Kle66**]. Exercise 2.53 is a theorem of Enflo from 1970 [**Enf70**]. Exercise 2.55 is a theorem of Latała and Oleszkiewicz [**LO94**]. In Exercise 2.56, part *(b)* is due to Mossel and O'Donnell [**MO05**]; part *(c)* was conjectured by Yang [**Yan04**] and proved by O'Donnell and Wright [**OW12**]. Exercise 2.57 is a polishing of the 1987 work by Chor and Geréb-Graus [**CGG87, CGG88**], a precursor of the KKL Theorem. The weaker Exercise 2.29 is also due to them and Noga Alon independently.

Spectral structure and learning

One reasonable way to assess the "complexity" of a Boolean function is in terms how complex its Fourier spectrum is. For example, functions with sufficiently simple Fourier spectra can be efficiently *learned* from examples. This chapter will be concerned with understanding the location, magnitude, and structure of a Boolean function's Fourier spectrum.

3.1. Low-degree spectral concentration

One way a Boolean function's Fourier spectrum can be "simple" is for it to be mostly concentrated at small degree.

Definition 3.1. We say that the Fourier spectrum of $f : \{-1,1\}^n \to \mathbb{R}$ is ϵ-*concentrated on degree up to* k if

$$\mathbf{W}^{>k}[f] = \sum_{\substack{S\subseteq[n] \\ |S|>k}} \widehat{f}(S)^2 \leq \epsilon.$$

For $f : \{-1,1\}^n \to \{-1,1\}$ we can express this condition using the spectral sample: $\mathbf{Pr}_{\boldsymbol{S}\sim\mathcal{S}_f}[|\boldsymbol{S}| > k] \leq \epsilon$.

It's possible to show such a concentration result combinatorially by showing that a function has small total influence:

Proposition 3.2. *For any* $f : \{-1,1\}^n \to \mathbb{R}$ *and* $\epsilon > 0$*, the Fourier spectrum of* f *is* ϵ*-concentrated on degree up to* $\mathbf{I}[f]/\epsilon$*.*

Proof. This follows immediately from Theorem 2.38, $\mathbf{I}[f] = \sum_{k=0}^{n} k \cdot \mathbf{W}^{k}[f]$. For $f : \{-1,1\}^n \to \{-1,1\}$, this is Markov's inequality applied to the cardinality of the spectral sample. □

For example, in Exercise 2.13 you showed that $\mathbf{I}[\mathrm{Tribes}_{w,2^w}] \leq O(\log n)$, where $n = w2^w$; thus this function's spectrum is .01-concentrated on degree up to $O(\log n)$, a rather low level. Proving this by explicitly calculating Fourier coefficients would be quite painful.

Another means of showing low-degree spectral concentration is through noise stability/sensitivity:

Proposition 3.3. *For any $f : \{-1,1\}^n \to \{-1,1\}$ and $\delta \in (0,1/2]$, the Fourier spectrum of f is ϵ-concentrated on degree up to $1/\delta$ for*

$$\epsilon = \tfrac{2}{1-e^{-2}}\mathbf{NS}_\delta[f] \leq 3\mathbf{NS}_\delta[f].$$

Proof. Using the Fourier formula from Theorem 2.49,

$$\begin{aligned} 2\mathbf{NS}_\delta[f] &= \mathop{\mathbf{E}}_{\boldsymbol{S}\sim\mathcal{S}_f}[1-(1-2\delta)^{|\boldsymbol{S}|}] \\ &\geq (1-(1-2\delta)^{1/\delta}) \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{S}\sim\mathcal{S}_f}[|\boldsymbol{S}| \geq 1/\delta] \\ &\geq (1-e^{-2}) \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{S}\sim\mathcal{S}_f}[|\boldsymbol{S}| \geq 1/\delta], \end{aligned}$$

where the first inequality used that $1-(1-2\delta)^k$ is a nonnegative nondecreasing function of k. The claim follows. □

As an example, Theorem 2.45 tells us that for $\delta > 0$ sufficiently small and n sufficiently large (as a function of δ), $\mathbf{NS}_\delta[\mathrm{Maj}_n] \leq \sqrt{\delta}$. Hence the Fourier spectrum of Maj_n is $3\sqrt{\delta}$-concentrated on degree up to $1/\delta$; equivalently, it is ϵ-concentrated on degree up to $9/\epsilon^2$. (We will give sharp constants for majority's spectral concentration in Chapter 5.3.) This example also shows there is no simple converse to Proposition 3.2; although Maj_n has its spectrum .01-concentrated on degree up to $O(1)$, its total influence is $\Theta(\sqrt{n})$.

Finally, suppose a function $f : \{-1,1\}^n \to \{-1,1\}$ has its Fourier spectrum 0-*concentrated* up to degree k; in other words, f has real degree $\deg(f) \leq k$. In this case f must be somewhat simple; indeed, if k is a constant, then f is a junta:

Theorem 3.4. *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ has $\deg(f) \leq k$. Then f is a $k2^{k-1}$-junta.*

The bound $k2^{k-1}$ cannot be significantly improved; see Exercise 3.24. The key to proving Theorem 3.4 is the following lemma, the proof of which is outlined in Exercise 3.4:

Lemma 3.5. *Suppose* $\deg(f) \le k$*, where* $f : \{-1,1\}^n \to \mathbb{R}$ *is not identically* 0*. Then* $\mathbf{Pr}[f(\boldsymbol{x}) \neq 0] \ge 2^{-k}$.

Since $\deg(\mathrm{D}_i f) \le k-1$ when $\deg(f) \le k$ (by the "differentiation" formula) and since $\mathbf{Inf}_i[f] = \mathbf{Pr}[\mathrm{D}_i f(\boldsymbol{x}) \neq 0]$ for Boolean-valued f, we immediately infer:

Proposition 3.6. *If* $f : \{-1,1\}^n \to \{-1,1\}$ *has* $\deg(f) \le k$ *then* $\mathbf{Inf}_i[f]$ *is either* 0 *or at least* 2^{1-k} *for all* $i \in [n]$.

We can now give the proof of Theorem 3.4. From Proposition 3.6 the number of coordinates which have nonzero influence on f is at most $\mathbf{I}[f]/2^{1-k}$, and this in turn is at most $k2^{k-1}$ by the following fact:

Fact 3.7. *For* $f : \{-1,1\}^n \to \{-1,1\}$, $\mathbf{I}[f] \le \deg(f)$.

Fact 3.7 is immediate from the Fourier formula for total influence.

We remark that the FKN Theorem (stated in Chapter 2.5) is a "robust" version of Theorem 3.4 for $k = 1$. In Chapter 9.6 we will see Friedgut's Junta Theorem, a related robust result showing that if $\mathbf{I}[f] \le k$ then f is ϵ-close to a $2^{O(k/\epsilon)}$-junta.

3.2. Subspaces and decision trees

In this section we treat the domain of a Boolean function as $\mathbb{F}_2^n$, an n-dimensional vector space over the field $\mathbb{F}_2$. As mentioned in Chapter 1.2, it can be natural to index the Fourier characters $\chi_S : \mathbb{F}_2^n \to \{-1,1\}$ not by subsets $S \subseteq [n]$ but by their 0-1 indicator vectors $\gamma \in \mathbb{F}_2^n$; thus

$$\chi_\gamma(x) = (-1)^{\gamma \cdot x},$$

with the dot product $\gamma \cdot x$ being carried out in $\mathbb{F}_2^n$. For example, in this notation we'd write χ_0 for the constantly 1 function and χ_{e_i} for the ith dictator. Fact 1.6 now becomes

$$\chi_\beta \chi_\gamma = \chi_{\beta+\gamma} \quad \forall \beta, \gamma. \tag{3.1}$$

Thus the characters form a group under multiplication, which is isomorphic to the group $\mathbb{F}_2^n$ under addition. To distinguish this group from the input domain we write it as $\widehat{\mathbb{F}_2^n}$; we also tend to identify the character with its index. Thus the Fourier expansion of $f : \mathbb{F}_2^n \to \mathbb{R}$ can be written as

$$f(x) = \sum_{\gamma \in \widehat{\mathbb{F}_2^n}} \widehat{f}(\gamma)\chi_\gamma(x).$$

The Fourier transform of f can be thought of as a function $\widehat{f} : \widehat{\mathbb{F}_2^n} \to \mathbb{R}$. We can measure its complexity with various norms.

Definition 3.8. The *Fourier (or spectral) p-norm* of $f : \{-1,1\}^n \to \mathbb{R}$ is

$$\hat{\|} f \hat{\|}_p = \left(\sum_{\gamma \in \widehat{\mathbb{F}_2^n}} |\widehat{f}(\gamma)|^p \right)^{1/p}.$$

Note that we use the "counting measure" on $\widehat{\mathbb{F}_2^n}$, and hence we have a nice rephrasing of Parseval's Theorem: $\|f\|_2 = \hat{\|} f \hat{\|}_2$. We make two more definitions relating to the simplicity of $\widehat{f}$:

Definition 3.9. The *Fourier (or spectral) sparsity* of $f : \{-1,1\}^n \to \mathbb{R}$ is

$$\mathrm{sparsity}(\widehat{f}) = |\mathrm{supp}(\widehat{f})| = \#\left\{\gamma \in \widehat{\mathbb{F}_2^n} : \widehat{f}(\gamma) \neq 0\right\}.$$

Definition 3.10. We say that $\widehat{f}$ is *ϵ-granular* if $\widehat{f}(\gamma)$ is an integer multiple of ϵ for all $\gamma \in \widehat{\mathbb{F}_2^n}$.

To gain some practice with this notation, let's look at the Fourier transforms of some indicator functions $1_A : \mathbb{F}_2^n \to \{0,1\}$ and probability density functions φ_A, where $A \subseteq \mathbb{F}_2^n$. First, suppose $A \leq \mathbb{F}_2^n$ is a *subspace*. Then one way to characterize A is by its *perpendicular subspace* $A^\perp$:

$$A^\perp = \{\gamma \in \widehat{\mathbb{F}_2^n} : \gamma \cdot x = 0 \text{ for all } x \in A\}.$$

It holds that $\dim A^\perp = n - \dim A$ (this is called the *codimension* of A) and that $A = (A^\perp)^\perp$.

Proposition 3.11. *If $A \leq \mathbb{F}_2^n$ has $\operatorname{codim} A = \dim A^\perp = k$, then*

$$1_A = \sum_{\gamma \in A^\perp} 2^{-k} \chi_\gamma, \qquad \varphi_A = \sum_{\gamma \in A^\perp} \chi_\gamma.$$

Proof. Let $\gamma_1, \dots, \gamma_k$ form a basis of $A^\perp$. Since $A = (A^\perp)^\perp$ it follows that $x \in A$ if and only if $\chi_{\gamma_i}(x) = 1$ for all $i \in [k]$. We therefore have

$$1_A(x) = \prod_{i=1}^k \left(\tfrac{1}{2} + \tfrac{1}{2}\chi_{\gamma_i}(x)\right) = 2^{-k} \sum_{\gamma \in \mathrm{span}\{\gamma_1, \dots, \gamma_k\}} \chi_\gamma(x)$$

as claimed, where the last equality used (3.1). The Fourier expansion of φ_A follows because $\mathbf{E}[1_A] = 2^{-k}$. □

More generally, suppose A is *affine subspace* (or *coset*) of $\mathbb{F}_2^n$; i.e., $A = H + a$ for some $H \leq \mathbb{F}_2^n$ and $a \in \mathbb{F}_2^n$, or equivalently

$$A = \{x \in \mathbb{F}_2^n : \gamma \cdot x = \gamma \cdot a \text{ for all } \gamma \in H^\perp\}.$$

Then it is easy (Exercise 3.11) to extend Proposition 3.11 to:

Proposition 3.12. *If $A = H + a$ is an affine subspace of codimension k, then*

$$\widehat{1_A}(\gamma) = \begin{cases} \chi_\gamma(a)2^{-k} & \text{if } \gamma \in H^\perp \\ 0 & \text{else;} \end{cases}$$

hence $\varphi_A = \sum_{\gamma \in H^\perp} \chi_\gamma(a)\chi_\gamma$. We have $\mathrm{sparsity}(\widehat{1_A}) = 2^k$*, $\widehat{1_A}$ is 2^{-k}-granular, $\hat{\|} 1_A \hat{\|}_\infty = 2^{-k}$, and $\hat{\|} 1_A \hat{\|} = 1$.*

In computer science terminology, any $f : \mathbb{F}_2^n \to \{0,1\}$ that is a conjunction of parity conditions is the indicator of an affine subspace (or the zero function). In the simple case that the parity conditions are all of the form "$x_i = a_i$", the function is a logical AND of *literals*, and we call the affine subspace a *subcube*.

Another class of Boolean functions with simple Fourier spectra are the ones computable by simple *decision trees*:

Definition 3.13. A *decision tree* T is a representation of a Boolean function $f : \mathbb{F}_2^n \to \mathbb{R}$. It consists of a rooted binary tree in which the internal nodes are labeled by coordinates $i \in [n]$, the outgoing edges of each internal node are labeled 0 and 1, and the leaves are labeled by real numbers. We insist that no coordinate $i \in [n]$ appears more than once on any root-to-leaf path.

On input $x \in \mathbb{F}_2^n$, the tree T constructs a *computation path* from the root node to a leaf. Specifically, when the computation path reaches an internal node labeled by coordinate $i \in [n]$ we say that T *queries* x_i; the computation path then follows the outgoing edge labeled by x_i. The output of T (and hence f) on input x is the label of the leaf reached by the computation path. We often identify a tree with the function it computes.

For decision trees, a picture is worth a thousand words; see Figure 3.1.

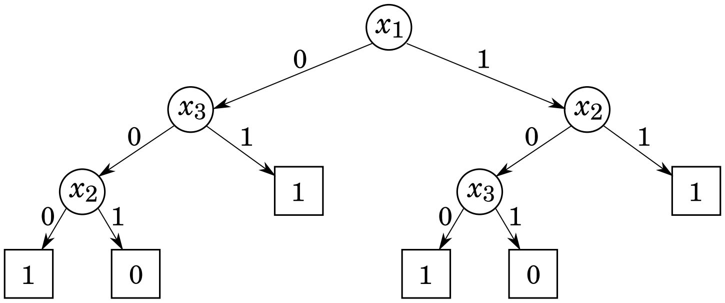


Figure 3.1. Decision tree computing Sort_3

(It's traditional to write x_i rather than i for the internal node labels.) For example, the computation path of the above tree on input $x = (0,1,0) \in \mathbb{F}_2^3$ starts at the root, queries x_1, proceeds left, queries x_3, proceeds left, queries

x_2, proceeds right, and reaches a leaf labeled 0. In fact, this tree computes the function Sort_3 defined by $\mathrm{Sort}_3(x) = 1$ if and only if $x_1 \le x_2 \le x_3$ or $x_1 \ge x_2 \ge x_3$.

Definition 3.14. The *size* s of a decision tree T is the total number of leaves. The *depth* k of T is the maximum length of any root-to-leaf path. For decision trees over $\mathbb{F}_2^n$ we have $k \le n$ and $s \le 2^k$. Given $f : \mathbb{F}_2^n \to \mathbb{R}$ we write $\mathrm{DT}(f)$ (respectively, $\mathrm{DT}_{\mathrm{size}}(f)$) for the least depth (respectively, size) of a decision tree computing f.

The example decision tree above has size 6 and depth 3.

Let T be a decision tree computing $f : \mathbb{F}_2^n \to \mathbb{R}$ and let P be one of its root-to-leaf paths. The set of inputs x that follow computation path P in T is precisely a subcube of $\mathbb{F}_2^n$, call it C_P. The function f is constant on C_P; we will call its value there $f(P)$. Further, since every input x follows a unique path in T, the subcubes $\{C_P : P \text{ a path in } T\}$ form a *partition* of $\mathbb{F}_2^n$. These observations yield the following "spectral simplicity" results for decision trees:

Fact 3.15. *Let $f : \mathbb{F}_2^n \to \mathbb{R}$ be computed by a decision tree T. Then*

$$f = \sum_{\textit{paths } P \textit{ of } T} f(P) \cdot 1_{C_P}.$$

Proposition 3.16. *Let $f : \mathbb{F}_2^n \to \mathbb{R}$ be computed by a decision tree T of size s and depth k. Then:*

- $\deg(f) \le k$;
- $\mathrm{sparsity}(\widehat{f}) \le s2^k \le 4^k$;
- $\hat{\|} f \hat{\|}_1 \le \|f\|_\infty \cdot s \le \|f\|_\infty \cdot 2^k$;
- $\widehat{f}$ *is* 2^{-k}*-granular assuming* $f : \mathbb{F}_2^n \to \mathbb{Z}$.

Proposition 3.17. *Let $f : \mathbb{F}_2^n \to \{-1,1\}$ be computable by a decision tree of size s and let $\epsilon \in (0,1]$. Then the spectrum of f is ϵ-concentrated on degree up to* $\log(s/\epsilon)$.

You are asked to prove these propositions in Exercises 3.21 and 3.22. Similar spectral simplicity results hold for some generalizations of the decision tree representation ("subcube partitions", "parity decision trees"); see Exercise 3.26.

3.3. Restrictions

A common operation on Boolean functions $f : \{-1,1\}^n \to \mathbb{R}$ is *restriction* to subcubes. Suppose $[n]$ is partitioned into two sets, J and $\overline{J} = [n] \setminus J$. If the inputs bits in $\overline{J}$ are fixed to constants, the result is a function $\{-1,1\}^J \to \mathbb{R}$. For example, if we take the function $\mathrm{Maj}_5 : \{-1,1\}^5 \to \{-1,1\}$ and restrict the 4th and 5th coordinates to be 1 and -1 respectively, we obtain the function

$\mathrm{Maj}_3 : \{-1,1\}^3 \to \{-1,1\}$. If we further restrict the 3rd coordinate to be -1, we obtain the two-bit function which is 1 if and only if both input bits are 1.

We introduce following notation:

Definition 3.18. Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $(J, \overline{J})$ be a partition of $[n]$. Let $z \in \{-1,1\}^{\overline{J}}$. Then we write $f_{J|z} : \{-1,1\}^J \to \mathbb{R}$ (pronounced "the restriction of f to J using z") for the subfunction of f given by fixing the coordinates in $\overline{J}$ to the bit values z. When the partition $(J, \overline{J})$ is understood we may write simply $f_{|z}$. If $y \in \{-1,1\}^J$ and $z \in \{-1,1\}^{\overline{J}}$ we will sometimes write (y,z) for the composite string in $\{-1,1\}^n$, even though y and z are not literally concatenated; with this notation, $f_{J|z}(y) = f(y,z)$.

Let's examine how restrictions affect the Fourier transform by considering an example.

Example 3.19. Let $f : \{-1,1\}^4 \to \{-1,1\}$ be the function defined by

$$f(x) = 1 \quad \iff \quad x_3 = x_4 = -1 \text{ or } x_1 \geq x_2 \geq x_3 \geq x_4 \text{ or } x_1 \leq x_2 \leq x_3 \leq x_4. \tag{3.2}$$

You can check that f has the Fourier expansion

$$\begin{aligned} f(x) = &+ \tfrac{1}{8} - \tfrac{1}{8}x_1 + \tfrac{1}{8}x_2 - \tfrac{1}{8}x_3 - \tfrac{1}{8}x_4 \\ &+ \tfrac{3}{8}x_1x_2 + \tfrac{1}{8}x_1x_3 - \tfrac{3}{8}x_1x_4 + \tfrac{3}{8}x_2x_3 - \tfrac{1}{8}x_2x_4 + \tfrac{5}{8}x_3x_4 \\ &+ \tfrac{1}{8}x_1x_2x_3 + \tfrac{1}{8}x_1x_2x_4 - \tfrac{1}{8}x_1x_3x_4 + \tfrac{1}{8}x_2x_3x_4 - \tfrac{1}{8}x_1x_2x_3x_4. \end{aligned} \tag{3.3}$$

Consider the restriction $x_3 = 1$, $x_4 = -1$, and let $f' = f_{\{1,2\}|(1,-1)}$ be the restricted function of x_1 and x_2. From the original definition (3.2) of f we see that $f'(x_1,x_2)$ is 1 if and only if $x_1 = x_2 = 1$. This is the $\min_2$ function of x_1 and x_2, which we know has Fourier expansion

$$f'(x_1,x_2) = \min_2(x_1,x_2) = -\tfrac{1}{2} + \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{2}x_1x_2. \tag{3.4}$$

We can of course obtain this expansion simply by plugging $x_3 = 1, x_4 = -1$ into (3.3). Now suppose we only wanted to know the coefficient on x_1 in the Fourier expansion of f'. We can find it as follows: Consider all monomials in (3.3) that contain x_1 and possibly also x_3, x_4; substitute $x_3 = 1$, $x_4 = -1$ into the associated terms; and sum the results. The relevant terms in (3.3) are $-\frac{1}{8}x_1$, $+\frac{1}{8}x_1x_3$, $-\frac{3}{8}x_1x_4$, $-\frac{1}{8}x_1x_3x_4$, and substituting in $x_3 = 1, x_4 = -1$ gives us $-\frac{1}{8} + \frac{1}{8} + \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$, as expected from (3.4).

Now we work out these ideas more generally. In the setting of Definition 3.18 the restricted function $f_{J|z}$ has $\{-1,1\}^J$ as its domain. Thus its Fourier coefficients are indexed by subsets of J. Let's introduce notation for the Fourier coefficients of a restricted function:

Definition 3.20. Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $(J, \overline{J})$ be a partition of $[n]$. Let $S \subseteq J$. Then we write $\mathrm{F}_{S|\overline{J}} f : \{-1,1\}^{\overline{J}} \to \mathbb{R}$ for the function $\widehat{f_{J|\bullet}}(S)$; i.e.,

$$\mathrm{F}_{S|\overline{J}} f(z) = \widehat{f_{J|z}}(S).$$

When the partition $(J, \overline{J})$ is understood we may write simply $\mathrm{F}_{S|} f$.

In Example 3.19 we considered $\overline{J} = \{3,4\}$, $S = \{1\}$, and $z = (1,-1)$. See Figure 3.2 for an illustration of a typical restriction scenario.

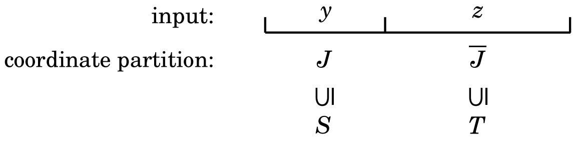


Figure 3.2. Notation for a typical restriction scenario. Note that J and $\overline{J}$ need not be literally contiguous.

In general, for a fixed partition $(J, \overline{J})$ of $[n]$ and a fixed $S \subseteq J$, we may wish to know what $\widehat{f_{J|z}}(S)$ is as a function of $z \in \{-1,1\}^{\overline{J}}$. This is precisely asking for the Fourier transform of $\mathrm{F}_{S|\overline{J}} f$. Since the function $\mathrm{F}_{S|\overline{J}} f$ has domain $\{-1,1\}^{\overline{J}}$, its Fourier transform has coefficients indexed by subsets of $\overline{J}$. The formula for this Fourier transform generalizes the computation we used at the end of Example 3.19:

Proposition 3.21. *In the setting of Definition 3.20 we have the Fourier expansion*

$$\mathrm{F}_{S|\overline{J}} f(z) = \sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T) z^T;$$

i.e.,

$$\widehat{\mathrm{F}_{S|\overline{J}} f}(T) = \widehat{f}(S \cup T).$$

Proof. (The $S = \emptyset$ case here is Exercise 1.15.) Every $U \subseteq [n]$ indexing f's Fourier coefficients can be written as a disjoint union $U = S \cup T$, where $S \subseteq J$ and $T \subseteq \overline{J}$. We can also decompose any $x \in \{-1,1\}^n$ into two substrings $y \in \{-1,1\}^J$ and $z \in \{-1,1\}^{\overline{J}}$. We have $x^U = y^S z^T$ and so

$$f(x) = \sum_{U \subseteq [n]} \widehat{f}(U) x^U = \sum_{\substack{S \subseteq J \\ T \subseteq \overline{J}}} \widehat{f}(S \cup T) y^S z^T = \sum_{S \subseteq J} \Big(\sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T) z^T \Big) y^S.$$

Thus when z is fixed, the resulting function of y indeed has $\sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T) z^T$ as its Fourier coefficient on the monomial y^S. □

Corollary 3.22. *Let $f : \{-1,1\}^n \to \mathbb{R}$, let $(J, \overline{J})$ be a partition of $[n]$, and fix $S \subseteq J$. Suppose $\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}$ is chosen uniformly at random. Then*

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

Proof. The first statement is immediate from Proposition 3.21, taking $T = \emptyset$ and unraveling the definition. As for the second statement,

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{z}}[\widehat{f_{J|\boldsymbol{z}}}(S)^2] &= \mathop{\mathbf{E}}_{\boldsymbol{z}}[\mathrm{F}_{S|\overline{J}} f(\boldsymbol{z})^2] && \text{(by definition)}\\ &= \sum_{T \subseteq \overline{J}} \widehat{\mathrm{F}_{S|\overline{J}} f}(T)^2 && \text{(Parseval)}\\ &= \sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T)^2 && \text{(Proposition 3.21)} \quad \square\end{aligned}$$

We move on to discussing a more general kind of restriction; namely, restricting a function $f : \mathbb{F}_2^n \to \mathbb{R}$ to an affine subspace $H + z$. This generalizes restriction to subcubes as we've seen so far, by considering $H = \mathrm{span}\{e_i : i \in J\}$ for a given subset $J \subseteq [n]$. For restrictions to a subspace $H \le \mathbb{F}_2^n$ we have a natural definition:

Definition 3.23. If $f : \mathbb{F}_2^n \to \mathbb{R}$ and $H \le \mathbb{F}_2^n$ is a subspace, we write $f_H : H \to \mathbb{R}$ for the restriction of f to H.

For restrictions to *affine* subspaces, we run into difficulties if we try to extend our notation for restrictions to subcubes. Unlike in the subcube case of $H = \mathrm{span}\{e_i : i \in J\}$, we don't in general have a canonical isomorphism between H and a coset $H + z$. Thus it's not natural to introduce notation such as $f_{H|z} : H \to \mathbb{R}$ for the function $h \mapsto f(h+z)$, because such a definition depends on the choice of representative for $H + z$. As an example consider $H = \{(0,0),(1,1)\} \le \mathbb{F}_2^2$, a 1-dimensional subspace (which satisfies $H^\perp = H$). Here the nontrivial coset is $H + (1,0) = H + (0,1) = \{(1,0),(0,1)\}$, which has no canonical representative.

To get around this difficulty we can view restriction to a coset $H + z$ as consisting of two steps: first, translation of the domain by a fixed representative z, and then restriction to the subspace H. Let's introduce some notation for the first operation:

Definition 3.24. Let $f : \mathbb{F}_2^n \to \mathbb{R}$ and let $z \in \mathbb{F}_2^n$. We define the function $f^{+z} : \mathbb{F}_2^n \to \mathbb{R}$ by $f^{+z}(x) = f(x+z)$.

By substituting $x = x + z$ into the Fourier expansion of f, we deduce:

Fact 3.25. *The Fourier coefficients of f^{+z} are given by $\widehat{f^{+z}}(\gamma) = (-1)^{\gamma \cdot z}\widehat{f}(\gamma)$; i.e.,*

$$f^{+z}(x) = \sum_{\gamma \in \widehat{\mathbb{F}_2^n}} \chi_\gamma(z)\widehat{f}(\gamma)\chi_\gamma(x).$$

(This fact also follows by noting that $f^{+z} = \varphi_{\{z\}} * f$; see Exercise 3.31.)

We can now give notation for the restriction of a function to an affine subspace:

Definition 3.26. Let $f : \mathbb{F}_2^n \to \mathbb{R}$, $z \in \mathbb{F}_2^n$, $H \leq \mathbb{F}_2^n$. We write $f_H^{+z} : H \to \mathbb{R}$ for the function $(f^{+z})_H$; namely, the restriction of f to coset $H + z$ with the representative z made explicit.

Finally, we would like to consider Fourier coefficients of restricted functions f_H^{+z}. These can be indexed by the cosets of $H^\perp$ in $\widehat{\mathbb{F}_2^n}$. However, we again have a notational difficulty since the only coset with a canonical representative is $H^\perp$ itself, with representative 0. There is no need to introduce extra notation for $\widehat{f_H^{+z}}(0)$, the average value of f on coset $H + z$, since it is just

$$\mathop{\mathbf{E}}_{\boldsymbol{h} \sim H}[f(\boldsymbol{h} + z)] = \langle \varphi_H, f^{+z} \rangle.$$

Applying Plancherel on the right-hand side, as well as Proposition 3.11 and Fact 3.25, we deduce the following classical fact:

Poisson Summation Formula. *Let $f : \mathbb{F}_2^n \to \mathbb{R}$, $H \leq \mathbb{F}_2^n$, $z \in \mathbb{F}_2^n$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{h} \sim H}[f(\boldsymbol{h} + z)] = \sum_{\gamma \in H^\perp} \chi_\gamma(z)\widehat{f}(\gamma).$$

3.4. Learning theory

Computational learning theory is an area of algorithms research devoted to the following task: Given a source of "examples" $(x, f(x))$ from an unknown function f, compute a "hypothesis" function h that is good at predicting $f(y)$ on future inputs y. In this book we will focus on just one possible formulation of the task:

Definition 3.27. In the model of *PAC ("Probably Approximately Correct") learning under the uniform distribution on* $\{-1,1\}^n$, a learning problem is identified with a *concept class* $\mathscr{C}$, which is just a collection of functions $f : \{-1,1\}^n \to \{-1,1\}$. A *learning algorithm* A for $\mathscr{C}$ is a randomized algorithm which has limited access to an unknown *target function* $f \in \mathscr{C}$. The two access models, in increasing order of strength, are:

- *random examples*, meaning A can draw pairs $(\boldsymbol{x}, f(\boldsymbol{x}))$ where $\boldsymbol{x} \in \{-1,1\}^n$ is uniformly random;

- *queries*, meaning A can request the value $f(x)$ for any $x \in \{-1,1\}^n$ of its choice.

In addition, A is given as input an *accuracy parameter* $\epsilon \in [0,1/2]$. The output of A is required to be (the circuit representation of) a *hypothesis* function $h : \{-1,1\}^n \to \{-1,1\}$. We say that A *learns* $\mathscr{C}$ *with error* ϵ if for any $f \in \mathscr{C}$, with high probability A outputs an h which is ϵ-close to f: i.e., satisfies $\mathrm{dist}(f,h) \leq \epsilon$.

In the above definition, the phrase "with high probability" can be fixed to mean, say, "except with probability at most 1/10". (As is common with randomized algorithms, the choice of constant 1/10 is unimportant; see Exercise 3.40.)

For us, the main desideratum of a learning algorithm is efficient *running time*. One can easily learn *any* function f to error 0 in time $\widetilde{O}(2^n)$ (see Exercise 3.33); however, this is not very efficient. If the concept class $\mathscr{C}$ contains very complex functions, then such exponential running time is necessary; however, if $\mathscr{C}$ contains only relatively "simple" functions, then more efficient learning may be possible. For example, the results of Section 3.5 show that the concept class

$$\mathscr{C} = \{f : \mathbb{F}_2^n \to \{-1,1\} \mid \mathrm{DT}_{\mathrm{size}}(f) \leq s\}$$

can be learned with queries to error ϵ by an algorithm running in time $\mathrm{poly}(s, n, 1/\epsilon)$.

A common way of trying to learn an unknown target $f : \{-1,1\}^n \to \{-1,1\}$ is by discovering "most of" its Fourier spectrum. To formalize this, let's generalize Definition 3.1:

Definition 3.28. Let $\mathscr{F}$ be a collection of subsets $S \subseteq [n]$. We say that the Fourier spectrum of $f : \{-1,1\}^n \to \mathbb{R}$ is *ϵ-concentrated on $\mathscr{F}$* if

$$\sum_{\substack{S \subseteq [n] \\ S \notin \mathscr{F}}} \widehat{f}(S)^2 \leq \epsilon.$$

For $f : \{-1,1\}^n \to \{-1,1\}$ we can express this condition using the spectral sample: $\mathbf{Pr}_{\boldsymbol{S} \sim \mathscr{S}_f}[\boldsymbol{S} \notin \mathscr{F}] \leq \epsilon$.

Most functions don't have their Fourier spectrum concentrated on a small collection (see Exercise 3.35). But for those that do, we may hope to discover "most of" their Fourier coefficients. The main result of this section is a kind of "meta-algorithm" for learning an unknown target f. It reduces the problem of learning f to the problem of identifying a collection of characters on which f's Fourier spectrum is concentrated.

Theorem 3.29. *Assume learning algorithm A has (at least) random example access to target $f : \{-1,1\}^n \to \{-1,1\}$. Suppose that A can – somehow – identify a* *collection $\mathscr{F}$ of subsets on which f's Fourier spectrum is $\epsilon/2$-concentrated. Then using* $\mathrm{poly}(|\mathscr{F}|, n, 1/\epsilon)$ *additional time, A can with high probability output a hypothesis h that is ϵ-close to f.*

The idea of the theorem is that A will estimate all of f's Fourier coefficients in $\mathscr{F}$, obtaining a good approximation to f's Fourier expansion. Then A's hypothesis will be the *sign* of this approximate Fourier expansion.

The first tool we need to prove Theorem 3.29 is the ability to accurately estimate any fixed Fourier coefficient:

Proposition 3.30. *Given access to random examples from $f : \{-1,1\}^n \to \{-1,1\}$, there is a randomized algorithm which takes as input $S \subseteq [n]$, $0 < \delta, \epsilon \leq 1/2$, and outputs an estimate $\widetilde{f}(S)$ for $\widehat{f}(S)$ that satisfies*

$$|\widetilde{f}(S) - \widehat{f}(S)| \leq \epsilon$$

except with probability at most δ. The running time is $\mathrm{poly}(n, 1/\epsilon) \cdot \log(1/\delta)$.

Proof. We have $\widehat{f}(S) = \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})]$. Given random examples $(\boldsymbol{x}, f(\boldsymbol{x}))$, the algorithm can compute $f(\boldsymbol{x})\chi_S(\boldsymbol{x}) \in \{-1,1\}$ and therefore empirically estimate $\mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})]$. A standard application of the Chernoff bound implies that $O(\log(1/\delta)/\epsilon^2)$ examples are sufficient to obtain an estimate within $\pm\epsilon$ with probability at least $1-\delta$. □

The second observation we need to prove Theorem 3.29 is the following:

Proposition 3.31. *Suppose that $f : \{-1,1\}^n \to \{-1,1\}$ and $g : \{-1,1\}^n \to \mathbb{R}$ satisfy $\|f-g\|_2^2 \leq \epsilon$. Let $h : \{-1,1\}^n \to \{-1,1\}$ be defined by $h(x) = \mathrm{sgn}(g(x))$, with $\mathrm{sgn}(0)$ chosen arbitrarily from $\{-1,1\}$. Then $\mathrm{dist}(f,h) \leq \epsilon$.*

Proof. Since $|f(x)-g(x)|^2 \geq 1$ whenever $f(x) \neq \mathrm{sgn}(g(x))$, we conclude

$$\mathrm{dist}(f,h) = \Pr_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq h(\boldsymbol{x})] = \mathop{\mathbf{E}}_{\boldsymbol{x}}[\mathbf{1}_{f(\boldsymbol{x})\neq \mathrm{sgn}(g(\boldsymbol{x}))}] \leq \mathop{\mathbf{E}}_{\boldsymbol{x}}[|f(\boldsymbol{x})-g(\boldsymbol{x})|^2] = \|f-g\|_2^2. \quad \square$$

(See Exercise 3.34 for an improvement to this argument.)

We can now prove Theorem 3.29:

Proof of Theorem 3.29. For each $S \in \mathscr{F}$ the algorithm uses Proposition 3.30 to produce an estimate $\widetilde{f}(S)$ for $\widehat{f}(S)$ which satisfies $|\widetilde{f}(S) - \widehat{f}(S)| \leq \sqrt{\epsilon}/(2\sqrt{|\mathscr{F}|})$ except with probability at most $1/(10|\mathscr{F}|)$. Overall this requires $\mathrm{poly}(|\mathscr{F}|, n, 1/\epsilon)$ time, and by the union bound, except with probability at most $1/10$ all $|\mathscr{F}|$ estimates have the desired accuracy. Finally, A forms the real-valued function $g = \sum_{S\in\mathscr{F}} \widetilde{f}(S)\chi_S$ and outputs hypothesis $h = \mathrm{sgn}(g)$. By Proposition 3.31, it suffices to show that $\|f - g\|_2^2 \leq \epsilon$. And indeed,

$$\begin{aligned}
\|f-g\|_2^2 &= \sum_{S\subseteq[n]} \widehat{f-g}(S)^2 && \text{(Parseval)}\\
&= \sum_{S\in\mathscr{F}} (\widehat{f}(S) - \widetilde{f}(S))^2 + \sum_{S\notin\mathscr{F}} \widehat{f}(S)^2\\
&\leq \sum_{S\in\mathscr{F}} \left(\frac{\sqrt{\epsilon}}{2\sqrt{|\mathscr{F}|}}\right)^2 + \epsilon/2 && \text{(estimates, concentration assumption)}\\
&= \epsilon/4 + \epsilon/2 \quad \leq \quad \epsilon,
\end{aligned}$$

as desired. □

As we described, Theorem 3.29 reduces the algorithmic task of learning f to the algorithmic task of identifying a collection $\mathscr{F}$ on which f's Fourier spectrum is concentrated. In Section 3.5 we will describe the Goldreich–Levin algorithm, a sophisticated way to find such an $\mathscr{F}$ assuming query access to f. For now, though, we observe that for several interesting concept classes we don't need to do any algorithmic searching for $\mathscr{F}$; we can just take $\mathscr{F}$ to be all sets of small cardinality. This works whenever all functions in $\mathscr{C}$ have low-degree spectral concentration.

The "Low-Degree Algorithm". *Let $k \geq 1$ and let $\mathscr{C}$ be a concept class for which every function $f : \{-1,1\}^n \to \{-1,1\}$ in $\mathscr{C}$ is $\epsilon/2$-concentrated up to degree k. Then $\mathscr{C}$ can be learned from random examples only with error ϵ in time* $\mathrm{poly}(n^k, 1/\epsilon)$.

Proof. Apply Theorem 3.29 with $\mathscr{F} = \{S \subseteq [n] : |S| \leq k\}$. We have $|\mathscr{F}| = \sum_{j=0}^{k} \binom{n}{j} \leq O(n^k)$. □

The Low-Degree Algorithm reduces the *algorithmic* problem of learning $\mathscr{C}$ from random examples to the *analytic* task of showing low-degree spectral concentration for the functions in $\mathscr{C}$. Using the results of Section 3.1 we can quickly obtain some learning-theoretic results. For example:

Corollary 3.32. *For $t \geq 1$, let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathbf{I}[f] \leq t\}$. Then $\mathscr{C}$ is learnable from random examples with error ϵ in time $n^{O(t/\epsilon)}$.*

Proof. Use the Low-Degree Algorithm with $k = 2t/\epsilon$; the result follows from Proposition 3.2. □

Corollary 3.33. *Let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid f$ is monotone$\}$. Then $\mathscr{C}$ is learnable from random examples with error ϵ in time $n^{O(\sqrt{n}/\epsilon)}$.*

Proof. Follows from the previous corollary and Theorem 2.33. □

You might be concerned that a running time such as $n^{O(\sqrt{n})}$ does not seem very efficient. Still, it's much better than the trivial running time of $\widetilde{O}(2^n)$. Further, as we will see in the next section, learning algorithms are sometimes used in attacks on cryptographic schemes, and in this context even subexponential-time algorithms are considered dangerous.

Continuing with applications of the Low-Degree Algorithm:

Corollary 3.34. *For $\delta \in (0, 1/2]$, let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathbf{NS}_\delta[f] \leq \epsilon/6\}$. Then $\mathscr{C}$ is learnable from random examples with error ϵ in time $\mathrm{poly}(n^{1/\delta}, 1/\epsilon)$.*

Proof. Follows from Proposition 3.3. □

Corollary 3.35. *Let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathrm{DT}_{\mathrm{size}}(f) \leq s\}$. Then $\mathscr{C}$ is learnable from random examples with error ϵ in time $n^{O(\log(s/\epsilon))}$.*

Proof. Follows from Proposition 3.17. □

With a slight extra twist one can also *exactly* learn the class of degree-k functions in time $\mathrm{poly}(n^k)$; see Exercise 3.36:

Theorem 3.36. *Let $k \geq 1$ and let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \deg(f) \leq k\}$ (e.g., $\mathscr{C}$ contains all depth-k decision trees). Then $\mathscr{C}$ is learnable from random examples with error 0 in time $n^k \cdot \mathrm{poly}(n, 2^k)$.*

3.5. Highlight: the Goldreich–Levin Algorithm

We close this chapter by briefly describing a topic which is in some sense the "opposite" of learning theory: *cryptography*. At the highest level, cryptography is concerned with constructing functions which are computationally easy to compute but computationally difficult to invert. Intuitively, think about the task of encrypting secret messages: You would like a scheme where it's easy to take any message x and produce an encrypted version $e(x)$, but where it's hard for an adversary to compute x given $e(x)$. Indeed, even with examples $e(x^{(1)}), \dots, e(x^{(m)})$ of several encryptions, it should be hard for an adversary to learn anything about the encrypted messages, or to predict ("forge") the encryption of future messages.

A basic task in cryptography is building stronger cryptographic functions from weaker ones. Often the first example in "Cryptography 101" is the *Goldreich–Levin Theorem*, which is used to build a "pseudorandom generator" from a "one-way permutation". We sketch the meaning of these terms and the analysis of the construction in Exercise 3.45; for now, suffice it to say that the key to the analysis of Goldreich and Levin's construction is a *learning algorithm*. Specifically, the Goldreich–Levin learning algorithm solves the following problem: Given *query* access to a target function $f : \mathbb{F}_2^n \to \mathbb{F}_2$, find all of the linear functions (in the sense of Chapter 1.6) with which f is at least slightly correlated. Equivalently, find all of the noticeably large Fourier coefficients of f.

Goldreich–Levin Theorem. *Given query access to a target $f : \{-1,1\}^n \to \{-1,1\}$ as well as input $0 < \tau \leq 1$, there is a* $\mathrm{poly}(n, 1/\tau)$*-time algorithm that with high probability outputs a list $L = \{U_1, \dots, U_\ell\}$ of subsets of $[n]$ such that:*

- $|\widehat{f}(U)| \geq \tau \implies U \in L$;
- $U \in L \implies |\widehat{f}(U)| \geq \tau/2$.

(By Parseval's Theorem, the second guarantee implies that $|L| \leq 4/\tau^2$.)

Although the Goldreich–Levin Theorem was originally developed for cryptography, it was soon put to use for learning theory. Recall that the "meta-algorithm" of Theorem 3.29 reduces learning an unknown target $f : \{-1,1\}^n \to \{-1,1\}$ to identifying a collection $\mathscr{F}$ of sets on which f's Fourier spectrum is $\epsilon/2$-concentrated. Using the Goldreich–Levin Algorithm, a learner with query access to f can "collect up" its largest Fourier coefficients until only $\epsilon/2$ Fourier weight remains unfound. This strategy straightforwardly yields the following result (see Exercise 3.39):

Theorem 3.37. *Let $\mathscr{C}$ be a concept class such that every $f : \{-1,1\}^n \to \{-1,1\}$ in $\mathscr{C}$ has its Fourier spectrum $\epsilon/4$-concentrated on a collection of at most M sets. Then $\mathscr{C}$ can be learned using queries with error ϵ in time* $\mathrm{poly}(M, n, 1/\epsilon)$.

The algorithm of Theorem 3.37 is often called the *Kushilevitz–Mansour Algorithm*. Much like the Low-Degree Algorithm, it reduces the computational problem of learning $\mathscr{C}$ (using queries) to the analytic problem of proving that the functions in $\mathscr{C}$ have concentrated Fourier spectra. The advantage of the Kushilevitz–Mansour Algorithm is that it works so long as the Fourier spectrum of f is concentrated on *some* small collection of sets; the Low-Degree Algorithm requires that the concentration specifically be on the low-degree characters. The disadvantage of the Kushilevitz–Mansour Algorithm is that it requires query access to f, rather than just random examples. An example concept class for which the Kushilevitz–Mansour Algorithm works well is the set of all f for which $\hat{\lVert} f \hat{\rVert}_1$ is not too large:

Theorem 3.38. *Let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \hat{\lVert} f \hat{\rVert}_1 \leq s\}$ (e.g., $\mathscr{C}$ contains any f computable by a decision tree of size at most s). Then $\mathscr{C}$ is learnable from queries with error ϵ in time* $\mathrm{poly}(n, s, 1/\epsilon)$.

This is proved in Exercise 3.38.

Let's now return to the Goldreich–Levin Algorithm itself, which seeks the Fourier coefficients $\widehat{f}(U)$ with magnitude at least τ. Given any candidate $U \subseteq [n]$, Proposition 3.30 lets us easily distinguish whether the associated coefficient is large, $|\widehat{f}(U)| \geq \tau$, or small, $|\widehat{f}(U)| \leq \tau/2$. The trouble is that there are 2^n potential candidates. The Goldreich–Levin Algorithm overcomes this difficulty using a divide-and-conquer strategy that measures the Fourier weight of f on various collections of sets. Let's make a definition:

Definition 3.39. Let $f : \{-1,1\}^n \to \mathbb{R}$ and $S \subseteq J \subseteq [n]$. We write

$$\mathbf{W}^{S|\overline{J}}[f] = \sum_{T \subseteq \overline{J}} \widehat{f}(S \cup T)^2$$

for the Fourier weight of f on sets whose restriction to J is S.

The crucial tool for the Goldreich–Levin Algorithm is Corollary 3.22, which says that

$$\mathbf{W}^{S|\overline{J}}[f] = \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}} [\widehat{f_{J|\boldsymbol{z}}}(S)^2]. \tag{3.5}$$

This identity lets a learning algorithm with query access to f efficiently estimate any $\mathbf{W}^{S|\overline{J}}[f]$ of its choosing. Intuitively, query access to f allows query access to $f_{J|z}$ for any $z \in \{-1,1\}^{\overline{J}}$; with this one can estimate any $\widehat{f_{J|z}}(S)$ and hence (3.5). More precisely:

Proposition 3.40. *For any $S \subseteq J \subseteq [n]$ an algorithm with query access to $f : \{-1,1\}^n \to \{-1,1\}$ can compute an estimate of $\mathbf{W}^{S|\overline{J}}[f]$ that is accurate to within $\pm\epsilon$ (except with probability at most δ) in time* $\mathrm{poly}(n, 1/\epsilon) \cdot \log(1/\delta)$.

Proof. From (3.5),

$$\begin{aligned}\mathbf{W}^{S|\overline{J}}[f] = \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}} [\widehat{f_{J|\boldsymbol{z}}}(S)^2] &= \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}} \left[\mathop{\mathbf{E}}_{\boldsymbol{y} \sim \{-1,1\}^{J}} [f(\boldsymbol{y},\boldsymbol{z})\chi_S(\boldsymbol{y})]^2 \right] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}} \mathop{\mathbf{E}}_{\boldsymbol{y},\boldsymbol{y}' \sim \{-1,1\}^{J}} [f(\boldsymbol{y},\boldsymbol{z})\chi_S(\boldsymbol{y}) \cdot f(\boldsymbol{y}',\boldsymbol{z})\chi_S(\boldsymbol{y}')],\end{aligned}$$

where $\boldsymbol{y}$ and $\boldsymbol{y}'$ are independent. As in Proposition 3.30, $f(\boldsymbol{y},\boldsymbol{z})\chi_S(\boldsymbol{y}) \cdot f(\boldsymbol{y}',\boldsymbol{z})\chi_S(\boldsymbol{y}')$ is a ±1-valued random variable that the algorithm can sample from using queries to f. A Chernoff bound implies that $O(\log(1/\delta)/\epsilon^2)$ samples are sufficient to estimate its mean with accuracy ϵ and confidence $1-\delta$. □

We're now ready to prove the Goldreich–Levin Theorem.

Proof of the Goldreich–Levin Theorem. We begin with an overview of how the algorithm works. Initially, all 2^n sets U are (implicitly) put in a single "bucket". The algorithm then repeats the following loop:

- Select any bucket $\mathscr{B}$ containing 2^m sets, $m \geq 1$.

- Split it into two buckets $\mathscr{B}_1$, $\mathscr{B}_2$ of 2^{m-1} sets each.
- "Weigh" each $\mathscr{B}_i$, $i = 1,2$; i.e., estimate $\sum_{U \in \mathscr{B}_i} \widehat{f}(U)^2$.
- Discard $\mathscr{B}_1$ or $\mathscr{B}_2$ if its weight estimate is at most $\tau^2/2$.

The algorithm stops once all buckets contain just 1 set; it then outputs the list of these sets.

We now fill in the details. First we argue the correctness of the algorithm, assuming all weight estimates are accurate (this assumption is removed later). On one hand, any set U with $|\widehat{f}(U)| \geq \tau$ will never be discarded, since it always contributes weight at least $\tau^2 \geq \tau^2/2$ to the bucket it's in. On the other hand, no set U with $|\widehat{f}(U)| \leq \tau/2$ can end up in a singleton bucket because such a bucket, when created, would have weight only $\tau^2/4 \leq \tau^2/2$ and thus be discarded. Notice that this correctness proof does not rely on the weight estimates being exact; it suffices for them to be accurate to within $\pm\tau^2/4$.

The next detail concerns running time. Note that any "active" (undiscarded) bucket has weight at least $\tau^2/4$, even assuming the weight estimates are only accurate to within $\pm\tau^2/4$. Therefore Parseval tells us there can only ever be at most $4/\tau^2$ active buckets. Since a bucket can be split only n times, it follows that the algorithm repeats its main loop at most $4n/\tau^2$ times. Thus as long as the buckets can be maintained and accurately weighed in $\mathrm{poly}(n, 1/\tau)$ time, the overall running time will be $\mathrm{poly}(n, 1/\tau)$ as claimed.

Finally, we describe the bucketing system. The buckets are indexed (and thus maintained implicitly) by an integer $0 \leq k \leq n$ and a subset $S \subseteq [k]$. The bucket $\mathscr{B}_{k,S}$ is defined by

$$\mathscr{B}_{k,S} = \Big\{ S \cup T : T \subseteq \{k+1, k+2, \dots, n\} \Big\}.$$

Note that $|\mathscr{B}_{k,S}| = 2^{n-k}$. The initial bucket is $\mathscr{B}_{0,\emptyset}$. The algorithm always splits a bucket $\mathscr{B}_{k,S}$ into the two buckets $\mathscr{B}_{k+1,S}$ and $\mathscr{B}_{k+1,S\cup\{k+1\}}$. The final singleton buckets are of the form $\mathscr{B}_{n,S} = \{S\}$. Finally, the weight of bucket $\mathscr{B}_{k,S}$ is precisely $\mathbf{W}^{S|\{k+1,\dots,n\}}[f]$. Thus it can be estimated to accuracy $\pm\tau^2/4$ with confidence $1-\delta$ in time $\mathrm{poly}(n, 1/\tau) \cdot \log(1/\delta)$ using Proposition 3.40. Since the main loop is executed at most $4n/\tau^2$ times, the algorithm overall needs to make at most $8n/\tau^2$ weighings; by setting $\delta = \tau^2/(80n)$ we ensure that *all* weighings are accurate with high probability (at least 9/10). The overall running time is therefore indeed $\mathrm{poly}(n, 1/\tau)$. □

3.6. Exercises and notes

3.1 Let $M : \mathbb{F}_2^n \to \mathbb{F}_2^n$ be an invertible linear transformation. Given $f : \mathbb{F}_2^n \to \mathbb{R}$, let $f \circ M : \mathbb{F}_2^n \to \mathbb{R}$ be defined by $f \circ M(x) = f(Mx)$. Show that $\widehat{f \circ M}(\gamma) =$

$\widehat{f}(M^{-\top}\gamma)$. What if M is an invertible *affine* transformation? What if M is not invertible?

3.2 Show that $\frac{2}{1-e^{-2}}$ is smallest constant (not depending on δ or n) that can be taken in Proposition 3.3.

3.3 Generalize Proposition 3.3 by showing that any $f : \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on degree up to $1/\delta$ for $\epsilon = (\mathbf{E}[f^2] - \mathbf{Stab}_{1-\delta}[f])/(1-1/e)$.

3.4 Prove Lemma 3.5 by induction on n. (Hint: If one of the subfunctions $f(x_1,\dots,x_n,\pm1)$ is identically 0, show that the other has degree at most $k-1$.)

3.5 Verify for all $p \in [1,\infty]$ that $\hat{\|}\cdot\hat{\|}_p$ is a norm on the vector space of functions $f : \mathbb{F}_2^n \to \mathbb{R}$.

3.6 Show that $\hat{\|} fg \hat{\|}_1 \leq \hat{\|} f \hat{\|}_1 \hat{\|} g \hat{\|}_1$ for all $f, g : \mathbb{F}_2^n \to \mathbb{R}$.

3.7 Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $J \subseteq [n]$, $z \in \{-1,1\}^{\overline{J}}$.

(*a*) Show that restriction reduces spectral 1-norm: $\hat{\|} f_{J|z} \hat{\|}_1 \leq \hat{\|} f \hat{\|}_1$.

(*b*) Show that it also reduces Fourier sparsity: $\mathrm{sparsity}(\widehat{f_{J|z}}) \leq \mathrm{sparsity}(\widehat{f})$.

3.8 Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $0 < p \leq q \leq \infty$. Show that $\hat{\|} f \hat{\|}_p \geq \hat{\|} f \hat{\|}_q$. (Cf. Exercise 1.13.)

3.9 Let $f : \{-1,1\}^n \to \mathbb{R}$. Show that $\hat{\|} f \hat{\|}_\infty \leq \|f\|_1$ and $\|f\|_\infty \leq \hat{\|} f \hat{\|}_1$. (These are easy special cases of the *Hausdorff–Young Inequality*.)

3.10 Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is monotone. Show that $|\widehat{f}(S)| \leq \widehat{f}(i)$ whenever $i \in S \subseteq [n]$. Deduce that $\hat{\|} f \hat{\|}_\infty = \max_S\{|\widehat{f}(S)|\}$ is achieved by an S of cardinality 0 or 1. (Hint: Apply the previous exercise to f's derivatives.)

3.11 Prove Proposition 3.12.

3.12 Verify Parseval's Theorem for the Fourier expansion of subspaces given in Proposition 3.11.

3.13 Let $f : \mathbb{F}_2^n \to \{0,1\}$ be the indicator of $A \subseteq \mathbb{F}_2^n$. We know that $\hat{\|} f \hat{\|}_1 = 1$ if A is an affine subspace. So assume that A is *not* an affine subspace.

(*a*) Show that there exists an affine subspace B of dimension 2 on which f takes the value 1 exactly 3 times.

(*b*) Let b be the point in B where f is 0 and let $\psi = \varphi_B - (1/2)\varphi_b$. Show that $\hat{\|} \psi \hat{\|}_\infty = 1/2$.

(*c*) Show that $\langle \psi, f \rangle = 3/4$ and deduce $\hat{\|} f \hat{\|}_1 \geq 3/2$.

3.14 Suppose $f : \{-1,1\}^n \to \mathbb{R}$ satisfies $\mathbf{E}[f^2] \leq 1$. Show that $\hat{\|} f \hat{\|}_1 \leq 2^{n/2}$, and show that for any even n the upper bound can be achieved by a function $f : \{-1,1\}^n \to \{-1,1\}$.

3.15 Given $f : \mathbb{F}_2^n \to \mathbb{R}$, define its *(fractional) sparsity* to be $\mathrm{sparsity}(f) = |\mathrm{supp}(f)|/2^n = \mathbf{Pr}_{\boldsymbol{x} \in \mathbb{F}_2^n}[f(\boldsymbol{x}) \neq 0]$. In this exercise you will prove the *uncertainty principle*: If f is nonzero, then $\mathrm{sparsity}(f) \cdot \mathrm{sparsity}(\widehat{f}) \geq 1$.:

(*a*) Show that we may assume $\|f\|_1 = 1$.
(*b*) Suppose $\mathscr{F} = \{\gamma : \widehat{f}(\gamma) \neq 0\}$. Show that $\hat{\|} f \hat{\|}_2^2 \leq |\mathscr{F}|$.
(*c*) Suppose $\mathscr{G} = \{x : f(x) \neq 0\}$. Show that $\|f\|_2^2 \geq 2^n/|\mathscr{G}|$, and deduce the uncertainty principle.
(*d*) Identify all cases of equality.

3.16 Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $\epsilon > 0$. Show that f is ϵ-concentrated on a collection $\mathscr{F} \subseteq 2^{[n]}$ with $|\mathscr{F}| \leq \hat{\|} f \hat{\|}_1^2/\epsilon$.

3.17 Suppose the Fourier spectrum of $f : \{-1,1\}^n \to \mathbb{R}$ is ϵ_1-concentrated on $\mathscr{F}$ and that $g : \{-1,1\}^n \to \mathbb{R}$ satisfies $\|f - g\|_2^2 \leq \epsilon_2$. Show that the Fourier spectrum of g is $2(\epsilon_1 + \epsilon_2)$-concentrated on $\mathscr{F}$.

3.18 Show that every function $f : \mathbb{F}_2^n \to \mathbb{R}$ is computed by a decision tree with depth at most n and size at most 2^n.

3.19 Let $f : \mathbb{F}_2^n \to \mathbb{R}$ be computable by a decision tree of size s and depth k Show that $-f$ and the Boolean dual $f^\dagger$ are also computable by decision trees of size s and depth k.

3.20 For each function in Exercise 1.1 with 4 or fewer inputs, give a decision tree computing it. Try primarily to use the least possible depth, and secondarily to use the least possible size.

3.21 Prove Proposition 3.16.

3.22 Let $f : \mathbb{F}_2^n \to \{-1,1\}$ be computed by a decision tree T of size s and let $\epsilon \in (0,1]$. Suppose each path in T is truncated (if necessary) so that its length does not exceed $\log(s/\epsilon)$; new leaves with labels -1 and 1 may be created in an arbitrary way as necessary. Show that the resulting decisions tree T' computes a function that is ϵ-close to f. Deduce Proposition 3.17.

3.23 A *decision list* is a decision tree in which every internal node has an outgoing edge to at least one leaf. Show that any function computable by a decision list is a linear threshold function.

3.24 A *read-once* decision tree is one in which every internal node queries a distinct variable. Bearing this in mind, show that the bound $k2^{k-1}$ in Theorem 3.4 cannot be reduced below $2^k - 1$.

3.25 Suppose that f is computed by a read-once decision tree in which every root-to-leaf path has length k and every internal node at the deepest level has one child (leaf) labeled -1 one one child labeled 1. Compute the influence of each coordinate on f, and compute $\mathbf{I}[f]$.

3.26 The following are generalizations of decision trees:

Subcube partition: This is defined by a collection $C_1, \dots, C_s$ of subcubes that form a partition of $\mathbb{F}_2^n$, along with values $b_1, \dots, b_s \in \mathbb{R}$. It computes the function $f : \mathbb{F}_2^n \to \mathbb{R}$ which has value b_i on all inputs in C_i.

The subcube partition's size is s and its "codimension" k (analogous to depth) is the maximum codimension of the cubes C_i.

Parity decision tree: This is similar to a decision tree except that the internal nodes are labeled by vectors $\gamma \in \mathbb{F}_2^n$. At such a node the computation path on input x follows the edge labeled $\gamma \cdot x$. We insist that for each root-to-leaf path, the vectors appearing in its internal nodes are linearly independent. Size s and depth k are defined as with normal decision trees.

Affine subspace partition: This is similar to a subcube partition except the subcubes may be C_i may be arbitrary affine subspaces.

(a) Show that subcube partition size/codimension and parity decision tree size/depth generalize normal decision tree size/depth, and are generalized by affine subspace partition size/codimension.

(b) Show that Proposition 3.16 holds also for the generalizations, except that the statement about degree need not hold for parity decision trees and affine subspace partitions.

(c) Show that the class of functions with affine subspace partition size at most s is learnable from queries with error ϵ in time $\mathrm{poly}(n, s, 1/\epsilon)$.

3.27 Define $\mathrm{Equ}_3 : \{-1,1\}^3 \to \{-1,1\}$ by $\mathrm{Equ}_3(x) = -1$ if and only if $x_1 = x_2 = x_3$.

(a) Show that $\deg(\mathrm{Equ}_3) = 2$.

(b) Show that $\mathrm{DT}(\mathrm{Equ}_3) = 3$.

(c) Show that Equ_3 is computable by a parity decision tree of codimension 2.

(d) For $d \in \mathbb{N}$, define $f\{-1,1\}^{3^d} \to \{-1,1\}$ by $f = \mathrm{Equ}_3^{\otimes d}$ (using the notation from Definition 2.6). Show that $\deg(f) = 2^d$ but $\mathrm{DT}(f) = 3^d$.

3.28 Let $f : \{-1,1\}^n \to \mathbb{R}$ and $J \subseteq [n]$. Define $f^{\subseteq J} : \{-1,1\}^n \to \mathbb{R}$ by $f(x) = \mathbf{E}_{\boldsymbol{y} \sim \{-1,1\}^{\overline{J}}}[f(x_J, \boldsymbol{y})]$, where $x_J \in \{-1,1\}^J$ is the projection of x to coordinates J. Verify the Fourier expansion

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

3.29 Let $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ be a probability density function corresponding to probability distribution ϕ on $\mathbb{F}_2^n$. Let $J \subseteq [n]$.

(a) Consider the marginal probability distribution of ϕ on coordinates J. What is its probability density function (a function $\mathbb{F}_2^J \to \mathbb{R}^{\geq 0}$) in terms of φ?

(b) Consider the probability distribution of ϕ conditioned on a substring $z \in \mathbb{F}_2^{\overline{J}}$. Assuming it's well defined, what is its probability density function in terms of φ?

3.30 Suppose $f : \{-1,1\}^n \to \mathbb{R}$ is computable by a decision tree that has a leaf at depth k labeled b. Show that $\hat{\lVert} f \hat{\rVert}_\infty \geq |b|/2^k$. (Hint: You may find Exercise 3.28 helpful.)

3.31 Prove Fact 3.25 by using Theorem 1.27 and Exercise 1.1(*d*).

3.32 (*a*) Suppose $f : \mathbb{F}_2^n \to \mathbb{R}$ has sparsity$(\widehat{f}) < 2^n$. Show that for any $\gamma \in$ supp$(\widehat{f})$ there exists nonzero $\beta \in \widehat{\mathbb{F}_2^n}$ such that $f_{\beta^\perp}$ has $\widehat{f}(\gamma)$ as a Fourier coefficient.

(*b*) Prove by induction on n that if $f : \mathbb{F}_2^n \to \{-1,1\}$ has sparsity$(\widehat{f}) = s > 1$ then $\widehat{f}$ is $2^{1-\lfloor \log s \rfloor}$-granular. (Hint: Distinguish the cases $s = 2^n$ and $s < 2^n$. In the latter case use part (*a*).)

(*c*) Prove that there are no functions $f : \{-1,1\}^n \to \{-1,1\}$ with sparsity$(\widehat{f}) \in \{2,3,5,6,7,9\}$.

3.33 Show that one can learn *any* target $f : \{-1,1\}^n \to \{-1,1\}$ with error 0 from random examples only in time $\widetilde{O}(2^n)$.

3.34 Improve Proposition 3.31 as follows. Suppose $f : \{-1,1\}^n \to \{-1,1\}$ and $g : \{-1,1\}^n \to \mathbb{R}$ satisfy $\|f - g\|_1 \le \epsilon$. Pick $\boldsymbol{\theta} \in [-1,1]$ uniformly at random and define $\boldsymbol{h} : \{-1,1\}^n \to \{-1,1\}$ by $\boldsymbol{h}(x) = \mathrm{sgn}(g(x) - \boldsymbol{\theta})$. Show that $\mathbf{E}[\mathrm{dist}(f, \boldsymbol{h})] \le \epsilon/2$.

3.35 (*a*) For n even, find a function $f : \{-1,1\}^n \to \{-1,1\}$ that is not 1/2-concentrated on any $\mathscr{F} \subseteq 2^{[n]}$ with $|\mathscr{F}| < 2^{n-1}$. (Hint: Exercise 1.1.)

(*b*) Let $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ be a random function as in Exercise 1.7. Show that with probability at least 1/2, $\boldsymbol{f}$ is not 1/4-concentrated on degree up to $\lfloor n/2 \rfloor$.

3.36 Prove Theorem 3.36. (Hint: In light of Exercise 1.11 you may round off certain estimates with confidence.)

3.37 Show that each of the following classes $\mathscr{C}$ (ordered by inclusion) can be learned exactly (i.e., with error 0) using queries in time poly$(n, 2^k)$:

(*a*) $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid f$ is a k-junta$\}$. (Hint: Estimate influences.)

(*b*) $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathrm{DT}(f) \le k\}$.

(*c*) $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathrm{sparsity}(\widehat{f}) \le 2^{O(k)}\}$. (Hint: Exercise 3.32.)

3.38 Prove Theorem 3.38. (Hint: Exercise 3.16.)

3.39 Deduce Theorem 3.37 from the Goldreich–Levin Algorithm.

3.40 Suppose A learns $\mathscr{C}$ from random examples with error $\epsilon/2$ in time T – with probability at least 9/10.

(*a*) After producing hypothesis h on target $f : \{-1,1\}^n \to \{-1,1\}$, show that A can "check" whether h is a good hypothesis in time poly$(n, T, 1/\epsilon) \cdot \log(1/\delta)$. Specifically, except with probability at most δ, A should output 'YES' if $\mathrm{dist}(f,h) \le \epsilon/2$ and 'NO' if $\mathrm{dist}(f,h) > \epsilon$. (Hint: Time poly$(T)$ may be required for A to evaluate $h(x)$.)

(*b*) Show that for any $\delta \in (0, 1/2]$, there is a learning algorithm that learns $\mathscr{C}$ with error ϵ in time poly$(n, T, \epsilon) \cdot \log(1/\delta)$ – with probability at least $1 - \delta$.

3.41 (*a*) Our description of the Low-Degree Algorithm with degree k and error ϵ involved using a new batch of random examples to estimate each low-degree Fourier coefficient. Show that one can instead simply draw a single batch $\mathscr{E}$ of $\mathrm{poly}(n^k, 1/\epsilon)$ examples and use $\mathscr{E}$ to estimate each of the low-degree coefficients.

(*b*) Show that when using the above form of the Low-Degree Algorithm, the final hypothesis $h : \{-1,1\}^n \to \{-1,1\}$ is of the form

$$h(y) = \mathrm{sgn}\left(\sum_{(x,f(x))\in\mathscr{E}} w(\Delta(y,x)) \cdot f(x)\right),$$

for some function $w : \{0,1,\dots,n\} \to \mathbb{R}$. In other words, the hypothesis on a given y is equal to a weighted vote over all examples seen, where an example's weight depends only on its Hamming distance to y. Simplify your expression for w as much as you can.

3.42 Extend the Goldreich–Levin Algorithm so that it works also for functions $f : \{-1,1\}^n \to [-1,1]$. (The learning model for targets $f : \{-1,1\}^n \to [-1,1]$ assumes that $f(x)$ is always a rational number expressible by $\mathrm{poly}(n)$ bits.)

3.43 (*a*) Assume $\gamma, \gamma' \in \widehat{\mathbb{F}_2^n}$ are distinct. Show that $\mathbf{Pr}_{\boldsymbol{x}}[\gamma \cdot \boldsymbol{x} = \gamma' \cdot \boldsymbol{x}] = 1/2$.

(*b*) Fix $\gamma \in \widehat{\mathbb{F}_2^n}$ and suppose $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)} \sim \mathbb{F}_2^n$ are drawn uniformly and independently. Show that if $m = Cn$ for C a sufficiently large constant then with high probability, the only $\gamma' \in \widehat{\mathbb{F}_2^n}$ satisfying $\gamma' \cdot \boldsymbol{x}^{(i)} = \gamma \cdot \boldsymbol{x}^{(i)}$ for all $i \in [m]$ is $\gamma' = \gamma$.

(*c*) Essentially improve on Exercise 1.27 by showing that the concept class of all linear functions $\mathbb{F}_2^n \to \mathbb{F}_2$ can be learned from random examples only, with error 0, in time $\mathrm{poly}(n)$. (Remark: If $\omega \in \mathbb{R}$ is such that $n \times n$ matrix multiplication can be done in $O(n^\omega)$ time, then the learning algorithm also requires only $O(n^\omega)$ time.)

3.44 Let $\tau \geq 1/2 + \epsilon$ for some constant $\epsilon > 0$. Give an algorithm simpler than Goldreich and Levin's that solves the following problem with high probability: Given query access to $f : \{-1,1\}^n \to \{-1,1\}$, in time $\mathrm{poly}(n, 1/\epsilon)$ find the unique $U \subseteq [n]$ such that $|\widehat{f}(U)| \geq \tau$, assuming it exists. (Hint: Use Proposition 1.31 and Exercise 1.27.)

3.45 Informally: a "one-way permutation" is a bijective function $f : \mathbb{F}_2^n \to \mathbb{F}_2^n$ that is easy to compute on all inputs but hard to invert on more than a negligible fraction of inputs; a "pseudorandom generator" is a function $g : \mathbb{F}_2^k \to \mathbb{F}_2^m$ for $m > k$ whose output on a random input "looks unpredictable" to any efficient algorithm. Goldreich and Levin proposed the following construction of the latter from the former: for $k = 2n$, $m = 2n+1$, define

$$g(r,s) = (r, f(s), r \cdot s),$$

where $r, s \in \mathbb{F}_2^n$. When g's input $(\boldsymbol{r}, \boldsymbol{s})$ is uniformly random, then so is the first $2n$ bits of its output (using the fact that f is a bijection). The key to the analysis is showing that the final bit, $\boldsymbol{r} \cdot \boldsymbol{s}$, is highly unpredictable to efficient algorithms even *given* the first $2n$ bits $(\boldsymbol{r}, f(\boldsymbol{s}))$. This is proved by contradiction.

(a) Suppose that an adversary has a deterministic, efficient algorithm A good at predicting the bit $\boldsymbol{r} \cdot \boldsymbol{s}$:

$$\mathop{\mathbf{Pr}}_{\boldsymbol{r},\boldsymbol{s} \sim \mathbb{F}_2^n}[A(\boldsymbol{r}, f(\boldsymbol{s})) = \boldsymbol{r} \cdot \boldsymbol{s}] \geq \frac{1}{2} + \gamma.$$

Show there exists $B \subseteq \mathbb{F}_2^n$ with $|B|/2^n \geq \frac{1}{2}\gamma$ such that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{r} \sim \mathbb{F}_2^n}[A(\boldsymbol{r}, f(s)) = \boldsymbol{r} \cdot s] \geq \frac{1}{2} + \frac{1}{2}\gamma$$

for all $s \in B$.

(b) Switching to ±1 notation in the output, deduce $\widehat{A_{|f(s)}}(s) \geq \gamma$ for all $s \in B$.

(c) Show that the adversary can efficiently compute s given $f(s)$ (with high probability) for any $s \in B$. If γ is nonnegligible, this contradicts the assumption that f is "one-way". (Hint: Use the Goldreich–Levin Algorithm.)

(d) Deduce the same conclusion even if A is a randomized algorithm.

Notes. The fact that the Fourier characters $\chi_\gamma : \mathbb{F}_2^n \to \{-1, 1\}$ form a group isomorphic to $\mathbb{F}_2^n$ is not a coincidence; the analogous result holds for any finite abelian group and is a special case of the theory of Pontryagin duality in harmonic analysis. We will see further examples of this in Chapter 8.

Regarding spectral structure, Karpovsky [**Kar76**] proposed sparsity$(\widehat{f})$ as a measure of complexity for the function f. Brandman's thesis [**Bra87**] (see also [**BOH90**]) is an early work connecting decision tree and subcube partition complexity to Fourier analysis. The notation introduced for restrictions in Section 3.3 is not standard; unfortunately there is no standard notation. The uncertainty principle from Exercise 3.15 dates back to Matolcsi and Szücs [**MS73**]. The result of Exercise 3.13 is due to Green and Sanders [**GS08**], with inspiration from Saeki [**Sae68**]. The main result of Green and Sanders is the sophisticated theorem that any $f : \mathbb{F}_2^n \to \{0, 1\}$ with $\hat{\|} f \hat{\|}_1 \leq s$ can be expressed as $\sum_{i=1}^{L} \pm 1_{H_i}$, where $L \leq 2^{2^{\mathrm{poly}(s)}}$ and each $H_i \leq \mathbb{F}_2^n$.

Theorem 3.4 is due to Nisan and Szegedy [**NS94**]. That work also showed a nontrivial kind of converse to the first statement in Proposition 3.16: Any $f : \{-1, 1\}^n \to \{-1, 1\}$ is computable by a decision tree of depth at most poly(deg(f)). The best upper bound currently known is $\deg(f)^3$ due to Midrijānis [**Mid04**]. Nisan and Szegedy also gave the example in Exercise 3.27 showing the dependence cannot be linear.

The field of computational learning theory was introduced by Valiant in 1984 [**Val84**]; for a good survey with focus on learning under the uniform distribution, see the thesis by Jackson [**Jac95**]. Linial, Mansour, and Nisan [**LMN93**] pioneered the Fourier approach to learning, developing the Low-Degree Algorithm. We present their strong results on constant-depth circuits in Chapter 4. The noise sensitivity approach to the Low-Degree Algorithm is from Klivans, O'Donnell, and Servedio [**KOS04**]. Corollary 3.33 is due to Bshouty and Tamon [**BT96**] who also gave certain matching lower bounds. Goldreich and Levin's work dates from 1989 [**GL89**]. Besides its applications to cryptography and learning, it is important in coding theory and complexity as a *local list-decoding algorithm* for the Hadamard code. The Kushilevitz–Mansour algorithm is from their 1993 paper [**KM93**]; they also are responsible for the results of Exercise 3.37(*b*) and 3.38. The results of Exercise 3.32 and 3.37(*c*) are from Gopalan et al. [**GOS$^+$11**].

Chapter 4

DNF formulas and small-depth circuits

In this chapter we investigate Boolean functions representable by small DNF formulas and constant-depth circuits; these are significant generalizations of decision trees. Besides being natural from a computational point of view, these representation classes are close to the limit of what complexity theorists can "understand" (e.g., prove explicit lower bounds for). One reason for this is that functions in these classes have strong Fourier concentration properties.

4.1. DNF formulas

One of the commonest ways of representing a Boolean function $f : \{0,1\}^n \to \{0,1\}$ is by a DNF formula:

Definition 4.1. A *DNF (disjunctive normal form) formula* over Boolean variables $x_1,\dots,x_n$ is defined to be a logical OR of *terms*, each of which is a logical AND of *literals*. A *literal* is either a variable x_i or its logical negation $\overline{x}_i$. We insist that no term contains both a variable and its negation. The number of literals in a term is called its *width*. We often identify a DNF formula with the Boolean function $f : \{0,1\}^n \to \{0,1\}$ it computes.

Example 4.2. Recall the function Sort_3, defined by $\mathrm{Sort}_3(x_1,x_2,x_3) = 1$ if and only if $x_1 \leq x_2 \leq x_3$ or $x_1 \geq x_2 \geq x_3$. We can represent it by a DNF formula as follows:

$$\mathrm{Sort}_3(x_1,x_2,x_3) \;=\; (x_1 \wedge x_2) \;\vee\; (\overline{x}_2 \wedge \overline{x}_3) \;\vee\; (\overline{x}_1 \wedge x_3).$$

The DNF representation says that the bits are sorted if either the first two bits are 1, or the last two bits are 0, or the first bit is 0 and the last bit is 1.

The complexity of a DNF formula is measured by its size and width:

Definition 4.3. The *size* of a DNF formula is its number of terms. The *width* is the maximum width of its terms. Given $f : \{-1,1\}^n \to \{-1,1\}$ we write $\mathrm{DNF}_{\mathrm{size}}(f)$ (respectively, $\mathrm{DNF}_{\mathrm{width}}(f)$) for the least size (respectively, width) of a DNF formula computing f.

The DNF formula for Sort_3 from Example 4.2 has size 3 and width 2. Every function $f : \{0,1\}^n \to \{0,1\}$ can be computed by a DNF of size at most 2^n and width at most n (Exercise 4.1).

There is also a "dual" notion to DNF formulas:

Definition 4.4. A *CNF (conjunctive normal form) formulas* is a logical AND of *clauses*, each of which is a logical OR of literals. Size and width are defined as for DNFs.

Some functions can be represented much more compactly by CNFs than DNFs (see Exercise 4.14). On the other hand, if we take a CNF computing f and switch its ANDs and ORs, the result is a DNF computing the dual function $f^\dagger$ (see Exercises 1.8 and 4.2). Since f and $f^\dagger$ have essentially the same Fourier expansion, there isn't much difference between CNFs and DNFs when it comes to Fourier analysis. We will therefore focus mainly on DNFs.

DNFs and CNFs are more powerful than decision trees for representing Boolean-valued functions, as the following proposition shows:

Proposition 4.5. *Let $f : \{0,1\}^n \to \{0,1\}$ be computable by a decision tree T of size s and depth k. Then f is computable by a DNF (and also a CNF) of size at most s and width at most k.*

Proof. Take each path in T from the root to a leaf labeled 1 and form the logical AND of the literals describing the path. These are the terms of the required DNF. (For the CNF clauses, take paths to label 0 and negate all literals describing the path.) □

Example 4.6. If we perform this conversion on the decision tree computing Sort_3 in Figure 3.1 we get the DNF

$$(\overline{x}_1 \wedge \overline{x}_3 \wedge \overline{x}_2) \vee (\overline{x}_1 \wedge x_3) \vee (x_1 \wedge \overline{x}_2 \wedge \overline{x}_3) \vee (x_2 \wedge x_3).$$

This has size 4 (indeed at most the decision tree size 6) and width 3 (indeed at most the decision tree depth 3). It is not as simple as the equivalent DNF from Example 4.2, though; DNF representation is not unique.

The class of functions computable by small DNFs is intensively studied in learning theory. This is one reason why the problem of analyzing spectral concentration for DNFs is important. Let's begin with the simplest method for this: understanding low-degree concentration via total influence. We will switch to ±1 notation.

Proposition 4.7. *Suppose that* $f : \{-1,1\}^n \to \{-1,1\}$ *has* $\mathrm{DNF}_{\mathrm{width}}(f) \le w$. *Then* $\mathbf{I}[f] \le 2w$.

Proof. We use Exercise 2.10, which states that

$$\mathbf{I}[f] = 2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\#\ (-1)\text{-pivotal coordinates for } f \text{ on } \boldsymbol{x}],$$

where coordinate i is "(-1)-pivotal" on input x if $f(x) = -1$ (logical True) but $f(x^{\oplus i}) = 1$ (logical False). It thus suffices to show that on *every* input x there are at most w coordinates which are (-1)-pivotal. To have any (-1)-pivotal coordinates at all on x we must have $f(x) = -1$ (True); this means that at least one term T in f's width-w DNF representation must be made True by x. But now if i is a (-1)-pivotal coordinate then either x_i or $\overline{x}_i$ must appear in T; otherwise, T would still be made true by $x^{\oplus i}$. Thus the number of (-1)-pivotal coordinates on x is at most the number of literals in T, which is at most w. □

Since $\mathbf{I}[f^\dagger] = \mathbf{I}[f]$ the proposition is also true for CNFs of width at most w. The proposition is very close to being tight: The parity function $\chi_{[w]} : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{I}[\chi_{[w]}] = w$ and $\mathrm{DNF}_{\mathrm{width}}(\chi_{[w]}) \le w$ (the latter being true for all w-juntas). In fact, the proposition can be improved to give the tight upper bound w (Exercise 4.17).

Using Proposition 3.2 we deduce:

Corollary 4.8. *Let* $f : \{-1,1\}^n \to \{-1,1\}$ *have* $\mathrm{DNF}_{\mathrm{width}}(f) \le w$. *Then for* $\epsilon > 0$, *the Fourier spectrum of f is ϵ-concentrated on degree up to* $2w/\epsilon$.

The dependence here on w is of the correct order (by the example of the parity $\chi_{[w]}$ again), but the dependence on ϵ can be significantly improved as we will see in Section 4.4.

There's usually more interest in DNF *size* than in DNF width; for example, learning theorists are often interested in the class of n-variable DNFs of size poly(n). The following fact (similar to Exercise 3.22) helps relate the two, suggesting $O(\log n)$ as an analogous width bound:

Proposition 4.9. *Let* $f : \{-1,1\}^n \to \{-1,1\}$ *be computable by a DNF (or CNF) of size s and let* $\epsilon \in (0,1]$. *Then f is ϵ-close to a function g computable by a DNF of width* $\log(s/\epsilon)$.

Proof. Take the DNF computing f and delete all terms with more than $\log(s/\epsilon)$ literals; let g be the function computed by the resulting DNF. For any deleted term T, the probability a random input $\boldsymbol{x} \sim \{-1,1\}^n$ makes T true is at most $2^{-\log(s/\epsilon)} = \epsilon/s$. Taking a union bound over the (at most s) such terms shows that $\mathbf{Pr}[g(\boldsymbol{x}) \ne f(\boldsymbol{x})] \le \epsilon$. (A similar proof works for CNFs.) □

By combining Proposition 4.9 and Corollary 4.8 we can deduce (using Exercise 3.17) that DNFs of size s have Fourier spectra ϵ-concentrated up to degree $O(\log(s/\epsilon)/\epsilon)$. Again, the dependence on ϵ will be improved in Section 4.4. We will also later show in Section 4.3 that size-s DNFs have total influence at most $O(\log s)$, something we cannot deduce immediately from Proposition 4.7.

In light of the Kushilevitz–Mansour learning algorithm it would also be nice to show that poly(n)-size DNFs have their Fourier spectra concentrated on small collections (not necessarily low-degree). In Section 4.4 we will show they are ϵ-concentrated on collections of size $n^{O(\log\log n)}$ for any constant $\epsilon > 0$. It has been conjectured that this can be improved to poly(n):

Mansour's Conjecture. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF of size $s > 1$ and let $\epsilon \in (0,1/2]$. Strong conjecture: f's Fourier spectrum is ϵ-concentrated on a collection $\mathscr{F}$ with $|\mathscr{F}| \leq s^{O(\log(1/\epsilon))}$. Weaker conjecture: if $s \leq$ poly(n) and $\epsilon > 0$ is any fixed constant, then we have the bound $|\mathscr{F}| \leq$ poly(n).*

4.2. Tribes

In this section we study the *tribes* DNF formulas, which serve as an important examples and counterexamples in analysis of Boolean functions. Perhaps the most notable feature of the tribes function is that (for a suitable choice of parameters) it is essentially unbiased and yet all of its influences are quite tiny.

Recall from Chapter 2.1 that the function $\mathrm{Tribes}_{w,s} : \{-1,1\}^{sw} \to \{-1,1\}$ is defined by its width-w, size-s DNF representation:

$$\mathrm{Tribes}_{w,s}(x_1,\dots,x_w,\dots,x_{(s-1)w+1},\dots,x_{sw}) \\ = (x_1 \wedge \cdots \wedge x_w) \;\vee\; \cdots \;\vee\; (x_{(s-1)w+1} \wedge \cdots \wedge x_{sw}).$$

(We are using the notation where -1 represents logical True and 1 represents logical False.) As is computed in Exercise 2.13 we have:

Fact 4.10. $\mathbf{Pr}_{\boldsymbol{x}}[\mathrm{Tribes}_{w,s}(\boldsymbol{x}) = -1] = 1-(1-2^{-w})^s$.

The most interesting setting of parameters makes this probability as close to 1/2 as possible (a slightly different choice than the one in Exercise 2.13):

Definition 4.11. For $w \in \mathbb{N}^+$, let $s = s_w$ be the largest integer such that $1-(1-2^{-w})^s \leq 1/2$. Then for $n = n_w = sw$ we define $\mathrm{Tribes}_n : \{-1,1\}^n \to \{-1,1\}$ to be $\mathrm{Tribes}_{w,s}$. Note this is only defined only for certain n: 1, 4, 15, 40, …

Here $s \approx \ln(2)2^w$, hence $n \approx \ln(2)w2^w$ and therefore $w \approx \log n - \log\ln n$ and $s \approx n/\log n$. A slightly more careful accounting (Exercise 4.5) yields:

Proposition 4.12. *For the* Tribes_n *function as in Definition 4.11:*

- *$s = \ln(2)2^w - \Theta_w(1)$;*
- *$n = \ln(2)w2^w - \Theta(w)$, thus $n_{w+1} = (2+o(1))n_w$;*
- *$w = \log n - \log\ln n + o_n(1)$, and $2^w = \frac{n}{\ln n}(1+o_n(1))$;*
- *$\mathbf{Pr}[\mathrm{Tribes}_n(\boldsymbol{x}) = -1] = 1/2 - O\left(\frac{\log n}{n}\right)$.*

Thus with this setting of parameters Tribes_n is essentially unbiased. Regarding its influences:

Proposition 4.13. *$\mathbf{Inf}_i[\mathrm{Tribes}_n] = \frac{\ln n}{n}(1 \pm o(1))$ for each $i \in [n]$ and hence $\mathbf{I}[\mathrm{Tribes}_n] = (\ln n)(1 \pm o(1))$.*

Proof. Thinking of $\mathrm{Tribes}_n = \mathrm{Tribes}_{w,s}$ as a voting rule, voter i is pivotal if and only if: (a) all other voters in i's "tribe" vote -1 (True); (b) all other tribes produce the outcome 1 (False). The probability of this is indeed

$$2^{-(w-1)} \cdot (1-2^{-w})^{s-1} = \tfrac{2}{2^w-1} \cdot \mathbf{Pr}[\mathrm{Tribes}_n = 1] = \tfrac{\ln n}{n}(1 \pm o(1)),$$

where we used Fact 4.10 and then Proposition 4.12. □

Thus if we are interested in (essentially) unbiased voting rules in which every voter has small influence, Tribes_n is a much stronger example than Maj_n where each voter has influence $\Theta(1/\sqrt{n})$. You may wonder if the maximum influence can be even *smaller* than $\Theta\left(\frac{\ln n}{n}\right)$ for unbiased voting rules. Certainly it can't be smaller than $\frac{1}{n}$, since the Poincaré Inequality says that $\mathbf{I}[f] \geq 1$ for unbiased f. In fact the famous KKL Theorem shows that the Tribes_n example is tight up to constants:

Kahn–Kalai–Linial (KKL) Theorem. *For any $f : \{-1,1\}^n \to \{-1,1\}$,*

$$\mathbf{MaxInf}[f] = \max_{i \in [n]}\{\mathbf{Inf}_i[f]\} \geq \mathbf{Var}[f] \cdot \Omega\left(\frac{\log n}{n}\right).$$

We prove the KKL Theorem in Chapter 9.

We conclude this section by recording a formula for the Fourier coefficients of $\mathrm{Tribes}_{w,s}$. The proof is Exercise 4.6.

Proposition 4.14. *Suppose we index the Fourier coefficients of* $\mathrm{Tribes}_{w,s}\{-1,1\}^{sw} \to \{-1,1\}$ *by sets $T = (T_1,\dots,T_s) \subseteq [sw]$, where T_i is the intersection of T with the ith "tribe". Then*

$$\widehat{\mathrm{Tribes}_{w,s}}(T) = \begin{cases} 2(1-2^{-w})^s - 1 & \text{if } T = \emptyset, \\ 2(-1)^{k+|T|}2^{-kw}(1-2^{-w})^{s-k} & \text{if } k = \#\{i : T_i \neq \emptyset\} > 0. \end{cases}$$

4.3. Random restrictions

In this section we describe the method of applying *random restrictions*. This is a very "Fourier-friendly" way of simplifying a Boolean function. As motivation, let's consider the problem of bounding total influence for size-s DNFs. One plan is to use the results from Section 4.1: size-s DNFs are .01-close to width-$O(\log s)$ DNFs, which in turn have total influence $O(\log s)$. This suggests that size-s DNFs themselves have total influence $O(\log s)$. To prove this though we'll need to reverse the steps of the plan; instead of truncating DNFs to a fixed width and arguing that a random input is unlikely to notice, we'll first pick a random (partial) input and argue that this is likely to make the width small.

Let's formalize the notion of a random partial input, or restriction:

Definition 4.15. For $\delta \in [0,1]$, we say that $\boldsymbol{J}$ is a *δ-random subset* of N if it is formed by including each element of N independently with probability δ. We define a *δ-random restriction on* $\{-1,1\}^n$ to be a pair $(\boldsymbol{J} \mid \boldsymbol{z})$, where first $\boldsymbol{J}$ is chosen to be a δ-random subset of $[n]$ and then $\boldsymbol{z} \sim \{-1,1\}^{\overline{\boldsymbol{J}}}$ is chosen uniformly at random. We say that coordinate $i \in [n]$ is *free* if $i \in \boldsymbol{J}$ and is *fixed* if $i \notin \boldsymbol{J}$. An equivalent definition is that each coordinate i is (independently) free with probability δ and fixed to ±1 with probability $(1-\delta)/2$ each.

Given $f : \{-1,1\}^n \to \mathbb{R}$ and a random restriction $(\boldsymbol{J} \mid \boldsymbol{z})$, we can form the restricted function $f_{\boldsymbol{J}|\boldsymbol{z}} : \{-1,1\}^{\boldsymbol{J}} \to \mathbb{R}$ as usual. However, it's inconvenient that the domain of this function depends on the random restriction. Thus when dealing with random restriction we usually invoke the following convention:

Definition 4.16. Given $f : \{-1,1\}^n \to \mathbb{R}$, $I \subseteq [n]$, and $z \in \{-1,1\}^{\overline{I}}$, we may identify the restricted function $f_{I|z} : \{-1,1\}^I \to \mathbb{R}$ with its extension $f_{I|z} : \{-1,1\}^n \to \mathbb{R}$ in which the input coordinates $\{-1,1\}^{\overline{I}}$ are ignored.

As mentioned, random restrictions interact nicely with Fourier expansions:

Proposition 4.17. *Fix $f : \{-1,1\}^n \to \mathbb{R}$ and $S \subseteq [n]$. Then if $(\boldsymbol{J} \mid \boldsymbol{z})$ is a δ-random restriction on $\{-1,1\}^n$,*

$$\mathbf{E}[\widehat{f_{\boldsymbol{J}|\boldsymbol{z}}}(S)] = \mathbf{Pr}[S \subseteq \boldsymbol{J}] \cdot \widehat{f}(S) = \delta^{|S|}\widehat{f}(S),$$

and

$$\mathbf{E}[\widehat{f_{\boldsymbol{J}|\boldsymbol{z}}}(S)^2] = \sum_{U \subseteq [n]} \mathbf{Pr}[U \cap \boldsymbol{J} = S] \cdot \widehat{f}(U)^2 = \sum_{U \supseteq S} \delta^{|S|}(1-\delta)^{|U \setminus S|}\widehat{f}(U)^2,$$

where we are treating $f_{\boldsymbol{J}|\boldsymbol{z}}$ as a function $\{-1,1\}^n \to \mathbb{R}$.

Proof. Suppose first that $J \subseteq [n]$ is fixed. When we think of restricted functions $f_{J|\boldsymbol{z}}$ as having domain $\{-1,1\}^n$, Corollary 3.22 may be stated as saying that for any $S \subseteq [n]$,

$$\mathop{\mathbf{E}}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[\widehat{f_{J|\boldsymbol{z}}}(S)] = \widehat{f}(S)\cdot \mathbf{1}_{S\subseteq J},$$

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

The proposition now follows by taking the expectation over $\boldsymbol{J}$. □

Corollary 4.18. *Fix $f : \{-1,1\}^n \to \mathbb{R}$ and $i \in [n]$. If $(\boldsymbol{J} \mid \boldsymbol{z})$ is a δ-random restriction, then $\mathbf{E}[\mathbf{Inf}_i[f_{\boldsymbol{J}|\boldsymbol{z}}]] = \delta\mathbf{Inf}_i[f]$. Hence also $\mathbf{E}[\mathbf{I}[f_{\boldsymbol{J}|\boldsymbol{z}}]] = \delta\mathbf{I}[f]$.*

Proof. We have

$$\begin{aligned}\mathbf{E}[\mathbf{Inf}_i[f_{\boldsymbol{J}|\boldsymbol{z}}]] = \mathbf{E}\left[\sum_{S\ni i}\widehat{f_{\boldsymbol{J}|\boldsymbol{z}}}(S)^2\right] &= \sum_{S\ni i}\sum_{U\subseteq[n]}\mathbf{Pr}[U\cap\boldsymbol{J}=S]\widehat{f}(U)^2\\ &= \sum_{U\subseteq[n]}\mathbf{Pr}[U\cap\boldsymbol{J}\ni i]\widehat{f}(U)^2 = \sum_{U\ni i}\delta\widehat{f}(U)^2 = \delta\mathbf{Inf}_i[f],\end{aligned}$$

where the second equality used Proposition 4.17. □

(Proving Corollary 4.18 via Proposition 4.17 is a bit more elaborate than necessary; see Exercise 4.9.)

Corollary 4.18 lets us bound the total influence of a function f by bounding the (expected) total influence of a random restriction of f. This is useful if f is computable by a DNF formula of small size, since a random restriction is very likely to make this DNF have small width. This is a consequence of the following lemma:

Lemma 4.19. *Let T be a DNF term over $\{-1,1\}^n$ and fix $w \in \mathbb{N}^+$. Let $(\boldsymbol{J} \mid \boldsymbol{z})$ be a $(1/2)$-random restriction on $\{-1,1\}^n$. Then $\mathbf{Pr}[width(T_{\boldsymbol{J}|\boldsymbol{z}}) \geq w] \leq (3/4)^w$.*

Proof. We may assume the initial width of T is at least w, as otherwise its restriction under $(\boldsymbol{J} \mid \boldsymbol{z})$ cannot have width at least w. Now if any literal appearing in T is fixed to False by the random restriction, the restricted term $T_{\boldsymbol{J}|\boldsymbol{z}}$ will be constantly False and thus have width $0 < w$. Each literal is fixed to False with probability 1/4; hence the probability no literal in T is fixed to False is at most $(3/4)^w$. □

We can now bound the total influence of small DNF formulas.

Theorem 4.20. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF of size s. Then $\mathbf{I}[f] \leq O(\log s)$.*

Proof. Let $(\boldsymbol{J} \mid \boldsymbol{z})$ be a (1/2)-random restriction on $\{-1,1\}^n$. Let $\boldsymbol{w} = \mathrm{DNF}_{\mathrm{width}}(f_{\boldsymbol{J}|\boldsymbol{z}})$. By a union bound and Lemma 4.19 we have that $\mathbf{Pr}[\boldsymbol{w} \geq w] \leq s(3/4)^w$. Hence

$$\begin{aligned}\mathbf{E}[\boldsymbol{w}] = \sum_{w=1}^{\infty} \mathbf{Pr}[\boldsymbol{w} \geq w] &\leq 3\log s + \sum_{w > 3\log s} s(3/4)^w \\ &\leq 3\log s + 4s(3/4)^{3\log s} \leq 3\log s + 4/s^{0.2} = O(\log s).\end{aligned}$$

From Proposition 4.7 we obtain $\mathbf{E}[\mathbf{I}[f_{\boldsymbol{J}|\boldsymbol{z}}]] \leq 2 \cdot O(\log s) = O(\log s)$. And so from Corollary 4.18 we conclude $\mathbf{I}[f] = 2\,\mathbf{E}[\mathbf{I}[f_{\boldsymbol{J}|\boldsymbol{z}}]] \leq O(\log s)$. □

4.4. Håstad's Switching Lemma and the spectrum of DNFs

Let's further investigate how random restrictions can simplify DNF formulas. Suppose f is computable by a DNF formula of width w, and we apply to it a δ-random restriction with $\delta \ll 1/w$. For each term T in the DNF, one of three things may happen to it under the random restriction. First and by far most likely, one of its literals may be fixed to False, allowing us to delete it. If this doesn't happen, the second possibility is that all of T's literals are made True, in which case the whole DNF reduces to the constantly True function. With $\delta \ll 1/w$, this is in turn much more likely than the third possibility, which is that at least one of T's literals is left free, but all the fixed literals are made True. Only in this third case is T not trivialized by the random restriction.

This reasoning might suggest that f is likely to become a constant function under the random restriction. Indeed, this is true, as the following theorem shows:

Baby Switching Lemma. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF or CNF of width at most w and let $(\boldsymbol{J} \mid \boldsymbol{z})$ be a δ-random restriction. Then*

$$\mathbf{Pr}[f_{\boldsymbol{J}|\boldsymbol{z}} \textit{ is not a constant function}] \leq 5\delta w.$$

This is in fact the $k = 1$ case of the following much more powerful theorem:

Håstad's Switching Lemma. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF or CNF of width at most w and let $(\boldsymbol{J} \mid \boldsymbol{z})$ be a δ-random restriction. Then for any $k \in \mathbb{N}$,*

$$\mathbf{Pr}[\mathrm{DT}(f_{\boldsymbol{J}|\boldsymbol{z}}) \geq k] \leq (5\delta w)^k.$$

What is remarkable about this result is that it has no dependence on the *size* of the DNF, or on n. In words, Håstad's Switching Lemma says that when $\delta \ll 1/w$, it's exponentially unlikely (in k) that applying a δ-random restriction to a width-w DNF does not convert ("switch") it to a decision tree of depth less than k. The result is called a "lemma" for historical reasons; in fact, its proof requires some work. You are asked to prove the Baby Switching Lemma in Exercise 4.19; for Håstad's Switching Lemma, consult Håstad's original proof [**Hås87**] or the alternate proof of Razborov [**Raz93, Bea94**].

Since we have strong results about the Fourier spectra of decision trees (Proposition 3.16), and since we know random restrictions interact nicely with Fourier coefficients (Proposition 4.17), Håstad's Switching Lemma allows us to prove some strong results about Fourier concentration of narrow DNF formulas. We start with an intermediate result which will be of use:

Lemma 4.21. *Let $f : \{-1,1\}^n \to \{-1,1\}$ and let $(\boldsymbol{J} \mid \boldsymbol{z})$ be a δ-random restriction, $\delta > 0$. Fix $k \in \mathbb{N}^+$ and write $\epsilon = \mathbf{Pr}[\mathrm{DT}(f_{\boldsymbol{J}|\boldsymbol{z}}) \geq k]$. Then the Fourier spectrum of f is 3ϵ-concentrated on degree up to $3k/\delta$.*

Proof. The key observation is that $\mathrm{DT}(f_{\boldsymbol{J}|\boldsymbol{z}}) < k$ implies $\deg(f_{\boldsymbol{J}|\boldsymbol{z}}) < k$ (Proposition 3.16), in which case the Fourier weight of $f_{\boldsymbol{J}|\boldsymbol{z}}$ at degree k and above is 0. Since this weight at most 1 in all cases we conclude

$$\mathop{\mathbf{E}}_{(\boldsymbol{J}|\boldsymbol{z})}\Bigg[\sum_{\substack{S\subseteq[n]\\|S|\geq k}} \widehat{f_{\boldsymbol{J}|\boldsymbol{z}}}(S)^2\Bigg] \leq \epsilon.$$

Using Proposition 4.17 we have

$$\mathop{\mathbf{E}}_{(\boldsymbol{J}|\boldsymbol{z})}\Bigg[\sum_{\substack{S\subseteq[n]\\|S|\geq k}} \widehat{f_{\boldsymbol{J}|\boldsymbol{z}}}(S)^2\Bigg] = \sum_{\substack{S\subseteq[n]\\|S|\geq k}} \mathop{\mathbf{E}}_{(\boldsymbol{J}|\boldsymbol{z})}[\widehat{f_{\boldsymbol{J}|\boldsymbol{z}}}(S)^2] = \sum_{U\subseteq[n]} \mathop{\mathbf{Pr}}_{(\boldsymbol{J}|\boldsymbol{z})}[|U \cap \boldsymbol{J}| \geq k]\cdot \widehat{f}(U)^2.$$

The distribution of random variable $|U \cap \boldsymbol{J}|$ is $\mathrm{Binomial}(|U|,\delta)$. When $|U| \geq 3k/\delta$ this random variable has mean at least $3k$, and a Chernoff bound shows $\mathbf{Pr}[|U \cap \boldsymbol{J}| < k] \leq \exp(-\frac{2}{3}k) \leq 2/3$. Thus

$$\epsilon \geq \sum_{U\subseteq[n]} \mathop{\mathbf{Pr}}_{(\boldsymbol{J}|\boldsymbol{z})}[|U \cap \boldsymbol{J}| \geq k]\cdot \widehat{f}(U)^2 \geq \sum_{|U|\geq 3k/\delta} (1-2/3)\cdot \widehat{f}(U)^2$$

and hence $\sum_{|U|\geq 3k/\delta} \widehat{f}(U)^2 \leq 3\epsilon$ as claimed. □

We can now improve the dependence on ϵ in Corollary 4.8's low-degree spectral concentration for DNFs:

Theorem 4.22. *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is computable by a DNF of width w. Then f's Fourier spectrum is ϵ-concentrated on degree up to $O(w\log(1/\epsilon))$.*

Proof. This follows immediately from Håstad's Switching Lemma and Lemma 4.21, taking $\delta = \frac{1}{10w}$ and $k = C\log(1/\epsilon)$ for a sufficiently large constant C. □

In Lemma 4.21, instead of using the fact that depth-k decision trees have no Fourier weight above degree k, we could have used the fact that their Fourier 1-norm is at most 2^k. As you are asked to show in Exercise 4.11, this would yield:

Lemma 4.23. *Let $f : \{-1,1\}^n \to \{-1,1\}$ and let $(\boldsymbol{J} \mid \boldsymbol{z})$ be a δ-random restriction. Then*

$$\sum_{U \subseteq [n]} \delta^{|U|} \cdot |\widehat{f}(U)| \leq \mathop{\mathbf{E}}_{(\boldsymbol{J}|\boldsymbol{z})}[2^{\mathrm{DT}(f_{\boldsymbol{J}|\boldsymbol{z}})}].$$

We can combine this with the Switching Lemma to deduce that width-w DNFs have small Fourier 1-norm at low degree:

Theorem 4.24. *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is computable by a DNF of width w. Then for any k,*

$$\sum_{|U| \leq k} |\widehat{f}(U)| \leq 2 \cdot (20w)^k.$$

Proof. Apply Håstad's Switching Lemma to f with $\delta = \frac{1}{20w}$ to deduce

$$\mathop{\mathbf{E}}_{(\boldsymbol{J}|\boldsymbol{z})}[2^{\mathrm{DT}(f_{\boldsymbol{J}|\boldsymbol{z}})}] \leq \sum_{d=0}^{\infty} \left(\tfrac{5}{20}\right)^d \cdot 2^d = 2.$$

Thus from Lemma 4.23 we get

$$2 \geq \sum_{U \subseteq [n]} \left(\tfrac{1}{20w}\right)^{|U|} \cdot |\widehat{f}(U)| \geq \left(\tfrac{1}{20w}\right)^k \cdot \sum_{|U| \leq k} |\widehat{f}(U)|,$$

as needed. □

Our two theorems about the Fourier structure of DNF are *almost* enough to prove Mansour's Conjecture:

Theorem 4.25. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF of width $w \geq 2$. Then for any $\epsilon \in (0,1/2]$, the Fourier spectrum of f is ϵ-concentrated on a collection $\mathscr{F}$ with $|\mathscr{F}| \leq w^{O(w \log(1/\epsilon))}$.*

Proof. Let $k = Cw\log(4/\epsilon)$ and let $g = f^{\leq k}$. If C is a large enough constant, then Theorem 4.22 tells us that $\|f - g\|_2^2 \leq \epsilon/4$. Furthermore, Theorem 4.24 gives $\hat{\|}g\hat{\|}_1 \leq w^{O(w\log(1/\epsilon))}$. By Exercise 3.16, g is $(\epsilon/4)$-concentrated on some collection $\mathscr{F}$ with $|\mathscr{F}| \leq 4\hat{\|}g\hat{\|}_1^2/\epsilon \leq w^{O(w\log(1/\epsilon))}$. And so by Exercise 3.17, f is ϵ-concentrated on this same collection. □

For the interesting case of DNFs of width $O(\log n)$ and constant ϵ, we get concentration on a collection of cardinality $O(\log n)^{O(\log n)} = n^{O(\log\log n)}$, nearly polynomial. Using Proposition 4.9 (and Exercise 3.17) we get the same deduction for DNFs of size poly(n); more generally, for size s we have ϵ-concentration on a collection of cardinality at most $(s/\epsilon)^{O(\log\log(s/\epsilon)\log(1/\epsilon))}$.

4.5. Highlight: LMN's work on constant-depth circuits

Having derived strong results about the Fourier spectrum of small DNFs and CNFs, we will now extend to the case of *constant-depth circuits*. We begin by describing how Håstad applied his Switching Lemma to constant-depth circuits. We then describe some Fourier-theoretic consequences coming from a very early (1989) work in analysis of Boolean functions by Linial, Mansour, and Nisan (LMN).

To define constant-depth circuits it is best to start with a picture. Here is an example of a depth-3 circuit:

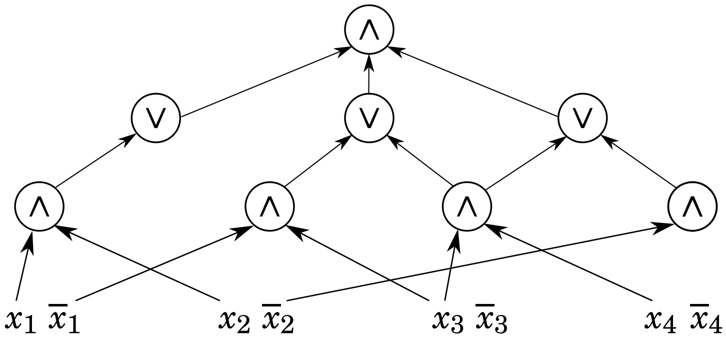


Figure 4.1. Example of a depth-3 circuit, with the layer 0 nodes at the bottom and the layer 3 node at the top

This circuit computes the function

$$x_1x_2 \wedge (\overline{x}_1x_3 \vee x_3x_4) \wedge (x_3x_4 \vee \overline{x}_2),$$

where we suppressed the $\wedge$ in concatenated literals. To be precise:

Definition 4.26. For an integer $d \geq 2$, we define a *depth-d circuit* over Boolean variables $x_1, \dots, x_n$ as follows: It is a directed acyclic graph in which the nodes ("gates") are arranged in $d+1$ layers, with all arcs ("wires") going from layer $j-1$ to layer j for some $j \in [d]$. There are exactly $2n$ nodes in layer 0 (the "inputs") and exactly 1 node in layer d (the "output"). The nodes in layer 0 are labeled by the $2n$ literals. The nodes in layers 1, 3, 5, etc. have the same label, either $\wedge$ or $\vee$, and the nodes in layers 2, 4, 6, etc. have the other label. Each node "computes" a function $\{-1,1\}^n \to \{-1,1\}$: the literals compute themselves and the $\wedge$ (respectively, $\vee$) nodes compute the logical AND (respectively, OR) of the functions computed by their incoming nodes. The circuit itself is said to compute the function computed by its output node.

In particular, DNFs and CNFs are depth-2 circuits. We extend the definitions of size and width appropriately:

Definition 4.27. The *size* of a depth-d circuit is defined to be the number of nodes in layers 1 through $d-1$. Its *width* is the maximum in-degree of any node at layer 1. (As with DNFs and CNFs, we insist that no node at layer 1 is connected to a variable or its negation more than once.)

The layering we assume in our definition of depth-d circuits can be achieved with a factor-$2d$ size overhead for any "unbounded fan-in AND/OR/NOT circuit". We will not discuss any other type of Boolean circuit in this section.

We now show that Håstad's Switching Lemma can be usefully applied not just to DNFs and CNFs but more generally to constant-depth circuits:

Lemma 4.28. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a depth-d circuit of size s and width w, and let $\epsilon \in (0,1]$. Set*

$$\delta = \frac{1}{10w}\left(\frac{1}{10\ell}\right)^{d-2}, \quad \textit{where } \ell = \log(2s/\epsilon).$$

Then if $(\boldsymbol{J} \mid \boldsymbol{z})$ is a δ-random restriction, $\mathbf{Pr}[\mathrm{DT}(f_{\boldsymbol{J}|\boldsymbol{z}}) \geq \log(2/\epsilon)] \leq \epsilon$.

Proof. The $d = 2$ case is immediate from Håstad's Switching Lemma, so we assume $d \geq 3$.

The first important observation is that random restrictions "compose". That is, making a δ_1-random restriction followed by a δ_2-random restriction to the free coordinates is equivalent to making a $\delta_1\delta_2$-random restriction. Thus we can think of $(\boldsymbol{J} \mid \boldsymbol{z})$ as being produced as follows:

(1) make a $\frac{1}{10w}$-random restriction;

(2) make $d-3$ subsequent $\frac{1}{10\ell}$-random restrictions;

(3) make a final $\frac{1}{10\ell}$-random restriction.

Without loss of generality, assume the nodes at layer 2 of the circuit are labeled $\vee$. Thus any node g at layer 2 computes a DNF of width at most w. By Håstad's Switching Lemma, after the initial $\frac{1}{10w}$-random restriction g can be replaced by a decision tree of depth at most ℓ except with probability at most $2^{-\ell}$. In particular, it can be replaced by a CNF of width at most ℓ, using Proposition 4.5. If we write s_2 for the number of nodes at layer 2, a union bound lets us conclude:

$$\mathop{\mathbf{Pr}}_{\substack{\frac{1}{10w}\text{-random} \\ \text{restriction}}}[\text{not all nodes at layer 2 replaceable by width-}\ell\text{ CNFs}] \leq s_2 \cdot 2^{-\ell}. \tag{4.1}$$

We now come to the second important observation: If all nodes at layer 2 can be switched to width-ℓ CNFs, then layers 2 and 3 can be "compressed", producing a depth-$(d-1)$ circuit of width at most ℓ. More precisely, we can form an equivalent circuit by shortening all length-2 paths from layer 1 to layer 3 into single arcs, and then deleting the nodes at layer 2. We give an illustration of this in Figure 4.2:

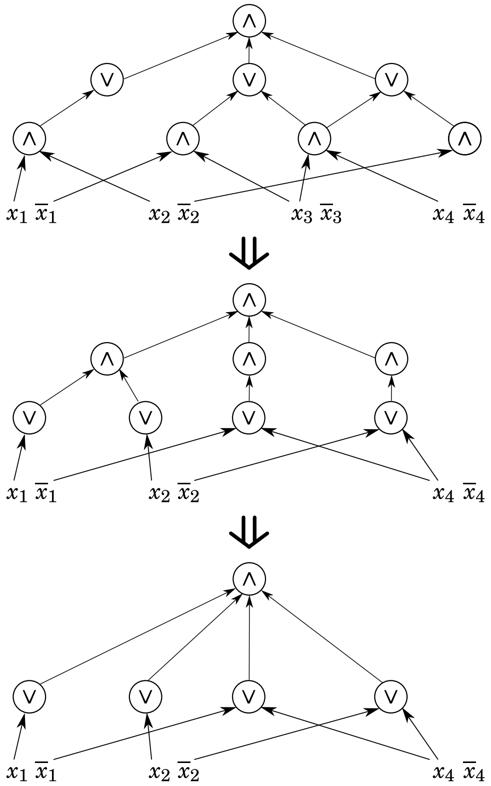


Figure 4.2. At top is the initial circuit. Under the restriction fixing x_3 = True, all three DNFs at layer 2 may be replaced by CNFs of width at most 2. Finally, the nodes at layers 2 and 3 may be compressed.

Assuming the event in (4.1) does not occur, the initial $\frac{1}{10w}$-random restriction reduces the circuit to having depth-$(d-1)$ and width at most ℓ. The number of $\wedge$-nodes at the new layer 2 is at most s_3, the number of nodes at layer 3 in the *original* circuit.

Next we make a $\frac{1}{10\ell}$-random restriction. As before, by Håstad's Switching Lemma this reduces all width-ℓ CNFs at the new layer 2 to depth-ℓ decision trees (hence width-ℓ DNFs), except with probability at most $s_3 \cdot 2^{-\ell}$. We may then compress layers and reduce depth again.

Proceeding for all $\frac{1}{10\ell}$-random restrictions except the final one, a union bound gives

$$\mathop{\mathbf{Pr}}_{\substack{\frac{1}{10w}(\frac{1}{10\ell})^{d-3}\text{-random} \\ \text{restriction}}}[\text{circuit does not reduce to depth 2 and width } \ell] \\ \leq s_2 \cdot 2^{-\ell} + s_3 \cdot 2^{-\ell} + \cdots + s_{d-1} \cdot 2^{-\ell} \leq s \cdot 2^{-\ell} = \epsilon/2.$$

Assuming the event above does not occur, Håstad's Switching Lemma tells us that the final $\frac{1}{10\ell}$-random restriction reduces the circuit to a decision tree of depth less than $\log(2/\epsilon)$ except with probability at most $\epsilon/2$. This completes the proof. □

We may now obtain the main theorem of Linial, Mansour, and Nisan:

LMN Theorem. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a depth-d circuit of size $s > 1$ and let $\epsilon \in (0,1/2]$. Then f's Fourier spectrum is ϵ-concentrated up to degree $O(\log(s/\epsilon))^{d-1} \cdot \log(1/\epsilon)$.*

Proof. If the circuit for f also had width at most w, we could deduce 3ϵ-concentration up to degree $30w \cdot (10\log(2s/\epsilon))^{d-2} \cdot \log(2/\epsilon)$ by combining Lemma 4.28 with Lemma 4.21. But if we simply delete all layer-1 nodes of width at least $\log(s/\epsilon)$, the resulting circuit computes a function which is ϵ-close to f, as in the proof of Proposition 4.9. Thus (using Exercise 3.17) f's spectrum is $O(\epsilon)$-concentrated up to degree $O(\log(2s/\epsilon))^{d-1} \cdot \log(2/\epsilon)$, and the result follows by adjusting constants. □

Remark 4.29. Håstad [**Hås01a**] has slightly sharpened the degree in the LMN Theorem to $O(\log(s/\epsilon))^{d-2} \cdot \log(s) \cdot \log(1/\epsilon)$.

In Exercise 4.20 you are asked to use a simpler version of this proof, along the lines of Theorem 4.20, to show the following:

Theorem 4.30. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a depth-d circuit of size s. Then $\mathbf{I}[f] \le O(\log s)^{d-1}$.*

These rather strong Fourier concentration results for constant-depth circuits have several applications. By introducing the Low-Degree Algorithm for learning, Linial–Mansour–Nisan gave as their main application:

Theorem 4.31. *Let $\mathscr{C}$ be the class of functions $f : \{-1,1\}^n \to \{-1,1\}$ computable depth-d $\mathrm{poly}(n)$-size circuits. Then $\mathscr{C}$ can be learned from random examples with error any $\epsilon = 1/\mathrm{poly}(n)$ in time $n^{O(\log n)^d}$.*

In complexity theory the class of poly-size, constant-depth circuits is referred to as AC^0. Thus the above theorem may be summarized as "AC^0 is learnable in quasipolynomial time". In fact, under a strong enough assumption about the intractability of factoring certain integers, it is known that quasipolynomial time is *required* to learn AC^0 circuits, even with query access [**Kha93**].

The original motivation of the line of work leading to Håstad's Switching Lemma was to show that the parity function $\chi_{[n]}$ cannot be computed in AC^0. Håstad even showed that AC^0 cannot even approximately compute parity. We can derive this result from the LMN Theorem:

Corollary 4.32. *Fix any constant $\epsilon_0 > 0$. Suppose C is a depth-d circuit over $\{-1,1\}^n$ with $\mathbf{Pr}_{\boldsymbol{x}}[C(\boldsymbol{x}) = \chi_{[n]}(x)] \geq 1/2 + \epsilon_0$. Then the size of C is at least $2^{\Omega(n^{1/(d-1)})}$.*

Proof. The hypothesis on C implies $\widehat{C}([n]) \geq 2\epsilon_0$. The result then follows by taking $\epsilon = 2\epsilon_0^2$ in the LMN Theorem. □

This corollary is close to being tight, since the parity $\chi_{[n]}$ *can* be computed by a depth-d circuit of size $n2^{n^{1/(d-1)}}$ for any $d \geq 2$; see Exercise 4.12. The simpler result Theorem 4.30 is often handier for showing that certain functions can't be computed by AC^0 circuits. For example, we know that $\mathbf{I}[\mathrm{Maj}_n] = \Theta(\sqrt{n})$; hence any constant-depth circuit computing Maj_n must have size at least $2^{n^{\Omega(1)}}$.

Finally, Linial, Mansour, and Nisan gave an application to cryptography. Informally, a function $f : \{-1,1\}^m \times \{-1,1\}^n \to \{-1,1\}$ is said to be a "pseudorandom function generator with seed length m" if, for any efficient algorithm A,

$$\left| \mathop{\mathbf{Pr}}_{\boldsymbol{s}\sim\{-1,1\}^m}[A(f(\boldsymbol{s},\cdot)) = \text{"accept"}] - \mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim\{-1,1\}^{\{-1,1\}^n}}[A(\boldsymbol{g}) = \text{"accept"}] \right| \leq 1/n^{\omega(1)}.$$

Here the notation $A(h)$ means that A has query access to target function h, and $\boldsymbol{g} \sim \{-1,1\}^{\{-1,1\}^n}$ means that $\boldsymbol{g}$ is a uniformly random n-bit function. In other words, for almost all "seeds" $\boldsymbol{s}$ the function $f(\boldsymbol{s},\cdot) : \{-1,1\}^n \to \{-1,1\}$ is nearly indistinguishable (to efficient algorithms) from a truly random function. Theorem 4.30 shows that pseudorandom function generators cannot be computed by AC^0 circuits. To see this, consider the algorithm $A(h)$ which chooses $\boldsymbol{x} \sim \{-1,1\}^n$ and $\boldsymbol{i} \in [n]$ uniformly at random, queries $h(\boldsymbol{x})$ and $h(\boldsymbol{x}^{\oplus \boldsymbol{i}})$, and accepts if these values are unequal. If h is a uniformly random function, $A(h)$ will accept with probability 1/2. In general, $A(h)$ accepts with probability $\mathbf{I}[h]/n$. Thus Theorem 4.30 implies that if h is computable in AC^0 then $A(h)$ accepts with probability at most $\mathrm{polylog}(n)/n \ll 1/2$.

4.6. Exercises and notes

4.1 Show that every function $f : \{0,1\}^n \to \{0,1\}$ can be represented by a DNF formula of size at most 2^n and width at most n.

4.2 Suppose we have a certain CNF computing $f : \{0,1\}^n \to \{0,1\}$. Switch ANDs with ORs in the CNF. Show that the result is a DNF computing the Boolean dual $f^\dagger : \{0,1\}^n \to \{0,1\}$.

4.3 A DNF formula is said to be *monotone* if its terms contain only unnegated variables. Show that monotone DNFs compute monotone functions and that any monotone function can be computed by a monotone DNF, but that a nonmonotone DNF may compute a monotone function.

4.4 Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF of size s.

(*a*) Show there exists $S \subseteq [n]$ with $|S| \le \log(s) + O(1)$ and $|\widehat{f}(S)| \ge \Omega(1/s)$. (Hint: Use Proposition 4.9 and Exercise 3.30.)

(*b*) Let $\mathscr{C}$ be the concept class of functions $: \{-1,1\}^n \to \{-1,1\}$ computable by DNF formulas of size at most s. Show that $\mathscr{C}$ is learnable using queries with error $\frac{1}{2} - \Omega(1/s)$ in time $\mathrm{poly}(n,s)$. (Such a result, with error bounded away from $\frac{1}{2}$, is called *weak learning*.)

4.5 Verify Proposition 4.12.

4.6 Verify Proposition 4.14.

4.7 For each n that is an input length for Tribes_n, show that there exists a function $f : \{-1,1\}^n \to \{-1,1\}$ that is truly unbiased ($\mathbf{E}[f] = 0$) and has $\mathbf{Inf}_i[f] \le O\big(\frac{\log n}{n}\big)$ for all $i \in [n]$.

4.8 Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is computed by a *read-once* DNF (meaning no variable is involved in more than one term) in which all terms have width exactly w. Compute $\hat{\lVert} f \hat{\rVert}_1$ exactly. Deduce that $\hat{\lVert} \mathrm{Tribes}_n \hat{\rVert}_1 = 2^{\frac{n}{\log n}(1 \pm o(1))}$ and that there are n-variable width-2 DNFs with Fourier 1-norm $\Omega(\sqrt{3/2}^{\,n})$.

4.9 Give a direct (Fourier-free) proof of Corollary 4.18. (Hint: Condition on whether $i \in \boldsymbol{J}$.)

4.10 Tighten the constant factor on $\log s$ in Theorem 4.20 as much as you can (avenues of improvement include the argument in Lemma 4.19, the choice of δ, and Exercise 4.17).

4.11 Prove Lemma 4.23.

4.12 (*a*) Show that the parity function $\chi_{[n]} : \{-1,1\}^n \to \{-1,1\}$ can be computed by a DNF (or a CNF) of size 2^{n-1}.

(*b*) Show that the bound 2^{n-1} above is exactly tight. (Hint: Show that every term must have width exactly n.)

(*c*) Show that there is a depth-3 circuit of size $O(n^{1/2}) \cdot 2^{n^{1/2}}$ computing $\chi_{[n]}$. (Hint: Break up the input into $n^{1/2}$ blocks of size $n^{1/2}$ and use (*a*) twice. How can you compress the result from depth 4 to depth 3?)

(*d*) More generally, show there is a depth-d circuit of size $O(n^{1-1/(d-1)}) \cdot 2^{n^{1/(d-1)}}$ computing $\chi_{[n]}$.

4.13 In this exercise we define the most standard class of Boolean circuits. A *(De Morgan) circuit C* over Boolean variables $x_1, \dots, x_n$ is a directed acyclic graph in which each node ("gate") is labeled with either an x_i or with $\wedge$, $\vee$, or $\neg$ (logical NOT). Each x_i is used as label exactly once; the associated nodes are called "input" gates and must have in-degree 0. Each $\wedge$ and $\vee$ node must have in-degree 2, and each $\neg$ node must have in-degree 1. Each node "computes" a Boolean function of the inputs as in Definition 4.26. Finally, one node of C is designated as the "output" gate, and C itself is said to compute the function computed by the output node. For this type of circuit we define its *size*, denoted size(C), to be the number of nodes.

Show that each of the following n-input functions can be computed by De Morgan circuits of size $O(n)$:

(a) The logical AND function.
(b) The parity function.
(c) The complete quadratic function from Exercise 1.1.

4.14 Show that computing $\mathrm{Tribes}_{w,s}$ by a CNF formula requires size at least w^s.

4.15 Show that there is a universal constant $\epsilon_0 > 0$ such that the following holds: Every $\frac{3}{4}n$-junta $g : \{-1,1\}^n \to \{-1,1\}$ is ϵ_0-far from Tribes_n (assuming $n > 1$). (Hint: Letting J denote the coordinates on which g depends, show that if J has non-full intersection with at least $\frac{1}{4}$ of the tribes/terms then when $\boldsymbol{x} \sim \{-1,1\}^J$, there is a constant chance that $\mathbf{Var}[f_{|\boldsymbol{x}}] \geq \Omega(1)$.)

4.16 Using the KKL Theorem, show that if $f : \{-1,1\}^n \to \{-1,1\}$ is a transitive-symmetric function with $\mathbf{Var}[f] \geq \Omega(1)$, then $\mathbf{I}[f] \geq \Omega(\log n)$.

4.17 Let $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ be computable by a CNF C of width w. In this exercise you will show that $\mathbf{I}[f] \leq w$.

Consider the following randomized algorithm that tries to produce an input $\boldsymbol{x} \in f^{-1}(\mathsf{True})$. First, choose a random permutation $\boldsymbol{\pi} \in S_n$. Then for $i = 1, \dots, n$: If the single-literal clause $x_{\boldsymbol{\pi}(i)}$ appears in C, then set $x_{\boldsymbol{\pi}(i)} = \mathsf{True}$, syntactically simplify C under this setting, and say that coordinate $\boldsymbol{\pi}(i)$ is "forced". Similarly, if the single-literal clause $\overline{x}_{\boldsymbol{\pi}(i)}$ appears in C, then set $x_{\boldsymbol{\pi}(i)} = \mathsf{False}$, syntactically simplify C, and say that $\boldsymbol{\pi}(i)$ is "forced". If neither holds, set $\boldsymbol{x}_{\pi(i)}$ uniformly at random. If C ever contains two single-literal clauses x_j and $\overline{x}_j$, the algorithm "gives up" and outputs $\boldsymbol{x} = \perp$.

(a) Show that if $\boldsymbol{x} \neq \perp$, then $f(\boldsymbol{x}) = \mathsf{True}$.
(b) For $x \in f^{-1}(\mathsf{True})$ let $p(x) = \mathbf{Pr}[\boldsymbol{x} = x]$. For $j \in [n]$ let $\boldsymbol{I}_j$ be the indicator random variable for the event that coordinate $j \in [n]$ is forced. Show that $p(x) = \mathbf{E}[\prod_{j=1}^n (1/2)^{1-\boldsymbol{I}_j}]$.
(c) Deduce $2^n p(x) \geq 2^{\sum_{j=1}^n \mathbf{E}[\boldsymbol{I}_j]}$.
(d) Show that for every x with $f(x) = \mathsf{True}$, $f(x^{\oplus j}) = \mathsf{False}$ it holds that $\mathbf{E}[\boldsymbol{I}_j \mid \boldsymbol{x} = x] \geq 1/w$.
(e) Deduce $\mathbf{I}[f] \leq w$.

4.18 Given Boolean variables $x_1, \dots, x_n$, a "random monotone term of width $w \in \mathbb{N}^+$" is defined to be the logical AND of $x_{\boldsymbol{i}_1}, \dots, x_{\boldsymbol{i}_w}$, where $\boldsymbol{i}_1, \dots, \boldsymbol{i}_w$ are chosen independently and uniformly at random from $[n]$. (If the $\boldsymbol{i}_j$'s are not all distinct then the resulting term will in fact have width strictly less than w.) A "random monotone DNF of width w and size s" is defined to be the logical OR of s independent random monotone terms. For this exercise we assume n is a sufficiently large perfect square, and we let $\boldsymbol{\varphi}$ be a random monotone DNF of width $\sqrt{n}$ and size $2^{\sqrt{n}}$.

(a) Fix an input $x \in \{-1,1\}^n$ and define $u = (\sum_{i=1}^n x_i)/\sqrt{n} \in [-\sqrt{n}, \sqrt{n}]$. Let $\boldsymbol{V}_j$ be the event that the jth term of $\boldsymbol{\varphi}$ is made 1 (logical False) by x. Compute $\mathbf{Pr}[\boldsymbol{V}_j]$ and $\mathbf{Pr}[\boldsymbol{\varphi}(x) = 1]$, and show that the latter is at least 10^{-9} assuming $|u| \leq 2$.

(b) Let $\boldsymbol{U}_j$ be the event that the jth term of $\boldsymbol{\varphi}$ has exactly one 1 on input x. Show that $\mathbf{Pr}[\boldsymbol{U}_j \mid \boldsymbol{V}_j] \geq \Omega(w2^{-w})$ assuming $|u| \leq 2$.

(c) Suppose we condition on $\boldsymbol{\varphi}(x) = 1$; i.e., $\cup_j \boldsymbol{V}_j$. Argue that the events $\boldsymbol{U}_j$ are independent. Further, argue that for the $\boldsymbol{U}_j$'s that do occur, the indices of their uniquely-1 variables are independent and uniformly random among the 1's of x.

(d) Show that $\mathbf{Pr}[\mathrm{sens}_{\boldsymbol{\varphi}}(x) \geq c\sqrt{n} \mid \boldsymbol{\varphi}(x) = 1] \geq 1 - 10^{-10}$ for $c > 0$ a sufficiently small constant.

(e) Show that $\mathbf{Pr}_{\boldsymbol{x}}[|(\sum_{i=1}^n \boldsymbol{x}_i)/\sqrt{n}| \leq 2] \geq \Omega(1)$.

(f) Deduce that there exists a monotone function $f : \{-1,1\}^n \to \{-1,1\}$ with the property that $\mathbf{Pr}_{\boldsymbol{x}}[\mathrm{sens}_f(\boldsymbol{x}) \geq c'\sqrt{n}] \geq c'$ for some universal constant $c' > 0$.

(g) Both Maj_n and the function f from the previous exercise have average sensitivity $\Theta(\sqrt{n})$. Contrast the "way" in which this occurs for the two functions.

4.19 In this exercise you will prove the Baby Switching Lemma with constant 3 in place of 5. Let $\phi = T_1 \vee T_2 \vee \cdots \vee T_s$ be a DNF of width $w \geq 1$ over variables $x_1, \dots, x_n$. We may assume $\delta \leq 1/3$, else the theorem is trivial.

(a) Suppose $R = (J \mid z)$ is a "bad" restriction, meaning that $\phi_{J|z}$ is not a constant function. Let i be minimal such that $(T_i)_{J|z}$ is neither constantly True or False, and let j be minimal such that x_j or $\overline{x}_j$ appears in this restricted term. Show there is a unique restriction $R' = (J \setminus \{j\} \mid z')$ extending R that doesn't falsify T_i.

(b) Suppose we enumerate all bad restrictions R, and for each we write the associated R' as in (a). Show that no restriction is written more than w times.

(c) If $(\boldsymbol{J} \mid \boldsymbol{z})$ is a δ-random restriction and R and R' are as in (a), show that $\mathbf{Pr}[(\boldsymbol{J} \mid \boldsymbol{z}) = R] = \frac{2\delta}{1-\delta} \mathbf{Pr}[(\boldsymbol{J} \mid \boldsymbol{z}) = R']$.

(d) Complete the proof by showing $\mathbf{Pr}[(\boldsymbol{J} \mid \boldsymbol{z}) \text{ is bad}] \leq 3\delta w$.

4.20 In this exercise you will prove Theorem 4.30. Say that a "(d, w, s')-circuit" is a depth-d circuit with width at most w and with at most s' nodes at layers 2 through d (i.e., excluding layers 0 and 1).

(a) Show by induction on $d \geq 2$ that any $f : \{-1,1\}^n \to \{-1,1\}$ computable by a (d, w, s')-circuit satisfies $\mathbf{I}[f] \leq wO(\log s')^{d-2}$.

(b) Deduce Theorem 4.30.

Notes. Mansour's Conjecture dates from 1994 [Man94]. Even the weaker version would imply that the Kushilevitz–Mansour algorithm learns the class of poly(n)-size DNF with any constant error, using queries, in time poly(n). In fact, this learning result was subsequently obtained in a celebrated work of Jackson [Jac97], using a different method (which begins with Exercise 4.4). Nevertheless, the Mansour Conjecture remains important for learning theory since Gopalan, Kalai, and Klivans [GKK08] have shown that it implies the same learning result in the more challenging and realistic model of "agnostic learning". Theorems 4.24 and 4.25 are also due to Mansour [Man95].

The method of random restrictions dates back to Subbotovskaya [Sub61]. Håstad's Switching Lemma [Hås87] and his Lemma 4.28 are the culmination of a line of work due to Furst, Saxe, and Sipser [FSS84], Ajtai [Ajt83], and Yao [Yao85]. Linial, Mansour, and Nisan [LMN89, LMN93] proved Lemma 4.21, which allowed them to deduce the LMN Theorem and its consequences. An additional cryptographic application of the LMN Theorem is found in Goldmann and Russell [GR00]. The strongest lower bound currently known for approximately computing parity in AC^0 is due to Impagliazzo, Matthews, and Paturi [IMP12] and independently to Håstad [Hås12].

Theorem 4.20 and its generalization Theorem 4.30 are due to Boppana [Bop97]; Linial, Mansour, and Nisan had given the weaker bound $O(\log s)^d$. Exercise 4.17 is due to Amano [Ama11], and Exercise 4.18 is due to Talagrand [Tal96].

Chapter 5

Majority and threshold functions

This chapter is devoted to linear threshold functions, their generalization to higher degrees, and their exemplar the majority function. The study of LTFs leads naturally to the introduction of the Central Limit Theorem and Gaussian random variables – important tools in analysis of Boolean functions. We will first use these tools to analyze the Fourier spectrum of the Maj_n function, which in some sense "converges" as $n \to \infty$. We'll then extend to analyzing the degree-1 Fourier weight, noise stability, and total influence of general linear threshold functions.

5.1. Linear threshold functions and polynomial threshold functions

Recall from Chapter 2.1 that a linear threshold function (abbreviated LTF) is a Boolean-valued function $f : \{-1,1\}^n \to \{-1,1\}$ that can be represented as

$$f(x) = \mathrm{sgn}(a_0 + a_1x_1 + \cdots + a_nx_n) \qquad (5.1)$$

for some constants $a_0, a_1, \ldots, a_n \in \mathbb{R}$. (For definiteness we'll take $\mathrm{sgn}(0) = 1$. If we're using the representation $f : \{-1,1\}^n \to \{0,1\}$, then f is an LTF if it can be represented as $f(x) = 1_{\{a_0+a_1x_1+\cdots+a_nx_n>0\}}$.) Examples include majority, AND, OR, dictators, and decision lists (Exercise 3.23). Besides representing "weighted majority" voting schemes, LTFs play an important role in learning theory and in circuit complexity.

There is also a geometric perspective on LTFs. Writing $\ell(x) = a_0 + a_1x_1 + \cdots + a_nx_n$, we can think of ℓ as an affine function $\mathbb{R}^n \to \mathbb{R}$. Then $\mathrm{sgn}(\ell(x))$ is the ±1-indicator of a *halfspace* in $\mathbb{R}^n$. A Boolean LTF is thus the restriction of such a halfspace-indicator to the discrete cube $\{-1,1\}^n \subset \mathbb{R}^n$. Equivalently, a function $f : \{-1,1\}^n \to \{-1,1\}$ is an LTF if and only if it has a "linear separator"; i.e., a hyperplane in $\mathbb{R}^n$ that separates the points f labels 1 from the points f labels -1.

An LTF $f : \{-1,1\}^n \to \{-1,1\}$ can have several different representations as in (5.1) – in fact it always has infinitely many. This is clear from the geometric viewpoint; any small enough perturbation to a linear separator will not change the way it partitions the discrete cube. Because we can make these perturbations, we may ensure that $a_0 + a_1x_1 + \cdots + a_nx_n \neq 0$ for every $x \in \{-1,1\}^n$. We'll usually insist that LTF representations have this property so that the nuisance of sgn(0) doesn't arise. We also observe that we can scale all of the coefficients in an LTF representation by the same positive constant without changing the LTF. These observations can be used to show it's always possible to take the a_i's to be integers (Exercise 5.1). However, we will most often scale so that $\sum_{i=1}^n a_i^2 = 1$; this is convenient when using the Central Limit Theorem.

The most elegant result connecting LTFs and Fourier expansions is Chow's Theorem, which says that a Boolean LTF is completely determined by its degree-0 and degree-1 Fourier coefficients. In fact, it's determined not just within the class of LTFs but within the class of all Boolean functions:

Theorem 5.1. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and let $g : \{-1,1\}^n \to \{-1,1\}$ be any function. If $\widehat{g}(S) = \widehat{f}(S)$ for all $|S| \leq 1$, then $g = f$.*

Proof. Let $f(x) = \mathrm{sgn}(\ell(x))$, where $\ell : \{-1,1\}^n \to \mathbb{R}$ has degree at most 1 and is never 0 on $\{-1,1\}^n$. For any $x \in \{-1,1\}^n$ we have $f(x)\ell(x) = |\ell(x)| \geq g(x)\ell(x)$, with equality if and only if $f(x) = g(x)$ (here we use $\ell(x) \neq 0$). Using this observation along with Plancherel's Theorem (twice) we have

$$\sum_{|S|\leq 1} \widehat{f}(S)\widehat{\ell}(S) = \mathbf{E}[f(\boldsymbol{x})\ell(\boldsymbol{x})] \geq \mathbf{E}[g(\boldsymbol{x})\ell(\boldsymbol{x})] = \sum_{|S|\leq 1} \widehat{g}(S)\widehat{\ell}(S).$$

But by assumption, the left-hand and right-hand sides above are equal. Thus the inequality must be an equality for every value of $\boldsymbol{x}$; i.e., $f(x) = g(x)\ \forall x$. □

In light of Chow's Theorem, the $n+1$ numbers $\widehat{g}(\emptyset), \widehat{g}(\{1\}), \dots, \widehat{g}(\{n\})$ are sometimes called the *Chow parameters* of the Boolean function g.

As we will show in Section 5.5, linear threshold functions are very noise-stable; hence they have a lot of their Fourier weight at low degrees. Here is a simple result along these lines:

Theorem 5.2. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF. Then* $\mathbf{W}^{\leq 1}[f] \geq 1/2$.

Proof. Writing $f(x) = \mathrm{sgn}(\ell(x))$ we have

$$\|\ell\|_1 = \mathbf{E}[|\ell(\boldsymbol{x})|] = \langle f, \ell\rangle = \langle f^{\leq 1}, \ell\rangle \leq \|f^{\leq 1}\|_2 \|\ell\|_2 = \sqrt{\mathbf{W}^{\leq 1}[f]} \cdot \|\ell\|_2,$$

where the third equality follows from Plancherel and the inequality is Cauchy–Schwarz. Assume first that $\ell(x) = a_1x_1 + \cdots + a_nx_n$ (i.e., $\ell(x)$ has no constant term). The Khintchine–Kahane Inequality (Exercise 2.55) states that $\|\ell\|_1 \geq \frac{1}{\sqrt{2}}\|\ell\|_2$, and hence we deduce

$$\tfrac{1}{\sqrt{2}}\|\ell\|_2 \leq \sqrt{\mathbf{W}^{\leq 1}[f]} \cdot \|\ell\|_2.$$

The conclusion $\mathbf{W}^{\leq 1}[f] \geq 1/2$ follows immediately (since $\|\ell\|_2$ cannot be 0). The case when $\ell(x)$ has a constant term is handled in Exercise 5.5. □

From Exercise 2.22 we know that $\mathbf{W}^{\leq 1}[\mathrm{Maj}_n] = \mathbf{W}^{1}[\mathrm{Maj}_n] \geq 2/\pi$ for all n; it is reasonable to conjecture that majority is extremal for Theorem 5.2. This is an open problem.

Conjecture 5.3. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF. Then $\mathbf{W}^{\leq 1}[f] \geq 2/\pi$.*

A natural generalization of linear threshold functions is *polynomial threshold functions*:

Definition 5.4. A function $f : \{-1,1\}^n \to \{-1,1\}$ is called a *polynomial threshold function (PTF)* of degree at most k if it is expressible as $f(x) = \mathrm{sgn}(p(x))$ for some real polynomial $p : \{-1,1\}^n \to \mathbb{R}$ of degree at most k.

Example 5.5. Let $f : \{-1,1\}^4 \to \{-1,1\}$ be the 4-bit equality function, which is 1 if and only if all input bits are equal. Then f is a degree-2 PTF because it has the representation $f(x) = \mathrm{sgn}(-3 + x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4)$.

Every Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ is a PTF of degree at most n, since we can take the sign of its Fourier expansion. Thus we are usually interested in the case when the degree k is "small", say, $k = O_n(1)$. Low-degree PTFs arise frequently in learning theory, for example, as hypotheses in the Low-Degree Algorithm and many other practical learning algorithms. Indeed, any function with low noise sensitivity is close to being a low-degree PTF; by combining Propositions 3.3 and 3.31 we immediately obtain:

Proposition 5.6. *Let $f : \{-1,1\}^n \to \{-1,1\}$ and let $\delta \in (0,1/2]$. Then f is $(3\mathbf{NS}_\delta[f])$-close to a PTF of degree $1/\delta$.*

For a kind of converse to this proposition, see Section 5.5.

PTFs also arise in circuit complexity, wherein a PTF representation

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{s} a_i x^{T_i}\right)$$

is thought of as a "threshold-of-parities circuit": i.e., a depth-2 circuit with s "parity gates" x^{T_i} at layer 1 and a single "(linear) threshold gate" at layer 2. From this point of view, the size of the circuit corresponds to the *sparsity* of the PTF representation:

Definition 5.7. We say a PTF representation $f(x) = \mathrm{sgn}(p(x))$ has *sparsity* at most s if $p(x)$ is a multilinear polynomial with at most s terms.

For example, the PTF representation of the 4-bit equality function from Example 5.5 has sparsity 7.

Let's extend the two theorems about LTFs we proved above to the case of PTFs. The generalization of Chow's Theorem is straightforward; its proof is left as Exercise 5.9:

Theorem 5.8. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a PTF of degree at most k and let $g : \{-1,1\}^n \to \{-1,1\}$ be any function. If $\widehat{g}(S) = \widehat{f}(S)$ for all $|S| \leq k$, then $g = f$.*

We also have the following extension of Theorem 5.2:

Theorem 5.9. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a degree-k PTF. Then $\mathbf{W}^{\leq k}[f] \geq e^{-2k}$.*

Proof. Writing $f(x) = \mathrm{sgn}(p(x))$ for p of degree k, we again have

$$\|p\|_1 = \mathbf{E}[|p(\boldsymbol{x})|] = \langle f, p\rangle = \langle f^{\leq k}, p\rangle \leq \|f^{\leq k}\|_2 \|p\|_2 = \sqrt{\mathbf{W}^{\leq k}[f]} \cdot \|p\|_2.$$

To complete the proof we need the fact that $\|p\|_2 \leq e^k \|p\|_1$ for any degree-k polynomial $p : \{-1,1\}^n \to \mathbb{R}$. We will prove this much later in Theorem 9.22 of Chapter 9 on hypercontractivity. $\square$

The e^{-2k} in this theorem cannot be improved beyond 2^{1-k}; see Exercise 5.11.

We close this section by discussing PTF sparsity. We begin with a (simpler) variant of Theorem 5.9, which is useful for proving PTF sparsity lower bounds:

Theorem 5.10. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be expressible as a PTF over the collection of monomials $\mathscr{F} \subseteq 2^{[n]}$; i.e., $f(x) = \mathrm{sgn}(p(x))$ for some polynomial $p(x) = \sum_{S \in \mathscr{F}} \widehat{p}(S) x^S$. Then $\sum_{S \in \mathscr{F}} |\widehat{f}(S)| \geq 1$.*

Proof. Define $g : \{-1,1\}^n \to \mathbb{R}$ by $g(x) = \sum_{S \in \mathscr{F}} \widehat{f}(S) x^S$. Since $\hat{\|} p \hat{\|}_\infty \leq \|p\|_1$ (Exercise 3.9) we have

$$\hat{\|} p \hat{\|}_\infty \leq \|p\|_1 = \mathbf{E}[f(\boldsymbol{x})p(\boldsymbol{x})] = \sum_{S \subseteq [n]} \widehat{f}(S)\widehat{p}(S) = \sum_{S \in \mathscr{F}} \widehat{g}(S)\widehat{p}(S) \leq \hat{\|} g \hat{\|}_1 \hat{\|} p \hat{\|}_\infty,$$

and hence $\hat{\|} g \hat{\|}_1 \geq 1$ as claimed. $\square$

We can use this result to show that the "inner product mod 2 function" (see Exercise 1.1) requires huge threshold-of-parities circuits:

Corollary 5.11. *Any PTF representation of the inner product mod 2 function* $\mathrm{IP}_{2n} : \mathbb{F}_2^{2n} \to \{-1,1\}$ *has sparsity at least* 2^n.

Proof. This follows immediately from Theorem 5.10 and the fact that $|\widehat{\mathrm{IP}}_{2n}(S)| = 2^{-n}$ for all $S \subseteq [2n]$ (Exercise 1.1). □

We can also show that any function $f : \{-1,1\}^n \to \{-1,1\}$ with small Fourier 1-norm $\hat{\lVert} f \hat{\rVert}_1$ has a sparse PTF representation. In fact a stronger result holds: such a function can be additively *approximated* by a sparse polynomial:

Theorem 5.12. *Let* $f : \{-1,1\}^n \to \mathbb{R}$ *be nonzero, let* $\delta > 0$*, and let* $s \geq 4n \hat{\lVert} f \hat{\rVert}_1^2/\delta^2$ *be an integer. Then there is a multilinear polynomial* $q : \{-1,1\}^n \to \mathbb{R}$ *of sparsity at most* s *such that* $\|f - q\|_\infty < \delta$.

Proof. The proof is by the probabilistic method. Let $\boldsymbol{T} \subseteq [n]$ be randomly chosen according to the distribution $\mathbf{Pr}[\boldsymbol{T} = T] = \frac{|\widehat{f}(T)|}{\hat{\lVert} f \hat{\rVert}_1}$. Let $\boldsymbol{T}_1, \dots, \boldsymbol{T}_s$ be independent draws from this distribution and define the multilinear polynomial

$$\boldsymbol{p}(x) = \sum_{i=1}^{s} \mathrm{sgn}(\widehat{f}(\boldsymbol{T}_i)) x^{\boldsymbol{T}_i}.$$

When $x \in \{-1,1\}^n$ is fixed, each monomial $\mathrm{sgn}(\widehat{f}(\boldsymbol{T}_i)) x^{\boldsymbol{T}_i}$ becomes a ±1-valued random variable with expectation

$$\sum_{T \subseteq [n]} \frac{|\widehat{f}(T)|}{\hat{\lVert} f \hat{\rVert}_1} \cdot \mathrm{sgn}(\widehat{f}(T)) x^T = \frac{1}{\hat{\lVert} f \hat{\rVert}_1} \sum_{T \subseteq [n]} \widehat{f}(T) x^T = \frac{f(x)}{\hat{\lVert} f \hat{\rVert}_1}.$$

Thus by a Chernoff bound, for any $\epsilon > 0$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{T}_1,\dots,\boldsymbol{T}_s}\left[\left| \boldsymbol{p}(x) - \frac{f(x)}{\hat{\lVert} f \hat{\rVert}_1} s \right| \geq \epsilon s \right] \leq 2\exp(-\epsilon^2 s/2).$$

Selecting $\epsilon = \delta / \hat{\lVert} f \hat{\rVert}_1$ and using $s \geq 4n \hat{\lVert} f \hat{\rVert}_1^2/\delta^2$, the probability is at most $2\exp(-2n) < 2^{-n}$. Taking a union bound over all 2^n choices of $x \in \{-1,1\}^n$, we conclude that there exists some $p(x) = \sum_{i=1}^{s} \mathrm{sgn}(\widehat{f}(T_i)) x^{T_i}$ such that for all $x \in \{-1,1\}^n$,

$$\left| p(x) - \frac{f(x)}{\hat{\lVert} f \hat{\rVert}_1} s \right| < \epsilon s = \frac{\delta}{\hat{\lVert} f \hat{\rVert}_1} s \quad \implies \quad \left| \frac{\hat{\lVert} f \hat{\rVert}_1}{s} \cdot p(x) - f(x) \right| < \delta.$$

Thus we may take $q = \frac{\hat{\lVert} f \hat{\rVert}_1}{s} \cdot p$. □

Corollary 5.13. *Let* $f : \{-1,1\}^n \to \{-1,1\}$. *Then* f *is expressible as a PTF of sparsity at most* $s = \lceil 4n \hat{\lVert} f \hat{\rVert}_1^2 \rceil$. *Indeed,* f *can be represented as a majority of* s *parities or negated-parities.*

Proof. Apply the previous theorem with $\delta = 1$; we then have $f(x) = \mathrm{sgn}(q(x))$. Since this is also equivalent to $\mathrm{sgn}(p(x))$, the terms $\mathrm{sgn}(\widehat{f}(T_i)) x^{T_i}$ are the required parities/negated-parities. □

Though functions computable by small DNFs need not have small Fourier 1-norm, it is a further easy corollary that they can be computed by sparse PTFs: see Exercise 5.13. We also remark that there is no good converse to Corollary 5.13: the Maj_n function has a PTF (indeed, an LTF) of sparsity n but has exponentially large Fourier 1-norm (Exercise 5.26).

5.2. Majority, and the Central Limit Theorem

Majority is one of the more important functions in Boolean analysis, and its study motivates the introduction of one of the more important tools: the Central Limit Theorem (CLT). In this section we will show how the CLT can be used to estimate the total influence and the noise stability of Maj_n. Though we already determined $\mathbf{I}[\mathrm{Maj}_n] \sim \sqrt{2/\pi}\sqrt{n}$ in Exercise 2.22 using binomial coefficients and Stirling's Formula, computations using the CLT are more flexible and extend to other linear threshold functions.

We begin with a reminder about the CLT. Suppose $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ are independent random variables and $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$. Roughly speaking, the CLT says that so long as no $\boldsymbol{X}_i$ is too dominant in terms of variance, the distribution of $\boldsymbol{S}$ is close to that of a Gaussian random variable with the same mean and variance. Recall:

Notation 5.14. We write $\boldsymbol{Z} \sim \mathrm{N}(0,1)$ denote that $\boldsymbol{Z}$ is a standard Gaussian random variable. We use the notation

$$\varphi(z) = \tfrac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(t) = \int_{-\infty}^{t} \phi(z)\,dz, \quad \overline{\Phi}(t) = \Phi(-t) = \int_{t}^{\infty} \phi(z)\,dz$$

for the pdf, cdf, and complementary cdf of this random variable. More generally, if $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is a positive semidefinite matrix, we write $\boldsymbol{Z} \sim \mathrm{N}(\mu,\Sigma)$ to denote that $\boldsymbol{Z}$ is a d-dimensional random vector with mean μ and covariance matrix Σ.

We give a precise statement of the CLT below in the form of the *Berry–Esseen Theorem*. The CLT also extends to the *multidimensional* case (sums of independent random vectors); we give a precise statement in Exercise 5.33. In Chapter 11 we will show one way to prove such CLTs.

Let's see how we can use the CLT to obtain the estimate $\mathbf{I}[\mathrm{Maj}_n] \sim \sqrt{2/\pi}\sqrt{n}$. Recall the proof of Theorem 2.33, which shows that Maj_n maximizes $\sum_{i=1}^n \widehat{f}(i)$ among all $f : \{-1,1\}^n \to \{-1,1\}$. In it we saw that

$$\mathbf{I}[\mathrm{Maj}_n] = \sum_{i=1}^{n} \widehat{\mathrm{Maj}_n}(i) = \mathop{\mathbf{E}}_{\boldsymbol{x}}[\mathrm{Maj}_n(\boldsymbol{x})(\sum_i \boldsymbol{x}_i)] = \mathop{\mathbf{E}}_{\boldsymbol{x}}[|\sum_i \boldsymbol{x}_i|]. \tag{5.2}$$

When using the CLT, it's convenient to define majority (equivalently) as

$$\mathrm{Maj}_n(x) = \mathrm{sgn}\Big(\sum_{i=1}^{n} \tfrac{1}{\sqrt{n}} x_i\Big).$$

This motivates writing (5.2) as

$$\mathbf{I}[\mathrm{Maj}_n] = \sqrt{n} \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[|\sum_i \tfrac{1}{\sqrt{n}} \boldsymbol{x}_i|]. \tag{5.3}$$

If we introduce $\boldsymbol{S} = \sum_{i=1}^n \frac{1}{\sqrt{n}} \boldsymbol{x}_i$, then $\boldsymbol{S}$ has mean 0 and variance $\sum_i (1/\sqrt{n})^2 = 1$. Thus the CLT tells us that the distribution of $\boldsymbol{S}$ is close (for large n) to that of a standard Gaussian, $\boldsymbol{Z} \sim \mathrm{N}(0,1)$. So as $n \to \infty$ we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}}[|\boldsymbol{S}|] \sim \mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mathrm{N}(0,1)}[|\boldsymbol{Z}|] = 2\int_0^\infty z \cdot \tfrac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = -\sqrt{2/\pi} e^{-z^2/2} \Big|_0^\infty = \sqrt{2/\pi}, \tag{5.4}$$

which when combined with (5.3) gives us the estimate $\mathbf{I}[\mathrm{Maj}_n] \sim \sqrt{2/\pi}\sqrt{n}$.

To make this kind of estimate more precise we state the Berry–Esseen Theorem, which is a strong version of the CLT giving explicit error bounds rather than just limiting statements.

Berry–Esseen (Central Limit) Theorem. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent random variables with $\mathbf{E}[\boldsymbol{X}_i] = 0$ and $\mathbf{Var}[\boldsymbol{X}_i] = \sigma_i^2$, and assume $\sum_{i=1}^n \sigma_i^2 = 1$. Let $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{X}_i$ and let $\boldsymbol{Z} \sim \mathrm{N}(0,1)$ be a standard Gaussian. Then for all $u \in \mathbb{R}$,*

$$|\mathbf{Pr}[\boldsymbol{S} \le u] - \mathbf{Pr}[\boldsymbol{Z} \le u]| \le c\gamma,$$

where

$$\gamma = \sum_{i=1}^n \|\boldsymbol{X}_i\|_3^3$$

and c is a universal constant. (For definiteness, c = .56 is acceptable.)

Remark 5.15. If all of the $\boldsymbol{X}_i$'s satisfy $|\boldsymbol{X}_i| \le \epsilon$ with probability 1, then we can use the bound

$$\gamma = \sum_{i=1}^n \mathbf{E}[|\boldsymbol{X}_i|^3] \le \epsilon \cdot \sum_{i=1}^n \mathbf{E}[|\boldsymbol{X}_i|^2] = \epsilon \cdot \sum_{i=1}^n \sigma_i^2 = \epsilon.$$

See Exercises 5.16 and 5.17 for some additional observations.

Our most frequent use of the Berry–Esseen Theorem will be in analyzing random sums

$$\boldsymbol{S} = \sum_{i=1}^n a_i \boldsymbol{x}_i,$$

where $\boldsymbol{x} \sim \{-1,1\}^n$ and the constants $a_i \in \mathbb{R}$ are normalized so that $\sum_i a_i^2 = 1$. For majority, all of the a_i's were equal to $\frac{1}{\sqrt{n}}$. But from Remark 5.15 we see that $\boldsymbol{S}$ is close in distribution to a standard Gaussian so long as each $|a_i|$ is small. For example, in Exercise 5.31 you are asked to show the following:

Theorem 5.16. *Let $a_1, \dots, a_n \in \mathbb{R}$ satisfy $\sum_i a_i^2 = 1$ and $|a_i| \le \epsilon$ for all i. Then*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[|\sum_i a_i \boldsymbol{x}_i|] - \sqrt{2/\pi} \right| \le C\epsilon,$$

where C is a universal constant.

Theorem 5.16 justifies (5.4) with an error bound of $O(1/\sqrt{n})$, yielding the more precise estimate $\mathbf{I}[\mathrm{Maj}_n] = \sqrt{2/\pi}\sqrt{n} \pm O(1)$ (cf. Exercise 2.22, which gives an even better error bound).

Now let's turn to the noise stability of majority. Theorem 2.45 stated the formula

$$\lim_{n\to\infty} \mathbf{Stab}_\rho[\mathrm{Maj}_n] = \tfrac{2}{\pi}\arcsin\rho = 1 - \tfrac{2}{\pi}\arccos\rho. \tag{5.5}$$

Let's now spend some time justifying this using the multidimensional CLT. (For complete details, see Exercise 5.33.) By definition,

$$\mathbf{Stab}_\rho[\mathrm{Maj}_n] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[\mathrm{Maj}_n(\boldsymbol{x})\cdot\mathrm{Maj}_n(\boldsymbol{y})] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[\mathrm{sgn}(\sum_i \tfrac{1}{\sqrt{n}}\boldsymbol{x}_i)\cdot\mathrm{sgn}(\sum_i \tfrac{1}{\sqrt{n}}\boldsymbol{y}_i)]. \tag{5.6}$$

For each $i \in [n]$ let's stack $\frac{1}{\sqrt{n}}\boldsymbol{x}_i$ and $\frac{1}{\sqrt{n}}\boldsymbol{y}_i$ into a 2-dimensional vector and then write

$$\vec{\boldsymbol{S}} = \sum_{i=1}^{n} \begin{bmatrix} \frac{1}{\sqrt{n}}\boldsymbol{x}_i \\ \frac{1}{\sqrt{n}}\boldsymbol{y}_i \end{bmatrix} \in \mathbb{R}^2. \tag{5.7}$$

We are summing n independent random vectors, so the multidimensional CLT tells us that the distribution of $\vec{\boldsymbol{S}}$ is close to that of a 2-dimensional Gaussian $\vec{\boldsymbol{Z}}$ with the same mean and covariance matrix, namely (see Exercise 5.19)

$$\vec{\boldsymbol{Z}} \sim \mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

Continuing from (5.6),

$$\begin{aligned}\mathbf{Stab}_\rho[\mathrm{Maj}_n] &= \mathbf{E}[\mathrm{sgn}(\vec{\boldsymbol{S}}_1)\cdot\mathrm{sgn}(\vec{\boldsymbol{S}}_2)]\\ &= \mathbf{Pr}[\mathrm{sgn}(\vec{\boldsymbol{S}}_1) = \mathrm{sgn}(\vec{\boldsymbol{S}}_2)] - \mathbf{Pr}[\mathrm{sgn}(\vec{\boldsymbol{S}}_1) \neq \mathrm{sgn}(\vec{\boldsymbol{S}}_2)]\\ &= 2\,\mathbf{Pr}[\mathrm{sgn}(\vec{\boldsymbol{S}}_1) = \mathrm{sgn}(\vec{\boldsymbol{S}}_2)] - 1 = 4\,\mathbf{Pr}[\vec{\boldsymbol{S}} \in Q_{--}] - 1,\end{aligned}$$

where Q_{--} denotes the lower-left quadrant of $\mathbb{R}^2$ and the last step uses the symmetry $\mathbf{Pr}[\vec{\boldsymbol{S}} \in Q_{++}] = \mathbf{Pr}[\vec{\boldsymbol{S}} \in Q_{--}]$. Since Q_{--} is convex, the 2-dimensional CLT lets us deduce

$$\lim_{n\to\infty} \mathbf{Pr}[\vec{\boldsymbol{S}} \in Q_{--}] = \mathbf{Pr}[\vec{\boldsymbol{Z}} \in Q_{--}].$$

So to justify the noise stability formula (5.5) for majority, it remains to verify

$$4\,\mathbf{Pr}[\vec{\boldsymbol{Z}} \in Q_{--}] - 1 = 1 - \tfrac{2}{\pi}\arccos\rho \quad\iff\quad \mathbf{Pr}[\vec{\boldsymbol{Z}} \in Q_{--}] = \frac{1}{2} - \frac{1}{2}\frac{\arccos\rho}{\pi}.$$

And this in turn is a 19th-century identity known as *Sheppard's Formula*:

Sheppard's Formula. *Let $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ be standard Gaussian random variables with correlation $\mathbf{E}[\boldsymbol{z}_1\boldsymbol{z}_2] = \rho \in [-1,1]$. Then*

$$\mathbf{Pr}[\boldsymbol{z}_1 \le 0, \boldsymbol{z}_2 \le 0] = \frac{1}{2} - \frac{1}{2}\frac{\arccos\rho}{\pi}.$$

Proving Sheppard's Formula is a nice exercise using the rotational symmetry of a pair of independent standard Gaussians; we defer the proof till Example 11.19 in Chapter 11.1. This completes the justification of formula (5.5) for the limiting noise stability of majority.

You may have noticed that once we applied the 2-dimensional CLT to (5.6), the remainder of the derivation had nothing to do with majority. In fact, the same analysis works for *any* linear threshold function $\mathrm{sgn}(a_1x_1+\cdots+a_nx_n)$, the only difference being the "error term" arising from the CLT. As in Theorem 5.16, this error is small so long as no coefficient a_i is too dominant:

Theorem 5.17. *Let $f:\{-1,1\}^n \to \{-1,1\}$ be an unbiased LTF, $f(x)=\mathrm{sgn}(a_1x_1+\cdots+a_nx_n)$ with $\sum_i a_i^2 = 1$ and $|a_i| \le \epsilon$ for all i. Then for any $\rho \in (-1,1)$,*

$$\left|\mathbf{Stab}_\rho[f] - \tfrac{2}{\pi}\arcsin\rho\right| \le O\left(\tfrac{\epsilon}{\sqrt{1-\rho^2}}\right).$$

You are asked to prove Theorem 5.17 in Exercise 5.33. In the particular case of Maj_n where $a_i = \frac{1}{\sqrt{n}}$ for all i we can make a slightly stronger claim (see Exercise 5.23):

Theorem 5.18. *For any $\rho \in [0,1)$, $\mathbf{Stab}_\rho[\mathrm{Maj}_n]$ is a decreasing function of n, with*

$$\tfrac{2}{\pi}\arcsin\rho \le \mathbf{Stab}_\rho[\mathrm{Maj}_n] \le \tfrac{2}{\pi}\arcsin\rho + O\left(\tfrac{1}{\sqrt{1-\rho^2}\sqrt{n}}\right).$$

We end this section by mentioning another way in which the majority function is extremal: among all unbiased functions with small influences, it has (essentially) the largest noise stability.

Majority Is Stablest Theorem. *Fix $\rho \in (0,1)$. Then for any $f:\{-1,1\}^n \to [-1,1]$ with $\mathbf{E}[f]=0$ and $\mathbf{MaxInf}[f] \le \tau$,*

$$\mathbf{Stab}_\rho[f] \le \tfrac{2}{\pi}\arcsin\rho + o_\tau(1) = 1 - \tfrac{2}{\pi}\arccos\rho + o_\tau(1).$$

For sufficiently small ρ, we'll prove this in Section 5.4. The proof of the full Majority Is Stablest Theorem will have to wait until Chapter 11.

5.3. The Fourier coefficients of Majority

In this section we will analyze the Fourier coefficients of Maj_n. In fact, we give an explicit formula for them in Theorem 5.19 below. But most of the time this formula is not too useful; instead, it's better to understand the Fourier coefficients of Maj_n asymptotically as $n \to \infty$.

Let's begin with a few basic observations. First, Maj_n is a symmetric function and hence $\widehat{\mathrm{Maj}_n}(S)$ only depends on $|S|$ (Exercise 1.30). Second, Maj_n is an odd function and hence $\widehat{\mathrm{Maj}_n}(S)=0$ whenever $|S|$ is even (Exercise 1.8).

It remains to determine the Fourier coefficients $\widehat{\mathrm{Maj}_n}(S)$ for $|S|$ odd. By symmetry, $\widehat{\mathrm{Maj}_n}(S)^2 = \mathbf{W}^k[\mathrm{Maj}_n]/\binom{n}{k}$ for all $|S| = k$, so if we are content to know the magnitudes of Maj_n's Fourier coefficients, it suffices to determine the quantities $\mathbf{W}^k(\mathrm{Maj}_n)$.

In fact, for each $k \in \mathbb{N}$ the quantity $\mathbf{W}^k(\mathrm{Maj}_n)$ converges to a fixed constant as $n \to \infty$. We can deduce this using our analysis of the noise stability of majority. From the previous section we know that for all $|\rho| \leq 1$,

$$\lim_{n\to\infty} \mathbf{Stab}_\rho[\mathrm{Maj}_n] = \tfrac{2}{\pi}\arcsin\rho = \tfrac{2}{\pi}\left(\rho + \tfrac{1}{6}\rho^3 + \tfrac{3}{40}\rho^5 + \tfrac{5}{112}\rho^7 + \cdots\right), \tag{5.8}$$

where we have used the power series for arcsin,

$$\arcsin z = \sum_{k \text{ odd}} \frac{2}{k2^k}\binom{k-1}{\frac{k-1}{2}} \cdot z^k, \tag{5.9}$$

valid for $|\rho| \leq 1$ (see Exercise 5.18). Comparing (5.8) with the formula

$$\mathbf{Stab}_\rho[\mathrm{Maj}_n] = \sum_{k\geq 0} \mathbf{W}^k[\mathrm{Maj}_n] \cdot \rho^k$$

suggests the following: For each fixed $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbf{W}^k[\mathrm{Maj}_n] = [\rho^k](\tfrac{2}{\pi}\arcsin\rho) = \begin{cases} \frac{4}{\pi k 2^k}\binom{k-1}{\frac{k-1}{2}} & \text{if } k \text{ odd}, \\ 0 & \text{if } k \text{ even}. \end{cases} \tag{5.10}$$

(Here $[z^k]F(z)$ denotes the coefficient on z^k in power series $F(z)$.) Indeed, we prove this identity below in Theorem 5.22. The noise stability method that suggests it can also be made formal (Exercise 5.25).

Identity (5.10) is one way to formulate precisely the statement that the "Fourier spectrum of Maj_n converges". Introducing notation such as "$\mathbf{W}^k(\mathrm{Maj})$" for the quantity in (5.10), we have the further asymptotics

$$\begin{aligned} \text{for } k \text{ odd}, \qquad & \mathbf{W}^k(\mathrm{Maj}) \sim \left(\tfrac{2}{\pi}\right)^{3/2} k^{-3/2}, \\ & \mathbf{W}^{>k}(\mathrm{Maj}) \sim \left(\tfrac{2}{\pi}\right)^{3/2} k^{-1/2} \qquad \text{as } k \to \infty. \end{aligned} \tag{5.11}$$

(See Exercise 5.27.) The estimates (5.11), together with the precise value $\mathbf{W}^1(\mathrm{Maj}) = \frac{2}{\pi}$, are usually all you need to know about the Fourier coefficients of majority.

Nevertheless, let's now compute the Fourier coefficients of Maj_n exactly.

Theorem 5.19. *If $|S|$ is even, then $\widehat{\mathrm{Maj}_n}(S) = 0$. If $|S| = k$ is odd,*

$$\widehat{\mathrm{Maj}_n}(S) = (-1)^{\frac{k-1}{2}} \frac{\binom{\frac{n-1}{2}}{\frac{k-1}{2}}}{\binom{n-1}{k-1}} \cdot \frac{2}{2^n}\binom{n-1}{\frac{n-1}{2}}.$$

Proof. The first statement holds because Maj_n is an odd function; henceforth we assume $|S| = k$ is odd. The trick will be to compute the Fourier expansion of majority's *derivative* $\mathrm{D}_n\mathrm{Maj}_n = \mathrm{Half}_{n-1} : \{-1,1\}^{n-1} \to \{0,1\}$, the 0-1 indicator of the set of $(n-1)$-bit strings with exactly half of their coordinates equal to -1. By the derivative formula and the fact that Maj_n is symmetric, $\widehat{\mathrm{Maj}_n}(S) = \widehat{\mathrm{Half}_{n-1}}(T)$ for any $T \subseteq [n-1]$ with $|T| = k-1$. So writing $n-1 = 2m$ and $k-1 = 2j$, it suffices to show

$$\widehat{\mathrm{Half}_{2m}}([2j]) = (-1)^j \frac{\binom{m}{j}}{\binom{2m}{2j}} \cdot \tfrac{1}{2^{2m}}\binom{2m}{m}. \tag{5.12}$$

By the probabilistic definition of T_ρ, for any $\rho \in [-1,1]$ we have

$$\mathrm{T}_\rho\mathrm{Half}_{2m}(1,1,\dots,1) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_\rho((1,1,\dots,1))}[\mathrm{Half}_{2m}(\boldsymbol{x})] = \mathbf{Pr}[\boldsymbol{x} \text{ has } m \text{ 1's and } m\ -1\text{'s}],$$

where each coordinate of $\boldsymbol{x}$ is 1 with probability $\frac{1}{2}+\frac{1}{2}\rho$. Thus

$$\mathrm{T}_\rho\mathrm{Half}_{2m}(1,1,\dots,1) = \binom{2m}{m}(\tfrac{1}{2}+\tfrac{1}{2}\rho)^m(\tfrac{1}{2}-\tfrac{1}{2}\rho)^m = \tfrac{1}{2^{2m}}\binom{2m}{m}(1-\rho^2)^m. \tag{5.13}$$

On the other hand, by the Fourier formula for T_ρ and the fact that Half_{2m} is symmetric we have

$$\mathrm{T}_\rho\mathrm{Half}_{2m}(1,1,\dots,1) = \sum_{U\subseteq[2m]} \widehat{\mathrm{Half}_{2m}}(U)\rho^{|U|} = \sum_{i=0}^{2m}\binom{2m}{i}\widehat{\mathrm{Half}_{2m}}([i])\rho^i. \tag{5.14}$$

Since we have equality (5.13) = (5.14) between two degree-$2m$ polynomials of ρ on all of $[-1,1]$, we can equate coefficients. In particular, for $i = 2j$ we have

$$\binom{2m}{2j}\widehat{\mathrm{Half}_{2m}}([2j]) = \tfrac{1}{2^{2m}}\binom{2m}{m}\cdot[\rho^{2j}](1-\rho^2)^m = \tfrac{1}{2^{2m}}\binom{2m}{m}\cdot(-1)^j\binom{m}{j},$$

confirming (5.12). □

You are asked to prove the following corollaries in Exercises 5.20, 5.22:

Corollary 5.20. $\widehat{\mathrm{Maj}_n}(S) = \widehat{\mathrm{Maj}_n}(T)$ *whenever* $|S|+|T| = n+1$. *Hence also* $\mathbf{W}^{n-k+1}[\mathrm{Maj}_n] = \frac{k}{n-k+1}\mathbf{W}^k[\mathrm{Maj}_n]$.

Corollary 5.21. *For any odd k, $\mathbf{W}^k[\mathrm{Maj}_n]$ is a strictly decreasing function of n (for $n \geq k$ odd).*

We can now prove the identity (5.10):

Theorem 5.22. *For each fixed odd k,*

$$\mathbf{W}^k[\mathrm{Maj}_n] \searrow [\rho^k](\tfrac{2}{\pi}\arcsin\rho) = \tfrac{4}{\pi k 2^k}\binom{k-1}{\frac{k-1}{2}}$$

as $n \geq k$ tends to ∞ (through the odd numbers). Further, we have the error bound

$$[\rho^k](\tfrac{2}{\pi}\arcsin\rho) \leq \mathbf{W}^k[\mathrm{Maj}_n] \leq (1+2k/n)\cdot[\rho^k](\tfrac{2}{\pi}\arcsin\rho) \tag{5.15}$$

for all $k < n/2$. (For $k > n/2$ you can use Corollary 5.20.)

Proof. Corollary 5.21 tells us that $\mathbf{W}^k[\mathrm{Maj}_n]$ is decreasing in n; hence we only need to justify (5.15). Using the formula from Theorem 5.19 we have

$$\frac{\mathbf{W}^k[\mathrm{Maj}_n]}{[\rho^k](\frac{2}{\pi}\arcsin\rho)} = \frac{\binom{n}{k}\frac{4}{2^{2n}}\binom{n-1}{\frac{n-1}{2}}^2 \binom{\frac{n-1}{2}}{\frac{k-1}{2}}^2 \Big/ \binom{n-1}{k-1}^2}{\frac{4}{\pi k 2^k}\binom{k-1}{\frac{k-1}{2}}} = \tfrac{\pi}{2}n\cdot 2^{k-n}\binom{n-k}{\frac{n-k}{2}}\cdot 2^{1-n}\binom{n-1}{\frac{n-1}{2}},$$

where the second identity is verified by expanding all binomial coefficients to factorials. By Stirling's approximation we have $2^{-m}\binom{m}{m/2} \nearrow \sqrt{\frac{2}{\pi m}}$, meaning that the ratio of the left side to the right side increases to 1 as $m\to\infty$. Thus

$$\frac{\mathbf{W}^k[\mathrm{Maj}_n]}{[\rho^k](\frac{2}{\pi}\arcsin\rho)} \nearrow \frac{n}{\sqrt{n-k}\sqrt{n-1}} = (1-\tfrac{k+1}{n}+\tfrac{k}{n^2})^{-1/2},$$

and the right-hand side is at most $1+2k/n$ for $1\le k\le n/2$ by Exercise 5.24. □

Finally, we can deduce the asymptotics (5.11) from this theorem (see Exercise 5.27):

Corollary 5.23. *Let $k\in\mathbb{N}$ be odd and assume $n = n(k)\ge 2k^2$. Then*

$$\mathbf{W}^k(\mathrm{Maj}_n) = \left(\tfrac{2}{\pi}\right)^{3/2} k^{-3/2}\cdot(1\pm O(1/k)),$$
$$\mathbf{W}^{>k}(\mathrm{Maj}_n) = \left(\tfrac{2}{\pi}\right)^{3/2} k^{-1/2}\cdot(1\pm O(1/k)),$$

and hence the Fourier spectrum of Maj_n *is ϵ-concentrated on degree up to* $\frac{8}{\pi^3}\epsilon^{-2}+O_\epsilon(1)$.

5.4. Degree-1 weight

In this section we prove two theorems about the degree-1 Fourier weight of Boolean functions:

$$\mathbf{W}^1[f] = \sum_{i=1}^n \widehat{f}(i)^2.$$

This important quantity can be given a combinatorial interpretation thanks to the noise stability formula $\mathbf{Stab}_\rho[f] = \sum_{k\ge 0}\rho^k\cdot\mathbf{W}^k[f]$:

$$\text{For } f:\{-1,1\}^n\to\mathbb{R},\quad \mathbf{W}^1[f] = \frac{d}{d\rho}\mathbf{Stab}_\rho[f]\Big|_{\rho=0}.$$

Thinking of $\|f\|_2$ as constant and $\rho\to 0$, the noise stability formula implies

$$\mathbf{Stab}_\rho[f] = \mathbf{E}[f]^2 + \mathbf{W}^1[f]\rho \pm O(\rho^2),$$

or equivalently,

$$\underset{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}{\mathbf{Cov}}[f(\boldsymbol{x}),f(\boldsymbol{y})] = \mathbf{W}^1[f]\rho \pm O(\rho^2).$$

In other words, for $f : \{-1,1\}^n \to \{-1,1\}$ the degree-1 weight quantifies the extent to which $\mathbf{Pr}[f(\boldsymbol{x}) = f(\boldsymbol{y})]$ increases when $\boldsymbol{x}$ and $\boldsymbol{y}$ go from being uncorrelated to being slightly correlated.

There is an additional viewpoint if we think of f as the indicator of a subset $A \subseteq \{-1,1\}^n$ and its noise sensitivity $\mathbf{NS}_\delta[f]$ as a notion of A's "surface area", or "noisy boundary size". For nearly maximal noise rates – i.e., $\delta = \frac{1}{2} - \frac{1}{2}\rho$ where ρ is small – we have that A's noisy boundary size is "small" if and only if $\mathbf{W}^1[f]$ is "large" (vis-à-vis A's measure).

Two examples suggest themselves when thinking of subsets of the Hamming cube with small "boundary": subcubes and Hamming balls.

Proposition 5.24. *Let $f : \mathbb{F}_2^n \to \{0,1\}$ be the indicator of a subcube of codimension $k \geq 1$ (e.g., the AND_k function). Then $\mathbf{E}[f] = 2^{-k}$, $\mathbf{W}^1[f] = k2^{-2k}$.*

Proposition 5.25. *Fix $t \in \mathbb{R}$. Consider the sequence of LTFs $f_n : \{-1,1\}^n \to \{0,1\}$ defined by $f_n(x) = 1$ if and only if $\sum_{i=1}^n \frac{1}{\sqrt{n}} x_i > t$. (That is, f_n is the indicator of the Hamming ball $\{x : \Delta(x,(1,\dots,1)) < \frac{n}{2} - \frac{t}{2}\sqrt{n}\}$.) Then*

$$\lim_{n\to\infty} \mathbf{E}[f_n] = \overline{\Phi}(t), \qquad \lim_{n\to\infty} \mathbf{W}^1[f_n] = \phi(t)^2.$$

You are asked to verify these facts in Exercises 5.29, 5.30. Regarding Proposition 5.25, it's natural for $\phi(t)$ to arise since $\mathbf{W}^1[f_n]$ is related to the influences of f_n, and coordinates are influential for f_n if and only if $\sum_{i=1}^n \frac{1}{\sqrt{n}} x_i \approx t$. If we write $\alpha = \lim_{n\to\infty} \mathbf{E}[f_n]$ then this proposition can be thought of as saying that $\mathbf{W}^1[f_n] \to \mathcal{U}(\alpha)^2$, where $\mathcal{U}$ is defined as follows:

Definition 5.26. The *Gaussian isoperimetric function* $\mathcal{U} : [0,1] \to [0, \frac{1}{\sqrt{2\pi}}]$ is defined by $\mathcal{U} = \phi \circ \Phi^{-1}$. This function is symmetric about 1/2; i.e., $\mathcal{U} = \phi \circ \overline{\Phi}^{-1}$.

The name of this function will be explained when we study the Gaussian Isoperimetric Inequality in Chapter 11.4. For now we'll just use the following fact:

Proposition 5.27. *For $\alpha \to 0^+$, $\mathcal{U}(\alpha) \sim \alpha\sqrt{2\ln(1/\alpha)}$.*

Proof. Write $\alpha = \overline{\Phi}(t)$, where $t \to \infty$. We use the well-known fact that $\overline{\Phi}(t) \sim \phi(t)/t$. Thus

$$\alpha \sim \frac{1}{\sqrt{2\pi}t}\exp(-t^2/2) \quad \Longrightarrow \quad t \sim \sqrt{2\ln(1/\alpha)},$$
$$\phi(t) \sim \overline{\Phi}(t)\cdot t \quad \Longrightarrow \quad \mathcal{U}(\alpha) \sim \alpha \cdot t \sim \alpha\sqrt{2\ln(1/\alpha)}. \qquad \square$$

Given Propositions 5.24 and 5.25, let's consider the degree-1 Fourier weight of subcubes and Hamming balls asymptotically as their "volume" $\alpha = \mathbf{E}[f]$ tends to 0. For the subcubes we have $\mathbf{W}^1[f] = \alpha^2 \log(1/\alpha)$. For the

Hamming balls we have $\mathbf{W}^1[f_n] \to \mathscr{U}(\alpha)^2 \sim 2\alpha^2 \ln(1/\alpha)$. So in both cases we have an upper bound of $O(\alpha^2 \log(1/\alpha))$.

You should think of this upper bound $O(\alpha^2 \log(1/\alpha))$ as being unusually small. The obvious a priori upper bound, given that $f : \{-1,1\}^n \to \{0,1\}$ has $\mathbf{E}[f] = \alpha$, is

$$\mathbf{W}^1[f] \leq \mathbf{Var}[f] = \alpha(1-\alpha) \sim \alpha.$$

Yet subcubes and Hamming balls have degree-1 weight which is almost quadratically smaller. In fact the first theorem we will show in this section is the following:

Level-1 Inequality. *Let $f : \{-1,1\}^n \to \{0,1\}$ have mean $\mathbf{E}[f] = \alpha \leq 1/2$. Then*

$$\mathbf{W}^1[f] \leq O(\alpha^2 \log(1/\alpha)).$$

(For the case $\alpha \geq 1/2$, replace f by $1-f$.)

Thus *all* small subsets of $\{-1,1\}^n$ have unusually small $\mathbf{W}^1[f]$; or equivalently (in some sense), unusually large "noisy boundary". This is another key illustration of the idea that the Hamming cube is a "small-set expander".

Remark 5.28. The bound in the Level-1 Inequality has a sharp form, $\mathbf{W}^1[f] \leq 2\alpha^2 \ln(1/\alpha)$. Thus Hamming balls are in fact the "asymptotic maximizers" of $\mathbf{W}^1[f]$ among sets of small volume α. Also, the inequality holds more generally for $f : \{-1,1\}^n \to [-1,1]$ with $\alpha = \mathbf{E}[|f|]$.

Remark 5.29. The name "Level-1 Inequality" is not completely standard; e.g., in additive combinatorics the result would be called *Chang's Inequality*. We use this name because we will also generalize to "Level-k Inequalities" in Chapter 9.5.

So far we considered maximizing degree-1 weight among subsets of the Hamming cube of a fixed small volume, α. The second theorem in this section is concerned with what happens when there is no volume constraint. In this case, maximizing examples tend to have volume $\alpha = 1/2$; switching the notation to $f : \{-1,1\}^n \to \{-1,1\}$, this corresponds to f being unbiased ($\mathbf{E}[f] = 0$). The unbiased Hamming ball is Maj_n, which we know has $\mathbf{W}^1[\mathrm{Maj}_n] \to \frac{2}{\pi}$. This is quite large. But unbiased subcubes are just the dictators χ_i and their negations; these have $\mathbf{W}^1[\pm\chi_i] = 1$ which is obviously maximal.

Thus the question of which $f : \{-1,1\}^n \to \{-1,1\}$ maximizes $\mathbf{W}^1[f]$ has a trivial answer. But this answer is arguably unsatisfactory, since dictators (and their negations) are not "really" functions of n bits. Indeed, when we studied social choice in Chapter 2 we were motivated to rule out functions f having a coordinate with unfairly large influence. And in fact Proposition 2.58 showed that if all $\widehat{f}(i)$ are equal (and hence small) then $\mathbf{W}^1[f] \leq \frac{2}{\pi} + o_n(1)$. The second theorem of this section significantly generalizes Proposition 2.58:

The $\frac{2}{\pi}$ Theorem. *Let $f : \{-1,1\}^n \to \{-1,1\}$ satisfy $|\widehat{f}(i)| \leq \epsilon$ for all $i \in [n]$. Then*

$$\mathbf{W}^1[f] \leq \tfrac{2}{\pi} + O(\epsilon). \tag{5.16}$$

Further, if $\mathbf{W}^1[f] \geq \frac{2}{\pi} - \epsilon$, then f is $O(\sqrt{\epsilon})$-close to the LTF $\mathrm{sgn}(f^{=1})$.

Functions f with $|\widehat{f}(i)| \leq \epsilon$ for all $i \in [n]$ are called $(\epsilon,1)$-*regular*; see Chapter 6.1. So the $\frac{2}{\pi}$ Theorem says (roughly speaking) that within the class of $(\epsilon,1)$-regular functions, the maximal degree-1 weight is $\frac{2}{\pi}$, and any function achieving this is an unbiased LTF. Further, from Theorem 5.17 we know that *all* unbiased LTFs which are $(\epsilon,1)$-regular achieve this.

Remark 5.30. Since we have $\mathbf{Stab}_\rho[f] \approx \mathbf{W}^1[f]\rho$ and $\frac{2}{\pi}\arcsin\rho \approx \frac{2}{\pi}\rho$ when ρ is small, the $\frac{2}{\pi}$ Theorem gives the Majority Is Stablest Theorem in the limit $\rho \to 0^+$.

Let's now discuss how we'll prove our two theorems about degree-1 weight. Let $f : \{-1,1\}^n \to \{0,1\}$ and $\alpha = \mathbf{E}[f]$; we think of α as small for the Level-1 Inequality and $\alpha = 1/2$ for the $\frac{2}{\pi}$ Theorem. By Plancherel, $\mathbf{W}^1[f] = \mathbf{E}[f(\boldsymbol{x})L(\boldsymbol{x})]$, where

$$L(x) = f^{=1}(x) = \widehat{f}(1)x_1 + \cdots + \widehat{f}(n)x_n.$$

To upper-bound $\mathbf{E}[f(\boldsymbol{x})L(\boldsymbol{x})]$, consider that as $\boldsymbol{x}$ varies the real number $L(\boldsymbol{x})$ may be rather large or small, but $f(\boldsymbol{x})$ is always 0 or 1. Given that $f(\boldsymbol{x})$ is 1 on only a α fraction of $\boldsymbol{x}$'s, the "worst case" for $\mathbf{E}[f(\boldsymbol{x})L(\boldsymbol{x})]$ would be if $f(x)$ were 1 precisely on the α fraction of x's where $L(x)$ is largest. In other words,

$$\mathbf{W}^1[f] = \mathbf{E}[f(\boldsymbol{x})L(\boldsymbol{x})] \leq \mathbf{E}[\mathbf{1}_{\{L(\boldsymbol{x}) \geq t\}} \cdot L(\boldsymbol{x})], \tag{5.17}$$

where t is chosen so that

$$\mathbf{Pr}[L(\boldsymbol{x}) \geq t] \approx \alpha. \tag{5.18}$$

But now we can analyze (5.17) quite effectively using tools such as Hoeffding's bound and the CLT, since $L(\boldsymbol{x})$ is just a linear combination of independent ±1 random bits. In particular $L(\boldsymbol{x})$ has mean 0 and standard deviation $\sigma = \sqrt{\mathbf{W}^1[f]}$ so by the CLT it acts like the Gaussian $\boldsymbol{Z} \sim \mathrm{N}(0,\sigma^2)$, at least if we assume all $|\widehat{f}(i)|$ are small. If we are thinking of $\alpha = 1/2$, then $t = 0$ and we get

$$\sigma^2 = \mathbf{W}^1[f] \leq \mathbf{E}[\mathbf{1}_{\{L(\boldsymbol{x}) \geq 0\}} \cdot L(\boldsymbol{x})] \approx \mathbf{E}[\mathbf{1}_{\{\boldsymbol{Z} \geq 0\}} \cdot \boldsymbol{Z}] = \tfrac{1}{\sqrt{2\pi}}\sigma;$$

This implies $\sigma^2 \lessapprox \frac{1}{2\pi}$, as claimed in the $\frac{2}{\pi}$ Theorem (after adjusting f's range to $\{-1,1\}$). If we are instead thinking of α as small then (5.18) suggest taking $t \sim \sigma\sqrt{2\ln(1/\alpha)}$ so that $\mathbf{Pr}[\boldsymbol{Z} \geq t] \approx \alpha$. Then a calculation akin to the one in Proposition 5.27 implies

$$\mathbf{W}^1[f] \leq \mathbf{E}[\mathbf{1}_{\{L(\boldsymbol{x}) \geq t\}} \cdot L(\boldsymbol{x})] \approx \alpha \cdot \sigma\sqrt{2\ln(1/\alpha)},$$

from which the Level-1 Inequality follows. In fact, we don't even need all $|\widehat{f}(i)|$ small for this latter analysis; for large t it's possible to upper-bound (5.17) using only Hoeffding's bound:

Lemma 5.31. *Let $\ell(x) = a_1x_1 + \cdots + a_nx_n$, where $\sum_i a_i^2 = 1$. Then for any $s \geq 1$,*

$$\mathbf{E}[\mathbf{1}_{\{|\ell(\boldsymbol{x})|>s\}} \cdot |\ell(\boldsymbol{x})|] \leq (2s+2)\exp(-\tfrac{s^2}{2}).$$

Proof. We have

$$\begin{aligned}\mathbf{E}[\mathbf{1}_{\{|\ell(\boldsymbol{x})|>s\}} \cdot |\ell(\boldsymbol{x})|] &= s\,\mathbf{Pr}[|\ell(\boldsymbol{x})| > s] + \int_s^\infty \mathbf{Pr}[|\ell(\boldsymbol{x})| > u]\,du \\ &\leq 2s\exp(-\tfrac{s^2}{2}) + \int_s^\infty 2\exp(-\tfrac{u^2}{2})\,du,\end{aligned}$$

using Hoeffding's bound. But for $s \geq 1$,

$$\int_s^\infty 2\exp(-\tfrac{u^2}{2})\,du \leq \int_s^\infty u \cdot 2\exp(-\tfrac{u^2}{2})\,du = 2\exp(-\tfrac{s^2}{2}). \qquad \square$$

We now give formal proofs of the two theorems, commenting that rather than $L(x)$ it's more convenient to work with

$$\ell(x) = \tfrac{1}{\sigma}f^{=1}(x) = \tfrac{\widehat{f}(1)}{\sigma}x_1 + \cdots + \tfrac{\widehat{f}(n)}{\sigma}x_n.$$

Proof of the Level-1 Inequality. Following Remark 5.28 we let $f : \{-1,1\}^n \to [-1,1]$ and $\alpha = \mathbf{E}[|f|]$. We may assume $\sigma = \sqrt{\mathbf{W}^1[f]} > 0$. Writing $\ell = \frac{1}{\sigma}f^{=1}$ we have $\langle f, \ell\rangle = \frac{1}{\sigma}\langle f, f^{=1}\rangle = \frac{1}{\sigma}\mathbf{W}^1[f] = \sigma$ and hence

$$\sigma = \langle f, \ell\rangle = \mathbf{E}[\mathbf{1}_{\{|\ell(\boldsymbol{x})|\leq s\}} \cdot f(\boldsymbol{x})\ell(\boldsymbol{x})] + \mathbf{E}[\mathbf{1}_{\{|\ell(\boldsymbol{x})|>s\}} \cdot f(\boldsymbol{x})\ell(\boldsymbol{x})]$$

holds for any $s \geq 1$. The first expectation above is at most $\mathbf{E}[s|f(\boldsymbol{x})|] = \alpha s$, and the second is at most $(2+2s)\exp(-s^2/2) \leq 4s\exp(-s^2/2)$ by Lemma 5.31. Hence

$$\sigma \leq \alpha s + 4s\exp(-s^2/2).$$

The optimal choice of s is $s = (\sqrt{2} + o_\alpha(1))\sqrt{\ln(1/\alpha)}$, yielding

$$\sigma \leq (\sqrt{2} + o(1))\alpha\sqrt{\ln(1/\alpha)}.$$

Squaring this establishes the claim $\sigma^2 \leq (2 + o_\alpha(1))\alpha^2\ln(1/\alpha)$. $\square$

Proof of the $\frac{2}{\pi}$ Theorem. We may assume $\sigma = \sqrt{\mathbf{W}^1[f]} \geq 1/2$: for the theorem's first statement this is because otherwise there is nothing to prove; for the theorem's second statement this is because we may assume ϵ sufficiently small.

We start by proving (5.16). Let $\ell = \frac{1}{\sigma}f^{=1}$, so $\|\ell\|_2 = 1$ and $|\widehat{\ell}(i)| \leq 2\epsilon$ for all $i \in [n]$. We have

$$\sigma = \langle f, \ell\rangle \leq \mathbf{E}[|\ell|] \leq \sqrt{\tfrac{2}{\pi}} + C\epsilon \tag{5.19}$$

for some constant C, where we used Theorem 5.16. Squaring this proves (5.16). We observe that (5.16) therefore holds even for $f : \{-1,1\}^n \to [-1,1]$.

Now suppose we also have $\mathbf{W}^1[f] \geq \frac{2}{\pi} - \epsilon$; i.e.,

$$\sigma \geq \sqrt{\tfrac{2}{\pi} - \epsilon} \geq \sqrt{\tfrac{2}{\pi}} - 2\epsilon.$$

Thus the first inequality in (5.19) must be close to tight; specifically,

$$(C+2)\epsilon \geq \mathbf{E}[|\ell|] - \langle f, \ell \rangle = \mathbf{E}[(\operatorname{sgn}(\ell(\boldsymbol{x})) - f(\boldsymbol{x}))\cdot \ell(\boldsymbol{x})]. \qquad (5.20)$$

By the Berry–Esseen Theorem (and Remark 5.15, Exercise 5.16),

$$\mathbf{Pr}[|\ell| \leq K\sqrt{\epsilon}] \leq \mathbf{Pr}[|\mathrm{N}(0,1)| \leq K\sqrt{\epsilon}] + .56\cdot 2\epsilon \leq \tfrac{1}{\sqrt{2\pi}}\cdot 2K\sqrt{\epsilon} + 1.12\epsilon \leq 2K\sqrt{\epsilon}$$

for any constant $K \geq 1$. We therefore have the implication

$$\begin{aligned}\mathbf{Pr}[f \neq \operatorname{sgn}(\ell)] \geq 3K\sqrt{\epsilon} &\implies \mathbf{Pr}[f(\boldsymbol{x}) \neq \operatorname{sgn}(\ell(\boldsymbol{x})) \;\wedge\; |\ell(\boldsymbol{x})| > K\sqrt{\epsilon}] \geq K\sqrt{\epsilon} \\ &\implies \mathbf{E}[(\operatorname{sgn}(\ell(\boldsymbol{x})) - f(\boldsymbol{x}))\cdot \ell(\boldsymbol{x})] \geq K\sqrt{\epsilon}\cdot 2(K\sqrt{\epsilon}) = 2K^2\epsilon.\end{aligned}$$

This contradicts (5.20) for $K = \sqrt{C+2}$, say. Thus $\mathbf{Pr}[f \neq \operatorname{sgn}(\ell)] \leq 3\sqrt{C+2}\sqrt{\epsilon}$, completing the proof. □

For an interpolation between these two theorems, see Exercise 5.44.

We conclude this section with an application of the Level-1 Inequality. First, a quick corollary which we leave for Exercise 5.37:

Corollary 5.32. *Let $f : \{-1,1\}^n \to \{-1,1\}$ have $|\mathbf{E}[f]| \geq 1-\delta \geq 0$. Then $\mathbf{W}^1[f] \leq 4\delta^2\log(2/\delta)$.*

In Chapter 2.5 we stated the FKN Theorem, which says that if $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{W}^1[f] \geq 1-\delta$ then it must be $O(\delta)$-close to a dictator or negated-dictator. The following theorem shows that once the FKN Theorem is proved, it can be strengthened to give an essentially optimal (Exercise 5.36) closeness bound:

Theorem 5.33. *Suppose the FKN Theorem holds with closeness bound $C\delta$, where $C \geq 1$ is a universal constant. Then in fact it holds with bound $\delta/4+\eta$, where $\eta = 16C^2\delta^2\max(\log(1/C\delta), 1)$.*

Proof. Suppose $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{W}^1[f] \geq 1-\delta \geq 0$. By assumption f is $C\delta$-close to $\pm\chi_i$ for some $i \in [n]$, say $i = n$. Thus we have

$$|\widehat{f}(n)| \geq 1-2C\delta$$

and our task is to show that in fact $|\widehat{f}(n)| \geq 1-\delta/2-2\eta$. We may assume $\delta \leq \frac{1}{10C}$ as otherwise $1-\delta/2-2\eta < 0$ (Exercise 5.38) and there is nothing to prove. By employing the trick from Exercise 2.49 we may also assume $\mathbf{E}[f] = 0$.

Consider the restriction of f given by fixing coordinate n to $b \in \{-1,1\}$; i.e., $f_{[n-1]|b}$. For both choices of b we have $|\mathbf{E}[f_{[n-1]|b}]| \geq 1-2C\delta$ and so Corollary 5.32 implies $\mathbf{W}^1[f_{[n-1]|b}] \leq 16C^2\delta^2\log(1/C\delta)$. Thus

$$16C^2\delta^2\log(1/C\delta) \geq \mathop{\mathbf{E}}_{\boldsymbol{b}}[\mathbf{W}^1[f_{[n-1]|\boldsymbol{b}}]] = \sum_{j<n}(\widehat{f}(\{j\})^2 + \widehat{f}(\{j,n\})^2) \geq \sum_{j<n}\widehat{f}(j)^2,$$

by Corollary 3.22. It follows that

$$\widehat{f}(n)^2 = \mathbf{W}^1[f] - \sum_{j<n} \widehat{f}(j)^2 \geq 1 - \delta - 16C^2\delta^2 \log(1/C\delta),$$

and the proof is completed by the fact that

$$1 - \delta - 16C^2\delta^2 \log(1/C\delta) \geq (1 - \delta/2 - 2\eta)^2$$

when $\delta \leq \frac{1}{10C}$ (Exercise 5.38). □

5.5. Highlight: Peres's Theorem and uniform noise stability

Theorem 5.17 says that if f is an unbiased linear threshold function $f(x) = \mathrm{sgn}(a_1x_1 + \cdots + a_nx_n)$ in which all a_i's are "small", then the noise stability $\mathbf{Stab}_\rho[f]$ is at least (roughly) $\frac{2}{\pi}\arcsin\rho$. Rephrasing in terms of noise sensitivity, this means $\mathbf{NS}_\delta[f]$ is at most (roughly) $\frac{2}{\pi}\sqrt{\delta} + O(\delta^{3/2})$ (see the statement of Theorem 2.45). On the other hand, if some a_i were particularly *large* then f would be pushed in the direction of the dictator function χ_i, which has $\mathbf{NS}_\delta[\chi_i] = \delta \ll \sqrt{\delta}$. This observation suggests that *all* unbiased LTFs f should have $\mathbf{NS}_\delta[f] \leq O(\sqrt{\delta})$. The unbiasedness assumption also seems inessential, since biasing a function should tend to decrease its noise sensitivity.

Indeed, the idea here is correct, as was shown by Peres in 1999:

Peres's Theorem. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any linear threshold function. Then $\mathbf{NS}_\delta[f] \leq O(\sqrt{\delta})$.*

Pleasantly, the proof is quite simple and uses no heavy tools like the Central Limit Theorem. Before getting to it, let's make some remarks. First, Peres's Theorem shows that the class of all linear threshold functions is what's called *uniformly noise-stable*.

Definition 5.34. Let $\mathscr{B}$ be a class of Boolean-valued functions. We say that $\mathscr{B}$ is *uniformly noise-stable* if there exists $\epsilon : [0,1/2] \to [0,1]$ with $\epsilon(\delta) \to 0$ as $\delta \to 0^+$ such that $\mathbf{NS}_\delta[f] \leq \epsilon(\delta)$ holds for all $f \in \mathscr{B}$.

This definition is only interesting for infinite classes $\mathscr{B}$. (Any class containing functions of only finitely many input lengths is vacuously uniformly noise-stable; see Exercise 5.34.) By Proposition 5.6 we see that functions in a uniformly noise-stable class have "almost all of their Fourier weight at constant degree"; i.e., for all $\epsilon > 0$ there is a $k \in \mathbb{N}$ such that $\mathbf{W}^{>k}[f] \leq \epsilon$ for all $f \in \mathscr{B}$. In particular, from Corollary 3.34 we get that if $\mathscr{B}$ is a uniformly noise-stable class then its restriction to n-input functions is learnable from random examples to any constant error in poly(n) time.

Let's make these observations more concrete in the context of linear threshold functions. Peres's Theorem immediately gives that LTFs have their Fourier spectrum ϵ-concentrated up to degree $O(1/\epsilon^2)$ (Proposition 3.3) and hence the class of LTFs is learnable from random examples with error ϵ in time $n^{O(1/\epsilon^2)}$ (Corollary 3.34). The latter result is not too impressive since it's been long known that LTFs are learnable in time $\mathrm{poly}(n, 1/\epsilon)$ using linear programming. However, the noise sensitivity approach is much more flexible. Consider the concept class

$$\mathscr{C} = \{h = g(f_1, \dots, f_s) \mid f_1, \dots, f_s : \{-1,1\}^n \to \{-1,1\} \text{ are LTFs}\}.$$

For each $h : \{-1,1\}^n \to \{-1,1\}$ in $\mathscr{C}$, Peres's Theorem and a union bound (Exercise 2.44) imply that $\mathbf{NS}_\delta[h] \leq O(s\sqrt{\delta})$. Thus from Corollary 3.34 we get that the class $\mathscr{C}$ is learnable in time $n^{O(s^2/\epsilon^2)}$. This is the only known way of showing even that an AND of two LTFs is learnable with error .01 in time $\mathrm{poly}(n)$.

The trick for proving Peres's Theorem is to employ a fairly general technique for bounding noise sensitivity using *average sensitivity* (total influence):

Theorem 5.35. *Let $\delta \in (0, 1/2]$ and let $A : \mathbb{N}^+ \to \mathbb{R}$. Let $\mathscr{B}$ be a class of Boolean-valued functions closed under negation and identification of input variables. Suppose that each $f \in \mathscr{B}$ with domain $\{-1,1\}^n$ has $\mathbf{I}[f] \leq A(n)$. Then each $f \in \mathscr{B}$ has $\mathbf{NS}_\delta[f] \leq \frac{1}{m}A(m)$, where $m = \lfloor 1/\delta \rfloor$.*

Proof. Fix any $f : \{-1,1\}^n \to \{-1,1\}$ from $\mathscr{B}$. Since noise sensitivity is an increasing function of the noise parameter (see the discussion surrounding Proposition 2.51) we may replace δ by $1/m$. Thus our task is to upper-bound $\mathbf{NS}_{1/m}[f] = \mathbf{Pr}[f(\boldsymbol{x}) \neq f(\boldsymbol{y})]$ where $\boldsymbol{x} \sim \{-1,1\}^n$ is uniformly random and $\boldsymbol{y} \in \{-1,1\}^n$ is formed from $\boldsymbol{x}$ by negating each bit independently with probability $1/m$. The rough idea of the proof is that this is equivalent to randomly partitioning $\boldsymbol{x}$'s bits into m parts and then negating a randomly chosen part.

More precisely, let $z \in \{-1,1\}^n$ and let $\pi : [n] \to [m]$ be a partition of $[n]$ into m parts. Define

$$g_{z,\pi} : \{-1,1\}^m \to \{-1,1\}, \quad g_{z,\pi}(w) = f(z \circ w^\pi),$$

where $\circ$ denotes entry-wise multiplication and $w^\pi = (w_{\pi(1)}, \dots, w_{\pi(n)}) \in \{-1,1\}^n$. Since $g_{z,\pi}$ is derived from f by negating and identifying input variables it follows that $g_{z,\pi} \in \mathscr{B}$. So by assumption $g_{z,\pi}$ has total influence $\mathbf{I}[g_{z,\pi}] \leq A(m)$ and hence *average* influence $\boldsymbol{\mathscr{E}}[g_{z,\pi}] \leq \frac{1}{m}A(m)$ (see Exercise 2.43(a)).

Now suppose $\boldsymbol{z} \sim \{-1,1\}^n$ and $\boldsymbol{\pi} : [n] \to [m]$ are chosen uniformly at random. We certainly have

$$\mathop{\mathbf{E}}_{\boldsymbol{z},\boldsymbol{\pi}}[\boldsymbol{\mathscr{E}}[g_{\boldsymbol{z},\boldsymbol{\pi}}]] \leq \frac{1}{m}A(m).$$

To complete the proof we will show that the left-hand side above is precisely $\mathbf{NS}_{1/m}[f]$. Recall that in the experiment for average influence $\boldsymbol{\mathscr{E}}[g]$ we choose

$\boldsymbol{w} \sim \{-1,1\}^m$ and $\boldsymbol{j} \sim [m]$ uniformly at random and check if $g(\boldsymbol{w}) \neq g(\boldsymbol{w}^{\oplus \boldsymbol{j}})$. Thus

$$\mathop{\mathbf{E}}_{\boldsymbol{z},\boldsymbol{\pi}}[\mathscr{E}[g_{\boldsymbol{z},\boldsymbol{\pi}}]] = \mathop{\mathbf{Pr}}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{j}}[g_{\boldsymbol{z},\boldsymbol{\pi}}(\boldsymbol{w}) \neq g_{\boldsymbol{z},\boldsymbol{\pi}}(\boldsymbol{w}^{\oplus \boldsymbol{j}})] = \mathop{\mathbf{Pr}}_{\boldsymbol{w},\boldsymbol{\pi},\boldsymbol{j},\boldsymbol{z}}[f(\boldsymbol{z} \circ \boldsymbol{w}^{\boldsymbol{\pi}}) \neq f(\boldsymbol{z} \circ (\boldsymbol{w}^{\oplus \boldsymbol{j}})^{\boldsymbol{\pi}})].$$

It is not hard to see that the joint distribution of $\boldsymbol{z} \circ \boldsymbol{w}^{\boldsymbol{\pi}}$, $\boldsymbol{z} \circ (\boldsymbol{w}^{\oplus \boldsymbol{j}})^{\boldsymbol{\pi}}$ is the same as that of $\boldsymbol{x}$, $\boldsymbol{y}$. To be precise, define $\boldsymbol{J} = \boldsymbol{\pi}^{-1}(\boldsymbol{j})$, distributed as a random subset of $[n]$ in which each coordinate is included with probability $1/m$, and define $\boldsymbol{\lambda} \in \{-1,1\}^n$ by $\boldsymbol{\lambda}_i = -1$ if and only if $i \in \boldsymbol{J}$. Then

$$\mathop{\mathbf{Pr}}_{\boldsymbol{w},\boldsymbol{\pi},\boldsymbol{j},\boldsymbol{z}}[f(\boldsymbol{z} \circ \boldsymbol{w}^{\boldsymbol{\pi}}) \neq f(\boldsymbol{z} \circ (\boldsymbol{w}^{\oplus \boldsymbol{j}})^{\boldsymbol{\pi}})] = \mathop{\mathbf{Pr}}_{\boldsymbol{w},\boldsymbol{\pi},\boldsymbol{j},\boldsymbol{z}}[f(\boldsymbol{z} \circ \boldsymbol{w}^{\boldsymbol{\pi}}) \neq f(\boldsymbol{z} \circ \boldsymbol{w}^{\boldsymbol{\pi}} \circ \boldsymbol{\lambda})].$$

But for every outcome of $\boldsymbol{w}$, $\boldsymbol{\pi}$, $\boldsymbol{j}$ (and hence $\boldsymbol{J}$, $\boldsymbol{\lambda}$), we may replace $\boldsymbol{z}$ with $\boldsymbol{z} \circ \boldsymbol{w}^{\boldsymbol{\pi}}$ since they have the same distribution, namely uniform on $\{-1,1\}^n$. Then the above becomes

$$\mathop{\mathbf{Pr}}_{\boldsymbol{w},\boldsymbol{\pi},\boldsymbol{j},\boldsymbol{z}}[f(\boldsymbol{z}) \neq f(\boldsymbol{z} \circ \boldsymbol{\lambda})] = \mathbf{NS}_{1/m}[f],$$

as claimed. □

Peres's Theorem is now a simple corollary of Theorem 5.35.

Proof of Peres's Theorem. Let $\mathscr{B}$ be the class of all linear threshold functions. This class is indeed closed under negating and identifying variables. Since each linear threshold function on m bits is *unate* (i.e., monotone up to negation of some input coordinates, see Exercises 2.5, 2.6), its total influence is at most $\sqrt{m}$ (see Exercise 2.23). Applying Theorem 5.35 we get that for any LTF f and any $\delta \in (0,1/2]$,

$$\begin{aligned}\mathbf{NS}_\delta[f] &\leq \tfrac{1}{m}\sqrt{m} = 1/\sqrt{m} \qquad (\text{for } m = \lfloor 1/\delta \rfloor)\\ &\leq O(\sqrt{\delta}).\end{aligned}$$

□

Remark 5.36. Our proof of Peres's Theorem attains the upper bound $\sqrt{1/\lfloor 1/\delta \rfloor}$. This is at most $\sqrt{3/2}\sqrt{\delta}$ for all $\delta \in (0,1/2]$ and it's also $\sqrt{\delta} + O(\delta^{3/2})$ for small δ. To further improve the constant we can use Theorem 2.33 in place of Exercise 2.23; it implies that all unate m-bit functions have total influence at most $\sqrt{2/\pi}\sqrt{m} + O(m^{-1/2})$. This lets us obtain the bound $\mathbf{NS}_\delta[f] \leq \sqrt{2/\pi}\sqrt{\delta} + O(\delta^{3/2})$ for all LTF f.

Recall from Theorem 2.45 that $\mathbf{NS}_\delta[\mathrm{Maj}_n] \sim \frac{2}{\pi}\sqrt{\delta}$ for large n. Thus the constant $\sqrt{2/\pi}$ in the bound from Remark 5.36 is fairly close to optimal. It seems quite likely that majority's $\frac{2}{\pi}$ is the correct constant here. There is still slack in Peres's proof because the random functions $g_{\boldsymbol{z},\boldsymbol{\pi}}$ arising in Theorem 5.35 are unlikely to be majorities, even if f is. The most elegant possible result in this direction would be to prove the following conjecture of Benjamini, Kalai, and Schramm:

Majority Is Least Stable Conjecture. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a linear threshold function, n odd. Then for all $\rho \in [0,1]$, $\mathbf{Stab}_\rho[f] \geq \mathbf{Stab}_\rho[\mathrm{Maj}_n]$.*

(This is a precise statement about majority's noise stability within the class of LTFs; the Majority Is Stablest Theorem refers to its noise stability within the class of small-influence functions.)

A challenging problem in this area is to extend Peres's Theorem to *polynomial threshold functions*. Let

$$\mathscr{P}_{n,k} = \{f : \{-1,1\}^n \to \{-1,1\} \mid f \text{ is a PTF of degree at most } k\}, \quad \mathscr{P}_k = \bigcup_n \mathscr{P}_{n,k}.$$

Peres's Theorem shows that the class $\mathscr{P}_1$ (i.e., LTFs) is uniformly noise-stable. Is the same true of $\mathscr{P}_2$? What about $\mathscr{P}_{100}$? More quantitatively, what upper bound can we prove on $\mathbf{NS}_\delta[f]$ for $f \in \mathscr{P}_k$? Since $\mathscr{P}_k$ is closed under negating and identifying variables, a natural approach to bounding the noise sensitivity of PTFs is to again use Theorem 5.35. For example, if we could show that $\mathbf{I}[f] = o(n)$ for all $f \in \mathscr{P}_k$ we could conclude that $\mathbf{NS}_\delta[f] = o_\delta(1)$ for all $f \in \mathscr{P}_k$; i.e., that $\mathscr{P}_k$ is uniformly noise-stable. (In fact, the total influence approach to bounding noise sensitivity is not just sufficient but is also necessary; see Exercise 5.40.) More ambitiously, if we could show that $\mathbf{I}[f] \leq O_k(1)\sqrt{n}$ for all $f \in \mathscr{P}_{n,k}$ then it would follow that $\mathbf{NS}_\delta[f] \leq O_k(1)\sqrt{\delta}$ for all $f \in \mathscr{P}_k$, strictly generalizing Peres's Theorem. In fact, a conjecture of Gotsman and Linial dating back to 1990 proposes an even more refined bound:

Gotsman–Linial Conjecture. *Let $f \in \mathscr{P}_{n,k}$. Then $\mathbf{I}[f] \leq O_k(1)\sqrt{n}$. More strongly, $\mathbf{I}[f] \leq O(k)\sqrt{n}$. Most strongly, the $f \in \mathscr{P}_{n,k}$ of maximal total influence is the symmetric one $f(x) = \mathrm{sgn}(p(x_1+\cdots+x_n))$, where p is a degree-k univariate polynomial which alternates sign on the $k+1$ values of $x_1+\cdots+x_n$ closest to 0.*

The strongest form of the Gotsman–Linial Conjecture is true when $k = 1$, by Theorem 2.33. However, even for $k = 2$ there was no progress on the conjecture for close to 20 years. At that point two independent works [**DHK⁺10, HKM10**] showed that every $f \in \mathscr{P}_{n,k}$ satisfies both $\mathbf{I}[f] \leq O(n^{1-1/2^k})$ and $\mathbf{I}[f] \leq 2^{O(k)} n^{1-1/O(k)}$. The former (essentially weaker) bound has the advantage of an elementary proof; see Exercise 5.45. It also suffices to show that $\mathscr{P}_k$, the class of degree-k PTFs, is indeed uniformly noise-stable. This gives a nice kind of converse to Proposition 5.6, which showed that every function in a uniformly noise-stable class is close to being a constant-degree PTF.

The latest progress on the Gotsman–Linial Conjecture is the following theorem of Kane [**Kan12**], which comes quite close to proving it:

Theorem 5.37. *Every $f \in \mathscr{P}_{n,k}$ satisfies $\mathbf{I}[f] \leq \sqrt{n} \cdot (2^k \log n)^{O(k \log k)}$. It follows (via Theorem 5.35) that for a* fixed *$k \in \mathbb{N}^+$, every $f \in \mathscr{P}_k$ satisfies $\mathbf{NS}_\delta[f] \leq \sqrt{\delta} \cdot \mathrm{polylog}(1/\delta)$.*

5.6. Exercises and notes

5.1 (*a*) Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is an LTF. Show that it can be expressed as $f(x) = \mathrm{sgn}(a_0 + a_1x_1 + \cdots a_nx_n)$ where the a_i's are integers. (Hint: First obtain rational a_i's by a perturbation.)

(*b*) Show also that a degree-d PTF has a representation in which all of the degree-d polynomial's coefficients are integers.

5.2 Let $f(x) = \mathrm{sgn}(a_0 + a_1x_1 + \cdots a_nx_n)$ be an LTF.

(*a*) Show that if $a_0 = 0$, then $\mathbf{E}[f] = 0$. (Hint: Show that f is in fact an odd function.)

(*b*) Show that if $a_0 \geq 0$, then $\mathbf{E}[f] \geq 0$. Show that the converse need not hold.

(*c*) Suppose $g : \{-1,1\}^n \to \{-1,1\}$ is an LTF with $\mathbf{E}[f] = 0$. Show that g can be represented as $g(x) = \mathrm{sgn}(c_1x_1 + \cdots + c_nx_n)$.

5.3 Suppose $f(x) = \mathrm{sgn}(a_0 + a_1x_1 + \cdots a_nx_n)$ is an LTF with $|a_1| \geq |a_2| \geq \cdots \geq |a_n|$. Show that $\mathbf{Inf}_1[f] \geq \mathbf{Inf}_2[f] \geq \cdots \geq \mathbf{Inf}_n[f]$. (Hint: Why does it suffice to prove this for $n = 2$?)

5.4 (*a*) Show that the number of functions $f : \{-1,1\}^n \to \{-1,1\}$ that are LTFs is at most $2^{n^2+O(n)}$. (Hint: Chow's Theorem.)

(*b*) More generally, show that the number of functions $f : \{-1,1\}^n \to \{-1,1\}$ that are degree-k PTFs is at most $2^{n^{k+1}+O(n)}$.

5.5 (*a*) Suppose $\ell : \{-1,1\}^n \to \mathbb{R}$ is defined by $\ell(x) = a_0 + a_1x_1 + \cdots + a_nx_n$. Define $\widetilde{\ell} : \{-1,1\}^{n+1} \to \mathbb{R}$ by $\widetilde{\ell}(x_0,\ldots,x_n) = a_0x_0 + a_1x_1 + \cdots a_nx_n$. Show that $\|\widetilde{\ell}\|_1 = \|\ell\|_1$ and $\|\widetilde{\ell}\|_2^2 = \|\ell\|_2^2$.

(*b*) Complete the proof of Theorem 5.2.

5.6 Let $f : \{-1,1\}^n \to \{-1,1\}$ be an unbiased linear threshold function. Show that $\mathbf{Inf}_i[f] \geq \frac{1}{\sqrt{2n}}$ for some $i \in [n]$, improving the KKL Theorem for LTFs.

5.7 Consider the following "correlation distillation" problem (cf. Exercise 2.56). For each $i \in [n]$ there is a number $\rho_i \in [-1,1]$ and an independent sequence of pairs of ρ_i-correlated bits, $(\boldsymbol{a}_1^{(1)}, \boldsymbol{b}_2^{(1)})$, $(\boldsymbol{a}_1^{(2)}, \boldsymbol{b}_2^{(2)})$, $(\boldsymbol{a}_1^{(3)}, \boldsymbol{b}_2^{(3)})$, etc. Party A on Earth has access to the stream of bits $\boldsymbol{a}_1^{(1)}$, $\boldsymbol{a}_1^{(2)}$, $\boldsymbol{a}_1^{(3)}$, ... and a party B on Venus has access to the stream $\boldsymbol{b}_1^{(1)}$, $\boldsymbol{b}_1^{(2)}$, $\boldsymbol{b}_1^{(3)}$, ... Neither party knows the numbers $\rho_1,\ldots,\rho_n$. The goal is for B to estimate these correlations. To assist in this, A can send a small number of bits to B. A reasonable strategy is for A to send $f(\boldsymbol{a}^{(1)})$, $f(\boldsymbol{a}^{(2)})$, $f(\boldsymbol{a}^{(3)})$, ... to B, where $f : \{-1,1\}^n \to \{-1,1\}$ is some Boolean function. Using this information B can try to estimate $\mathbf{E}[f(\boldsymbol{a})\boldsymbol{b}_i]$ for each i.

(*a*) Show that $\mathbf{E}[f(\boldsymbol{a})\boldsymbol{b}_i] = \widehat{f}(i)\rho_i$.

(*b*) This motivates choosing an f for which all $\widehat{f}(i)$ are large. If we also insist all $\widehat{f}(i)$ be equal, show that majority functions f maximize this common value.

5.8 For $n \geq 2$, let $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ be a randomly chosen function (as in Exercise 1.7). Show that $\hat{\|}\boldsymbol{f}\hat{\|}_\infty \leq 2\sqrt{n}2^{-n/2}$ except with probability at most 2^{-n}.

5.9 Prove Theorem 5.8.

5.10 (*a*) Give as simple a proof as you can that the parity function $\chi_{[n]} : \{-1,1\}^n \to \{-1,1\}$ is not a PTF of degree $n-1$.

(*b*) Show that if $f : \{-1,1\}^n \to \{-1,1\}$ is not $\pm\chi_{[n]}$, then it *is* a PTF of degree $n-1$. (Hint: Consider $f^{\leq n-1}$.)

5.11 For each $k \in \mathbb{N}^+$, show that there is a degree-k PTF f with $\mathbf{W}^{\leq k}[f] < 2^{1-k}$.

5.12 In this exercise you will show that threshold-of-parities circuits can be effectively simulated by threshold-of-threshold circuits, but not the converse.

(*a*) Let $f : \{-1,1\}^n \to \{-1,1\}$ be a symmetric function. Show that f is computable as the *sum* of at most $2n$ LTFs, plus a constant.

(*b*) Deduce that if $f : \{-1,1\}^n \to \{-1,1\}$ is computable by a size-s threshold-of-parities circuit, then it is also computable by a size-$2ns$ threshold-of-thresholds circuit.

(*c*) Show that the complete quadratic function $\mathrm{CQ}_n : \mathbb{F}_2^n \to \{-1,1\}$ (see Exercise 1.1) is computable by a size-$2n$ threshold-of-thresholds circuit.

(*d*) Assume n even. Show that any threshold-of-parities circuit for CQ_n requires size $2^{n/2}$.

5.13 Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a DNF of size s. Show that f has a PTF representation of sparsity $O(ns^2)$. (Hint: Approximate the ANDs using Theorem 5.12.)

5.14 In contrast to the previous exercise, show that there is a function $f : \{-1,1\}^n \to \{-1,1\}$ computable by a depth-3 AC^0 circuit (see Chapter 4.5) but requiring threshold-of-parities circuits of size at least $n^{\log n}$. (Hint: Involve the inner product mod 2 function and Exercise 4.12.)

5.15 Let $\mathscr{F}$ be a nonempty collection of subsets $S \subseteq [n]$. For each $a \in \{-1,1\}^n$, write $1_{\{a\}} : \{-1,1\}^n \to \{0,1\}$ for the indicator of $\{a\}$, write $1_{\{a\}}^{\mathscr{F}} : \{-1,1\}^n \to \mathbb{R}$ for $\sum_{S\in\mathscr{F}} \widehat{1_{\{a\}}}(S)\chi_S$, and write $\psi_a = \frac{2^n}{|\mathscr{F}|} \cdot 1_{\{a\}}^{\mathscr{F}}$.

(*a*) Show that $\psi_a(a) = 1$ and $\mathbf{E}[\psi_a^2] = \frac{1}{|\mathscr{F}|}$. Show also that for all $x \in \{-1,1\}^n$, $\psi_a(x) = \psi_x(a)$ and $\sum_{a:a\neq x} \psi_a(x)^2 = \frac{2^n}{|\mathscr{F}|} - 1$.

(*b*) Fix $0 < \epsilon < 1$ and suppose that $|\mathscr{F}| \geq (1-\frac{\epsilon^2}{6n})2^n$. Let $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ be a random function as in Exercise 1.7. Show that for each $x \in \{-1,1\}^n$, except with probability at most 4^{-n} it holds that $|\sum_{a:a\neq x} \boldsymbol{f}(a)\psi_a(x)| < \epsilon$.

(*c*) Deduce that for all but a 2^{-n} fraction of functions $f : \{-1,1\}^n \to \{-1,1\}$, there a multilinear polynomial $q : \{-1,1\}^n \to \mathbb{R}$ supported on the monomials $\{\chi_S : S \in \mathscr{F}\}$ such that $\|f - q\|_\infty < \epsilon$.

(*d*) Deduce that all but a 2^{-n} fraction of functions $f : \{-1,1\}^n \to \{-1,1\}$ have PTF representation of degree at most $n/2 + O(\sqrt{n\log n})$.

5.16 (*a*) Show that in the Berry–Esseen Theorem we can also conclude

$$|\mathbf{Pr}[\boldsymbol{S} < u] - \mathbf{Pr}[\boldsymbol{Z} < u]| \le c\gamma.$$

(Hint: You'll need that $\lim_{\delta\to 0^+}\mathbf{Pr}[\boldsymbol{Z} \le u-\delta] = \mathbf{Pr}[\boldsymbol{Z} \le u]$.)

(*b*) Deduce that if $I \subseteq \mathbb{R}$ is any interval, we can also conclude

$$|\mathbf{Pr}[\boldsymbol{S} \in I] - \mathbf{Pr}[\boldsymbol{Z} \in I]| \le 2c\gamma.$$

5.17 Show that the assumptions $\mathbf{E}[\boldsymbol{X}_i] = 0$ and $\sum_{i=1}^n \mathbf{Var}[\boldsymbol{X}_i] = 1$ in the Berry–Esseen Theorem are not restrictive, as follows. Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be independent random variables with finite means and variances. Let $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{X}_i$ and let $\boldsymbol{Z} \sim \mathrm{N}(\mu,\sigma^2)$, where $\mu = \sum_{i=1}^n \mathbf{E}[\boldsymbol{X}_i]$ and $\sigma^2 = \sum_{i=1}^n \mathbf{Var}[\boldsymbol{X}_i]$. Assuming $\sigma^2 > 0$, show that for all $u \in \mathbb{R}$,

$$|\mathbf{Pr}[\boldsymbol{S} \le u] - \mathbf{Pr}[\boldsymbol{Z} \le u]| \le c\epsilon/\sigma^3,$$

where

$$\epsilon = \sum_{i=1}^n \|\boldsymbol{X}_i - \mathbf{E}[\boldsymbol{X}_i]\|_3^3.$$

5.18 (*a*) Use the generalized Binomial Theorem to compute the power series for $(1-z^2)^{-1/2}$, valid for $|z| < 1$.

(*b*) Integrate to obtain the power series for $\arcsin z$ given in (5.9), valid for $|z| < 1$.

(*c*) Confirm that equality holds also for $z = \pm 1$.

5.19 Verify that the random vector $\vec{\boldsymbol{S}}$ defined in (5.7) has $\mathbf{E}[\vec{\boldsymbol{S}}_1] = \mathbf{E}[\vec{\boldsymbol{S}}_2] = 0$, $\mathbf{E}[\vec{\boldsymbol{S}}_1^2] = \mathbf{E}[\vec{\boldsymbol{S}}_2^2] = 1$, $\mathbf{E}[\vec{\boldsymbol{S}}_1\vec{\boldsymbol{S}}_2] = \rho$; i.e., $\mathbf{E}[\vec{\boldsymbol{S}}] = \begin{bmatrix}0\\0\end{bmatrix}$ and $\mathbf{Cov}[\vec{\boldsymbol{S}}] = \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}$.

5.20 Prove Corollary 5.20.

5.21 Fix n odd. Using Theorem 5.19 show that $|\widehat{\mathrm{Maj}_n}(S)|$ is a decreasing function of $|S|$ for odd $1 \le |S| \le \frac{n-1}{2}$. Deduce (using also Corollary 5.20) that $\hat{\|}\mathrm{Maj}_n\hat{\|}_\infty = \mathrm{Maj}_n(\{1\}) \sim \frac{\sqrt{2/\pi}}{\sqrt{n}}$.

5.22 Prove Corollary 5.21.

5.23 Prove Theorem 5.18. (Hint: Corollary 5.21.)

5.24 Complete the proof of Theorem 5.22 by showing that $(1 - \frac{k+1}{n} + \frac{k}{n^2})^{-1/2} \le 1 + 2k/n$ for all $1 \le k \le n/2$.

5.25 Using just the facts that $\mathbf{Stab}_\rho[\mathrm{Maj}_n] \to \frac{2}{\pi}\arcsin\rho$ for all $\rho \in [-1,1]$ and that $\mathbf{Stab}_\rho[\mathrm{Maj}_n] = \sum_{k\ge 0}\mathbf{W}^k[\mathrm{Maj}_n]\rho^k$, deduce that $\lim_{n\to\infty}\mathbf{W}^k[\mathrm{Maj}_n] \to [\rho^k](\frac{2}{\pi}\arcsin\rho)$ for all $k \in \mathbb{N}$. (Hint: By induction on k, always taking ρ "small enough".)

5.26 (*a*) For $0 \le j \le m$ integers, show that $\hat{\|}\mathrm{Maj}_{2m+1}^{=2j+1}\hat{\|}_1 = \binom{m}{j}\frac{1}{2j+1}\cdot\frac{2m+1}{2^{2m}}\binom{2m}{m}$.

(*b*) Deduce that $\hat{\|}\mathrm{Maj}_{2m+1}\hat{\|}_1 = \mathbf{E}\left[\frac{1}{2\boldsymbol{X}+1}\right] \cdot \frac{2m+1}{2^m}\binom{2m}{m}$, where $\boldsymbol{X} \sim \mathrm{Binomial}(m, 1/2)$.

(*c*) Deduce that $\hat{\|}\mathrm{Maj}_n\hat{\|}_1 \sim \frac{2}{\sqrt{\pi}}\frac{1}{\sqrt{n}}2^{n/2}$.

5.27 (*a*) Show that for each odd $k \in \mathbb{N}$,

$$\left(\tfrac{2}{\pi}\right)^{3/2} k^{-3/2} \leq [\rho^k](\tfrac{2}{\pi}\arcsin\rho) \leq \left(\tfrac{2}{\pi}\right)^{3/2} k^{-3/2}(1 + O(1/k)).$$

(Hint: Stirling's approximation.)

(*b*) Prove Corollary 5.23. (Hint: For the second statement you'll need to approximate the sum $\sum_{\text{odd } j>k}\left(\frac{2}{\pi}\right)^{3/2} j^{-3/2}$ by an integral.)

5.28 For integer $0 \leq j \leq n$, define $\mathcal{K}_j : \{-1,1\}^n \to \mathbb{R}$ by $\mathcal{K}_j(x) = \sum_{|S|=j} x^S$. Since $\mathcal{K}_j$ is symmetric, the value $\mathcal{K}_j(x)$ depends only on the number z of -1's in x; or equivalently, on $\sum_{i=1}^n x_i$. Thus we may define $K_j : \{0,1,\dots,n\} \to \mathbb{R}$ by $K_j(z) = \mathcal{K}_j(x)$ for any x with $\sum_i x_i = n - 2z$.

(*a*) Show that $K_j(z)$ can be expressed as a degree-j polynomial in z. It is called the *Kravchuk (*or *Krawtchouk) polynomial* of degree j. (The dependence on n is usually implicit.)

(*b*) Show that $\sum_{j=0}^n \mathcal{K}_j(x) = 2^n \cdot 1_{(1,\dots,1)}(x)$.

(*c*) Show for $\rho \in [-1,1]$ that $\sum_{j=0}^n \mathcal{K}_j(x)\rho^j = 2^n\,\mathbf{Pr}[\boldsymbol{y} = (1,\dots,1)]$, where $\boldsymbol{y} = N_\rho(x)$.

(*d*) Deduce the generating function identity $K_j(z) = [\rho^j]((1-\rho)^z(1+\rho)^{n-z})$.

5.29 Prove Proposition 5.24.

5.30 Prove Proposition 5.25 using the Central Limit Theorem. (Hint for $\mathbf{W}^1[f_n]$: use symmetry to show it equals the square of $\mathbf{E}[f_n(\boldsymbol{x})\sum \frac{1}{\sqrt{n}}\boldsymbol{x}_i]$.)

5.31 Consider the setting of Theorem 5.16. Let $\boldsymbol{S} = \sum_i a_i \boldsymbol{x}_i$ where $\boldsymbol{x} \sim \{-1,1\}^n$, and let $\boldsymbol{Z} \sim \mathrm{N}(0,1)$.

(*a*) Show that $\mathbf{Pr}[|\boldsymbol{S}| \geq t], \mathbf{Pr}[|\boldsymbol{Z}| \geq t] \leq 2\exp(-t^2/2)$ for all $t \geq 0$.

(*b*) Recalling $\mathbf{E}[|\boldsymbol{Y}|] = \int_0^\infty \mathbf{Pr}[|\boldsymbol{Y}| \geq t]\,dt$ for any random variable $\boldsymbol{Y}$, use the Berry–Esseen Theorem (and Remark 5.15, Exercise 5.16) to show

$$\Big|\mathbf{E}[|\boldsymbol{S}|] - \mathbf{E}[|\boldsymbol{Z}|]\Big| \leq O(\epsilon T + \exp(-T^2/2))$$

for any $T \geq 1$.

(*c*) Deduce $|\mathbf{E}[|\boldsymbol{S}|] - \sqrt{2/\pi}| \leq O(\epsilon\sqrt{\log(1/\epsilon)})$.

(*d*) Improve $O(\epsilon\sqrt{\log(1/\epsilon)})$ to the bound $O(\epsilon)$ stated in Theorem 5.16 by using the *nonuniform Berry–Esseen Theorem*, which states that the bound $c\gamma$ in the Berry–Esseen Theorem can be improved to $C\gamma \cdot \frac{1}{1+|u|^3}$ for some constant C.

5.32 Consider the sequence of LTFs defined in Proposition 5.25. Show that

$$\lim_{n\to\infty} \mathbf{Stab}_\rho[f_n] = \Lambda_\rho(\mu).$$

Here $\mu = \overline{\Phi}(t)$ and $\Lambda_\rho(\mu)$ is the *Gaussian quadrant probability* defined by $\Lambda_\rho(\mu) = \mathbf{Pr}[\boldsymbol{z}_1 > t, \boldsymbol{z}_2 > t]$, where $\boldsymbol{z}_1, \boldsymbol{z}_2$ are standard Gaussians with correlation $\mathbf{E}[\boldsymbol{z}_1\boldsymbol{z}_2] = \rho$. Verify also that $\Lambda_\rho(\alpha) = \mathbf{Pr}[\boldsymbol{z}_1 \le t, \boldsymbol{z}_2 \le t]$ where $\alpha = \Phi(t)$.

5.33 In this exercise you will complete the justification of Theorem 5.17 using the following multidimensional Berry-Esseen Theorem:

Theorem 5.38. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent $\mathbb{R}^d$-valued random vectors, each having mean zero. Write $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{X}_i$ and assume $\Sigma = \mathbf{Cov}[\boldsymbol{S}]$ is invertible. Let $\boldsymbol{Z} \sim \mathrm{N}(0, \Sigma)$ be a d-dimensional Gaussian with the same mean and covariance matrix as $\boldsymbol{S}$. Then for all convex sets $U \subseteq \mathbb{R}^d$,*

$$|\mathbf{Pr}[\boldsymbol{S} \in U] - \mathbf{Pr}[\boldsymbol{Z} \in U]| \le C d^{1/4} \gamma,$$

where C is a universal constant, $\gamma = \sum_{i=1}^n \mathbf{E}[\|\Sigma^{-1/2}\boldsymbol{X}_i\|_2^3]$, and $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$.

(*a*) Let $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ where $\rho \in (-1,1)$. Show that

$$\Sigma^{-1} = \begin{bmatrix} 1 & -\rho \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-\rho^2)^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\rho & 1 \end{bmatrix}.$$

(*b*) Compute $y^\top \Sigma^{-1} y$ for $y = \begin{bmatrix} \pm a \\ \pm a \end{bmatrix} \in \mathbb{R}^2$.

(*c*) Complete the proof of Theorem 5.17.

5.34 Let $\mathscr{B}$ be a class of Boolean-valued functions, all of input length at most n. Show that $\mathbf{NS}_\delta[f] \le n\delta$ for all $f \in \mathscr{B}$ and hence $\mathscr{B}$ is uniformly noise-stable (in a sense, vacuously). (Hint: Exercise 2.42.)

5.35 Give a simple proof of the following fact, which is a robust form of the edge-isoperimetric inequality (for volume 1/2) and a weak form of the FKN Theorem: If $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{E}[f] = 0$ and $\mathbf{I}[f] \le 1 + \delta$, then f is $O(\delta)$-close to $\pm\chi_i$ for some $i \in [n]$. In fact, you should be able to achieve δ-closeness (which can be further improved using Theorem 5.33). (Hint: Upper- and lower-bound $\sum_i \widehat{f}(i)^2 \le (\max_i |\widehat{f}(i)|)(\sum_i |\widehat{f}(i)|)$ using Proposition 3.2 and Exercise 2.5(*a*).)

5.36 Show that Theorem 5.33 is essentially optimal by exhibiting functions $f : \{-1,1\}^n \to \{-1,1\}$ with $\widehat{f}(1) = 1 - \delta/2$ and $\mathbf{W}^1[f] \ge 1 - \delta + \Omega(\delta^2 \log(1/\delta))$, for a sequence of δ tending to 0.

5.37 Prove Corollary 5.32.

5.38 Fill in the details of the proof of Theorem 5.33.

5.39 Show that if $f : \{-1,1\}^n \to \{-1,1\}$ is an LTF, then $\frac{d}{d\delta}\mathbf{NS}_\delta[f] \le O(1/\sqrt{\delta})$. (Hint: The only fact needed about LTFs is the corollary of Peres's Theorem that $\mathbf{W}^{\ge k}[f] \le O(1/\sqrt{k})$ for all k.)

5.40 As discussed in Section 5.5, Theorem 5.35 implies that an upper bound on the total influence of degree-k PTFs is sufficient to derive an upper bound on their noise sensitivity. This exercise asks you to show necessity as well. More precisely, suppose $\mathbf{NS}_\delta[f] \leq \epsilon(\delta)$ for all $f \in \mathscr{P}_k$. Show that $\mathbf{I}[f] \leq O(\epsilon(1/n) \cdot n)$ for all $f \in \mathscr{P}_{n,k}$. Deduce that $\mathscr{P}_k$ is uniformly noise-stable if and only if $\mathbf{I}[f] = o(n)$ for all $f \in \mathscr{P}_{n,k}$ and that $\mathbf{NS}_\delta[f] \leq O(k\sqrt{\delta})$ for all $f \in \mathscr{P}_k$ if and only if $\mathbf{I}[f] \leq O(k\sqrt{n})$ for all $f \in \mathscr{P}_{n,k}$. (Hint: Exercise 2.43(*a*).)

5.41 Estimate carefully the asymptotics of $\mathbf{I}[f]$, where $f \in \mathrm{PTF}_{n,k}$ is as in the strongest form of the Gotsman–Linial Conjecture.

5.42 Let $A \subseteq \{-1,1\}^n$ have cardinality $\alpha 2^n$, $\alpha \leq 1/2$. Thinking of $\{-1,1\}^n \subset \mathbb{R}^n$, let $\mu_A \in \mathbb{R}^n$ be the center of mass of A. Show that μ_A is close to the origin in Euclidean distance: $\|\mu_A\|_2 \leq O(\sqrt{\log(1/\alpha)})$.

5.43 Show that the Gaussian isoperimetric function satisfies $\mathscr{U}'' = -1/\mathscr{U}$ on $(0,1)$. Deduce that $\mathscr{U}$ is concave.

5.44 Fix $\alpha \in (0,1/2)$. Let $f : \{-1,1\}^n \to [-1,1]$ satisfy $\mathbf{E}[|f|] \leq \alpha$ and $|\widehat{f}(i)| \leq \epsilon$ for all $i \in [n]$. Show that $\mathbf{W}^1[f] \leq \mathscr{U}(\alpha) + C\epsilon$, where $\mathscr{U}$ is the Gaussian isoperimetric function and where the constant C may depend on α. (Hint: You will need the nonuniform Berry–Esseen Theorem from Exercise 5.31.)

5.45 In this exercise you will show by induction on k that $\mathbf{Inf}[f] \leq 2n^{1-1/2^k}$ for all degree-k PTFs $f : \{-1,1\}^n \to \{-1,1\}$. The $k = 0$ case is trivial. So for $k > 0$, suppose $f = \mathrm{sgn}(p)$ where $p : \{-1,1\}^n \to \mathbb{R}$ is a degree-k polynomial that is never 0.

(*a*) Show for $i \in [n]$ that $\mathbf{E}[f(\boldsymbol{x})\boldsymbol{x}_i \mathrm{sgn}(\mathrm{D}_i p(\boldsymbol{x}))] = \mathbf{Inf}_i[f]$. (Hint: First use the decomposition $f = x_i \mathrm{D}_i f + \mathrm{E}_i f$ to reach $\mathbf{E}[\mathrm{D}_i f \cdot \mathrm{sgn}(\mathrm{D}_i p)]$; then show that $\mathrm{D}_i f = \mathrm{sgn}(\mathrm{D}_i p)$ whenever $\mathrm{D}_i f \neq 0$.)

(*b*) Conclude that $\mathbf{I}[f] \leq \mathbf{E}[|\sum_i \boldsymbol{x}_i \mathrm{sgn}(\mathrm{D}_i p(\boldsymbol{x}))|]$. Remark: When $k = 2$ and thus each $\mathrm{sgn}(\mathrm{D}_i p)$ is an LTF, it is conjectured that this bound is still $O(\sqrt{n})$.

(*c*) Apply Cauchy–Schwarz and deduce

$$\mathbf{I}[f] \leq \sqrt{n + \sum_{i \neq j} \mathbf{E}[\boldsymbol{x}_i \boldsymbol{x}_j \mathrm{sgn}(\mathrm{D}_i p(\boldsymbol{x}))\mathrm{sgn}(\mathrm{D}_j p(\boldsymbol{x}))]}.$$

(*d*) Use Exercise 2.19 and the AM-GM inequality to obtain $\mathbf{I}[f] \leq \sqrt{n + \sum_i \mathbf{I}[\mathrm{sgn}(\mathrm{D}_i p)]}$.

(*e*) Complete the induction.

(*f*) Finally, deduce that the class of degree-k PTFs is uniformly noise-stable, specifically, that every degree-k PTF f satisfies $\mathbf{NS}_\delta[f] \leq 3\delta^{1/2^k}$ for all $\delta \in (0,1/2]$. (Hint: Theorem 5.35.)

Notes. Chow's Theorem was proved by independently by Chow [**Cho61**] and by Tannenbaum [**Tan61**] in 1961; see also Elgot [**Elg61**]. The generalization to PTFs (Theorem 5.8) is due to Bruck [**Bru90**], as is Theorem 5.10 and

Exercise 5.12. Theorems 5.2 and 5.9 are from Gotsman and Linial [GL94] and may be called the Gotsman–Linial Theorems; this work also contains the Gotsman–Linial Conjecture and Exercise 5.11. Conjecture 5.3 should be considered folklore. Corollary 5.13 was proved by Bruck and Smolensky [BS92]; they also essentially proved Theorem 5.12 (but see [SB91]). Exercise 5.13 is usually credited to Krause and Pudlák [KP97]. The upper bound in Exercise 5.4 is asymptotically sharp [Zue89]. Exercise 5.15 is from O'Donnell and Servedio [OS08].

Theorem 2.33 and Proposition 2.58, discussed in Section 5.2, were essentially proved by Titsworth in 1962 [Tit62]; see also [Tit63]. More precisely, Titsworth solved a version of the problem from Exercise 5.7. His motivation was in fact the construction of "interplanetary ranging systems" for measuring deep space distances, e.g., the distance from Earth to Venus. The connection between ranging systems and Boolean functions was suggested by his advisor, Solomon Golomb. Titsworth [Tit62] was also the first to compute the Fourier expansion of Maj_n. His approach involved generating functions and contour integration. Other approaches have used special properties of binomial coefficients [Bra87] or of Kravchuk polynomials [Kal02]. The asymptotics of $\mathbf{W}^k[\mathrm{Maj}_n]$ described in Section 5.3 may have first appeared in Kalai [Kal02], with the error bounds being from O'Donnell [O'D03]. Kravchuk polynomials were introduced by Kravchuk [Kra29].

The Berry–Esseen Theorem is due independently to Berry [Ber41] and Esseen [Ess42]. Shevtsova [She13] has the record for the smallest known constant B that works therein: roughly .5514. The nonuniform version described in Exercise 5.31 is due to Bikelis [Bik66]. The multidimensional version Theorem 5.38 stated in Exercise 5.33 is due to Bentkus [Ben04]. Sheppard proved his formula in 1899 [She99]. The results of Theorem 5.18 may have appeared first in O'Donnell [O'D04, O'D03].

The Level-1 Inequality should probably be considered folklore; it was perhaps first published in Talagrand [Tal96] and we have followed his proof. The first half of the $\frac{2}{\pi}$ Theorem is from Khot et al. [KKMO07]; the second half is from Matulef et al. [MORS10]. Theorem 5.33, which improves the FKN Theorem to achieve "closeness" $\delta/4$, was independently obtained by Jendrej, Oleszkiewicz, and Wojtaszczyk [JOW12], as was Exercise 5.36 showing optimality of this closeness. The closeness achieved in the original proof of the FKN Theorem [FKN02] was $\delta/2$; that proof (like ours) relies on having a separate proof of closeness $O(\delta)$. Kindler and Safra [KS02, Kin02] gave a self-contained proof of the $\delta/2$ bound relying only on the Hoeffding bound. The content of Exercise 5.35 was communicated to the author by Eric Blais. The result of Exercise 5.44 is from [KKMO07]; Exercise 5.42 was suggested by Rocco Servedio.

Peres's Theorem was published in 2004 [**Per04**] but was mentioned as early as 1999 by Benjamini, Kalai, and Schramm [**BKS99**]. The work [**BKS99**] introduced the definition of uniform noise stability and showed that the class of all LTFs satisfies it; however, their upper bound on the noise sensitivity of LTFs was $O(\delta^{1/4})$, worse than Peres's. The proof of Peres's Theorem that we presented is a simplification due to Parikshit Gopalan and incorporates an idea of Diakonikolas et al. [**DHK⁺10, HKM10**]. Regarding the total influence of PTFs, the work of Kane [**Kan12**] shows that every degree-k PTF on n variables has $\mathbf{I}[f] \leq \mathrm{poly}(k)n^{1-1/O(k)}$, which is better than Theorem 5.37 for certain superconstant values of k. Exercise 5.39 was suggested by Nitin Saurabh.

Chapter 6

Pseudorandomness and $\mathbb{F}_2$-polynomials

In this chapter we discuss various notions of pseudorandomness for Boolean functions; by this we mean properties of a fixed Boolean function that are in some way characteristic of randomly chosen functions. We will see some deterministic constructions of pseudorandom probability density functions with small support; these have algorithmic application in the field of derandomization. Finally, several of the results in the chapter will involve interplay between the representation of $f : \{0,1\}^n \to \{0,1\}$ as a polynomial over the reals and its representation as a polynomial over $\mathbb{F}_2$.

6.1. Notions of pseudorandomness

The most obvious spectral property of a truly random function $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ is that all of its Fourier coefficients are very small (as we saw in Exercise 5.8). Let's switch notation to $\boldsymbol{f} : \{-1,1\}^n \to \{0,1\}$; in this case $\boldsymbol{f}(\emptyset)$ will not be very small but rather very close to 1/2. Generalizing:

Proposition 6.1. *Let $n > 1$ and let $\boldsymbol{f} : \{-1,1\}^n \to \{0,1\}$ be a p-biased random function; i.e., each $\boldsymbol{f}(x)$ is 1 with probability p and 0 with probability $1-p$, independently for all $x \in \{-1,1\}^n$. Then except with probability at most 2^{-n}, all of the following hold:*

$$|\widehat{\boldsymbol{f}}(\emptyset) - p| \le 2\sqrt{n}2^{-n/2}, \qquad \forall S \neq \emptyset \quad |\widehat{\boldsymbol{f}}(S)| \le 2\sqrt{n}2^{-n/2}.$$

Proof. We have $\widehat{\boldsymbol{f}}(S) = \sum_x \frac{1}{2^n} x^S \boldsymbol{f}(x)$, where the random variables $\boldsymbol{f}(x)$ are independent. If $S = \emptyset$, then the coefficients $\frac{1}{2^n} x^S$ sum to 1 and the mean of $\widehat{\boldsymbol{f}}(S)$ is p; otherwise the coefficients sum to 0 and the mean of $\widehat{\boldsymbol{f}}(S)$ is 0. Either way we may apply the Hoeffding bound to conclude that

$$\mathbf{Pr}[|\widehat{\boldsymbol{f}}(S) - \mathbf{E}[\widehat{\boldsymbol{f}}(S)]| \geq t] \leq 2\exp(-t^2 \cdot 2^{n-1})$$

for any $t > 0$. Selecting $t = 2\sqrt{n}2^{-n/2}$, the above bound is $2\exp(-2n) \leq 4^{-n}$. The result follows by taking a union bound over all $S \subseteq [n]$. □

This proposition motivates the following basic notion of "pseudorandomness":

Definition 6.2. A function $f : \{-1,1\}^n \to \mathbb{R}$ is *ϵ-regular* (sometimes called *ϵ-uniform*) if $|\widehat{f}(S)| \leq \epsilon$ for all $S \neq \emptyset$.

Remark 6.3. By Exercise 3.9, every function f is ϵ-regular for $\epsilon = \|f\|_1$. We are often concerned with $f : \{-1,1\}^n \to [-1,1]$, in which case we focus on $\epsilon \leq 1$.

Example 6.4. Proposition 6.1 states that a random p-biased function is $(2\sqrt{n}2^{-n/2})$-regular with very high probability. A function is 0-regular if and only if it is constant (even though you might not think of a constant function as very "random"). If $A \subseteq \mathbb{F}_2^n$ is an affine subspace of codimension k then 1_A is 2^{-k}-regular (Proposition 3.12). For n even the inner product mod 2 function and the complete quadratic function, $\mathrm{IP}_n, \mathrm{CQ}_n : \mathbb{F}_2^n \to \{0,1\}$, are $2^{-n/2-1}$-regular (Exercise 1.1). On the other hand, the parity functions $\chi_S : \{-1,1\}^n \to \{-1,1\}$ are not ϵ-regular for any $\epsilon < 1$ (except for $S = \emptyset$). By Exercise 5.21, Maj_n is $\frac{1}{\sqrt{n}}$-regular.

The notion of regularity can be particularly useful for probability density functions; in this case it is traditional to use an alternate name:

Definition 6.5. If $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ is a probability density which is ϵ-regular, we call it an *ϵ-biased density*. Equivalently, φ is ϵ-biased if and only if $|\mathbf{E}_{\boldsymbol{x}\sim\varphi}[\chi_\gamma(\boldsymbol{x})]| \leq \epsilon$ for all $\gamma \in \widehat{\mathbb{F}_2^n} \setminus \{0\}$; thus one can think of "ϵ-biased" as meaning "at most ϵ-biased on subspaces". Note that the marginal of such a distribution on any set of coordinates $J \subseteq [n]$ is also ϵ-biased. If φ is $\varphi_A = 1_A/\mathbf{E}[1_A]$ for some $A \subseteq \mathbb{F}_2^n$ we call A an *ϵ-biased set*.

Example 6.6. For φ a probability density we have $\|\varphi\|_1 = \mathbf{E}[\varphi] = 1$, so every density is 1-biased. The density corresponding to the uniform distribution on $\mathbb{F}_2^n$, namely $\varphi \equiv 1$, is the only 0-biased density. Densities corresponding to the uniform distribution on smaller affine subspaces are "maximally biased": if $A \subseteq \mathbb{F}_2^n$ is an affine subspace of codimension less than n, then φ_A is not ϵ-biased for any $\epsilon < 1$ (Proposition 3.12 again). If $E = \{(0,\dots,0),(1,\dots,1)\}$, then E is a 1/2-biased set (an easy computation, see also Exercise 1.1(*h*)).

There is a "combinatorial" property of functions f that is roughly equivalent to ϵ-regularity. Recall from Exercise 1.29 that $\hat{\|} f \hat{\|}_4^4$ has an equivalent non-Fourier formula: $\mathbf{E}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}}[f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z})f(\boldsymbol{x}+\boldsymbol{y}+\boldsymbol{z})]$. We show (roughly speaking) that f is regular if and only if this expectation is not much bigger than $\mathbf{E}[f]^4 = \mathbf{E}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{w}}[f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z})f(\boldsymbol{w})]$:

Proposition 6.7. *Let* $f : \mathbb{F}_2^n \to \mathbb{R}$*. Then*

(1) If f is ϵ-regular, then $\widehat{\|}f\widehat{\|}_4^4 - \mathbf{E}[f]^4 \le \epsilon^2 \cdot \mathbf{Var}[f]$.

(2) If f is not ϵ-regular, then $\widehat{\|}f\widehat{\|}_4^4 - \mathbf{E}[f]^4 \ge \epsilon^4$.

Proof. If f is ϵ-regular, then

$$\widehat{\|}f\widehat{\|}_4^4 - \mathbf{E}[f]^4 = \sum_{S \neq \emptyset} \widehat{f}(S)^4 \le \max_{S \neq \emptyset}\{\widehat{f}(S)^2\} \cdot \sum_{S \neq \emptyset} \widehat{f}(S)^2 \le \epsilon^2 \cdot \mathbf{Var}[f].$$

On the other hand, if f is not ϵ-regular, then $|\widehat{f}(T)| \ge \epsilon$ for some $T \neq \emptyset$; hence $\widehat{\|}f\widehat{\|}_4^4$ is at least $\widehat{f}(\emptyset)^4 + \widehat{f}(T)^4 \ge \mathbf{E}[f]^4 + \epsilon^4$. □

The condition of ϵ-regularity – that *all* non-empty-set coefficients are small – is quite strong. As we saw when investigating the $\frac{2}{\pi}$ Theorem in Chapter 5.4 it's also interesting to consider f that merely have $|\widehat{f}(i)| \le \epsilon$ for all $i \in [n]$; for monotone f this is the same saying $\mathbf{Inf}_i[f] \le \epsilon$ for i. This suggests two weaker possible notions of pseudorandomness: having all low-degree Fourier coefficients small, and having all influences small. We will consider both possibilities, starting with the second.

Now a randomly chosen $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ will *not* have all of its influences small; in fact as we saw in Exercise 2.12, each $\mathbf{Inf}_i[\boldsymbol{f}]$ is 1/2 in expectation. However, for any $\delta > 0$ it will have all of its $(1-\delta)$-*stable* influences exponentially small (recall Definition 2.52). In Exercise 6.2 you will show:

Fact 6.8. *Fix* $\delta \in [0,1]$ *and let* $\boldsymbol{f} : \{-1,1\}^n \to \{-1,1\}$ *be a randomly chosen function. Then for any* $i \in [n]$*,*

$$\mathbf{E}[\mathbf{Inf}_i^{(1-\delta)}[\boldsymbol{f}]] = \frac{(1-\delta/2)^n}{2-\delta}.$$

This motivates a very important notion of pseudorandomness in the analysis of Boolean functions: having all stable-influences small. Recalling the discussion surrounding Proposition 2.54, we can also describe this as having no "notable" coordinates.

Definition 6.9. We say that $f : \{-1,1\}^n \to \mathbb{R}$ has (ϵ,δ)-*small stable influences*, or *no* (ϵ,δ)-*notable coordinates*, if $\mathbf{Inf}_i^{(1-\delta)}[f] \le \epsilon$ for each $i \in [n]$. This condition gets stronger as ϵ and δ decrease: when $\delta = 0$, meaning $\mathbf{Inf}_i[f] \le \epsilon$ for all i, we simply say f has ϵ-*small influences*.

Example 6.10. Besides random functions, important examples of Boolean-valued functions with no notable coordinates are constants, majority, and large parities. Constant functions are the ultimate in this regard: they have $(0,0)$-small stable influences. (Indeed, constant functions are the only ones with 0-small influences.) The Maj_n function has $\frac{1}{\sqrt{n}}$-small influences. To see the distinction between influences and stable influences, consider the parity functions χ_S. Any parity function χ_S (with $S \neq \emptyset$) has at least one coordinate with maximal influence, 1. But if $|S|$ is "large" then all of its *stable* influences will be small: We have $\mathbf{Inf}_i^{(1-\delta)}[\chi_S]$ equal to $(1-\delta)^{|S|-1}$ when $i \in S$ and equal to 0 otherwise; i.e., χ_S has $((1-\delta)^{|S|-1}, \delta)$-small stable influences. In particular, χ_S has (ϵ, δ)-small stable influences whenever $|S| \geq \frac{\ln(e/\epsilon)}{\delta}$.

The prototypical example of a function $f : \{-1,1\}^n \to \{-1,1\}$ that does *not* have small stable influences is an unbiased k-junta. Such a function has $\mathbf{Var}[f] = 1$ and hence from Fact 2.53 the sum of its $(1-\delta)$-stable influences is at least $(1-\delta)^{k-1}$. Thus $\mathbf{Inf}_i^{(1-\delta)}[f] \geq (1-\delta)^{k-1}/k$ for at least one i; hence f does *not* have $((1-\delta)^k/k, \delta)$-small stable influences for any $\delta \in (0,1)$. A somewhat different example is the function $f(x) = x_0 \mathrm{Maj}_n(x_1, \dots, x_n)$, which has $\mathbf{Inf}_0^{(1-\delta)}[f] \geq 1 - \sqrt{\delta}$; see Exercise 6.5(*d*).

Let's return to considering the interesting condition that $|\widehat{f}(i)| \leq \epsilon$ for all $i \in [n]$. We will call this condition $(\epsilon, 1)$-*regularity*. It is equivalent to saying that $f^{\leq 1}$ is ϵ-regular, or that f has at most ϵ "correlation" with every dictator: $|\langle f, \pm\chi_i \rangle| \leq \epsilon$ for all i. Our third notion of pseudorandomness extends this condition to higher degrees:

Definition 6.11. A function $f : \{-1,1\}^n \to \mathbb{R}$ is (ϵ, k)-*regular* if $|\widehat{f}(S)| \leq \epsilon$ for all $0 < |S| \leq k$; equivalently, if $f^{\leq k}$ is ϵ-regular. For $k = n$ (or $k = \infty$), this condition coincides with ϵ-regularity. When $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ is an (ϵ, k)-regular probability density, it is more usual to call φ (and the associated probability distribution) (ϵ, k)-*wise independent*.

Below we give two alternate characterizations of (ϵ, k)-regularity; however, they are fairly "rough" in the sense that they have exponential losses on k. This can be acceptable if k is thought of as a constant. The first characterization is that f is (ϵ, k)-regular if and only if fixing k input coordinates changes f's mean by at most $O(\epsilon)$. The second characterization is the condition that f has $O(\epsilon)$ covariance with every k-junta.

Proposition 6.12. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $\epsilon \geq 0$, $k \in \mathbb{N}$.*

(1) If f is (ϵ, k)-regular then any restriction of at most k coordinates changes f's mean by at most $2^k \epsilon$.

(2) If f is not (ϵ, k)-regular then some restriction to at most k coordinates changes f's mean by more than ϵ.

Proposition 6.13. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $\epsilon \geq 0$, $k \in \mathbb{N}$.*

(1) *If f is (ϵ,k)-regular, then* $\mathbf{Cov}[f,h] \le \hat{\lVert} h \hat{\rVert}_1 \epsilon$ *for any* $h : \{-1,1\}^n \to \mathbb{R}$ *with* $\deg(h) \le k$. *In particular,* $\mathbf{Cov}[f,h] \le 2^{k/2}\epsilon$ *for any k-junta* $h : \{-1,1\}^n \to \{-1,1\}$.

(2) *If f is not (ϵ,k)-regular, then* $\mathbf{Cov}[f,h] > \epsilon$ *for some k-junta* $h : \{-1,1\}^n \to \{-1,1\}$.

We will prove Proposition 6.12, leaving the proof of Proposition 6.13 to the exercises.

Proof of Proposition 6.12. For the first statement, suppose f is (ϵ,k)-regular and let $J \subseteq [n]$, $z \in \{-1,1\}^J$, where $|J| \le k$. Then the statement holds because

$$\mathbf{E}[f_{\overline{J}|z}] = \widehat{f}(\emptyset) + \sum_{\emptyset \neq T \subseteq J} \widehat{f}(T) z^T$$

(Exercise 1.15) and each of the at most 2^k terms $|\widehat{f}(T)z^T| = |\widehat{f}(T)|$ is at most ϵ.

For the second statement, suppose that $|\widehat{f}(J)| > \epsilon$, where $0 < |J| \le k$. Then a given restriction $z \in \{-1,1\}^J$ changes f's mean by

$$h(z) = \sum_{\emptyset \neq T \subseteq J} \widehat{f}(T) z^T.$$

We need to show that $\|h\|_\infty > \epsilon$, and this follows from

$$\|h\|_\infty = \|h\chi_J\|_\infty \ge |\mathbf{E}[h\chi_J]| = |\widehat{h}(J)| = |\widehat{f}(J)| > \epsilon. \qquad \square$$

Taking $\epsilon = 0$ in the above two propositions we obtain:

Corollary 6.14. *For* $f : \{-1,1\}^n \to \mathbb{R}$, *the following are equivalent:*

(1) *f is $(0,k)$-regular.*

(2) *Every restriction of at most k coordinates leaves f's mean unchanged.*

(3) $\mathbf{Cov}[f,h] = 0$ *for every k-junta* $h : \{-1,1\}^n \to \{-1,1\}$.

If f is a probability density, condition (3) is equivalent to $\mathbf{E}_{\boldsymbol{x}\sim f}[h(\boldsymbol{x})] = \mathbf{E}[h]$ *for every k-junta* $h : \{-1,1\}^n \to \{-1,1\}$.

For such functions, additional terminology is used:

Definition 6.15. If $f : \{-1,1\}^n \to \{-1,1\}$ is $(0,k)$-regular, it is also called *kth-order correlation immune*. If f is in addition unbiased, then it is called *k-resilient*. Finally, if $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\ge 0}$ is a $(0,k)$-regular probability density, then we call φ (and the associated probability distribution) *k-wise independent*.

Example 6.16. Any parity function $\chi_S : \{-1,1\}^n \to \{-1,1\}$ with $|S| = k+1$ is k-resilient. More generally, so is $\chi_S \cdot g$ for any $g : \{-1,1\}^n \to \{-1,1\}$ that does not depend on the coordinates in S. For a good example of a correlation immune function that is not resilient, consider $h : \{-1,1\}^{3m} \to \{-1,1\}$ defined by $h = \chi_{\{1,\dots,2m\}} \wedge \chi_{\{m+1,\dots,3m\}}$. This h is not unbiased, being True on only a

1/4-fraction of inputs. However, its bias does not change unless at least $2m$ input bits are fixed; hence h is $(2m-1)$th-order correlation immune.

We conclude this section with Figure 6.1, indicating how our various notions of pseudorandomness compare:

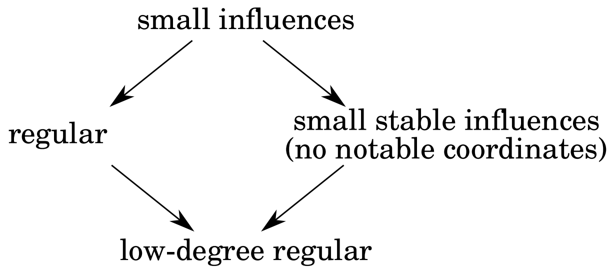


Figure 6.1. Comparing notions of pseudorandomness: arrows go from stronger notions to (strictly) weaker ones

For precise quantitative statements, counterexamples showing that no other relationships are possible, and explanations for why these notions essentially coincide for monotone functions, see Exercise 6.5.

6.2. $\mathbb{F}_2$-polynomials

We began our study of Boolean functions in Chapter 1.2 by considering their polynomial representations over the real field. In this section we take a brief look at their polynomial representations over the field $\mathbb{F}_2$, with False, True being represented by $0, 1 \in \mathbb{F}_2$ as usual. Note that in the field $\mathbb{F}_2$, the arithmetic operations $+$ and $\cdot$ correspond to logical XOR and logical AND, respectively.

Example 6.17. Consider the logical parity (XOR) function on n bits, $\chi_{[n]}$. To represent it over the reals (as we have done so far) we encode False, True by $\pm 1 \in \mathbb{R}$; then $\chi_{[n]} : \{-1,1\}^n \to \{-1,1\}$ has the polynomial representation $\chi_{[n]}(x) = x_1 x_2 \cdots x_n$. Suppose instead we encode False, True by $0, 1 \in \mathbb{F}_2$; then $\chi_{[n]} : \mathbb{F}_2^n \to \mathbb{F}_2$ has the polynomial representation $\chi_{[n]}(x) = x_1 + x_2 + \cdots + x_n$. Notice this polynomial has degree 1, whereas the representation over the reals has degree n.

In general, let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be any Boolean function. Just as in Chapter 1.2 we can find a (multilinear) polynomial representation for it by interpolation. The indicator function $1_{\{a\}} : \mathbb{F}_2^n \to \mathbb{F}_2$ for $a \in \mathbb{F}_2^n$ can be written as

$$1_{\{a\}}(x) = \prod_{i:a_i=1} x_i \prod_{i:a_i=0} (1-x_i), \tag{6.1}$$

a degree-n multilinear polynomial. (We could have written $1+x_i$ rather than $1-x_i$ since these are the same in $\mathbb{F}_2$.) Hence f has the multilinear polynomial expression

$$f(x) = \sum_{a \in \mathbb{F}_2^n} f(a) 1_{\{a\}}(x). \tag{6.2}$$

After simplification, this may be put in the form

$$f(x) = \sum_{S \subseteq [n]} c_S x^S, \tag{6.3}$$

where $x^S = \prod_{i \in S} x_i$ as usual, and each coefficient c_S is in $\mathbb{F}_2$. We call (6.3) the $\mathbb{F}_2$*-polynomial representation of* f. As an example, if $f = \chi_{[3]}$ is the parity function on 3 bits, its interpolation is

$$\begin{aligned}\chi_{[3]}(x) &= (1-x_1)(1-x_2)x_3 + (1-x_1)x_2(1-x_3) + x_1(1-x_2)(1-x_3) + x_1x_2x_3 \\ &= x_1 + x_2 + x_3 - 2(x_1x_2 + x_1x_3 + x_2x_3) + 4x_1x_2x_3 \qquad (6.4) \\ &= x_1 + x_2 + x_3\end{aligned}$$

as expected. We also have uniqueness of the $\mathbb{F}_2$-polynomial representation; the quickest way to see this is to note that there are 2^{2^n} functions $\mathbb{F}_2^n \to \mathbb{F}_2$ and also 2^{2^n} possible choices for the coefficients c_S. Summarizing:

Proposition 6.18. *Every $f : \mathbb{F}_2^n \to \mathbb{F}_2$ has a unique $\mathbb{F}_2$-polynomial representation as in* (6.3).

Example 6.19. The logical AND function $\mathrm{AND}_n : \mathbb{F}_2^n \to \mathbb{F}_2$ has the simple expansion $\mathrm{AND}_n(x) = x_1x_2\cdots x_n$. The inner product mod 2 function has the degree-2 expansion $\mathrm{IP}_{2n}(x_1,\dots,x_n,y_1,\dots,y_n) = x_1y_1 + x_2y_2 + \cdots + x_ny_n$.

Since the $\mathbb{F}_2$-polynomial representation is unique we may define $\mathbb{F}_2$-degree:

Definition 6.20. The $\mathbb{F}_2$*-degree* of a Boolean function $f : \{\mathsf{False}, \mathsf{True}\}^n \to \{\mathsf{False}, \mathsf{True}\}$, denoted $\deg_{\mathbb{F}_2}(f)$, is the degree of its $\mathbb{F}_2$-polynomial representation. We reserve the notation $\deg(f)$ for the degree of f's Fourier expansion.

We can also give a formula for the coefficients of the $\mathbb{F}_2$-polynomial representation:

Proposition 6.21. *Suppose $f : \mathbb{F}_2^n \to \mathbb{F}_2$ has $\mathbb{F}_2$-polynomial representation $f(x) = \sum_{S \subseteq [n]} c_S x^S$. Then $c_S = \sum_{\mathrm{supp}(x) \subseteq S} f(x)$.*

Corollary 6.22. *Let $f : \{\mathsf{False}, \mathsf{True}\}^n \to \{\mathsf{False}, \mathsf{True}\}$. Then $\deg_{\mathbb{F}_2}(f) = n$ if and only if $f(x) = \mathsf{True}$ for an odd number of inputs x.*

The proof of Proposition 6.21 is left for Exercise 6.10; Corollary 6.22 is just the case $S = [n]$. You can also directly see that $c_{[n]} = \sum_x f(x)$ by observing what happens with the monomial $x_1x_2\cdots x_n$ in the interpolation (6.1), (6.2).

Given a generic Boolean function $f : \{\mathsf{False}, \mathsf{True}\}^n \to \{\mathsf{False}, \mathsf{True}\}$ it's natural to ask about the relationship between its Fourier expansion (i.e., polynomial representation over $\mathbb{R}$) and its $\mathbb{F}_2$-polynomial representation. In fact you can easily derive the $\mathbb{F}_2$-representation from the $\mathbb{R}$-representation. Suppose $p(x)$ is the Fourier expansion of f; i.e., f's $\mathbb{R}$-multilinear representation when we interpret False, True as $\pm 1 \in \mathbb{R}$. From Exercise 1.9, $q(x) = \frac{1}{2} - \frac{1}{2}p(1-2x_1, \dots, 1-2x_n)$ is the unique $\mathbb{R}$-multilinear representation for f when we interpret False, True as $0, 1 \in \mathbb{R}$. But we can also obtain $q(x)$ by carrying out the interpolation in (6.1), (6.2) over $\mathbb{Z}$. Thus the $\mathbb{F}_2$ representation of f is obtained simply by reducing $q(x)$'s (integer) coefficients modulo 2.

We saw an example of this derivation above with $\chi_{[3]}$. The ± 1-representation is $x_1 x_2 x_3$. The representation over $\{0, 1\} \in \mathbb{Z} \subseteq \mathbb{R}$ is $\frac{1}{2} - \frac{1}{2}(1-2x_1)(1-2x_2)(1-2x_3)$, which when expanded equals (6.4) and has integer coefficients. Finally, we obtain the $\mathbb{F}_2$ representation $x_1 + x_2 + x_3$ by reducing the coefficients of (6.4) modulo 2.

One thing to note about this transformation from Fourier expansion to $\mathbb{F}_2$-representation is that it can only decrease degree. As noted in Exercise 1.11, the first step, forming $q(x) = \frac{1}{2} - \frac{1}{2}p(1-2x_1, \dots, 1-2x_n)$, does not change the degree at all (except if $p(x) \equiv 1$, $q(x) \equiv 0$). And the second step, reducing q's coefficients modulo 2, cannot increase the degree. We conclude:

Proposition 6.23. *Let $f : \{-1,1\}^n \to \{-1,1\}$. Then $\deg_{\mathbb{F}_2}(f) \leq \deg(f)$.*

Here is an interesting consequence of this proposition. Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is k-resilient; i.e., $\widehat{f}(S) = 0$ for all $|S| \leq k < n$. Let $g = \chi_{[n]} \cdot f$; thus $\widehat{g}(S) = \widehat{f}([n] \setminus S)$ and hence $\deg(g) \leq n-k-1$. From Proposition 6.23 we deduce $\deg_{\mathbb{F}_2}(g) \leq n-k-1$. But if we interpret $f, g : \mathbb{F}_2^n \to \mathbb{F}_2$, then $g = x_1 + \cdots + x_n + f$ and hence $\deg_{\mathbb{F}_2}(g) = \deg_{\mathbb{F}_2}(f)$ (unless f is parity or its negation). Thus:

Proposition 6.24. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be k-resilient, $k < n-1$. Then $\deg_{\mathbb{F}_2}(f) \leq n-k-1$.*

This proposition was shown by Siegenthaler, a cryptographer who was studying stream ciphers; his motivation is discussed further in the notes in Section 6.6. More generally, Siegenthaler proved the following result (the proof does not require Fourier analysis):

Siegenthaler's Theorem. *Proposition 6.24 holds. Further, if f is merely kth-order correlation immune, then we still have $\deg_{\mathbb{F}_2}(f) \leq n-k$ (for $k < n$).*

Proof. Pick a monomial x^J of maximal degree $d = \deg_{\mathbb{F}_2}(f)$ in f's $\mathbb{F}_2$-polynomial representation; we may assume $d > 1$ else we are done. Make an arbitrary restriction to the $n-d$ coordinates outside of J, forming function $g : \mathbb{F}_2^J \to \mathbb{F}_2$. The monomial x^J still appears in g's $\mathbb{F}_2$-polynomial representation; thus by Corollary 6.22, g is 1 for an odd number of inputs.

Let us first show Proposition 6.24. Assuming f is k-resilient, it is unbiased. But g is 1 for an odd number of inputs so it cannot be unbiased (since 2^{d-1} is even for $d > 1$). Thus the restriction changed f's bias, and we must have $n-d > k$, hence $d \le n-k-1$.

Suppose now f is merely kth-order correlation immune. Pick an arbitrary input coordinate for g and suppose its two possible restrictions give subfunctions g_0 and g_1. Since g has an odd number of 1's, one of g_0 has an odd number of 1's and the other has an even number. In particular, g_0 and g_1 have different biases. One of these biases must differ from f's. Thus $n-d+1 > k$, hence $d \le n-k$. □

We end this section by mentioning another bound related to correlation immunity:

Theorem 6.25. *Suppose $f : \{-1,1\}^n \to \{-1,1\}$ is kth-order correlation immune but* not *k-resilient (i.e., $\mathbf{E}[f] \neq 0$). Then $k+1 \le \frac{2}{3}n$.*

The proof of this theorem (left to Exercise 6.14) uses the Fourier expansion rather than the $\mathbb{F}_2$-representation. The bounds in both Siegenthaler's Theorem and Theorem 6.25 can be sharp in many cases; see Exercise 6.15.

6.3. Constructions of various pseudorandom functions

In this section we give some constructions of Boolean functions with strong pseudorandomness properties. We begin by discussing *bent* functions:

Definition 6.26. A function $f : \mathbb{F}_2^n \to \{-1,1\}$ (with n even) is called *bent* if $|\widehat{f}(\gamma)| = 2^{-n/2}$ for all $\gamma \in \widehat{\mathbb{F}_2^n}$.

Bent functions are $2^{-n/2}$-regular. If the definition of ϵ-regularity were changed so that even $|\widehat{f}(0)|$ needed to be at most ϵ, then bent functions would be the most regular possible functions. This is because $\sum_\gamma \widehat{f}(\gamma)^2 = 1$ for any $f : \mathbb{F}_2^n \to \{-1,1\}$ and hence at least one $|\widehat{f}(\gamma)|$ must be at least $2^{-n/2}$. In particular, bent functions are those that are maximally distant from the class of affine functions, $\{\pm\chi_\gamma : \gamma \in \widehat{\mathbb{F}_2^n}\}$.

We have encountered some bent functions already. The canonical example is the inner product mod 2 function, $\mathrm{IP}_n(x) = \chi(x_1x_{n/2+1} + x_2x_{n/2+2} + \cdots + x_{n/2}x_n)$. (Recall the notation $\chi(b) = (-1)^b$.) For $n = 2$ this is just the AND_2 function $\frac{1}{2} + \frac{1}{2}x_1 + \frac{1}{2}x_2 - \frac{1}{2}x_1x_2$, which is bent by inspection. For general n, the bentness is a consequence of the following fact (proved in Exercise 6.16):

Proposition 6.27. *Let $f : \mathbb{F}_2^n \to \{-1,1\}$ and $g : \mathbb{F}_2^{n'} \to \{-1,1\}$ be bent. Then $f \oplus g : \mathbb{F}_2^{n+n'} \to \{-1,1\}$ defined by $(f \oplus g)(x,x') = f(x)g(x')$ is also bent.*

Another example of a bent function is the complete quadratic function $\mathrm{CQ}_n(x) = \chi(\sum_{1\le i<j\le n} x_i x_j)$ from Exercise 1.1. Actually, in some sense it is the "same" example, as we now explain.

Proposition 6.28. *Let $f : \mathbb{F}_2^n \to \{-1,1\}$ be bent. Then $\pm\chi_\gamma \cdot f$ is bent for any $\gamma \in \widehat{\mathbb{F}_2^n}$, as is $f \circ M$ for any invertible linear transformation $M : \mathbb{F}_2^n \to \mathbb{F}_2^n$.*

Proof. Multiplying by -1 does not change bentness, and both $\chi_\gamma \cdot f$ and $f \circ M$ have the same Fourier coefficients as f up to a permutation (see Exercise 3.1). □

We claim that CQ_n arises from $f = \mathrm{IP}_n$ as in Proposition 6.28. In the case $n = 4$, this is because $\sum_{1\le i<j\le 4} x_i x_j = (x_1+x_3)(x_2+x_3)+(x_1+x_2+x_3)x_4+x_3$ over $\mathbb{F}_2$; thus

$$\mathrm{CQ}_4(x) = \mathrm{IP}_4(Mx)\cdot\chi_{(0,0,1,0)}(x), \quad \text{where } M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ is invertible.}$$

The general case is left to Exercise 6.20. In fact, every bent f with $\deg_{\mathbb{F}_2}(f) \le 2$ arises by applying Proposition 6.28 to the inner product mod 2 function; see Exercise 6.19. There are other large families of bent functions; however, the problem of classifying all bent functions is open and seems difficult. We content ourselves by describing one more family:

Proposition 6.29. *Let $f : \mathbb{F}_2^{2n} \to \{-1,1\}$ be defined by $f(x,y) = \mathrm{IP}_{2n}(x,y)g(y)$ where $g : \{-1,1\}^n \to \{-1,1\}$ is arbitrary. Then f is bent.*

Proof. We will think of $y \in \widehat{\mathbb{F}_2^n}$, so $\mathrm{IP}_{2n}(x,y) = \chi_y(x)$. We'll also write a generic $\gamma \in \widehat{\mathbb{F}_2^{2n}}$ as (γ_1,γ_2). Then indeed

$$\begin{aligned}\widehat{f}(\gamma) = \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}}[\chi_{\boldsymbol{y}}(\boldsymbol{x})g(\boldsymbol{y})\chi_{(\gamma_1,\gamma_2)}(\boldsymbol{x},\boldsymbol{y})] &= \mathop{\mathbf{E}}_{\boldsymbol{y}}\left[g(\boldsymbol{y})\chi_{\gamma_2}(\boldsymbol{y})\mathop{\mathbf{E}}_{\boldsymbol{x}}[\chi_{\boldsymbol{y}+\gamma_1}(\boldsymbol{x})]\right] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{y}}[g(\boldsymbol{y})\chi_{\gamma_2}(\boldsymbol{y})\mathbf{1}_{\{\boldsymbol{y}+\gamma_1=0\}}] = 2^{-n}g(\gamma_1)\chi_{\gamma_2}(\gamma_1) = \pm 2^{-n}. \quad \square\end{aligned}$$

We next discuss explicit constructions of small ϵ-biased sets, which are of considerable use in the field of algorithmic derandomization. The most basic step in a randomized algorithm is drawing a string $\boldsymbol{x} \sim \mathbb{F}_2^n$ from the uniform distribution; however, this has the "cost" of generating n independent, random bits. But sometimes it's not necessary that $\boldsymbol{x}$ precisely have the uniform distribution; it may suffice that $\boldsymbol{x}$ be drawn from an ϵ-biased density. If we can deterministically find an ϵ-biased (multi-)set A of cardinality, say, 2^ℓ, then we can generate $\boldsymbol{x} \sim \varphi_A$ using just ℓ independent random bits. We will see some example derandomizations of this nature in Section 6.4; for now we discuss constructions.

Fix $\ell \in \mathbb{N}^+$ and recall that there exists a finite field $\mathbb{F}_{2^\ell}$ with exactly 2^ℓ elements. It is easy to find an explicit representation for $\mathbb{F}_{2^\ell}$ – a complete addition and multiplication table, say – in time $2^{O(\ell)}$. (In fact, one can compute within $\mathbb{F}_{2^\ell}$ even in deterministic poly(ℓ) time.) The field elements $x \in \mathbb{F}_{2^\ell}$ are naturally encoded by distinct ℓ-bit vectors; we will write $\mathrm{enc} : \mathbb{F}_{2^\ell} \to \mathbb{F}_2^\ell$ for this encoding. The encoding is linear; i.e., it satisfies $\mathrm{enc}(0) = (0,\dots,0)$ and $\mathrm{enc}(x+y) = \mathrm{enc}(x) + \mathrm{enc}(y)$ for all $x, y \in \mathbb{F}_{2^\ell}$.

Theorem 6.30. *There is a deterministic algorithm that, given $n \geq 1$ and $0 < \epsilon \leq 1/2$, runs in* $\mathrm{poly}(n/\epsilon)$ *time and outputs a multiset $A \subseteq \mathbb{F}_2^n$ of cardinality at most $16(n/\epsilon)^2$ with the property that φ_A is an ϵ-biased density.*

Proof. It suffices to obtain cardinality $(n/\epsilon)^2$ under the assumption that $\epsilon = 2^{-t}$ and $n = 2^{\ell-t}$ are integer powers of 2. We will describe a probability density φ on $\mathbb{F}_2^n$ by giving a procedure for drawing a string $\boldsymbol{y} \sim \varphi$ which uses 2ℓ independent random bits. A will be the multiset of $2^{2\ell} = (n/\epsilon)^2$ possible outcomes for $\boldsymbol{y}$. It will be clear that A can be generated in deterministic polynomial time. The goal will be to show that φ is 2^{-t}-biased.

To draw $\boldsymbol{y} \sim \varphi$, first choose $\boldsymbol{r}, \boldsymbol{s} \sim \mathbb{F}_{2^\ell}$ independently and uniformly. This uses 2ℓ independent random bits. Then define the ith coordinate of $\boldsymbol{y}$ by

$$\boldsymbol{y}_i = \langle \mathrm{enc}(\boldsymbol{r}^i), \mathrm{enc}(\boldsymbol{s})\rangle, \quad i \in [n],$$

where the inner product $\langle\cdot,\cdot\rangle$ takes place in $\mathbb{F}_2^\ell$. Fixing $\gamma \in \widehat{\mathbb{F}_2^n} \setminus \{0\}$, we need to argue that $|\mathbf{E}[\chi_\gamma(\boldsymbol{y})]| \leq 2^{-t}$. Now over $\mathbb{F}_2^\ell$,

$$\langle \gamma, \boldsymbol{y}\rangle = \sum_{i=1}^n \gamma_i \langle \mathrm{enc}(\boldsymbol{r}^i), \mathrm{enc}(\boldsymbol{s})\rangle = \left\langle \sum_{i=1}^n \gamma_i \mathrm{enc}(\boldsymbol{r}^i), \mathrm{enc}(\boldsymbol{s})\right\rangle = \langle \mathrm{enc}(\sum_{i=1}^n \gamma_i \boldsymbol{r}^i), \mathrm{enc}(\boldsymbol{s})\rangle,$$

where the last step used linearity of enc. Thus

$$\mathbf{E}[\chi_\gamma(\boldsymbol{y})] = \mathbf{E}[(-1)^{\langle\gamma,\boldsymbol{y}\rangle}] = \mathop{\mathbf{E}}_{\boldsymbol{r}}\left[\mathop{\mathbf{E}}_{\boldsymbol{s}}[(-1)^{\langle \mathrm{enc}(p_\gamma(\boldsymbol{r})), \mathrm{enc}(\boldsymbol{s})\rangle}]\right], \tag{6.5}$$

where $p_\gamma : \mathbb{F}_{2^\ell} \to \mathbb{F}_{2^\ell}$ is the polynomial $a \mapsto \gamma_1 a + \gamma_2 a^2 + \cdots + \gamma_n a^n$. This polynomial is of degree at most n, and is nonzero since $\gamma \neq 0$. Hence it has at most n roots (zeroes) over the field $\mathbb{F}_{2^\ell}$. Whenever $\boldsymbol{r}$ is one of these roots, $\mathrm{enc}(p_\gamma(\boldsymbol{r})) = 0$ and the inner expectation in (6.5) is 1. But whenever $\boldsymbol{r}$ is not a root of p_γ we have $\mathrm{enc}(p_\gamma(\boldsymbol{r})) \neq 0$ and so the inner expectation is 0. (We are using Fact 1.7 here.) We deduce that

$$0 \leq \mathbf{E}[\chi_\gamma(\boldsymbol{y})] \leq \mathbf{Pr}[\boldsymbol{r} \text{ is a root of } p_\gamma] \leq \frac{n}{2^\ell} = 2^{-t},$$

which is stronger than what we need. □

The bound of $O(n/\epsilon)^2$ in this theorem is fairly close to being optimally small; see Exercise 6.24 and the notes for this chapter.

Another useful tool in derandomization is that of k-wise independent distributions. Sometimes a randomized algorithm using n independent random bits will still work assuming only that every subset of k of the bits is independent. Thus as with ϵ-biased sets, it's worthwhile to come up with deterministic constructions of small sets $A \subset \mathbb{F}_2^n$ such that the density function φ_A is k-wise independent (i.e., $(0,k)$-regular). The best known examples have the additional pleasant feature that A is a linear subspace of $\mathbb{F}_2^n$; in this case, k-wise independence is easy to characterize:

Proposition 6.31. *Let H be an $m \times n$ matrix over $\mathbb{F}_2$ and let $A \leq \mathbb{F}_2^n$ be the span of H's rows. Then φ_A is k-wise independent if and only if any sum of at most k columns of H is nonzero in $\mathbb{F}_2^m$. (We exclude the "empty" sum.)*

Proof. Since $\varphi_A = \sum_{\gamma \in A^\perp} \chi_\gamma$ (Proposition 3.11), φ_A is k-wise independent if and only if $|\gamma| > k$ for every $\gamma \in A^\perp \setminus \{0\}$. But $\gamma \in A^\perp$ if and only if $H\gamma = 0$. □

Here is a simple construction of such a matrix with $m \sim k \log n$:

Theorem 6.32. *Let $k, \ell \in \mathbb{N}^+$ and assume $n = 2^\ell \geq k$. Then for $m = (k-1)\ell + 1$, there is a matrix $H \in \mathbb{F}_2^{m \times n}$ such that any sum of at most k columns of H is nonzero in $\mathbb{F}_2^m$.*

Proof. Write $\alpha_1, \dots, \alpha_n$ for the elements of the finite field $\mathbb{F}_n$, and consider the following matrix $H' \in \mathbb{F}_n^{k \times n}$:

$$H' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \cdots & \alpha_n^{k-1} \end{bmatrix}.$$

Any submatrix of H' formed by choosing k columns is a Vandermonde matrix and is therefore nonsingular. Hence any subset of k columns of H' is linearly independent in $\mathbb{F}_n^k$. In particular, any sum of at most k columns of H' is nonzero in $\mathbb{F}_n^k$. Now form $H \in \mathbb{F}_2^{m \times n}$ from H' by replacing each entry α_j^i $(i > 0)$ with $\mathrm{enc}(\alpha_j^i)$, thought of as a column vector in $\mathbb{F}_2^\ell$. Since enc is a linear map we may conclude that any sum of at most k columns of H is nonzero in $\mathbb{F}_2^m$. □

Corollary 6.33. *There is a deterministic algorithm that, given integers $1 \leq k \leq n$, runs in $\mathrm{poly}(n^k)$ time and outputs a subspace $A \leq \mathbb{F}_2^n$ of cardinality at most $2^k n^{k-1}$ such that φ_A is k-wise independent.*

Proof. It suffices to assume $n = 2^\ell$ is a power of 2 and then obtain cardinality $2n^{k-1} = 2^{(k-1)\ell+1}$. In this case, the algorithm constructs H as in Theorem 6.32 and takes A to be the span of its rows. The fact that φ_A is k-wise independent is immediate from Proposition 6.31. □

For constant k this upper bound of $O(n^{k-1})$ is close to optimal. It can be improved to $O(n^{\lfloor k/2 \rfloor})$, but there is a lower bound of $\Omega(n^{\lfloor k/2 \rfloor})$ for constant k; see Exercises 6.27, 6.28.

We conclude this section by noting that taking an ϵ-biased density within a k-wise independent subspace yields an (ϵ, k)-wise independent density:

Lemma 6.34. *Suppose $H \in \mathbb{F}_2^{m \times n}$ is such that any sum of at most k columns of H is nonzero in $\mathbb{F}_2^m$. Let φ be an ϵ-biased density on $\mathbb{F}_2^m$. Consider drawing $\boldsymbol{y} \sim \varphi$ and setting $\boldsymbol{z} = \boldsymbol{y}^\top H \in \mathbb{F}_2^n$. Then the density of $\boldsymbol{z}$ is (ϵ, k)-wise independent.*

Proof. Suppose $\gamma \in \widehat{\mathbb{F}_2^n}$ has $0 < |\gamma| \leq k$. Then $H\gamma$ is nonzero by assumption and hence $|\mathbf{E}[\chi_\gamma(\boldsymbol{z})]| = |\mathbf{E}_{\boldsymbol{y} \sim \varphi}[(-1)^{\boldsymbol{y}^\top H \gamma}]| \leq \epsilon$ since φ is ϵ-biased. □

As a consequence, combining the constructions of Theorem 6.30 and Theorem 6.32 gives an (ϵ, k)-wise independent distribution that can be sampled from using only $O(\log k + \log\log(n) + \log(1/\epsilon))$ independent random bits:

Theorem 6.35. *There is a deterministic algorithm that, given integers $1 \leq k \leq n$ and also $0 < \epsilon \leq 1/2$, runs in time $\mathrm{poly}(n/\epsilon)$ and outputs a multiset $A \subseteq \mathbb{F}_2^n$ of cardinality $O(k \log(n)/\epsilon)^2$ (a power of 2) such that φ_A is (ϵ, k)-wise independent.*

6.4. Applications in learning and testing

In this section we describe some applications of our study of pseudorandomness.

We begin with a notorious open problem from learning theory, that of learning juntas. Let $\mathscr{C} = \{f : \mathbb{F}_2^n \to \mathbb{F}_2 \mid f \text{ is a } k\text{-junta}\}$; we will always assume that $k \leq O(\log n)$. In the query access model, it is quite easy to learn $\mathscr{C}$ exactly (i.e., with error 0) in $\mathrm{poly}(n)$ time (Exercise 3.37(*a*)). However, in the model of random examples, it's not obvious how to learn $\mathscr{C}$ more efficiently than in the $n^k \cdot \mathrm{poly}(n)$ time required by the Low-Degree Algorithm (see Theorem 3.36). Unfortunately, this is superpolynomial as soon as $k > \omega(1)$. The state of affairs is the same in the case of depth-k decision trees (a superclass of $\mathscr{C}$), and is similar in the case of $\mathrm{poly}(n)$-size DNFs and CNFs. Thus if we wish to learn, say, $\mathrm{poly}(n)$-size decision trees or DNFs from random examples only, a necessary prerequisite is doing the same for $O(\log n)$-juntas.

Whether or not $\omega(1)$-juntas can be learned from random examples in polynomial time is a longstanding open problem. Here we will show a modest improvement on the n^k-time algorithm:

Theorem 6.36. *For $k \leq O(\log n)$, the class $\mathscr{C} = \{f : \mathbb{F}_2^n \to \mathbb{F}_2 \mid f \text{ is a } k\text{-junta}\}$ can be exactly learned from random examples in time $n^{(3/4)k} \cdot \mathrm{poly}(n)$.*

(The 3/4 in this theorem can in fact be replaced by $\omega/(\omega+1)$, where ω is any number such that $n \times n$ matrices can be multiplied in time $O(n^{\omega})$.)

The first observation we will use to prove Theorem 6.36 is that to learn k-juntas, it suffices to be able to identify a single coordinate that is relevant (see Definition 2.18). The proof of this is fairly simple and is left for Exercise 6.31:

Lemma 6.37. *Theorem 6.36 follows from the existence of a learning algorithm that, given random examples from a nonconstant k-junta $f : \mathbb{F}_2^n \to \mathbb{F}_2$, finds at least one relevant coordinate for f (with probability at least $1-\delta$) in time $n^{(3/4)k} \cdot \mathrm{poly}(n) \cdot \log(1/\delta)$.*

Assume then that we have random example access to a (nonconstant) k-junta $f : \mathbb{F}_2^n \to \mathbb{F}_2$. As in the Low-Degree Algorithm we will estimate the Fourier coefficients $\widehat{f}(S)$ for all $1 \leq |S| \leq d$, where $d \leq k$ is a parameter to be chosen later. Using Proposition 3.30 we can ensure that all estimates are accurate to within $(1/3)2^{-k}$, except with probability most $\delta/2$, in time $n^d \cdot \mathrm{poly}(n) \cdot \log(1/\delta)$. (Recall that $2^k \leq \mathrm{poly}(n)$.) Since f is a k-junta, all of its Fourier coefficients are either 0 or at least 2^{-k} in magnitude; hence we can exactly identify the sets S for which $\widehat{f}(S) \neq 0$. For any such S, *all* of the coordinates $i \in S$ are relevant for f (Exercise 2.11). So unless $\widehat{f}(S) = 0$ for all $1 \leq |S| \leq d$, we can find a relevant coordinate for f in time $n^d \cdot \mathrm{poly}(n) \cdot \log(1/\delta)$ (except with probability at most $\delta/2$).

To complete the proof of Theorem 6.36 it remains to handle the case that $\widehat{f}(S) = 0$ for all $1 \leq |S| \leq d$; i.e., f is dth-order correlation immune. In this case, by Siegenthaler's Theorem we know that $\deg_{\mathbb{F}_2}(f) \leq k-d$. (Note that $d < k$ since f is not constant.) But there is a learning algorithm running in time in time $O(n)^{3\ell} \cdot \log(1/\delta)$ that exactly learns any $\mathbb{F}_2$-polynomial of degree at most ℓ (except with probability at most $\delta/2$). Roughly speaking, the algorithm draws $O(n)^{\ell}$ random examples and then solves an $\mathbb{F}_2$-linear system to determine the coefficients of the unknown polynomial; see Exercise 6.30 for details. Thus in time $n^{3(k-d)} \cdot \mathrm{poly}(n) \cdot \log(1/\delta)$ this algorithm will exactly determine f, and in particular find a relevant coordinate.

By choosing $d = \left\lceil \frac{3}{4}k \right\rceil$ we balance the running time of the two algorithms. Regardless of whether f is dth-order correlation immune, at least one of the two algorithms will find a relevant coordinate for f (except with probability at most $\delta/2 + \delta/2 = \delta$) in time $n^{(3/4)k} \cdot \mathrm{poly}(n) \cdot \log(1/\delta)$. This completes the proof of Theorem 6.36.

Our next application of pseudorandomness involves using ϵ-biased distributions to give a *deterministic* version of the Goldreich–Levin Algorithm (and hence the Kushilevitz–Mansour learning algorithm) for functions f with small $\hat{\lVert} f \hat{\rVert}_1$. We begin with a basic lemma showing that you can get a good estimate for the mean of such functions using an ϵ-biased distribution:

Lemma 6.38. *If $f : \{-1,1\}^n \to \mathbb{R}$ and $\varphi : \{-1,1\}^n \to \mathbb{R}$ is an ϵ-biased density, then*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\varphi}[f(\boldsymbol{x})] - \mathbf{E}[f] \right| \leq \hat{\lVert} f \hat{\rVert}_1 \epsilon.$$

This lemma follows from Proposition 6.13.(1), but we provide a separate proof:

Proof. By Plancherel,

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\varphi}[f(\boldsymbol{x})] = \langle \varphi, f\rangle = \widehat{f}(\emptyset) + \sum_{S\neq\emptyset} \widehat{\varphi}(S)\widehat{f}(S),$$

and the difference of this from $\mathbf{E}[f] = \widehat{f}(\emptyset)$ is, in absolute value, at most

$$\sum_{S\neq\emptyset} |\widehat{\varphi}(S)| \cdot |\widehat{f}(S)| \leq \epsilon \cdot \sum_{S\neq\emptyset} |\widehat{f}(S)| \leq \hat{\lVert} f \hat{\rVert}_1 \epsilon. \qquad \square$$

Since $\hat{\lVert} f^2 \hat{\rVert}_1 \leq \hat{\lVert} f \hat{\rVert}_1^2$ (Exercise 3.6), we also have the following immediate corollary:

Corollary 6.39. *If $f : \{-1,1\}^n \to \mathbb{R}$ and $\varphi : \{-1,1\}^n \to \mathbb{R}$ is an ϵ-biased density, then*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\varphi}[f(\boldsymbol{x})^2] - \mathbf{E}[f^2] \right| \leq \hat{\lVert} f \hat{\rVert}_1^2 \epsilon.$$

We can use the first lemma to get a deterministic version of Proposition 3.30, the learning algorithm that estimates a specified Fourier coefficient.

Proposition 6.40. *There is a* deterministic *algorithm that, given query access to a function $f : \{-1,1\}^n \to \mathbb{R}$ as well as $U \subseteq [n]$, $0 < \epsilon \leq 1/2$, and $s \geq 1$, outputs an estimate $\widetilde{f}(U)$ for $\widehat{f}(U)$ satisfying*

$$|\widetilde{f}(U) - \widehat{f}(U)| \leq \epsilon,$$

provided $\hat{\lVert} f \hat{\rVert}_1 \leq s$. The running time is $\mathrm{poly}(n, s, 1/\epsilon)$.

Proof. It suffices to handle the case $U = \emptyset$ because for general U, the algorithm can simulate query access to $f \cdot \chi_U$ with $\mathrm{poly}(n)$ overhead, and $\widehat{f \cdot \chi_U}(\emptyset) = \widehat{f}(U)$. The algorithm will use Theorem 6.30 to construct an (ϵ/s)-biased density φ that is uniform over a (multi-)set of cardinality $O(n^2 s^2/\epsilon^2)$. By enumerating over this set and using queries to f, it can deterministically output the estimate $\widetilde{f}(\emptyset) = \mathbf{E}_{\boldsymbol{x}\sim\varphi}[f(\boldsymbol{x})]$ in time $\mathrm{poly}(n,s,1/\epsilon)$. The error bound now follows from Lemma 6.38. □

The other key ingredient needed for the Goldreich–Levin Algorithm was Proposition 3.40, which let us estimate

$$\mathbf{W}^{S|\overline{J}}[f] = \sum_{T\subseteq\overline{J}} \widehat{f}(S\cup T)^2 = \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[\widehat{f_{J|\boldsymbol{z}}}(S)^2] \tag{6.6}$$

for any $S \subseteq J \subseteq [n]$. Observe that for any $z \in \{-1,1\}^{\overline{J}}$ we can use Proposition 6.40 to deterministically estimate $\widehat{f_{J|z}}(S)$ to accuracy $\pm\epsilon$. The reason is that we can simulate query access to the restricted function $\widehat{f_{J|z}}$, the (ϵ/s)-biased density φ remains (ϵ/s)-biased on $\{-1,1\}^{J}$, and most importantly $\hat{\lVert} f_{J|z} \hat{\rVert}_1 \leq \hat{\lVert} f \hat{\rVert}_1 \leq s$ by Exercise 3.7. It is not much more difficult to deterministically estimate (6.6):

Proposition 6.41. *There is a* deterministic *algorithm that, given query access to a function $f : \{-1,1\}^n \to \{-1,1\}$ as well as $S \subseteq J \subseteq [n]$, $0 < \epsilon \leq 1/2$, and $s \geq 1$, outputs an estimate β for $\mathbf{W}^{S|\overline{J}}[f]$ that satisfies*

$$|\mathbf{W}^{S|\overline{J}}[f] - \beta| \leq \epsilon,$$

provided $\hat{\lVert} f \hat{\rVert}_1 \leq s$. The running time is $\mathrm{poly}(n, s, 1/\epsilon)$.

Proof. Recall the notation $\mathrm{F}_{S|\overline{J}} f$ from Definition 3.20; by (6.6), the algorithm's task is to estimate $\mathbf{E}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[(\mathrm{F}_{S|\overline{J}} f)^2(\boldsymbol{z})]$. If $\varphi : \{-1,1\}^{\overline{J}} \to \mathbb{R}^{\geq 0}$ is an $\frac{\epsilon}{4s^2}$-biased density, Corollary 6.39 tells us that

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\varphi}[(\mathrm{F}_{S|\overline{J}} f)^2(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[(\mathrm{F}_{S|\overline{J}} f)^2(\boldsymbol{z})] \right| \leq \hat{\lVert} \mathrm{F}_{S|\overline{J}} f \hat{\rVert}_1^2 \cdot \tfrac{\epsilon}{4s^2} \leq \hat{\lVert} f \hat{\rVert}_1^2 \cdot \tfrac{\epsilon}{4s^2} \leq \tfrac{\epsilon}{4}, \quad (6.7)$$

where the second inequality is immediate from Proposition 3.21. We now show the algorithm can approximately compute $\mathbf{E}_{\boldsymbol{z}\sim\varphi}[(\mathrm{F}_{S|\overline{J}} f)^2(\boldsymbol{z})]$. For each $z \in \{-1,1\}^{\overline{J}}$, the algorithm can use φ to deterministically estimate $(\mathrm{F}_{S|\overline{J}} f)(z) = \widehat{f_{J|z}}(S)$ to within $\pm s \cdot \frac{\epsilon}{4s^2} \leq \frac{\epsilon}{4}$ in $\mathrm{poly}(n, s, 1/\epsilon)$ time, just as was described in the text following (6.6). Since $|\widehat{f_{J|z}}(S)| \leq 1$, the square of this estimate is within, say, $\frac{3\epsilon}{4}$ of $(\mathrm{F}_{S|\overline{J}} f)^2(z)$. Hence by enumerating over the support of φ, the algorithm can in deterministic $\mathrm{poly}(n, s, 1/\epsilon)$ time estimate $\mathbf{E}_{\boldsymbol{z}\sim\varphi}[(\mathrm{F}_{S|\overline{J}} f)^2(\boldsymbol{z})]$ to within $\pm\frac{3\epsilon}{4}$, which by (6.7) gives an estimate to within $\pm\epsilon$ of the desired quantity $\mathbf{E}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[(\mathrm{F}_{S|\overline{J}} f)^2(\boldsymbol{z})]$. □

Propositions 6.40 and 6.41 are the only two ingredients needed for a derandomization of the Goldreich–Levin Algorithm. We can therefore state a derandomized version of its corollary Theorem 3.38 on learning functions with small Fourier 1-norm:

Theorem 6.42. *Let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \hat{\lVert} f \hat{\rVert}_1 \leq s\}$. Then $\mathscr{C}$ is* deterministically *learnable from queries with error ϵ in time* $\mathrm{poly}(n, s, 1/\epsilon)$.

Since any $f : \{-1,1\}^n \to \{-1,1\}$ with $\mathrm{sparsity}(\widehat{f}) \leq s$ also has $\hat{\lVert} f \hat{\rVert}_1 \leq s$, we may also deduce from Exercise 3.37(c):

Theorem 6.43. *Let $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathrm{sparsity}(\widehat{f}) \leq 2^{O(k)}\}$. Then $\mathscr{C}$ is* deterministically *learnable exactly* (0 *error) from queries in time* $\mathrm{poly}(n, 2^k)$.

Example functions that fall into the concept classes of these theorems are decision trees of size at most s, and decision trees of depth at most k, respectively.

We conclude this section by discussing a derandomized version of the Blum–Luby–Rubinfeld linearity test from Chapter 1.6:

Derandomized BLR Test. *Given query access to* $f : \mathbb{F}_2^n \to \mathbb{F}_2$*:*

(1) Choose $\boldsymbol{x} \sim \mathbb{F}_2^n$ *and* $\boldsymbol{y} \sim \varphi$*, where* φ *is an* ϵ*-biased density.*

(2) Query f *at* $\boldsymbol{x}$*,* $\boldsymbol{y}$*, and* $\boldsymbol{x}+\boldsymbol{y}$*.*

(3) "Accept" if $f(\boldsymbol{x}) + f(\boldsymbol{y}) = f(\boldsymbol{x}+\boldsymbol{y})$*.*

Whereas the original BLR Test required exactly $2n$ independent random bits, the above derandomized version needs only $n + O(\log(n/\epsilon))$. This is very close to minimum possible; a test using only, say, $.99n$ random bits would only be able to inspect a $2^{-.01n}$ fraction of f's values.

If f is $\mathbb{F}_2$-linear then it is still accepted by the Derandomized BLR Test with probability 1. As for the approximate converse, we'll have to make a slight concession: We'll show that any function accepted with probability close to 1 must be close to an *affine* function, i.e., satisfy $\deg_{\mathbb{F}_2}(f) \le 1$. This concession is necessary: the function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ might be 1 everywhere except on the (tiny) support of φ. In that case the acceptance criterion $f(\boldsymbol{x}) + f(\boldsymbol{y}) = f(\boldsymbol{x}+\boldsymbol{y})$ will almost always be $1+0=1$; yet f is very far from every linear function. It is, however, very close to the affine function 1.

Theorem 6.44. *Suppose the Derandomized BLR Test accepts* $f : \mathbb{F}_2^n \to \mathbb{F}_2$ *with probability* $\frac{1}{2} + \frac{1}{2}\theta$*. Then* f *has correlation at least* $\sqrt{\theta^2 - \epsilon}$ *with some affine* $g : \mathbb{F}_2^n \to \mathbb{F}_2$*; i.e.,* $\mathrm{dist}(f,g) \le \frac{1}{2} - \frac{1}{2}\sqrt{\theta^2-\epsilon}$*.*

Remark 6.45. The bound in this theorem works well both when θ is close to 0 and when θ is close to 1; e.g., for $\theta = 1-2\delta$ we get that if f is accepted with probability $1-\delta$, then f is nearly δ-close to an affine function, provided $\epsilon \ll \delta$.

Proof. As in the analysis of the BLR Test (Theorem 1.30) we encode f's outputs by $\pm 1 \in \mathbb{R}$. Using the first few lines of that analysis we see that our hypothesis is equivalent to

$$\theta \le \mathop{\mathbf{E}}_{\substack{\boldsymbol{x}\sim\mathbb{F}_2^n \\ \boldsymbol{y}\sim\varphi}}[f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{x}+\boldsymbol{y})] = \mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[f(\boldsymbol{y})\cdot(f*f)(\boldsymbol{y})].$$

By Cauchy–Schwarz,

$$\mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[f(\boldsymbol{y})\cdot(f*f)(\boldsymbol{y})] \le \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[f(\boldsymbol{y})^2]}\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[(f*f)^2(\boldsymbol{y})]} = \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[(f*f)^2(\boldsymbol{y})]},$$

and hence

$$\theta^2 \leq \mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[(f*f)^2(\boldsymbol{y})] \leq \mathbf{E}[(f*f)^2] + \hat{\lVert} f*f \hat{\rVert}_1 \epsilon = \sum_{\gamma\in\widehat{\mathbb{F}_2^n}} \widehat{f}(\gamma)^4 + \epsilon,$$

where the inequality is Corollary 6.39 and we used $\widehat{f*f}(\gamma) = \widehat{f}(\gamma)^2$. The conclusion of the proof is as in the original analysis (cf. Proposition 6.7, Exercise 1.29):

$$\theta^2 - \epsilon \leq \sum_{\gamma\in\widehat{\mathbb{F}_2^n}} \widehat{f}(\gamma)^4 \leq \max_{\gamma\in\widehat{\mathbb{F}_2^n}}\{\widehat{f}(\gamma)^2\}\cdot \sum_{\gamma\in\widehat{\mathbb{F}_2^n}} \widehat{f}(\gamma)^2 = \max_{\gamma\in\widehat{\mathbb{F}_2^n}}\{\widehat{f}(\gamma)^2\},$$

and hence there exists γ^* such that $|\widehat{f}(\gamma^*)| \geq \sqrt{\theta^2-\epsilon}$. □

6.5. Highlight: Fooling $\mathbb{F}_2$-polynomials

Recall that a density φ is said to be ϵ-biased if its correlation with every $\mathbb{F}_2$-linear function f is at most ϵ in magnitude. In the lingo of pseudorandomness, one says that φ *fools* the class of $\mathbb{F}_2$-linear functions:

Definition 6.46. Let $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ be a density function and let $\mathscr{C}$ be a class of functions $\mathbb{F}_2^n \to \mathbb{R}$. We say that φ *ϵ-fools* $\mathscr{C}$ if

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[f(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[f(\boldsymbol{x})] \right| \leq \epsilon$$

for all $f \in \mathscr{C}$.

Theorem 6.30 implies that using just $O(\log(n/\epsilon))$ independent random bits, one can generate a density that ϵ-fools the class of $f : \mathbb{F}_2^n \to \{-1,1\}$ with $\deg_{\mathbb{F}_2}(f) \leq 1$. A natural problem in the field of derandomization is: How many independent random bits are needed to generate a density which ϵ-fools all functions of $\mathbb{F}_2$-degree at most d? A naive hope might be that ϵ-biased densities automatically fool functions of $\mathbb{F}_2$-degree $d > 1$. The next example shows that this hope fails badly, even for $d = 2$:

Example 6.47. Recall the inner product mod 2 function, $\mathrm{IP}_n : \mathbb{F}_2^n \to \{0,1\}$, which has $\mathbb{F}_2$-degree 2. Let $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ be the density of the uniform distribution on the support of IP_n. Now IP_n is an extremely regular function (see Example 6.4), and indeed φ is a roughly $2^{-n/2}$-biased density (see Exercise 6.7). But φ is very bad at fooling at least one function of $\mathbb{F}_2$-degree 2, namely IP_n itself:

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[\mathrm{IP}_n(\boldsymbol{x})] \approx 1/2, \qquad \mathop{\mathbf{E}}_{\boldsymbol{y}\sim\varphi}[\mathrm{IP}_n(\boldsymbol{y})] = 1.$$

The problem of using few random bits to fool n-bit, $\mathbb{F}_2$-degree-d functions was first taken up by Luby, Veličković, and Wigderson [**LVW93**]. They showed how to generate a fooling distribution using $\exp(O(\sqrt{d\log(n/d)+\log(1/\epsilon)}))$ independent random bits. There was no improvement on this for 14 years, at which point Bogdanov and Viola [BV07] achieved $O(\log(n/\epsilon))$ random bits for $d = 2$ and $O(\log n) + \exp(\mathrm{poly}(1/\epsilon))$ random bits for $d = 3$. In general, they suggested that $\mathbb{F}_2$-degree-d functions might be fooled by the sum of d independent draws from a small-bias distribution. Soon thereafter Lovett [Lov08] showed that a sum of 2^d independent draws from a small-bias distribution suffices, implying that $\mathbb{F}_2$-degree-d functions can be fooled using just $2^{O(d)} \cdot \log(n/\epsilon)$ random bits. More precisely, if φ is any ϵ-biased density on $\mathbb{F}_2^n$, Lovett showed that

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{y}^{(1)},\dots,\boldsymbol{y}^{(2^d)}\sim\varphi}[f(\boldsymbol{y}^{(1)}+\cdots+\boldsymbol{y}^{(2^d)})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[f(\boldsymbol{x})] \right| \le O(\epsilon^{1/4^d}).$$

In other words, the 2^d-fold convolution φ^{*2^d} density fools functions of $\mathbb{F}_2$-degree d.

The current state of the art for this problem is Viola's Theorem, which shows that the original idea of Bogdanov and Viola [BV07] works: Summing d independent draws from an ϵ-biased distribution fools $\mathbb{F}_2$-degree-d polynomials.

Viola's Theorem. *Let φ be any ϵ-biased density on $\mathbb{F}_2^n$, $0 \le \epsilon \le 1$. Let $d \in \mathbb{N}^+$ and define $\epsilon_d = 9\epsilon^{1/2^{d-1}}$. Then the class of all $f : \mathbb{F}_2^n \to \{-1,1\}$ with $\deg_{\mathbb{F}_2}(f) \le d$ is ϵ_d-fooled by the d-fold convolution φ^{*d}; i.e.,*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{y}^{(1)},\dots,\boldsymbol{y}^{(d)}\sim\varphi}[f(\boldsymbol{y}^{(1)}+\cdots+\boldsymbol{y}^{(d)})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[f(\boldsymbol{x})] \right| \le 9\epsilon^{1/2^{d-1}}.$$

In light of Theorem 6.30, Viola's Theorem implies that one can ϵ-fool n-bit functions of $\mathbb{F}_2$-degree d using only $O(d\log n) + O(d2^d\log(1/\epsilon))$ independent random bits.

The proof of Viola's Theorem is an induction on d. To reduce the case of degree $d+1$ to degree d, Viola makes use of a simple concept: directional derivatives.

Definition 6.48. Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ and let $y \in \mathbb{F}_2^n$. The *directional derivative* $\Delta_y f : \mathbb{F}_2^n \to \mathbb{F}_2$ is defined by

$$\Delta_y f(x) = f(x+y) - f(x).$$

Over $\mathbb{F}_2$ we may equivalently write $\Delta_y f(x) = f(x+y) + f(x)$.

As expected, taking a derivative reduces degree by 1:

Fact 6.49. *For any $f : \mathbb{F}_2^n \to \mathbb{F}_2$ and $y \in \mathbb{F}_2^n$ we have* $\deg_{\mathbb{F}_2}(\Delta_y f) \le \deg_{\mathbb{F}_2}(f) - 1$.

In fact, we'll prove a slightly stronger statement:

Proposition 6.50. *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ have $\deg_{\mathbb{F}_2}(f) = d$ and fix $y, y' \in \mathbb{F}_2^n$. Define $g : \mathbb{F}_2^n \to \mathbb{F}_2$ by $g(x) = f(x+y) - f(x+y')$. Then $\deg_{\mathbb{F}_2}(g) \leq d-1$.*

Proof. In passing from the $\mathbb{F}_2$-polynomial representation of $f(x)$ to that of $g(x)$, each monomial x^S of maximal degree d is replaced by $(x+y)^S - (x+y')^S$. Upon expansion the monomials x^S cancel, leaving a polynomial of degree at most $d-1$. □

We are now ready to give the proof of Viola's Theorem.

Proof of Viola's Theorem. The proof is by induction on d. The $d = 1$ case is immediate (even without the factor of 9) because φ is ϵ-biased. Assume that the theorem holds for general $d \geq 1$ and let $f : \mathbb{F}_2^n \to \{-1,1\}$ have $\deg_{\mathbb{F}_2}(f) \leq d+1$. We split into two cases, depending on whether the bias of f is large or small.

Case 1: $\mathbf{E}[f]^2 > \epsilon_d$. In this case,

$$\begin{aligned}
&\sqrt{\epsilon_d} \cdot \left| \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*(d+1)}}[f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[f(\boldsymbol{x})] \right| \\
&< |\mathbf{E}[f]| \cdot \left| \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*(d+1)}}[f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[f(\boldsymbol{x})] \right| \\
&= \left| \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim \mathbb{F}_2^n, \boldsymbol{z} \sim \varphi^{*(d+1)}}[f(\boldsymbol{x}')f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{x}', \boldsymbol{x} \sim \mathbb{F}_2^n}[f(\boldsymbol{x}')f(\boldsymbol{x})] \right| \\
&= \left| \mathop{\mathbf{E}}_{\boldsymbol{y} \sim \mathbb{F}_2^n, \boldsymbol{z} \sim \varphi^{*(d+1)}}[f(\boldsymbol{z}+\boldsymbol{y})f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{y}, \boldsymbol{x} \sim \mathbb{F}_2^n}[f(\boldsymbol{x}+\boldsymbol{y})f(\boldsymbol{x})] \right| \\
&= \left| \mathop{\mathbf{E}}_{\boldsymbol{y} \sim \mathbb{F}_2^n, \boldsymbol{z} \sim \varphi^{*(d+1)}}[\Delta_{\boldsymbol{y}} f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{y}, \boldsymbol{x} \sim \mathbb{F}_2^n}[\Delta_{\boldsymbol{y}} f(\boldsymbol{x})] \right| \\
&\leq \mathop{\mathbf{E}}_{\boldsymbol{y} \sim \mathbb{F}_2^n}\left[\left| \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*(d+1)}}[\Delta_{\boldsymbol{y}} f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[\Delta_{\boldsymbol{y}} f(\boldsymbol{x})] \right| \right].
\end{aligned}$$

For each outcome $\boldsymbol{y} = y$ the directional derivative $\Delta_y f$ has $\mathbb{F}_2$-degree at most d (Fact 6.49). By induction we know that φ^{*d} ϵ_d-fools any such polynomial, and it follows from Exercise 6.29 that $\varphi^{*(d+1)}$ does too. Thus each quantity in the expectation over $\boldsymbol{y}$ is at most ϵ_d, and we conclude

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*(d+1)}}[f(\boldsymbol{z})] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{F}_2^n}[f(\boldsymbol{x})] \right| \leq \frac{\epsilon_d}{\sqrt{\epsilon_d}} = \sqrt{\epsilon_d} = \tfrac{1}{3}\epsilon_{d+1} \leq \epsilon_{d+1}.$$

Case 2: $\mathbf{E}[f]^2 \leq \epsilon_d$. In this case we want to show that $\mathbf{E}_{\boldsymbol{w} \sim \varphi^{*(d+1)}}[f(\boldsymbol{w})]^2$ is nearly as small. By Cauchy–Schwarz,

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{w} \sim \varphi^{*(d+1)}}[f(\boldsymbol{w})]^2 &= \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*d}}\left[\mathop{\mathbf{E}}_{\boldsymbol{y} \sim \varphi}[f(\boldsymbol{z}+\boldsymbol{y})] \right]^2 \leq \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*d}}\left[\mathop{\mathbf{E}}_{\boldsymbol{y} \sim \varphi}[f(\boldsymbol{z}+\boldsymbol{y})]^2 \right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*d}}\left[\mathop{\mathbf{E}}_{\boldsymbol{y}, \boldsymbol{y}' \sim \varphi}[f(\boldsymbol{z}+\boldsymbol{y})f(\boldsymbol{z}+\boldsymbol{y}')] \right] = \mathop{\mathbf{E}}_{\boldsymbol{y}, \boldsymbol{y}' \sim \varphi}\left[\mathop{\mathbf{E}}_{\boldsymbol{z} \sim \varphi^{*d}}[f(\boldsymbol{z}+\boldsymbol{y})f(\boldsymbol{z}+\boldsymbol{y}')] \right].
\end{aligned}$$

For each outcome of $\boldsymbol{y} = y$, $\boldsymbol{y}' = y'$, the function $f(\boldsymbol{z}+y)f(\boldsymbol{z}+y')$ is of $\mathbb{F}_2$-degree at most d in the variables $\boldsymbol{z}$, by Proposition 6.50. Hence by induction we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{y},\boldsymbol{y}'\sim\varphi}\Big[\mathop{\mathbf{E}}_{\boldsymbol{z}\sim\varphi^{*d}}[f(\boldsymbol{z}+\boldsymbol{y})f(\boldsymbol{z}+\boldsymbol{y}')]\Big] &\le \mathop{\mathbf{E}}_{\boldsymbol{y},\boldsymbol{y}'\sim\varphi}\Big[\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[f(\boldsymbol{x}+\boldsymbol{y})f(\boldsymbol{x}+\boldsymbol{y}')]\Big]+\epsilon_d\\ &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{F}_2^n}[(\varphi*f)(\boldsymbol{x})^2]+\epsilon_d\\ &= \sum_{\gamma\in\widehat{\mathbb{F}_2^n}}\widehat{\varphi}(\gamma)^2\widehat{f}(\gamma)^2+\epsilon_d\\ &\le \widehat{f}(0)^2+\epsilon^2\sum_{\gamma\neq 0}\widehat{f}(\gamma)^2+\epsilon_d\\ &\le 2\epsilon_d+\epsilon^2,\end{aligned}$$

where the last step used the hypothesis of Case 2. We have thus shown

$$\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\varphi^{*(d+1)}}[f(\boldsymbol{w})]^2 \le 2\epsilon_d+\epsilon^2 \le 3\epsilon_d \le 4\epsilon_d,$$

and hence $|\mathbf{E}[f(\boldsymbol{w})]| \le 2\sqrt{\epsilon_d}$. Since we are in Case 2, $|\mathbf{E}[f]| \le \sqrt{\epsilon_d}$, and so

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\varphi^{*(d+1)}}[f(\boldsymbol{w})]-\mathbf{E}[f]\right| \le 3\sqrt{\epsilon_d} = \epsilon_{d+1},$$

as needed. □

We end this section by discussing the tightness of parameters in Viola's Theorem. First, if we ignore the error parameter, then the result is sharp: a counting argument (see [BV07]) shows that the d-fold convolution of ϵ-biased densities cannot in general fool functions of $\mathbb{F}_2$-degree $d+1$. More explicitly, for any $d\in\mathbb{N}^+$, $\ell \ge 2d+1$, Lovett and Tzur [LT09] gave an explicit $\frac{\ell}{2^n}$-biased density on $\mathbb{F}_2^{(\ell+1)n}$ and an explicit function $f : \mathbb{F}_2^{(\ell+1)n} \to \{-1,1\}$ of degree $d+1$ for which

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\varphi^{*d}}[f(\boldsymbol{w})]-\mathbf{E}[f]\right| \ge 1-\frac{2d}{2^n}.$$

Regarding the error parameter in Viola's Theorem, it is not known whether the quantity $\epsilon^{1/2^{d-1}}$ can be improved, even in the case $d=2$. However, obtaining even a modest improvement to $\epsilon^{1/1.99^d}$ (for d as large as $\log n$) would constitute a major advance since it would imply progress on the notorious problem of "correlation bounds for polynomials"; see Viola [Vio09].

6.6. Exercises and notes

6.1 Let $\boldsymbol{f}$ be chosen as in Proposition 6.1. Compute $\mathbf{Var}[\widehat{\boldsymbol{f}}(S)]$ for each $S\subseteq[n]$.

6.2 Prove Fact 6.8.

6.3 Show that any nonconstant k-junta has $\mathbf{Inf}_i^{(1-\delta)}[f] \ge (1/2-\delta/2)^{k-1}/k$ for at least one coordinate i.

6.4 Let $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ be an ϵ-biased density. For each $d \in \mathbb{N}^+$ show that the d-fold convolution φ^{*d} is an ϵ^d-biased density.

6.5 (*a*) Show that if $f : \{-1,1\}^n \to \mathbb{R}$ has ϵ-small influences, then it is $\sqrt{\epsilon}$-regular.

(*b*) Show that for all even n there exists $f : \{-1,1\}^n \to \{-1,1\}$ that is $2^{-n/2}$-regular but does not have ϵ-small influences for any $\epsilon < 1/2$.

(*c*) Show that there is a function $f : \{-1,1\}^n \to \{-1,1\}$ with $((1-\delta)^{n-1}, \delta)$-small stable influences that is not ϵ-regular for any $\epsilon < 1$.

(*d*) Verify that the function $f(x) = x_0 \mathrm{Maj}_n(x_1, \dots, x_n)$ from Example 6.10 satisfies $\mathbf{Inf}_0^{(1-\delta)}[f] = \mathbf{Stab}_{1-\delta}[\mathrm{Maj}_n]$ for $\delta \in (0,1)$, and thus does not have (ϵ,δ)-small stable influences unless $\epsilon \geq 1 - \sqrt{\delta}$.

(*e*) Show that the function $f : \{-1,1\}^{n+1} \to \{-1,1\}$ from part (*d*) is $\frac{1}{\sqrt{n}}$-regular.

(*f*) Suppose $f : \{-1,1\}^n \to \mathbb{R}$ has (ϵ,δ)-small stable influences. Show that f is (η, k)-regular for $\eta = \sqrt{\epsilon/(1-\delta)^{k-1}}$.

(*g*) Show that f has $(\epsilon, 1)$-small stable influences if and only if f is $(\sqrt{\epsilon}, 1)$-regular.

(*h*) Let $f : \{-1,1\}^n \to \{-1,1\}$ be monotone. Show that if f is $(\epsilon,1)$-regular then f is ϵ-regular and has ϵ-small influences.

6.6 (*a*) Let $f : \{-1,1\}^n \to \mathbb{R}$. Let $(J, \overline{J})$ be a partition of $[n]$ and let $z \in \{-1,1\}^{\overline{J}}$. For $\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}$ uniformly random, give a formula for $\mathbf{Var}_{\boldsymbol{z}}[\mathbf{E}[f_{J|\boldsymbol{z}}]]$ in terms of f's Fourier coefficients. (Hint: Direct application of Corollary 3.22.)

(*b*) Using the above formula and the probabilistic method, give an alternate proof of the second statement of Proposition 6.12.

6.7 Let $\varphi : \mathbb{F}_2^n \to \mathbb{R}^{\geq 0}$ be the density corresponding to the uniform distribution on the support of $\mathrm{IP}_n : \mathbb{F}_2^n \to \{0,1\}$. Show that φ is ϵ-biased for $\epsilon = 2^{-n/2}/(1 - 2^{-n/2})$, but not for smaller ϵ.

6.8 Prove Proposition 6.13.

6.9 Compute the $\mathbb{F}_2$-polynomial representation of the equality function $\mathrm{Equ}_n : \{0,1\}^n \to \{0,1\}$, defined by $\mathrm{Equ}_n(x) = 1$ if and only if $x_1 = x_2 = \cdots = x_n$.

6.10 (*a*) Let $f : \{0,1\}^n \to \mathbb{R}$ and let $q(x) = \sum_{S \subseteq [n]} c_S x^S$ be the (unique) multilinear polynomial representation of f over $\mathbb{R}$. Show that $c_S = \sum_{R \subseteq S} (-1)^{|S|-|R|} f(R)$, where we identify $R \subseteq [n]$ with its 0-1 indicator string. This formula is sometimes called *Möbius inversion*.

(*b*) Prove Proposition 6.21.

6.11 (Cf. Lemma 3.5.) Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be nonzero and suppose $\deg_{\mathbb{F}_2}(f) \leq k$. Show that $\mathbf{Pr}[f(\boldsymbol{x}) \neq 0] \geq 2^{-k}$. (Hint: As in the similar Exercise 3.4, use induction on n.)

6.12 Let $f : \{-1,1\}^n \to \{0,1\}$.

(*a*) Show that $\deg_{\mathbb{F}_2}(f) \leq \log(\text{sparsity}(\widehat{f}))$. (Hint: You will need Exercise 3.7, Corollary 6.22, and Exercise 1.3.)

(*b*) Suppose $\widehat{f}$ is 2^{-k}-granular. Show that $\deg_{\mathbb{F}_2}(f) \leq k$. (This is a stronger result than part (*a*), by Exercise 3.32.)

6.13 Let $f : \{-1,1\}^n \to \{-1,1\}$ be bent, $n > 2$. Show that $\deg_{\mathbb{F}_2}(f) \leq n/2$. (Note that the upper bound $n/2+1$ follows from Exercise 6.12(*b*).)

6.14 In this exercise you will prove Theorem 6.25.

(*a*) Suppose $p(x) = c_0 + c_S x^S + r(x)$ is a real multilinear polynomial over $x_1, \dots, x_n$ with $c_0, c_S \neq 0$, $|S| > \frac{2}{3}n$, and $|T| > \frac{2}{3}n$ for all monomials x^T appearing in $r(x)$. Show that after expansion and multilinear reduction (meaning $x_i^2 \mapsto 1$), $p(x)^2$ contains the term $2c_0c_Sx^S$.

(*b*) Deduce Theorem 6.25.

6.15 In this exercise you will explore the sharpness of Siegenthaler's Theorem and Theorem 6.25.

(*a*) For all n and $k < n-1$, find an $f : \{0,1\}^n \to \{0,1\}$ that is k-resilient and has $\deg_{\mathbb{F}_2}(f) = n-k-1$.

(*b*) For all $n \geq 3$, find an $f : \{0,1\}^n \to \{0,1\}$ that is 1st-order correlation immune and has $\deg_{\mathbb{F}_2}(f) = n-1$.

(*c*) For all n divisible by 3, find a biased $f : \{0,1\}^n \to \{0,1\}$ that is $(\frac{2}{3}n-1)$th-order correlation immune.

6.16 Prove Proposition 6.27.

6.17 Bent functions come in pairs: Show that if $f : \mathbb{F}_2^n \to \{-1,1\}$ is bent, then $2^{n/2}\widehat{f}$ is also a bent function (with domain $\widehat{\mathbb{F}_2^n}$).

6.18 Extend Proposition 6.29 to show that if π is any permutation on $\mathbb{F}_2^n$, then $f(x,y) = \mathrm{IP}_{2n}(x,\pi(y))g(y)$ is bent.

6.19 *Dickson's Theorem* says the following: Any polynomial $p : \mathbb{F}_2^n \to \mathbb{F}_2$ of degree at most 2 can be expressed as

$$p(x) = \ell_0(x) + \sum_{j=1}^{k} \ell_j(x)\ell'_j(x), \tag{6.8}$$

where ℓ_0 is an affine function and $\ell_1, \ell'_1, \dots, \ell_k, \ell'_k$ are linearly independent linear functions. Here k depends only on p and is called the "rank" of p. Show that for n even, $g : \mathbb{F}_2^n \to \{-1,1\}$ defined by $g(x) = \chi(p(x))$ is bent if and only if $k = n/2$, if and only if g arises from IP_n as in Proposition 6.28.

6.20 Without appealing to Dickson's Theorem, prove that the complete quadratic $x \mapsto \sum_{1 \leq i < j \leq n} x_i x_j$ can be expressed as in (6.8), with $k = \lfloor n/2 \rfloor$. (Hint: Induction on n, with different steps depending on the parity of n.)

6.21 Define $\mathrm{mod}_3 : \{-1,1\}^n \to \{0,1\}$ by $\mathrm{mod}_3(x) = 1$ if and only if $\sum_{j=1}^n x_i$ is divisible by 3. Derive the Fourier expansion

$$\mathrm{mod}_3(x) = \tfrac{1}{3} + \tfrac{2}{3}(-1/2)^n \sum_{\substack{S\subseteq[n]\\ |S| \text{ even}}} (-1)^{(|S| \bmod 4)/2} \sqrt{3}^{|S|} x^S$$

and conclude that mod_3 is $\frac{2}{3}(\frac{\sqrt{3}}{2})^n$-regular. (Hint: Consider $\prod_{j=1}^n(-\frac{1}{2} + \frac{\sqrt{-3}}{2})x_j$.)

6.22 In Theorem 6.30, show that given $\boldsymbol{r}, \boldsymbol{s}$ any fixed bit $\boldsymbol{y}_i$ can be obtained in deterministic poly(ℓ) time.

6.23 (*a*) Slightly modify the construction in Theorem 6.30 to obtain a $(2^{-t} - 2^{-\ell})$-biased density. (Hint: Arrange for p_γ to have degree at most $n-1$.)

(*b*) Since $\mathbb{F}_{2^\ell}$ is a dimension-ℓ vector space over $\mathbb{F}_2$, it has some basis $v_1, \dots, v_\ell$. Suppose we modify the construction in Theorem 6.30 so that φ is a density on $\mathbb{F}_2^{n\ell}$, with $\boldsymbol{y}_{ij} = \langle \mathrm{enc}(v_j \boldsymbol{r}^i), \mathrm{enc}(\boldsymbol{s})\rangle$ for $i \in [n], j \in [\ell]$. Show that φ remains 2^{-t}-biased.

6.24 Fix $\epsilon \in (0,1)$ and $n \in \mathbb{N}$. Let $A \subseteq \mathbb{F}_2^n$ be a randomly chosen multiset in which $\lceil Cn/\epsilon^2 \rceil$ elements are included, independently and uniformly. Show that if C is a large enough constant, then A is ϵ-biased except with probability at most 2^{-n}.

6.25 Consider the problem of computing the matrix multiplication $C = AB$, where $A, B \in \mathbb{F}_2^{n\times n}$. There is an algorithm [LG14] for solving this problem in time $O(n^\omega)$, where $\omega < 2.373$; however, the algorithm is very complicated. Suppose you are given A, B, and the outcome C' of running this algorithm; you want to test that indeed $C' = AB$.

(*a*) Give an algorithm using n random bits and time $O(n^2)$ with the following property: If $C' = AB$, then the algorithm "accepts" with probability 1; if $C' \neq AB$, then the algorithm "accepts" with probability at most 1/2. (Hint: Compute $C'\boldsymbol{x}$ and $AB\boldsymbol{x}$ for a random $\boldsymbol{x} \in \mathbb{F}_2^n$.)

(*b*) Show how to reduce the number of random bits used to $O(\log n)$ at the expense of making the false acceptance probability 2/3, while keeping the running time $O(n^2)$. (You may use the fact that in Theorem 6.30, the time required to compute $\boldsymbol{y}$ given $\boldsymbol{r}$ and $\boldsymbol{s}$ is $n \cdot \mathrm{polylog}(\ell)$.)

6.26 Simplify the exposition and analysis of Theorem 6.32 and Corollary 6.33 in the case of $k = 2$, and show that you can take m to be one less (i.e., $m = \ell$).

6.27 Consider the matrix $H' \in \mathbb{F}_n^{k\times n}$ constructed in Theorem 6.32, and suppose we delete all rows corresponding to even (nonzero) powers of the α_j's. Show that H' retains the property that any sum of at most k columns of H' is nonzero in $\mathbb{F}_n^k$. (Hint: Prove and use that $(\sum_j \beta_j)^2 = \sum_j \beta_j^2$ for any sequence of $\beta_j \in \mathbb{F}_n$.) Deduce that the cardinality of A in Corollary 6.33 can be decreased to $2(2n)^{\lfloor k/2 \rfloor}$.

6.28 Let $A \subseteq \{-1,1\}^n$ be a multiset and suppose that the probability density ϕ_A is k-wise independent. In this exercise you will prove the lower bound $|A| \geq \Omega(n^{\lfloor k/2 \rfloor})$ (for k constant).

(a) Suppose $\mathscr{F} \subseteq 2^{[n]}$ is a collection of subsets of $[n]$ such that $|S \cup T| \leq k$ for all $S, T \in \mathscr{F}$. For each $S \in \mathscr{F}$ define $\chi_S^A \in \{-1,1\}^{|A|} \subseteq \mathbb{R}^{|A|}$ to be the real vector with entries indexed by A whose ath entry is $a^S = \prod_{i \in S} a_i$. Show that the set of vectors $\{\frac{1}{\sqrt{|A|}}\chi_S^A : S \in \mathscr{F}\}$ is orthonormal and hence $|A| \geq |\mathscr{F}|$.

(b) Show that we can find $\mathscr{F}$ satisfying $|\mathscr{F}| \geq \sum_{j=0}^{k/2}\binom{n}{j}$ if k is even and $|\mathscr{F}| \geq \sum_{j=0}^{(k-1)/2}\binom{n}{j} + \binom{n-1}{(k-1)/2}$ if k is odd.

6.29 Let $\mathscr{C}$ be a class of functions $\mathbb{F}_2^n \to \mathbb{R}$ that is closed under translation; i.e., $f^{+z} \in \mathscr{C}$ whenever $f \in \mathscr{C}$ and $z \in \mathbb{F}_2^n$ (recall Definition 3.24). An example is the class of functions of $\mathbb{F}_2$-degree at most d. Show that if ψ is a density that ϵ-fools $\mathscr{C}$, then $\psi * \varphi$ also ϵ-fools $\mathscr{C}$ for any density φ.

6.30 Fix an integer $\ell \geq 1$. In this exercise you will generalize Exercise 3.43 by showing how to exactly learn $\mathbb{F}_2$-polynomials of degree at most ℓ.

(a) Fix $p : \mathbb{F}_2^n \to \mathbb{F}_2$ with $\deg_{\mathbb{F}_2}(p) \leq \ell$ and suppose that $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)} \sim \mathbb{F}_2^n$ are drawn uniformly and independently from $\mathbb{F}_2^n$. Assume that $m \geq C \cdot 2^\ell(n^\ell + \log(1/\delta))$ for $0 < \delta \leq 1/2$ and C a sufficiently large constant. Show that except with probability at most δ, the only $q : \mathbb{F}_2^n \to \mathbb{F}_2$ with $\deg_{\mathbb{F}_2}(q) \leq \ell$ that satisfies $q(\boldsymbol{x}^{(i)}) = p(\boldsymbol{x}^{(i)})$ for all $i \in [m]$ is $q = p$. (Hint: Exercise 6.11 with $q - p$.)

(b) Show that the concept class of all polynomials $\mathbb{F}_2^n \to \mathbb{F}_2$ of degree at most ℓ can be learned from random examples only, with error 0, in time $O(n)^{3\ell}$. (Remark: As in Exercise 3.43, since the key step is solving a linear system, the learning algorithm can also be done in $O(n)^{\omega\ell}$ time, assuming matrix multiplication can be done in $O(n^\omega)$ time.)

(c) Extend this learning algorithm so that in running time $O(n)^{3\ell} \cdot \log(1/\delta)$ it achieves success probability at least $1 - \delta$. (Hint: Similar to Exercise 3.40.)

6.31 In this exercise you will prove Lemma 6.37.

(a) Give a $\mathrm{poly}(n, 2^k) \cdot \log(1/\delta)$-time learning algorithm that, given random examples from a k-junta $\mathbb{F}_2^n \to \mathbb{F}_2$, determines (except with probability at most δ) if f is a constant function, and if so, which one.

(b) Given access to random examples from a k-junta $f : \mathbb{F}_2^n \to \mathbb{F}_2$, let $P \subseteq [n]$ be a set of relevant coordinates for f and let $z \in \mathbb{F}_2^P$. Show

how to obtain M independent random examples from the $(k-|P|)$-junta $f_{\overline{P}|z}$ in time $\mathrm{poly}(n,2^k)\cdot M\cdot \log(1/\delta)$ (except with probability at most δ).

(c) Complete the proof of Lemma 6.37. (Hint: Build a depth-k decision tree for f.)

6.32 (a) Improve the bound in Lemma 6.38 to $\hat{\lVert} f \hat{\rVert}_1\epsilon - |\widehat{f}(\emptyset)|\epsilon$ and the bound in Corollary 6.39 to $\hat{\lVert} f \hat{\rVert}_1^2\epsilon - \lVert f\rVert_2^2\epsilon$.

(b) Improve the bound in Theorem 6.44 to $\sqrt{\theta^2-\epsilon}/\sqrt{1-\epsilon}$.

6.33 Improve on Theorem 6.44 by a factor of roughly 2 in the case of acceptance probability near 1. Specifically, show that if f passes the Derandomized BLR Test with probability $1-\delta$, then there exists $\gamma^* \in \widehat{\mathbb{F}_2^n}$ with $|\widehat{f}(\gamma^*)| \geq \sqrt{1-2\delta-\epsilon}/\sqrt{1-\epsilon}$.

6.34 Fix an integer $k \in \mathbb{N}^+$. Let $(f_s)_{s\in\{0,1\}^k}$ be a collection of functions indexed by length-k binary sequences, each $f_s : \mathbb{F}_2^n \to \mathbb{R}$. Define the *kth Gowers "inner product"* $\langle (f_s)_s\rangle_{U^k} \in \mathbb{R}$ by

$$\langle (f_s)_s\rangle_{U^k} = \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}_1,\dots,\boldsymbol{y}_k}\left[\prod_{s\in\{0,1\}^k} f_s(\boldsymbol{x}+\sum_{i:s_i=1}\boldsymbol{y}_i)\right],$$

where the $k+1$ random vectors $\boldsymbol{x},\boldsymbol{y}_1,\dots,\boldsymbol{y}_k$ are independent and uniformly distributed on $\mathbb{F}_2^n$. Define the *kth Gowers norm* of a function $f : \mathbb{F}_2^n \to \mathbb{R}$ by

$$\lVert f\rVert_{U^k} = \langle (f,f,\dots,f)\rangle_{U^k}^{1/2^k},$$

where $(f,f,\dots,f)$ denotes that all 2^k functions in the collection equal f. (You will later verify that $\langle (f,f,\dots,f)\rangle_{U^k}$ is always nonnegative.)

(a) Check that $\langle f_0,f_1\rangle_{U^1} = \mathbf{E}[f_0]\mathbf{E}[f_1]$ and therefore $\lVert f\rVert_{U^1}^2 = \mathbf{E}[f]^2$.

(b) Check that

$$\langle f_{00},f_{10},f_{01},f_{11}\rangle_{U^2} = \sum_{\gamma\in\widehat{\mathbb{F}_2^n}} \widehat{f_{00}}(\gamma)\widehat{f_{10}}(\gamma)\widehat{f_{01}}(\gamma)\widehat{f_{11}}(\gamma)$$

and therefore $\lVert f\rVert_{U^2}^4 = \hat{\lVert} f\hat{\rVert}_4^4$. (Cf. Exercise 1.29(*b*).)

(c) Show that

$$\langle (f_s)_s\rangle_{U^k} = \mathop{\mathbf{E}}_{\boldsymbol{y}_1,\dots,\boldsymbol{y}_{k-1}}\left[\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\prod_{s:s_k=0} f_s(\boldsymbol{x}+\sum_{i:s_i=1}\boldsymbol{y}_i)\right]\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}'}\left[\prod_{s:s_k=1} f_s(\boldsymbol{x}'+\sum_{i:s_i=1}\boldsymbol{y}_i)\right]\right], \tag{6.9}$$

where $\boldsymbol{x}'$ is independent of $\boldsymbol{x},\boldsymbol{y}_1,\dots,\boldsymbol{y}_{k-1}$ and uniformly distributed.

(d) Show that $\langle (f,f,\dots,f)\rangle_{U^k}$ is always nonnegative, as promised.

(e) Using (6.9) and Cauchy–Schwarz, show that

$$\langle (f_s)_s\rangle_{U^k} \leq \sqrt{\langle (f_{(s_1,\dots,s_{k-1},0)})_S\rangle_{U^k}}\sqrt{\langle (f_{(s_1,\dots,s_{k-1},1)})_s\rangle_{U^k}}.$$

(*f*) Show that

$$\langle (f_s)_s \rangle_{U^k} \leq \prod_{s \in \{0,1\}^k} \|f_s\|_{U^k}. \tag{6.10}$$

(*g*) Fixing $f : \mathbb{F}_2^n \to \mathbb{R}$, show that $\|f\|_{U^k} \leq \|f\|_{U^{k+1}}$. (Hint: Consider $(f_s)_{s \in \{0,1\}^{k+1}}$ defined by $f_s = f$ if $s_{k+1} = 0$ and $f_s = 1$ if $s_{k+1} = 1$.)

(*h*) Show that $\|\cdot\|_{U^k}$ satisfies the triangle inequality and is therefore a seminorm. (Hint: First show that

$$\|f_0 + f_1\|_{U^k}^{2^k} = \sum_{S \subseteq \{0,1\}^k} \langle (f_{\mathbf{1}[s \in S]})_{s \in \{0,1\}^k} \rangle_{U^k}$$

and then use (6.10).)

(*i*) Show that $\|\cdot\|_{U^k}$ is in fact a norm for all $k \geq 2$; i.e., $\|f\|_{U^k} = 0 \implies f = 0$.

Notes. The $\mathbb{F}_2$-polynomial representation of a Boolean function f is often called its algebraic normal form. It seems to have first been explicitly introduced by Zhegalkin in 1927 [**Zhe27**].

For functions $f : \mathbb{Z}_n \to \mathbb{R}$, the idea of ϵ-regularity as a pseudorandomness notion dates back to Chung and Graham [**CG92**], as does the equivalent combinatorial condition Proposition 6.7. (In the context of quasirandom graphs, the ideas date further back to Thomason [**Tho87**] and to Chung, Graham, and Wilson [**CGW89**].) The idea of treating functions with small (stable) influences as being "generic" has its origins in the work of Kahn, Kalai, and Linial [**KKL88**]. The notion was brought to the fore in work on hardness of approximation – implicitly, by Håstad [**Hås96, Hås99**], and later more explicitly by Khot, Kindler, Mossel, and O'Donnell [**KKMO07**].

The notion of ϵ-biased sets (and also (ϵ, k)-wise independent distributions) was introduced by Naor and Naor [**NN93**] (see also the independent work of Peralta [**Per90**]). The construction in Theorem 6.30 is due to Alon, Goldreich, Håstad, and Peralta [**AGHP92**] (as is Exercise 6.23). As noted by Naor and Naor [**NN93**], ϵ-biased sets are closely related to error-correcting codes over $\mathbb{F}_2$; indeed, they are equivalent to linear error-correcting in which all pairs of codewords have relative distance in $[\frac{1}{2} - \frac{1}{2}\epsilon, \frac{1}{2} + \frac{1}{2}\epsilon]$. In particular, the construction in Theorem 6.30 is the concatenation of the well-known Reed–Solomon and Hadamard codes (see, e.g., MacWilliams and Sloane [**MS77**] for definitions). The nonconstructive upper bound in Exercise 6.24 is essentially the Gilbert–Varshamov bound and is close to known lower bound of $\Omega(\frac{n}{\epsilon^2 \log(1/\epsilon)})$ (assuming $\epsilon \geq 2^{-\Omega(n)}$), which follows from the work of McEliece, Rodemich, Rumsey, and Welch [**MRRW77**] (see [**MS77**]). Additionally, constructive upper bounds of $O(\frac{n}{\epsilon^3})$ and $O(\frac{n^{5/4}}{\epsilon^{5/2}})$ are known using tools from coding theory; see the work of Ben-Aroya and Ta-Shma [**BT09**] and Matthews and Peachey [**MP11**].

The probabilistic notion of correlation immunity – i.e., condition (2) of Corollary 6.14 – was first introduced by Siegenthaler [**Sie84**]; we further discuss his work below. Independently and shortly thereafter, Chor, Friedman, Goldreich, Håstad, Rudich, and Smolensky [**CFG$^+$85**] introduced the definition of resilience and also connected it to $(0,k)$-regularity of the Fourier spectrum; i.e., they proved Corollary 6.14. (In the cryptography literature, Corollary 6.14 is called the Xiao–Massey Theorem [**XM88**].) The work [**CFG$^+$85**] also essentially contains Theorem 6.25 and the relevant function from Example 6.16; cf. the work of Mossel et al. [**MOS04**].

The problem of constructing explicit k-wise distributions of small support arose in different guises in different areas – in the study of orthogonal arrays (in statistics), error-correcting codes, and algorithmic derandomization. Alon, Babai, and Itai [**ABI85**] gave the construction in Theorem 6.32 – in fact, the stronger one from Exercise 6.27 – based on the analysis of dual BCH codes in MacWilliams and Sloane [**MS77**]. The lower bound from Exercise 6.28 is essentially due to Rao [**Rao47**]; see also independent proofs [**CFG$^+$85, ABI85**].

Siegenthaler's Theorem dates from 1984 [**Sie84**]. His motivation was the study of cryptographic stream ciphers in cryptography. In this application, a short random sequence of bits ("secret key") is transformed via some scheme into a very long sequence of pseudorandom bits ("keystream"), which can then be used as a one-time pad for encryption. A basic component of most schemes is a linear feedback shift register (LFSR), which can efficiently generate long, fairly statistically-uniform sequences. However, due to its $\mathbb{F}_2$-linearity, it suffers from some simple cryptanalytic attacks. An early idea for combating this is to take n independent LFSR streams and combine them via some function $f : \mathbb{F}_2^n \to \mathbb{F}_2$. Effective attacks are possible in such a scheme if f is correlated with any of its input bits – or indeed (as Siegenthaler pointed out) any input pair, triple, etc. This led Siegenthaler to define the probabilistic notion of correlation-immunity. Although $\chi_{[n]}$ is the maximally correlation-immune function, it is not suitable as a LFSR combining function precisely because of its $\mathbb{F}_2$-linearity; the same is true of any function of low $\mathbb{F}_2$-degree. Siegenthaler precisely captured this tradeoff between correlation-immunity and $\mathbb{F}_2$-degree in his theorem.

Bent functions were named and first studied by Rothaus around 1966; he didn't publish the notion until 1976, however [**Rot76**], at which point there were already several works on subject, see, e.g., [**Dil72**]. Bent functions have application in cryptography and coding theory; see, e.g., Carlet's survey [**Car10**]. The basic constructions presented in Section 6.3 are due to Rothaus; the class of bent functions described in Exercise 6.18 is called the Maiorana–McFarland family. Dickson's Theorem is from a 1901 publication [Dic01, Theorem 199]; see also MacWilliams and Sloane [MS77, Theorem 15.4].

Theorem 6.36 is from Mossel et al. [MOS04]; there is an improved algorithm for learning k-juntas that runs in time roughly $n^{.6024k}\mathrm{poly}(n)$, due to Gregory Valiant [Val12]. Avrim Blum offers a prize of $1,000 for solving the case of $k = \log\log n$ in poly(n) time [Blu03]. Theorem 6.42 is due to Kushilevitz and Mansour [KM93]. The Derandomized BLR Test and Theorem 6.44 (and Exercise 6.32) are due to Ben-Sasson, Sudan, Vadhan, and Wigderson [BSSVW03].

The result of Exercise 6.11 is due to Muller [Mul54a, Theorem 6]; deriving Exercise 6.30 from it and from Blumer et al. [BEHW87] is folklore. The result of Exercise 6.12(*a*) is due to Bernasconi and Codenotti [BC99]; Exercise 6.13 is from MacWilliams and Sloane [MS77]. In Exercise 6.25, part (*a*) is due to Freivalds [Fre79] and part (*b*) to Naor and Naor [NN93]. The Gowers norm and results of Exercise 6.34 are from Gowers [Gow01]. Our proof of the second statement in Proposition 6.12 was suggested by Noam Lifshitz.

Chapter 7

Property testing, PCPPs, and CSPs

In this chapter we study several closely intertwined topics: property testing, probabilistically checkable proofs of proximity (PCPPs), and constraint satisfaction problems (CSPs). All of our work will be centered around the task of testing whether an unknown Boolean function is a dictator. We begin by extending the BLR Test to give a 3-query property testing algorithm for the class of dictator functions. This in turn allows us to give a 3-query testing algorithm for *any* property, so long as the right "proof" is provided. We then introduce CSPs, which are in fact identical to string testing algorithms. Finally, we explain how dictator tests can be translated into computational complexity results for CSPs, and we sketch the proofs of some of Håstad's optimal inapproximability results.

7.1. Dictator testing

In Chapter 1.6 we described the BLR property testing algorithm: Given query access to an unknown function $f : \{0,1\}^n \to \{0,1\}$, this algorithm queries f on a few random inputs and approximately determines whether f has the property of being linear over $\mathbb{F}_2$. The field of *property testing* for Boolean functions is concerned with coming up with similar algorithms for other properties. In general, a "property" can be any collection $\mathscr{C}$ of n-bit Boolean functions; it's the same as the notion of "concept class" from learning theory. Indeed, before running an algorithm to try to learn an unknown $f \in \mathscr{C}$, one might first run a property testing algorithm to try to verify that indeed $f \in \mathscr{C}$.

Let's encapsulate the key aspects of the BLR linearity test with some definitions:

Definition 7.1. An *r-query function testing algorithm* for Boolean functions $f : \{0,1\}^n \to \{0,1\}$ is a randomized algorithm that:

- chooses r (or fewer) strings $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)} \in \{0,1\}^n$ according to some probability distribution;
- queries $f(\boldsymbol{x}^{(1)}), \dots, f(\boldsymbol{x}^{(r)})$;
- based on the outcomes, decides (deterministically) whether to "accept" f.

Definition 7.2. Let $\mathscr{C}$ be a "property" of n-bit Boolean functions, i.e., a collection of functions $\{0,1\}^n \to \{0,1\}$. We say a function testing algorithm is a *local tester for $\mathscr{C}$* (with *rejection rate* $\lambda > 0$) if it satisfies the following:

- If $f \in \mathscr{C}$, then the tester accepts with probability 1.
- For all $0 \le \epsilon \le 1$, if $\mathrm{dist}(f, \mathscr{C}) > \epsilon$ (in the sense of Definition 1.29), then the tester rejects f with probability greater than $\lambda \cdot \epsilon$.

  Equivalently, if the tester accepts f with probability at least $1 - \lambda \cdot \epsilon$, then f is ϵ-close to $\mathscr{C}$; i.e., $\exists g \in \mathscr{C}$ such that $\mathrm{dist}(f, g) \le \epsilon$.

By taking $\epsilon = 0$ in the above definition you see that any local tester gives a characterization of $\mathscr{C}$: a function is in $\mathscr{C}$ if and only if it is accepted by the tester with probability 1. But a local tester furthermore gives a "robust" characterization: Any function accepted with probability *close* to 1 must be *close* to satisfying $\mathscr{C}$.

Example 7.3. By Theorem 1.30, the BLR Test is a 3-query local tester for the property $\mathscr{C} = \{f : \mathbb{F}_2^n \to \mathbb{F}_2 \mid f \text{ is linear}\}$ (with rejection rate 1).

Remark 7.4. To be pedantic, the BLR linearity test is actually a family of local testers, one for each value of n. This is a common scenario: We will usually be interested in testing natural *families* of properties $(\mathscr{C}_n)_{n \in \mathbb{N}^+}$, where $\mathscr{C}_n$ contains functions $\{0,1\}^n \to \{0,1\}$. In this case we need to describe a family of testers, one for each n. Generally, these testers will "act the same" for all values of n and will have the property that the rejection rate $\lambda > 0$ is a universal constant independent of n.

There are a number of standard variations of Definition 7.2 that one could consider. One variation is to allow for an *adaptive* testing algorithm, meaning that the algorithm can decide how to generate $\boldsymbol{x}^{(t)}$ based on the query outcomes $f(\boldsymbol{x}^{(1)}), \dots, f(\boldsymbol{x}^{(t-1)})$. However, in this book we will only consider nonadaptive testing. Another variation is to relax the requirement that ϵ-far functions be rejected with probability $\Omega(\epsilon)$; one could allow for smaller rates such as $\Omega(\epsilon^2)$, or $\Omega(\epsilon/\log n)$. For simplicity, we will stick with the strict demand that the rejection probability be linear in ϵ. Finally, the most common definition of property testing allows the number of queries to be a function $r(\epsilon)$ of ϵ but requires that any function ϵ-far from $\mathscr{C}$ be rejected with probability at least 1/2. This is easier to achieve than satisfying Definition 7.2; see Exercise 7.1.

So far we have seen that the property of being linear over $\mathbb{F}_2$ is locally testable. We'll now spend some time discussing local testability of an even simpler property, the property of being a *dictator*. In other words, we'll consider the property

$$\mathscr{D} = \{f : \{0,1\}^n \to \{0,1\} \mid f(x) = x_i \text{ for some } i \in [n]\}.$$

As we will see, dictatorship is in some ways the most important property to be able to test.

We begin with a reminder: Even though $\mathscr{D}$ is a subclass of the linear functions and we have a local tester for linearity, this doesn't mean we automatically have a local tester for dictatorship. (This is in contrast to learning theory, where a learning algorithm for a concept class automatically works for any subclass.) The reason is that the non-dictator linear functions – i.e., χ_S for $|S| \neq 1$ – are at distance $\frac{1}{2}$ from $\mathscr{D}$ but are accepted by any linearity test with probability 1.

Still, we could use a linearity test as a first component of a test for dictatorship; this essentially reduces the problem to testing if an unknown *linear* function is a dictator. Historically, the first local testers for dictatorship [**BGS95, PRS01**] worked this way; after testing linearity, they chose $\boldsymbol{x}, \boldsymbol{y} \sim \{0,1\}^n$ uniformly and independently, set $\boldsymbol{z} = \boldsymbol{x} \wedge \boldsymbol{y}$ (the bitwise logical AND), and tested whether $f(\boldsymbol{z}) = f(\boldsymbol{x}) \wedge f(\boldsymbol{y})$. The idea is that the only parity functions that satisfy this "AND test" with probability 1 are the dictators (and the constant 0). The analysis of the test takes a bit of work; see Exercise 7.8 for details.

Here we will describe a simpler dictatorship test. Recall we have already seen an important result that characterizes dictatorship: Arrow's Theorem, from Chapter 2.5. Furthermore the robust version of Arrow's Theorem (Corollary 2.60) involves evaluating a 3-candidate Condorcet election under the impartial culture assumption, and this is the same as querying the election rule f on 3 correlated random inputs. This suggests a dictatorship testing component we call the "NAE Test":

NAE Test. *Given query access to $f : \{-1,1\}^n \to \{-1,1\}$:*

- *Choose $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{-1,1\}^n$ by letting each triple $(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ be drawn independently and uniformly at random from among the 6 triples satisfying the not-all-equal predicate $\mathrm{NAE}_3 : \{-1,1\}^3 \to \{0,1\}$.*
- *Query f at $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$.*

- *Accept if* $\mathrm{NAE}_3(f(\boldsymbol{x}), f(\boldsymbol{y}), f(\boldsymbol{z}))$ *is satisfied.*

The NAE Test by itself is *almost* a 3-query local tester for the property of being a dictator. Certainly if f is a dictator then the NAE Test accepts with probability 1. Furthermore, in Chapter 2.5 we proved:

Theorem 7.5 (Restatement of Corollary 2.60)**.** *If the NAE Test accepts f with probability $1-\epsilon$, then $\mathbf{W}^1[f] \geq 1-\frac{9}{2}\epsilon$, and hence f is $O(\epsilon)$-close to $\pm\chi_i$ for some $i \in [n]$ by the FKN Theorem.*

There are two slightly unsatisfactory aspects to this theorem. First, it gives a local tester only for the property of being a dictator or a negated-dictator. Second, though the deduction $\mathbf{W}^1[f] \geq 1-\frac{9}{2}\epsilon$ requires only simple Fourier analysis, the conclusion that f is close to a (negated-)dictator relies on the non-trivial FKN Theorem. Fortunately we can fix both issues simply by adding in the BLR Test:

Theorem 7.6. *Given query access to $f : \{-1,1\}^n \to \{-1,1\}$, perform both the BLR Test and the NAE Test. This is a 6-query local tester for the property of being a dictator (with rejection rate .1).*

Proof. The first condition in Definition 7.2 is easy to check: If $f : \{-1,1\}^n \to \{-1,1\}$ is a dictator, then both tests accept f with probability 1. To check the second condition, fix $0 \leq \epsilon \leq 1$ and assume the overall test accepts f with probability at least $1-.1\epsilon$. Our goal is to show that f is ϵ-close to some dictator.

Since the overall test accepts with probability at least $1-.1\epsilon$, both the BLR and the NAE tests must individually accept f with probability at least $1-.1\epsilon$. By the analysis of the NAE Test we deduce that $\mathbf{W}^1[f] \geq 1-\frac{9}{2}\cdot .1\epsilon = 1-.45\epsilon$. By the analysis of the BLR Test (Theorem 1.30) we deduce that f is $.1\epsilon$-close to some parity function; i.e., $\widehat{f}(S^*) \geq 1-.2\epsilon$ for some $S^* \subseteq [n]$. Now if $|S^*| \neq 1$ we would have

$$1 = \sum_{k=0}^{n} \mathbf{W}^k[f] \geq (1-.45\epsilon) + (1-.2\epsilon)^2 \geq 2-.85\epsilon > 1,$$

a contradiction. Thus we must have $|S^*| = 1$ and hence f is $.1\epsilon$-close to the dictator χ_{S^*}, stronger than what we need. □

As you can see, we haven't been particularly careful about obtaining the largest possible rejection rate. Instead, we will be more interested in using as few queries as possible (while maintaining some positive constant rejection rate). Indeed we now show a small trick which lets us reduce our 6-query local tester for dictatorship down to a 3-query one. This is best possible since dictatorship can't be locally tested with 2 queries (see Exercise 7.6).

BLR+NAE Test. *Given query access to $f : \{-1,1\}^n \to \{-1,1\}$:*

- *With probability* 1/2*, perform the BLR Test on* f.
- *With probability* 1/2*, perform the NAE Test on* f.

Theorem 7.7. *The BLR+NAE Test is a* 3*-query local tester for the property of being a dictator (with rejection rate* .05*).*

Proof. The only observation we need to make is that if the BLR+NAE Test accepts with probability $1-.05\epsilon$ then both the BLR and the NAE tests individually must accept f with probability at least $1-.1\epsilon$. The result then follows from the analysis of Theorem 7.6. □

Remark 7.8. In general, this trick lets us take the *maximum* of the query complexities when we combine tests, rather than the sum (at the expense of worsening the rejection rate). Suppose we wish to combine $t = O(1)$ different testing algorithms, where the ith tester uses r_i queries. We make an overall test that performs each subtest with probability $1/t$. This gives a $\max(r_1,\dots,r_t)$-query testing algorithm with the following guarantee: If the overall test accepts f with probability $1-\frac{\lambda}{t}\epsilon$ then *every* subtest must accept f with probability at least $1-\lambda\epsilon$.

We can now explain one reason why dictatorship is a particularly important property to be able to test locally. Given the BLR Test for linear functions it still took us a little thought to find a local test for the subclass $\mathcal{D}$ of dictators. But given our dictatorship test, it's easy to give a 3-query local tester for *any* subclass of $\mathcal{D}$. (On a related note, Exercise 7.15 asks you to give a 3-query local tester for any affine subspace of the linear functions.)

Theorem 7.9. *Let $\mathscr{S}$ be any subclass of n-bit dictators; i.e., let $S \subseteq [n]$ and let*

$$\mathscr{S} = \{\chi_i : \{0,1\}^n \to \{0,1\} \mid i \in S\}.$$

Then there is a 3*-query local tester for $\mathscr{S}$ (with rejection rate* .01*).*

Proof. Let $1_S \in \{0,1\}^n$ denote the indicator string for the subset S. Given access to $f : \{0,1\}^n \to \{0,1\}$, the test is as follows:

- With probability 1/2, perform the BLR+NAE Test on f.
- With probability 1/2, apply the local correcting routine of Proposition 1.31 to f on string 1_S; accept if and only if the output value is 1.

This test always makes either 2 or 3 queries, and whenever $f \in \mathscr{S}$ it accepts with probability 1. Now let $0 \le \epsilon \le 1$ and suppose the test accepts f with probability at least $1-\lambda\epsilon$, where $\lambda = .01$. Our goal will be to show that f is ϵ-close to a dictator χ_i with $i \in S$.

Since the overall test accepts f with probability at least $1-\lambda\epsilon$, the BLR+NAE Test must accept f with probability at least $1-2\lambda\epsilon$. By Theorem 7.7 we may deduce that f is $40\lambda\epsilon$-close to some dictator χ_i. Our goal is to show that $i \in S$; this will complete the proof because $40\lambda\epsilon \leq \epsilon$ (by our choice of $\lambda = .01$).

So suppose by way of contradiction that $i \not\in S$; i.e., $\chi_i(1_S) = 0$. Since f is $40\lambda\epsilon$-close to the parity function χ_i, Proposition 1.31 tells us that

$$\mathbf{Pr}[\text{locally correcting } f \text{ on input } 1_S \text{ produces the output } \chi_i(1_S) = 0] \geq 1 - 80\lambda\epsilon.$$

On the other hand, since the overall test accepts f with probability at least $1-\lambda\epsilon$, the second subtest must accept f with probability at least $1-2\lambda\epsilon$. This means

$$\mathbf{Pr}[\text{locally correcting } f \text{ on input } 1_S \text{ produces the output } 0] \leq 2\lambda\epsilon.$$

But this is a contradiction, since $2\lambda\epsilon < 1 - 80\lambda\epsilon$ for all $0 \leq \epsilon \leq 1$ (by our choice of $\lambda = .01$). Hence $i \in S$ as desired. □

7.2. Probabilistically Checkable Proofs of Proximity

In the previous section we saw that every subproperty of the dictatorship property has a 3-query local tester. In this section we will show that *any property whatsoever* has a 3-query local tester – if an appropriate "proof" is provided.

To make sense of this statement let's first generalize the setting in which we study property testing. Definitions 7.1 and 7.2 are concerned with testing a Boolean function $f : \{0,1\}^n \to \{0,1\}$ by querying its values on various inputs. If we think of f's truth table as a Boolean string of length $N = 2^n$, then a testing algorithm simply queries various coordinates of this string. It makes sense to generalize to the notion of testing properties of N-bit *strings*, for any length N. Here a property $\mathscr{C}$ will just be a collection $\mathscr{C} \subseteq \{0,1\}^N$ of strings, and we'll be concerned with the relative Hamming distance $\mathrm{dist}(w,w') = \frac{1}{N}\Delta(w,w')$ between strings. For simplicity, we'll begin to write n instead of N.

Definition 7.10. An *r-query string testing algorithm* for strings $w \in \{0,1\}^n$ is a randomized algorithm that:

- chooses r (or fewer) indices $\boldsymbol{i}_1, \dots, \boldsymbol{i}_r \in [n]$ according to some probability distribution;
- queries $w_{\boldsymbol{i}_1}, \dots, w_{\boldsymbol{i}_r}$;
- based on the outcomes, decides (deterministically) whether to "accept" w.

We may also generalize this definition to testing strings $w \in \Omega^n$ over finite alphabets Ω of cardinality larger than 2.

Definition 7.11. Let $\mathscr{C} \subseteq \{0,1\}^n$ be a "property" of n-bit Boolean strings. We say a string testing algorithm is a *local tester for $\mathscr{C}$* (with *rejection rate* $\lambda > 0$) if it satisfies the following:

- If $w \in \mathscr{C}$, then the tester accepts with probability 1.
- For all $0 \le \epsilon \le 1$, if $\mathrm{dist}(w, \mathscr{C}) > \epsilon$, then the tester rejects w with probability greater than $\lambda \cdot \epsilon$.

  Equivalently, if the tester accepts w with probability at least $1 - \lambda \cdot \epsilon$, then w is ϵ-close to $\mathscr{C}$; i.e., $\exists w' \in \mathscr{C}$ such that $\mathrm{dist}(w, w') \le \epsilon$.

Example 7.12. Let $\mathscr{Z} = \{(0, 0, \dots, 0)\} \subseteq \{0,1\}^n$ be the property of being the all-zeroes string. Then the following is a 1-query local tester for $\mathscr{Z}$ (with rejection rate 1): Pick a uniformly random index $\boldsymbol{i}$ and accept if $w_{\boldsymbol{i}} = 0$.

Let $\mathscr{E} = \{(0, 0, \dots, 0), (1, 1, \dots, 1)\} \subseteq \{0,1\}^n$ be the property of having all coordinates equal. Then the following is a 2-query local tester for $\mathscr{E}$: Pick two independent and uniformly random indices $\boldsymbol{i}$ and $\boldsymbol{j}$ and accept if $w_{\boldsymbol{i}} = w_{\boldsymbol{j}}$. In Exercise 7.4 you are asked to show that if $\mathrm{dist}(w, \mathscr{E}) = \epsilon$, then this tester rejects w with probability $\frac{1}{2} - \frac{1}{2}(1-2\epsilon)^2 \ge \epsilon$.

Let $\mathscr{O} = \{w \in \mathbb{F}_2^n : w \text{ has an odd number of 1's}\}$. This property does *not* have a local tester making few queries. In fact, in Exercise 7.5 you are asked to show that any local tester for $\mathscr{O}$ must make the maximum number of queries, n.

As the last example shows, not every property has a local tester making a small number of queries; indeed, most properties of n-bit strings do not. This is rather too bad: Imagine that for any large n and any complicated property $\mathscr{C} \subseteq \{0,1\}^n$ there were an $O(1)$-query local tester. Then if anyone supplied you with a string w claiming it satisfied $\mathscr{C}$, you wouldn't have to laboriously check this yourself, nor would you have to trust the supplier; you could simply spot-check w in a constant number of coordinates and become convinced that w is (close to being) in $\mathscr{C}$.

But what if, in addition to $w \in \{0,1\}^n$, you could require the supplier to give you some additional side information $\Pi \in \{0,1\}^\ell$ about w so as to assist you in testing that $w \in \mathscr{C}$? One can think of Π as a kind of "proof" that w satisfies $\mathscr{C}$. In this case it's possible that you can spot-check w and Π together in a constant number of coordinates and become convinced that w is (close to being) in $\mathscr{C}$ – all without having to "trust" the supplier of the string w and the purported proof Π. These ideas lead to the notion of *probabilistically checkable proofs of proximity* (PCPPs).

Definition 7.13. Let $\mathscr{C} \subseteq \{0,1\}^n$ be a property of n-bit Boolean strings and let $\ell \in \mathbb{N}$. We say that $\mathscr{C}$ has an *r-query, length-ℓ probabilistically checkable proof of proximity (PCPP) system* (with rejection rate $\lambda > 0$) when the following holds: There exists an r-query testing algorithm T for $(n+\ell)$-bit strings, thought of as pairs $w \in \{0,1\}^n$ and $\Pi \in \{0,1\}^\ell$, such that:

- ("Completeness.") If $w \in \mathscr{C}$, then there exists a "proof" $\Pi \in \{0,1\}^\ell$ such that T accepts with probability 1.
- ("Soundness.") For all $0 \le \epsilon \le 1$, if $\mathrm{dist}(w,\mathscr{C}) > \epsilon$, then for *every* "proof" $\Pi \in \{0,1\}^\ell$ the tester T rejects with probability greater than $\lambda \cdot \epsilon$.

  Equivalently, if there exists $\Pi \in \{0,1\}^\ell$ that causes T to accept with probability at least $1 - \lambda \cdot \epsilon$, then w must be ϵ-close to $\mathscr{C}$.

PCPP systems are also known as assisted testers, locally testable proofs, or assignment testers.

Remark 7.14. A word on the three parameters: We are usually interested in fixing the number of queries r to a very small universal constant (such as 3) while trying to keep the proof length $\ell = \ell(n)$ relatively small (e.g., $\mathrm{poly}(n)$ is a good goal). We are usually not very concerned with the rejection rate λ so long as it's a positive universal constant (independent of n).

Example 7.15. In Example 7.12 we stated that $\mathscr{O} = \{w \in \mathbb{F}_2^n : w_1 + \cdots + w_n = 1\}$ has no local tester making fewer than n queries. But it's easy to give a 3-query PCPP system for $\mathscr{O}$ with proof length $n-1$ (and rejection rate 1). The idea is to require the proof string Π to contain the partial sums of w:

$$\Pi_j = \sum_{i=1}^{j+1} w_i \pmod 2.$$

The tester will perform one of the following checks, uniformly at random:

$$\begin{aligned} \Pi_1 &= w_1 + w_2 \\ \Pi_2 &= \Pi_1 + w_3 \\ \Pi_3 &= \Pi_2 + w_4 \\ &\cdots \\ \Pi_{n-1} &= \Pi_{n-2} + w_n \\ \Pi_{n-1} &= 1 \end{aligned}$$

Evidently the tester always makes at most 3 queries. Further, in the "completeness" case $w \in \mathscr{O}$, if Π is a correct list of partial sums then the tester will accept with probability 1. It remains to analyze the "soundness" case, $w \notin \mathscr{O}$. Here we are significantly aided by the fact that $\mathrm{dist}(w,\mathscr{O})$ must be exactly $1/n$ (since every string is at Hamming distance either 0 or 1 from $\mathscr{O}$). Thus to confirm the claimed rejection rate of 1, we only need to observe that if $w \notin \mathscr{O}$ then at least one of the tester's n checks must fail.

This example generalizes to give a very efficient PCPP system for testing that w satisfies any fixed $\mathbb{F}_2$-linear equation. What about testing that w satisfies a fixed system of $\mathbb{F}_2$-linear equations? This interesting question is explored in Exercise 7.16, which serves as a good warmup for our next result.

We now extend Theorem 7.9 to show the rather remarkable fact that *any* property of n-bit strings has a 3-query PCPP system. (The proof length, however, is enormous.)

Theorem 7.16. *Let $\mathscr{C} \subseteq \{0,1\}^n$ be any class of strings. Then there is a 3-query, length-2^{2^n} PCPP system for $\mathscr{C}$ (with rejection rate $.001$).*

Proof. Let $N = 2^n$ and fix an arbitrary bijection $\iota : \{0,1\}^n \to [N]$. The tester will interpret the string $w \in \{0,1\}^n$ to be tested as an index $\iota(w) \in [N]$ and will interpret the 2^N-length proof Π as a function $\Pi : \{0,1\}^N \to \{0,1\}$. The idea is for the tester to require that Π be the dictator function corresponding to index $\iota(w)$; i.e., $\chi_{\iota(w)} : \{0,1\}^N \to \{0,1\}$.

Now under the identification ι, we can think of the string property $\mathscr{C}$ as a subclass of all N-bit dictators, namely

$$\mathscr{C}' = \{\chi_{\iota(w')} : \{0,1\}^N \to \{0,1\} \mid w' \in \mathscr{C}\}.$$

In particular, $\mathscr{C}'$ is a property of N-bit functions. We can now state the twofold goal of the tester:

(1) check that $\Pi \in \mathscr{C}'$;

(2) given that Π is indeed some dictator $\chi_{\iota(w')} : \{0,1\}^N \to \{0,1\}$ with $w' \in \mathscr{C}$, check that $w' = w$.

To accomplish the latter the tester would like to check $w_{\boldsymbol{j}} = w'_{\boldsymbol{j}}$ for a random $\boldsymbol{j} \in [n]$. The tester can query any w_j directly but accessing w'_j requires a little thought. The trick is to prepare the string

$$X^{(j)} \in \{0,1\}^N \text{ defined by } X^{(j)}_{\iota(y)} = y_j.$$

and then to locally correct Π on $X^{(j)}$ (using Proposition 1.31).

Thus the tester is defined as follows:

(1) With probability 1/2, locally test the function property $\mathscr{C}'$ using Theorem 7.9.

(2) With probability 1/2, pick $\boldsymbol{j} \sim [n]$ uniformly at random; locally correct Π on the string $X^{(\boldsymbol{j})}$ and accept if the outcome equals $w_{\boldsymbol{j}}$.

Note that the tester makes 3 queries in both of the subtests.

Verifying "completeness" of this PCPP system is easy: if $w \in \mathscr{C}$ and Π is indeed the (truth table of) $\chi_{\iota(w)} : \{0,1\}^N \to \{0,1\}$ then the test will accept with probability 1. It remains to verify the "soundness" condition. Fix $w \in \{0,1\}^n$, $\Pi : \{0,1\}^N \to \{0,1\}$, and $0 \le \epsilon \le 1$ and suppose that the tester accepts (w, Π) with probability at least $1 - \lambda\epsilon$, where $\lambda = .001$. Our goal is to show that w is ϵ-close to some string $w' \in \mathscr{C}$.

Since the overall test accepts with probability at least $1-\lambda\epsilon$, subtest (1) above accepts with probability at least $1-2\lambda\epsilon$. Thus by Theorem 7.9, Π must be $200\lambda\epsilon$-close to some dictator $\chi_{\iota(w')}$ with $w' \in \mathscr{C}$. Since dictators are parity functions, Proposition 1.31 tells us that

$$\forall j,\ \mathbf{Pr}[\text{locally correcting } \Pi \text{ on } X^{(j)} \text{ produces } \chi_{\iota(w')}(X^{(j)}) = w'_j] \geq 1-400\lambda\epsilon \geq 1/2, \tag{7.1}$$

where we used $400\lambda\epsilon < 400\lambda \leq 1/2$ by the choice $\lambda = .001$.

On the other hand, since the overall test accepts with probability at least $1-\lambda\epsilon$, subtest (2) above rejects with probability at most $2\lambda\epsilon$. This means

$$\mathop{\mathbf{E}}_{\boldsymbol{j}\sim[n]}\Big[\mathbf{Pr}[\text{locally correcting } \Pi \text{ on } X^{(\boldsymbol{j})} \textit{ doesn't} \text{ produce } w_{\boldsymbol{j}}]\Big] \leq 2\lambda\epsilon.$$

By Markov's inequality we deduce that except for at most a $4\lambda\epsilon$ fraction of coordinates $j \in [n]$ we have

$$\mathbf{Pr}[\text{locally correcting } \Pi \text{ on } X^{(j)} \textit{ doesn't} \text{ produce } w_j] < 1/2.$$

Combining this information with (7.1) we deduce that $w_j = w'_j$ except for at most a $4\lambda\epsilon \leq \epsilon$ fraction of coordinates $j \in [n]$. Since $w' \in \mathscr{C}$ we conclude that $\mathrm{dist}(w, C) \leq \epsilon$, as desired. □

You may feel that the doubly-exponential proof length 2^{2^n} in this theorem is quite bad, but bear in mind there are 2^{2^n} different properties $\mathscr{C}$. Actually, giving a PCPP system for *every* property is a bit overzealous since most properties are not interesting or natural. A more reasonable goal would be to give efficient PCPP systems for all "explicit" properties. A good way to formalize this is to consider properties decidable by polynomial-size circuits. Here we use the definition of general (De Morgan) circuits from Exercise 4.13. Given an n-variable circuit C we consider the set of strings which it "accepts" to be a property,

$$\mathscr{C} = \{w \in \{0,1\}^n : C(w) = 1\}. \tag{7.2}$$

For properties computed by modest-sized circuits C we may hope for PCPP systems with proof length much less than 2^{2^n}. We saw such a case in Example 7.15.

Another advantage of considering "explicit" properties is that we can define a notion of *constructing* a PCPP system, "given" a property. A theorem of the form "for each explicit property $\mathscr{C}$ there exists an efficient PCPP system..." may not be useful, practically speaking, if its proof is nonconstructive. We can formalize the issue as follows:

Definition 7.17. A *PCPP reduction* is an algorithm which takes as input a circuit C and outputs the *description* of a PCPP system for the string property $\mathscr{C}$ decided by C as in (7.2), where n is the number of inputs to C. If the output PCPP system always makes r queries, has proof length $\ell(n, \mathrm{size}(C))$

(for some function ℓ), and has rejection rate $\lambda > 0$, we say that the PCPP reduction has the same parameters. Finally, the PCPP reduction should run in time $\mathrm{poly}(\mathrm{size}(C), \ell)$.

(We haven't precisely specified what it means to output the description of a PCPP system; this will be explained more carefully in Section 7.3. In brief it means to list – for each possible outcome of the tester's randomness – which bits are queried and what predicate of them is used to decide acceptance.)

Looking back at the results on testing subclasses of dictatorship (Theorem 7.9) and PCPPs for any property (Theorem 7.16) we can see they have the desired sort of "constructive" proofs. In Theorem 7.9 the local tester's description depends in a very simple way on the input 1_S. As for Theorem 7.16, it suffices to note that given an n-input circuit C we can write down its truth table (and hence the property it decides) in time $\mathrm{poly}(\mathrm{size}(C)) \cdot 2^n$, whereas the allowed running time is at least $\mathrm{poly}(\mathrm{size}(C), 2^{2^n})$. Hence we may state:

Theorem 7.18. *There exists a 3-query PCPP reduction with proof length 2^{2^n} (and rejection rate .001).*

In Exercise 7.18 you are asked to improve this result as follows:

Theorem 7.19. *There exists a 3-query PCPP reduction with proof length $2^{\mathrm{poly}(\mathrm{size}(C))}$ (and positive rejection rate).*

(The fact that we again have just 3 queries is explained by Exercise 7.12; there is a generic reduction from any constant number of queries down to 3.)

Indeed, there is a much more dramatic improvement:

The PCPP Theorem. *There exists a 3-query PCPP reduction with proof length $\mathrm{poly}(\mathrm{size}(C))$ (and positive rejection rate).*

This is (a slightly strengthened version of) the famous "PCP Theorem" [**FGL⁺96**, **AS98**, **ALM⁺98**] from the field of computational complexity, which is discussed later in this chapter. Though the PCPP Theorem is far stronger than Theorem 7.18, the latter is not unnecessary; it's actually an ingredient in Dinur's proof of the PCP Theorem [**Din07**], being applied only to circuits of "constant" size. The current state of the art for PCPP length [**Din07**, **BS08**] is highly efficient:

Theorem 7.20. *There exists a 3-query PCPP reduction with proof length $\mathrm{size}(C) \cdot \mathrm{polylog}(\mathrm{size}(C))$ (and positive rejection rate).*

7.3. CSPs and computational complexity

This section is about the computational complexity of constraint satisfaction problems (CSPs), a fertile area of application for analysis of Boolean functions.

To study it we need to introduce a fair bit of background material; in fact, this section will mainly consist of definitions.

In brief, a CSP is an algorithmic task in which a large number of "variables" must be assigned "labels" so as to satisfy given "local constraints". We start by informally describing some examples:

Example 7.21.

- In the "Max-3-Sat" problem, given is a CNF formula of width at most 3 over Boolean variables $x_1,\dots,x_n$. The task is to find a setting of the inputs that satisfies (i.e., makes True) as many clauses as possible.
- In the "Max-Cut" problem, given is an undirected graph $G=(V,E)$. The task is to find a "cut" – i.e., a partition of V into two parts – so that as many edges as possible "cross the cut".
- In the "Max-E3-Lin" problem, given is a system of linear equations over $\mathbb{F}_2$, each equation involving exactly 3 variables. The system may in general be overdetermined; the task is to find a solution which satisfies as many equations as possible.
- In the "Max-3-Coloring" problem, given is an undirected graph $G=(V,E)$. The task is to color each vertex either red, green, or blue so as to make as many edges as possible bichromatic.

Let's rephrase the last two of these examples so that the descriptions have more in common. In Max-E3-Lin we have a set of variables V, to be assigned labels from the domain $\Omega=\mathbb{F}_2$. Each constraint is of the form $v_1+v_2+v_3=0$ or $v_1+v_2+v_3=1$, where $v_1,v_2,v_3\in V$. In Max-3-Coloring we have a set of variables (vertices) V to be assigned labels from the domain $\Omega=\{\text{red},\text{green},\text{blue}\}$. Each constraint (edge) is a pair of variables, constrained to be labeled by unequal colors.

We now make formal definitions which encompass all of the above examples:

Definition 7.22. A constraint satisfaction problem (CSP) over *domain* Ω is defined by a finite set of *predicates* ("types of constraints") Ψ, with each $\psi\in\Psi$ being of the form $\psi:\Omega^r\to\{0,1\}$ for some *arity* r (possibly different for different predicates). We say that the *arity* of the CSP is the maximum arity of its predicates.

Such a CSP is associated with an algorithmic task called "Max-CSP(Ψ)", which we will define below. First, though, let us see how the CSPs from Example 7.21 fit into the above definition.

- Max-3-Sat: Domain $\Omega=\{\mathsf{True},\mathsf{False}\}$; Ψ contains 14 predicates: the 8 logical OR functions on 3 literals (variables/negated-variables), the 4 logical OR functions on 2 literals, and the 2 logical OR functions on 1 literal.

- Max-Cut: Domain $\Omega = \{-1,1\}$; $\Psi = \{\neq\}$, the "not-equal" predicate $\neq$: $\{-1,1\}^2 \to \{0,1\}$.
- Max-E3-Lin: Domain $\Omega = \mathbb{F}_2$; Ψ contains two 3-ary predicates, $(x_1,x_2,x_3) \mapsto x_1 + x_2 + x_3$ and $(x_1,x_2,x_3) \mapsto x_1 + x_2 + x_3 + 1$.
- Max-3-Coloring: Domain $\Omega = \{\text{red},\text{green},\text{blue}\}$; Ψ contains just the single not-equal predicate $\neq$: $\Omega^2 \to \{0,1\}$.

Remark 7.23. Let us add a few words about traditional CSP terminology. *Boolean* CSPs refer to the case $|\Omega| = 2$. If $\psi : \{-1,1\}^r \to \{0,1\}$ is a Boolean predicate we sometimes write "Max-ψ" to refer to the CSP where all constraints are of the form ψ applied to *literals*; i.e., $\Psi = \{\psi(\pm v_1,\dots,\pm v_r)\}$. As an example, Max-E3-Lin could also be called Max-$\chi_{[3]}$. The "E3" in the name Max-E3-Lin refers to the fact that all constraints involve "E"xactly 3 variables. Thus e.g. Max-3-Lin is the generalization in which 1- and 2-variable equations are allowed. Conversely, Max-E3-Sat is the special case of Max-3-Sat where each clause must be of width exactly 3 (a CSP which could also be called Max-OR_3).

To formally define the algorithmic task Max-CSP(Ψ), we begin by defining its input:

Definition 7.24. An *instance* (or *input*) $\mathscr{P}$ of Max-CSP(Ψ) over variable set V is a list (multiset) of *constraints*. Each constraint $C \in \mathscr{P}$ is a pair $C = (S,\psi)$, where $\psi \in \Psi$ and where the *scope* $S = (v^1,\dots,v^r)$ is a tuple of *distinct* variables from V, with r being the arity of ψ. We always assume that each $v \in V$ participates in at least one constraint scope. The *size* of an instance is the number of bits required to represent it; writing $n = |V|$ and treating $|\Omega|$, $|\Psi|$ and the arity of Ψ as constants, the size is between n and $O(|\mathscr{P}|\log n)$.

Remark 7.25. Let's look at how the small details of Definition 7.24 affect input graphs for Max-Cut. Since an instance is a multiset of constraints, this means we allow graphs with parallel edges. Since each scope must consist of distinct variables, this means we disallow graphs with self-loops. Finally, since each variable must participate in at least one constraint, this means input graphs must have no isolated vertices (though they may be disconnected).

Given an assignment of labels for the variables, we are interested in the number of constraints that are "satisfied". The reason we explicitly allow duplicate constraints in an instance is that we may want some constraints to be more important than others. In fact it's more convenient to normalize by looking at the *fraction* of satisfied constraints, rather than the number. Equivalently, we can choose a constraint $\boldsymbol{C} \sim \mathscr{P}$ uniformly at random and look at the *probability* that it is satisfied. It will actually be quite useful to think of a CSP instance $\mathscr{P}$ as a probability distribution on constraints. (Indeed, we could have more generally defined *weighted CSPs* in which the constraints are given arbitrary nonnegative weights summing to 1; however, we don't want to worry about the issue of representing, say, irrational weights with finitely many bits.)

Definition 7.26. An *assignment* (or *labeling*) for instance $\mathscr{P}$ of Max-CSP(Ψ) is just a mapping $F : V \to \Omega$. For constraint $C = (S, \psi) \in \mathscr{P}$ we say that F *satisfies* C if $\psi(F(S)) = 1$. Here we use shorthand notation: if $S = (v^1, \dots, v^r)$ then $F(S)$ denotes $(F(v^1), \dots, F(v^r))$. The *value* of F, denoted $\mathrm{Val}_{\mathscr{P}}(F)$, is the fraction of constraints in $\mathscr{P}$ that F satisfies:

$$\mathrm{Val}_{\mathscr{P}}(F) = \mathop{\mathbf{E}}_{(\boldsymbol{S},\boldsymbol{\psi})\sim\mathscr{P}}[\boldsymbol{\psi}(F(\boldsymbol{S}))] \in [0,1]. \tag{7.3}$$

The *optimum* value of $\mathscr{P}$ is

$$\mathrm{Opt}(\mathscr{P}) = \max_{F:V\to\Omega} \{\mathrm{Val}_{\mathscr{P}}(F)\}.$$

If $\mathrm{Opt}(\mathscr{P}) = 1$, we say that $\mathscr{P}$ is *satisfiable*.

Remark 7.27. In the literature on CSPs there is sometimes an unfortunate blurring between a variable and its assignment. For example, a Max-E3-Lin instance may be written as

$$\begin{aligned} x_1 + x_2 + x_3 &= 0 \\ x_1 + x_5 + x_6 &= 0 \\ x_3 + x_4 + x_6 &= 1; \end{aligned}$$

then a particular assignment $x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1, x_5 = 1, x_6 = 1$ may be given. Now there is confusion: Does x_2 represent the name of a variable or does it represent 1? Because of this we prefer to display CSP instances with the name of the assignment F present in the constraints. That is, the above instance would be described as finding $F : \{x_1, \dots, x_6\} \to \mathbb{F}_2$ so as to satisfy as many as possible of the following:

$$\begin{aligned} F(x_1) + F(x_2) + F(x_3) &= 0 \\ F(x_1) + F(x_5) + F(x_6) &= 0 \\ F(x_3) + F(x_4) + F(x_6) &= 1, \end{aligned}$$

Finally, we define the algorithmic task associated with a CSP:

Definition 7.28. The algorithmic task Max-CSP(Ψ) is defined as follows: The input is an instance $\mathscr{P}$. The goal is to output an assignment F with as large a value as possible.

Having defined CSPs, let us make a connection to the notion of a string testing algorithm from the previous section. The connection is this: *CSPs* *and string testing algorithms are the same object*. Indeed, consider a CSP instance $\mathscr{P}$ over domain Ω with n variables V. Fix an assignment $F : V \to \Omega$; we can also think of F as a string in Ω^n (under some ordering of V). Now think of a testing algorithm which chooses a constraint $(\boldsymbol{S}, \boldsymbol{\psi}) \sim \mathscr{P}$ at random, “queries” the string entry $F(v)$ for each $v \in \boldsymbol{S}$, and accepts if and only if the predicate $\boldsymbol{\psi}(F(\boldsymbol{S}))$ is satisfied. This is indeed an r-query string testing algorithm, where r is the arity of the CSP; the probability the tester accepts is precisely $\mathrm{Val}_{\mathscr{P}}(F)$.

Conversely, let T be some randomized testing algorithm for strings in Ω^n. Assume for simplicity that T’s randomness comes from the uniform distribution over some sample space U. Now suppose we enumerate all outcomes in U, and for each we write the tuple of indices S that T queries and the predicate $\psi : \Omega^{|S|} \to \{0,1\}$ that T uses to make its subsequent accept/reject decision. Then this list of scope/predicates pairs is precisely an instance of an n-variable CSP over Ω. The arity of the CSP is equal to the (maximum) number of queries that T makes and the predicates for the CSP are precisely those used by the tester in making its accept/reject decisions. Again, the probability that T accepts a string $F \in \Omega^n$ is equal to the value of F as an assignment for the CSP. (Our actual definition of string testers allowed any form of randomness, including, say, irrational probabilities; thus technically not every string tester can be viewed as a CSP. However, it does little harm to ignore this technicality.)

In particular, this equivalence between string testers and CSPs lets us properly define “outputting the description of a PCPP system” as in Definition 7.17 of PCPP reductions.

Example 7.29. The PCPP system for $\mathscr{O} = \{w \in \mathbb{F}_2 : w_1 + \cdots + w_n = 1\}$ given in Example 7.15 can be thought of as an instance of the Max-3-Lin CSP over the $2n-1$ variables $\{w_1, \dots, w_n, \Pi_1, \dots, \Pi_{n-1}\}$. The BLR linearity test for functions $\mathbb{F}_2^n \to \mathbb{F}_2$ can also be thought of as instance of Max-3-Lin over 2^n variables (recall that function testers are string testers). In this case we identify the variable set with $\mathbb{F}_2^n$; if $n = 2$ then the variables are named $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$; and, if we write $F : \mathbb{F}_2^2 \to \mathbb{F}_2$ for the assignment, the instance is

$$\begin{array}{llll}
F(0,0)+F(0,0)+F(0,0)=0 & F(0,1)+F(0,0)+F(0,1)=0 & F(1,0)+F(0,0)+F(1,0)=0 & F(1,1)+F(0,0)+F(1,1)=0 \\
F(0,0)+F(0,1)+F(0,1)=0 & F(0,1)+F(0,1)+F(0,0)=0 & F(1,0)+F(0,1)+F(1,1)=0 & F(1,1)+F(0,1)+F(1,0)=0 \\
F(0,0)+F(1,0)+F(1,0)=0 & F(0,1)+F(1,0)+F(1,1)=0 & F(1,0)+F(1,0)+F(0,0)=0 & F(1,1)+F(1,0)+F(0,1)=0 \\
F(0,0)+F(1,1)+F(1,1)=0 & F(0,1)+F(1,1)+F(1,0)=0 & F(1,0)+F(1,1)+F(0,1)=0 & F(1,1)+F(1,1)+F(0,0)=0.
\end{array}$$

Cf. Remark 7.27; also, note the duplicate constraints.

We end this section by discussing the computational complexity of finding high-value assignments for a given CSP – equivalently, finding strings that make a given string tester accept with high probability. Consider, for example, the task of Max-Cut on n-vertex graphs. Of course, given a Max-Cut instance one can always find the optimal solution in time roughly 2^n, just by trying all possible cuts. Unfortunately, this is not very efficient, even for slightly large values of n. In computational complexity theory, an algorithm is generally deemed "efficient" if it runs in time poly(n). For some subfamilies of graphs there are poly(n)-time algorithms for finding the maximum cut, e.g., bipartite graphs (Exercise 7.14) or planar graphs. However, it seems very unlikely that there is a poly(n)-time algorithm that is guaranteed to find an optimal Max-Cut assignment given any input graph. This statement is formalized by a basic theorem from the field of computational complexity:

Theorem 7.30. *The task of finding the maximum cut in a given input graph is "*NP*-hard".*

We will not formally define NP-hardness in this book (though see Exercise 7.13 for some more explanation). Roughly speaking it means "at least as hard as the Circuit-Sat problem", where "Circuit-Sat" is the following task: Given an n-variable Boolean circuit C, decide whether or not C is satisfiable (i.e., there exists $w \in \{0,1\}^n$ such that $C(w) = 1$). It is widely believed that Circuit-Sat does not have a polynomial-time algorithm (this is the "$\mathsf{P} \neq \mathsf{NP}$" conjecture). In fact it is also believed that Circuit-Sat does not have a $2^{o(n)}$-time algorithm.

For essentially all CSPs, including Max-E3-Sat, Max-E3-Lin, and Max-3-Coloring, finding an optimal solution is NP-hard. This motivates considering a relaxed goal:

Definition 7.31. Let $0 \leq \alpha \leq \beta \leq 1$. We say that algorithm A is an (α,β)-*approximation algorithm* for Max-CSP(Ψ) (pronounced "α out of β approximation") if it has the following guarantee: on any instance with optimum value at least β, algorithm A outputs an assignment of value at least α. In case A is a randomized algorithm, we only require that its output has value at least α in expectation.

A mnemonic here is that when the βest assignment has value β, the αlgorithm gets value α.

Example 7.32. Consider the following algorithm for Max-E3-Lin: Given an instance, output either the assignment $F \equiv 0$ or the assignment $F \equiv 1$, whichever has higher value. Since either 0 or 1 occurs on at least half of the instance's "right-hand sides", the output assignment will always have value at least $\frac{1}{2}$. Thus this is an efficient $(\frac{1}{2},\beta)$-approximation algorithm for any β. In the case $\beta = 1$ one can do better: performing Gaussian elimination is an efficient $(1,1)$-approximation algorithm for Max-E3-Lin (or indeed Max-r-Lin for any r).

As a far more sophisticated example, Goemans and Williamson [GW95] showed that there is an efficient (randomized) algorithm which $(.878\beta, \beta)$-approximates Max-Cut for every β.

Not only is finding the optimal solution of a Max-E3-Sat instance NP-hard, it's even NP-hard on *satisfiable* instances. In other words:

Theorem 7.33. $(1,1)$-*approximating Max-E3Sat is* NP-*hard. The same is true of Max-3-Coloring.*

On the other hand, it's easy to $(1,1)$-approximate Max-3-Lin (Example 7.32) or Max-Cut (Exercise 7.14). Nevertheless, the "textbook" NP-hardness results for these problems imply the following:

Theorem 7.34. (β,β)-*approximating Max-E3-Lin is* NP-*hard for any fixed* $\beta \in (\frac{1}{2}, 1)$. *The same is true of Max-Cut.*

In some ways, saying that $(1,1)$-distinguishing Max-E3-Sat is NP-hard is not necessarily that disheartening. For example, if $(1-\delta, 1)$-approximating Max-E3-Sat were possible in polynomial time for every $\delta > 0$, you might consider that "good enough". Unfortunately, such a state of affairs is very likely ruled out:

Theorem 7.35. *There exists a positive universal constant* $\delta_0 > 0$ *such that* $(1-\delta_0, 1)$-*approximating Max-E3-Sat is* NP-*hard.*

In fact, Theorem 7.35 is *equivalent* to the "PCP Theorem" mentioned in Section 7.2. It follows straightforwardly from the PCPP Theorem, as we now sketch:

Proof sketch. Let δ_0 be the rejection rate in the PCPP Theorem. We want to show that $(1-\delta_0, 1)$-approximating Max-E3-Sat is at least as hard as the Circuit-Sat problem. Equivalently, we want to show that if there is an efficient algorithm A for $(1-\delta_0, 1)$-approximating Max-E3-Sat then there is an efficient algorithm B for Circuit-Sat. So suppose A exists and let C be a Boolean circuit given as input to B. Algorithm B first applies to C the PCPP reduction given by the PCPP Theorem. The output is some arity-3 CSP instance $\mathscr{P}$ over variables $w_1, \dots, w_n, \Pi_1, \dots, \Pi_\ell$, where $\ell \leq \mathrm{poly}(\mathrm{size}(C))$. By Exercise 7.12 we may assume that $\mathscr{P}$ is an instance of Max-E3-Sat. From the definition of a PCPP system, it is easy to check (Exercise 7.19) the following: If C is satisfiable then $\mathrm{Opt}(\mathscr{P}) = 1$; and, if C is not satisfiable then $\mathrm{Opt}(\mathscr{P}) < 1 - \delta_0$. Algorithm B now runs the supposed $(1-\delta_0, 1)$-approximation algorithm A on $\mathscr{P}$ and outputs "C is satisfiable" if and only if A finds an assignment of value at least $1-\delta_0$. □

7.4. Highlight: Håstad's hardness theorems

In Theorem 7.35 we saw that it is NP-hard to $(1-\delta_0, 1)$-approximate Max-E3Sat for some positive but inexplicit constant δ_0. You might wonder how large δ_0 can be. The natural limit here is $\frac{1}{8}$ because there is a very simple algorithm that satisfies a $\frac{7}{8}$-fraction of the constraints in any Max-E3Sat instance:

Proposition 7.36. *Consider the Max-E3-Sat algorithm that outputs a uniformly random assignment F. This is a $(\frac{7}{8}, \beta)$-approximation for any β.*

Proof. In instance $\mathscr{P}$, each constraint is a logical OR of exactly 3 literals and will therefore be satisfied by F with probability exactly $\frac{7}{8}$. Hence in expectation the algorithm will satisfy a $\frac{7}{8}$-fraction of the constraints. □

(It's also easy to "derandomize" this algorithm, giving a deterministic guarantee of at least $\frac{7}{8}$ of the constraints; see Exercise 7.21.)

This algorithm is of course completely brainless – it doesn't even "look at" the instance it is trying to approximately solve. But rather remarkably, it achieves the best possible approximation guarantee among all efficient algorithms (assuming $\mathsf{P} \neq \mathsf{NP}$). This is a consequence of the following 1997 theorem of Håstad [**Hås01b**], improving significantly on Theorem 7.35:

Håstad's 3-Sat Hardness. *For any constant $\delta > 0$, it is NP-hard to $(\frac{7}{8}+\delta, 1)$-approximate Max-E3-Sat.*

Håstad gave similarly optimal hardness-of-approximation results for several other problems, including Max-E3-Lin:

Håstad's 3-Lin Hardness. *For any constant $\delta > 0$, it is NP-hard to $(\frac{1}{2}+\delta, 1-\delta)$-approximate Max-E3-Lin.*

In this hardness theorem, both the "α" and "β" parameters are optimal; as we saw in Example 7.32 one can efficiently $(\frac{1}{2}, \beta)$-approximate and also $(1,1)$-approximate Max-E3-Lin.

The goal of this section is to sketch the proof of the above theorems, mainly Håstad's 3-Lin Hardness Theorem. Let's begin by considering the 3-Sat hardness result. If our goal is to increase the inexplicit constant δ_0 in Theorem 7.35, it makes sense to look at how the constant arises. From the proof of Theorem 7.35 we see that it's just the rejection rate in the PCPP Theorem. We didn't prove that theorem, but let's consider its length-2^{2^n} analogue, Theorem 7.18. The key ingredient in the proof of Theorem 7.18 is the dictator test. Indeed, if we strip away the few local correcting and consistency checks, we see that the dictator test component controls both the rejection rate *and* the type of predicates output by the PCPP reduction. This observation suggests that to get a strong hardness-of-approximation result for, say, Max-E3-Lin, we should seek a local tester for dictatorship which (a) has a large rejection rate, and (b) makes its accept/reject decision using 3-variable linear equation predicates.

This approach (which of course needs to be integrated with efficient "PCPP technology") was suggested in a 1995 paper of Bellare, Goldreich, and Sudan [**BGS95**]. Using it, they managed to prove NP-hardness of $(1-\delta_0, 1)$-approximating Max-E3-Sat with the explicit constant $\delta_0 = .026$. Håstad's key conceptual contribution (originally from [**Hås96**]) was showing that given known PCPP technology, it suffices to construct a certain kind of *relaxed* dictator test. Roughly speaking, dictators should still be accepted with probability 1 (or close to 1), but only functions which are "very unlike" dictators need to be rejected with substantial probability. Since this is a weaker requirement than in the standard definition of a local tester, we can potentially achieve a much higher rejection rate, and hence a much stronger hardness-of-approximation result.

For these purposes, the most useful formalization of being "very unlike a dictator" turns out to be "having no notable coordinates" in the sense of Definition 6.9. We make the following definition which is appropriate for Boolean CSPs.

Definition 7.37. Let Ψ be a finite set of predicates over the domain $\Omega = \{-1,1\}$. Let $0 < \alpha < \beta \leq 1$ and let $\lambda : [0,1] \to [0,1]$ satisfy $\lambda(\epsilon) \to 0$ as $\epsilon \to 0$. Suppose that for each $n \in \mathbb{N}^+$ there is a local tester for functions $f : \{-1,1\}^n \to \{-1,1\}$ with the following properties:

- If f is a dictator then the test accepts with probability at least β.
- If f has no (ϵ,ϵ)-notable coordinates – i.e., $\mathbf{Inf}_i^{(1-\epsilon)}[f] \leq \epsilon$ for all $i \in [n]$ – then the test accepts with probability at most $\alpha + \lambda(\epsilon)$.
- The tester's accept/reject decision uses predicates from Ψ; i.e., the tester can be viewed as an instance of Max-CSP(Ψ).

Then, abusing terminology, we call this family of testers an *(α,β)-Dictator-vs.-No-Notables test using predicate set Ψ.*

Remark 7.38. For very minor technical reasons, the above definition should actually be slightly amended. In this section we freely ignore the amendments, but for the sake of correctness we state them here. One is a strengthening, one is a weakening.

- The second condition should be required even for functions $f : \{-1,1\}^n \to [-1,1]$; what this means is explained in Exercise 7.22.

- When the tester makes accept/reject decisions by applying $\psi \in \Psi$ to query results $f(\boldsymbol{x}^{(1)}), \dots, f(\boldsymbol{x}^{(r)})$, it is *allowed* that the query strings are not all distinct. (See Exercise 7.31.)

Remark 7.39. It's essential in this definition that the "error term" $\lambda(\epsilon) = o_\epsilon(1)$ be independent of n. On the other hand, we otherwise care very little about the rate at which it tends to 0; this is why we didn't mind using the same parameter ϵ in the "(ϵ, ϵ)-notable" hypothesis.

Just as the dictator test was the key component in our PCPP reduction (Theorem 7.18), Dictator-vs.-No-Notables tests are the key to obtaining strong hardness-of-approximation results. The following result (essentially proved in Khot et al. [KKMO07]) lets you obtain hardness results from Dictator-vs.-No-Notables tests in a black-box way:

Theorem 7.40. *Fix a CSP over domain $\Omega = \{-1, 1\}$ with predicate set Ψ. Suppose there exists an (α, β)-Dictator-vs.-No-Notables test using predicate set Ψ. Then for all $\delta > 0$, it is "UG-hard" to $(\alpha + \delta, \beta - \delta)$-approximate Max-*CSP($\Psi$).

In other words, the distinguishing parameters of a Dictator-vs.-No-Notables test automatically translate to the distinguishing parameters of a hardness result (up to an arbitrarily small δ).

The advantage of Theorem 7.40 is that it reduces a problem about computational complexity to a purely Fourier-analytic problem, and a constructive one at that. The theorem has two disadvantages, however. The first is that instead of NP-hardness – the gold standard in complexity theory – it merely gives "UG-hardness", which roughly means "at least as hard as the Unique-Games problem". We leave the definition of the Unique-Games problem to Exercise 7.27, but suffice it to say it's not as universally believed to be hard as Circuit-Sat is. The second disadvantage of Theorem 7.40 is that it only has $\beta - \delta$ rather than β. This can be a little disappointing, especially when you are interested in hardness for satisfiable instances ($\beta = 1$), as in Håstad's 3-Sat Hardness. In his work, Håstad showed that both disadvantages can be erased provided you construct something similar to, but more complicated than, an (α, β)-Dictator-vs.-No-Notables test. This is how the Håstad 3-Sat and 3-Lin Hardness Theorems are proved. Describing this extra complication is beyond the scope of this book; therefore we content ourselves with the following theorems:

Theorem 7.41. *For any $0 < \delta < \frac{1}{8}$, there exists a $(\frac{7}{8} + \delta, 1)$-Dictator-vs.-No-Notables test which uses logical OR functions on 3 literals as its predicates.*

Theorem 7.42. *For any $0 < \delta < \frac{1}{2}$, there exists a $(\frac{1}{2}, 1 - \delta)$-Dictator-vs.-No-Notables test using 3-variable $\mathbb{F}_2$-linear equations as its predicates.*

Theorem 7.42 will be proved below, while the proof of Theorem 7.41 is left for Exercise 7.29. By applying Theorem 7.40 we immediately deduce the following weakened versions of Håstad's Hardness Theorems:

Corollary 7.43. *For any $\delta > 0$, it is UG-hard to $(\frac{7}{8}+\delta, 1-\delta)$-approximate Max-E3-Sat.*

Corollary 7.44. *For any $\delta > 0$, it is UG-hard to $(\frac{1}{2}+\delta, 1-\delta)$-approximate Max-E3-Lin.*

Remark 7.45. For Max-E3-Lin, we don't mind the fact that Theorem 7.40 has $\beta-\delta$ instead of β because our Dictator-vs.-No-Notables test only accepts dictators with probability $1-\delta$ anyway. Note that the $1-\delta$ in Theorem 7.42 cannot be improved to 1; see Exercise 7.7.)

To prove a result like Theorem 7.42 there are two components: the design of the test, and its analysis. We begin with the design. Since we are looking for a test using 3-variable linear equation predicates, the BLR Test naturally suggests itself; indeed, all of its checks are of the form $f(x)+f(y)+f(z)=0$. It also accepts dictators with probability 1. Unfortunately it's not true that it accepts functions with no notable coordinates with probability close to $\frac{1}{2}$. There are two problems: the constant 0 function and "large" parity functions are both accepted with probability 1, despite having no notable coordinates. The constant 1 function is easy to deal with: we can replace the BLR Test by the "Odd BLR Test".

Odd BLR Test. *Given query access to $f : \mathbb{F}_2^n \to \mathbb{F}_2$:*

- *Choose $\boldsymbol{x} \sim \mathbb{F}_2^n$ and $\boldsymbol{y} \sim \mathbb{F}_2^n$ independently.*
- *Choose $\boldsymbol{b} \sim \mathbb{F}_2$ uniformly at random and set $\boldsymbol{z} = \boldsymbol{x}+\boldsymbol{y}+(\boldsymbol{b},\boldsymbol{b},\dots,\boldsymbol{b}) \in \mathbb{F}_2^n$.*
- *Accept if $f(\boldsymbol{x})+f(\boldsymbol{y})+f(\boldsymbol{z}) = \boldsymbol{b}$.*

Note that this test uses both kinds of 3-variable linear equations as its predicates. For the test's analysis, we as usual switch to ±1 notation and think of testing $f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z}) = \boldsymbol{b}$. It is easy to show the following (see the proof of Theorem 7.42, or Exercise 7.15 for a generalization):

Proposition 7.46. *The Odd BLR Test accepts $f : \{-1,1\}^n \to \{-1,1\}$ with probability*

$$\tfrac{1}{2} + \tfrac{1}{2} \sum_{\substack{S \subseteq [n] \\ |S| \text{ odd}}} \widehat{f}(S)^3 \leq \tfrac{1}{2} + \tfrac{1}{2} \max_{\substack{S \subseteq [n] \\ |S| \text{ odd}}} \{\widehat{f}(S)\}.$$

This twist rules out the constant 1 function; it passes the Odd BLR Test with probability $\frac{1}{2}$. It remains to deal with large parity functions. Håstad's innovation here was to add a small amount of *noise* to the Odd BLR Test. Specifically, given a small $\delta > 0$ we replace $\boldsymbol{z}$ in the above test with $\boldsymbol{z}' \sim$

$N_{1-\delta}(\boldsymbol{z})$; i.e., we flip each of its bits with probability $\delta/2$. If f is a dictator, then there is only a $\delta/2$ chance this will affect the test. On the other hand, if f is a parity of large cardinality, the cumulative effect of the noise will destroy its chance of passing the linearity test. Note that parities of small odd cardinality will also pass the test with probability close to 1; however, we don't need to worry about them since they have notable coordinates. We can now present Håstad's Dictator-vs.-No-Notables test for Max-E3-Lin.

Proof of Theorem 7.42. Given a parameter $0 < \delta < 1$, define the following test, which uses Max-E3-Lin predicates:

Håstad$_\delta$ Test. *Given query access to* $f : \{-1,1\}^n \to \{-1,1\}$*:*

- *Choose* $\boldsymbol{x}, \boldsymbol{y} \sim \{-1,1\}^n$ *uniformly and independently.*
- *Choose bit* $\boldsymbol{b} \sim \{-1,1\}$ *uniformly and set* $\boldsymbol{z} = \boldsymbol{b} \cdot (\boldsymbol{x} \circ \boldsymbol{y}) \in \{-1,1\}^n$ *(where* $\circ$ *denotes entry-wise multiplication).*
- *Choose* $\boldsymbol{z}' \sim N_{1-\delta}(\boldsymbol{z})$*.*
- *Accept if* $f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z}') = \boldsymbol{b}$*.*

We will show that this is a $(\frac{1}{2}, 1-\delta/2)$-Dictator-vs.-No-Notables test. First, let us analyze the test assuming $\boldsymbol{b} = 1$.

$$\begin{aligned}
\mathbf{Pr}[\text{Håstad}_\delta \text{ Test accepts } f \mid \boldsymbol{b} = 1] &= \mathbf{E}[\tfrac{1}{2} + \tfrac{1}{2} f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z}')] \\
&= \tfrac{1}{2} + \tfrac{1}{2}\mathbf{E}[f(\boldsymbol{x}) \cdot f(\boldsymbol{y}) \cdot \mathrm{T}_{1-\delta} f(\boldsymbol{x} \circ \boldsymbol{y})]] \\
&= \tfrac{1}{2} + \tfrac{1}{2}\mathop{\mathbf{E}}_{\boldsymbol{x}}[f(\boldsymbol{x}) \cdot (f * \mathrm{T}_{1-\delta} f)(\boldsymbol{x})] \\
&= \tfrac{1}{2} + \tfrac{1}{2}\sum_{S \subseteq [n]} \widehat{f}(S) \cdot \widehat{f * \mathrm{T}_{1-\delta} f}(S) \\
&= \tfrac{1}{2} + \tfrac{1}{2}\sum_{S \subseteq [n]} (1-\delta)^{|S|} \widehat{f}(S)^3.
\end{aligned}$$

On the other hand, when $\boldsymbol{b} = -1$ we take the expectation of $\frac{1}{2} - \frac{1}{2} f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z}')$ and note that $\boldsymbol{z}'$ is distributed as $N_{-(1-\delta)}(\boldsymbol{x} \circ \boldsymbol{y})$. Thus

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

Averaging the above two results we deduce

$$\mathbf{Pr}[\text{Håstad}_\delta \text{ Test accepts } f] = \tfrac{1}{2} + \tfrac{1}{2}\sum_{|S| \text{ odd}} (1-\delta)^{|S|} \widehat{f}(S)^3. \tag{7.4}$$

(Incidentally, by taking $\delta = 0$ here we obtain the proof of Proposition 7.46.)

From (7.4) we see that if f is a dictator, $f = \chi_S$ with $|S| = 1$, then it is accepted with probability $1 - \delta/2$. (It's also easy to see this directly from the definition of the test.) To complete the proof that we have a $(\frac{1}{2}, 1-\delta/2)$-Dictator-vs.-No-Notables test, we need to bound the probability that f is accepted given that it has (ϵ,ϵ)-small stable influences. More precisely, assuming

$$\mathbf{Inf}_i^{(1-\epsilon)}[f] = \sum_{S \ni i} (1-\epsilon)^{|S|-1} \widehat{f}(S)^2 \leq \epsilon \quad \text{for all } i \in [n] \tag{7.5}$$

we will show that

$$\mathbf{Pr}[\text{Håstad}_\delta \text{ Test accepts } f] \leq \tfrac{1}{2} + \tfrac{1}{2}\sqrt{\epsilon}, \quad \text{provided } \epsilon \leq \delta. \tag{7.6}$$

This is sufficient because we can take $\lambda(\epsilon)$ in Definition 7.37 to be

$$\lambda(\epsilon) = \begin{cases} \frac{1}{2}\sqrt{\epsilon} & \text{for } \epsilon \leq \delta, \\ \frac{1}{2} & \text{for } \epsilon > \delta. \end{cases}$$

Now to obtain (7.6), we continue from (7.4):

$$\begin{aligned} \mathbf{Pr}[\text{Håstad}_\delta \text{ Test accepts } f] &\leq \tfrac{1}{2} + \tfrac{1}{2} \max_{|S| \text{ odd}} \{(1-\delta)^{|S|} \widehat{f}(S)\} \cdot \sum_{|S| \text{ odd}} \widehat{f}(S)^2 \\ &\leq \tfrac{1}{2} + \tfrac{1}{2} \max_{|S| \text{ odd}} \{(1-\delta)^{|S|} \widehat{f}(S)\} \\ &\leq \tfrac{1}{2} + \tfrac{1}{2} \sqrt{\max_{|S| \text{ odd}} \{(1-\delta)^{2|S|} \widehat{f}(S)^2\}} \\ &\leq \tfrac{1}{2} + \tfrac{1}{2} \sqrt{\max_{|S| \text{ odd}} \{(1-\delta)^{|S|-1} \widehat{f}(S)^2\}} \\ &\leq \tfrac{1}{2} + \tfrac{1}{2} \sqrt{\max_{i \in [n]} \{\mathbf{Inf}_i^{(1-\delta)}[f]\}}, \end{aligned}$$

where we used that $|S|$ odd implies S nonempty. And the above is indeed at most $\frac{1}{2} + \frac{1}{2}\sqrt{\epsilon}$ provided $\epsilon \leq \delta$, by (7.5). □

7.5. Exercises and notes

7.1 Suppose there is an r-query local tester for property $\mathscr{C}$ with rejection rate λ. Show that there is a testing algorithm that, given inputs $0 < \epsilon, \delta \leq 1/2$, makes $O(\frac{r \log(1/\delta)}{\lambda \epsilon})$ (nonadaptive) queries to f and satisfies the following:

- If $f \in \mathscr{C}$, then the tester accepts with probability 1.
- If f is ϵ-far from $\mathscr{C}$, then the tester accepts with probability at most δ.

7.2 Let $\mathscr{M} = \{(x,y) \in \{0,1\}^{2n} : x = y\}$, the property that a string's first half matches its second half. Give a 2-query local tester for $\mathscr{M}$ with rejection rate 1. (Hint: Locally test that $x \oplus y = (0,0,\dots,0)$.)

7.3 Reduce the proof length in Example 7.15 to $n-2$.

7.4 Verify the claim from Example 7.12 regarding the 2-query tester for the property that a string has all its coordinates equal. (Hint: Use ±1 notation.)

7.5 Let $\mathcal{O} = \{w \in \mathbb{F}_2^n : w$ has an odd number of 1's$\}$. Let T be any $(n-1)$-query string testing algorithm that accepts every $w \in \mathcal{O}$ with probability 1. Show that T in fact accepts *every* string $v \in \mathbb{F}_2^n$ with probability 1 (even though $\mathrm{dist}(w, \mathcal{O}) = \frac{1}{n} > 0$ for half of all strings w). Thus locally testing $\mathcal{O}$ requires n queries.

7.6 Let T be a 2-query testing algorithm for functions $\{-1,1\}^n \to \{-1,1\}$. Suppose that $\mathcal{T}$ accepts every dictator with probability 1. Show that it also accepts $\mathrm{Maj}_{n'}$ with probability 1 for every odd $n' \leq n$. This shows that there is no 2-query local tester for dictatorship assuming $n > 2$. (Hint: You'll need to enumerate all predicates on up to 2 bits.)

7.7 For every $\alpha < 1$, show that there is no $(\alpha, 1)$-Dictator-vs.-No-Notables test using Max-E3-Lin predicates. (Hint: Consider large odd parities.)

7.8 (*a*) Consider the following 3-query testing algorithm for $f : \{0,1\}^n \to \{0,1\}$. Let $\boldsymbol{x}, \boldsymbol{y} \sim \{0,1\}^n$ be independent and uniformly random, define $\boldsymbol{z} \in \{0,1\}^n$ by $\boldsymbol{z}_i = \boldsymbol{x}_i \wedge \boldsymbol{y}_i$ for each $i \in [n]$, and accept if $f(\boldsymbol{x}) \wedge f(\boldsymbol{y}) = f(\boldsymbol{z})$. Let p_k be the probability that this test accepts a parity function $\chi_S : \{0,1\}^n \to \{0,1\}$ with $|S| = k$. Show that $p_0 = p_1 = 1$ and that in general $p_k \leq \frac{1}{2} + 2^{-|S|}$. In fact, you might like to show that $p_k = \frac{1}{2} + (\frac{3}{4} - \frac{1}{4}(-1)^k)2^{-k}$. (Hint: It suffices to consider $k = n$ and then compute the correlation of $\chi_{\{1,\dots,n\}} \wedge \chi_{\{n+1,\dots,2n\}}$ with the bent function IP_{2n}.)

(*b*) Show how to obtain a 3-query local tester for dictatorship by combining the following subtests: (i) the Odd BLR Test; (ii) the test from part (*a*).

7.9 Obtain the largest explicit rejection rate in Theorem 7.7 that you can. You might want to return to the Fourier expressions arising in Theorem 1.30 and 2.56, as well as Exercise 1.28. Can you improve your bound by doing the BLR and NAE Tests with probabilities other than 1/2, 1/2?

7.10 (*a*) Say that A is an (α, β)-*distinguishing* algorithm for Max-CSP(Ψ) if it outputs 'YES' on instances with value at least β and outputs 'NO' on instances with value strictly less than α. (On each instance with value in $[\alpha, \beta)$, algorithm A may have either output.) Show that if there is an efficient (α, β)-approximation algorithm for Max-CSP(Ψ), then there is also an efficient (α, β)-distinguishing algorithm for Max-CSP(Ψ).

(*b*) Consider Max-CSP(Ψ), where Ψ be a class of predicates that is closed under restrictions (to nonconstant functions); e.g., Max-3-Sat. Show that if there is an efficient $(1,1)$-distinguishing algorithm, then there is also an efficient $(1,1)$-approximation algorithm. (Hint: Try out all labels for the first variable and use the distinguisher.)

7.11 (*a*) Let ϕ be a CNF of size s and width $w \geq 3$ over variables $x_1, \dots, x_n$. Show that there is an "equivalent" CNF ϕ' of size at most $(w-2)s$ and width 3 over the variables $x_1,\dots,x_n$ plus auxiliary variables $\Pi_1,\dots,\Pi_\ell$, with $\ell \leq (w-3)s$. Here "equivalent" means that for every x such that $\phi(x) = \mathsf{True}$ there exists Π such that $\phi'(x,\Pi) = \mathsf{True}$; and, for every x such that $\phi(x) = \mathsf{False}$ we have $\phi'(x,\Pi) = \mathsf{False}$ for all Π.

(b) Extend the above so that every clause in ϕ' has width *exactly* 3 (the size may increase by $O(s)$).

7.12 Suppose there exists an r-query PCPP reduction $\mathscr{R}_1$ with rejection rate λ. Show that there exists a 3-query PCPP reduction $\mathscr{R}_2$ with rejection rate at least $\lambda/(r2^r)$. The proof length of $\mathscr{R}_2$ should be at most $r2^r \cdot m$ plus the proof length of $\mathscr{R}_1$ (where m is the description-size of $\mathscr{R}_1$'s output) and the predicates output by the reduction should all be logical ORs applied to exactly three literals. (Hint: Exercises 4.1, 7.11.)

7.13 (a) Give a polynomial-time algorithm R that takes as input a general Boolean circuit C and outputs a width-3 CNF formula ϕ with the following guarantee: C is satisfiable if and only if ϕ is satisfiable. (Hint: Introduce a variable for each gate in C.)

(b) The previous exercise in fact formally justifies the following statement: "(1,1)-distinguishing Max-3-Sat is NP-hard". (See Exercise 7.10 for the definition of (1,1)-distinguishing.) Argue that, indeed, if (1,1)-distinguishing (or (1,1)-approximating) Max-3-Sat is in polynomial time, then so is Circuit-Sat.

(c) Prove Theorem 7.33. (Hint: Exercise 7.11(b).)

7.14 Describe an efficient (1,1)-approximation algorithm for Max-Cut.

7.15 (a) Let H be any subspace of $\mathbb{F}_2^n$ and let $\mathscr{H} = \{\chi_\gamma : \mathbb{F}_2^n \to \{-1,1\} \mid \gamma \in H^\perp\}$. Give a 3-query local tester for $\mathscr{H}$ with rejection rate 1. (Hint: Similar to BLR, but with $\langle \varphi_H * f, f * f\rangle$.)

(b) Generalize to the case that H is any affine subspace of $\mathbb{F}_2^n$.

7.16 Let A be any affine subspace of $\mathbb{F}_2^n$. Construct a 3-query, length-2^n PCPP system for A with rejection rate a positive universal constant. (Hint: Given $w \in \mathbb{F}_2^n$, the tester should expect the proof $\Pi \in \{-1,1\}^{2^n}$ to encode the truth table of χ_w. Use Exercise 7.15 and also a consistency check based on local correcting of Π at $e_{\boldsymbol{i}}$, where $\boldsymbol{i} \in [n]$ is uniformly random.)

7.17 (a) Give a 3-query, length-$O(n)$ PCPP system (with rejection rate a positive universal constant) for the class $\{w \in \mathbb{F}_2^n : \mathrm{IP}_n(w) = 1\}$, where IP_n is the inner product mod 2 function (n even).

(b) Do the same for the complete quadratic function CQ_n from Exercise 1.1. (Hint: Exercise 4.13.)

7.18 In this exercise you will prove Theorem 7.19.

(a) Let $D \in \mathbb{F}_2^{n\times n}$ be a nonzero matrix and suppose $\boldsymbol{x},\boldsymbol{y} \sim \mathbb{F}_2^n$ are uniformly random and independent. Show that $\mathbf{Pr}[\boldsymbol{y}^\top D\boldsymbol{x} \neq 0] \geq \frac{1}{4}$.

(*b*) Let $\gamma \in \mathbb{F}_2^n$ and $\Gamma \in \mathbb{F}_2^{n\times n}$. Suppose $\boldsymbol{x}, \boldsymbol{y} \sim \mathbb{F}_2^n$ are uniformly random and independent. Show that $\mathbf{Pr}[(\gamma^\top \boldsymbol{x})(\gamma^\top \boldsymbol{y}) = \Gamma \bullet (\boldsymbol{x}\boldsymbol{y}^\top)]$ is 1 if $\Gamma = \gamma\gamma^\top$ and is at most $\frac{3}{4}$ otherwise. Here we use the notation $B \bullet C = \sum_{i,j} B_{ij}C_{ij}$ for matrices $B, C \in \mathbb{F}_2^{n\times n}$.

(*c*) Suppose you are given query access to *two* functions $\ell : \mathbb{F}_2^n \to \mathbb{F}_2$ and $q : \mathbb{F}_2^{n\times n} \to \mathbb{F}_2$. Give a 4-query testing algorithm with the following two properties (for some universal constant $\lambda > 0$): (i) if $\ell = \chi_\gamma$ and $q = \chi_{\gamma\gamma^\top}$ for some $\gamma \in \mathbb{F}_2^n$, the test accepts with probability 1; (ii) for all $0 \le \epsilon \le 1$, if the test accepts with probability at least $1 - \gamma \cdot \epsilon$, then there exists some $\gamma \in \mathbb{F}_2^n$ such that ℓ is ϵ-close to χ_γ and q is ϵ-close to $\chi_{\gamma\gamma^\top}$. (Hint: Apply the BLR Test to ℓ and q, and use part (*b*) with local correcting on q.)

(*d*) Let L be a list of homogenous degree-2 polynomial equations over variables $w_1, \dots, w_n \in \mathbb{F}_2$. (Each equation is of the form $\sum_{i,j=1}^n c_{ij} w_i w_j = b$ for constants $b, c_{ij} \in \mathbb{F}_2$; we remark that $w_i^2 = w_i$.) Define the string property $\mathscr{L} = \{w \in \mathbb{F}_2^n : w \text{ satisfies all equations in L}\}$. Give a 4-query, length-$(2^n + 2^{n^2})$ PCPP system for $\mathscr{L}$ (with rejection rate a positive universal constant). (Hint: The tester should expect the truth table of χ_w and $\chi_{ww^\top}$. You will need part (*c*) as well as Exercise 7.15 applied to "q".)

(*e*) Complete the proof of Theorem 7.19. (Hints: given $w \in \{0,1\}^n$, the tester should expect a proof consisting of all gate values $\bar{w} \in \{0,1\}^{\mathrm{size}(C)}$ in C's computation on w, as well as truth tables of $\chi_{\bar{w}}$ and $\chi_{\bar{w}\bar{w}^\top}$. Show that $\bar{w}$ being a valid computation of C is encodable with a list of homogeneous degree-2 polynomial equations. Add a consistency check between w and $\bar{w}$ using local correcting, and reduce the number of queries to 3 using Exercise 7.12.)

7.19 Verify the connection between $\mathrm{Opt}(\mathscr{P})$ and C's satisfiability stated in the proof sketch of Theorem 7.35. (Hint: Every string w is 1-far from the empty property.)

7.20 A *randomized assignment* for an instance $\mathscr{P}$ of a CSP over domain Ω is a mapping $\boldsymbol{F}$ that labels each variable in V with a *probability distribution* over domain elements. Given a constraint (S, ψ) with $S = (v_1, \dots, v_r)$, we write $\psi(\boldsymbol{F}(S)) \in [0,1]$ for the expected value of $\psi(\boldsymbol{F}(v_1), \dots, \boldsymbol{F}(v_r))$. This is simply the probability that ψ is satisfied when one actually draws from the domain-distributions assigned by $\boldsymbol{F}$. Finally, we define the *value* of $\boldsymbol{F}$ to be $\mathrm{Val}_{\mathscr{P}}(\boldsymbol{F}) = \mathbf{E}_{(\boldsymbol{S},\boldsymbol{\psi})\sim\mathscr{P}}[\boldsymbol{\psi}(\boldsymbol{F}(\boldsymbol{S}))]$.

(*a*) Suppose that A is a deterministic algorithm that produces a randomized assignment of value α on a given instance $\mathscr{P}$. Show a simple modification to A that makes it a randomized algorithm that produces a (normal) assignment whose value is α in expectation. (Thus, in constructing approximation algorithms we may allow ourselves to output randomized assignments.)

(*b*) Let A be the deterministic Max-E3-Sat algorithm that on every instance outputs the randomized assignment that assigns the uniform distribution on $\{0,1\}$ to each variable. Show that this is a $(\frac{7}{8},\beta)$-approximation algorithm for any β. Show also that the same algorithm is a $(\frac{1}{2},\beta)$-approximation algorithm for Max-3-Lin.

(*c*) When the domain Ω is $\{-1,1\}$, we may model a randomized assignment as a function $f : V \to [-1,1]$; here $f(v) = \mu$ is interpreted as the unique probability distribution on $\{-1,1\}$ which has mean μ. Now given a constraint (S,ψ) with $S = (v_1,\dots,v_r)$, show that the value of f on this constraint is in fact $\psi(f(v_1),\dots,f(v_r))$, where we identify $\psi : \{-1,1\}^r \to \{0,1\}$ with its multilinear (Fourier) expansion. (Hint: Exercise 1.4.)

(*d*) Let Ψ be a collection of predicates over domain $\{-1,1\}$. Let $\nu = \min_{\psi\in\Psi}\{\widehat{\psi}(\emptyset)\}$. Show that outputting the randomized assignment $f \equiv 0$ is an efficient (ν,β)-approximation algorithm for Max-CSP(Ψ).

7.21 Let $\boldsymbol{F}$ be a randomized assignment of value α for CSP instance $\mathscr{P}$ (as in Exercise 7.20). Give an efficient deterministic algorithm that outputs a usual assignment F of value at least α. (Hint: Try all possible labelings for the first variable and compute the expected value that would be achieved if $\boldsymbol{F}$ were used for the remaining variables. Pick the best label for the first variable and repeat.)

7.22 Given a local tester for functions $f : \{-1,1\}^n \to \{-1,1\}$, we can interpret it also as a tester for functions $f : \{-1,1\}^n \to [-1,1]$; simply view the tester as a CSP and view the acceptance probability as the value of f when treated as a randomized assignment (as in Exercise 7.20(*c*)). Equivalently, whenever the tester "queries" $f(x)$, imagine that what is returned is a random bit $\boldsymbol{b} \in \{-1,1\}$ whose mean is $f(x)$. This interpretation completes Definition 7.37 of Dictator-vs.-No-Notables tests for functions $f : \{-1,1\}^n \to [-1,1]$ (see Remark 7.38). Given this definition, verify that the Håstad$_\delta$ Test is indeed a $(\frac{1}{2}, 1-\delta)$-Dictator-vs.-No-Notables test. (Hint: Show that (7.4) still holds for functions $f : \{-1,1\}^n \to [-1,1]$. There is only one subsequent inequality that uses that f's range is $\{-1,1\}$, and it still holds with range $[-1,1]$.)

7.23 Let Ψ be a finite set of predicates over domain $\Omega = \{-1,1\}$ that is closed under negating variables. (An example is the scenario of Max-ψ from Remark 7.23.) In this exercise you will show that Dictator-vs.-No-Notables tests using Ψ may assume $f : \{-1,1\}^n \to [-1,1]$ is odd without loss of generality.

(a) Let T be an (α,β)-Dictator-vs.-No-Notables test using predicate set Ψ that works under the assumption that $f : \{-1,1\}^n \to [-1,1]$ is odd. Modify T as follows: Whenever it is about to query $f(x)$, with probability $\frac{1}{2}$ let it use $f(x)$ and with probability $\frac{1}{2}$ let it use $-f(-x)$. Call the modified test T'. Show that the probability T' accepts an arbitrary $f : \{-1,1\}^n \to [-1,1]$ is equal to the probability T accepts f^{odd} (recall Exercise 1.8).

(b) Prove that T' is an (α,β)-Dictator-vs.-No-Notables test using predicate set Ψ for functions $f : \{-1,1\}^n \to [-1,1]$.

7.24 This problem is similar to Exercise 7.23 in that it shows you may assume that Dictator-vs.-No-Notables tests are testing "smoothed" functions of the form $\mathrm{T}_{1-\delta}h$ for $h : \{-1,1\}^n \to [-1,1]$, so long as you are willing to lose $O(\delta)$ in the probability that dictators are accepted.

(a) Let U be an (α,β)-Dictator-vs.-No-Notables test using an arity-r predicate set Ψ (over domain $\{-1,1\}$) which works under the assumption that the function $f : \{-1,1\}^n \to [-1,1]$ being tested is of the form $\mathrm{T}_{1-\delta}h$ for $h : \{-1,1\}^n \to [-1,1]$. Modify U as follows: whenever it is about to query $f(x)$, let it draw $\boldsymbol{y} \sim N_{1-\delta}(x)$ and use $f(\boldsymbol{y})$ instead. Call the modified test U'. Show that the probability U' accepts an arbitrary $h : \{-1,1\}^n \to [-1,1]$ is equal to the probability U accepts $\mathrm{T}_{1-\delta}h$.

(b) Prove that U' is an $(\alpha,\beta - r\delta/2)$-Dictator-vs.-No-Notables test using predicate set Ψ.

7.25 Give a slightly alternate proof of Theorem 7.42 by using the original BLR Test analysis and applying Exercises 7.23, 7.24.

7.26 Show that when using Theorem 7.40, it suffices to have a "Dictators-vs.-No-Influentials test", meaning replacing $\mathbf{Inf}_i^{(1-\epsilon)}[f]$ in Definition 7.37 with just $\mathbf{Inf}_i[f]$. (Hint: Exercise 7.24.)

7.27 For $q \in \mathbb{N}^+$, *Unique-Games(q)* refers to the arity-2 CSP with domain $\Omega = [q]$ in which all $q!$ "bijective" predicates are allowed; here ψ is "bijective" if there is a bijection $\pi : [q] \to [q]$ such that $\psi(i,j) = 1$ iff $\pi(j) = i$. Show that $(1,1)$-approximating Unique-Games(q) can be done in polynomial time. (The *Unique Games Conjecture* of Khot [**Kho02**] states that for all $\delta > 0$ there exists $q \in \mathbb{N}^+$ such that $(\delta, 1-\delta)$-approximating Unique-Games(q) is NP-hard.)

7.28 In this problem you will show that Corollary 7.43 actually follows directly from Corollary 7.44.

(a) Consider the $\mathbb{F}_2$-linear equation $v_1 + v_2 + v_3 = 0$. Exhibit a list of 4 clauses (i.e., logical ORs of literals) over the variables such that if the equation is satisfied, then so are all 4 clauses, but if the equation is not satisfied, then at most 3 of the clauses are. Do the same for the equation $v_1 + v_2 + v_3 = 1$.

(*b*) Suppose that for every $\delta > 0$ there is an efficient algorithm for $(\frac{7}{8} + \delta, 1-\delta)$-approximating Max-E3-Sat. Give, for every $\delta > 0$, an efficient algorithm for $(\frac{1}{2} + \delta, 1-\delta)$-approximating Max-E3-Lin.

(*c*) Alternatively, show how to transform any (α, β)-Dictator-vs.-No-Notables test using Max-E3-Lin predicates into a $(\frac{3}{4} + \frac{1}{4}\alpha, \beta)$-Dictator-vs.-No-Notables test using Max-E3-Sat predicates.

7.29 In this exercise you will prove Theorem 7.41.

(*a*) Recall the predicate OXR from Exercise 1.1. Fix a small $0 < \delta < 1$. The remainder of the exercise will be devoted to constructing a $(\frac{3}{4} + \delta/4, 1)$-Dictator-vs.-No-Notables test using Max-OXR predicates. Show how to convert this to a $(\frac{7}{8} + \delta/8, 1)$-Dictator-vs.-No-Notables test using Max-E3-Sat predicates. (Hint: Similar to Exercise 7.28(*c*).)

(*b*) By Exercise 7.23, it suffices to construct a $(\frac{3}{4} + \delta/4, 1)$-Dictator-vs.-No-Notables test using the OXR predicate assuming $f : \{-1,1\}^n \to [-1,1]$ is odd. Håstad tests $\mathrm{OXR}(f(\boldsymbol{x}), f(\boldsymbol{y}), f(\boldsymbol{z}))$ where $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{-1,1\}^n$ are chosen randomly as follows: For each $i \in [n]$ (independently), with probability $1-\delta$ choose $(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ uniformly subject to $\boldsymbol{x}_i\boldsymbol{y}_i\boldsymbol{z}_i = -1$, and with probability δ choose $(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ uniformly subject to $\boldsymbol{y}_i\boldsymbol{z}_i = -1$. Show that the probability this test accepts an odd $f : \{-1,1\}^n \to [-1,1]$ is

$$\frac{3}{4} - \frac{1}{4}\mathbf{Stab}_{-\delta}[f] - \frac{1}{4}\sum_{S \subseteq [n]} \widehat{f}(S)^2 \mathop{\mathbf{E}}_{\boldsymbol{J} \subseteq_{1-\delta} S}[(-1)^{|\boldsymbol{J}|}\widehat{f}(\boldsymbol{J})], \tag{7.7}$$

where $\boldsymbol{J} \subseteq_{1-\delta} S$ denotes that $\boldsymbol{J}$ is a $(1-\delta)$-random subset of S in the sense of Definition 4.15. In particular, show that dictators are accepted with probability 1.

(*c*) Upper-bound (7.7) by

$$\frac{3}{4} + \delta/4 + \frac{1}{4}\sqrt{(1-\delta)^t} + \frac{1}{4}\sum_{|S| \leq t} \widehat{f}(S)^2 \mathop{\mathbf{E}}_{\boldsymbol{J} \subseteq_{1-\delta} S}[|\widehat{f}(\boldsymbol{J})|],$$

or something stronger. (Hint: Cauchy–Schwarz.)

(*d*) Complete the proof that this is a $(\frac{3}{4} + \delta/4, 1)$-Dictator-vs.-No-Notables test, assuming f is odd.

7.30 In this exercise you will prove Theorem 7.40. Assume there exists an (α, β)-Dictator-vs.-No-Notables test T using predicate set Ψ over domain $\{-1,1\}$. We define a certain efficient algorithm R, which takes as input an instance $\mathcal{G}$ of Unique-Games(q) and outputs an instance $\mathscr{P}$ of Max-CSP(Ψ). For simplicity we refer to the variables V of the Unique-Games instance $\mathcal{G}$ as "vertices" and its constraints as "edges". We also assume that when $\mathcal{G}$ is viewed as an undirected graph, it is regular. (By a result of Khot–Regev [KR08] this assumption is without loss of generality for the purposes of the Unique Games Conjecture.) The Max-CSP(Ψ) instance $\mathscr{P}$ output by algorithm R will have variable set $V \times \{-1,1\}^q$, and we write assignments for it as collections of functions $(f_v)_{v \in V}$, where $f : \{-1,1\}^q \to \{-1,1\}$. The draw of a random of constraint for $\mathscr{P}$ is defined as follows:

- Choose $\boldsymbol{u} \in V$ uniformly at random.
- Draw a random constraint from the test T; call it $\boldsymbol{\psi}(f(\boldsymbol{x}^{(1)}), \dots, f(\boldsymbol{x}^{(\boldsymbol{r})}))$.
- Choose $\boldsymbol{r}$ random "neighbors" $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{\boldsymbol{r}}$ of $\boldsymbol{u}$ in $\mathscr{G}$, independently and uniformly. (By a neighbor of $\boldsymbol{u}$, we mean a vertex v such that either $(\boldsymbol{u}, v)$ or $(v, \boldsymbol{u})$ is the scope of a constraint in $\mathscr{G}$.) Since $\mathscr{G}$'s constraints are bijective, we may assume that the associated scopes are $(\boldsymbol{u}, \boldsymbol{v}_1), \dots, (\boldsymbol{u}, \boldsymbol{v}_{\boldsymbol{r}})$ with bijections $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{\boldsymbol{r}} : [q] \to [q]$.
- Output the constraint $\boldsymbol{\psi}(f_{\boldsymbol{v}_1}^{\boldsymbol{\pi}_1}(\boldsymbol{x}^{(1)}), \dots, \boldsymbol{\psi}(f_{\boldsymbol{v}_{\boldsymbol{r}}}^{\boldsymbol{\pi}_{\boldsymbol{r}}}(\boldsymbol{x}^{(\boldsymbol{r})}))$, where we use the permutation notation f^{π} from Exercise 1.30.

(a) Suppose $\mathrm{Opt}(\mathscr{G}) \geq 1 - \delta$. Show that there is an assignment for $\mathscr{P}$ with value at least $\beta - O(\delta)$ in which each f_v is a dictator. (You will use regularity of $\mathscr{G}$ here.) Thus $\mathrm{Opt}(\mathscr{P}) \geq \beta - O(\delta)$.

(b) Given an assignment $F = (f_v)_{v \in V}$ for $\mathscr{P}$, introduce for each $u \in V$ the function $g_u : \{-1,1\}^q \to [-1,1]$ defined by $g(x) = \mathbf{E}_{\boldsymbol{v}}[f_{\boldsymbol{v}}^{\boldsymbol{\pi}}(x)]$, where $\boldsymbol{v}$ is a random neighbor of $\boldsymbol{u}$ in $\mathscr{G}$ and $\boldsymbol{\pi}$ is the associated constraint's permutation. Show that $\mathrm{Val}_{\mathscr{P}}(F) = \mathbf{E}_{\boldsymbol{u} \in V}[\mathrm{Val}_T(g_{\boldsymbol{u}})]$ (using the definition from Exercise 7.22).

(c) Fix an $\epsilon > 0$ and suppose that $\mathrm{Val}_{\mathscr{P}}(F) \geq s + 2\lambda(\epsilon)$, where λ is the "rejection rate" associated with T. Show that for at least a $\lambda(\epsilon)$-fraction of vertices $u \in V$, the set $\mathrm{NbrNotable}_u = \{i \in [q] : \mathbf{Inf}_i^{(1-\epsilon)}[g_u] > \epsilon\}$ is nonempty.

(d) Show that for any $u \in V$ and $i \in [q]$ we have $\mathbf{E}[\mathbf{Inf}^{(1-\epsilon)}_{\boldsymbol{\pi}^{-1}(i)}[f_{\boldsymbol{v}}]] \geq \mathbf{Inf}_i^{(1-\epsilon)}[g_u]$, where $\boldsymbol{v}$ is a random neighbor of u and $\boldsymbol{\pi}$ is the associated constraint's permutation. (Hint: Exercise 2.48.)

(e) For $v \in V$, define also the set $\mathrm{Notable}_u = \{i \in [q] : \mathbf{Inf}_i^{(1-\epsilon)}[f_v] \geq \epsilon/2\}$. Show that if $i \in \mathrm{NbrNotable}_u$, then $\mathbf{Pr}_{\boldsymbol{v}}[\boldsymbol{\pi}^{-1}(i) \in \mathrm{Notable}_{\boldsymbol{v}}] \geq \epsilon/2$, where $\boldsymbol{v}$ and $\boldsymbol{\pi}$ are as in the previous part.

(f) Show that for every $u \in V$ we have $|\mathrm{Notable}_u \cup \mathrm{NbrNotable}_u| \leq O(1/\epsilon^2)$. (Hint: Proposition 2.54.)

(g) Consider the following randomized assignment for $\mathscr{G}$ (see Exericse 7.20): for each $u \in V$, give it the uniform distribution on $\mathrm{Notable}_u \cup \mathrm{NbrNotable}_u$ (if this set is nonempty; otherwise, give it an arbitrary labeling). Show that this randomized assignment has value $\Omega(\lambda(\epsilon)\epsilon^5)$.

(h) Conclude Theorem 7.40, where "UG-hard" means "NP-hard assuming the Unique Games Conjecture".

7.31 Technically, Exercise 7.30 has a small bug: Since a Dictator-vs.-No-Notables test using predicate set Ψ is allowed to use duplicate query strings in its predicates (see Remark 7.38), the reduction in the previous exercise does not necessarily output instances of Max-CSP(Ψ) because our definition of CSPs requires that each scope consist of distinct variables. In this exercise you will correct this bug. Let $M \in \mathbb{N}^+$ and suppose we modify the algorithm R from Exercise 7.30 to a new algorithm R', producing an instance $\mathscr{P}'$ with variable set $V \times [M] \times \{-1,1\}^q$. We now think of assignments to $\mathscr{P}'$ as M-tuples of functions $f_v^1,\dots,f_v^M$, one tuple for each $v \in V$. Further, thinking of $\mathscr{P}$ as a function tester, we have $\mathscr{P}'$ act as follows: Whenever $\mathscr{P}$ is about to query $f_v(x)$, we have $\mathscr{P}'$ instead query $f_v^{\boldsymbol{j}}(x)$ for a uniformly random $\boldsymbol{j} \in [M]$.

(*a*) Show that $\mathrm{Opt}(\mathscr{P}) = \mathrm{Opt}(\mathscr{P}')$.

(*b*) Show that if we delete all constraints in $\mathscr{P}'$ for which the scope contains duplicates, then $\mathrm{Opt}(\mathscr{P}')$ changes by at most $1/M$.

(*c*) Show that the deleted version of $\mathscr{P}'$ is a genuine instance of Max-CSP(Ψ). Since the constant $1/M$ can be arbitrarily small, this corrects the bug in Exercise 7.30's proof of Theorem 7.40.

Notes. The study of property testing was initiated by Rubinfeld and Sudan [**RS96**] and significantly expanded by Goldreich, Goldwasser, and Ron [**GGR98**]; the stricter notion of local testability was introduced (in the context of error-correcting codes) by Friedl and Sudan [**FS95**]. The first local tester for dictatorship was given by Bellare, Goldreich, and Sudan [**BGS95, BGS98**] (as in Exercise 7.8); it was later rediscovered by Parnas, Ron, and Samorodnitsky [**PRS01, PRS02**]. The relevance of Arrow's Theorem to testing dictatorship was pointed out by Kalai [**Kal02**].

The idea of assisting testers by providing proofs grew out of complexity-theoretic research on interactive proofs and PCPs; see the early work Ergün, Kumar, and Rubinfeld [**EKR99**] and the references therein. The specific definition of PCPPs was introduced independently by Ben-Sasson, Goldreich, Harsha, Sudan, and Vadhan [**BSGH$^+$04**] and by Dinur and Reingold [**DR04**] in 2004. Both of these works obtained the PCPP Theorem, relying on the fact that previous literature essentially already gave PCPP reductions of exponential (or greater) proof length: Ben-Sasson et al. [**BSGH$^+$04**] observed that Theorem 7.19 can be obtained from Arora et. al. [**ALM$^+$98**] (their proof is Exercise 7.18), while Dinur and Reingold [**DR04**] pointed out that the slightly easier Theorem 7.18 can be extracted from the work of Bellare, Goldreich, and Sudan [**BGS98**]. The proof we gave for Theorem 7.16 is inspired by the presentation in Dinur [**Din07**].

The PCP Theorem and its stronger forms (the PCPP Theorem and Theorem 7.20) have a somewhat remarkable consequence. Suppose a researcher claims to prove a famous mathematical conjecture, say, "P ≠ NP". To ensure maximum confidence in correctness, a journal might request the researcher submit a formalized proof, suitable for a mechanical proof-checking system. If the submitted formalized proof w is a Boolean string of length n, the proof-checker will be implementable by a circuit C of size $O(n)$. Notice that the string property $\mathscr{C}$ decided by C is nonempty if and only if there exists a (length-n) proof of $\mathsf{P} \neq \mathsf{NP}$. Suppose the journal applies Theorem 7.20 to C and requires the researcher submit the additional proof Π of length $n \cdot \mathrm{polylog}(n)$. Now the journal can run a rather amazing testing algorithm, which reads just 3 bits of the submitted proof (w, Π). If the researcher's proof of $\mathsf{P} \neq \mathsf{NP}$ is correct then the test will accept with probability 1. On the other hand, if the test accepts with probability at least $1-\gamma$ (where γ is the rejection rate in Theorem 7.20), then w must be 1-close to the set of strings accepted by C. This doesn't necessarily mean that w is a correct proof of $\mathsf{P} \neq \mathsf{NP}$ – but it does mean that $\mathscr{C}$ is nonempty, and hence a correct proof of $\mathsf{P} \neq \mathsf{NP}$ exists! By querying a larger constant number of bits from (w, Π) as in Exercise 7.1, say, $\lceil 30/\gamma \rceil$ bits, the journal can become 99.99% convinced that indeed $\mathsf{P} \neq \mathsf{NP}$.

CSPs are very widely studied in computer science; it is impossible to survey the topic here. In the case of Boolean CSPs various monographs [**CKS01, KSTW01**] contain useful background regarding complexity theory and approximation algorithms. The notion of approximation algorithms and the derandomized $(\frac{7}{8}, 1)$-approximation algorithm for Max-E3-Sat (Proposition 7.36, Exercise 7.21) are due to Johnson [**Joh74**]. Incidentally, there is also an efficient $(\frac{7}{8}, 1)$-approximation algorithm for Max-3-Sat [**KZ97**], but both the algorithm and its analysis are extremely difficult, the latter requiring computer assistance [**Zwi02**].

Håstad's hardness theorems appeared in 2001 [**Hås01b**], building on earlier work [**Hås96, Hås99**]. Håstad [**Hås01b**] also proved NP-hardness of $(\frac{1}{p} + \delta, 1-\delta)$-approximating Max-E3-Lin(mod p) (for p prime) and of $(\frac{7}{8}, 1)$-approximating Max-CSP($\{\mathrm{NAE}_4\}$), both of which are optimal. Using tools due to Trevisan et al. [**TSSW00**], Håstad also showed NP-hardness of $(\frac{11}{16} + \delta, \frac{3}{4})$-approximating Max-Cut, which is still the best known such result. The best known inapproximability result for Unique-Games(q) is NP-hardness of $(\frac{3}{8} + q^{-\Theta(1)}, \frac{1}{2})$-approximation [**OW12**]. Khot's influential Unique Games Conjecture dates from 2002 [**Kho02**]; the peculiar name has its origins in a work of Feige and Lovász [**FL92**]. The generic Theorem 7.40, giving UG-hardness from Dictator-vs.-No-Notables tests, is essentially from Khot et al. [**KKMO07**]; the first explicit proof appearing in print may be due to Austrin [**Aus08**]. (We remark that the terminology "Dictator-vs.-No-Notables test" is not standard.) If one is willing to assume the Unique Games Conjecture, there is an almost-complete theory of optimal inapproximability due to Raghavendra [**Rag09**]. Many more inapproximability results, with and without the Unique Games Conjecture, are known; for some surveys, see those of Khot [**Kho05, Kho10a, Kho10b**].

As mentioned, Exercise 7.8 is due to Bellare, Goldreich, and Sudan [**BGS95**] and to Parnas, Ron, and Samorodnitsky [**PRS01**]. The technique described in

Exercise 7.21 is known as the Method of Conditional Expectations. The trick in Exercise 7.23 is closely related to the notion of "folding" from the theory of PCPs. The bug described in Exercise 7.31 is rarely addressed in the literature; the trick used to overcome it appears in, e.g., Arora et al. [**ABH$^+$05**].

Chapter 8

Generalized domains

So far we have studied functions $f : \{0,1\}^n \to \mathbb{R}$. What about, say, $f : \{0,1,2\}^n \to \mathbb{R}$? In fact, very little of what we've done so far depends on the domain being $\{0,1\}^n$; what it has mostly depended on is our viewing the domain as a *product probability distribution*. Indeed, much of analysis of Boolean functions carries over to the case of functions $f : \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ where the domain has a product probability distribution $\pi_1 \otimes \cdots \otimes \pi_n$. There are two main exceptions: the "derivative" operator D_i does not generalize to the case when $|\Omega_i| > 2$ (though the Laplacian operator L_i does), and the important notion of hypercontractivity (introduced in Chapter 9) depends strongly on the probability distributions π_i.

In this chapter we focus on the case where all the Ω_i's are the same, as are the π_i's. (This is just to save on notation; it will be clear that everything we do holds in the more general setting.) Important classic cases include functions on the *p-biased hypercube* (Section 8.4) and functions on abelian groups (Section 8.5). For the issue of generalizing the *range* of functions – e.g., studying functions $f : \{0,1,2\}^n \to \{0,1,2\}$ – see Exercise 8.33.

8.1. Fourier bases for product spaces

We will now begin to discuss functions on (finite) product probability spaces.

Definition 8.1. Let (Ω, π) be a finite probability space with $|\Omega| \geq 2$ and assume π has full support. For $n \in \mathbb{N}^+$ we write $L^2(\Omega^n, \pi^{\otimes n})$ for the (real) inner product space of functions $f : \Omega^n \to \mathbb{R}$, with inner product

$$\langle f, g \rangle = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \pi^{\otimes n}}[f(\boldsymbol{x})g(\boldsymbol{x})].$$

Here $\pi^{\otimes n}$ denotes the product probability distribution on Ω^n.

Example 8.2. A simple example to keep in mind is $\Omega = \{a,b,c\}$ with $\pi(a) = \pi(b) = \pi(c) = 1/3$. Here a, b, and c are simply abstract set elements.

We can (and will) generalize to nondiscrete probability spaces, and to complex inner product spaces. However, we will keep to the above definition for now.

Notation 8.3. We will write $\pi_{1/2}$ for the uniform probability distribution on $\{-1,1\}$. Thus so far in this book we have been studying functions in $L^2(\{-1,1\}^n, \pi_{1/2}^{\otimes n})$. For simplicity, we will write this as $L^2(\{-1,1\}^n)$.

Notation 8.4. Much of the notation we used for $L^2(\{-1,1\}^n)$ extends naturally to the case of $L^2(\Omega^n, \pi^{\otimes n})$: e.g., $\|f\|_p = \mathbf{E}_{\boldsymbol{x}\sim\pi^{\otimes n}}[|f(\boldsymbol{x})|^p]^{1/p}$, or the restriction notation from Chapter 3.3.

As we described in Chapter 1.4, the essence of Boolean Fourier analysis is in deriving combinatorial properties of a Boolean function $f : \{-1,1\}^n \to \mathbb{R}$ from its coefficients over a particular basis of $L^2(\{-1,1\}^n)$, the basis of parity functions. We would like to achieve the same thing more generally for functions in $L^2(\Omega^n, \pi^{\otimes n})$. We begin by considering vector space bases more generally.

Definition 8.5. Let $|\Omega| = m$. The *indicator basis* (or *standard basis*) for $L^2(\Omega, \pi)$ is just the set of m indicator functions $(1_x)_{x\in\Omega}$, where

$$1_x(y) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{if } y \neq x. \end{cases}$$

Fact 8.6. *The indicator basis is indeed a basis for* $L^2(\Omega,\pi)$ *since the functions* $(1_x)_{x\in\Omega}$ *are nonzero, spanning, and orthogonal. Hence* $\dim(L^2(\Omega,\pi)) = m$.

We will usually fix Ω and π and then consider $L^2(\Omega^n, \pi^{\otimes n})$ for $n \in \mathbb{N}^+$. Applying the above definition gives us an indicator basis $(1_x)_{x\in\Omega^n}$ for the m^n-dimensional space $L^2(\Omega^n, \pi^{\otimes n})$. The representation of $f \in L^2(\Omega,\pi)$ in this basis is just $f = \sum_{x\in\Omega} f(x)1_x$. This is not very interesting; the coefficients are just the values of f so they don't tell us anything new about the function. We would like a different basis that will generate useful "Fourier formulas" as in Chapter 1.4.

For inspiration, let's look critically at the familiar case of $L^2(\{-1,1\}^n)$. Here we used the basis of all parity functions, $\chi_S(x) = \prod_{i\in S} x_i$. It will be helpful to think of the basis function $\chi_S : \{-1,1\}^n \to \mathbb{R}$ as follows: Identify S with its 0-1 indicator vector and write

$$\chi_S(x) = \prod_{i=1}^{n} \phi_{S_i}(x_i), \qquad \text{where} \quad \phi_0 \equiv 1, \quad \phi_1 = id.$$

(Here id is just the identity map $id(b) = b$.) We will identify three properties of this basis which we'd like to generalize.

First, the parity basis is a *product basis*. We can break down its "product structure" as follows: For each coordinate $i \in [n]$ of the product domain $\{-1,1\}^n$, the set $\{1, id\}$ is a basis for the 2-dimensional space $L^2(\{-1,1\}, \pi_{1/2})$. We then get a basis for the 2^n-dimensional product space $L^2(\{-1,1\}^n)$ by taking all possible n-fold products. More generally, suppose we are given an inner product space $L^2(\Omega, \pi)$ with $|\Omega| = m$. Let $\phi_0, \dots, \phi_{m-1}$ be any basis for this space. Then the set of all products $\phi_{i_1}\phi_{i_2}\cdots\phi_{i_n}$ $(0 \le i_j < m)$ forms a basis for the space $L^2(\Omega^n, \pi^{\otimes n})$.

Second, it is convenient that the parity basis is *orthonormal*. We will later check that if a basis $\phi_0, \dots, \phi_{m-1}$ for $L^2(\Omega, \pi)$ is orthonormal, then so too is the associated product basis for $L^2(\Omega^n, \pi^{\otimes n})$. This relies on the fact that $\pi^{\otimes n}$ is the product distribution. For example, the parity basis for $L^2(\{-1,1\}^n)$ is orthonormal because the basis $\{1, id\}$ for $L^2(\{-1,1\}, \pi_{1/2})$ is orthonormal: $\mathbf{E}[1^2] = \mathbf{E}[\boldsymbol{x}_i^2] = 1$, $\mathbf{E}[1 \cdot \boldsymbol{x}_i] = 0$. Orthonormality is the property that makes Parseval's Theorem hold; in the general context, this means that if $f \in L^2(\Omega, \pi)$ has the representation $\sum_{i=0}^{m-1} c_i \phi_i$ then $\mathbf{E}[f^2] = \sum_{i=0}^{m-1} c_i^2$.

Finally, the parity basis contains the constant function 1. This fact leads to several of our pleasant Fourier formulas. In particular, when you take an orthonormal basis $\phi_0, \dots, \phi_{m-1}$ for $L^2(\Omega, \pi)$ which has $\phi_0 \equiv 1$, then $0 = \langle \phi_0, \phi_i \rangle = \mathbf{E}_{\boldsymbol{x} \sim \pi}[\phi_i(\boldsymbol{x})]$ for all $i > 0$. Hence if $f \in L^2(\Omega, \pi)$ has the expansion $f = \sum_{i=0}^{m-1} c_i \phi_i$, then $\mathbf{E}[f] = c_0$ and $\mathbf{Var}[f] = \sum_{i>0} c_i^2$.

We encapsulate the second and third properties with a definition:

Definition 8.7. A *Fourier basis* for an inner product space $L^2(\Omega, \pi)$ is an orthonormal basis $\phi_0, \dots, \phi_{m-1}$ with $\phi_0 \equiv 1$.

Example 8.8. For each $n \in \mathbb{N}^+$, the 2^n parity functions $(\chi_S)_{S \subseteq [n]}$ form a Fourier basis for $L^2(\{-1,1\}^n, \pi_{1/2}^{\otimes n})$.

Remark 8.9. A Fourier basis for $L^2(\Omega, \pi)$ always exists because you can extend the set $\{1\}$ to a basis and then perform the Gram–Schmidt process. On the other hand, Fourier bases are not unique. Even in the case of $L^2(\{-1,1\}, \pi_{1/2})$ there are two possibilities: the basis $\{1, id\}$ and the basis $\{1, -id\}$.

Example 8.10. In the case of $\Omega = \{a, b, c\}$ with $\pi(a) = \pi(b) = \pi(c) = 1/3$, one possible Fourier basis (see Exercise 8.4) is

$$\phi_0 \equiv 1, \quad \begin{array}{ll} \phi_1(a) = +\sqrt{2} & \phi_2(a) = 0 \\ \phi_1(b) = -\sqrt{2}/2 & \phi_2(b) = +\sqrt{6}/2, \\ \phi_1(c) = -\sqrt{2}/2, & \phi_2(c) = -\sqrt{6}/2. \end{array}$$

As mentioned, given a Fourier basis for $L^2(\Omega,\pi)$ you can construct a Fourier basis for any $L^2(\Omega^n,\pi^{\otimes n})$ by "taking all n-fold products". To make this precise we need some notation.

Definition 8.11. An n-dimensional *multi-index* is a tuple $\alpha \in \mathbb{N}^n$. We write

$$\operatorname{supp}(\alpha) = \{i : \alpha_i \neq 0\}, \quad \#\alpha = |\operatorname{supp}(\alpha)|, \quad |\alpha| = \sum_{i=1}^{n} \alpha_i.$$

We may write $\alpha \in \mathbb{N}^n_{<m}$ when we want to emphasize that each $\alpha_i \in \{0,1,\dots,m-1\}$.

Definition 8.12. Given functions $\phi_0,\dots,\phi_{m-1} \in L^2(\Omega,\pi)$ and a multi-index $\alpha \in \mathbb{N}^n_{<m}$, we define $\phi_\alpha \in L^2(\Omega^n,\pi^{\otimes n})$ by

$$\phi_\alpha(x) = \prod_{i=1}^{n} \phi_{\alpha_i}(x_i).$$

Now we can show that products of Fourier bases are Fourier bases.

Proposition 8.13. *Let $\phi_0,\dots,\phi_{m-1}$ be a Fourier basis for $L^2(\Omega,\pi)$. Then the collection $(\phi_\alpha)_{\alpha\in\mathbb{N}^n_{<m}}$ is a Fourier basis for $L^2(\Omega^n,\pi^{\otimes n})$ (with the understanding that $\alpha = (0,0,\dots,0)$ indexes the constant function 1).*

Proof. First we check orthonormality. For any multi-indices $\alpha,\beta \in \mathbb{N}^n_{<m}$ we have

$$\begin{aligned}
\langle \phi_\alpha,\phi_\beta\rangle &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\phi_\alpha(\boldsymbol{x})\cdot\phi_\beta(\boldsymbol{x})] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\Bigl[\prod_{i=1}^{n}\phi_{\alpha_i}(\boldsymbol{x}_i)\cdot\prod_{i=1}^{n}\phi_{\beta_i}(\boldsymbol{x}_i)\Bigr] \\
&= \prod_{i=1}^{n}\mathop{\mathbf{E}}_{\boldsymbol{x}_i\sim\pi}[\phi_{\alpha_i}(\boldsymbol{x}_i)\cdot\phi_{\beta_i}(\boldsymbol{x}_i)] \qquad \text{(since } \pi^{\otimes n} \text{ is a product distribution)} \\
&= \prod_{i=1}^{n}\mathbf{1}_{\{\alpha_i=\beta_i\}} \qquad \text{(since } \{\phi_0,\dots,\phi_{m-1}\} \text{ is orthonormal)} \\
&= \mathbf{1}_{\{\alpha=\beta\}}.
\end{aligned}$$

This confirms that the collection $(\phi_\alpha)_{\alpha\in\mathbb{N}^n_{<m}}$ is orthonormal, and consequently linearly independent. It is therefore also a basis because it has cardinality m^n, which we know is the dimension of $L^2(\Omega^n,\pi^{\otimes n})$ (see Fact 8.6). □

Given a product Fourier basis as in Proposition 8.13, we can express any $f \in L^2(\Omega^n,\pi^{\otimes n})$ as a linear combination of basis functions. We will write $\widehat{f}(\alpha)$ for the "Fourier coefficient" on ϕ_α in this expression.

Definition 8.14. Having *fixed* a Fourier basis $\phi_0, \dots, \phi_{m-1}$ for $L^2(\Omega, \pi)$, every $f \in L^2(\Omega^n, \pi^{\otimes n})$ is uniquely expressible as

$$f = \sum_{\alpha \in \mathbb{N}^n_{<m}} \widehat{f}(\alpha)\phi_\alpha.$$

This is the *Fourier expansion* of f with respect to the basis. The real number $\widehat{f}(\alpha)$ is called the *Fourier coefficient of f on α* and it satisfies

$$\widehat{f}(\alpha) = \langle f, \phi_\alpha \rangle.$$

Example 8.15. Fix the Fourier basis as in Example 8.10. Let $f : \{a,b,c\}^2 \to \{0,1\}$ be the function which is 1 if and only if both inputs are c. Then you can check (Exercise 8.5) that

$$f = \tfrac{1}{9} - \tfrac{\sqrt{2}}{18}\phi_{(1,0)} - \tfrac{\sqrt{6}}{18}\phi_{(2,0)} - \tfrac{\sqrt{2}}{18}\phi_{(0,1)} - \tfrac{\sqrt{6}}{18}\phi_{(0,2)} + \tfrac{1}{18}\phi_{(1,1)} + \tfrac{\sqrt{12}}{36}\phi_{(2,1)} + \tfrac{\sqrt{12}}{36}\phi_{(1,2)} + \tfrac{1}{6}\phi_{(2,2)}.$$

The notation $\widehat{f}(\alpha)$ may seem poorly chosen because it doesn't show the dependence on the basis. However, the Fourier formulas we develop in the next section will have the property that *they are the same for every product Fourier basis*. We will show a basis-independent way of developing the formulas in Section 8.3.

8.2. Generalized Fourier formulas

In this section we will revisit a number of combinatorial/probabilistic notions and show that for functions $f \in L^2(\Omega^n, \pi^{\otimes n})$, these notions have familiar Fourier formulas that don't depend on the Fourier basis.

The orthonormality of Fourier bases gives us some formulas almost immediately:

Proposition 8.16. *Let $f, g \in L^2(\Omega^n, \pi^{\otimes n})$. Then for any fixed product Fourier basis, the following formulas hold:*

$$\begin{aligned}
\mathbf{E}[f] &= \widehat{f}(0) \\
\mathbf{E}[f^2] &= \sum_{\alpha \in \mathbb{N}^n_{<m}} \widehat{f}(\alpha)^2 && \text{(Parseval)} \\
\mathbf{Var}[f] &= \sum_{\alpha \neq 0} \widehat{f}(\alpha)^2 \\
\langle f, g \rangle &= \sum_{\alpha \in \mathbb{N}^n_{<m}} \widehat{f}(\alpha)\widehat{g}(\alpha) && \text{(Plancherel)} \\
\mathbf{Cov}[f,g] &= \sum_{\alpha \neq 0} \widehat{f}(\alpha)\widehat{g}(\alpha).
\end{aligned}$$

Proof. We verify Plancherel's Theorem, from which the other identities follow (Exercise 8.6):

$$\begin{aligned}\langle f,g\rangle &= \Big\langle \sum_{\alpha\in\mathbb{N}^n_{<m}} \widehat{f}(\alpha)\phi_\alpha, \sum_{\beta\in\mathbb{N}^n_{<m}} \widehat{g}(\beta)\phi_\beta\Big\rangle\\ &= \sum_{\alpha,\beta\in\mathbb{N}^n_{<m}} \widehat{f}(\alpha)\widehat{g}(\beta)\langle\phi_\alpha,\phi_\beta\rangle\\ &= \sum_{\alpha\in\mathbb{N}^n_{<m}} \widehat{f}(\alpha)\widehat{g}(\alpha)\end{aligned}$$

by orthonormality of $(\phi_\alpha)_{\alpha\in\mathbb{N}^n_{<m}}$. □

We now give a definition that will be the key for developing basis-independent Fourier expansions. In the case of $L^2(\{-1,1\})$ this definition appeared already in Exercise 3.28.

Definition 8.17. Let $J\subseteq[n]$ and write $\overline{J}=[n]\setminus J$. Given $f\in L^2(\Omega^n,\pi^{\otimes n})$, the *projection of f on coordinates J* is the function $f^{\subseteq J}\in L^2(\Omega^n,\pi^{\otimes n})$ defined by

$$f^{\subseteq J}(x)=\mathop{\mathbf{E}}_{\boldsymbol{x}'\sim\pi^{\otimes\overline{J}}}[f(x_J,\boldsymbol{x}')],$$

where $x_J\in\Omega^J$ denotes the values of x in the J-coordinates. In other words, $f^{\subseteq J}(x)$ is the expectation of f when the $\overline{J}$-coordinates of x are rerandomized. Note that we take $f^{\subseteq J}$ to have Ω^n as its domain, even though it only depends on the coordinates in J.

Forming $f^{\subseteq J}$ is indeed the application of a projection linear operator to f, namely the *expectation over $\overline{J}$ operator*, $\mathrm{E}_{\overline{J}}$. We take this as the definition of the operator: $\mathrm{E}_{\overline{J}}f=f^{\subseteq J}$. When $\overline{J}=\{i\}$ is a singleton we write simply E_i.

Remark 8.18. This definition of E_i is consistent with Definition 2.23. You are asked to verify that $\mathrm{E}_{\overline{J}}$ is indeed a projection, self-adjoint linear operator in Exercise 8.7.

Proposition 8.19. *Let $J\subseteq[n]$ and $f\in L^2(\Omega^n,\pi^{\otimes n})$. Then for any fixed product Fourier basis,*

$$f^{\subseteq J}=\sum_{\substack{\alpha\in\mathbb{N}^n_{<m}\\ \mathrm{supp}(\alpha)\subseteq J}}\widehat{f}(\alpha)\phi_\alpha.$$

Proof. Since $\mathrm{E}_{\overline{J}}$ is a linear operator, it suffices to verify for all α that

$$\phi_\alpha^{\subseteq J}=\begin{cases}\phi_\alpha & \text{if } \mathrm{supp}(\alpha)\subseteq J,\\ 0 & \text{otherwise.}\end{cases}$$

If $\mathrm{supp}(\alpha)\subseteq J$, then ϕ_α does not depend on the coordinates outside J; hence indeed $\phi_\alpha^{\subseteq J}=\phi_\alpha$. So suppose $\mathrm{supp}(\alpha)\not\subseteq J$. Since $\phi_\alpha(x)=\left(\prod_{i\in J}\phi_{\alpha_i}(x_i)\right)\left(\prod_{i\in\overline{J}}\phi_{\alpha_i}(x_i)\right)$, we can write $\phi_\alpha=\phi_{\alpha_J}\cdot\phi_{\alpha_{\overline{J}}}$, where ϕ_{α_J} depends only on the coordinates in J,

$\phi_{\alpha_{\overline{J}}}$ depends only on the coordinates in $\overline{J}$, and $\mathbf{E}[\phi_{\alpha_{\overline{J}}}] = 0$ precisely because $\mathrm{supp}(\alpha) \not\subseteq J$. Thus for every $x \in \Omega^n$,

$$\phi_\alpha^{\subseteq J}(x) = \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim \pi^{\otimes \overline{J}}}[\phi_{\alpha_J}(x_J)\phi_{\alpha_{\overline{J}}}(\boldsymbol{x}')] = \phi_{\alpha_J}(x_J) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim \pi^{\otimes \overline{J}}}[\phi_{\alpha_{\overline{J}}}(\boldsymbol{x}')] = 0$$

as needed. □

Corollary 8.20. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ and fix a product Fourier basis. If f depends only on the coordinates in $J \subseteq [n]$ then $\widehat{f}(\alpha) = 0$ whenever* $\mathrm{supp}(\alpha) \not\subseteq J$.

Proof. This follows from Proposition 8.19 because $f = f^{\subseteq J}$. □

Corollary 8.21. *Let $i \in [n]$ and $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then for any fixed product Fourier basis,*

$$\mathrm{E}_i f = \sum_{\alpha : \alpha_i = 0} \widehat{f}(\alpha)\phi_\alpha.$$

Let us now define influences for functions $f \in L^2(\Omega^n, \pi^{\otimes n})$. In the case of $\Omega = \{-1, 1\}$, our definition of $\mathbf{Inf}_i[f]$ from Chapter 2.2 was $\mathbf{E}[\mathrm{D}_i f]$. However, the notion of a derivative operator does not make sense for more general domains Ω. In fact, even in the case of $\Omega = \{-1, 1\}$ it isn't a basis-invariant notion: the choice of $\frac{f(x^{(i \mapsto 1)}) - f(x^{(i \mapsto -1)})}{2}$ rather than $\frac{f(x^{(i \mapsto -1)}) - f(x^{(i \mapsto 1)})}{2}$ is inherently arbitrary. Instead we can fall back on the *Laplacian* operators, and take the identity $\mathbf{Inf}_i[f] = \langle f, \mathrm{L}_i f \rangle$ from Proposition 2.26 as a definition.

Definition 8.22. Let $i \in [n]$ and $f \in L^2(\Omega^n, \pi^{\otimes n})$. The *ith coordinate Laplacian operator* L_i is the self-adjoint, projection linear operator defined by

$$\mathrm{L}_i f = f - \mathrm{E}_i f.$$

The *influence of coordinate i on f* is defined to be

$$\mathbf{Inf}_i[f] = \langle f, \mathrm{L}_i f \rangle = \langle \mathrm{L}_i f, \mathrm{L}_i f \rangle.$$

The *total influence* of f is defined to be $\mathbf{I}[f] = \sum_{i=1}^n \mathbf{Inf}_i[f]$.

You can think of $\mathrm{L}_i f$ as "the part of f which depends on the ith coordinate".

Proposition 8.23. *Let $i \in [n]$ and $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then for any fixed product Fourier basis,*

$$\mathrm{L}_i f = \sum_{\alpha : \alpha_i \neq 0} \widehat{f}(\alpha)\phi_\alpha, \quad \mathbf{Inf}_i[f] = \sum_{\alpha : \alpha_i \neq 0} \widehat{f}(\alpha)^2, \quad \mathbf{I}[f] = \sum_\alpha \#\alpha \cdot \widehat{f}(\alpha)^2,$$

Proof. The first formula is immediate from Corollary 8.21, the second from Plancherel, and the third from summing over i. □

Exercise 8.9 asks you to verify the following formulas (cf. Exercise 2.21), which are often useful for computations:

Proposition 8.24. *Let $i \in [n]$ and $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then*

$$\mathbf{Inf}_i[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}} [\mathop{\mathbf{Var}}_{\boldsymbol{x}'_i\sim\pi}[f(\boldsymbol{x}_1,\dots,\boldsymbol{x}_{i-1},\boldsymbol{x}'_i,\boldsymbol{x}_{i+1},\dots,\boldsymbol{x}_n)]].$$

If furthermore f's range is $\{-1,1\}$, then

$$\mathbf{Inf}_i[f] = \mathbf{E}[|\mathrm{L}_i f|] = 2 \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim\pi^{\otimes n}\\ \boldsymbol{x}'_i\sim\pi}}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}_1,\dots,\boldsymbol{x}_{i-1},\boldsymbol{x}'_i,\boldsymbol{x}_{i+1},\dots,\boldsymbol{x}_n)].$$

Example 8.25. Let's continue Example 8.15, in which $\{a,b,c\}$ has the uniform distribution and $f : \{a,b,c\}^2 \to \{0,1\}$ is 1 if and only if both inputs are c. We compute $\mathbf{Inf}_1[f]$ two ways. Using Proposition 8.24 we have $\mathbf{Var}[f(\boldsymbol{x}_1,a)] = \mathbf{Var}[f(\boldsymbol{x}_1,b)] = 0$ and $\mathbf{Var}[f(\boldsymbol{x}_1,c)] = \frac{1}{3}\cdot\frac{2}{3} = \frac{2}{9}$ (because $f(\boldsymbol{x}_1,c)$ is Bernoulli with parameter $\frac{1}{3}$); thus $\mathbf{Inf}_1[f] = \frac{1}{3}\cdot\frac{2}{9} = \frac{2}{27}$. Alternatively, using the formula from Proposition 8.23 as well as the Fourier expansion from Example 8.15, we can compute $\mathbf{Inf}_1[f] = (-\frac{\sqrt{2}}{18})^2 + (-\frac{\sqrt{6}}{18})^2 + (\frac{1}{18})^2 + (\frac{\sqrt{12}}{36})^2 + (\frac{\sqrt{12}}{36})^2 + (\frac{1}{6})^2 = \frac{2}{27}$.

Next, we straightforwardly extend our definitions of the noise operator and noise stability to general product spaces.

Definition 8.26. Fix a finite product probability space $(\Omega^n, \pi^{\otimes n})$. For $\rho \in [0,1]$ and $x \in \Omega^n$ we write $\boldsymbol{y} \sim N_\rho(x)$ to denote that $\boldsymbol{y} \in \Omega^n$ is randomly chosen as follows: For each $i \in [n]$ independently,

$$\boldsymbol{y}_i = \begin{cases} x_i & \text{with probability } \rho,\\ \text{drawn from } \pi & \text{with probability } 1-\rho.\end{cases}$$

If $\boldsymbol{x} \sim \pi^{\otimes n}$ and $\boldsymbol{y} \sim N_\rho(\boldsymbol{x})$, we say that $(\boldsymbol{x},\boldsymbol{y})$ is a *ρ-correlated pair under $\pi^{\otimes n}$*. (This definition is symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$.)

Definition 8.27. For a fixed space $L^2(\Omega^n, \pi^{\otimes n})$ and $\rho \in [0,1]$, the *noise operator with parameter ρ* is the linear operator T_ρ on functions $f \in L^2(\Omega^n, \pi^{\otimes n})$ defined by

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

The *noise stability of f at ρ* is

$$\mathbf{Stab}_\rho[f] = \langle f, \mathrm{T}_\rho f\rangle = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\ \rho\text{-correlated}\\ \text{under } \pi^{\otimes n}}}[f(\boldsymbol{x})f(\boldsymbol{y})].$$

Proposition 8.28. *Let $\rho \in [0,1]$ and let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then for any fixed product Fourier basis,*

$$\mathrm{T}_\rho f = \sum_{\alpha\in\mathbb{N}^n_{<m}} \rho^{\#\alpha}\widehat{f}(\alpha)\phi_\alpha, \qquad \mathbf{Stab}_\rho[f] = \sum_{\alpha\in\mathbb{N}^n_{<m}} \rho^{\#\alpha}\widehat{f}(\alpha)^2.$$

Proof. Let $\boldsymbol{J}$ denote a ρ-random subset of $[n]$; i.e., $\boldsymbol{J}$ is formed by including each $i \in [n]$ independently with probability ρ. Then by definition $T_\rho f(x) = \mathbf{E}_{\boldsymbol{J}}[f^{\subseteq \boldsymbol{J}}(x)]$, and so from Proposition 8.19 we get

$$T_\rho f(x) = \mathop{\mathbf{E}}_{\boldsymbol{J}}[f^{\subseteq \boldsymbol{J}}(x)] = \mathop{\mathbf{E}}_{\boldsymbol{J}}\Big[\sum_{\substack{\alpha \in \mathbb{N}^n_{<m} \\ \mathrm{supp}(\alpha) \subseteq \boldsymbol{J}}} \widehat{f}(\alpha)\phi_\alpha(x)\Big] = \sum_{\alpha \in \mathbb{N}^n_{<m}} \rho^{\#\alpha} \widehat{f}(\alpha)\phi_\alpha(x),$$

since for a fixed α, the probability of $\mathrm{supp}(\alpha) \subseteq \boldsymbol{J}$ is $\rho^{\#\alpha}$. The formula for $\mathbf{Stab}_\rho[f]$ now follows from Plancherel. □

Remark 8.29. The first formula in this proposition may be used to extend the definition of $\mathrm{T}_\rho f$ to values of ρ outside $[0,1]$.

We also define ρ-stable influences. The factor of ρ^{-1} in our definition is for consistency with the $L^2(\{-1,1\}^n)$ case.

Definition 8.30. For $f \in L^2(\Omega^n, \pi^{\otimes n})$, $\rho \in (0,1]$, and $i \in [n]$, the *ρ-stable influence* of i on f is

$$\mathbf{Inf}_i^{(\rho)}[f] = \rho^{-1}\mathbf{Stab}_\rho[\mathrm{L}_i f] = \sum_{\alpha : \alpha_i \neq 0} \rho^{\#\alpha - 1}\widehat{f}(\alpha)^2.$$

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

Just as in the case of $L^2(\{-1,1\}^n)$ we can use stable influences to define the "notable" coordinates of a function, of which there is a bounded quantity. A verbatim repetition of the proof of Proposition 2.54 yields the following generalization:

Proposition 8.31. *Suppose $f \in L^2(\Omega^n, \pi^{\otimes n})$ has $\mathbf{Var}[f] \le 1$. Given $0 < \delta < 1$, $0 < \epsilon \le 1$, let $J = \{i \in [n] : \mathbf{Inf}_i^{(1-\delta)}[f] \ge \epsilon\}$. Then $|J| \le \frac{1}{\delta\epsilon}$.*

We end this section by discussing the "degree" of functions on general product spaces. For $f \in L^2(\{-1,1\}^n)$ the Fourier expansion is a real polynomial; this yields an obvious definition for degree. But for general $f \in L^2(\Omega^n, \pi^{\otimes n})$ the domain is just an abstract set so we need to look for a more intrinsic definition. We take our cue from Exercise 1.10(*b*):

Definition 8.32. Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be nonzero. The *degree* of f, written $\deg(f)$, is the least $k \in \mathbb{N}$ such that f is a sum of k-juntas (functions depending on at most k coordinates).

Proposition 8.33. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be nonzero. Then for any fixed product Fourier basis we have $\deg(f) = \max\{\#\alpha : \widehat{f}(\alpha) \neq 0\}$.*

Proof. The inequality $\deg(f) \le \max\{\#\alpha : \widehat{f}(\alpha) \neq 0\}$ is immediate from the Fourier expansion:

$$f = \sum_{\alpha : \widehat{f}(\alpha) \neq 0} \widehat{f}(\alpha)\phi_\alpha$$

and each function $\widehat{f}(\alpha)\phi_\alpha$ depends on at most $\#\alpha$ coordinates. For the reverse inequality, suppose $f = g_1 + \cdots + g_m$ where each g_i depends on at most k coordinates. By Corollary 8.20 each g_i has its Fourier support on functions ϕ_α with $\#\alpha \leq k$. But $\widehat{f}(\alpha) = \widehat{g_1}(\alpha) + \cdots + \widehat{g_m}(\alpha)$, so the same is true of f. □

8.3. Orthogonal decomposition

In this section we describe a basis-free kind of "Fourier expansion" for functions on general product domains. We will refer to it as the *orthogonal decomposition* of $f \in L^2(\Omega^n, \pi^{\otimes n})$, though it goes by several other names in the literature: e.g., *Hoeffding decomposition*, *Efron–Stein decomposition*, or *ANOVA decomposition*. The general idea is to express

$$f = \sum_{S \subseteq [n]} f^{=S} \tag{8.1}$$

where each function $f^{=S} \in L^2(\Omega^n, \pi^{\otimes n})$ gives the "contribution to f coming from coordinates S (but not from any subset of S)".

To make this more precise, let's start with the familiar case of $f : \{-1,1\}^n \to \mathbb{R}$. Here it is possible to define the functions $f^{=S} : \{-1,1\}^n \to \mathbb{R}$ simply by $f^{=S} = \widehat{f}(S)\chi_S$. (Later we will give an equivalent definition that doesn't involve the Fourier basis.) This definition satisfies (8.1) as well as the following two properties:

(1) $f^{=S}$ depends only on the coordinates in S.

(2) If $T \subsetneq S$ and g is a function depending only on the coordinates in T, then $\langle f^{=S}, g\rangle = 0$.

These properties describe what we mean precisely when we say that $f^{=S}$ is the "contribution to f coming from coordinates S (but not from any subset of S)". Furthermore, the decomposition (8.1) is *orthogonal*, meaning $\langle f^{=S}, f^{=T}\rangle = 0$ whenever $S \neq T$.

To make this definition basis-free, recall the "projection of f onto coordinates J", $f^{\subseteq J}$, from Exercise 3.28 and Definition 8.17. You can think of $f^{\subseteq J}$ as the "contribution to f coming from coordinates J (collectively)". It has a probabilistic definition not depending on any basis, and with the definition $f^{=S} = \widehat{f}(S)\chi_S$ we have from Exercise 3.28 or Proposition 8.19 that

$$f^{\subseteq J} = \sum_{S \subseteq J} f^{=S}. \tag{8.2}$$

It is precisely by inverting (8.2) that we can give a basis-free definition of the functions $f^{=S}$.

Let's do this inversion for a general $f \in L^2(\Omega^n, \pi^{\otimes n})$. The projection functions $f^{\subseteq J} \in L^2(\Omega^n, \pi^{\otimes n})$ can be defined as in Definition 8.17. If we want (8.2) to hold for $J = \emptyset$ then we should define

$$f^{=\emptyset} = f^{\subseteq \emptyset}$$

(which is the constant function equal to $\mathbf{E}[f]$). Given this, if we want (8.2) to hold for singleton sets $J = \{j\}$, then we need

$$f^{\subseteq \{j\}} = f^{=\emptyset} + f^{=\{j\}} \quad \iff \quad f^{=\{j\}} = f^{\subseteq \{j\}} - f^{\subseteq \emptyset}.$$

In other words,

$$f^{=\{j\}}(x) = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \pi^{\otimes n}}[f \mid \boldsymbol{x}_j = x_j] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \pi^{\otimes n}}[f(\boldsymbol{x})].$$

Notice this function only depends on the input value x_j; it measures the change in expectation of f if you know the value x_j. Moving on to sets of cardinality 2, if we want (8.2) to hold for $J = \{i, j\}$, then we need

$$\begin{aligned} f^{\subseteq \{i,j\}} &= f^{=\emptyset} + f^{=\{i\}} + f^{=\{j\}} + f^{=\{i,j\}} \\ &= f^{\subseteq \emptyset} + (f^{\subseteq \{i\}} - f^{\subseteq \emptyset}) + (f^{\subseteq \{j\}} - f^{\subseteq \emptyset}) + f^{=\{i,j\}} \end{aligned}$$

and hence

$$f^{=\{i,j\}} = f^{\subseteq \{i,j\}} - f^{\subseteq \{i\}} - f^{\subseteq \{j\}} + f^{\subseteq \emptyset}.$$

It's clear that we can continue this and define all the functions $f^{=S}$ by the principle of inclusion-exclusion. To show this definition leads to an orthogonal decomposition we will need the following lemma:

Lemma 8.34. *Let $f, g \in L^2(\Omega^n, \pi^{\otimes n})$. Assume that f does not depend on any coordinate outside $I \subseteq [n]$, and g does not depend on any coordinate outside $J \subseteq [n]$. Then $\langle f, g \rangle = \langle f^{\subseteq I \cap J}, g^{\subseteq I \cap J} \rangle$.*

Proof. We may assume without loss of generality that $I \cup J = [n]$. Given any $x \in \Omega^n$ we can break it into the parts $(x_{I \cap J}, x_{I \setminus J}, x_{J \setminus I})$. We then have

$$\langle f, g \rangle = \mathop{\mathbf{E}}_{\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{I \setminus J}, \boldsymbol{x}_{J \setminus I}}[f(\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{I \setminus J}) \cdot g(\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{J \setminus I})],$$

where we have abused notation slightly by writing f and g as functions just of the coordinates on which they actually depend. Since $\boldsymbol{x}_{I \setminus J}$ and $\boldsymbol{x}_{J \setminus I}$ are independent, the above equals

$$\mathop{\mathbf{E}}_{\boldsymbol{x}_{I \cap J}}\left[\mathop{\mathbf{E}}_{\boldsymbol{x}_{I \setminus J}}[f(\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{I \setminus J})] \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}_{J \setminus I}}[g(\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{J \setminus I})]\right].$$

But now $\mathbf{E}_{\boldsymbol{x}_{I \setminus J}}[f(\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{I \setminus J})]$ is nothing more than $f^{\subseteq I \cap J}(\boldsymbol{x}_{I \cap J})$, and similarly $\mathbf{E}_{\boldsymbol{x}_{J \setminus I}}[g(\boldsymbol{x}_{I \cap J}, \boldsymbol{x}_{J \setminus I})] = g^{\subseteq I \cap J}(\boldsymbol{x}_{I \cap J})$. Thus the above equals

$$\mathop{\mathbf{E}}_{\boldsymbol{x}_{I \cap J}}[f^{\subseteq I \cap J}(\boldsymbol{x}_{I \cap J}) \cdot g^{\subseteq I \cap J}(\boldsymbol{x}_{I \cap J})] = \langle f^{\subseteq I \cap J}, g^{\subseteq I \cap J} \rangle. \qquad \square$$

We can now give the main theorem on orthogonal decomposition:

Theorem 8.35. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then f has a unique decomposition as*

$$f = \sum_{S \subseteq [n]} f^{=S}$$

where the functions $f^{=S} \in L^2(\Omega^n, \pi^{\otimes n})$ satisfy the following:

(1) *$f^{=S}$ depends only on the coordinates in S.*

(2) *If $T \subsetneq S$ and $g \in L^2(\Omega^n, \pi^{\otimes n})$ depends only on the coordinates in T, then $\langle f^{=S}, g\rangle = 0$.*

This decomposition has the following additional properties:

(3) *Condition (2) additionally holds whenever $S \not\subseteq T$.*

(4) *The decomposition is* orthogonal*: $\langle f^{=S}, f^{=T}\rangle = 0$ for $S \neq T$.*

(5) *$\sum_{S \subseteq T} f^{=S} = f^{\subseteq T}$.*

(6) *For each $S \subseteq [n]$, the mapping $f \mapsto f^{=S}$ is a linear operator.*

Proof. We first show the existence of a decomposition satisfying (1)–(6). We then show uniqueness for decompositions satisfying (1) and (2). As suggested above, for each $S \subseteq [n]$ we define

$$f^{=S} = \sum_{J \subseteq S} (-1)^{|S|-|J|} f^{\subseteq J},$$

where the functions $f^{\subseteq J} \in L^2(\Omega^n, \pi^{\otimes n})$ are as in Definition 8.17. Since each $f^{\subseteq J}$ depends only on the coordinates in J, condition (1) certainly holds. It is also immediate that condition (5) holds by inclusion-exclusion; you are asked to prove this explicitly in Exercise 8.14. Condition (6) also follows because each $f \mapsto f^{\subseteq J}$ is a linear operator, as discussed after Definition 8.17.

We now verify (2). Assume $T \subsetneq S$ and that $g \in L^2(\Omega^n, \pi^{\otimes n})$ only depends on the coordinates in T. We have

$$\langle f^{=S}, g\rangle = \sum_{J \subseteq S} (-1)^{|S|-|J|} \langle f^{\subseteq J}, g\rangle. \tag{8.3}$$

Take any $i \in S \setminus T$ and pair up the summands in (8.3) as J', J'', where $J' \not\ni i$ and $J'' = J' \cup \{i\}$. By Lemma 8.34 we have

$$\langle f^{\subseteq J''}, g\rangle = \langle f^{\subseteq J'' \cap T}, g^{\subseteq T}\rangle = \langle f^{\subseteq J' \cap T}, g^{\subseteq T}\rangle,$$

the latter equality using $i \notin T$. But the signs $(-1)^{|S|-|J'|}$ and $(-1)^{|S|-|J''|}$ are opposite, so the summands in (8.3) cancel in pairs. This shows the sum is 0, confirming (2).

We complete the existence proof by noting that (2) $\implies$ (3) $\implies$ (4) (assuming (1)). The first implication is because $\langle f^{=S}, g\rangle = \langle f^{=S}, g^{\subseteq S \cap T}\rangle$ when g depends only on the coordinates in T (Lemma 8.34), and $S \cap T \subsetneq S$ when $S \not\subseteq T$. The second implication is because $S \neq T$ implies either $S \not\subseteq T$ or $T \not\subseteq S$.

It remains to prove the uniqueness statement. Suppose f has two representations satisfying (1) and (2). By subtracting them we get a decomposition of the 0 function that satisfies (1) and (2); our goal is to show that each function in this decomposition is the 0 function. We can do this by showing that any decomposition satisfying (1) and (2) also satisfies "Parseval's Theorem": $\langle f, f\rangle = \sum_{S\subseteq[n]} \|f^{=S}\|_2^2$. But this is an easy consequence of (4), which we just noted is itself a consequence of (1) and (2). □

We can connect the orthogonal decomposition of f to its expansion under Fourier bases as follows:

Proposition 8.36. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ have orthogonal decomposition $f = \sum_{S\subseteq[n]} f^{=S}$. Fix any Fourier basis $\phi_0,\dots,\phi_{m-1}$ for $L^2(\Omega,\pi)$. Then*

$$f^{=S} = \sum_{\substack{\alpha\in\mathbb{N}^n_{<m}\\ \mathrm{supp}(\alpha)=S}} \widehat{f}(\alpha)\phi_\alpha. \tag{8.4}$$

Proof. This follows easily from the uniqueness part of Theorem 8.35. If we take (8.4) as the definition of functions $f^{=S}$, it is immediate that $\sum_S f^{=S} = f$ and that $f^{=S}$ depends only on the coordinates in S. Further, if g depends only on coordinates $T \subsetneq S$, then $f^{=S}$ and g have disjoint Fourier support by Corollary 8.20; hence $\langle f^{=S}, g\rangle = 0$ by Plancherel (Proposition 8.16). □

Example 8.37. Let's compute the orthogonal decomposition of the function $f : \{a,b,c\}^2 \to \{0,1\}$ from Example 8.15. Recall that in this example $\{a,b,c\}$ has the uniform distribution and $f(x_1,x_2) = 1$ if and only if $x_1 = x_2 = c$. First,

$$f^{=\emptyset} = \mathbf{E}[f] = \tfrac{1}{9}.$$

Next, for $i = 1,2$ we have that $f^{\subseteq\{i\}}(x)$ is $\frac{1}{3}$ if $x_i = c$ and 0 otherwise; hence

$$f^{=\{i\}}(x_1,x_2) = \begin{cases} +\frac{2}{9} & \text{if } x_i = c,\\ -\frac{1}{9} & \text{else.}\end{cases}$$

Finally, it's easiest to compute $f^{=\{1,2\}}$ as $f - f^{=\emptyset} - f^{=\{1\}} - f^{=\{2\}}$; this yields

$$f^{=\{1,2\}}(x_1,x_2) = \begin{cases} +\frac{4}{9} & \text{if } x_1 = x_2 = c,\\ -\frac{2}{9} & \text{if exactly one of } x_1,\ x_2 \text{ is } c,\\ +\frac{1}{9} & \text{if } x_1, x_2 \neq c.\end{cases}$$

You can check (Exercise 8.20) that this is consistent with Proposition 8.36 and the Fourier expansion from Example 8.15.

We can write all of the Fourier formulas from Section 8.2 in terms of the orthogonal decomposition; e.g.,

$$\langle f, g\rangle = \sum_{S\subseteq[n]} \langle f^{=S}, g^{=S}\rangle, \quad \mathbf{Inf}_i[f] = \sum_{S\ni i} \|f^{=S}\|_2^2, \quad \mathrm{T}_\rho f = \sum_{S\subseteq[n]} \rho^{|S|} f^{=S}.$$

These formulas can be proved either by using the connection from Proposition 8.36 or by reasoning directly from the defining Theorem 8.35; see Exercise 8.18. The orthogonal decomposition also gives us the natural way of stratifying f by degree; we end this section by generalizing some more definitions from Chapter 1.4:

Definition 8.38. For $f \in L^2(\Omega^n, \pi^{\otimes n})$ and $k \in \mathbb{N}$ we define the *degree k part of f* to be $f^{=k} = \sum_{|S|=k} f^{=S}$ and the *weight of f at degree k* to be $\mathbf{W}^k[f] = \|f^{=k}\|_2^2$. We also use notation like $f^{\leq k} = \sum_{|S|\leq k} f^{=S}$ and $\mathbf{W}^{>k}[f] = \sum_{|S|>k} \|f^{=S}\|_2^2$.

8.4. p-biased analysis

Perhaps the most common generalized domain in analysis of Boolean functions is the case of the hypercube with "biased" bits. In this setting we think of a random input in $\{-1,1\}^n$ as having each bit independently equal to -1 (True) with probability $p \in (0,1)$ and equal to 1 (False) with probability $q = 1-p$. (We could also consider different parameters p_i for each coordinate; see Exercise 8.24.) In the notation of the chapter this means $L^2(\Omega^n, \pi_p^{\otimes n})$, where $\Omega = \{-1,1\}$ and π_p is the distribution on Ω defined by $\pi_p(-1) = p$, $\pi_p(1) = q$. This context is often referred to as *p-biased Fourier analysis*, though it would be more consistent with our terminology if it were called "μ-biased", where

$$\mu = \mathop{\mathbf{E}}_{\boldsymbol{x}_i \sim \pi_p}[\boldsymbol{x}_i] = q - p = 1 - 2p.$$

One of the more interesting features of the setting is that we can fix a combinatorial Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ and then consider its properties for various p between 0 and 1; we will discuss this further later in this section. We will also sometimes use the abbreviated notation $\mathbf{Pr}_{\pi_p}[\cdot]$ in place of $\mathbf{Pr}_{\boldsymbol{x} \sim \pi_p^{\otimes n}}[\cdot]$, and similarly $\mathbf{E}_{\pi_p}[\cdot]$.

The p-biased hypercube is one of the generalized domains where it can pay to look at an explicit Fourier basis. In fact, since we have $|\Omega| = 2$ there is a *unique* Fourier basis $\{\phi_0, \phi_1\}$ (up to negating ϕ_1). For notational simplicity we'll write ϕ instead of ϕ_1 and use "set notation" rather than multi-index notation:

Definition 8.39. In the context of p-biased Fourier analysis we define the basis function $\phi : \{-1,1\} \to \mathbb{R}$ by

$$\phi(x_i) = \frac{x_i - \mu}{\sigma},$$

where

$$\mu = \mathop{\mathbf{E}}_{\boldsymbol{x}_i \sim \pi_p}[\boldsymbol{x}_i] = q - p = 1 - 2p, \quad \sigma = \mathop{\mathbf{stddev}}_{\boldsymbol{x}_i \sim \pi_p}[\boldsymbol{x}_i] = \sqrt{4pq} = 2\sqrt{p}\sqrt{1-p}.$$

Note that $\sigma^2 = 1 - \mu^2$. We also have the formula $\phi(1) = \sqrt{p/q}$, $\phi(-1) = -\sqrt{q/p}$.

We will use the notation μ and σ throughout this section. It's clear that $\{1, \phi\}$ is indeed a Fourier basis for $L^2(\{-1,1\}, \pi_p)$ because $\mathbf{E}[\phi(\boldsymbol{x}_i)] = 0$ and $\mathbf{E}[\phi(\boldsymbol{x}_i)^2] = 1$ by design.

Definition 8.40. In the context of $L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ we define the product Fourier basis functions $(\phi_S)_{S\subseteq[n]}$ by

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

Given $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ we write $\widehat{f}(S)$ for the associated Fourier coefficient; i.e.,

$$\widehat{f}(S) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[f(\boldsymbol{x})\phi_S(\boldsymbol{x})].$$

Thus we have the biased Fourier expansion

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

Although the notation is very similar to that of the classic uniform-distribution Fourier analysis, we caution that in general,

$$\phi_S\phi_T \neq \phi_{S\triangle T}.$$

Example 8.41. Let $\chi_i \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ be the ith dictator function, $\chi_i(x) = x_i$, viewed under the p-biased distribution. We have

$$\phi(x_i) = \frac{x_i - \mu}{\sigma} \quad\implies\quad x_i = \mu + \sigma\phi(x_i),$$

and the latter is evidently f's (biased) Fourier expansion. That is,

$$\widehat{\chi_i}(\emptyset) = \mu, \quad \widehat{\chi_i}(\{i\}) = \sigma, \quad \widehat{\chi_i}(S) = 0 \text{ otherwise}.$$

This example lets us see a link between a function's "usual" Fourier expansion and its biased Fourier expansion. (For more on this, see Exercise 8.25.) Let's abuse notation a little by writing simply ϕ_i instead of $\phi(x_i)$. We have the formulas

$$\phi_i = \frac{x_i - \mu}{\sigma} \quad\iff\quad x_i = \mu + \sigma\phi_i, \tag{8.5}$$

and we can go from the usual Fourier expansion to the biased Fourier expansion simply by plugging in the latter.

Example 8.42. Recall the "selection function" $\mathrm{Sel} : \{-1,1\}^3 \to \{-1,1\}$ from Exercise 1.1(*j*); $\mathrm{Sel}(x_1, x_2, x_2)$ outputs x_2 if $x_1 = -1$ and outputs x_3 if $x_1 = 1$. The usual Fourier expansion of Sel is

$$\mathrm{Sel}(x_1,x_2,x_3) = \tfrac{1}{2}x_2 + \tfrac{1}{2}x_3 - \tfrac{1}{2}x_1x_2 + \tfrac{1}{2}x_1x_3.$$

Using the substitution from (8.5) we get

$$\begin{aligned}\mathrm{Sel}(x_1,x_2,x_3) &= \tfrac{1}{2}(\mu+\sigma\phi_2) + \tfrac{1}{2}(\mu+\sigma\phi_3) - \tfrac{1}{2}(\mu+\sigma\phi_1)(\mu+\sigma\phi_2) + \tfrac{1}{2}(\mu+\sigma\phi_1)(\mu+\sigma\phi_3)\\ &= \mu + (\tfrac{1}{2} - \tfrac{1}{2}\mu)\sigma\,\phi_2 + (\tfrac{1}{2} + \tfrac{1}{2}\mu)\sigma\,\phi_3 - \tfrac{1}{2}\sigma^2\,\phi_1\phi_2 + \tfrac{1}{2}\sigma^2\,\phi_1\phi_3. \end{aligned} \tag{8.6}$$

Thus if we write $\mathrm{Sel}^{(p)}$ for the selection function thought of as an element of $L^2(\{-1,1\}^3, \pi_p^{\otimes 3})$, we have

$$\widehat{\mathrm{Sel}^{(p)}}(\emptyset) = \mu, \quad \widehat{\mathrm{Sel}^{(p)}}(2) = (\tfrac{1}{2} - \tfrac{1}{2}\mu)\sigma, \quad \widehat{\mathrm{Sel}^{(p)}}(3) = (\tfrac{1}{2} + \tfrac{1}{2}\mu)\sigma,$$
$$\widehat{\mathrm{Sel}^{(p)}}(\{1,2\}) = -\tfrac{1}{2}\sigma^2, \quad \widehat{\mathrm{Sel}^{(p)}}(\{1,3\}) = \tfrac{1}{2}\sigma^2, \quad \widehat{\mathrm{Sel}^{(p)}}(S) = 0 \text{ else.}$$

By the Fourier formulas of Section 8.2 we can deduce, e.g., that $\mathbf{E}[\mathrm{Sel}^{(p)}] = \mu$, $\mathbf{Inf}_1[\mathrm{Sel}^{(p)}] = (-\frac{1}{2}\sigma^2)^2 + (\frac{1}{2}\sigma^2)^2 = \frac{1}{2}\sigma^4$, etc.

Let's codify a piece of notation from this example:

Notation 8.43. Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $p \in (0,1)$. We write $f^{(p)}$ for the function when viewed as an element of $L^2(\{-1,1\}^n, \pi_p^{\otimes n})$.

We now discuss derivative operators. We would like to define an operator D_i on $L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ that acts like differentiation on the biased Fourier expansion. For example, referring to (8.6) we would like to have

$$\mathrm{D}_3\mathrm{Sel}^{(p)} = (\tfrac{1}{2} + \tfrac{1}{2}\mu)\sigma + \tfrac{1}{2}\sigma^2\phi_1.$$

In general we are seeking $\frac{\partial}{\partial \phi_i}$ which, by basic calculus and the relationship (8.5), satisfies

$$\frac{\partial}{\partial \phi_i} = \frac{\partial x_i}{\partial \phi_i} \cdot \frac{\partial}{\partial x_i} = \sigma \cdot \frac{\partial}{\partial x_i}.$$

Recognizing $\frac{\partial}{\partial x_i}$ as the "usual" ith derivative operator, we are led to the following:

Definition 8.44. For $i \in [n]$, the *ith (discrete) derivative* operator D_i on $L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ is defined by

$$\mathrm{D}_i f(x) = \sigma \cdot \frac{f(x^{(i\mapsto 1)}) - f(x^{(i\mapsto -1)})}{2}.$$

Note that this defines a different operator for each value of p. We sometimes write the above definition as

$$\mathrm{D}_{\phi_i} = \sigma \cdot \mathrm{D}_{x_i}.$$

With respect to the biased Fourier expansion of $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ the operator D_i satisfies

$$\mathrm{D}_i f = \sum_{S \ni i} \widehat{f}(S)\,\phi_{S\setminus\{i\}}. \tag{8.7}$$

Given this definition we can derive some additional formulas for influences, including a generalization of Proposition 2.21:

Proposition 8.45. *Suppose $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ is Boolean-valued (i.e., has range $\{-1,1\}$). Then*

$$\mathbf{Inf}_i[f] = \sigma^2 \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{\oplus i})]$$

for each $i \in [n]$, and

$$\mathbf{I}[f] = \sigma^2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[\mathrm{sens}_f(\boldsymbol{x})].$$

If furthermore f is monotone, then $\mathbf{Inf}_i[f] = \sigma \widehat{f}(i)$.

Proof. Using Definition 8.44's notation we have

$$\mathbf{Inf}_i[f] = \mathop{\mathbf{E}}_{\pi_p}[(\mathrm{D}_{\phi_i} f)^2] = \sigma^2 \mathop{\mathbf{E}}_{\pi_p}[(\mathrm{D}_{x_i} f)^2].$$

Since $(\mathrm{D}_{x_i} f)^2$ is the 0-1 indicator that i is pivotal for f, the first formula follows. The second formula follows by summing over i. Finally, when f is monotone we furthermore have that $(\mathrm{D}_{x_i} f)^2 = \mathrm{D}_{x_i} f$ and hence

$$\mathbf{Inf}_i[f] = \sigma^2 \mathop{\mathbf{E}}_{\pi_p}[\mathrm{D}_{x_i} f] = \sigma \mathop{\mathbf{E}}_{\pi_p}[\mathrm{D}_{\phi_i} f] = \sigma \widehat{f}(i),$$

as claimed. □

The remainder of this section is devoted to the topic of *threshold phenomena* in Boolean functions. Much of the motivation for this comes from theory of random graphs, which we now briefly introduce.

Definition 8.46. Given an undirected graph G on $v \geq 2$ vertices, we identify it with the string in $\{\mathsf{True}, \mathsf{False}\}^{\binom{v}{2}}$ which indicates which edges are present (True) and which are absent (False). We write $\mathcal{G}(v,p)$ for the distribution $\pi_p^{\otimes\binom{v}{2}}$; this is called the *Erdős–Rényi random graph model*. Note that if we permute the v vertices of a graph, this induces a permutation on the $\binom{v}{2}$ edges. A (v-vertex) *graph property* is a Boolean function $f : \{\mathsf{True}, \mathsf{False}\}^{\binom{v}{2}} \to \{\mathsf{True}, \mathsf{False}\}$ that is invariant under all $v!$ such permutations of its input; colloquially, this means that f "does not depend on the names of the vertices".

Graph properties are always transitive-symmetric functions in the sense of Definition 2.10.

Example 8.47. The following are all v-vertex graph properties:

$$\begin{aligned}
\mathrm{Conn}(G) &= \mathsf{True} \text{ if } G \text{ is connected;}\\
\mathrm{3Col}(G) &= \mathsf{True} \text{ if } G \text{ is 3-colorable;}\\
\mathrm{Clique}_k(G) &= \mathsf{True} \text{ if } G \text{ is contains a clique on at least } k \text{ vertices;}\\
\mathrm{Maj}_n(G) &= \mathsf{True} \text{ (assuming } n = \tbinom{v}{2} \text{ is odd) if } G \text{ has at least } \tbinom{v}{2}/2 \text{ edges;}\\
\chi_{[n]}(G) &= \mathsf{True} \text{ if } G \text{ has an odd number of edges.}
\end{aligned}$$

Note that each of these actually defines a family of Boolean functions, one for each value of v; this is the typical situation in the study of graph properties. An example of a function $f : \{\mathsf{True}, \mathsf{False}\}^{\binom{v}{2}} \to \{\mathsf{True}, \mathsf{False}\}$ that is *not* a graph property is the one defined by $f(G) = \mathsf{True}$ if vertex #1 has at least one neighbor; this f is not invariant under permuting the vertices.

Graph properties which are *monotone* are particularly nice to study; these are the ones for which adding edges can never make the property go from True to False. The properties Conn, Clique$_k$, and Maj$_n$ defined above are all monotone, as is ¬3Col. Now suppose we take a monotone graph property, say, Conn. A typical question in random graph theory would be, "how many edges does a graph need to have before it is likely to be connected?" Or more precisely, how does $\mathbf{Pr}_{\boldsymbol{G}\sim\mathcal{G}(v,p)}[\mathrm{Conn}(\boldsymbol{G}) = \mathsf{True}]$ vary as p increases from 0 to 1?

There's no need to ask this question just for graph properties. Given any monotone Boolean function $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ it is intuitively clear that when p increases from 0 to 1 this causes $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ to increase from 0 to 1 (unless f is a constant function). As illustration, we show a plot of $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ versus p for the dictator function, AND$_2$, and Maj$_{101}$.

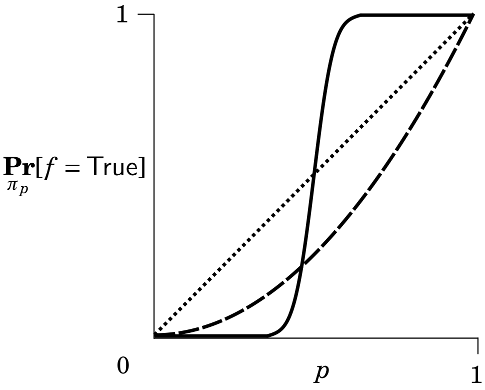


Figure 8.1. Plot of $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ versus p for f a dictator (dotted), $f = \mathrm{AND}_2$ (dashed), and $f = \mathrm{Maj}_{101}$ (solid)

The *Margulis–Russo Formula* quantifies the rate at which $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ increases with p; specifically, it relates the slope of the curve at p to the total influence of f under $\pi_p^{\otimes n}$. To prove the formula we switch to ±1 notation.

Margulis–Russo Formula. *Let $f : \{-1,1\}^n \to \mathbb{R}$. Recalling Notation 8.43 and the relation $\mu = 1 - 2p$, we have*

$$\frac{d}{d\mu}\mathbf{E}[f^{(p)}] = \frac{1}{\sigma} \cdot \sum_{i=1}^{n} \widehat{f^{(p)}}(i). \tag{8.8}$$

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

$$\frac{d}{dp} \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[f(\boldsymbol{x}) = -1] = \frac{d}{d\mu}\mathbf{E}[f^{(p)}] = \frac{1}{\sigma^2}\cdot \mathbf{I}[f^{(p)}]. \tag{8.9}$$

Proof. Treating f as a multilinear polynomial over $x_1,\dots,x_n$ we have

$$\mathbf{E}[f^{(p)}] = \mathrm{T}_\mu f(1,\dots,1) = f(\mu,\dots,\mu)$$

(this also follows from Exercise 1.4). By basic calculus,

$$\frac{d}{d\mu} f(\mu,\dots,\mu) = \sum_{i=1}^n \mathrm{D}_{x_i} f(\mu,\dots,\mu).$$

But

$$\mathrm{D}_{x_i} f(\mu,\dots,\mu) = \mathbf{E}[\mathrm{D}_{x_i} f^{(p)}] = \frac{1}{\sigma}\mathbf{E}[\mathrm{D}_{\phi_i} f^{(p)}] = \frac{1}{\sigma}\widehat{f^{(p)}}(i),$$

completing the proof of (8.8). As for (8.9), the second equality follows immediately from Proposition 8.45. The first equality holds because $\mu = 1-2p$ and $\mathbf{E}[f] = 1 - 2\mathbf{Pr}[f = -1]$; the two factors of -2 cancel. □

Remark 8.48. If $f : \{\mathsf{True},\mathsf{False}\}^n \to \{\mathsf{True},\mathsf{False}\}$ is a nonconstant monotone function, the Margulis–Russo Formula implies that $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ is a strictly increasing function of p, because $\mathbf{I}[f^{(p)}]$ is always positive.

Looking again at Figure 8.1 we see that the plot for Maj_{101} looks very much like a step function, jumping from nearly 0 to nearly 1 around the critical value $p = 1/2$. For Maj_n, this "sharp threshold at $p = 1/2$" becomes more and more pronounced as n increases. This is clearly suggested by the Margulis–Russo Formula: the derivative of the curve at $p = 1/2$ is equal to $\mathbf{I}[\mathrm{Maj}_n]$ (the usual, uniform-distribution total influence), which has the very large value $\Theta(\sqrt{n})$ (Theorem 2.33). Such sharp thresholds exist for many Boolean functions; we give some examples:

Example 8.49. In Exercise 8.23 you are asked to show that for every $\epsilon > 0$ there is a C such that

$$\mathop{\mathbf{Pr}}_{\pi_{1/2-C/\sqrt{n}}}[\mathrm{Maj}_n = \mathsf{True}] \le \epsilon, \qquad \mathop{\mathbf{Pr}}_{\pi_{1/2+C/\sqrt{n}}}[\mathrm{Maj}_n = \mathsf{True}] \ge 1-\epsilon.$$

Regarding the Erdős–Rényi graph model, the following facts are known:

$$\mathop{\mathbf{Pr}}_{\boldsymbol{G}\sim\mathcal{G}(v,p)}[\mathrm{Clique}_{\log v}(\boldsymbol{G}) = \mathsf{True}] \xrightarrow[v\to\infty]{} \begin{cases} 0 & \text{if } p < 1/4, \\ 1 & \text{if } p > 1/4. \end{cases}$$

$$\mathop{\mathbf{Pr}}_{\boldsymbol{G}\sim\mathcal{G}(v,p)}[\mathrm{Conn}(\boldsymbol{G}) = \mathsf{True}] \xrightarrow[v\to\infty]{} \begin{cases} 0 & \text{if } p < \frac{\ln v}{v}(1 - \frac{\log\log v}{\log v}), \\ 1 & \text{if } p > \frac{\ln v}{v}(1 + \frac{\log\log v}{\log v}). \end{cases}$$

In the above examples you can see that the "jump" occurs at various values of p. To investigate this phenomenon, we first single out the value for which $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}] = 1/2$:

Definition 8.50. Let $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ be monotone and nonconstant. The *critical probability* for f, denoted p_c, is the unique value in $(0,1)$ for which $\mathbf{Pr}_{\boldsymbol{x} \sim \pi_p^{\otimes n}}[f(\boldsymbol{x}) = \mathsf{True}] = 1/2$. We also write $q_c = 1 - p_c$, $\mu_c = q_c - p_c = 1 - 2p_c$, and $\sigma_c = \sqrt{4p_c q_c}$.

In Exercise 8.27 you are asked to verify that p_c is well defined.

Looking at the connectivity property from Example 8.49 we see that not only does $\mathbf{Pr}_{\pi_p}[\mathsf{Conn} = \mathsf{True}]$ jump from near 0 to near 1 in an interval of the form $p_c \pm o(1)$, it actually makes the jump in an interval of the form $p_c(1 \pm o(1))$. This latter phenomenon is (roughly speaking) what is meant by a "sharp threshold". To investigate this further, suppose that f is a (nonconstant) monotone function and Δ is the derivative of $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ at $p = p_c$. Intuitively, we would expect $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ to jump from near 0 to near 1 in an interval of around p_c of width about $1/\Delta$. Thus a "sharp threshold" should roughly correspond to the case that $1/\Delta$ is small even compared to $\min(p_c, q_c)$. The Margulis–Russo Formula says that $\Delta = \frac{1}{\sigma_c^2}\mathbf{I}[f^{(p_c)}]$, and since $\min(p_c, q_c)$ is proportional to $4p_c q_c = \sigma_c^2$ it follows that $1/\Delta$ is "small" compared to $\min(p_c, q_c)$ if and only if $\mathbf{I}[f^{(p_c)}]$ is "large". Thus we have a neat criterion:

Sharp threshold principle: *Let $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ be monotone. Then, roughly speaking, $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}]$ has a "sharp threshold" if and only if f has "large" ("superconstant") total influence under its critical probability distribution.*

Of course this should all be made a bit more precise; see Exercise 8.28 for details. In light of this principle, we may try to prove that a given f has a sharp threshold by proving that $\mathbf{I}[f^{(p_c)}]$ is not "small". This strongly motivates the problem of "characterizing" Boolean-valued functions $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ for which $\mathbf{I}[f]$ is small. Friedgut's Junta Theorem, mentioned at the end of Chapter 3.1 and proved in Chapter 9.6, tells us that in the uniform distribution case $p = 1/2$, the only way $\mathbf{I}[f]$ can be small is if f is close to a junta. In particular, any monotone graph property with $p_c = 1/2$ must have a very large derivative $\frac{d}{dp}\mathbf{Pr}_{\pi_p}[f = \mathsf{True}]$ at $p = p_c$: since the function is transitive-symmetric, all n coordinates are equally influential and it can't be close to a junta. These results also hold so long as p is bounded away from 0 and 1; see Chapter 10.3. However, many interesting monotone graph properties have p_c very close to 0: e.g., connectivity, as we saw in Example 8.49. Characterizing the functions $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ with small $\mathbf{I}[f]$ when $p = o_n(1)$ is a trickier task; see the work of Friedgut, Bourgain, and Hatami described in Chapter 10.5.

8.5. Abelian groups

The previous section covered the case of $f \in L^2(\Omega^n, \pi^{\otimes n})$ with $|\Omega| = 2$; there, we saw it could be helpful to look at explicit Fourier bases. When $|\Omega| \geq 3$ this is often *not* helpful, especially if the only "operation" on the domain is equality. For example, if $f : \{\mathsf{Red}, \mathsf{Green}, \mathsf{Blue}\}^n \to \mathbb{R}$, then it's best to just work abstractly with the orthogonal decomposition. However, if there is a notion of, say, "addition" in Ω, then there is a natural, canonical Fourier basis for $L^2(\Omega, \pi)$ when π is the uniform distribution.

More precisely, suppose the domain Ω is a finite abelian group G, with operation $+$ and identity 0. We will consider the domain G under the uniform probability distribution π; this is quite natural because π is *translation-invariant*: $\pi(X) = \pi(t + X)$ for any $X \subseteq G$, $t \in G$. In this setting it is more convenient to allow functions with range the complex numbers; thus we come to the following definition:

Definition 8.51. Let G be a finite abelian group with operation $+$ and identity 0. For $n \in \mathbb{N}^+$ we write $L^2(G^n)$ for the complex inner product space of functions $f : G^n \to \mathbb{C}$, with inner product

$$\langle f, g \rangle = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim G^n}[f(\boldsymbol{x})\overline{g(\boldsymbol{x})}].$$

Here and throughout this section $\boldsymbol{x} \sim G^n$ denotes that $\boldsymbol{x}$ is drawn from the uniform distribution on G^n.

Everything we have done in this chapter for the real inner product space $L^2(\Omega^n, \pi^{\otimes n})$ generalizes easily to the case of a complex inner product; the main difference is that Plancherel's Theorem becomes

$$\langle f, g \rangle = \sum_{\alpha \in \mathbb{N}^n_{<m}} \widehat{f}(\alpha)\overline{\widehat{g}(\alpha)} = \sum_{S \subseteq [n]} \langle f^{=S}, g^{=S} \rangle.$$

See Exercise 8.32 for more.

A natural Fourier basis for $L^2(G)$ comes from a natural family of functions $G \to \mathbb{C}$, namely the *characters*. These are defined to be the group homomorphisms from G to $\mathbb{C}^\times$, where $\mathbb{C}^\times$ is the abelian group of nonzero complex numbers under multiplication.

Definition 8.52. A *character* of the (finite) group G is a function $\chi : G \to \mathbb{C}^\times$ which is a homomorphism; i.e., satisfies $\chi(x + y) = \chi(x)\chi(y)$. Since G is finite there is some $m \in \mathbb{N}^+$ such that $0 = x + x + \cdots + x$ (m times) for each $x \in G$. Thus $1 = \chi(0) = \chi(x)^m$, meaning the range of χ is in fact contained in the mth roots of unity. In particular, $|\chi(x)| = 1$ for all $x \in G$.

We have the following easy facts:

Fact 8.53. *If χ and ϕ are characters of G, then so are $\overline{\chi}$ and $\phi \cdot \chi$.*

Proposition 8.54. *Let χ be a character of G. Then either $\chi \equiv 1$ or $\mathbf{E}[\chi] = 0$.*

Proof. If $\chi \not\equiv 1$, pick some $y \in G$ such that $\chi(y) \neq 1$. Since $\boldsymbol{x} + y$ is uniformly distributed on G when $\boldsymbol{x} \sim G$,

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim G}[\chi(\boldsymbol{x})] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim G}[\chi(\boldsymbol{x}+y)] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim G}[\chi(\boldsymbol{x})\chi(y)] = \chi(y) \mathop{\mathbf{E}}_{\boldsymbol{x}\sim G}[\chi(\boldsymbol{x})].$$

Since $\chi(y) \neq 1$ it follow that $\mathbf{E}[\chi(\boldsymbol{x})]$ must be 0. □

Proposition 8.55. *The set of all characters of G is orthonormal. (As a consequence, G has at most $\dim(L^2(G)) = |G|$ characters.)*

Proof. First, if χ is a character, then $\langle \chi, \chi \rangle = \mathbf{E}[|\chi|^2] = 1$ because $|\chi| \equiv 1$. Next, if ϕ is another character distinct from χ then $\langle \phi, \chi \rangle = \mathbf{E}[\phi \cdot \overline{\chi}]$. But $\phi \cdot \overline{\chi}$ is a character by Fact 8.53, and $\phi \cdot \overline{\chi} = \phi/\chi \not\equiv 1$ because ϕ and χ are distinct; here we used $\overline{\chi} = 1/\chi$ because $|\chi| \equiv 1$. Thus $\langle \phi, \chi \rangle = 0$ by Proposition 8.54. □

As we will see next, G in fact has exactly $|G|$ characters. It thus follows from Proposition 8.55 that the set of all characters (which includes the constant 1 function) constitutes a Fourier basis for $L^2(G)$.

To check that each finite abelian group G has $|G|$ distinct characters, we begin with the case of a cyclic group, $\mathbb{Z}_m$ for some m. In this case we know that every character's range will be contained in the mth roots of unity.

Definition 8.56. Fix an integer $m \geq 2$ and write ω for the mth root of unity $\exp(2\pi i/m)$. For $0 \leq j < m$, we define $\chi_j : \mathbb{Z}_m \to \mathbb{C}$ by $\chi_j(x) = \omega^{jx}$. It is easy to see that these are distinct characters of $\mathbb{Z}_m$.

Thus the functions $\chi_0 \equiv 1, \chi_1, \dots, \chi_{m-1}$ form a Fourier basis for $L^2(\mathbb{Z}_m)$. Furthermore, Proposition 8.13 tells us that we can get a Fourier basis for $L^2(\mathbb{Z}_m^n)$ by taking all products of these functions.

Definition 8.57. Continuing Definition 8.56, let $n \in \mathbb{N}^+$. For $\alpha \in \mathbb{N}_{<m}^n$ we define $\chi_\alpha : \mathbb{Z}_m^n \to \mathbb{C}$ by

$$\chi_\alpha(x) = \prod_{j=1}^n \chi_{\alpha_j}(x_j).$$

These functions are easily seen to be (all of the) characters of the group $\mathbb{Z}_m^n$, and they constitute a Fourier basis of $L^2(\mathbb{Z}_m^n)$.

Most generally, by the Fundamental Theorem of Finitely Generated Abelian Groups we know that any finite abelian G is a direct product of cyclic groups of prime-power order. In Exercise 8.35 you are asked to check that you get all of the characters of G – and hence a Fourier basis for $L^2(G)$ – by taking all products of the associated cyclic groups' characters. In the remainder of the section we mostly stick to groups of the form $\mathbb{Z}_m^n$ for simplicity.

Returning to the characters $\chi_0, \dots, \chi_{m-1}$ from Definition 8.56, it is easy to see (using $\omega^m = 1$) that they satisfy $\chi_j \cdot \chi_{j'} = \chi_{j+j' \text{ (mod } m)}$ and also $1/\chi_j = \overline{\chi_j} = \chi_{-j \text{ (mod } m)}$. Thus the characters themselves form a group under multiplication, isomorphic to $\mathbb{Z}_m$. As in Chapter 3.2, we index them using the notation $\widehat{\mathbb{Z}_m}$. More generally, indexing the Fourier basis/characters of $L^2(\mathbb{Z}_m^n)$ by $\widehat{\mathbb{Z}_m^n}$ instead of multi-indices, we have:

Fact 8.58. *The characters $(\chi_\alpha)_{\alpha \in \widehat{\mathbb{Z}_m^n}}$ of $\mathbb{Z}_m^n$ form a group under multiplication:*

- $\chi_\alpha \cdot \chi_\beta = \chi_{\alpha+\beta}$,
- $1/\chi_\alpha = \overline{\chi_\alpha} = \chi_{-\alpha}$.

As mentioned, the salient feature of $L^2(G)$ distinguishing it from other spaces $L^2(\Omega, \pi)$ is that there is a notion of addition on the domain. This means that *convolution* plays a major role in its analysis. We generalize the definition from the setting of $\mathbb{F}_2^n$:

Definition 8.59. Let $f, g \in L^2(G)$. Their *convolution* is the function $f * g \in L^2(G)$ defined by

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

Exercise 8.36 asks you to check that convolution is associative and commutative, and that the following generalization of Theorem 1.27 holds:

Theorem 8.60. *Let $f, g \in L^2(G)$. Then $\widehat{f * g}(\alpha) = \widehat{f}(\alpha)\widehat{g}(\alpha)$.*

We conclude this section by mentioning vector space domains. When doing Fourier analysis over the group $\mathbb{Z}_m^n$, it is natural for subgroups to arise. Things are simplest when the only subgroups of $\mathbb{Z}_m$ are the trivial ones, $\{0\}$ and $\mathbb{Z}_m$; in this case, all subgroups will be isomorphic to $\mathbb{Z}_m^{n'}$ for some $n' \leq n$. Of course, this simple situation occurs if and only if m is equal to some prime p. In that case, $\mathbb{Z}_p$ can be thought of as a field, $\mathbb{Z}_p^n$ as an n-dimensional vector space over this field, and its subgroups as subspaces. We use the notation $\mathbb{F}_p^n$ in this setting and write $\widehat{\mathbb{F}_p^n}$ to index the Fourier basis/characters; this generalizes the notation introduced for $p = 2$ in Chapter 3.2. Indeed, all of the notions from Chapters 3.2 and 3.3 regarding affine subspaces and restrictions thereto generalize easily to $L^2(\mathbb{F}_p^n)$.

8.6. Highlight: Randomized decision tree complexity

A decision tree T for $f : \{-1,1\}^n \to \{-1,1\}$ can be thought of as a deterministic algorithm which, given adaptive query access to the bits of an unknown string

$x \in \{-1,1\}^n$, outputs $f(x)$. For example, to describe a natural decision tree for $f = \mathrm{Maj}_3$ in words: "Query x_1, then x_2. If they are equal, output their value; otherwise, query and output x_3." For a worst-case input (one where $x_1 \neq x_2$) this algorithm has a *cost* of 3, meaning it makes 3 queries. The cost of the worst-case input is the depth of the decision tree.

As is often the case with algorithms it can be advantageous to allow randomization. For example, consider using the following randomized query algorithm for Maj_3: "Choose two distinct input coordinates at random and query them. If they are equal, output their value; otherwise, query and output the third input coordinate." Now for *every* input there is at least a 1/3 chance that the algorithm will finish after only 2 queries. Indeed, if we define the cost of an input x to be the expected number of queries the algorithm makes on it, it is easy to see that the worst-case inputs for this algorithm have cost $(1/3)\cdot 2 + (2/3)\cdot 3 = 8/3 < 3$.

Let's formalize the notion of a randomized decision tree:

Definition 8.61. Given $f : \{-1,1\}^n \to \mathbb{R}$, a *(zero-error) randomized decision tree* $\mathscr{T}$ computing f is formally defined to be a probability distribution over (deterministic) decision trees that compute f. The *cost* of $\mathscr{T}$ on input $x \in \{-1,1\}^n$ is defined to be the expected number of queries $\boldsymbol{T}$ makes on x when $\boldsymbol{T} \sim \mathscr{T}$. The cost of $\mathscr{T}$ itself is defined to be the maximum cost of any input. Finally, the *(zero-error) randomized decision tree complexity* of f, denoted $\mathrm{RDT}(f)$, is the minimum cost of a randomized decision tree computing f.

We can get further savings from randomization if we are willing to assume that the input x is chosen randomly. For example, if $\boldsymbol{x} \sim \{-1,1\}^3$ is uniformly random then any of the deterministic decision trees for Maj_3 will make 2 queries with probability 1/2 and 3 queries with probability 1/2, for an overall expected $5/2 < 8/3 < 3$ queries.

Definition 8.62. Let $\mathscr{T}$ be a randomized decision tree. We define

$$\delta_i(\mathscr{T}) = \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim\{-1,1\}^n,\\ \boldsymbol{T}\sim\mathscr{T}}}[\boldsymbol{T} \text{ queries } \boldsymbol{x}_i],$$

$$\Delta(\mathscr{T}) = \sum_{i=1}^n \delta_i(\mathscr{T}) = \mathop{\mathbf{E}}_{\substack{\boldsymbol{x}\sim\{-1,1\}^n,\\ \boldsymbol{T}\sim\mathscr{T}}}[\#\text{ of coordinates queried by } \boldsymbol{T} \text{ on } \boldsymbol{x}]. \tag{8.10}$$

Given $f : \{-1,1\}^n \to \mathbb{R}$, we define $\Delta(f)$ to be the minimum of $\Delta(\mathscr{T})$ over all randomized decision trees $\mathscr{T}$ computing f.

We can also generalize these definitions for functions $f \in L^2(\Omega, \pi^{\otimes n})$. A deterministic decision tree over domain Ω is the natural generalization in which each internal query node has $|\Omega|$ outgoing edges, labeled by the elements of Ω. We write $\delta_i^{(\pi)}(\mathscr{T}), \Delta^{(\pi)}(\mathscr{T}), \Delta^{(\pi)}(f)$ for the generalizations to trees over Ω; in the case of $L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ we use the superscript (p) instead of (π_p) for brevity.

It follows immediately from the definitions that for any $f \in L^2(\Omega^n, \pi^{\otimes n})$,

$$\Delta^{(\pi)}(f) \leq \mathrm{RDT}(f) \leq \mathrm{DT}(f).$$

Remark 8.63. In the definition of $\Delta^{(\pi)}(f)$ it is equivalent if we only allow deterministic decision trees; this is because in (8.10) we can always choose the "best" deterministic T in the support of $\mathscr{T}$.

Example 8.64. It follows from our discussions that $\mathrm{RDT}(\mathrm{Maj}_3) \leq 8/3$ and $\Delta(\mathrm{Maj}_3) \leq 5/2$; indeed, it's not hard to show that both of these bounds are equalities. In Exercise 8.38 you are asked to generalize to the recursive majority of 3 function on $n = 3^d$ inputs; it satisfies $\mathrm{DT}(\mathrm{Maj}_3^{\otimes d}) = 3^d = n$, but

$$\mathrm{RDT}(\mathrm{Maj}_3^{\otimes d}) \leq (8/3)^d = n^{\log_3(8/3)} \approx n^{.89},$$
$$\Delta(\mathrm{Maj}_3^{\otimes d}) \leq (5/2)^d = n^{\log_3(5/2)} \approx n^{.83}.$$

Incidentally, these bounds are not asymptotically sharp; estimating $\mathrm{RDT}(\mathrm{Maj}_3^{\otimes d})$ in particular is a well-studied open problem.

Example 8.65. In Exercise 8.39 you are asked to show that for the logical OR function, $\Delta^{(p)}(\mathrm{OR}_n) = \frac{1-(1-p)^n}{p}$, which is roughly 2 for $p = 1/2$ but is asymptotic to $n/(2\ln 2)$ at the critical probability p_c.

Example 8.64 illustrates a mildly surprising phenomenon: using randomness it's possible to evaluate certain unbiased n-bit functions f while reading only a $1/n^{\Theta(1)}$ fraction of the input bits. This is even more interesting when f is transitive-symmetric like $\mathrm{Maj}_3^{\otimes d}$. In that case it's not hard to show (Exercise 8.37) that any randomized decision tree $\mathscr{T}$ computing f can be converted to one where $\Delta(\mathscr{T})$ remains the same but all $\delta_i(\mathscr{T})$ are equal to $\Delta(f)/n$. Then f can be evaluated despite the fact that *each* input bit is only queried with probability $1/n^{\Theta(1)}$.

In this section we explore the limits of this phenomenon. In particular, a longstanding conjecture of Yao [Yao77] says that this is *not* possible for monotone graph properties:

Yao's Conjecture. *Let* $f : \{-1,1\}^n \to \{-1,1\}$ *be a nonconstant monotone v-vertex graph property, where* $n = \binom{v}{2}$. *Then* $\mathrm{RDT}(f) \geq \Omega(n)$.

Toward this conjecture we will present a lower bound due to O'Donnell, Saks, Schramm, and Servedio [OSSS05]. (Two other incomparable bounds are discussed in the notes for this chapter.) It has the advantages that it works for the more general class of transitive-symmetric functions and that it even lower-bounds $\Delta^{(p_c)}(f)$:

Theorem 8.66. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a nonconstant monotone transitive-symmetric function with critical probability p_c. Then*

$$\Delta^{(p_c)}(f) \geq (n/\sigma_c)^{2/3}.$$

Theorem 8.66 is essentially sharp in several interesting cases. Whenever the critical probability p_c is $\Theta(1/n)$ or $1-\Theta(1/n)$ then $\sigma_c = \Theta(1/\sqrt{n})$ and Theorem 8.66 gives the strongest possible bound, $\Delta^{(p_c)}(f) \geq \Omega(n)$. This occurs, e.g., for the OR_n function (Example 8.65). Furthermore, Theorem 8.66 can be tight up to a logarithmic factor when $p_c = 1/2$ as the following theorem of Benjamini, Schramm, and Wilson shows:

Theorem 8.67. [**BSW05**]. *There exists an infinite family of monotone transitive-symmetric functions $f_n : \{-1,1\}^n \to \{-1,1\}$ with critical probability $p_c = 1/2$ and $\Delta(f) \leq O(n^{2/3} \log n)$.*

Theorem 8.66 follows easily from two inequalities [**OS06, OS07**], [**OSSS05**], which we now present:

OS Inequality. *Let $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$. Then $\sum_{i=1}^n \widehat{f}(i) \leq \|f\|_2 \cdot \sqrt{\Delta^{(p)}(f)}$. In particular, if f has range $\{-1,1\}$ and is monotone, then $\mathbf{I}[f] \leq \sigma\sqrt{\Delta^{(p)}(f)}$.*

OSSS Inequality. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ have range $\{-1,1\}$ and let $\mathcal{T}$ be any randomized decision tree computing f. Then*

$$\mathbf{Var}[f] \leq \sum_{i=1}^n \delta_i^{(\pi)}(\mathcal{T}) \cdot \mathbf{Inf}_i[f].$$

Remark 8.68. An interesting corollary of the OSSS Inequality is that

$$\mathbf{MaxInf}[f] \geq \mathbf{Var}[f]/\Delta^{(\pi)}(f) \geq \mathbf{Var}[f]/\mathrm{DT}(f) \geq \mathbf{Var}[f]/\deg(f)^3,$$

the last inequality assuming $\Omega = \{-1,1\}$. See Exercise 8.44.

These two inequalities can be thought of as strengthenings of basic Fourier inequalities which take into account the decision tree complexity of f. The OS Inequality essentially generalizes the result that majority functions maximizes $\sum_{i=1}^n \widehat{f}(i)$; i.e., Theorem 2.33. The OSSS Inequality is a generalization of the Poincaré Inequality, discounting the influences of coordinates that are rarely read.

We will first derive the query complexity lower bound Theorem 8.66 from the OS and OSSS Inequalities. We will then prove the latter two inequalities.

Proof of Theorem 8.66. We consider f to be an element of $L^2(\{-1,1\}^n, \pi_{p_c}^{\otimes n})$. Let $\mathcal{T}$ be a randomized decision tree achieving $\Delta^{(p_c)}(f)$. In the OSSS Inequality, we have $\mathbf{Var}[f] = 1$ since p_c is the critical probability and $\mathbf{Inf}_i[f] = \mathbf{I}[f]/n$

for each $i \in [n]$ since f is transitive-symmetric. Thus

$$1 \le \sum_{i=1} \delta_i^{(p_c)}(\mathscr{T}) \cdot \frac{\mathbf{I}[f]}{n} \quad \Longrightarrow \quad n \le \Delta^{(p_c)}(f) \cdot \mathbf{I}[f] \le \sigma \Delta^{(p_c)}(f)^{3/2},$$

where we used the OS Inequality. The theorem follows by rearranging. □

Now we prove the OS and OSSS Inequalities, starting with the latter. We will need a simple lemma that uses the decomposition $f = \mathrm{E}_i f + \mathrm{L}_i f$.

Lemma 8.69. *Let $f, g \in L^2(\Omega^n, \pi^{\otimes n})$ and let $j \in [n]$. Given $\omega \in \Omega$, write $f_{|\omega}$ for the restriction of f in which the jth coordinate is fixed to value ω, and similarly for g. Then*

$$\mathbf{Cov}[f,g] = \mathop{\mathbf{E}}_{\substack{\boldsymbol{\omega},\boldsymbol{\omega}' \sim \pi \\ \text{independent}}} [\mathbf{Cov}[f_{|\boldsymbol{\omega}}, g_{|\boldsymbol{\omega}'}]] + \langle \mathrm{L}_j f, \mathrm{L}_j g \rangle.$$

Proof. Since the covariances and Laplacians are unchanged when constants are added, we may assume without loss of generality that $\mathbf{E}[f] = \mathbf{E}[g] = 0$. Then $\mathbf{Cov}[f,g] = \langle f, g \rangle$ and

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{\omega},\boldsymbol{\omega}'}[\mathbf{Cov}[f_{|\boldsymbol{\omega}}, g_{|\boldsymbol{\omega}'}]] &= \mathop{\mathbf{E}}_{\boldsymbol{\omega},\boldsymbol{\omega}'}[\langle f_{|\boldsymbol{\omega}}, g_{|\boldsymbol{\omega}'} \rangle - \mathbf{E}[f_{|\boldsymbol{\omega}}]\mathbf{E}[g_{|\boldsymbol{\omega}'}]] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{\omega},\boldsymbol{\omega}'}[\langle f_{|\boldsymbol{\omega}}, g_{|\boldsymbol{\omega}'} \rangle] - \mathbf{E}[f]\mathbf{E}[g] = \mathop{\mathbf{E}}_{\boldsymbol{\omega},\boldsymbol{\omega}'}[\langle f_{|\boldsymbol{\omega}}, g_{|\boldsymbol{\omega}'} \rangle] = \langle \mathrm{E}_j f, \mathrm{E}_j g \rangle.\end{aligned}$$

Thus the stated equality reduces to the basic (Exercise 8.8) identity

$$\langle f, g \rangle = \langle \mathrm{E}_j f, \mathrm{E}_j g \rangle + \langle \mathrm{L}_j f, \mathrm{L}_j g \rangle.$$

□

Proof of the OSSS Inequality. More generally we show that if $g : \{-1,1\}^n \to \{-1,1\}$ is also an element of $L^2(\Omega^n, \pi^{\otimes n})$, then

$$\mathbf{Cov}[f,g] \le \sum_{i=1}^n \delta_i^{(\pi)}(\mathscr{T}) \cdot \mathbf{Inf}_i[g]. \tag{8.11}$$

The result then follow by taking $g = f$. We may also assume that $\mathscr{T} = T$ is a single deterministic tree computing f; this is because (8.11) is linear in the quantities $\delta_i^{(\pi)}(\mathscr{T})$.

We prove (8.11) by induction on the structure of T. If T is depth-0, then f must be a constant function; hence $\mathbf{Cov}[f,g] = 0$ and (8.11) is trivial. Otherwise, let $j \in [n]$ be the coordinate queried at the root of T. For each $\omega \in \Omega$, write T_ω for the subtree of T given by the ω-labeled child of the root. By applying Lemma 8.69 and induction (noting that T_ω computes the restricted function $f_{|\omega}$), we get

$$\begin{aligned}
\mathbf{Cov}[f,g] &= \mathop{\mathbf{E}}_{\substack{\boldsymbol{\omega},\boldsymbol{\omega}'\sim\pi\\ \text{independent}}}[\mathbf{Cov}[f_{|\boldsymbol{\omega}}, g_{|\boldsymbol{\omega}'}]] + \langle \mathrm{L}_j f, \mathrm{L}_j g\rangle \\
&\le \mathop{\mathbf{E}}_{\boldsymbol{\omega},\boldsymbol{\omega}'\sim\pi}\Big[\sum_{i\neq j} \delta_i^{(\pi)}(T_{\boldsymbol{\omega}})\cdot \mathbf{Inf}_i[g_{\boldsymbol{\omega}'}]\Big] + \langle \mathrm{L}_j f, \mathrm{L}_j g\rangle \\
&= \sum_{i\neq j} \delta_i^{(\pi)}(T)\cdot \mathbf{Inf}_i[g] + \langle f, \mathrm{L}_j g\rangle \qquad \text{(in part since } \mathbf{E}[\mathrm{L}_j g] = 0) \\
&\le \sum_{i\neq j} \delta_i^{(\pi)}(T)\cdot \mathbf{Inf}_i[g] + \mathbf{E}[|\mathrm{L}_j g|] \qquad \text{(since } |f|\le 1) \\
&= \sum_{i=1}^{n} \delta_i^{(\pi)}(T)\cdot \mathbf{Inf}_i[g],
\end{aligned}$$

where the last step used $\delta_j^{(\pi)}(T) = 1$ and Proposition 8.24. This completes the inductive proof of (8.11). □

Finally, we prove the OS Inequality. For this we require a definition.

Definition 8.70. Let (Ω,π) be a finite probability space and T a deterministic decision tree over Ω. The *decision tree process* associated to T generates a random string $\boldsymbol{x}$ distributed according to π (and some additional random variables), as follows:

(1) Start at the root node of T; say it queries coordinate j_1. Choose $\boldsymbol{x}_{j_1} \sim \pi$ and follow the outgoing edge labeled by the outcome.

(2) Suppose the node of T which is reached queries coordinate $\boldsymbol{j}_2$. Choose $\boldsymbol{x}_{\boldsymbol{j}_2} \sim \pi$ and follow the outgoing edge labeled by the outcome.

(3) Repeat until a leaf node is reached. At this point, define $\boldsymbol{J} = \{j_1, \boldsymbol{j}_2, \boldsymbol{j}_3, \dots\} \subseteq [n]$ to be the set of coordinates queried.

(4) Draw the as-yet-unqueried coordinates, denoted $\boldsymbol{x}_{\overline{\boldsymbol{J}}}$, from $\pi^{\otimes\overline{\boldsymbol{J}}}$.

Despite the fact that the coordinates $\boldsymbol{x}_i$ are drawn in a random, dependent order, it's not hard to see (Exercise 8.42) that the final string $\boldsymbol{x} = (\boldsymbol{x}_{\boldsymbol{J}}, \boldsymbol{x}_{\overline{\boldsymbol{J}}})$ is distributed according the product distribution $\pi^{\otimes n}$.

Proof of the OS Inequality. We will prove the claim $\sum_{i=1}^n \widehat{f}(i) \le \|f\|_2 \cdot \sqrt{\Delta^{(p)}(f)}$; the "in particular" statement follows immediately from Proposition 8.45. Fix a deterministic decision tree T achieving $\Delta^{(p)}(f)$ (see Remark 8.63) and let $\boldsymbol{x} = (\boldsymbol{x}_{\boldsymbol{J}}, \boldsymbol{x}_{\overline{\boldsymbol{J}}})$ be drawn from the associated decision tree process. Using the notation ϕ from Definition 8.39 we have

$$\sum_{i=1}^{n} \widehat{f}(i) = \mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x}_{\boldsymbol{J}},\boldsymbol{x}_{\overline{\boldsymbol{J}}}}[f(\boldsymbol{x})\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)] = \mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x}_{\boldsymbol{J}}}[f(\boldsymbol{x}_{\boldsymbol{J}})\mathop{\mathbf{E}}_{\boldsymbol{x}_{\overline{\boldsymbol{J}}}}[\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)]].$$

Here we abused notation slightly by writing $f(\boldsymbol{x_J})$; in the decision tree process, f's value is determined once $\boldsymbol{x_J}$ is. Since $\mathbf{E}[\phi(\boldsymbol{x}_i)] = 0$ for each $i \not\in \boldsymbol{J}$ we may continue:

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}[f(\boldsymbol{x_J})\mathop{\mathbf{E}}_{\boldsymbol{x_{\overline{J}}}}[\sum_{i=1}^n \phi(\boldsymbol{x}_i)]] &= \mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}[f(\boldsymbol{x_J})\sum_{i=1}^n \mathbf{1}_{\{i\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)] \\ &\le \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}[f(\boldsymbol{x_J})^2]}\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}\left[\left(\sum_{i=1}^n \mathbf{1}_{\{i\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)\right)^2\right]},\end{aligned}$$

by Cauchy–Schwarz. Now $\sqrt{\mathbf{E}_{\boldsymbol{J},\boldsymbol{x_J}}[f(\boldsymbol{x_J})^2]}$ is simply $\|f\|_2$ since T computes f. To complete the proof it suffices to show that

$$\mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}\left[\left(\sum_{i=1}^n \mathbf{1}_{\{i\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)\right)^2\right] = \Delta^{(p)}(f).$$

To see this, expand the square:

$$\mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}\left[\left(\sum_{i=1}^n \mathbf{1}_{\{i\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)\right)^2\right] = \sum_{i=1}^n \mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}[\mathbf{1}_{\{i\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)^2] + \sum_{i\neq i'} \mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}[\mathbf{1}_{\{i,i'\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_{i'})].$$

Conditioned on $i \in \boldsymbol{J}$ the quantity $\mathbf{E}[\phi(\boldsymbol{x}_i)^2]$ is simply 1. Thus

$$\sum_{i=1}^n \mathop{\mathbf{E}}_{\boldsymbol{J},\boldsymbol{x_J}}[\mathbf{1}_{\{i\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)^2] = \sum_{i=1}^n \mathbf{Pr}[i \in \boldsymbol{J}] = \Delta^{(p)}(f).$$

It remains to show that $\mathbf{E}_{\boldsymbol{J},\boldsymbol{x_J}}[\mathbf{1}_{\{i,i'\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_{i'})] = 0$ whenever $i \neq i'$. Suppose we condition on the event that $i, i' \in \boldsymbol{J}$ and we further condition on i being queried before i' is queried. Certainly this may affect the conditional distribution of $\boldsymbol{x}_i$, but the conditional distribution of $\boldsymbol{x}_{i'}$ remains π_p; hence $\mathbf{E}[\phi(\boldsymbol{x}_{i'})] = 0$ under this conditioning. Of course the same argument holds when we condition on i' being queried before i. It follows that $\mathbf{E}_{\boldsymbol{J},\boldsymbol{x_J}}[\mathbf{1}_{\{i,i'\in\boldsymbol{J}\}}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_{i'})]$ is indeed 0, completing the proof. □

8.7. Exercises and notes

8.1 Explain how to generalize the definitions and results in Sections 8.1 and 8.2 to general finite product spaces $L^2(\Omega_1 \times \cdots \times \Omega_n, \pi_1 \times \cdots \times \pi_n)$.

8.2 Verify that Definition 8.1 indeed defines a real inner product space. (Where is the fact that π has full support used?)

8.3 Verify the formula for $\widehat{f}(\alpha)$ in Definition 8.14.

8.4 Verify that ϕ_0, ϕ_1, ϕ_2 from Example 8.10 indeed constitute a Fourier basis for $\Omega = \{a, b, c\}$ with the uniform distribution.

8.5 Verify the Fourier expansion in Example 8.15.

8.6 Complete the proof of Proposition 8.16.

8.7 Prove that the expectation over I operator, E_I, is a linear operator on $L^2(\Omega^n, \pi^{\otimes n})$ (i.e., $\mathrm{E}_I(f+g) = \mathrm{E}_I f + \mathrm{E}_I g$), a projection (i.e., $\mathrm{E}_I \circ \mathrm{E}_I = \mathrm{E}_I$), and self-adjoint (i.e., $\langle f, \mathrm{E}_I g\rangle = \langle \mathrm{E}_I f, g\rangle$). Deduce that T_ρ is also self-adjoint.

8.8 Show for any $f, g \in L^2(\Omega^n, \pi^{\otimes n})$ and $j \in [n]$ that $f = \mathrm{E}_j f + \mathrm{L}_j f$ and that $\langle f, g\rangle = \langle \mathrm{E}_j f, \mathrm{E}_j g\rangle + \langle \mathrm{L}_j f, \mathrm{L}_j g\rangle$.

8.9 Prove Proposition 8.24. (Hint: Exercise 1.17.)

8.10 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ have range $\{-1,1\}$. Proposition 8.24 tells us that $\|\mathrm{L}_i f\|_1 = \|\mathrm{L}_i f\|_2^2 = \mathbf{Inf}_i[f]$.

(*a*) Show that $\|\mathrm{L}_i f\|_p^p \le 2^p \mathbf{Inf}_i[f]$ for any $p \ge 1$.

(*b*) In case $1 \le p \le 2$, show that in fact $\|\mathrm{L}_i f\|_p^p \le \mathbf{Inf}_i[f]$. (Hint: Use the general form of Hölder's inequality to bound $\|\mathrm{L}_i f\|_p$ in terms of $\|\mathrm{L}_i f\|_1$ and $\|\mathrm{L}_i f\|_2$.)

8.11 Generalize all of Exercise 2.35 to the setting of $L^2(\Omega^n, \pi^{\otimes n})$. Caution: the two statements referring to $\rho \in [-1,1]$ should refer only to $\rho \in [0,1]$ in this more general setting.

8.12 Assume $|\Omega| = m$ and let π denote the uniform distribution on Ω.

(*a*) For $x \in \Omega^n$ and $\boldsymbol{y} \sim N_\rho(x)$, write a formula for $\mathbf{Pr}[\boldsymbol{y}_i = \omega]$ in terms of ρ (there are two cases depending on whether or not $x_i = \omega$).

(*b*) Verify that your formula defines a valid probability distribution on Ω even when $-\frac{1}{m-1} \le \rho < 0$. We may therefore extend the definition of N_ρ to this case. (Cf. the second half of Definition 2.40.)

(*c*) Verify that for $\boldsymbol{x} \sim \pi^{\otimes n}$ and $\boldsymbol{y} \sim N_\rho(\boldsymbol{x})$, the distribution of $(\boldsymbol{x}, \boldsymbol{y})$ is symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$.

(*d*) Show that when $\boldsymbol{y} \sim N_{-\frac{1}{m-1}}(x)$, each $\boldsymbol{y}_i$ is uniformly distributed on $\Omega \setminus \{x_i\}$.

(*e*) Verify that the formula for T_ρ from Proposition 8.28 continues to hold for $-\frac{1}{m-1} \le \rho < 0$. (Hint: Use the fact that it holds for $\rho \in [0,1]$ and that the formula in part (*a*) is a polynomial in ρ.)

8.13 Show that Definition 8.30 extends by continuity to

$$\mathbf{Inf}_i^{(0)}[f] = \sum_{\substack{\#\alpha=1 \\ \alpha_i \neq 0}} \widehat{f}(\alpha)^2.$$

Extend also Proposition 8.31 to the case of $\delta = 1$.

8.14 Prove explicitly that condition 5 holds in Theorem 8.35.

8.15 Prove that condition 6 must hold in Theorem 8.35 directly from the uniqueness statement (i.e., without appealing to the explicit construction).

8.16 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Prove directly from the defining Theorem 8.35 that $(f^{=S})^{\subseteq T}$ is equal to $f^{=S}$ if $S \subseteq T$ and is equal to 0 otherwise.

8.17 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ and let $\boldsymbol{x} \sim \pi^{\otimes n}$. In this exercise you should think about how the (conditional) expectation of f changes as the random variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are revealed one at a time.

(a) Recalling that $f^{\subseteq [t]}(\boldsymbol{x})$ depends only on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_t$, show that the sequence of random variables $(f^{\subseteq [t]}(\boldsymbol{x}))_{t=0\dots n}$ is a martingale (where $f^{\subseteq [0]}$ denotes $f^{\emptyset}$); i.e.,

$$\mathbf{E}[f^{\subseteq [t]}(\boldsymbol{x}) \mid f^{\subseteq [0]}(\boldsymbol{x}), \dots, f^{\subseteq [t-1]}(\boldsymbol{x})] = f^{\subseteq [t-1]}(\boldsymbol{x}) \quad \forall t \in [n].$$

(This is the *Doob martingale* for f.)

(b) For each $t \in [n]$ define

$$\mathrm{d}_t f = f^{\subseteq [t]} - f^{\subseteq [t-1]} = \sum_{\substack{S \subseteq [n] \\ \max(S) = t}} f^{=S}.$$

Show that $\mathbf{E}[\mathrm{d}_t f(\boldsymbol{x}) \mid f^{\subseteq [0]}(\boldsymbol{x}), \dots, f^{\subseteq [t-1]}(\boldsymbol{x})] = 0$. (Here $(\mathrm{d}_t f)_{t=1\dots n}$ is the *martingale difference sequence* for f.)

8.18 For $f, g \in L^2(\Omega^n, \pi^{\otimes n})$, prove the following directly from Theorem 8.35:

$$\langle f, g \rangle = \sum_{S \subseteq [n]} \langle f^{=S}, g^{=S} \rangle$$

$$\mathbf{Inf}_i[f] = \sum_{S \ni i} \|f^{=S}\|_2^2$$

$$\mathbf{I}[f] = \sum_{k=0}^{n} k \cdot \mathbf{W}^k[f]$$

$$\mathrm{T}_\rho(f^{=S}) = (\mathrm{T}_\rho f)^{=S} = \rho^k f^{=S}$$

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

8.19 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ and let $S \subseteq [n]$. Show that $\|f^{=S}\|_\infty \le 2^{|S|} \|f\|_\infty$.

8.20 Explicitly verify that Proposition 8.36 holds for the function in Examples 8.15 and 8.37.

8.21 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ and let $i \in S \subseteq [n]$. Suppose we take $f^{=S}$ and restrict its ith coordinate to have value ω_i, forming the subfunction $g = (f^{=S})_{|\omega_i}$. Show that $g = g^{=S \setminus \{i\}}$. In particular, $\mathbf{E}[g] = 0$ assuming $|S| \ge 2$.

8.22 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be a symmetric function. Show that if $1 \le |S| \le |T| \le n$, then $\frac{1}{|S|} \mathbf{Var}[f^{\subseteq S}] \le \frac{1}{|T|} \mathbf{Var}[f^{\subseteq T}]$.

8.23 Prove the sharp threshold statement about the majority function made in Example 8.49. (Hint: Chernoff bound.) In the social choice literature, this fact is known as the *Condorcet Jury Theorem*.

8.24 Let $p_1, \dots, p_n \in (0,1)$ and let $\pi = \pi_{p_1} \otimes \cdots \pi_{p_n}$ be the associated product distribution on $\{-1,1\}^n$. Write $\mu_i = 1 - 2p_i$ and $\sigma_i = 2\sqrt{p_i}\sqrt{1-p_i}$. Generalize Proposition 8.45 to the setting of $L^2(\{-1,1\}^n, \pi)$.

8.25 Let $f : \{-1,1\}^n \to \mathbb{R}$ and consider the general product distribution setting of Exercise 8.24.

(*a*) For $S = \{i_1, \dots, i_k\} \subseteq [n]$, write D_{ϕ_S} for $\mathrm{D}_{\phi_{i_1}} \circ \cdots \circ \mathrm{D}_{\phi_{i_k}}$ and similarly D_{x_S}. Show that $\mathrm{D}_{\phi_S} = \prod_{i \in S} \sigma_i \cdot \mathrm{D}_{x_S}$.

(*b*) Writing $f^{(\mu)}$ for the function f viewed as an element of $L^2(\{-1,1\}^n, \pi)$, show that

$$\widehat{f^{(p)}}(S) = \prod_{i \in S} \sigma_i \cdot \mathrm{D}_{x_S} f(\mu_1, \dots, \mu_n).$$

(*c*) Show that $\hat{\|} f^{(p)} \hat{\|}_\infty \le \prod_{i \in S} \sigma_i \cdot \|f\|_\infty$.

8.26 (*a*) Generalize Exercise 2.10 by showing that for $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ with range $\{-1,1\}$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \pi_p^{\otimes n}}[i \text{ is } b\text{-pivotal for } f \text{ on } \boldsymbol{x}] = \pi_p(b)\mathbf{Inf}_i[f]$$

for $i \in [n]$ and $b \in \{-1,1\}$.

(*b*) Generalize Proposition 4.7 by showing that if $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathrm{DNF}_{\mathrm{width}}(f) \le w$, then $\mathbf{I}[f^{(p)}] \le 4qw \le 4w$, and if f has $\mathrm{CNF}_{\mathrm{width}}(f) \le w$, then $\mathbf{I}[f^{(p)}] \le 4pw \le 4w$.

8.27 Fix any $\alpha \in (0,1)$. Let $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ be a nonconstant monotone function. Show that there exists $p \in (0,1)$ such that $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}] = \alpha$. (Hint: Intermediate Value Theorem.)

8.28 Fix a small constant $0 < \epsilon < 1/2$. Let $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ be a nonconstant monotone function. Let p_0 (respectively, p_c, p_1) be the unique value of $p \in (0,1)$ such that $\mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = \mathsf{True}] = \epsilon$ (respectively, 1/2, $1-\epsilon$). (This is a valid definition by Exercise 8.27.) Define also $\sigma_c^2 = 4p_c(1-p_c)$. The *threshold interval* for f is defined to be $[p_0, p_1]$, and $\delta = p_1 - p_0$ is the *threshold width*. Now let $(f_n)_{n \in \mathbb{N}}$ be a sequence of nonconstant monotone Boolean functions (usually "naturally related", with f_n's input length an increasing function of n). Define the sequences $p_0(n)$, $p_c(n)$, $p_1(n)$, $\sigma_c^2(n)$, $\delta(n)$. We say that the family (f_n) has a *sharp threshold* if $\delta(n)/\sigma_c^2(n) \to 0$ as $n \to \infty$; otherwise, we say it has a *coarse threshold*. (Note: If $p_c(n) \le 1/2$ for all n, this is the same as saying that $\delta(n)/p_c(n) \to 0$.) Show that if (f_n) has a coarse threshold, then there exists $C < \infty$, an infinite sequence $n_1 < n_2 < n_3 < \cdots$, and a sequence $(p(n_i))_{i \in \mathbb{N}}$ such that:

- $\epsilon < \mathbf{Pr}_{\pi_{p(n_i)}}[f_{n_i}(\boldsymbol{x}) = \mathsf{True}] < 1 - \epsilon$ for all i;
- $\mathbf{I}[f_{n_i}^{(p(n_i))}] \le C$ for all i.

(Hint: Margulis–Russo and the Mean Value Theorem.)

8.29 Let $f : \{-1,1\}^n \to \{-1,1\}$ be a nonconstant monotone function and let $F : [0,1] \to [0,1]$ be the (strictly increasing) function defined by $F(p) = \mathbf{Pr}_{\pi_p}[f(\boldsymbol{x}) = -1]$. Let p_c be the critical probability such that $F(p_c) = 1/2$. Assume that $p_c \le 1/2$. (This is without loss of generality since we can replace f by $f^\dagger$. We often think of $p_c \ll 1/2$.) The goal of this exercise is to show a weak kind of threshold result: roughly speaking, $F(p) = o(1)$ when $p = o(p_c)$ and $F(p) = 1 - o(1)$ when $p = \omega(p_c)$.

(*a*) Using the Margulis–Russo Formula and the Poincaré Inequality show that for all $0 < p < 1$,

$$F'(p) \geq \frac{F(p)(1-F(p))}{p(1-p)}.$$

(*b*) Show that for all $p \leq p_c$ we have $F'(p) \geq \frac{F(p)}{2p}$ and hence $\frac{d}{dp}\ln F(p) \geq \frac{1}{2p}$.

(*c*) Deduce that for any $0 \leq p_0 \leq p_c$ we have $F(p_0) \leq \frac{1}{2}\sqrt{p_0/p_c}$; i.e., $F(p_0) \leq \epsilon$ if $p_0 \leq (2\epsilon)^2 p_c$.

(*d*) Show that the factor $(2\epsilon)^2$ can be improved to $\Theta(\tau)\epsilon^{1+\tau}$ for any small constant $\tau > 0$. (Hint: The quadratic dependence on ϵ arose because we used $1 - F(p) \geq 1/2$ for $p \leq p_c$; but from part (*c*) we have the improved bound $1 - F(p) \geq 1 - \tau$ once $p \leq (2\tau)^2 p_c$.)

(*e*) In the other direction, show that so long as $p_1 = \frac{1}{(2\epsilon)^2} p_c \leq 1/2$, we have $F(p_1) \geq 1 - \epsilon$. (Hint: Work with $\ln(1 - F(p))$.) In case $p_1 \leq 1/2$ does not hold, show that we at least have $F(1/2) \geq 1 - \sqrt{p_c/2}$.

(*f*) The bounds in part (*e*) are not very interesting when p_c is close to 1/2. Show that we also have $F(1-\delta) \geq 1 - \sqrt{\delta/2}$ (even when $p_c = 1/2$).

8.30 Consider the sequence of functions $f_n : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ defined for all odd $n \geq 3$ as follows: $f_n(x_1, \dots, x_n) = \mathrm{Maj}_3(x_1, x_2, \mathrm{Maj}_{n-2}(x_3, \dots, x_n))$.

(*a*) Show that f_n is monotone and has critical probability $p_c = 1/2$.

(*b*) Sketch a plot of $\mathbf{Pr}_{\pi_p}[f_n(\boldsymbol{x}) = \mathsf{True}]$ versus p (assuming n very large).

(*c*) Show that $\mathbf{I}[f_n] = \Theta(\sqrt{n})$.

(*d*) Show that the sequence f_n has a coarse threshold as defined in Exercise 8.28 (assuming $\epsilon < 1/4$).

8.31 (*a*) Consider the following probability distributions on strings $\boldsymbol{x} \in \mathbb{F}_2^n$:

(1) First choose $\boldsymbol{k} \sim \{0, 1, 2, \dots, n\}$ uniformly. Then choose $\boldsymbol{x}$ uniformly from the set of all strings of Hamming weight $\boldsymbol{k}$.

(2) First choose a uniformly random "path $\boldsymbol{\pi}$ from $(0, 0, \dots, 0)$ up to $(1, 1, \dots, 1)$"; i.e., let $\boldsymbol{\pi}$ be a uniformly random permutation from S_n and let $\boldsymbol{\pi}^{\leq i} \in \mathbb{F}_2^n$ denote the string whose jth coordinate is 1 if and only if $\pi(j) \leq i$. Then choose $\boldsymbol{k} \sim \{0, 1, 2, \dots, n\}$ uniformly and let $\boldsymbol{x}$ be the "$\boldsymbol{k}$th string on the path", namely $\boldsymbol{\pi}^{\leq \boldsymbol{k}}$.

(3) First choose $\boldsymbol{p} \sim [0, 1]$. Then choose $\boldsymbol{x} \sim \pi_{\boldsymbol{p}}^{\otimes n}$.

Show that these are in fact the same distribution. (Hint: Imagine choosing $n + 1$ indistinguishable points uniformly from $[0, 1]$ and then randomly assigning them the labels "p", 1, 2, . . . , n.)

(*b*) We denote by ν^n the distribution on $\mathbb{F}_2^{[n]}$ from part (*a*); more generally, we use the notation ν^N for the distribution on $\mathbb{F}_2^N$ where N is an abstract set of cardinality n. Given a nonempty $J \subseteq [n]$, show that if $\boldsymbol{x} \sim \nu^n$ and $\boldsymbol{x}_J \in \mathbb{F}_2^J$ denotes the restriction of $\boldsymbol{x}$ to coordinates J, then $\boldsymbol{x}_J$ has the distribution ν^J.

(*c*) Let $f : \mathbb{F}_2^n \to \mathbb{R}$ and fix $i \in [n]$. The *ith Shapley value* of f is defined to be

$$\mathbf{Shap}_i[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\nu^n}[f(\boldsymbol{x}^{(i\mapsto 1)}) - f(\boldsymbol{x}^{(i\mapsto 0)})].$$

Show that $\sum_{i=1}^n \mathbf{Shap}_i[f] = f(1,1,\dots,1) - f(0,0,\dots,0)$.

(*d*) Suppose $f : \mathbb{F}_2^n \to \{0,1\}$ is monotone. Show that $\mathbf{Shap}_i[f] = 4\int_0^1 \mathbf{Inf}_i[f^{(p)}]\,dp$.

8.32 Explain how to generalize the definitions and results in Sections 8.1, 8.2 to the case of the *complex* inner product space $L^2(\Omega^n, \pi^{\otimes n})$. In particular, verify the following formulas from Proposition 8.16:

$$\begin{aligned}
\mathbf{E}[f] &= \widehat{f}(0)\\
\mathbf{E}[|f|^2] = \mathbf{E}[\langle f, f\rangle] &= \sum_{\alpha\in\mathbb{N}^n_{<m}} \langle \widehat{f}(\alpha), \widehat{f}(\alpha)\rangle = \sum_{\alpha\in\mathbb{N}^n_{<m}} |\widehat{f}(\alpha)|^2\\
\mathbf{Var}[f] = \langle f - \mathbf{E}[f], f - \mathbf{E}[f]\rangle &= \sum_{\alpha\neq 0} |\widehat{f}(\alpha)|^2\\
\langle f, g\rangle &= \sum_{\alpha\in\mathbb{N}^n_{<m}} \langle \widehat{f}(\alpha), \widehat{g}(\alpha)\rangle = \sum_{\alpha\in\mathbb{N}^n_{<m}} \widehat{f}(\alpha)\overline{\widehat{g}(\alpha)}\\
\mathbf{Cov}[f,g] = \langle f - \mathbf{E}[f], g - \mathbf{E}[g]\rangle &= \sum_{\alpha\neq 0} \widehat{f}(\alpha)\overline{\widehat{g}(\alpha)}.
\end{aligned}$$

8.33 (*a*) As in Exercise 2.58, explain how to generalize the definitions and results in Sections 8.1, 8.2 to the case of functions $f : \Omega^n \to V$, where V is a real inner product space with inner product $\langle\cdot,\cdot\rangle_V$. Here the Fourier coefficients $\widehat{f}(\alpha)$ will be elements of V, and $\langle f, g\rangle$ is defined to be $\mathbf{E}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\langle f(\boldsymbol{x}), g(\boldsymbol{x})\rangle_V]$. In particular, verify the formulas from Proposition 8.16, including the Placherel Theorem $\langle f, g\rangle = \sum_\alpha \langle \widehat{f}(\alpha), \widehat{g}(\alpha)\rangle_V$.

(*b*) For Σ a finite set we write $\triangle_\Sigma$ for the set of all probability distributions over Σ (cf. Exercise 7.22). Writing $|\Sigma| = m$, we also identify $\triangle_\Sigma$ with the standard convex simplex in $\mathbb{R}^m$, namely $\{\mu \in \mathbb{R}^m : \mu_1 + \cdots + \mu_m = 1, \mu_i \geq 0\ \forall i\}$ (where we assume some fixed ordering of Σ). Finally, we identify the m elements of Σ with the constant distributions in $\triangle_\Sigma$; equivalently, the vertices of the form $(0,\dots,0,1,0,\dots,0)$. Given a function $f : \Omega^n \to \Sigma$, often the most useful way to treat it analytically is to interpret it as a function $f : \Omega^n \to \triangle_\Sigma \subset \mathbb{R}^m$ and then use the setting described in part (*a*), with $V = \mathbb{R}^m$. Using this idea, show that if $f : \Omega^n \to \Sigma$ and π is a distribution on Ω, then

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi^{\otimes n}, \boldsymbol{y}\sim N_\rho(\boldsymbol{x})}[f(\boldsymbol{x}) = f(\boldsymbol{y})].$$

(Here in $\mathbf{Stab}_\rho[f]$ we are interpreting f's range as $\triangle_\Sigma \subset \mathbb{R}^m$, whereas in the expression $f(\boldsymbol{x}) = f(\boldsymbol{y})$ we are treating f's range as the abstract set Σ.)

8.34 We say a function $f \in L^2(\Omega^n, \pi^{\otimes n})$ is a *linear threshold function* if it is expressible as $f(x) = \mathrm{sgn}(\ell(x))$, where $\ell : \Omega^n \to \mathbb{R}$ has degree at most 1 (in the sense of Definition 8.32).

(*a*) Given $\omega^{(+1)}, \omega^{(-1)} \in \Omega^n$ and $x \in \{-1,1\}^n$, we introduce the notation $\omega^{(x)}$ for the string $(\omega_1^{(x_1)}, \dots, \omega_n^{(x_n)}) \in \Omega^n$. Show that if $\boldsymbol{\omega}^{(+1)}, \boldsymbol{\omega}^{(-1)} \sim \pi^{\otimes n}$ are drawn independently and $(\boldsymbol{x}, \boldsymbol{y}) \sim \{-1,1\}^n \times \{-1,1\}^n$ is a ρ-correlated pair of binary strings, then $(\boldsymbol{\omega}^{(\boldsymbol{x})}, \boldsymbol{\omega}^{(\boldsymbol{y})})$ is a ρ-correlated pair under $\pi^{\otimes n}$.

(*b*) Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be a linear threshold function. Given a pair $\omega^{(+1)}, \omega^{(-1)} \in \Omega^n$, define $g_{\omega^{(+1)},\omega^{(-1)}} : \{-1,1\}^n \to \{-1,1\}$ by $g_{\omega^{(+1)},\omega^{(-1)}}(x) = f(\omega^{(x)})$. Show that $g_{\omega^{(+1)},\omega^{(-1)}}$ is a linear threshold function in the "usual" sense.

(*c*) Prove that Peres's Theorem (from Chapter 5.5) applies to linear threshold functions in $L^2(\Omega^n, \pi^{\otimes n})$, with the same bounds.

8.35 Let G be a finite abelian group. We know by the Fundamental Theorem of Finitely Generated Abelian Groups that $G \cong \mathbb{Z}_{m_1} \times \cdots \mathbb{Z}_{m_n}$ where each m_j is a prime power.

(*a*) Given $\alpha \in G$, define $\chi_\alpha : G \to \mathbb{C}$ by

$$\chi_\alpha(x) = \prod_{j=1}^{n} \exp(2\pi i x_j / m_j).$$

Show χ_α is a character of G and that the χ_α's are distinct functions for distinct α's. Deduce that the set of all χ_α's forms a Fourier basis for $L^2(G)$.

(*b*) Show that this set of characters forms a group under multiplication and that this group is isomorphic to G; i.e., generalize Fact 8.58. This is called the *dual group* of G and it is written $\widehat{G}$. We also identify the characters in $\widehat{G}$ with their indices α.

8.36 Verify that the convolution operation on $L^2(G)$ is associative and commutative, and that it satisfies $\widehat{f * g}(\alpha) = \widehat{f}(\alpha)\widehat{g}(\alpha)$ for all $\alpha \in \widehat{G}$. (See Exercise 8.35 for the definition of $\widehat{G}$.)

8.37 (*a*) Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be any transitive-symmetric function and let $\mathcal{T}$ be a randomized decision tree computing f. Show that there exists a randomized decision tree $\mathcal{T}'$ computing f with $\Delta^{(\pi)}(\mathcal{T}') = \Delta^{(\pi)}(\mathcal{T})$ and such that $\delta_i^{(\pi)}(\mathcal{T}')$ is the same for all $i \in [n]$. (Hint: Randomize over the automorphism group $\mathrm{Aut}(f)$ and use Exercise 2.47.)

(*b*) Given a randomized decision tree $\mathcal{T}$, let $\delta^{(\pi)}(\mathcal{T}) = \max_{i \in n}\{\delta_i^{(\pi)}(\mathcal{T})\}$. Given $f \in L^2(\{-1,1\}^n, \pi^{\otimes n})$, define $\delta^{(\pi)}(f)$ to be the minimum value of

$\delta_i^{(\pi)}(\mathcal{T})$ over all $\mathcal{T}$ which compute f; this is called the *revealment* of f. Show that if f is transitive-symmetric, then $\delta^{(\pi)}(f) = \frac{1}{n}\Delta^{(\pi)}(f)$.

8.38 (*a*) Show that $\mathrm{DT}(\mathrm{Maj}_3^{\otimes d}) = 3^d$, $\mathrm{RDT}(\mathrm{Maj}_3^{\otimes d}) \le (8/3)^d$, and $\Delta(\mathrm{Maj}_3^{\otimes d}) \le (5/2)^d$.

(*b*) Show that $\mathrm{RDT}(\mathrm{Maj}_3^{\otimes 2}) < (8/3)^2$. How small can you make your upper bound?

8.39 (*a*) Show that for every deterministic decision tree T computing the logical OR function on n bits,

$$\Delta^{(p)}(T) = p\cdot 1 + (1-p)p\cdot 2 + (1-p)^2 p\cdot 3 + \cdots$$
$$\cdots + (1-p)^{n-2}p\cdot(n-1) + (1-p)^{n-1}\cdot n = \frac{1-(1-p)^n}{p}.$$

Deduce $\Delta^{(p)}(\mathrm{OR}_n) = \frac{1-(1-p)^n}{p}$.

(*b*) Show that $\Delta^{(p_c)}(\mathrm{OR}_n) \sim n/(2\ln 2)$ as $n \to \infty$, where p_c denotes the critical probability for OR_n.

8.40 Let $\mathrm{NAND} : \{\mathsf{True}, \mathsf{False}\}^2 \to \{\mathsf{True}, \mathsf{False}\}$ be the function that outputs True unless both its inputs are True.

(*a*) Show that for d even, $\mathrm{NAND}^{\otimes d} = \mathrm{Tribes}_{2,2}^{\otimes d/2}$. (Thus the recursive NAND function is sometimes known as the AND-OR tree.)

(*b*) Show that $\mathrm{DT}(\mathrm{NAND}^{\otimes d}) = 2^d$.

(*c*) Show that $\mathrm{RDT}(\mathrm{NAND}) = 2$.

(*d*) For $b \in \{\mathsf{True}, \mathsf{False}\}$ and $\mathcal{T}$ a randomized decision tree computing a function f, let $\mathrm{RDT}_b(\mathcal{T})$ denote the maximum cost of $\mathcal{T}$ among inputs x with $f(x) = b$. Show that there is a randomized decision tree $\mathcal{T}$ computing NAND with $\mathrm{RDT}_{\mathsf{False}}(\mathcal{T}) = 3/2$.

(*e*) Show that $\mathrm{RDT}(\mathrm{NAND}^{\otimes 2}) \le 3$.

(*f*) Show that there is a family of randomized decision trees $(\mathcal{T}_d)_{d\in\mathbb{N}^+}$, with $\mathcal{T}_d$ computing $\mathrm{NAND}^{\otimes d}$, satisfying the inequalities

$$\begin{aligned}\mathrm{RDT}_{\mathsf{False}}(\mathcal{T}_d) &\le 2\mathrm{RDT}_{\mathsf{True}}(\mathcal{T}_{d-1})\\ \mathrm{RDT}_{\mathsf{True}}(\mathcal{T}_d) &\le \mathrm{RDT}_{\mathsf{False}}(\mathcal{T}_{d-1}) + (1/2)\mathrm{RDT}_{\mathsf{True}}(\mathcal{T}_{d-1}).\end{aligned}$$

(*g*) Deduce $\mathrm{RDT}(\mathrm{NAND}^{\otimes d}) \le (\frac{1+\sqrt{33}}{4})^d \approx n^{.754}$, where $n = 2^d$.

8.41 Let $\mathscr{C} = \{\text{monotone } f : \{-1,1\}^n \to \{-1,1\} \mid \mathrm{DT}(f) \le k\}$. Show that $\mathscr{C}$ is learnable from random examples with error ϵ in time $n^{O(\sqrt{k}/\epsilon)}$. (Hint: OS Inequality and Corollary 3.32.)

8.42 Verify that the decision tree process described in Definition 8.70 indeed generates strings distributed according to $\pi^{\otimes n}$. (Hint: Induction on the structure of the tree.)

8.43 Let T be a deterministic decision tree of size s. Show that $\Delta(T) \leq \log s$. (Hint: Let $\boldsymbol{P}$ be a random root-to-leaf path chosen as in the decision tree process. How can you bound the entropy of the random variable $\boldsymbol{P}$?)

8.44 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be a nonconstant function with range $\{-1,1\}$.

(*a*) Show that $\mathbf{MaxInf}[f] \geq \mathbf{Var}[f]/\Delta^{(\pi)}(f)$ (cf. the KKL Theorem from Chapter 4.2).

(*b*) In case $\Omega = \{-1,1\}$ show that $\mathbf{MaxInf}[f] \geq \mathbf{Var}[f]/\deg(f)^3$. (You should use the result of Midrijānis mentioned in the notes in Chapter 3.6.)

(*c*) Show that $\mathbf{I}[f] \geq \mathbf{Var}[f]/\delta^{(\pi)}(f)$, where $\delta^{(\pi)}(f)$ is the revealment of f, defined in Exercise 8.37(*b*).

8.45 Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ have range $\{-1,1\}$.

(*a*) Let $\mathcal{T}$ be a randomized decision computing f and let $i \in [n]$. Show that $\mathbf{Inf}_i[f] \leq \delta_i^{(\pi)}(f)$. (Hint: The decision tree process.)

(*b*) Suppose f is transitive-symmetric. Show that $\Delta^{(\pi)}(f) \geq \sqrt{\mathbf{Var}[f]/n}$. (Hint: Exercise 8.37(*b*).) This result can be sharp up to an $O(\sqrt{\log n})$ factor even for an $f : \{-1,1\}^n \to \{-1,1\}$ with $\mathbf{Var}[f] = 1$; see [BSW05].

8.46 In this exercise you will give an alternate proof of the OSSS Inequality that is sharp when $\mathbf{Var}[f] = 1$ and is weaker by only a factor of 2 when $\mathbf{Var}[f]$ is small. Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ have range $\{-1,1\}$. Given a randomized decision tree $\mathcal{T}$ we write $\mathrm{err}(\mathcal{T}) = \mathbf{Pr}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\mathcal{T}(\boldsymbol{x}) \neq f(\boldsymbol{x})]$.

(*a*) Let T be a depth-k deterministic decision tree (not necessarily computing f) whose root queries coordinate i. Let $\mathcal{T}$ be the distribution over deterministic trees of depth at most $k-1$ given by following a random outgoing edge from T's root (according to π). Show that $\mathrm{err}(\mathcal{T}) \leq \mathrm{err}(T) + \frac{1}{2}\mathbf{Inf}_i[f]$.

(*b*) Let $\mathcal{T}$ be a randomized decision tree of depth 0. Show that $\mathrm{err}(\mathcal{T}) \geq \min\{\mathbf{Pr}[f(\boldsymbol{x}) = 1], \mathbf{Pr}[f(\boldsymbol{x}) = -1]\}$.

(*c*) Prove by induction on depth that if $\mathcal{T}$ is any randomized decision tree, then $\frac{1}{2}\sum_{i=1}^{n} \delta_i^{(\pi)}(T) \cdot \mathbf{Inf}_i[f] \geq \min\{\mathbf{Pr}[f(\boldsymbol{x}) = 1], \mathbf{Pr}[f(\boldsymbol{x}) = -1]\} - \mathrm{err}(\mathcal{T})$. Verify that this yields the OSSS Inequality when $\mathbf{Var}[f] = 1$ and in general yields the OSSS Inequality up to a factor of 2.

8.47 Show that the OSSS Inequality fails for functions $f : \{-1,1\}^n \to \mathbb{R}$. (Hint: The simplest counterexample uses a decision tree with the shape in Figure 8.2.)

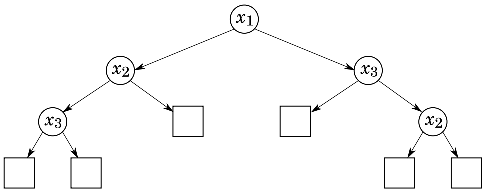


Figure 8.2. The basis for a counterexample to the OSSS Inequality when $f : \{-1,1\}^n \to \mathbb{R}$

Can you make the ratio of the left-hand side to the right-hand side equal to $\frac{130+20\sqrt{3}}{157}$? Larger?

Notes. The origins of the orthogonal decomposition described in Section 8.3 date back to the work of Hoeffding [Hoe48] (see also von Mises [vM47]). Hoeffding's work introduced *U-statistics*, i.e., functions f of independent random variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ of the form $\text{avg}_{1 \le i_1 < \cdots < i_k \le n}\, g(\boldsymbol{X}_{i_1}, \dots, \boldsymbol{X}_{i_k})$, where $g : \mathbb{R}^k \to \mathbb{R}$ is a symmetric function. Such functions are themselves symmetric. For these functions, Hoeffding introduced $f^{\subseteq S}$ (which, by symmetry, depends only on $|S|$) and proved certain inequalities (e.g., those in Exercise 8.22) relating $\mathbf{Var}[f]$ to the quantities $\|f^{\subseteq S}\|_2^2$, $\|f^{=S}\|_2^2$. Nonsymmetric functions f were considered only rarely in the subsequent three decades of statistics research. One notable exception comes in the work of Hájek [Háj68], who effectively introduced $f^{\le 1}$, known as the *Hájek projection* of f. Also, a work of Bourgain [Bou79] essentially describes the decomposition $f = \sum_k f^{=k}$. The first work that mentions the general orthogonal decomposition for not-necessarily-symmetric functions appears to be that of Efron and Stein [ES81] from the late 1970s. Efron and Stein's description is brief; the subsequent work of Karlin and Rinott [KR82] gives a more thorough development. Efron and Stein's main result was a proof of the statement $\mathbf{Var}[f] \le \mathbf{I}[f]$ for symmetric f; in the statistics literature this is known as the *Efron–Stein Inequality*. Steele [Ste86a] extended this to the case of nonsymmetric f by a simple proof that used the Fourier basis approach to orthogonal decomposition. This approach via Fourier bases originated in the work of Rubin and Vitale [RV80]; see also Takemura [Tak83] and Vitale [Vit84]. The terminology "Fourier basis" we use is not standard.

The p-biased hypercube distribution is strongly motivated by the Erdős–Rényi [ER59] theory of random graphs (see e.g., Bollobás and Riordan [BR08] for history) and by percolation theory (introduced in Broadbent and Hammersley [BH57]). Influences under the p-biased distribution – and their connection to threshold phenomena – were studied by Russo [Rus81, Rus82]. The former work proved the Margulis–Russo formula independently of Margulis, who had proven it earlier [Mar74]. Fourier analysis under the p-biased distribution seems to have been first introduced to the theoretical computer science literature by Furst, Jackson, and Smith [FJS91], who extended the LMN learning algorithm for AC^0 to this setting. Talagrand [Tal93, Tal94] developed p-biased Fourier for the study of threshold phenomena, strengthening Margulis and Russo's work and proving the KKL Theorem in the p-biased setting. Similar results were obtained by Friedgut and Kalai [FK96] using an earlier work of Bourgain, Kahn, Kalai, Linial, and Katznelson [BKK+92] that proved a version of the KKL Theorem in the setting of general product spaces. The statements about sharp thresholds for cliques and connectivity in Example 8.49 are essentially due to Matula and to Erdős–Rényi, respectively; see, e.g., Bollobás [Bol01]. Weak threshold results similar to the ones in Exercise 8.29 were proved by Bollobás and Thomason [BT87], using the Kruskal–Katona Theorem rather than the Poincaré Inequality.

Fourier analysis on finite abelian groups – and more generally, on locally compact abelian groups – is an enormous subject upon which we have touched only briefly. We cannot survey it here but refer instead to the standard textbook of Rudin [Rud62] and to the reader-friendly textbook of Terras [Ter99], which focuses on finite groups.

One of the earliest works on randomized decision tree complexity is that of Saks and Wigderson [SW86]; they proved the contents of Exercise 8.40. (We note that $\mathrm{RDT}(f)$ is usually denoted $R(f)$ in the literature, and $\mathrm{DT}(f)$ is usually denoted $D(f)$.) One basic lower bound in the area is that $\mathrm{RDT}(f) \geq \sqrt{\mathrm{DT}(f)}$ for any $f : \{-1,1\}^n \to \{-1,1\}$; in fact, this lower bound holds even for "nondeterministic decision tree complexity", as proved in [BI87, Tar89]. Yao's Conjecture is also sometimes attributed to Richard Karp. Regarding the recursive majority-of-3 function, Ravi Boppana was the first to point out that $\mathrm{RDT}(\mathrm{Maj}_3^{\otimes d}) = o(3^d)$ even though $\mathrm{DT}(\mathrm{Maj}_3^{\otimes d}) = 3^d$. Saks and Wigderson noted the bound $\mathrm{RDT}(\mathrm{Maj}_3^{\otimes d}) \leq (8/3)^d$ and also that it is not optimal. Following subsequent works [JKS03, She08] the best known upper bound is $O(2.65^d)$ [MNSX11] and the best known lower bound is $\Omega(2.55^d)$ [Leo12].

The proof of the OSSS Inequality we presented is essentially due to Lee [Lee10]; the alternate proof from Exercise 8.46 is due to Jain and Zhang [JZ11]. The Condorcet Jury Theorem (see Exercise 8.23) is from [dC85]. The Shapley value described in Exercise 8.31 was introduced by the Nobelist Shapley [Sha53]; for more, see Roth [Rot88]. Exercise 8.34 is from Blais, O'Donnell, and Wimmer [BOW10]. Exercises 8.37(*a*) and 8.45 are from Benjamini, Schramm, and Wilson [BSW05]; the term "revealment" was introduced by Schramm and Steif [SS10]. Exercise 8.47 is from [OSSS05]. Related to this, it is extremely interesting to ask whether something like the result of Exercise 8.44(*b*) holds for functions $f : \{-1,1\}^n \to [-1,1]$. It has been suggested that the answer is yes:

Aaronson–Ambainis Conjecture. [**Aar08**, **AA11**] *Let* $f : \{-1,1\}^n \to [-1,1]$. *Then* $\mathbf{MaxInf}[f] \geq \mathrm{poly}(\mathbf{Var}[f]/\deg(f))$.

If true, this conjecture would have significant consequences regarding the limitations of efficient quantum computation; see Aaronson and Ambainis [**AA11**]. The best result in the direction in the direction of the conjecture is $\mathbf{MaxInf}[f] \geq \mathrm{poly}(\mathbf{Var}[f]/2^{\deg(f)})$, due to Dinur et al. [**DFKO07**].

Chapter 9

Basics of hypercontractivity

In 1970, Bonami proved the following central result:

The Hypercontractivity Theorem. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $1 \le p \le q \le \infty$. Then $\|\mathrm{T}_\rho f\|_q \le \|f\|_p$ for $0 \le \rho \le \sqrt{\frac{p-1}{q-1}}$.*

As stated, this theorem may look somewhat opaque. In this chapter we consider some special cases of it that are easier to understand, easier to prove, and that encompass almost all of the theorem's uses. The proof of the full theorem is deferred to Chapter 10. The special cases in this chapter are the following:

Bonami Lemma. *Let $f : \{-1,1\}^n \to \mathbb{R}$ have degree k. Then $\|f\|_4 \le \sqrt{3}^k \|f\|_2$.*

The fundamental idea of this statement is that if $\boldsymbol{x} \sim \{-1,1\}^n$ and $f : \{-1,1\}^n \to \mathbb{R}$ has low degree then the random variable $f(\boldsymbol{x})$ is quite "reasonable"; e.g., it is "nicely" distributed around its mean. The Bonami Lemma has a very easy inductive proof and is already powerful enough to obtain many of the well-known applications of "hypercontractivity", including the KKL Theorem (proven at the end of this chapter) and the Invariance Principle.

(2, q)-Hypercontractivity Theorem. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $2 \le q \le \infty$. Then $\|\mathrm{T}_{1/\sqrt{q-1}} f\|_q \le \|f\|_2$. As a consequence, if f has degree at most k then $\|f\|_q \le \sqrt{q-1}^k \|f\|_2$.*

This theorem quantifies the extent to which T_ρ is a "smoothing" operator; equivalently, it gives even more control over the "reasonableness" of low-degree polynomials. Its consequences include a generalization of the Level-1

Inequality (from Chapter 5.4) to "Level-k Inequalities", as well as a Chernoff-like tail bound for low-degree polynomials of random bits.

(p,2)-Hypercontractivity Theorem. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $1 \leq p \leq 2$. Then $\|\mathrm{T}_{\sqrt{p-1}} f\|_2 \leq \|f\|_p$. Equivalently, $\mathbf{Stab}_\rho[f] \leq \|f\|_{1+\rho}^2$ for $0 \leq \rho \leq 1$.*

This theorem is actually "equivalent" to the $(2,q)$-Hypercontractivity Theorem by virtue of Hölder's inequality. When specialized to the case of $f : \{-1,1\}^n \to \{0,1\}$ it gives a precise quantification of the fact that the "noisy hypercube graph" is a "small-set expander". Qualitatively, this means that if $A \subseteq \{-1,1\}^n$ is "small", $\boldsymbol{x} \sim A$, and $\boldsymbol{y} \sim N_\rho(x)$, then $\boldsymbol{y}$ is very unlikely to be in A.

9.1. Low-degree polynomials are reasonable

As anyone who has worked in probability knows, a random variable can sometimes behave in rather "unreasonable" ways. It may be never close to its expectation. It might exceed its expectation almost always, or almost never. It might have finite 1st, 2nd, and 3rd moments, but an infinite 4th moment. All of this poor behavior can cause a lot of trouble – wouldn't it be nice to have a class of "reasonable" random variables?

A very simple condition on a random variable that guarantees some good behavior is that its 4th moment is not too large compared to its 2nd moment.

Definition 9.1. For a real number $B \geq 1$, we say that the real random variable $\boldsymbol{X}$ is *B-reasonable* if $\mathbf{E}[\boldsymbol{X}^4] \leq B\,\mathbf{E}[\boldsymbol{X}^2]^2$. (Equivalently, if $\|\boldsymbol{X}\|_4 \leq B^{1/4}\|\boldsymbol{X}\|_2$.)

The smaller B is, the more "reasonable" $\boldsymbol{X}$ is. This definition is scale-invariant (i.e., $c\boldsymbol{X}$ is B-reasonable if and only if $\boldsymbol{X}$ is, for $c \neq 0$) but not translation-invariant ($c+\boldsymbol{X}$ and $\boldsymbol{X}$ may not be equally reasonable). The latter fact can sometimes be awkward, a point we'll address further in Section 9.3. Indeed, we'll later encounter a few alternative conditions that also capture "reasonableness". For example, in Chapter 11 we'll consider the analogous 3rd moment condition, $\mathbf{E}[|\boldsymbol{X}|^3] \leq B\,\mathbf{E}[\boldsymbol{X}^2]^{3/2}$. Strictly speaking, the 4th moment condition is stronger: if $\boldsymbol{X}$ is B-reasonable, then

$$\mathbf{E}[|\boldsymbol{X}|^3] = \mathbf{E}[|\boldsymbol{X}| \cdot \boldsymbol{X}^2] \leq \sqrt{\mathbf{E}[\boldsymbol{X}^2]}\sqrt{\mathbf{E}[\boldsymbol{X}^4]} \leq \sqrt{B}\,\mathbf{E}[\boldsymbol{X}^2]^{3/2};$$

on the other hand, there exist random variables with finite 3rd moment and infinite 4th moment. However, such unusual random variables almost never arise for us, and morally speaking the 4th and 3rd moment conditions are about equally good proxies for reasonableness.

Example 9.2. If $\boldsymbol{x} \sim \{-1,1\}$ is uniformly random then $\boldsymbol{x}$ is 1-reasonable. If $\boldsymbol{g} \sim \mathrm{N}(0,1)$ is a standard Gaussian, then $\mathbf{E}[\boldsymbol{g}^4] = 3$, so $\boldsymbol{g}$ is 3-reasonable. If $\boldsymbol{u} \sim [-1,1]$ is uniform, then you can calculate that it is $\frac{9}{5}$-reasonable. In all of these examples B is a “small” constant, and we think of these random variables simply as “reasonable”. An example of an “unreasonable” random variable would be highly biased Bernoulli random variable; say, $\mathbf{Pr}[\boldsymbol{y}=1]=2^{-n}$, $\mathbf{Pr}[\boldsymbol{y}=0]=1-2^{-n}$, where n is large. This $\boldsymbol{y}$ is not B-reasonable unless $B \geq 2^n$.

Let’s give a few illustrations of why reasonable random variables are nice to work with. First, they have slightly better tail bounds than what you would get out of the Chebyshev inequality:

Proposition 9.3. *Let $\boldsymbol{X} \not\equiv 0$ be B-reasonable. Then $\mathbf{Pr}[|\boldsymbol{X}| \geq t\|\boldsymbol{X}\|_2] \leq B/t^4$ for all $t > 0$.*

Proof. This is immediate from Markov’s inequality:

$$\mathbf{Pr}[|\boldsymbol{X}| \geq t\|\boldsymbol{X}\|_2] = \mathbf{Pr}[\boldsymbol{X}^4 \geq t^4\|\boldsymbol{X}\|_2^4] \leq \frac{\mathbf{E}[\boldsymbol{X}^4]}{t^4\,\mathbf{E}[\boldsymbol{X}^2]^2} \leq \frac{B}{t^4}. \qquad \square$$

More interestingly, they also satisfy *anticoncentration* bounds; e.g., you can *upper*-bound the probability that they are near 0.

Proposition 9.4. *Let $\boldsymbol{X} \not\equiv 0$ be B-reasonable. Then $\mathbf{Pr}[|\boldsymbol{X}| > t\|\boldsymbol{X}\|_2] \geq (1-t^2)^2/B$ for all $t \in [0,1]$.*

Proof. Applying the Paley–Zygmund inequality (also called the “second moment method”) to $\boldsymbol{X}^2$, we get

$$\mathbf{Pr}[|\boldsymbol{X}| \geq t\|\boldsymbol{X}\|_2] = \mathbf{Pr}[\boldsymbol{X}^2 \geq t^2\,\mathbf{E}[\boldsymbol{X}^2]] \geq (1-t^2)^2\frac{\mathbf{E}[\boldsymbol{X}^2]^2}{\mathbf{E}[\boldsymbol{X}^4]} \geq \frac{(1-t^2)^2}{B}. \qquad \square$$

For a generalization of this proposition, see Exercise 9.12.

For a discrete random variable $\boldsymbol{X}$, a simple condition that guarantees reasonableness is that $\boldsymbol{X}$ takes on each of its values with nonnegligible probability:

Proposition 9.5. *Let $\boldsymbol{X}$ be a discrete random variable with probability mass function π. Write*

$$\lambda = \min(\pi) = \min_{x \in \mathrm{range}(\boldsymbol{X})}\{\mathbf{Pr}[\boldsymbol{X}=x]\}.$$

Then $\boldsymbol{X}$ is $(1/\lambda)$-reasonable.

Proof. Let $M = \|\boldsymbol{X}\|_\infty$. Since $\mathbf{Pr}[|\boldsymbol{X}| = M] \geq \lambda$ we get

$$\mathbf{E}[\boldsymbol{X}^2] \geq \lambda M^2 \quad\implies\quad M^2 \leq \mathbf{E}[\boldsymbol{X}^2]/\lambda.$$

On the other hand,

$$\mathbf{E}[\boldsymbol{X}^4] = \mathbf{E}[\boldsymbol{X}^2 \cdot \boldsymbol{X}^2] \leq M^2 \cdot \mathbf{E}[\boldsymbol{X}^2],$$

and thus $\mathbf{E}[\boldsymbol{X}^4] \leq (1/\lambda)\mathbf{E}[\boldsymbol{X}^2]^2$ as required. □

The converse to Proposition 9.5 is certainly not true. For example, if $\boldsymbol{X} = \frac{1}{\sqrt{n}}\boldsymbol{x}_1 + \cdots + \frac{1}{\sqrt{n}}\boldsymbol{x}_n$ where $\boldsymbol{x} \sim \{-1,1\}^n$, then $\boldsymbol{X}$ is very close to a standard Gaussian random variable (for n large) and is, unsurprisingly, 3-reasonable. On the other hand, the "λ" for this $\boldsymbol{X}$ is tiny, 2^{-n}.

This discussion raises the issue of how you might try to construct an *unreasonable* random variable out of independent uniform ±1 bits. By Proposition 9.5, at the very least you must use a lot of them. Furthermore, it also seems that they must be combined in a *high-degree* way. For example, to construct the unreasonable random variable $\boldsymbol{y}$ from Example 9.2 requires degree n: $\boldsymbol{y} = (1+\boldsymbol{x}_1)(1+\boldsymbol{x}_2)\cdots(1+\boldsymbol{x}_n)/2^n$.

Indeed, the idea that high degree is required for unreasonableness is correct, as the following crucial result shows:

The Bonami Lemma. *For each k, if $f : \{-1,1\}^n \to \mathbb{R}$ has degree at most k and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are independent, uniformly random ± 1 bits, then the random variable $f(\boldsymbol{x})$ is 9^k-reasonable, i.e.,*

$$\mathbf{E}[f^4] \leq 9^k\, \mathbf{E}[f^2]^2 \qquad \iff \qquad \|f\|_4 \leq \sqrt{3}^k \|f\|_2.$$

In other words, *low-degree polynomials of independent uniform ±1 bits are reasonable*. As we will explain later, the Bonami Lemma is a special case of more general results in the theory of "hypercontractivity". However, many key theorems using hypercontractivity – e.g., the KKL Theorem, the Invariance Principle – really need only the simple Bonami Lemma. (We should also note that the name "Bonami Lemma" is not standard; however, the result was first proved by Bonami and it's often used as a lemma, so the name fits. See the discussion in the notes in Section 9.7.)

One pleasant thing about the Bonami Lemma is that once you decide to prove it by induction on n, the proof practically writes itself. The only "non-automatic" step is an application of Cauchy–Schwarz.

Proof of the Bonami Lemma. We assume $k \geq 1$ as otherwise f must be constant and the claim is trivial. The proof is by induction on n. Again, if $n = 0$, then f must be constant and the claim is trivial. For $n \geq 1$ we can use the decomposition $f(x) = x_n \mathrm{D}_n f(x) + \mathrm{E}_n f(x)$ (Proposition 2.24), where $\deg(\mathrm{D}_n f) \leq k-1$, $\deg(\mathrm{E}_n f) \leq k$, and the polynomials $\mathrm{D}_n f(x)$ and $\mathrm{E}_n f(x)$ don't depend on x_n. For brevity we write $\boldsymbol{f} = f(\boldsymbol{x})$, $\boldsymbol{d} = \mathrm{D}_n f(\boldsymbol{x})$, and $\boldsymbol{e} = \mathrm{E}_n f(\boldsymbol{x})$. Now

$$\begin{aligned}\mathbf{E}[\boldsymbol{f}^4] &= \mathbf{E}[(\boldsymbol{x}_n\boldsymbol{d}+\boldsymbol{e})^4]\\ &= \mathbf{E}[\boldsymbol{x}_n^4\boldsymbol{d}^4] + 4\,\mathbf{E}[\boldsymbol{x}_n^3\boldsymbol{d}^3\boldsymbol{e}] + 6\,\mathbf{E}[\boldsymbol{x}_n^2\boldsymbol{d}^2\boldsymbol{e}^2] + 4\,\mathbf{E}[\boldsymbol{x}_n\boldsymbol{d}\boldsymbol{e}^3] + \mathbf{E}[\boldsymbol{e}^4]\\ &= \mathbf{E}[\boldsymbol{x}_n^4]\mathbf{E}[\boldsymbol{d}^4] + 4\,\mathbf{E}[\boldsymbol{x}_n^3]\mathbf{E}[\boldsymbol{d}^3\boldsymbol{e}] + 6\,\mathbf{E}[\boldsymbol{x}_n^2]\mathbf{E}[\boldsymbol{d}^2\boldsymbol{e}^2] + 4\,\mathbf{E}[\boldsymbol{x}_n]\mathbf{E}[\boldsymbol{d}\boldsymbol{e}^3] + \mathbf{E}[\boldsymbol{e}^4].\end{aligned}$$

In the last step we used the fact that $\boldsymbol{x}_n$ is independent of $\boldsymbol{d}$ and $\boldsymbol{e}$, since $\mathrm{D}_n f$ and $\mathrm{E}_n f$ do not depend on x_n. We now use $\mathbf{E}[\boldsymbol{x}_n] = \mathbf{E}[\boldsymbol{x}_n^3] = 0$ and $\mathbf{E}[\boldsymbol{x}_n^2] =$

$\mathbf{E}[\boldsymbol{x}_n^4] = 1$ to deduce

$$\mathbf{E}[\boldsymbol{f}^4] = \mathbf{E}[\boldsymbol{d}^4] + 6\,\mathbf{E}[\boldsymbol{d}^2\boldsymbol{e}^2] + \mathbf{E}[\boldsymbol{e}^4]. \tag{9.1}$$

A similar (and simpler) sequence of steps shows that

$$\mathbf{E}[\boldsymbol{f}^2] = \mathbf{E}[\boldsymbol{d}^2] + \mathbf{E}[\boldsymbol{e}^2]. \tag{9.2}$$

To upper-bound (9.1), recall that $\boldsymbol{d} = \mathrm{D}_n f(\boldsymbol{x})$ where $\mathrm{D}_n f$ is a multilinear polynomial of degree at most $k-1$ depending on $n-1$ variables. Thus we can apply the induction hypothesis to deduce $\mathbf{E}[\boldsymbol{d}^4] \leq 9^{k-1}\,\mathbf{E}[\boldsymbol{d}^2]^2$. Similarly, $\mathbf{E}[\boldsymbol{e}^4] \leq 9^k\,\mathbf{E}[\boldsymbol{e}^2]^2$ since $\deg(\mathrm{E}_n f) \leq k$. To bound $\mathbf{E}[\boldsymbol{d}^2\boldsymbol{e}^2]$ we apply Cauchy–Schwarz, getting $\sqrt{\mathbf{E}[\boldsymbol{d}^4]}\sqrt{\mathbf{E}[\boldsymbol{e}^4]}$ and letting us use induction again. Thus we have

$$\begin{aligned}\mathbf{E}[\boldsymbol{f}^4] &\leq 9^{k-1}\,\mathbf{E}[\boldsymbol{d}^2]^2 + 6\sqrt{9^{k-1}\,\mathbf{E}[\boldsymbol{d}^2]^2}\sqrt{9^k\,\mathbf{E}[\boldsymbol{e}^2]^2} + 9^k\,\mathbf{E}[\boldsymbol{e}^2]^2\\ &\leq 9^k\Bigl(\mathbf{E}[\boldsymbol{d}^2]^2 + 2\,\mathbf{E}[\boldsymbol{d}^2]\,\mathbf{E}[\boldsymbol{e}^2] + \mathbf{E}[\boldsymbol{e}^2]^2\Bigr) = 9^k\Bigl(\mathbf{E}[\boldsymbol{d}^2] + \mathbf{E}[\boldsymbol{e}^2]\Bigr)^2,\end{aligned}$$

where we used $9^{k-1}\,\mathbf{E}[\boldsymbol{d}^2]^2 \leq 9^k\,\mathbf{E}[\boldsymbol{d}^2]^2$. In light of (9.2), this completes the proof. □

Some aspects of the sharpness of the Bonami Lemma are explored in Exercises 9.2, 9.3, 9.37, and 9.38. Here we make one more observation. At the end of the proof we used the wasteful-looking inequality $9^{k-1}\,\mathbf{E}[\boldsymbol{d}^2]^2 \leq 9^k\,\mathbf{E}[\boldsymbol{d}^2]^2$. Tracing back through the proof, it's easy to see that it would still be valid even if we just had $\mathbf{E}[\boldsymbol{x}_i^4] \leq 9$ rather than $\mathbf{E}[\boldsymbol{x}_i^4] = 1$. For example, the Bonami Lemma holds not just if the $\boldsymbol{x}_i$'s are random bits, but if they are standard Gaussians, or are uniform on $[-1,1]$, or there are some of each. We leave the following as Exercise 9.4.

Corollary 9.6. *Let $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ be independent, not necessarily identically distributed, random variables satisfying $\mathbf{E}[\boldsymbol{x}_i] = \mathbf{E}[\boldsymbol{x}_i^3] = 0$. (This holds if, e.g., each $-\boldsymbol{x}_i$ has the same distribution as $\boldsymbol{x}_i$.) Assume also that each $\boldsymbol{x}_i$ is B-reasonable. Let $\boldsymbol{f} = F(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$, where F is a multilinear polynomial of degree at most k. Then $\boldsymbol{f}$ is* $\max(B,9)^k$*-reasonable.*

As a first application of the Bonami Lemma, let us combine it with Proposition 9.4 to show that a low-degree function is not too concentrated around its mean:

Theorem 9.7. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a nonconstant function of degree at most k; write $\mu = \mathbf{E}[f]$ and $\sigma = \sqrt{\mathbf{Var}[f]}$. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}}[|f(\boldsymbol{x}) - \mu| > \tfrac{1}{2}\sigma] \geq \tfrac{1}{16}9^{1-k}.$$

Proof. Let $g = \frac{1}{\sigma}(f - \mu)$, a function of degree at most k satisfying $\|g\|_2 = 1$. By the Bonami Lemma, g is 9^k-reasonable. The result now follows by applying Proposition 9.4 to g with $t = \frac{1}{2}$. □

Using this theorem, we can give a short proof of the FKN Theorem from Chapter 2.5: If $f : \{-1,1\}^n \to \{-1,1\}$ has $\mathbf{W}^1[f] = 1-\delta$ then f is $O(\delta)$-close to $\pm\chi_i$ for some $i \in [n]$.

Proof of the FKN Theorem. Write $\ell = f^{=1}$, so $\mathbf{E}[\ell^2] = 1-\delta$ by assumption. We may assume without loss of generality that $\delta \leq \frac{1}{1600}$. The goal of the proof is to show that $\mathbf{Var}[\ell^2]$ is small; specifically we'll show that $\mathbf{Var}[\ell^2] \leq 6400\delta$. This will complete the proof because (using Exercise 1.20 for the first equality below)

$$\tfrac{1}{2}\mathbf{Var}[\ell^2] = \sum_{i \neq j} \widehat{f}(i)^2 \widehat{f}(j)^2 = \Big(\sum_{i=1}^n \widehat{f}(i)^2\Big)^2 - \sum_{i=1}^n \widehat{f}(i)^4$$
$$= (1-\delta)^2 - \sum_{i=1}^n \widehat{f}(i)^4 \geq (1-2\delta) - \sum_{i=1}^n \widehat{f}(i)^4$$

and hence $\mathbf{Var}[\ell^2] \leq 6400\delta$ implies

$$1 - 3202\delta \leq \sum_{i=1}^n \widehat{f}(i)^4 \leq \max_i\{\widehat{f}(i)^2\} \sum_{i=1}^n \widehat{f}(i)^2 \leq \max_i\{\widehat{f}(i)^2\} \leq \max_i\{|\widehat{f}(i)|\},$$

as required.

To bound $\mathbf{Var}[\ell^2]$ we first apply Theorem 9.7 to the degree-2 function ℓ^2; this yields

$$\mathbf{Pr}\Big[\big|\ell^2 - (1-\delta)\big| \geq \tfrac{1}{2}\sqrt{\mathbf{Var}[\ell^2]}\Big] \geq \tfrac{1}{16}9^{1-2} = \tfrac{1}{144}.$$

Now suppose by way of contradiction that $\mathbf{Var}[\ell^2] > 6400\delta$; then the above implies

$$\tfrac{1}{144} \leq \mathbf{Pr}\Big[\big|\ell^2 - (1-\delta)\big| > 40\sqrt{\delta}\Big] \leq \mathbf{Pr}\Big[\big|\ell^2 - 1\big| > 39\sqrt{\delta}\Big]. \tag{9.3}$$

This says that $|\ell|$ is frequently far from 1. Since $|f| = 1$ always, we can deduce that $|f - \ell|^2$ is frequently large. More precisely, a short calculation (Exercise 9.5) shows that $(f-\ell)^2 \geq 169\delta$ whenever $|\ell^2 - 1| > 39\sqrt{\delta}$. But now (9.3) implies $\mathbf{E}[(f-\ell)^2] \geq \frac{1}{144} \cdot 169\delta > \delta$, a contradiction since $\mathbf{E}[(f-\ell)^2] = 1 - \mathbf{W}^1[f] = \delta$ by assumption. □

9.2. Small subsets of the hypercube are noise-sensitive

An immediate consequence of the Bonami Lemma is that for any $f : \{-1,1\}^n \to \mathbb{R}$ and $k \in \mathbb{N}$,

$$\|\mathrm{T}_{1/\sqrt{3}} f^{=k}\|_4 = \tfrac{1}{\sqrt{3}^k}\|f^{=k}\|_4 \leq \|f^{=k}\|_2. \tag{9.4}$$

This is a special case of the (2,4)-*Hypercontractivity Theorem* (whose name will be explained shortly), which says that the assumption of degree-k homogeneity is not necessary:

(2,4)-Hypercontractivity Theorem. *Let* $f : \{-1,1\}^n \to \mathbb{R}$. *Then*

$$\|\mathrm{T}_{1/\sqrt{3}} f\|_4 \le \|f\|_2.$$

It almost looks as though you could prove this theorem simply by summing (9.4) over k. In fact that proof strategy can be made to work given a few extra tricks (see Exercise 9.6), but it's just as easy to repeat the induction technique used for the Bonami Lemma.

Proof. We'll prove $\mathbf{E}[\mathrm{T}_{1/\sqrt{3}} f(\boldsymbol{x})^4] \le \mathbf{E}[f(\boldsymbol{x})^2]^2$ using the same induction as in the Bonami Lemma. Retaining the notation $\boldsymbol{d}$ and $\boldsymbol{e}$, and using the shorthand $\mathrm{T} = \mathrm{T}_{1/\sqrt{3}}$, we have

$$\mathrm{T}\boldsymbol{f} = \boldsymbol{x}_n \cdot \tfrac{1}{\sqrt{3}} \mathrm{T}\boldsymbol{d} + \mathrm{T}\boldsymbol{e}.$$

Similar computations to those in the Bonami Lemma proof yield

$$\begin{aligned}
\mathbf{E}[(\mathrm{T}\boldsymbol{f})^4] &= \left(\tfrac{1}{\sqrt{3}}\right)^4 \mathbf{E}[(\mathrm{T}\boldsymbol{d})^4] + 6\left(\tfrac{1}{\sqrt{3}}\right)^2 \mathbf{E}[(\mathrm{T}\boldsymbol{d})^2(\mathrm{T}\boldsymbol{e})^2] + \mathbf{E}[(\mathrm{T}\boldsymbol{e})^4] \\
&\le \mathbf{E}[(\mathrm{T}\boldsymbol{d})^4] + 2\,\mathbf{E}[(\mathrm{T}\boldsymbol{d})^2(\mathrm{T}\boldsymbol{e})^2] + \mathbf{E}[(\mathrm{T}\boldsymbol{e})^4] \\
&\le \mathbf{E}[(\mathrm{T}\boldsymbol{d})^4] + 2\sqrt{\mathbf{E}[(\mathrm{T}\boldsymbol{d})^4]}\sqrt{\mathbf{E}[(\mathrm{T}\boldsymbol{e})^4]} + \mathbf{E}[(\mathrm{T}\boldsymbol{e})^4] \\
&\le \mathbf{E}[\boldsymbol{d}^2]^2 + 2\,\mathbf{E}[\boldsymbol{d}^2]\mathbf{E}[\boldsymbol{e}^2] + \mathbf{E}[\boldsymbol{e}^2]^2 \\
&= \left(\mathbf{E}[\boldsymbol{d}^2] + \mathbf{E}[\boldsymbol{e}^2]\right)^2 = \mathbf{E}[\boldsymbol{f}^2]^2,
\end{aligned}$$

where the second inequality is Cauchy–Schwarz, the third is induction, and the final equality is a simple computation analogous to (9.2). □

The name "hypercontractivity" in this theorem describes the fact that not only is $\mathrm{T}_{1/\sqrt{3}}$ a "contraction" on $L^2(\{-1,1\}^n)$ – meaning $\|\mathrm{T}_{1/\sqrt{3}} f\|_2 \le \|f\|_2$ for all f (Exercise 2.33) – it's even a contraction when viewed as an operator from $L^2(\{-1,1\}^n)$ to $L^4(\{-1,1\}^n)$. You should think of hypercontractivity theorems as quantifying the extent to which T_ρ is a "smoothing", or "reasonable-izing" operator.

Unfortunately the quantity $\|\mathrm{T}_{1/\sqrt{3}} f\|_4$ in the (2,4)-Hypercontractivity Theorem does not have an obvious combinatorial meaning. On the other hand, the quantity

$$\|\mathrm{T}_{1/\sqrt{3}} f\|_2 = \sqrt{\langle \mathrm{T}_{1/\sqrt{3}} f, \mathrm{T}_{1/\sqrt{3}} f\rangle} = \sqrt{\langle f, \mathrm{T}_{1/\sqrt{3}} \mathrm{T}_{1/\sqrt{3}} f\rangle} = \sqrt{\mathbf{Stab}_{1/3}[f]},$$

does have a nice combinatorial meaning. And we can make this quantity appear in the Hypercontractivity Theorem via a simple trick from analysis, just using the fact that $\mathrm{T}_{1/\sqrt{3}}$ is a self-adjoint operator. We "flip the norms across 2" using Hölder's inequality:

(4/3, 2)-Hypercontractivity Theorem. *Let $f : \{-1,1\}^n \to \mathbb{R}$. Then*

$$\|\mathrm{T}_{1/\sqrt{3}} f\|_2 \le \|f\|_{4/3};$$

i.e.,

$$\mathbf{Stab}_{1/3}[f] \le \|f\|_{4/3}^2. \tag{9.5}$$

Proof. Writing $\mathrm{T} = \mathrm{T}_{1/\sqrt{3}}$ for brevity we have

$$\|\mathrm{T}f\|_2^2 = \langle \mathrm{T}f, \mathrm{T}f\rangle = \langle f, \mathrm{T}\mathrm{T}f\rangle \le \|f\|_{4/3}\|\mathrm{T}\mathrm{T}f\|_4 \le \|f\|_{4/3}\|\mathrm{T}f\|_2 \tag{9.6}$$

by Hölder's inequality and the (2,4)-Hypercontractivity Theorem. Dividing through by $\|\mathrm{T}f\|_2$ (which we may assume is nonzero) completes the proof. □

In the inequality (9.5) the left-hand side is a natural quantity. The right-hand side is just 1 when $f : \{-1,1\}^n \to \{-1,1\}$, which is not very interesting. But if we instead look at $f : \{-1,1\}^n \to \{0,1\}$ we get something very interesting:

Corollary 9.8. *Let $A \subseteq \{-1,1\}^n$ have volume α; i.e., let $1_A : \{-1,1\}^n \to \{0,1\}$ satisfy $\mathbf{E}[1_A] = \alpha$. Then*

$$\mathbf{Stab}_{1/3}[1_A] = \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim\{-1,1\}^n \\ \boldsymbol{y}\sim N_{1/3}(\boldsymbol{x})}}[\boldsymbol{x} \in A, \boldsymbol{y} \in A] \le \alpha^{3/2}.$$

Equivalently (for $\alpha > 0$),

$$\mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim A \\ \boldsymbol{y}\sim N_{1/3}(\boldsymbol{x})}}[\boldsymbol{y} \in A] \le \alpha^{1/2}.$$

Proof. This is immediate from inequality (9.5), since

$$\|1_A\|_{4/3}^2 = \left(\mathop{\mathbf{E}}_{\boldsymbol{x}}[|1_A(\boldsymbol{x})|^{4/3}]^{3/4}\right)^2 = \mathop{\mathbf{E}}_{\boldsymbol{x}}[1_A(\boldsymbol{x})]^{3/2} = \alpha^{3/2}.$$ □

See Section 9.5 for the generalization of this corollary to noise rates other than 1/3.

Example 9.9. Assume $\alpha = 2^{-k}$, $k \in \mathbb{N}^+$, and A is a subcube of codimension k; e.g., $1_A : \mathbb{F}_2^n \to \{0,1\}$ is the logical AND function on the first k coordinates. For every $x \in A$, when we form $\boldsymbol{y} \sim N_{1/3}(x)$ we'll have $\boldsymbol{y} \in A$ if and only if the first k coordinates of x do not change, which happens with probability $(2/3)^k = (2/3)^{\log(1/\alpha)} = \alpha^{\log(3/2)} \approx \alpha^{.585} \le \alpha^{1/2}$. In fact, the bound $\alpha^{1/2}$ in Corollary 9.8 is essentially sharp when A is a Hamming ball; see Exercise 9.24.

We can phrase Corollary 9.8 in terms of the *expansion* in a certain graph:

Definition 9.10. For $n \in \mathbb{N}^+$ and $\rho \in [-1,1]$, the n-dimensional *ρ-stable hypercube graph* is the edge-weighted, complete directed graph on vertex set $\{-1,1\}^n$ in which the weight on directed edge $(x,y) \in \{-1,1\}^n \times \{-1,1\}^n$ is equal to $\mathbf{Pr}[(\boldsymbol{x},\boldsymbol{y}) = (x,y)]$ when $(\boldsymbol{x},\boldsymbol{y})$ is a ρ-correlated pair. If $\rho = 1 - 2\delta$ for $\delta \in [0,1]$, we also call this the *δ-noisy hypercube graph*. Here the weight on (x,y) is

$\mathbf{Pr}[(\boldsymbol{x},\boldsymbol{y}) = (x,y)]$ where $\boldsymbol{x} \sim \{-1,1\}^n$ is uniform and $\boldsymbol{y}$ is formed from $\boldsymbol{x}$ by negating each coordinate independently with probability δ.

Remark 9.11. The edge weights in this graph are nonnegative and sum to 1. The graph is also "regular" in the sense that for each $x \in \{-1,1\}^n$ the sum of all the edge weight leaving (or entering) x is 2^{-n}. You can also consider the graph to be undirected, since the weight on (x,y) is the same as the weight on (y,x); in this viewpoint, the weight on the undirected edge (x,y) would be $2^{1-n}\delta^{\Delta(x,y)}(1-\delta)^{n-\Delta(x,y)}$. In fact, the graph is perhaps best thought of as the discrete-time Markov chain on state space $\{-1,1\}^n$ in which a step from state $x \in \{-1,1\}^n$ consists of moving to state $\boldsymbol{y} \sim N_\rho(x)$. This is a reversible chain with the uniform stationary distribution. Each discrete step is equivalent to running the "usual" *continuous-time* Markov chain on the hypercube for time $t = \ln(1/\rho)$ (assuming $\rho \in [0,1]$).

With this definition in place, we can see Corollary 9.8 as saying that the 1/3-stable (equivalently, 1/3-noisy) hypercube graph is a "small-set expander": given any small α-fraction of the vertices A, almost all of the edge weight touching A is on its boundary. More precisely, if we choose a random vertex $\boldsymbol{x} \in A$ and take a random edge out of $\boldsymbol{x}$ (with probability proportional to its edge weight), we end up outside A with probability at least $1-\alpha^{1/2}$. You can compare this with the discussion surrounding the Level-1 Inequality in Section 5.4, which is the analogous statement for the ρ-stable hypercube graph "in the limit $\rho \to 0^+$". The appropriate statement for general ρ is appears in Section 9.5 as the "Small-Set Expansion Theorem".

Corollary 9.8 would apply equally well if 1_A were replaced by a function $g : \{-1,1\}^n \to \{-1,0,1\}$, with α denoting $\mathbf{Pr}[g \neq 0] = \mathbf{E}[|g|] = \mathbf{E}[g^2]$. This situation occurs naturally when $g = \mathrm{D}_i f$ for some Boolean-valued $f : \{-1,1\}^n \to \{-1,1\}$. In this case $\mathbf{Stab}_{1/3}[g] = \mathbf{Inf}_i^{(1/3)}[f]$, the 1/3-stable influence of i on f. We conclude that for a Boolean-valued function, if the influence of i is small then its 1/3-stable influence is much smaller:

Corollary 9.12. *Let $f : \{-1,1\}^n \to \{-1,1\}$. Then $\mathbf{Inf}_i^{(1/3)}[f] \leq \mathbf{Inf}_i[f]^{3/2}$ for all i.*

We remark that the famous KKL Theorem (stated in Chapter 4.2) more or less follows by summing the above inequality over $i \in [n]$; if you're impatient to see its proof you can skip directly to Section 9.6 now.

Let's take one more look at the "small-set expansion result", Corollary 9.8. Since noise stability roughly measures how "low" a function's Fourier weight is, this corollary implies that a function $f : \{-1,1\}^n \to \{0,1\}$ with small mean α cannot have much of its Fourier weight at low degree. More precisely, for any

$k \in \mathbb{N}$ we have

$$\alpha^{3/2} \geq \mathbf{Stab}_{1/3}[f] \geq (1/3)^k \mathbf{W}^{\leq k}[f] \quad \Longrightarrow \quad \mathbf{W}^{\leq k}[f] \leq 3^k \alpha^{3/2}. \tag{9.7}$$

For $k = 1$ this gives $\mathbf{W}^{\leq 1}[f] \leq 3\alpha^{3/2}$, which is nontrivial but not as strong as the Level-1 Inequality from Section 5.4. But (9.7) also gives us "level-k inequalities" for larger values of k. For example,

$$\mathbf{W}^{\leq .25\log(1/\alpha)}[f] \leq \alpha^{-.25\log 3 + 3/2} \leq \alpha^{1.1} \ll \alpha = \|f\|_2^2;$$

i.e., almost all of f's Fourier weight is above degree $.25\log(1/\alpha)$. We will give slightly improved versions of these level-k inequalities in Section 9.5.

9.3. $(2,q)$- and $(p,2)$-hypercontractivity for a single bit

Although you can get a lot of mileage out of studying the 4-norm of random variables, it's also natural to consider other norms. For example, we would get improved versions of our concentration and anticoncentration results, Propositions 9.3 and 9.4, if we could bound the higher norms of a random variable in terms of its 2-norm. As we'll see, we can also get stronger "level-k inequalities" by bounding the $(2+\epsilon)$-norm of a Boolean function for small $\epsilon > 0$.

We started with the 4-norm due to the simplicity of the proofs of the Bonami Lemma and the $(2,4)$-Hypercontractivity Theorem. To generalize these results to other norms it's a bit more elegant to work with the latter. Partly this is because it's "formally stronger" (see Theorem 9.21). But the main reason is that the hypercontractivity version alleviates the inelegant issue that being "B-reasonable" is not translation-invariant. Thus instead of generalizing the condition that $\|\rho \boldsymbol{X}\|_4 \leq \|\boldsymbol{X}\|_2$ ("$\boldsymbol{X}$ is ρ^{-4}-reasonable") we'll generalize the condition that $\|a + \rho b\boldsymbol{X}\|_4 \leq \|a + b\boldsymbol{X}\|_2$ (cf. the $n = 1$ case of the $(2,4)$-Hypercontractivity Theorem).

Definition 9.13. Let $1 \leq p \leq q \leq \infty$ and let $0 \leq \rho < 1$. We say that a real random variable $\boldsymbol{X}$ (with $\|\boldsymbol{X}\|_q < \infty$) is (p,q,ρ)-*hypercontractive* if

$$\|a + \rho b\boldsymbol{X}\|_q \leq \|a + b\boldsymbol{X}\|_p \quad \text{for all constants } a, b \in \mathbb{R}.$$

Remark 9.14. By homogeneity, it suffices to check the condition for $a = 1$, $b \in \mathbb{R}$ or for $a \in \mathbb{R}$, $b = 1$ (cf. Exercise 9.9(a)). It's also true (Exercise 9.11) that if $\boldsymbol{X}$ is (p,q,ρ)-hypercontractive then it is (p,q,ρ')-hypercontractive for $\rho' < \rho$ as well.

In Exercise 9.10 you will show that if $\boldsymbol{X}$ is hypercontractive then $\mathbf{E}[\boldsymbol{X}]$ must be 0. Thus hypercontractivity, like reasonableness, is not a translation-invariant notion. Nevertheless, the fact that the definition involves translation by an arbitrary a greatly facilitates proofs by induction. For example, an elegant property we gain from the definition is the following (Exercise 10.2):

Proposition 9.15. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent (p,q,ρ)-hypercontractive random variables. Then $\boldsymbol{X}+\boldsymbol{Y}$ is also (p,q,ρ)-hypercontractive.*

The $n=1$ case of our $(2,4)$-Hypercontractivity Theorem precisely says that a single uniformly random ±1 bit $\boldsymbol{x}$ is $(2,4,1/\sqrt{3})$-hypercontractive; the $(4/3,2)$-Hypercontractivity Theorem says that $\boldsymbol{x}$ is also $(4/3,2,1/\sqrt{3})$-hypercontractive. We'll spend the remainder of this section generalizing these facts to $(2,q,\rho)$- and $(p,2,\rho)$-hypercontractivity for other values of p and q. We remark that in our study of hypercontractivity we'll focus mainly on the cases of $p=2$ or $q=2$. The study of hypercontractivity for $p,q\neq 2$ and for random variables other than uniform ±1 bits is deferred to Chapter 10.

We now consider hypercontractivity of a uniformly random ±1 bit $\boldsymbol{x}$. We know that $\boldsymbol{x}$ is $(2,q,1/\sqrt{3})$-hypercontractive for $q=4$; what about other values of q? Things are most pleasant when q is an even integer because then you don't need to take the absolute value when computing $\|a+\rho b\boldsymbol{X}\|_q$. So let's try $q=6$.

Proposition 9.16. *For $\boldsymbol{x}$ a uniform ±1 bit, we have $\|a+\rho b\boldsymbol{x}\|_6 \le \|a+b\boldsymbol{x}\|_2$ for all $a,b\in\mathbb{R}$ if (and only if) $\rho\le 1/\sqrt{5}$. That is, $\boldsymbol{x}$ is $(2,6,1/\sqrt{5})$-hypercontractive.*

Proof. Raising the inequality to the 6th power, we need to show

$$\mathbf{E}[(a+\rho b\boldsymbol{x})^6] \le \mathbf{E}[(a+b\boldsymbol{x})^2]^3. \tag{9.8}$$

The result is trivial when $a=0$; otherwise, we may assume $a=1$ by homogeneity. We expand both quantities inside expectations and use the fact that $\mathbf{E}[\boldsymbol{x}^k]$ is 0 when k is odd and 1 when k is even. Thus (9.8) is equivalent to

$$1+15\rho^2b^2+15\rho^4b^4+\rho^6b^6 \le (1+b^2)^3 = 1+3b^2+3b^4+b^6. \tag{9.9}$$

Comparing the two sides term-by-term we see that the coefficient on b^2 is the limiting factor: in order for (9.9) to hold for all $b\in\mathbb{R}$ it is sufficient that $15\rho^2\le 3$; i.e., $\rho\le 1/\sqrt{5}$. By considering $b\to 0$ it's also easy to see that this condition is necessary. □

If you repeat this analysis for the case of $q=8$ you'll find that again the limiting factor is the coefficient on b^2, and that $\boldsymbol{x}$ is $(2,8,\rho)$-hypercontractive if (and only if) $\binom{8}{2}\rho^2\le\binom{4}{1}$; i.e., $\rho\le 1/\sqrt{7}$. In light of this it is natural to guess that the following is true:

Theorem 9.17. *Let $\boldsymbol{x}$ be a uniform ±1 bit and let $q\in(2,\infty]$. Then $\|a+\rho b\boldsymbol{x}\|_q\le \|a+b\boldsymbol{x}\|_2$ for all $a,b\in\mathbb{R}$ assuming $\rho\le 1/\sqrt{q-1}$.*

Equivalent statements are that $\|a+(1/\sqrt{q-1})b\boldsymbol{x}\|_q^2\le a^2+b^2$, that $\boldsymbol{x}$ is $(2,q,1/\sqrt{q-1})$-hypercontractive, and that $\|\mathrm{T}_{1/\sqrt{q-1}}f\|_q\le\|f\|_2$ holds for any $f:\{-1,1\}\to\mathbb{R}$.

For q an even integer it is not hard (see Exercise 9.36) to prove Theorem 9.17 just as we did for $q = 6$. Indeed, the proof works even under more general moment conditions on $\boldsymbol{x}$, as in Corollary 9.6. Unfortunately, obtaining Theorem 9.17 for all real $q > 2$ takes some more tricks. A natural idea is to try forging ahead as in Proposition 9.16, using the series expansions for $(1+\rho bx)^q$ and $(1+b^2)^{q/2}$ provided by the Generalized Binomial Theorem. However, even when $|b| < 1$ (so that convergence is not an issue) there is a difficulty because the coefficients in the expansion of $(1+b^2)^{q/2}$ are sometimes negative.

Luckily, this issue of negative coefficients in the series expansion goes away if you try to prove the analogous $(p,2,\rho)$-hypercontractivity statement. Thus the slick proof of Theorem 9.17 proceeds by first proving that statement, then "flipping the norms across 2".

Theorem 9.18. *Let $\boldsymbol{x}$ be a uniform ±1 bit and let $1 \le p < 2$. Then $\|a+\rho b\boldsymbol{x}\|_2 \le \|a+b\boldsymbol{x}\|_p$ for all $a, b \in \mathbb{R}$ assuming $0 \le \rho \le \sqrt{p-1}$. That is, $\boldsymbol{x}$ is $(p,2,\sqrt{p-1})$-hypercontractive.*

Proof. By Remark 9.14 we may assume $a = 1$ and $\rho = \sqrt{p-1}$. By Exercise 9.7 we may also assume without loss of generality that $1+bx \ge 0$ for $x \in \{-1,1\}$; i.e., that $|b| \le 1$. It then suffices to prove the result for all $|b| < 1$ because the $|b| = 1$ case follows by continuity. Writing $b = \epsilon$ for the sake of intuition, we need to show

$$\begin{aligned} &\|1+\sqrt{p-1}\cdot\epsilon\boldsymbol{x}\|_2^p \le \|1+\epsilon\boldsymbol{x}\|_p^p \\ \iff\quad &\mathbf{E}[(1+\sqrt{p-1}\cdot\epsilon\boldsymbol{x})^2]^{p/2} \le \mathbf{E}[(1+\epsilon\boldsymbol{x})^p]. \end{aligned} \tag{9.10}$$

Here we were able to drop the absolute value on the right-hand side because $|\epsilon| < 1$. The left-hand side of (9.10) is

$$(1+(p-1)\epsilon^2)^{p/2} \le 1 + \tfrac{p(p-1)}{2}\epsilon^2, \tag{9.11}$$

where we used the inequality $(1+t)^\theta \le 1+\theta t$ for $t \ge 0$ and $0 \le \theta \le 1$ (easily proved by comparing derivatives in t). As for the right-hand side of (9.10), since $|\epsilon\boldsymbol{x}| < 1$ we may use the Generalized Binomial Theorem to show it equals

$$\begin{aligned} &\mathbf{E}\left[1 + p\epsilon\boldsymbol{x} + \tfrac{p(p-1)}{2!}\epsilon^2\boldsymbol{x}^2 + \tfrac{p(p-1)(p-2)}{3!}\epsilon^3\boldsymbol{x}^3 + \tfrac{p(p-1)(p-2)(p-3)}{4!}\epsilon^4\boldsymbol{x}^4 + \cdots\right] \\ &= 1 + p\epsilon\,\mathbf{E}[\boldsymbol{x}] + \tfrac{p(p-1)}{2!}\epsilon^2\,\mathbf{E}[\boldsymbol{x}^2] + \tfrac{p(p-1)(p-2)}{3!}\epsilon^3\,\mathbf{E}[\boldsymbol{x}^3] + \tfrac{p(p-1)(p-2)(p-3)}{4!}\epsilon^4\,\mathbf{E}[\boldsymbol{x}^4] + \cdots \\ &= 1 + \tfrac{p(p-1)}{2}\epsilon^2 + \tfrac{p(p-1)(p-2)(p-3)}{4!}\epsilon^4 + \tfrac{p(p-1)(p-2)(p-3)(p-4)(p-5)}{6!}\epsilon^6 + \cdots. \end{aligned}$$

In light of (9.11), to verify (9.10) it suffices to note that each "post-quadratic" term above,

$$\tfrac{p(p-1)(p-2)(p-3)\cdots(p-(2k+1))}{(2k)!}\epsilon^{2k},$$

is nonnegative. This follows from $1 \le p \le 2$: the numerator has two positive factors and an even number of negative factors. □

To deduce Theorem 9.17 from Theorem 9.18 we again just need to flip the norms across 2 using the fact that T_ρ is self-adjoint. This is accomplished by taking $\Omega = \{-1,1\}$, $\pi = \pi_{1/2}$, $q = 2$, $T = \mathrm{T}_{\sqrt{p-1}}$, and $C = 1$ in the following proposition (and noting that $1/\sqrt{p'-1} = \sqrt{p-1}$):

Proposition 9.19. *Let T be a self-adjoint operator on $L^2(\Omega,\pi)$, let $1 \le p, q \le \infty$, and let p', q' be their conjugate Hölder indices. Assume $\|Tf\|_q \le C\|f\|_p$ for all f. Then $\|Tg\|_{p'} \le C\|g\|_{q'}$ for all g.*

Proof. This follows from

$$\|Tg\|_{p'} = \sup_{\|f\|_p=1} \langle f, Tg\rangle = \sup_{\|f\|_p=1} \langle Tf, g\rangle \le \sup_{\|f\|_p=1} \|Tf\|_q\|g\|_{q'} \le C\|g\|_{q'},$$

where the first equality is the sharpness of Hölder's inequality, the second equality holds because T is self-adjoint, the subsequent inequality is Hölder's, and the final inequality uses the hypothesis $\|Tf\|_q \le C\|f\|_p$. □

At this point we have established that if $\boldsymbol{x}$ is a uniform ±1 bit, then it is $(2, q, 1/\sqrt{q-1})$-hypercontractive and $(p, 2, \sqrt{p-1})$-hypercontractive. In the next section we will give a very simple induction which transforms these facts into the full $(2,q)$- and $(p,2)$-Hypercontractivity Theorems stated at the beginning of the chapter.

9.4. Two-function hypercontractivity and induction

At this point we have established that if $f : \{-1,1\} \to \mathbb{R}$ then for any $p \le 2 \le q$,

$$\|\mathrm{T}_{\sqrt{p-1}} f\|_2 \le \|f\|_p, \qquad \|\mathrm{T}_{1/\sqrt{q-1}} f\|_q \le \|f\|_2.$$

We would like to extend these facts to the case of general $f : \{-1,1\}^n \to \mathbb{R}$; i.e., establish the $(p,2)$- and $(2,q)$-Hypercontractivity Theorems stated at the beginning of the chapter. A natural approach is induction.

In analysis of Boolean functions, there are two methods for proving statements about $f : \{-1,1\}^n \to \mathbb{R}$ by induction on n. One method, which might be called "induction by derivatives", uses the decomposition $f(x) = x_n \mathrm{D}_n f(x) + \mathrm{E}_n f(x)$. We saw this approach in our inductive proof of the Bonami Lemma. The other method, which might be called "induction by restrictions", goes via the subfunctions $f_{\pm1}$ obtained by restricting the nth coordinate of f to ±1. We saw this approach in our proof of the OSSS Inequality in Chapter 8.6. In both methods we reduce inductively from one function f to two functions: either $\mathrm{D}_n f$ and $\mathrm{E}_n f$, or f_{-1} and f_{+1}. Because of this, when trying to prove a fact by induction on n it's often helpful to try proving a generalized fact about *two* functions. Our proof of the OSSS Inequality gives a good example this technique.

So to facilitate induction, let's find a two-function version of the hypercontractivity statements we've proven so far. Perhaps the most natural statement we've seen is the noise-stability rephrasing of the (4/3, 2)-Hypercontractivity Theorem, namely $\mathbf{Stab}_{1/3}[f] \le \|f\|_{4/3}^2$. At least in the case $n = 1$, our work in the previous section (Theorem 9.18) generalizes this to $\mathbf{Stab}_{p-1}[f] \le \|f\|_p^2$ for $1 \le p \le 2$. I.e.,

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})f(\boldsymbol{y})] \le \|f\|_{1+\rho}^2$$

for $0 \le \rho \le 1$. Looking at this, you might naturally guess a (correct) generalization for two functions $f, g : \{-1,1\}^n \to \mathbb{R}$, namely

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] \le \|f\|_{1+\rho}\|g\|_{1+\rho}. \tag{9.12}$$

We have a nice interpretation of this inequality when $f, g : \{-1,1\}^n \to \{0,1\}$ are indicators of subsets $A, B \subseteq \{-1,1\}^n$ as in Corollary 9.8; it gives an upper bound on the probability of going from A to B in one step on the ρ-stable hypercube graph. This bound is sharp when A and B have the same volume, but for A and B of different sizes you might imagine it's helpful to measure f and g by different norms in (9.12). To see what we can expect, let's break up the ρ-correlation in (9.12) into two parts; say, write

$$\rho = \sqrt{rs}, \qquad 0 \le r, s \le 1,$$

and use

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \sqrt{rs}\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] = \mathbf{E}[\mathrm{T}_{\sqrt{r}}f \cdot \mathrm{T}_{\sqrt{s}}g].$$

Then Cauchy–Schwarz implies

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] = \mathbf{E}[\mathrm{T}_{\sqrt{r}}f \cdot \mathrm{T}_{\sqrt{s}}g] \le \|\mathrm{T}_{\sqrt{r}}f\|_2\|\mathrm{T}_{\sqrt{s}}g\|_2 \le \|f\|_{1+r}\|g\|_{1+s}, \tag{9.13}$$

where the last step used $(p,2)$-hypercontractivity – which we have so far only proven in the case $n = 1$ (Theorem 9.18). The inequality (9.13), restated below, is precisely the desired two-function version of the $(2,q)$- and $(p,2)$-Hypercontractive Theorems.

(Weak) Two-Function Hypercontractivity Theorem. *Let $f, g : \{-1,1\}^n \to \mathbb{R}$, let $0 \le r, s \le 1$, and assume $0 \le \rho \le \sqrt{rs} \le 1$. Then*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] \le \|f\|_{1+r}\|g\|_{1+s}.$$

We call this the "Weak" Two-Function Hypercontractivity Theorem because the hypothesis $r, s \le 1$ is not actually necessary; see Chapter 10.1. As mentioned, we have so far established this theorem in the case $n = 1$. However, the beauty of hypercontractivity in this form is that it extends to general n by an almost trivial induction. The form of the induction is "induction by restrictions". (It's also possible – but a little trickier – to extend the $(2,q)$-Hypercontractivity Theorem from $n=1$ to general n via "induction by derivatives"; see Exercise 9.16.) For future use, we will write the induction in more general notation.

Two-Function Hypercontractivity Induction Theorem. *Let $0 \le \rho \le 1$ and assume that*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] \le \|f\|_p \|g\|_q$$

holds for every $f,g \in L^2(\Omega,\pi)$. Then the inequality also holds for every $f,g \in L^2(\Omega^n,\pi^{\otimes n})$.

Proof. The proof is by induction on n, with the $n=1$ case holding by assumption. For $n>1$, let $f,g \in L^2(\Omega^n,\pi^{\otimes n})$ and let $(\boldsymbol{x},\boldsymbol{y})$ denote a ρ-correlated pair under $\pi^{\otimes n}$. We'll use the notation $\boldsymbol{x}=(\boldsymbol{x}',\boldsymbol{x}_n)$ where $\boldsymbol{x}'=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_{n-1})$, and similar notation for $\boldsymbol{y}$. Note that $(\boldsymbol{x}',\boldsymbol{y}')$ and $(\boldsymbol{x}_n,\boldsymbol{y}_n)$ are both ρ-correlated pairs (of length $n-1$ and 1, respectively). We'll also write $f_{x_n} = f_{[n-1]|x_n}$ for the restriction of f in which the last coordinate is fixed to value x_n, and similarly for g. Now

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})}[f(\boldsymbol{x})g(\boldsymbol{y})] = \mathop{\mathbf{E}}_{(\boldsymbol{x}_n,\boldsymbol{y}_n)}\mathop{\mathbf{E}}_{(\boldsymbol{x}',\boldsymbol{y}')}[f_{\boldsymbol{x}_n}(\boldsymbol{x}')g_{\boldsymbol{y}_n}(\boldsymbol{y}')] \le \mathop{\mathbf{E}}_{(\boldsymbol{x}_n,\boldsymbol{y}_n)}[\|f_{\boldsymbol{x}_n}\|_p\|g_{\boldsymbol{y}_n}\|_q]$$

by induction. If we write $F \in L^2(\Omega,\pi)$ for the function $x_n \mapsto \|f_{x_n}\|_p$ and similarly write $G(y_n)=\|g_{y_n}\|_q$, then we may continue the above as

$$\mathop{\mathbf{E}}_{(\boldsymbol{x}_n,\boldsymbol{y}_n)}[\|f_{\boldsymbol{x}_n}\|_p\|g_{\boldsymbol{y}_n}\|_q] = \mathop{\mathbf{E}}_{(\boldsymbol{x}_n,\boldsymbol{y}_n)}[F(\boldsymbol{x}_n)G(\boldsymbol{y}_n)] \le \|F\|_{p,\boldsymbol{x}_n}\|G\|_{q,\boldsymbol{x}_n},$$

where we used the base case of the induction. Finally,

$$\|F\|_{p,\boldsymbol{x}_n} = \mathop{\mathbf{E}}_{\boldsymbol{x}_n}[|F(\boldsymbol{x}_n)|^p]^{1/p} = \mathop{\mathbf{E}}_{\boldsymbol{x}_n}[\|f_{\boldsymbol{x}_n}\|_p^p]^{1/p} = \big(\mathop{\mathbf{E}}_{\boldsymbol{x}_n}\mathop{\mathbf{E}}_{\boldsymbol{x}'}|f_{\boldsymbol{x}_n}(\boldsymbol{x}')|^p]\big)^{1/p} = \|f\|_p$$

by definition, and similarly for $\|G\|_{q,\boldsymbol{x}_n}$. Thus we have established $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{y})] \le \|f\|_p\|g\|_q$, completing the induction. □

Remark 9.20. More generally, if we assume the inequality holds over each of $(\Omega_1,\pi_1),\dots,(\Omega_n,\pi_n)$, then it also holds over $(\Omega_1\times\cdots\times\Omega_n, \pi_1\otimes\cdots\otimes\pi_n)$; the only change needed to the proof is notational.

At this point, we have fully established the Weak Two-Function Hypercontractivity Theorem. By taking $g=f$ and $r=s=\rho$ in the theorem we obtain the full $(p,2)$-Hypercontractivity Theorem stated at the beginning of the chapter. Finally, by applying Proposition 9.19 we also obtain the $(2,q)$-Hypercontractivity Theorem for all $f:\{-1,1\}^n \to \mathbb{R}$.

9.5. Applications of hypercontractivity

With the $(2,q)$- and $(p,2)$-Hypercontractivity Theorems in hand, let's revisit some applications we saw in Sections 9.1 and 9.2. We begin by deducing a generalization of the Bonami Lemma:

Theorem 9.21. *Let $f : \{-1,1\}^n \to \mathbb{R}$ have degree at most k. Then $\|f\|_q \leq \sqrt{q-1}^k \|f\|_2$ for any $q \geq 2$.*

Proof. We have

$$\|f\|_q^2 = \|\mathrm{T}_{1/\sqrt{q-1}}\mathrm{T}_{\sqrt{q-1}} f\|_q^2 \leq \|\mathrm{T}_{\sqrt{q-1}} f\|_2^2$$

using the $(2,q)$-Hypercontractivity Theorem. (Here we are extending the definition of T_ρ to $\rho > 1$ via $\mathrm{T}_\rho f = \sum_j \rho^j f^{=j}$; see also Remark 8.29.) The result now follows since

$$\|\mathrm{T}_{\sqrt{q-1}} f\|_2^2 = \sum_{j=0}^{k} (q-1)^j \mathbf{W}^j[f] \leq (q-1)^k \sum_{j=0}^{k} \mathbf{W}^j[f] = (q-1)^k \|f\|_2^2. \qquad \square$$

Using a trick similar to the one in our proof of the $(4/3,2)$-Hypercontractivity Theorem you can use this to deduce $\|f\|_2 \leq (1/\sqrt{p-1})^k \|f\|_p$ when f has degree k for any $1 \leq p \leq 2$; see Exercise 9.14. However, a different trick yields a strictly better result, including a finite bound for $p = 1$:

Theorem 9.22. *Let $f : \{-1,1\}^n \to \mathbb{R}$ have degree at most k. Then $\|f\|_2 \leq e^k \|f\|_1$. More generally, for $1 \leq p \leq 2$ it holds that $\|f\|_2 \leq (e^{\frac{2}{p}-1})^k \|f\|_p$.*

Proof. We prove the statement about the 1-norm, leaving the case of general $1 \leq p \leq 2$ to Exercise 9.15. For $\epsilon > 0$, let $0 < \theta < 1$ be the solution of $\frac{1}{2} = \frac{\theta}{1} + \frac{1-\theta}{2+\epsilon}$ (namely, $\theta = \frac{1}{2}\frac{\epsilon}{1+\epsilon}$). Applying the general version of Hölder's inequality and then Theorem 9.21, we get

$$\|f\|_2 \leq \|f\|_{2+\epsilon}^{1-\theta} \|f\|_1^{\theta} \leq \sqrt{1+\epsilon}^{k(1-\theta)} \|f\|_2^{1-\theta} \|f\|_1^{\theta}.$$

Dividing by $\|f\|_2^{1-\theta}$ (which we may assume is nonzero) and then raising the result to the power of $1/\theta$ yields

$$\|f\|_2 \leq \left((1+\epsilon)^{\frac{1-\theta}{2\theta}}\right)^k \|f\|_1 = \left((1+\epsilon)^{\frac{1}{\epsilon}+\frac{1}{2}}\right)^k \|f\|_1.$$

The result follows by taking the limit as $\epsilon \to 0$. $\square$

In the linear case of $k = 1$, Theorems 9.21 and 9.22 taken together show that $c_p \|\sum_i a_i \boldsymbol{x}_i\|_2 \leq \|\sum_i a_i \boldsymbol{x}_i\|_p \leq C_p \|\sum_i a_i \boldsymbol{x}_i\|_2$ for some constants $0 < c_p < C_p$ depending only on $p \in [1,\infty)$. This fact is known as Khintchine's Inequality.

Theorem 9.21 can be used to get a strong concentration bound for degree-k Boolean functions. Chernoff tells us that the probability a linear form $\sum a_i \boldsymbol{x}_i$

exceeds t standard deviations decays like $\exp(-\Theta(t^2))$. The following theorem generalizes this to degree-k forms, with decay $\exp(-\Theta(t^{2/k}))$:

Theorem 9.23. *Let $f : \{-1,1\}^n \to \mathbb{R}$ have degree at most k. Then for any $t \geq \sqrt{2e}^k$ we have*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}[|f(\boldsymbol{x})| \geq t\|f\|_2] \leq \exp\left(-\tfrac{k}{2e}t^{2/k}\right).$$

Proof. We may assume $\|f\|_2 = 1$ without loss of generality. Let $q \geq 2$ be a parameter to be chosen later. By Markov's inequality,

$$\mathbf{Pr}[|f(\boldsymbol{x})| \geq t] = \mathbf{Pr}[|f(\boldsymbol{x})|^q \geq t^q] \leq \frac{\mathbf{E}[|f(\boldsymbol{x})|^q]}{t^q}.$$

By Theorem 9.21 we have

$$\mathbf{E}[|f(\boldsymbol{x})|^q] \leq (\sqrt{q-1}^k)^q \|f\|_2^q = (q-1)^{(k/2)q} \leq q^{(k/2)q}.$$

Thus $\mathbf{Pr}[|f(\boldsymbol{x})| \geq t] \leq (q^{k/2}/t)^q$. It's not hard to see that the q that minimizes this expression should be just slightly less than $t^{2/k}$. Specifically, by choosing $q = t^{2/k}/e \geq 2$ we get

$$\mathbf{Pr}[|f(\boldsymbol{x})| \geq t] \leq \exp(-(k/2)q) = \exp\left(-\tfrac{k}{2e}t^{2/k}\right)$$

as claimed. □

We can use Theorem 9.22 to get a "one-sided" analogue of Theorem 9.7, showing that a low-degree function exceeds its mean with noticeable probability:

Theorem 9.24. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a nonconstant function of degree at most k. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[f(\boldsymbol{x}) > \mathbf{E}[f]\right] \geq \tfrac{1}{4}e^{-2k}.$$

Proof. We may assume $\mathbf{E}[f] = 0$ without loss of generality. We then have

$$\tfrac{1}{2}\|f\|_1 = \tfrac{1}{2}\left(\mathbf{E}[f\cdot\mathbf{1}_{\{f(\boldsymbol{x})>0\}}] - \mathbf{E}[f\cdot(1-\mathbf{1}_{\{f(\boldsymbol{x})>0\}})]\right) = \mathbf{E}[f\cdot\mathbf{1}_{\{f(\boldsymbol{x})>0\}}];$$

hence,

$$\tfrac{1}{4}\|f\|_1^2 = \mathbf{E}[f\cdot\mathbf{1}_{\{f(\boldsymbol{x})>0\}}]^2 \leq \mathbf{E}[f^2]\cdot\mathbf{E}[\mathbf{1}_{\{f(\boldsymbol{x})>0\}}^2] \leq e^{2k}\|f\|_1^2\cdot\mathbf{Pr}[f(\boldsymbol{x}) > 0]$$

using Cauchy–Schwarz and Theorem 9.22. The result follows. □

Next we turn to noise stability. Using the $(p,2)$-Hypercontractivity Theorem we can immediately deduce the following generalization of Corollary 9.8:

Small-Set Expansion Theorem. *Let $A \subseteq \{-1,1\}^n$ have volume α; i.e., let $1_A : \{-1,1\}^n \to \{0,1\}$ satisfy $\mathbf{E}[1_A] = \alpha$. Then for any $0 \le \rho \le 1$,*

$$\mathbf{Stab}_\rho[1_A] = \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim\{-1,1\}^n \\ \boldsymbol{y}\sim N_\rho(\boldsymbol{x})}}[\boldsymbol{x} \in A, \boldsymbol{y} \in A] \le \alpha^{\frac{2}{1+\rho}}.$$

Equivalently (for $\alpha > 0$),

$$\mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim A \\ \boldsymbol{y}\sim N_\rho(\boldsymbol{x})}}[\boldsymbol{y} \in A] \le \alpha^{\frac{1-\rho}{1+\rho}}.$$

In other words, the δ-noisy hypercube is a small-set expander for any $\delta > 0$: the probability that one step from a random $\boldsymbol{x} \sim A$ stays inside A is at most $\alpha^{\delta/(1-\delta)}$. It's also possible to derive a "two-set" generalization of this fact using the Two-Function Hypercontractivity Theorem; we defer the discussion to Chapter 10.1 since the most general result requires the non-weak form of the theorem. We can also obtain the generalization of Corollary 9.12:

Corollary 9.25. *Let $f : \{-1,1\}^n \to \{-1,1\}$. Then for any $0 \le \rho \le 1$ we have $\mathbf{Inf}_i^{(\rho)}[f] \le \mathbf{Inf}_i[f]^{\frac{2}{1+\rho}}$ for all i.*

Finally, from the Small-Set Expansion Theorem we see that indicators of small-volume sets are not very noise-stable and hence can't have much of their Fourier weight at low levels. Indeed, using hypercontractivity we can deduce the Level-1 Inequality from Chapter 5.4 and also generalize it to higher degrees.

Level-k Inequalities. *Let $f : \{-1,1\}^n \to \{0,1\}$ have mean $\mathbf{E}[f] = \alpha$ and let $k \in \mathbb{N}^+$ be at most $2\ln(1/\alpha)$. Then*

$$\mathbf{W}^{\le k}[f] \le \left(\tfrac{2e}{k}\ln(1/\alpha)\right)^k \alpha^2.$$

In particular, defining $k_\epsilon = 2(1-\epsilon)\ln(1/\alpha)$ (for any $0 \le \epsilon \le 1$) we have

$$\mathbf{W}^{\le k_\epsilon}[f] \le \alpha^{\epsilon^2}.$$

Proof. By the Small-Set Expansion Theorem,

$$\mathbf{W}^{\le k}[f] \le \rho^{-k}\mathbf{Stab}_\rho[f] \le \rho^{-k}\alpha^{2/(1+\rho)} \le \rho^{-k}\alpha^{2(1-\rho)}$$

for any $0 < \rho \le 1$. Basic calculus shows the right-hand side is minimized when $\rho = \frac{k}{2\ln(1/\alpha)} \le 1$; substituting this into $\rho^{-k}\alpha^{2(1-\rho)}$ yields the first claim. The second claim follows after substituting $k = k_\epsilon$; see Exercise 9.19. □

For the case $k = 1$, a slightly different argument gives the sharp Level-1 Inequality $\mathbf{W}^1[f] \le 2\alpha^2\ln(1/\alpha)$; see Exercise 9.18.

9.6. Highlight: The Kahn–Kalai–Linial Theorem

Recalling the social choice setting of Chapter 2.1, consider a 2-candidate, n-voter election using a monotone voting rule $f : \{-1,1\}^n \to \{-1,1\}$. We assume the impartial culture assumption (that the votes are independent and uniformly random), but with a twist: one of the candidates, say $b \in \{-1,1\}$, is able to secretly bribe k voters, fixing their votes to b. (Since f is monotone, this is always the optimal way for the candidate to fix the bribed votes.) How much can this influence the outcome of the election? This question was posed by Ben-Or and Linial in a 1985 work [BL85, BL90]; more precisely, they were interested in designing (unbiased) voting rules f that minimize the effect of any bribed k-coalition.

Let's first consider $k = 1$. If voter i is bribed to vote for candidate b (but all other votes remain uniformly random), this changes the bias of f by $b\widehat{f}(i) = b\mathbf{Inf}_i[f]$. Here we used the assumption that f is monotone (i.e., Proposition 2.21). This led Ben-Or and Linial to the question of which unbiased $f : \{-1,1\}^n \to \{-1,1\}$ has the least possible maximum influence:

Definition 9.26. Let $f : \{-1,1\}^n \to \mathbb{R}$. The *maximum influence* of f is

$$\mathbf{MaxInf}[f] = \max\{\mathbf{Inf}_i[f] : i \in [n]\}.$$

Ben-Or and Linial constructed the (nearly) unbiased $\mathrm{Tribes}_n : \{-1,1\}^n \to \{-1,1\}$ function (from Chapter 4.2) and noted that it satisfies $\mathbf{MaxInf}[\mathrm{Tribes}_n] = O(\frac{\log n}{n})$. They further conjectured that every unbiased function f has $\mathbf{MaxInf}[f] = \Omega(\frac{\log n}{n})$. This conjecture was famously proved by Kahn, Kalai, and Linial [KKL88]:

Kahn–Kalai–Linial (KKL) Theorem. *For any $f : \{-1,1\}^n \to \{-1,1\}$,*

$$\mathbf{MaxInf}[f] \geq \mathbf{Var}[f] \cdot \Omega\Big(\frac{\log n}{n}\Big).$$

Notice that the theorem says something sensible even for very biased functions f, i.e., those with low variance. The variance of f is indeed the right "scaling factor" since

$$\frac{1}{n}\mathbf{Var}[f] \leq \mathbf{MaxInf}[f] \leq \mathbf{Var}[f]$$

holds trivially, by the Poincaré Inequality and Exercise 2.8.

Before proving the KKL Theorem, let's see an additional consequence for Ben-Or and Linial's problem.

Proposition 9.27. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be monotone and assume $\mathbf{E}[f] \geq -.99$. Then there exists a subset $J \subseteq [n]$ with $|J| \leq O(n/\log n)$ that if "bribed to vote 1" causes the outcome to be 1 almost surely; i.e.,*

$$\mathbf{E}[f_{\overline{J}|(1,\dots,1)}] \geq .99. \tag{9.14}$$

Similarly, if $\mathbf{E}[f] \leq .99$ *there exists* $J \subseteq [n]$ *with* $|J| \leq O(n/\log n)$ *such that* $\mathbf{E}[f_{\overline{J}|(-1,\dots,-1)}] \leq -.99$.

Proof. By symmetry it suffices to prove the result regarding bribery by candidate $+1$. The candidate executes the following strategy: First, bribe the voter i_1 with the largest influence on $f_0 = f$; then bribe the voter i_2 with the largest influence on $f_1 = f^{(i_1 \mapsto 1)}$; then bribe the voter i_3 with the largest influence on $f_2 = f^{(i_1, i_2 \mapsto 1)}$; etc. For each $t \in \mathbb{N}$ we have

$$\mathbf{E}[f_{t+1}] \geq \mathbf{E}[f_t] + \mathbf{MaxInf}[f_t].$$

If after t bribes the candidate has not yet achieved (9.14) we have $-.99 \leq \mathbf{E}[f_t] < .99$; thus $\mathbf{Var}[f_t] \geq \Omega(1)$ and the KKL Theorem implies that $\mathbf{MaxInf}[f_t] \geq \Omega(\frac{\log n}{n})$. Thus the candidate will achieve a bias of at least .99 after bribing at most $(.99 - (-.99))/\Omega(\frac{\log n}{n}) = O(n/\log n)$ voters. □

Thus in any monotone election scheme, there is always a candidate $b \in \{-1, 1\}$ and a $o(1)$-fraction of the voters that b can bribe such that the election becomes 99%-biased in b's favor. And if the election scheme was not terribly biased to begin with, then *both* candidates have this ability. For a more precise version of this result, see Exercise 9.27; for a nonmonotone version, see Exercise 9.28. Note also that although the Tribes_n function is essentially optimal for standing up to a single bribed voter, it is quite bad at standing up to bribed coalitions: by bribing just a single tribe (DNF term) – about $\log n$ voters – the outcome can be completely forced to True. Nevertheless, Proposition 9.27 is close to sharp: Ajtai and Linial [**AL93**] constructed an unbiased monotone function $f : \{-1,1\}^n \to \{-1,1\}$ such that bribing any set of at most $\epsilon n/\log^2 n$ voters changes the expectation by at most $O(\epsilon)$.

The remainder of this section is devoted to the proof of the KKL Theorem and some variants. As mentioned earlier, the proof quickly follows from summing Corollary 9.12 over all coordinates; but let's give a more leisurely description. We'll focus on the main case of interest: showing that $\mathbf{MaxInf}[f] \geq \Omega(\frac{\log n}{n})$ when f is unbiased (i.e., $\mathbf{Var}[f] = 1$). If f's total influence is at least, say, $.1 \log n$, then even the *average* influence is $\Omega(\frac{\log n}{n})$. So we may as well assume $\mathbf{I}[f] \leq .1 \log n$.

This leads us to the problem of characterizing (unbiased) functions with small total influence. (This is the same issue that arose at the end of Chapter 8.4 when studying sharp thresholds.) It's helpful to think about the case that the total influence is *very* small – say $\mathbf{I}[f] \leq K$ where $K = 10$ or $K = 100$, though we eventually want to handle $K = .1 \log n$. Let's think of f as the indicator of a volume-1/2 set $A \subset \{-1,1\}^n$, so $\frac{\mathbf{I}[f]}{n}$ is the fraction of Hamming cube edges on the boundary of A. The edge-isoperimetric inequality (or Poincaré Inequality) tells us that $\mathbf{I}[f] \geq 1$: at least a $\frac{1}{n}$ fraction of the cube's edges must be on A's boundary, with dictators and negated-dictators being the minimizers. Now what can we say if $\mathbf{I}[f] \le K$; i.e., A's boundary has only K times more edges than the minimum? Must f be "somewhat similar" to a dictator or negated-dictator? Kahn, Kalai, and Linial showed that the answer is yes: f must have a coordinate with influence at least $2^{-O(K)}$. This should be considered very large (and dictator-like), since a priori all of the influences could have been equal to $\frac{K}{n}$.

KKL Edge-Isoperimetric Theorem. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be nonconstant and let $\widetilde{\mathbf{I}}[f] = \mathbf{I}[f]/\mathbf{Var}[f] \ge 1$ (which is just $\mathbf{I}[f]$ if f is unbiased). Then*

$$\mathbf{MaxInf}[f] \ge \left(\frac{9}{\widetilde{\mathbf{I}}[f]^2}\right) \cdot 9^{-\widetilde{\mathbf{I}}[f]}.$$

This theorem is sharp for $\widetilde{\mathbf{I}}[f] = 1$ (cf. Exercises 1.19, 5.35), and it's nontrivial (in the unbiased case) for $\mathbf{I}[f]$ as large as $\Theta(\log n)$. This last fact lets us complete the proof of the KKL Theorem as originally stated:

Proof of the KKL Theorem from the Edge-Isoperimetric version. We may assume f is nonconstant. If $\widetilde{\mathbf{I}}[f] = \mathbf{I}[f]/\mathbf{Var}[f] \ge .1\log n$, then we are done: the total influence is at least $.1\,\mathbf{Var}[f]\cdot\log n$ and hence $\mathbf{MaxInf}[f] \ge .1\,\mathbf{Var}[f]\cdot\frac{\log n}{n}$. Otherwise, the KKL Edge-Isoperimetric Theorem implies

$$\mathbf{MaxInf}[f] \ge \Omega\left(\frac{1}{\log^2 n}\right)\cdot 9^{-.1\log n} = \widetilde{\Omega}(n^{-.1\log 9}) = \Omega(n^{-.317}) \gg \mathbf{Var}[f]\cdot\Omega\left(\frac{\log n}{n}\right).$$

□

(You are asked to be careful about the constant factors in Exercise 9.30.)

We now turn to proving the KKL Edge-Isoperimetric Theorem. The high-level idea is to look at the contrapositive: supposing all of f's influences are small, we want to show its total influence must be large. The assumption here is that each derivative $\mathrm{D}_i f$ is a $\{-1,0,1\}$-valued function which is nonzero only on a "small" set. Hence "small-set expansion" implies that each derivative has "unusually large" noise sensitivity. (We are really just repeating Corollary 9.12 in words here.) In turn this means that for each $i \in [n]$, the Fourier weight of f on coefficients containing i must be quite "high up". Since this holds for all i we deduce that *all* of f's Fourier weight must be quite "high up" – hence f must have "large" total influence. We now make this story formal:

Proof of the KKL Edge-Isoperimetric Theorem. We treat only the case that f is unbiased, leaving the general case to Exercise 9.29 (see also the version for product space domains in Chapter 10.3). The theorem is an immediate consequence of the following chain of inequalities:

$$3\cdot 3^{-\mathbf{I}[f]} \overset{\text{(a)}}{\le} 3\mathbf{Stab}_{1/3}[f] \overset{\text{(b)}}{\le} \mathbf{I}^{(1/3)}[f] \overset{\text{(c)}}{\le} \sum_{i=1}^{n} \mathbf{Inf}_i[f]^{3/2} \overset{\text{(d)}}{\le} \mathbf{MaxInf}[f]^{1/2}\cdot\mathbf{I}[f].$$

The key inequality is (c), which comes from summing Corollary 9.12 over all coordinates $i \in [n]$. Inequality (d) is immediate from $\mathbf{Inf}_i[f]^{3/2} \leq \mathbf{MaxInf}[f]^{1/2} \cdot \mathbf{Inf}_i[f]$. Inequality (b) is trivial from the Fourier formulas (recall Fact 2.53):

$$\mathbf{I}^{(1/3)}[f] = \sum_{|S|\geq 1} |S|(1/3)^{|S|-1}\widehat{f}(S)^2 \geq 3 \sum_{|S|\geq 1} (1/3)^{|S|}\widehat{f}(S)^2 = 3\mathbf{Stab}_{1/3}[f]$$

(the last equality using $\widehat{f}(\emptyset) = 0$). Finally, inequality (a) is quickly proved using the spectral sample: for $\boldsymbol{S} \sim \mathcal{S}_f$ we have

$$3\mathbf{Stab}_{1/3}[f] = 3 \sum_{S\subseteq[n]} (1/3)^{|S|}\widehat{f}(S)^2 = 3\,\mathbf{E}[3^{-|\boldsymbol{S}|}] \geq 3 \cdot 3^{-\mathbf{E}[|\boldsymbol{S}|]} = 3 \cdot 3^{-\mathbf{I}[f]}, \qquad (9.15)$$

the inequality following from convexity of $s \mapsto 3^{-s}$. We remark that it's essentially only this (9.15) that needs to be adjusted when f is not unbiased. □

We end this chapter by deriving an even stronger version of the KKL Edge-Isoperimetric Theorem, and deducing Friedgut's Junta Theorem (mentioned at the end of Chapter 3.1) as a consequence. The KKL Edge-Isoperimetric Theorem tells us that if f is unbiased and $\mathbf{I}[f] \leq K$ then f must look somewhat like a 1-junta, in the sense of having a coordinate with influence at least $2^{-O(K)}$. Friedgut's Junta Theorem shows that in fact f must essentially be a $2^{O(K)}$-junta. To obtain this conclusion, you really just have to sum Corollary 9.12 only over the coordinates which have small influence on f. It's also possible to get even stronger conclusions if f is known to have particularly good low-degree Fourier concentration. In aid of this, we'll start by proving the following somewhat technical-looking result:

Theorem 9.28. *Let $f : \{-1,1\}^n \to \{-1,1\}$. Given $0 < \epsilon \leq 1$ and $k \geq 0$, define*

$$\tau = \frac{\epsilon^2}{\mathbf{I}[f]^2} 9^{-k}, \qquad J = \{j \in [n] : \mathbf{Inf}_j[f] \geq \tau\}, \qquad \textit{so } |J| \leq (\mathbf{I}[f]^3/\epsilon^2)9^k.$$

Then f's Fourier spectrum is ϵ-concentrated on

$$\mathscr{F} = \{S : S \subseteq J\} \cup \{S : |S| > k\}.$$

In particular, suppose f's Fourier spectrum is also ϵ-concentrated on degree up to k. Then f's Fourier spectrum is 2ϵ-concentrated on

$$\mathscr{F}' = \{S : S \subseteq J, |S| \leq k\},$$

and f is ϵ-close to a $|J|$-junta $h : \{-1,1\}^J \to \{-1,1\}$.

Proof. Summing Corollary 9.12 just over $i \notin J$ we obtain

$$\sum_{i\notin J} \mathbf{Inf}_i^{(1/3)}[f] \leq \sum_{i\notin J} \mathbf{Inf}_i[f]^{3/2} \leq \max_{i\notin J}\{\mathbf{Inf}_i[f]^{1/2}\} \cdot \sum_{i\notin J} \mathbf{Inf}_i[f] \leq \tau^{1/2} \cdot \mathbf{I}[f] \leq 3^{-k}\epsilon,$$

where the last two inequalities used the definitions of J and τ, respectively. On the other hand,

$$\sum_{i \notin J} \mathbf{Inf}_i^{(1/3)}[f] = \sum_{i \notin J} \sum_{S \ni i} (1/3)^{|S|-1} \widehat{f}(S)^2 = \sum_S |S \cap \overline{J}| \cdot 3^{1-|S|} \widehat{f}(S)^2$$
$$\geq \sum_{S \notin \mathscr{F}} |S \cap \overline{J}| \cdot 3^{1-|S|} \widehat{f}(S)^2 \geq 3^{-k} \sum_{S \notin \mathscr{F}} \widehat{f}(S)^2.$$

Here the last inequality used that $S \notin \mathscr{F}$ implies $|S \cap \overline{J}| \geq 1$ and $3^{1-|S|} \geq 3^{-k}$. Combining these two deductions yields $\sum_{S \notin \mathscr{F}} \widehat{f}(S)^2 \leq \epsilon$, as claimed.

As for the second part of the theorem, when f's Fourier spectrum is 2ϵ-concentrated on $\mathscr{F}'$ it follows from Proposition 3.31 that f is 2ϵ-close to the Boolean-valued $|J|$-junta $\mathrm{sgn}(f^{\subseteq J})$. From Exercise 3.34 we may deduce that f is in fact ϵ-close to some $h : \{-1,1\}^J \to \{-1,1\}$. □

Remark 9.29. As you are asked to show in Exercise 9.31, by using Corollary 9.25 in place of Corollary 9.12, we can achieve junta size $(\mathbf{I}[f]^{2+\eta}/\epsilon^{1+\eta}) \cdot C(\eta)^k$ in Theorem 9.28 for any $\eta > 0$, where $C(\eta) = (2/\eta + 1)^2$.

In Theorem 9.28 we may always take $k = \mathbf{I}[f]/\epsilon$, by the "Markov argument" Proposition 3.2. Thus we obtain as a corollary:

Friedgut's Junta Theorem. *Let $f : \{-1,1\}^n \to \{-1,1\}$ and let $0 < \epsilon \leq 1$. Then f is ϵ-close to an $\exp(O(\mathbf{I}[f]/\epsilon))$-junta. Indeed, there is a set $J \subseteq [n]$ with $|J| \leq \exp(O(\mathbf{I}[f]/\epsilon))$ such that f's Fourier spectrum is 2ϵ-concentrated on $\{S \subseteq J : |S| \leq \mathbf{I}[f]/\epsilon\}$.*

As mentioned, we can get stronger results for functions that are ϵ-concentrated up to degree much less than $\mathbf{I}[f]/\epsilon$. Width-w DNFs, for example, are ϵ-concentrated on degree up to $O(w\log(1/\epsilon))$ (by Theorem 4.22). Thus:

Corollary 9.30. *Any width-w DNF is ϵ-close to a $(1/\epsilon)^{O(w)}$-junta.*

Uniformly noise-stable functions do even better. From Peres's Theorem we know that linear threshold functions are ϵ-concentrated up to degree $O(1/\epsilon^2)$. Thus Theorem 9.28 and Remark 9.29 imply:

Corollary 9.31. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a linear threshold function and let $0 < \epsilon, \eta \leq 1/2$. Then f is ϵ-close to a junta on $\mathbf{I}[f]^{2+\eta} \cdot (1/\eta)^{O(1/\epsilon^2)}$ coordinates.*

Assuming ϵ is a small universal constant we can take $\eta = 1/\log(O(\mathbf{I}[f]))$ and deduce that every LTF is ϵ-close to a junta on $\mathbf{I}[f]^2 \cdot \mathrm{polylog}(\mathbf{I}[f])$ coordinates. This is essentially best possible since $\mathbf{I}[\mathrm{Maj}_n] = \Theta(\sqrt{n})$, but Maj_n is not even .1-close to any $o(n)$-junta. By virtue of Theorem 5.37 on the uniform noise stability of PTFs, we can also get this conclusion for any constant-degree PTF.

One more interesting fact we may derive is that every Boolean function has a Fourier coefficient that is at least inverse-exponential in the square of its total influence:

Corollary 9.32. *Assume $f : \{-1,1\}^n \to \{-1,1\}$ satisfies $\mathbf{Var}[f] \geq 1/2$. Then there exists $S \subseteq [n]$ with $0 < |S| \leq O(\mathbf{I}[f])$ such that $\widehat{f}(S)^2 \geq \exp(-O(\mathbf{I}[f]^2))$.*

Proof. Taking $\epsilon = 1/8$ in Friedgut's Junta Theorem we get a J with $|J| \leq \exp(O(\mathbf{I}[f]))$ such that f has Fourier weight at least $1-2\epsilon = 3/4$ on $\mathscr{F} = \{S \subseteq J : |S| \leq 8\mathbf{I}[f]\}$. Since $\widehat{f}(\emptyset)^2 = 1 - \mathbf{Var}[f] \leq 1/2$ we conclude that f has Fourier weight at least $1/4$ on $\mathscr{F}' = \mathscr{F} \setminus \{\emptyset\}$. But $|\mathscr{F}'| \leq |J|^{8\mathbf{I}[f]} = \exp(O(\mathbf{I}[f]^2))$, so the result follows by the Pigeonhole Principle. (Here we used that $(1/4)\exp(-O(\mathbf{I}[f]^2)) = \exp(-O(\mathbf{I}[f]^2))$ because $\mathbf{I}[f] \geq \mathbf{Var}[f] \geq \frac{1}{2}$.) □

Remark 9.33. Of course, if $\mathbf{Var}[f] < 1/2$, then f has a large empty Fourier coefficient: $\widehat{f}(\emptyset)^2 \geq 1/2$. For a more refined version of Corollary 9.32, see Exercise 9.32.

It is an open question whether Corollary 9.32 can be improved to give a Fourier coefficient satisfying $\widehat{f}(S)^2 \geq \exp(-O(\mathbf{I}[f]))$; see Exercise 9.33 and the discussion of the Fourier Entropy–Influence Conjecture in Exercise 10.23.

9.7. Exercises and notes

9.1 For every $1 < b < B$ show that there is a b-reasonable random variable $\boldsymbol{X}$ such that $1 + \boldsymbol{X}$ is not B-reasonable.

9.2 For $k = 1$, improve the 9 in the Bonami Lemma to 3. More precisely, suppose $f : \{-1,1\}^n \to \mathbb{R}$ has degree at most 1 and that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are independent 3-reasonable random variables satisfying $\mathbf{E}[\boldsymbol{x}_i] = \mathbf{E}[\boldsymbol{x}_i^3] = 0$. (For example, the $\boldsymbol{x}_i$'s may be uniform ± 1 bits.) Show that $f(\boldsymbol{x})$ is also 3-reasonable. (Hint: By direct computation, or by running through the Bonami Lemma proof with $k = 1$ more carefully.)

9.3 Let k be a positive multiple of 3 and let $n \geq 2k$ be an integer. Define $f : \{-1,1\}^n \to \mathbb{R}$ by

$$f(x) = \sum_{\substack{S \subseteq [n] \\ |S| = k}} x^S.$$

(*a*) Show that

$$\mathbf{E}[f^4] \geq \frac{\binom{n}{k/3,k/3,k/3,k/3,k/3,k/3,n-2k}}{\binom{n}{k}^2} \mathbf{E}[f^2]^2,$$

where the numerator of the fraction is a multinomial coefficient – specifically, the number of ways of choosing six disjoint size-$k/3$ subsets of $[n]$. (Hint: Given such size-$k/3$ subsets, consider quadruples of size-k subsets that hit each size-$k/3$ subset twice.)

(*b*) Using Stirling's Formula, show that

$$\lim_{n\to\infty} \frac{\binom{n}{k/3,k/3,k/3,k/3,k/3,k/3,n-2k}}{\binom{n}{k}^2} = \Theta(k^{-2}9^k).$$

Deduce the following lower bound for the Bonami Lemma: $\|f\|_4 \geq \Omega(k^{-1/2}) \cdot \sqrt{3}^k \|f\|_2$. (In fact, $\|f\|_4 = \Theta(k^{-1/4}) \cdot \sqrt{3}^k \|f\|_2$ and such an upper bound holds for all f homogeneous of degree k; see Exercise and 9.38(*f*).)

9.4 Prove Corollary 9.6.

9.5 Let $0 \leq \delta \leq \frac{1}{1600}$ and let f, ℓ be real numbers satisfying $|\ell^2 - 1| > 39\sqrt{\delta}$ and $|f| = 1$. Show that $|f - \ell|^2 \geq 169\delta$. (This is a loose estimate; stronger ones are possible.)

9.6 Theorem 9.21 shows that the (2,4)-Hypercontractivity Theorem implies the Bonami Lemma. In this exercise you will show the reverse implication.

(*a*) Let $f : \{-1,1\}^n \to \mathbb{R}$. For a fixed $\delta \in (0,1)$, use the Bonami Lemma to show that

$$\|\mathrm{T}_{(1-\delta)/\sqrt{3}} f\|_4 \leq \sum_{k=0}^{\infty} (1-\delta)^k \|f^{=k}\|_2 \leq \tfrac{1}{\delta}\|f\|_2.$$

(*b*) For $g : \{-1,1\}^n \to \mathbb{R}$ and $d \in \mathbb{N}^+$, let $g^{\oplus d} : \{-1,1\}^{dn} \to \mathbb{R}$ be the function defined by $g^{\oplus d}(x^{(1)}, \dots, x^{(d)}) = g(x^{(1)})g(x^{(2)})\cdots g(x^{(d)})$ (where each $x^{(i)} \in \{-1,1\}^n$). Show that $\|\mathrm{T}_\rho(g^{\oplus d})\|_p = \|\mathrm{T}_\rho g\|_p^d$ holds for every $p \in \mathbb{R}^+$ and $\rho \in [-1,1]$. Note the special case $\rho = 1$.

(*c*) Deduce from parts (*a*) and (*b*) that in fact $\|\mathrm{T}_{(1-\delta)/\sqrt{3}} f\|_4 \leq \|f\|_2$. (Hint: Apply part (*a*) to $f^{\oplus d}$ for larger and larger d.)

(*d*) Deduce that in fact $\|\mathrm{T}_{1/\sqrt{3}} f\|_4 \leq \|f\|_2$; i.e., the (2,4)-Hypercontractivity Theorem follows from the Bonami Lemma. (Hint: Take the limit as $\delta \to 0^+$.)

9.7 Suppose we wish to show that $\|\mathrm{T}_\rho f\|_q \leq \|f\|_p$ for all $f : \{-1,1\}^n \to \mathbb{R}$. Show that it suffices to show this for all nonnegative f. (Hint: Exercise 2.34.)

9.8 Fix $k \in \mathbb{N}$. The goal of this exercise is to show that "projection to degree k is a bounded operator in all L^p norms, $p > 1$". Let $f : \{-1,1\}^n \to \mathbb{R}$.

(*a*) Let $q \geq 2$. Show that $\|f^{\leq k}\|_q \leq \sqrt{q-1}^k \|f\|_q$. (Hint: Use Theorem 9.21 to show the stronger statement $\|f^{\leq k}\|_q \leq \sqrt{q-1}^k \|f\|_2$.)

(*b*) Let $1 < q \leq 2$. Show that $\|f^{\leq k}\|_q \leq (1/\sqrt{q-1})^k \|f\|_q$. (Hint: Either give a similar direct proof using the $(p,2)$-Hypercontractivity Theorem, or explain how this follows from part (*a*) using the dual norm Proposition 9.19.)

9.9 Let $\boldsymbol{X}$ be (p,q,ρ)-hypercontractive.

(*a*) Show that $c\boldsymbol{X}$ is (p,q,ρ)-hypercontractive for any $c \in \mathbb{R}$.
(*b*) Show that $\rho \leq \frac{\|\boldsymbol{X}\|_p}{\|\boldsymbol{X}\|_q}$.

9.10 Let $\boldsymbol{X}$ be (p,q,ρ)-hypercontractive. (For simplicity you may want to assume $\boldsymbol{X}$ is a discrete random variable.)
(*a*) Show that $\mathbf{E}[\boldsymbol{X}]$ must be 0. (Hint: Taylor expand $\|1+\rho\epsilon\boldsymbol{X}\|_r$ to one term around $\epsilon = 0$; note that $\rho < 1$ by definition.)
(*b*) Show that $\rho \leq \sqrt{\frac{p-1}{q-1}}$. (Hint: Taylor expand $\|1+\rho\epsilon\boldsymbol{X}\|_r$ to two terms around $\epsilon = 0$.)

9.11 (*a*) Suppose $\mathbf{E}[\boldsymbol{X}] = 0$. Show that $\boldsymbol{X}$ is $(q,q,0)$-hypercontractive for all $q \geq 1$. (Hint: Use monotonicity of norms to reduce to the case $q = 1$.)
(*b*) Show further that $\boldsymbol{X}$ is (q,q,ρ)-hypercontractive for all $0 \leq \rho < 1$. (Hint: Write $(a+\rho\boldsymbol{X}) = (1-\rho)a + \rho(a+\boldsymbol{X})$ and employ the triangle inequality for $\|\cdot\|_q$.)
(*c*) Show that if $\boldsymbol{X}$ is (p,q,ρ)-hypercontractive, then it is also (p,q,ρ')-hypercontractive for all $0 \leq \rho' < \rho$. (Hint: Use the previous exercise along with Exercise 9.10(*a*).)

9.12 Let $\boldsymbol{X}$ be a (nonconstant) $(2,4,\rho)$-hypercontractive random variable. The goal of this exercise is to show the following anticoncentration result: For all $\theta \in \mathbb{R}$ and $0 < t < 1$,

$$\mathbf{Pr}[|\boldsymbol{X}-\theta| > t\|\boldsymbol{X}\|_2] \geq (1-t^2)^2\rho^4.$$

(*a*) Reduce to the case $\|\boldsymbol{X}\|_2 = 1$.
(*b*) Letting $\boldsymbol{Y} = (\boldsymbol{X}-\theta)^2$, show that $\mathbf{E}[\boldsymbol{Y}] = 1+\theta^2$ and $\mathbf{E}[\boldsymbol{Y}^2] \leq (\rho^{-2}+\theta^2)^2$.
(*c*) Using the Paley–Zygmund inequality, show that

$$\mathbf{Pr}[|\boldsymbol{X}-\theta| > t] \geq \left(\frac{\rho^2(1-t^2)+\rho^2\theta^2}{1+\rho^2\theta^2}\right)^2.$$

(*d*) Show that the right-hand side above is minimized for $\theta = 0$, thereby completing the proof.

9.13 Let $m \in \mathbb{N}^+$ and let $f : \{-1,1\}^n \to [m]$ be "unbiased", meaning $\mathbf{Pr}[f(\boldsymbol{x}) = i] = \frac{1}{m}$ for all $i \in [m]$. Let $0 \leq \rho \leq 1$ and let $(\boldsymbol{x},\boldsymbol{y})$ be a ρ-correlated pair. Show that $\mathbf{Pr}[f(\boldsymbol{x}) = f(\boldsymbol{y})] \leq (1/m)^{(1-\rho)/(1+\rho)}$. (More generally, you might show that this is an upper bound on $\mathbf{Stab}_\rho[f]$ for all $f : \{-1,1\}^n \to \triangle_m$ with $\mathbf{E}[f] = (\frac{1}{m},\dots,\frac{1}{m})$; see Exercise 8.33.)

9.14 (*a*) Let $f : \{-1,1\}^n \to \mathbb{R}$ have degree at most k. Prove that $\|f\|_2 \leq (1/\sqrt{p-1})^k\|f\|_p$ for any $1 \leq p \leq 2$ using the Hölder inequality strategy from our proof of the (4/3,2)-Hypercontractivity Theorem, together with Theorem 9.21.
(*b*) Verify that $\exp(\frac{2}{p}-1) < 1/\sqrt{p-1}$ for all $1 \leq p < 2$; i.e., the trickier Theorem 9.22 strictly improves on the bound from part (*a*).

9.15 Prove Theorem 9.22 in full generality. (Hint: Let θ be the solution of $\frac{1}{2} = \frac{\theta}{p} + \frac{1-\theta}{2+\epsilon}$. You will need to show that $\frac{1-\theta}{2\theta} = (\frac{2}{p} - 1)\frac{1}{\epsilon} + (\frac{1}{p} - \frac{1}{2})$.)

9.16 As mentioned, it's possible to deduce the $(2,q)$-Hypercontractivity Theorem from the $n = 1$ case using induction by derivatives. From this one can also obtain the $(p,2)$-Hypercontractivity Theorem via Proposition 9.19. Employing the notation $\boldsymbol{x} = (\boldsymbol{x}', \boldsymbol{x}_n)$, $\mathrm{T} = \mathrm{T}_{1/\sqrt{q-1}}$, $\boldsymbol{d} = \mathrm{D}_n f(\boldsymbol{x}')$, and $\boldsymbol{e} = \mathrm{E}_n f(\boldsymbol{x}')$, fill in details and justifications for the following proof sketch:

$$\|\mathrm{T}_{1/\sqrt{q-1}} f\|_q^2 = \mathop{\mathbf{E}}_{\boldsymbol{x}'}\Big[\mathop{\mathbf{E}}_{\boldsymbol{x}_n}\big[|\mathrm{T}\boldsymbol{e} + (1/\sqrt{q-1})\boldsymbol{x}_n \mathrm{T}\boldsymbol{d}|^q\big]\Big]^{2/q} \le \mathop{\mathbf{E}}_{\boldsymbol{x}'}\big[((\mathrm{T}\boldsymbol{e})^2 + (\mathrm{T}\boldsymbol{d})^2)^{q/2}\big]^{2/q}$$
$$= \|(\mathrm{T}\boldsymbol{e})^2 + (\mathrm{T}\boldsymbol{d})^2\|_{q/2} \le \|(\mathrm{T}\boldsymbol{e})^2\|_{q/2} + \|(\mathrm{T}\boldsymbol{d})^2\|_{q/2} = \|\mathrm{T}\boldsymbol{e}\|_q^2 + \|\mathrm{T}\boldsymbol{d}\|_q^2 \le \|\boldsymbol{e}\|_2^2 + \|\boldsymbol{d}\|_2^2 = \|f\|_2^2.$$

9.17 Deduce the $p < 2 < q$ cases of the Hypercontractivity Theorem from the $(2,q)$- and $(p,2)$-Hypercontractivity Theorems. (Hint: Use the semigroup property of T_ρ, Exercise 2.32.)

9.18 Let $f : \{-1,1\}^n \to \{0,1\}$ have $\mathbf{E}[f] = \alpha$.

(*a*) Show that $\mathbf{W}^1[f] \le \frac{1}{\rho}(\alpha^{2/(1+\rho)} - \alpha^2)$ for any $0 < \rho \le 1$.

(*b*) Deduce the sharp Level-1 Inequality $\mathbf{W}^1[f] \le 2\alpha^2 \ln(1/\alpha)$. (Hint: Take the limit $\rho \to 0^+$.)

9.19 In this exercise you will prove the second statement of the Level-k Inequalities.

(*a*) Show that choosing $k = k_\epsilon$ in the theorem yields

$$\mathbf{W}^{\le k_\epsilon}[f] \le \alpha^{2\epsilon - (2-2\epsilon)\ln(1/(1-\epsilon))}.$$

(*b*) Show that $2\epsilon - (2-2\epsilon)\ln(1/(1-\epsilon)) \ge \epsilon^2$ for all $0 \le \epsilon \le 1$.

9.20 Show that the KKL Theorem fails for functions $f : \{-1,1\}^n \to [-1,1]$, even under the assumption $\mathbf{Var}[f] \ge \Omega(1)$. (Hint: $f(x) = \mathrm{trunc}_{[-1,1]}(\frac{x_1+\cdots+x_n}{\sqrt{n}})$.)

9.21 (*a*) Show that $\mathscr{C} = \{f : \{-1,1\}^n \to \{-1,1\} \mid \mathbf{I}[f] \le O(\sqrt{\log n})\}$ is learnable from queries to any constant error $\epsilon > 0$ in time $\mathrm{poly}(n)$. (Hint: Theorem 9.28.)

(*b*) Show that $\mathscr{C} = \{\text{monotone } f : \{-1,1\}^n \to \{-1,1\} \mid \mathbf{I}[f] \le O(\sqrt{\log n})\}$ is learnable from random examples to any constant error $\epsilon > 0$ in time $\mathrm{poly}(n)$.

(*c*) Show that $\mathscr{C} = \{\text{monotone } f : \{-1,1\}^n \to \{-1,1\} \mid \mathrm{DT}_{\mathrm{size}}(f) \le \mathrm{poly}(n)\}$ is learnable from random examples to any constant error $\epsilon > 0$ in time $\mathrm{poly}(n)$. (Hint: the OS Inequality and Exercise 8.43.)

9.22 Deduce the following generalization of the $(2,q)$-Hypercontractivity Theorem: Let $f : \{-1,1\}^n \to \mathbb{R}$, $q \ge 2$, and assume $0 \le \rho \le 1$ satisfies $\rho^\lambda \le 1/\sqrt{q-1}$ for some $0 \le \lambda \le 1$. Then

$$\|\mathrm{T}_\rho f\|_q \le \|\mathrm{T}_\rho f\|_2^{1-\lambda} \|f\|_2^\lambda.$$

(Hint: Show $\|\mathrm{T}_\rho f\|_q^2 \le \sum_S (\rho^{2|S|} \widehat{f}(S)^2)^{1-\lambda} \cdot (\widehat{f}(S)^2)^\lambda$ and use Hölder.)

9.23 Let $f : \{-1,1\}^n \to [-1,1]$, let $0 \le \epsilon \le 1$, and assume $q \ge 2+2\epsilon$. Show that

$$\|\mathrm{T}_{1-\epsilon} f\|_q^q \le \|\mathrm{T}_{\frac{1}{\sqrt{1+2\epsilon}}} f\|_q^q \le (\|f\|_2^2)^{1+\epsilon}.$$

9.24 Recall the Gaussian quadrant probability $\Lambda_\rho(\mu)$ defined in Exercise 5.32 by $\Lambda_\rho(\mu) = \mathbf{Pr}[\boldsymbol{z}_1 > t, \boldsymbol{z}_2 > t]$, where $\boldsymbol{z}_1, \boldsymbol{z}_2$ are standard Gaussians with correlation $\mathbf{E}[\boldsymbol{z}_1\boldsymbol{z}_2] = \rho$ and t is defined by $\overline{\Phi}(t) = \mu$. The goal of this exercise is to show that for fixed $0 < \rho < 1$ we have the estimate

$$\Lambda_\rho(\mu) = \widetilde{\Theta}(\mu^{\frac{2}{1+\rho}}) \tag{9.16}$$

as $\mu \to 0$. In light of Exercise 5.32, this will show that the Small-Set Expansion Theorem for the ρ-stable hypercube graph is essentially sharp due to the example of Hamming balls of volume μ.

(*a*) First let's do an imprecise "heuristic" calculation. We have $\mathbf{Pr}[\boldsymbol{z}_1 > t] = \mathbf{Pr}[\boldsymbol{z}_1 \ge t] = \mu$ by definition. Conditioned on a Gaussian being at least t it is unlikely to be much more than t, so let's just pretend that $\boldsymbol{z}_1 = t$. Then the conditional distribution of $\boldsymbol{z}_2$ is $\rho t + \sqrt{1-\rho^2}\boldsymbol{y}$, where $\boldsymbol{y} \sim \mathrm{N}(0,1)$ is an independent Gaussian. Using the fact that $\overline{\Phi}(u) \sim \phi(u)/u$ as $u \to \infty$, deduce that $\mathbf{Pr}[\boldsymbol{z}_2 > t \mid \boldsymbol{z}_1 = t] = \widetilde{\Theta}(\mu^{\frac{1-\rho}{1+\rho}})$ and "hence" (9.16) holds.

(*b*) Let's now be rigorous. Recall that we are treating $0 < \rho < 1$ as fixed and letting $\mu \to 0$ (hence $t \to \infty$). Let $\phi_\rho(z_1, z_2)$ denote the joint pdf of $\boldsymbol{z}_1, \boldsymbol{z}_2$ so that

$$\Lambda_\rho(\mu) = \int_t^\infty \int_t^\infty \phi_\rho(z_1, z_2)\, dz_1\, dz_2.$$

Derive the following similar-looking integral:

$$\int_t^\infty \int_t^\infty (z_2 - \rho z_1)(z_1 - \rho t)\phi_\rho(z_1, z_2)\, dz_1\, dz_2 = \frac{(1-\rho^2)^{3/2}}{2\pi} \exp\left(-\frac{2}{1+\rho}\frac{t^2}{2}\right) \tag{9.17}$$

and show that the right-hand side is $\widetilde{\Theta}(\mu^{\frac{2}{1+\rho}})$.

(*c*) Show that

$$\mathbf{Pr}\left[\boldsymbol{z}_1 > \tfrac{t-1}{\rho}\right] = \int_{\frac{t-1}{\rho}}^\infty \phi(z_1)\, dz_1 = \widetilde{\Theta}(\mu^{\frac{1}{\rho^2}}) = o(\mu^{\frac{2}{1+\rho}}).$$

(*d*) Deduce (9.16). (Hint: Try to arrange that the extraneous factors $(z_2 - \rho)$, $(z_1 - \rho t)$ in (9.17) are both at least 1.)

9.25 Let $f : \{-1,1\}^n \to \{-1,1\}$, let $J \subseteq [n]$, and write $\overline{J} = [n] \setminus J$. Define the *coalitional influence* of J on f to be

$$\widetilde{\mathbf{Inf}}_J[f] = \mathop{\mathbf{Pr}}_{\boldsymbol{z} \sim \{-1,1\}^{\overline{J}}}[f_{J|\boldsymbol{z}} \text{ is not constant}].$$

Furthermore, for $b \in \{-1,+1\}$ define the *coalitional influence toward* b of J on f to be

$$\begin{aligned}\widetilde{\mathbf{Inf}}^{b}_{J}[f] &= \mathop{\mathbf{Pr}}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[f_{J|\boldsymbol{z}} \text{ can be made } b] - \mathbf{Pr}[f=b] \\ &= \mathop{\mathbf{Pr}}_{\boldsymbol{z}\sim\{-1,1\}^{\overline{J}}}[f_{J|\boldsymbol{z}} \not\equiv -b] - \mathbf{Pr}[f=b].\end{aligned}$$

For brevity, we'll sometimes write $\widetilde{\mathbf{Inf}}^{\pm}_{J}[f]$ rather than $\widetilde{\mathbf{Inf}}^{\pm 1}_{J}[f]$.

(a) Show that for coalitions of size 1 we have $\mathbf{Inf}_i[f] = \widetilde{\mathbf{Inf}}_{\{i\}}[f] = 2\widetilde{\mathbf{Inf}}^{\pm}_{\{i\}}[f]$.

(b) Show that $0 \le \widetilde{\mathbf{Inf}}^{\pm}_{J}[f] \le 1$.

(c) Show that $\widetilde{\mathbf{Inf}}_{J}[f] = \widetilde{\mathbf{Inf}}^{+}_{J}[f] + \widetilde{\mathbf{Inf}}^{-}_{J}[f]$.

(d) Show that if f is monotone, then

$$\widetilde{\mathbf{Inf}}^{b}_{J}[f] = \mathbf{Pr}[f_{\overline{J}|(b,\dots,b)} = b] - \mathbf{Pr}[f=b].$$

(e) Show that $\widetilde{\mathbf{Inf}}_{J}[\chi_{[n]}] = 1$ for all $J \neq \emptyset$.

(f) Supposing we write $t = |J|/\sqrt{n}$, show that $\widetilde{\mathbf{Inf}}^{\pm}_{J}[\mathrm{Maj}_n] = \Phi(t) - \frac{1}{2} \pm o(1)$ and hence $\widetilde{\mathbf{Inf}}_{J}[\mathrm{Maj}_n] = 2\Phi(t) - 1 \pm o(1)$. Thus $\widetilde{\mathbf{Inf}}_{J}[\mathrm{Maj}_n] = o(1)$ if $|J| = o(\sqrt{n})$ and $\widetilde{\mathbf{Inf}}_{J}[\mathrm{Maj}_n] = 1 - o(1)$ if $|J| = \omega(\sqrt{n})$. (Hint: Central Limit Theorem.)

(g) Show that $\max\{\widetilde{\mathbf{Inf}}^{\mathsf{True}}_{J}[\mathrm{Tribes}_n] : |J| \le \log n\} = 1/2 + \Theta(\frac{\log n}{n})$. On the other hand, show that $\max\{\widetilde{\mathbf{Inf}}^{\mathsf{False}}_{J}[\mathrm{Tribes}_n] : |J| \le k\} \le k \cdot O(\frac{\log n}{n})$. Deduce that for some positive constant c we have $\max\{\widetilde{\mathbf{Inf}}_{J}[\mathrm{Tribes}_n] : |J| \le cn/\log n\} \le .51$. (Hint: Refer to Proposition 4.12.)

9.26 Show that the exponential dependence on $\mathbf{I}[f]$ in Friedgut's Junta Theorem is necessary. (Hint: Exercise 4.15.)

9.27 Let $f : \{-1,1\}^n \to \{-1,1\}$ be a monotone function with $\mathbf{Var}[f] \ge \delta > 0$, and let $0 < \epsilon < 1/2$ be given.

(a) Improve Proposition 9.27 as follows: Show that there exists $J \subseteq [n]$ with $|J| \le O(\log \frac{1}{\epsilon\delta}) \cdot \frac{n}{\log n}$ such that $\mathbf{E}[f_{\overline{J}|(1,\dots,1)}] \ge 1 - \epsilon$. (Hint: How many bribes are required to move f's mean outside the interval $[1 - 2\eta, 1 - \eta]$?)

(b) Show that there exists $J \subseteq [n]$ with $|J| \le O(\log \frac{1}{\epsilon\delta}) \cdot \frac{n}{\log n}$ such that $\widetilde{\mathbf{Inf}}_{J}[f] \ge 1 - \epsilon$. (Hint: Use Exercise 9.25(*d*) and take the union of two influential sets.)

9.28 Let $f : \{-1,1\}^n \to \{-1,1\}$.

(a) Let $f^* : \{-1,1\}^n \to \{-1,1\}$ be the "monotonization" of f as defined in Exercise 2.52. Show that $\widetilde{\mathbf{Inf}}^{b}_{J}[f^*] \le \widetilde{\mathbf{Inf}}^{b}_{J}[f]$ for all $J \subseteq [n]$ and $b \in \{-1,1\}$, and hence also $\widetilde{\mathbf{Inf}}_{J}[f^*] \le \widetilde{\mathbf{Inf}}_{J}[f]$.

(*b*) Let $\mathbf{Var}[f] \geq \delta > 0$ and let $0 < \epsilon < 1/2$ be given. Show that there exists $J \subseteq [n]$ with $|J| \leq O(\log \frac{1}{\epsilon\delta}) \cdot \frac{n}{\log n}$ such that $\widetilde{\mathbf{Inf}}_J[f] \geq 1 - \epsilon$. (Hint: Combine part (*a*) with Exercise 9.27(*b*).)

9.29 Establish the general-variance case of the KKL Edge-Isoperimetric Theorem. (Hint: You'll need to replace (9.15) with

$$3 \sum_{|S| \geq 1} (1/3)^{|S|} \widehat{f}(S)^2 \geq 3\,\mathbf{Var}[f] \cdot 3^{-\mathbf{I}[f]/\mathbf{Var}[f]}.$$

Use the same convexity argument, but applied to the random variable $\boldsymbol{S}$ that takes on each outcome $\emptyset \neq S \subseteq [n]$ with probability $\widehat{f}(S)^2/\mathbf{Var}[f]$.)

9.30 The goal of this exercise is to attain the best known constant factor in the statement of the KKL Theorem.

(*a*) By using Corollary 9.25 in place of Corollary 9.12, obtain the following generalization of the KKL Edge-Isoperimetric Theorem: For any (nonconstant) $f : \{-1,1\}^n \to \{-1,1\}$ and $0 < \delta < 1$,

$$\mathbf{MaxInf}[f] \geq \left(\tfrac{1+\delta}{1-\delta}\right)^{\frac{1}{\delta}} \left(\tfrac{1}{\widetilde{\mathbf{I}}[f]}\right)^{\frac{1}{\delta}} \cdot \left(\tfrac{1-\delta}{1+\delta}\right)^{\frac{1}{\delta}\widetilde{\mathbf{I}}[f]},$$

where $\widetilde{\mathbf{I}}[f]$ denotes $\mathbf{I}[f]/\mathbf{Var}[f]$. (Hint: Write $\rho = \frac{1-\delta}{1+\delta}$.) Deduce that for any constant $C > e^2$ we have

$$\mathbf{MaxInf}[f] \geq \widetilde{\Omega}(C^{-\widetilde{\mathbf{I}}[f]}).$$

(*b*) More carefully, show that by taking $\delta = \frac{1}{2\widetilde{\mathbf{I}}[f]^{1/3}}$ we can achieve

$$\mathbf{MaxInf}[f] \geq \exp(-2\widetilde{\mathbf{I}}[f]) \cdot e^2 \cdot \left(\tfrac{1}{\widetilde{\mathbf{I}}[f]}\right)^{2\widetilde{\mathbf{I}}[f]^{1/3}} \cdot \exp(-\tfrac{1}{4}\widetilde{\mathbf{I}}[f]^{1/3}).$$

(Hint: Establish $\left(\frac{1-\delta}{1+\delta}\right)^{\frac{1}{\delta}} \geq \exp(-2-\delta^2)$ for $0 < \delta \leq 1/2$.)

(*c*) By distinguishing whether or not $\widetilde{\mathbf{I}}[f] \geq \frac{1}{2}(\ln n - \sqrt{\log n})$, establish the following form of the KKL Theorem: For any $f : \{-1,1\}^n \to \{-1,1\}$,

$$\mathbf{MaxInf}[f] \geq \tfrac{1}{2}\mathbf{Var}[f] \cdot \frac{\ln n}{n}(1 - o_n(1)).$$

9.31 Establish the claim in Remark 9.29.

9.32 Show that if $f : \{-1,1\}^n \to \{-1,1\}$ is nonconstant, then there exists $S \subseteq [n]$ with $0 < |S| \leq O(\mathbf{I}[f]/\mathbf{Var}[f])$ such that $\widehat{f}(S)^2 \geq \exp(-O(\mathbf{I}[f]^2/\mathbf{Var}[f]^2))$. (Hint: By mimicking Corollary 9.32's proof you should be able to establish the lower bound $\Omega(\mathbf{Var}[f]) \cdot \exp(-O(\mathbf{I}[f]^2/\mathbf{Var}[f]^2))$. To show that this quantity is also $\exp(-O(\mathbf{I}[f]^2/\mathbf{Var}[f]^2))$, use Theorem 2.39.)

9.33 Let $f : \{-1,1\}^n \to \{-1,1\}$ be a nonconstant *monotone* function. Improve on Corollary 9.32 by showing that there exists $S \neq \emptyset$ with $\widehat{f}(S)^2 \geq \exp(-O(\mathbf{I}[f]/\mathbf{Var}[f]))$. (Hint: You can even get $|S| \leq 1$; use the KKL Edge-Isoperimetric Theorem and Proposition 2.21.)

9.34 Let $f : \{-1,1\}^n \to \mathbb{R}$. Prove that $\|f\|_4 \leq \mathrm{sparsity}(\widehat{f})^{1/4}\|f\|_2$.

9.35 Let $q = 2r$ be a positive even integer, let $\rho = 1/\sqrt{q-1}$, and let $f_1, \dots, f_r : \{-1,1\}^n \to \mathbb{R}$. Generalize the $(2,q)$-Hypercontractivity Theorem by showing that

$$\mathbf{E}\left[\prod_{i=1}^{r} (\mathrm{T}_\rho f_i)^2\right] \leq \prod_{i=1}^{r} \mathbf{E}[f_i^2].$$

(Hint: Hölder's inequality.)

9.36 In this exercise you will give a simpler, stronger version of Theorem 9.17 under the assumption that $q = 2r$ is a positive even integer.

(*a*) Using the idea of Proposition 9.16, show that if $\boldsymbol{x}$ is a uniformly random ± 1 bit then $\boldsymbol{x}$ is $(2,q,\rho)$-hypercontractive if and only if $\rho \leq 1/\sqrt{q-1}$.

(*b*) Show the same statement for any random variable $\boldsymbol{x}$ satisfying $\mathbf{E}[\boldsymbol{x}^2] = 1$ and

$$\mathbf{E}[\boldsymbol{x}_i^{2j-1}] = 0, \quad \mathbf{E}[\boldsymbol{x}_i^{2j}] \leq (2r-1)^j \frac{\binom{r}{j}}{\binom{2r}{2j}} \quad \text{for all integers } 1 \leq j \leq r.$$

(*c*) Show that none of the even moment conditions in part (*b*) can be relaxed.

9.37 Let $q = 2r$ be a positive even integer and let $f : \{-1,1\}^n \to \mathbb{R}$ be homogeneous of degree $k \geq 1$ (i.e., $f = f^{=k}$). The goal of this problem is to improve slightly on the generalized Bonami Lemma, Theorem 9.21.

(*a*) Show that

$$\mathbf{E}[f^q] = \sum \widehat{f}(S_1) \cdots \widehat{f}(S_k) \leq \sum |\widehat{f}(S_1)| \cdots |\widehat{f}(S_k)|, \tag{9.18}$$

where the sum is over all tuples $S_1, \dots, S_k$ satisfying $S_1 \triangle \cdots \triangle S_k = \emptyset$.

(*b*) Let G denote the complete q-partite graph over vertex sets $V_1, \dots, V_q$, each of cardinality k. Let $\mathscr{M}$ denote the set of all perfect matchings in G. Show that the right-hand side of (9.18) is equal to

$$\frac{1}{(k!)^q} \sum_{M \in \mathscr{M}} \sum_{\ell : M \to [n]} |\widehat{f}(T_1(M,\ell))| \cdots |\widehat{f}(T_k(M,\ell))|, \tag{9.19}$$

where $T_j(M,\ell)$ denotes $\bigcup\{\ell(e) : e \in M, e \cap V_j \neq \emptyset\}$.

(*c*) Show that (9.19) is equal to

$$\frac{1}{(rk)! \cdot (k!)^q} \sum_{\overline{M} \in \overline{\mathscr{M}}} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_{rk}=1}^{n} |\widehat{f}(U_1(\overline{M}, i_1, \dots, i_{rk}))| \cdots |\widehat{f}(U_k(\overline{M}, i_1, \dots, i_{rk}))|, \tag{9.20}$$

where $\overline{\mathscr{M}}$ is the set of *ordered* perfect matchings of G, and now $U_j(\overline{M}, i_1, \dots, i_{rk})$ denotes $\bigcup\{i_t : \overline{M}(t) \cap V_j \neq \emptyset\}$.

(*d*) Show that for any $\overline{M} \in \overline{\mathscr{M}}$ we have

$$\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_{rk}=1}^{n} |\widehat{f}(U_1(\overline{M},i_1,\dots,i_{rk}))|\cdots|\widehat{f}(U_k(\overline{M},i_1,\dots,i_{rk}))| \leq \left(\sum_{j_1,\dots,j_k=1}^{n}\widehat{f}(\{j_1,\dots,j_k\})^2\right)^r$$

(Hint: Use Cauchy–Schwarz rk times.)

(*e*) Deduce that $\|f\|_q^q \leq \frac{1}{(rk)!\cdot(k!)^q}\cdot|\overline{\mathscr{M}}|\cdot(k!)^r\|f\|_2^{2r}$ and hence

$$\|f\|_q \leq \frac{|\mathscr{M}|^{1/q}}{\sqrt{k!}}\|f\|_2.$$

9.38 The goal of this problem is to estimate $|\mathscr{M}|$ from Exercise 9.37 so as to give a concrete improvement on Theorem 9.21.

(*a*) Show that for $q = 4$, $k = 2$ we have $|\mathscr{M}| = 60$.

(*b*) Show that $|\mathscr{M}| \leq (qk-1)!!$. (Hint: Show that $(qk-1)!!$ is the number of perfect matchings in the complete graph on qk vertices.) Deduce $\|f\|_q \leq \sqrt{q}^k\|f\|_2$.

(*c*) Show that $|\overline{\mathscr{M}}| \leq (\frac{2r-1}{r})^{rk}(rk)!^2$, and thereby deduce

$$\|f\|_q \leq C_{q,k}\cdot\sqrt{q-1}^k\|f\|_2,$$

where $C_{q,k} = \left(\frac{(rk)!}{k!^r r^{rk}}\right)^{1/q}$. (Hint: Suppose that the first t edges of the perfect matching have been chosen; show that there are $(\frac{2r-1}{r})(rk-t)^2$ choices for the next edge. The worst case is if the vertices used up so far are spread equally among the q parts.)

(*d*) Give a simple proof that $C_{q,k} \leq 1$, thereby obtaining Theorem 9.21.

(*e*) Show that in fact $C_{q,k} = \Theta(1)\cdot k^{-1/4+1/(2q)}$. (Hint: Stirling's Formula.)

(*f*) Can you obtain the improved estimate

$$\frac{|\mathscr{M}|^{1/q}}{\sqrt{k!}} = \Theta_q(1)\cdot k^{-1/4}\cdot\sqrt{q-1}^k?$$

(Hint: First exactly count – then estimate – the number of perfect matchings with exactly e_{ij} edges between parts i and j. Then sum your estimate over a range of the most likely values for e_{ij}.)

Notes. The history of the Hypercontractivity Theorem is complicated. Its earliest roots are in the work of Paley [Pal32] from 1932; he showed that for $1 < p < \infty$ there are constants $0 < c_p < C_p < \infty$ such that $c_p\|\mathrm{S}f\|_p \leq \|f\|_p \leq C_p\|\mathrm{S}f\|_p$ holds for any $f : \{-1,1\}^n \to \mathbb{R}$. Here $\mathrm{S}f = \sum_{t=1}^{n}\sqrt{\sum_{t=1}^{n}(\mathrm{d}_t f)^2}$ is the "square function" of f, and $\mathrm{d}_t f = \sum_{S:\max(S)=t}\widehat{f}(S)\chi_S$ is the martingale difference sequence for f defined in Exercise 8.17. The main task in Paley's work is to prove the statement when p is an even integer; other values of p follow by the Riesz(–Thorin) interpolation theorem. Using this result, Paley showed the following hypercontractivity result: If $f : \{-1,1\}^n \to \mathbb{R}$ is homogeneous of degree 2, then $c'_p \|f\|_2 \leq \|f\|_p \leq C'_p \|f\|_2$ for any $p \in \mathbb{R}^+$.

In 1968 Bonami [**Bon68**] stated the following variant of Theorem 9.21: If $f : \{-1,1\}^n \to \mathbb{R}$ is homogeneous of degree k, then for all $q \geq 2$, $\|f\|_q \leq c_k \sqrt{q} \|f\|_2$, where the constant c_k may be taken to be 1 if q is an even integer. She remarks that this theorem can be deduced from Paley's result but with a much worse (exponential) dependence on q. The proof she gives is combinatorial and actually only treats the case $k = 2$ and q an even integer; it is similar to Exercise 9.37.

Independently in 1969, Kiener [**Kie69**] published his Ph.D. thesis, which extended Paley's hypercontractivity result as follows: If $f : \{-1,1\}^n \to \mathbb{R}$ is homogeneous of degree k, then $c_{p,k} \|f\|_2 \leq \|f\|_p \leq C_{p,k} \|f\|_2$ for any $p \in \mathbb{R}^+$. The proof is an induction on k, and again the bulk of the work is the case of even integer p. Kiener also gave a long combinatorial proof showing that if $f : \{-1,1\}^n \to \mathbb{R}$ is homogeneous of degree 2, then $\mathbf{E}[f^4] \leq 51\,\mathbf{E}[f^2]^2$. (Exercise 9.38(*a*) improves this 51 to 15.)

Also independently in 1969, Schreiber [**Sch69**] considered multilinear polynomials f over a general orthonormal sequence $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of centered real (or complex) random variables. He showed that if f has degree at most k, then for any even integer $q \geq 4$ it holds that $\|f\|_q \leq C \|f\|_2$, where C depends only on k, q, and the q-norms of the $\boldsymbol{x}_i$'s. Again, the proof is very similar to Exercise 9.37; Schreiber does not estimate his analogue of $|\mathscr{M}|$ but merely notes that it's finite. Schreiber was interested mainly in the case that the $\boldsymbol{x}_i$'s are Gaussian; indeed, his 1969 work [**Sch69**] is a generalization of his earlier work [**Sch67**] specific to the Gaussian case.

In 1970, Bonami published her Ph.D. thesis [**Bon70**], which contains the full Hypercontractivity Theorem as stated at the beginning of the chapter. Her proof follows the standard template seen in essentially all proofs of hypercontractivity: first an elementary proof for the case $n = 1$ and then an induction to extend to general n. She also gives the sharper combinatorial result appearing in Exercises 9.37 and 9.38(*c*). (The stronger bound from Exercise 9.38(*f*) is due to Janson [**Jan97**, Remark 5.20].) As in Corollary 9.6, Bonami notes that her combinatorial proof can be extended to a general sequence of symmetric orthonormal random variables, at the expense of including factors of $\|\boldsymbol{x}_i\|_q$ into the bound. She points out that this includes the Gaussian case independently studied by Schreiber.

Bonami's work was published in French, and it remained unknown to most English-language mathematicians for about a decade. In the late 1960s and early 1970s, researchers in quantum field theory developed the theory of hypercontractivity for the Gaussian analogue of T_ρ, namely, the Ornstein–Uhlenbeck operator U_ρ. This is now recognized as essentially being a *special case* of hypercontractivity for bits, in light of the fact that $\frac{\boldsymbol{x}_1+\cdots+\boldsymbol{x}_n}{\sqrt{n}}$ tends to a Gaussian as $n \to \infty$ by the CLT (see Chapter 11.1). We summarize here some of the work in this setting. In 1966 Nelson [**Nel66**] showed that $\|\mathrm{U}_{1/\sqrt{q-1}} f\|_q \leq C_q \|f\|_2$ for all $q \geq 2$. Glimm [**Gli68**] gave the alternative result that for each $q \geq 2$ there is a sufficiently small $\rho_q > 0$ such that $\|\mathrm{U}_{\rho_q} f\|_q \leq \|f\|_2$. Segal [**Seg70**] observed that hypercontractive results can be proved by induction on the dimension n. In 1973 Nelson [**Nel73**] gave the full Hypercontractivity Theorem in the Gaussian setting: $\|\mathrm{U}_{\sqrt{(p-1)/(q-1)}} f\|_q \leq \|f\|_p$ for all $1 \leq p < q \leq \infty$. He also proved the combinatorial Exercise 9.37. The equivalence to the Two-Function Hypercontractivity Theorem is from the work of Neveu [**Nev76**].

In 1975 Gross [**Gro75**] introduced the notion of Log-Sobolev Inequalities (see Exercise 10.23) and showed how to deduce hypercontractivity inequalities from them. He established the Log-Sobolev Inequality for 1-bit functions, used induction (citing Segal) to obtain it for n-bit functions, and then used the CLT to transfer results to the Gaussian setting. (For some earlier results along these lines, see the works of Federbush and Gross [**Fed69, Gro72**].) This gave a new proof of Nelson's result and also independently established Bonami's full Hypercontractivity Theorem. Also in 1975, Beckner [**Bec75**] published his Ph.D. thesis, which proved a sharp form of the hypercontractive inequality for purely complex ρ. (It is unfortunate that the influential paper of Kahn, Kalai, and Linial [**KKL88**] miscredited the Hypercontractivity Theorem to Beckner.) The case of general complex ρ was subsequently treated by Weissler [**Wei79**], with the sharp result being obtained by Epperson [**Epp89**]. Weissler [**Wei80**] also appears to have been the first to make the connection between this line of work and Bonami's thesis.

Independently of all this work, the $(q,2)$-Hypercontractivity Theorem was reproved (without sharp constant) in the Banach spaces community by Rosenthal [**Ros76**] in 1975, using methods similar to those of Paley and Kiener. For additional early references, see Müller [**Mül05**, Chapter 1].

The *term* "hypercontractivity" was introduced in Simon and Høegh-Krohn [**SHK72**]; Definition 9.13 of a hypercontractive random variable is due to Krakowiak and Szulga [**KS88**]. The short inductive proof of the Bonami Lemma may have appeared first in Mossel, O'Donnell, and Oleszkiewicz [**MOO05a**]. Theorems 9.22 and 9.24 appear in Janson [**Jan97**]. Theorem 9.23 dates back to Pisier and Zinn and to Borell [**PZ78, Bor79**]. The Small-Set Expansion Theorem is due to Kahn, Kalai, and Linial [**KKL88**]; the Level-k Inequalities appear in several places but can probably be fairly credited to Kahn, Kalai, and Linial [**KKL88**] as well. The optimal constants for Khintchine's

Inequality were established by Haagerup [Haa82]; see also Nazarov and Podkorytov [NP00]. They always occur either when $\sum_i a_i \boldsymbol{x}_i$ is just $\frac{1}{\sqrt{2}}\boldsymbol{x}_1 + \frac{1}{\sqrt{2}}\boldsymbol{x}_2$ or in the limiting Gaussian case of $a_i \equiv \frac{1}{\sqrt{n}}$, $n \to \infty$.

Ben-Or and Linial's work [BL85, BL90] was motivated both by game theory and by the Byzantine Generals problem [LSP82] from distributed computing; the content of Exercise 9.25 is theirs. In turn it motivated the watershed paper by Kahn, Kalai, and Linial [KKL88]. (See also the intermediate work of Chor and Geréb-Graus [CGG87].) The "KKL Edge-Isoperimetric Theorem" (which is essentially a strengthening of the basic KKL Theorem) was first explicitly proved by Talagrand [Tal94] (possibly independently of Kahn, Kalai, and Linial [KKL88]?); he also treated the p-biased case. There is no known combinatorial proof of the KKL Theorem (i.e., one which does not involve real-valued functions). However, several slightly different analytic proofs are known; see Falik and Samorodnitsky [FS07], Rossignol [Ros06], and O'Donnell and Wimmer [OW13]. The explicit lower bound on the "KKL constant" achieved in Exercise 9.30 is the best known; it appeared first in Falik and Samorodnitsky [FS07]. It is still a factor of 2 away from the best known upper bound, achieved by the tribes function.

Friedgut's Junta Theorem dates from 1998 [Fri98]. The observation that its junta size can be improved for functions which have $\mathbf{W}^k[f] \leq \epsilon$ for $k \ll \mathbf{I}[f]/\epsilon$ was independently made by Li-Yang Tan in 2011; so was the consequence Corollary 9.31 and its extension to constant-degree PTFs. A stronger result than Corollary 9.31 is known: Diakonikolas and Servedio [DS09] showed that every LTF is ϵ-close to a $\mathbf{I}[f]^2\mathrm{poly}(1/\epsilon)$-junta. As for Corollary 9.30, it's incomparable with a result from Gopalan, Meka, and Reingold [GMR12], which shows that every width-w DNF is ϵ-close to a $(w\log(1/\epsilon))^{O(w)}$-junta.

Exercise 9.3 was suggested to the author by Krzysztof Oleszkiewicz. Exercise 9.12 is from Gopalan et al. [GOWZ10]. Exercise 9.21 appears in O'Donnell and Servedio [OS07]; Exercise 9.22 appears in O'Donnell and Wu [OW09]. The estimate in Exercise 9.24 is from de Klerk, Pasechnik, and Warners [dKPW04] (see also Rinott and Rotar' [RR01] and Khot et al. [KKMO07]). Exercises 9.27 and 9.28 are due to Kahn, Kalai, and Linial [KKL88]. Exercise 9.34 was suggested to the author by John Wright. Exercise 9.36 appears in Kauers et al. [KOTZ13].

Chapter 10

Advanced hypercontractivity

In this chapter we complete the proof of the Hypercontractivity Theorem for uniform ±1 bits. We then generalize the $(p,2)$ and $(2,q)$ statements to the setting of arbitrary product probability spaces, proving the following:

The General Hypercontractivity Theorem. *Let (Ω_1,π_1), …, (Ω_n,π_n) be finite probability spaces, in each of which every outcome has probability at least λ. Let $f \in L^2(\Omega_1 \times \cdots \times \Omega_n, \pi_1 \otimes \cdots \otimes \pi_n)$. Then for any $q > 2$ and $0 \leq \rho \leq \frac{1}{\sqrt{q-1}} \cdot \lambda^{1/2-1/q}$,*

$$\|\mathrm{T}_\rho f\|_q \leq \|f\|_2 \quad \textit{and} \quad \|\mathrm{T}_\rho f\|_2 \leq \|f\|_{q'}.$$

(And in fact, the upper bound on ρ can be slightly relaxed to the value stated in Theorem 10.18.)

We can thereby extend all the consequences of the basic Hypercontractivity Theorem for $f : \{-1,1\}^n \to \mathbb{R}$ to functions $f \in L^2(\Omega^n, \pi^{\otimes n})$, except with quantitatively worse parameters depending on "λ". We also introduce the technique of randomization/symmetrization and show how it can sometimes eliminate this dependence on λ. For example, it's used to prove Bourgain's Sharp Threshold Theorem, a characterization of Boolean-valued $f \in L^2(\Omega^n, \pi^{\otimes n})$ with low total influence that has no dependence at all on π.

10.1. The Hypercontractivity Theorem for uniform ±1 bits

In this section we'll prove the full Hypercontractivity Theorem for uniform ±1 bits stated at the beginning of Chapter 9:

The Hypercontractivity Theorem. *Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $1 \le p \le q \le \infty$. Then $\|\mathrm{T}_\rho f\|_q \le \|f\|_p$ for $0 \le \rho \le \sqrt{\frac{p-1}{q-1}}$.*

Actually, when neither p nor q is 2, the following equivalent form of theorem seems easier to interpret:

Two-Function Hypercontractivity Theorem. *Let $f, g : \{-1,1\}^n \to \mathbb{R}$, let $r, s \ge 0$, and assume $0 \le \rho \le \sqrt{rs} \le 1$. Then*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] \le \|f\|_{1+r}\|g\|_{1+s}.$$

As a reminder, the only difference between this theorem and its "weak" form (proven in Chapter 9.4) is that we don't assume $r, s \le 1$. Below we will show that the two theorems *are* equivalent, via Hölder's inequality. Given the Two-Function Hypercontractivity Induction Theorem from Chapter 9.4, this implies that to prove the Hypercontractivity Theorem for general n we only need to prove it for $n = 1$. This is an elementary but technical inequality, which we defer to the end of the section.

Before carrying out these proofs, let's take some time to interpret the Two-Function Hypercontractivity Theorem. One interpretation is simply as a generalization of Hölder's inequality. Consider the case that the strings $\boldsymbol{x}$ and $\boldsymbol{y}$ in the theorem are fully correlated; i.e., $\rho = 1$. Then the theorem states that

$$\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{x})] \le \|f\|_{1+r}\|g\|_{1+1/r} \tag{10.1}$$

because the condition $\sqrt{rs} = 1$ is equivalent to $s = 1/r$. This statement is identical to Hölder's inequality, since $(1+r)' = 1 + 1/r$. Hölder's inequality is often used to "break the correlation" between two random variables; in the absence of any information about how f and g correlate then we can at least bound $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{x})]$ by the product of certain norms of f and g. (If f and g have different "sizes", then Hölder lets us choose different norms for them; if f and g have roughly the same "size", then we can take $r = s = 1$ and get Cauchy–Schwarz.) Now suppose we are considering $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{y})]$ for ρ-correlated $\boldsymbol{x}, \boldsymbol{y}$ with $\rho < 1$. In this case we might hope to improve (10.1) by using smaller norms on the right-hand side; in the extreme case of independent $\boldsymbol{x}, \boldsymbol{y}$ (i.e., $\rho = 0$) we can use $\mathbf{E}[f(\boldsymbol{x})g(\boldsymbol{y})] = \mathbf{E}[f]\mathbf{E}[g] \le \|f\|_1\|g\|_1$. The Two-Function Hypercontractivity Theorem gives a precise interpolation between these two cases; the smaller the correlation ρ is, the smaller the norms we may take on the right-hand side.

In the case that f and g have range $\{0,1\}$, these ideas yield another interpretation of the Two-Function Hypercontractivity Theorem, namely a two-set generalization of the Small-Set Expansion Theorem:

Generalized Small-Set Expansion Theorem. *Let* $0 \le \rho \le 1$. *Let* $A, B \subseteq \{-1,1\}^n$ *have volumes* $\exp(-\frac{a^2}{2})$, $\exp(-\frac{b^2}{2})$ *and assume* $0 \le \rho a \le b \le a$. *Then*

$$\underset{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}{\mathbf{Pr}}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] \le \exp\left(-\tfrac{1}{2}\tfrac{a^2-2\rho ab+b^2}{1-\rho^2}\right).$$

Proof. Apply the Two-Function Hypercontractivity Theorem with $f = 1_A$, $g = 1_B$ and minimize the right-hand side by selecting $r = \rho\frac{a-\rho b}{b-\rho a}$, $s = \rho\frac{b-\rho a}{a-\rho b}$. □

Remark 10.1. When a and b are not too close the optimal choice of r in the proof exceeds 1. Thus the Generalized Small-Set Expansion Theorem really needs the full (non-weak) Two-Function Hypercontractivity Theorem; equivalently, the full Hypercontractivity Theorem.

Remark 10.2. This theorem is essentially sharp in the case that A and B are concentric Hamming balls; see Exercise 10.5. In the case $b = a$ we recover the Small-Set Expansion Theorem. In the case $b = \rho a$ we get only the trivial bound that $\mathbf{Pr}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] \le \exp(-\frac{a^2}{2}) = \mathbf{Pr}[\boldsymbol{x} \in A]$. However, not much better than this can be expected; in the concentric Hamming ball case it indeed holds that $\mathbf{Pr}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] \sim \mathbf{Pr}[\boldsymbol{x} \in A]$ whenever $b < \rho a$.

Remark 10.3. There is also a *reverse* form of the Hypercontractivity Theorem and its Two-Function version; see Exercises 10.6–10.9. It directly implies the following:

Reverse Small-Set Expansion Theorem. *Let* $0 \le \rho \le 1$. *Let* $A, B \subseteq \{-1,1\}^n$ *have volumes* $\exp(-\frac{a^2}{2})$, $\exp(-\frac{b^2}{2})$, *where* $a, b \ge 0$. *Then*

$$\underset{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}{\mathbf{Pr}}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] \ge \exp\left(-\tfrac{1}{2}\tfrac{a^2+2\rho ab+b^2}{1-\rho^2}\right).$$

We now turn to the proofs. We begin by showing that the Hypercontractivity Theorem and the Two-Function version are indeed equivalent. This is a consequence of the following general fact (take $T = \mathrm{T}_\rho$, $p = 1+r$, $q = 1+1/s$):

Proposition 10.4. *Let T be an operator on $L^2(\Omega,\pi)$ and let $1 \le p, q \le \infty$. Then*

$$\|Tf\|_q \le \|f\|_p \tag{10.2}$$

holds for all $f \in L^2(\Omega,\pi)$ if and only if

$$\langle Tf, g\rangle \le \|f\|_p \|g\|_{q'} \tag{10.3}$$

holds for all $f, g \in L^2(\Omega,\pi)$.

Proof. For the "only if" statement, $\langle Tf, g\rangle \le \|Tf\|_q \|g\|_{q'} \le \|f\|_p \|g\|_{q'}$ by Hölder's inequality and (10.2). As for the "if" statement, by Hölder's inequality and (10.3) we have

$$\|Tf\|_q = \sup_{\|g\|_{q'}=1} \langle Tf, g\rangle \le \sup_{\|g\|_{q'}=1} \|f\|_p \|g\|_{q'} = \|f\|_p. \qquad \square$$

Now suppose we prove the Hypercontractivity Theorem in the case $n = 1$. By the above proposition we deduce the Two-Function version in the case $n = 1$. Then the Two-Function Hypercontractivity Induction Theorem from Chapter 9.4 yields the general-n case of the Two-Function Hypercontractivity Theorem. Finally, applying the above proposition again we get the general-n case of the Hypercontractivity Theorem, thereby completing all needed proofs. These observations all hold in the context of more general product spaces, so let's record the following for future use:

Hypercontractivity Induction Theorem. *Let $0 \le \rho \le 1$, $1 \le p, q \le \infty$, and assume that $\|\mathrm{T}_\rho f\|_q \le \|f\|_p$ holds for every $f \in L^2(\Omega_1,\pi_1),\dots,L^2(\Omega_n,\pi_n)$. Then it also holds for every $f \in L^2(\Omega_1\times\cdots\times\Omega_n, \pi_1\otimes\cdots\otimes\pi_n)$.*

Remark 10.5. In traditional proofs of the Hypercontractivity Theorem for ±1 bits, this theorem is proven directly; it's a slightly tricky induction by derivatives (see Exercise 10.3). For more general product spaces the same direct induction strategy also works but the notation becomes quite complicated.

Our remaining task, therefore, is to prove the Hypercontractivity Theorem in the case $n = 1$; in other words, to show that a uniformly random ±1 bit is $(p, q, \sqrt{(p-1)/(q-1)})$-hypercontractive. This fact is often called the "Two-Point Inequality" because (for fixed p, q, and ρ) it's just an "elementary" inequality about two real variables.

Two-Point Inequality. *Let $1 \le p \le q \le \infty$ and let $0 \le \rho \le \sqrt{(p-1)/(q-1)}$. Then $\|\mathrm{T}_\rho f\|_q \le \|f\|_p$ for any $f : \{-1,1\} \to \mathbb{R}$. Equivalently (for $\rho \neq 1$), a uniformly random bit $\boldsymbol{x} \sim \{-1,1\}$ is (p,q,ρ)-hypercontractive; i.e., $\|a + \rho b\boldsymbol{x}\|_q \le \|a + b\boldsymbol{x}\|_p$ for all $a, b \in \mathbb{R}$.*

Proof. As in Section 9.3, our main task will be to prove the inequality for $1 \le p < q \le 2$. Having done this, the $2 \le p < q \le \infty$ cases follow from Proposition 9.19, the $p < 2 < q$ cases follow using the semigroup property of T_ρ (Exercise 9.17), and the $p = q$ cases follow from Exercise 2.33 (or continuity). The proof for $1 \le p < q \le 2$ will be very similar to that of Theorem 9.18 (the $q = 2$ case). As in that proof we may reduce to the case that $\rho = \sqrt{(p-1)/(q-1)}$, $a = 1$, and $b = \epsilon$ satisfies $|\epsilon| < 1$. It then suffices to show

$$\begin{aligned} & \|1 + \rho\epsilon\boldsymbol{x}\|_q^p \le \|1 + \epsilon\boldsymbol{x}\|_p^p \\ \iff\quad & \left(\tfrac{1}{2}(1+\rho\epsilon)^q + \tfrac{1}{2}(1-\rho\epsilon)^q\right)^{p/q} \le \tfrac{1}{2}(1+\epsilon)^p + \tfrac{1}{2}(1-\epsilon)^p \\ \iff\quad & \left(1 + \sum_{k=1}^{\infty} \tbinom{q}{2k}\rho^{2k}\epsilon^{2k}\right)^{p/q} \le 1 + \sum_{k=1}^{\infty} \tbinom{p}{2k}\epsilon^{2k}. \end{aligned} \tag{10.4}$$

Again we used $|\epsilon| < 1$ to drop the absolute value signs and justify the Generalized Binomial Theorem. For each of the binomial coefficients on the left in (10.4) we have

$$\binom{q}{2k} = \frac{q(q-1)(q-2)(q-3)\cdots(q-(2k-2))(q-(2k-1))}{(2k)!} = \frac{q(q-1)(2-q)(3-q)\cdots((2k-2)-q)((2k-1)-q)}{(2k)!} \geq 0.$$

(Here we reversed an even number of signs, since $1 \leq q \leq 2$. We will later do the same when expanding $\binom{p}{2k}$.) Thus we can again employ the inequality $(1+t)^\theta \leq 1+\theta t$ for $t \geq 0$ and $0 \leq \theta \leq 1$ to deduce that the left-hand side of (10.4) is at most

$$1+\sum_{k=1}^{\infty} \frac{p}{q}\binom{q}{2k}\rho^{2k}\epsilon^{2k} = 1+\sum_{k=1}^{\infty} \frac{p}{q}\left(\frac{p-1}{q-1}\right)^k \binom{q}{2k}\epsilon^{2k}.$$

We can now complete the proof of (10.4) by showing the following term-by-term inequality: for all $k \geq 1$,

$$\begin{aligned} &\frac{p}{q}\left(\frac{p-1}{q-1}\right)^k \binom{q}{2k} \leq \binom{p}{2k} \\ \iff \quad & \frac{p}{q}\left(\frac{p-1}{q-1}\right)^k \frac{q(q-1)(2-q)\cdots((2k-1)-q)}{(2k)!} \leq \frac{p(p-1)(2-p)\cdots((2k-1)-p)}{(2k)!} \\ \iff \quad & \frac{2-q}{\sqrt{q-1}}\cdot\frac{3-q}{\sqrt{q-1}}\cdots\frac{(2k-1)-q}{\sqrt{q-1}} \leq \frac{2-p}{\sqrt{p-1}}\cdot\frac{3-p}{\sqrt{p-1}}\cdots\frac{(2k-1)-p}{\sqrt{p-1}}. \end{aligned}$$

And indeed this inequality holds factor-by-factor. This is because $p<q$ and $\frac{j-r}{\sqrt{r-1}}$ is a decreasing function of $r \geq 1$ for all $j \geq 2$, as is evident from $\frac{d}{dr}\frac{j-r}{\sqrt{r-1}} = -\frac{j-2+r}{2(r-1)^{3/2}}$. □

Remark 10.6. The upper-bound $\rho \leq \sqrt{(p-1)/(q-1)}$ in this theorem is best possible; see Exercise 9.10(*b*).

10.2. Hypercontractivity of general random variables

Let's now study hypercontractivity for general random variables. By the end of this section we will have proved the General Hypercontractivity Theorem stated at the beginning of the chapter.

Recall Definition 9.13 which says that $\boldsymbol{X}$ is (p,q,ρ)-hypercontractive if $\mathbf{E}[|\boldsymbol{X}|^q] < \infty$ and

$$\|a+\rho b\boldsymbol{X}\|_q \leq \|a+b\boldsymbol{X}\|_p \quad \text{for all constants } a,b \in \mathbb{R}.$$

(By homogeneity, it's sufficient to check this either with a fixed to 1 or with b fixed to 1.) Let's also collect some additional basic facts regarding the concept:

Fact 10.7. *Suppose $\boldsymbol{X}$ is (p,q,ρ)-hypercontractive ($1 \leq p \leq q \leq \infty$, $0 \leq \rho < 1$). Then:*

(1) $\mathbf{E}[\boldsymbol{X}] = 0$ *(Exercise 9.10).*

(2) $c\boldsymbol{X}$ *is (p,q,ρ)-hypercontractive for any $c \in \mathbb{R}$ (Exercise 9.9).*

(3) $\boldsymbol{X}$ *is (p,q,ρ')-hypercontractive for any $0 \leq \rho' < \rho$ (Exercise 9.11).*

(4) $\rho \leq \sqrt{\frac{p-1}{q-1}}$ *and* $\rho \leq \frac{\|\boldsymbol{X}\|_p}{\|\boldsymbol{X}\|_q}$ *(Exercises 9.10, 9.9).*

Proposition 10.8. *Let $\boldsymbol{X}$ be $(2,q,\rho)$-hypercontractive. Then $\boldsymbol{X}$ is also $(q',2,\rho)$-hypercontractive, where q' is the conjugate Hölder index of q.*

Proof. The deduction is essentially the same as (9.6) from Chapter 9.2. Since $\mathbf{E}[\boldsymbol{X}] = 0$ (Fact 10.7(1)) we have

$$\|a+\rho b\boldsymbol{X}\|_2^2 = \mathbf{E}[a^2 + 2\rho ab\boldsymbol{X} + \rho^2 b^2 \boldsymbol{X}^2] = \mathbf{E}[(a+b\boldsymbol{X})(a+\rho^2 b\boldsymbol{X})].$$

By Hölder's inequality and then the $(2,q,\rho)$-hypercontractivity of $\boldsymbol{X}$ this is at most

$$\|a+b\boldsymbol{X}\|_{q'}\|a+\rho^2 b\boldsymbol{X}\|_q \le \|a+b\boldsymbol{X}\|_{q'}\|a+\rho b\boldsymbol{X}\|_2.$$

Dividing through by $\|a+\rho b\boldsymbol{X}\|_2$ (which can't be 0 unless $\boldsymbol{X} \equiv 0$) gives $\|a+\rho b\boldsymbol{X}\|_2 \le \|a+b\boldsymbol{X}\|_{q'}$ as needed. □

Remark 10.9. The converse does not hold; see Exercise 10.4.

Remark 10.10. As mentioned in Proposition 9.15, the sum of independent hypercontractive random variables is equally hypercontractive. Furthermore, low-degree polynomials of independent hypercontractive random variables are "reasonable". See Exercises 10.2 and 10.3.

Given $\boldsymbol{X}$, p, and q, computing the largest ρ for which $\boldsymbol{X}$ is (p,q,ρ)-hypercontractive can often be quite a chore. However, if you're not overly concerned about constant factors then things become much easier. Let's focus on the most useful case, $p = 2$ and $q > 2$. By Fact 10.7(2) we may assume $\|\boldsymbol{X}\|_2 = 1$. Then we can ask:

Question 10.11. *Let $\mathbf{E}[\boldsymbol{X}] = 0$, $\|\boldsymbol{X}\|_2 = 1$, and assume $\|\boldsymbol{X}\|_q < \infty$. For what ρ is $\boldsymbol{X}$ $(2,q,\rho)$-hypercontractive?*

In this section we'll answer the question by showing that $\rho = \Theta_q(1/\|\boldsymbol{X}\|_q)$ is sufficient. By the second part of Fact 10.7(4), $\rho \le 1/\|\boldsymbol{X}\|_q$ is also necessary. So for a mean-zero random variable $\boldsymbol{X}$, the largest ρ for which $\boldsymbol{X}$ is $(2,q,\rho)$-hypercontractive is always within a constant (depending only on q) of $\frac{\|\boldsymbol{X}\|_2}{\|\boldsymbol{X}\|_q}$.

Let's arrive at this result in steps, introducing the useful techniques of *symmetrization* and *randomization* along the way. When studying hypercontractivity of a random variable $\boldsymbol{X}$, things are much more convenient if $\boldsymbol{X}$ is a *symmetric* random variable, meaning $-\boldsymbol{X}$ has the same distribution as $\boldsymbol{X}$. One advantage of symmetric random variables $\boldsymbol{X}$ is that they have $\mathbf{E}[\boldsymbol{X}^k] = 0$ for all odd $k \in \mathbb{N}$. Using this it is easy to prove (Exercise 10.11) the following fact, similar to Corollary 9.6. (The proof similar to that of Proposition 9.16.)

Proposition 10.12. *Let $\boldsymbol{X}$ be a symmetric random variable with $\|\boldsymbol{X}\|_2 = 1$. Assume $\|\boldsymbol{X}\|_4 = C$ (hence $\boldsymbol{X}$ is "C^4-reasonable"). Then $\boldsymbol{X}$ is $(2,4,\rho)$-hypercontractive if and only if $\rho \le \min(\frac{1}{\sqrt{3}}, \frac{1}{C})$.*

Given a symmetric random variable $\boldsymbol{X}$, the *randomization* trick is to replace $\boldsymbol{X}$ by the identically distributed random variable $\boldsymbol{r}\boldsymbol{X}$, where $\boldsymbol{r} \sim \{-1,1\}$ is an independent uniformly random bit. This trick sometimes lets you reduce a probabilistic statement about $\boldsymbol{X}$ to a related one about $\boldsymbol{r}$.

Theorem 10.13. *Let $\boldsymbol{X}$ be a symmetric random variable with $\|\boldsymbol{X}\|_2 = 1$ and let $\|\boldsymbol{X}\|_q = C$, where $q > 2$. Then $\boldsymbol{X}$ is $(2,q,\rho)$-hypercontractive for $\rho = \frac{1}{C\sqrt{q-1}}$.*

Proof. Let $\boldsymbol{r} \sim \{-1,1\}$ be uniformly random and let $\widetilde{\boldsymbol{X}}$ denote $\boldsymbol{X}/C$. Then for any $a \in \mathbb{R}$,

$$\begin{aligned}
\|a+\rho\boldsymbol{X}\|_q^2 &= \|a+\rho\boldsymbol{r}\boldsymbol{X}\|_q^2 && \text{(by symmetry of } \boldsymbol{X}\text{)}\\
&= \mathop{\mathbf{E}}_{\boldsymbol{X}}\Big[\mathop{\mathbf{E}}_{\boldsymbol{r}}[|a+\rho\boldsymbol{r}\boldsymbol{X}|^q]\Big]^{2/q}\\
&\leq \mathop{\mathbf{E}}_{\boldsymbol{X}}\Big[\mathop{\mathbf{E}}_{\boldsymbol{r}}[|a+\tfrac{1}{C}\boldsymbol{r}\boldsymbol{X}|^2]^{q/2}\Big]^{2/q} && (\boldsymbol{r} \text{ is } (2,q,\tfrac{1}{\sqrt{q-1}})\text{-hypercontractive})\\
&= \mathop{\mathbf{E}}_{\boldsymbol{X}}[(a^2+\widetilde{\boldsymbol{X}}^2)^{q/2}]^{2/q} && \text{(Parseval)}\\
&= \|a^2+\widetilde{\boldsymbol{X}}^2\|_{q/2} && \text{(norm with respect to } \boldsymbol{X}\text{)}\\
&\leq a^2+\|\widetilde{\boldsymbol{X}}^2\|_{q/2} && \text{(triangle inequality for } \|\cdot\|_{q/2}\text{)}\\
&= a^2+\|\widetilde{\boldsymbol{X}}\|_q^2\\
&= a^2+1 = a^2+\mathbf{E}[\boldsymbol{X}^2] = \|a+\boldsymbol{X}\|_2^2,
\end{aligned}$$

where the last step also used $\mathbf{E}[\boldsymbol{X}]=0$. □

Next, if $\boldsymbol{X}$ is not symmetric then we can use a *symmetrization* trick to make it so. One way to do this is to replace $\boldsymbol{X}$ with the symmetric random variable $\boldsymbol{X}-\boldsymbol{X}'$, where $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$. In general $\boldsymbol{X}-\boldsymbol{X}'$ has similar properties to $\boldsymbol{X}$. In particular, if $\mathbf{E}[\boldsymbol{X}]=0$ we can compare norms using the following one-sided bound:

Lemma 10.14. *Let $\boldsymbol{X}$ be a random variable satisfying $\mathbf{E}[\boldsymbol{X}]=0$ and $\|\boldsymbol{X}\|_q<\infty$, where $q \geq 1$. Then for any $a \in \mathbb{R}$,*

$$\|a+\boldsymbol{X}\|_q \leq \|a+\boldsymbol{X}-\boldsymbol{X}'\|_q,$$

where $\boldsymbol{X}'$ denotes an independent copy of $\boldsymbol{X}$.

Proof. We have

$$\|a+\boldsymbol{X}\|_q^q = \mathbf{E}[|a+\boldsymbol{X}|^q] = \mathbf{E}[|a+\boldsymbol{X}-\mathbf{E}[\boldsymbol{X}']|^q],$$

where we used the fact that $\mathbf{E}[\boldsymbol{X}' \mid \boldsymbol{X}] \equiv 0$. But now

$$\mathbf{E}[|a+\boldsymbol{X}-\mathbf{E}[\boldsymbol{X}']|^q] = \mathbf{E}[|\mathbf{E}[a+\boldsymbol{X}-\boldsymbol{X}']|^q] \leq \mathbf{E}[|a+\boldsymbol{X}-\boldsymbol{X}'|^q] = \|a+\boldsymbol{X}-\boldsymbol{X}'\|_q^q,$$

where we used convexity of $t \mapsto |t|^q$ □

A combination of the randomization and symmetrization tricks is to replace an arbitrary random variable $\boldsymbol{X}$ by $\boldsymbol{r}\boldsymbol{X}$, where $\boldsymbol{r} \sim \{-1,1\}$ is an independent uniformly random bit. This often lets you extend results about symmetric random variables to the case of general mean-zero random variables. For example, the following hypercontractivity lemma lets us reduce to the case of a symmetric random variable while only "spending" a factor of $\frac{1}{2}$:

Lemma 10.15. *Let $\boldsymbol{X}$ be a random variable satisfying $\mathbf{E}[\boldsymbol{X}] = 0$ and $\|\boldsymbol{X}\|_q < \infty$, where $q \geq 1$. Then for any $a \in \mathbb{R}$,*

$$\|a + \tfrac{1}{2}\boldsymbol{X}\|_q \leq \|a + \boldsymbol{r}\boldsymbol{X}\|_q,$$

where $\boldsymbol{r} \sim \{-1,1\}$ is an independent uniformly random bit.

Proof. Letting $\boldsymbol{X}'$ be an independent copy of $\boldsymbol{X}$ we have

$$\begin{aligned}
\|a + \tfrac{1}{2}\boldsymbol{X}\|_q &\leq \|a + \tfrac{1}{2}\boldsymbol{X} - \tfrac{1}{2}\boldsymbol{X}'\|_q && \text{(Lemma 10.14 applied to } \tfrac{1}{2}\boldsymbol{X}\text{)}\\
&\leq \|a + \boldsymbol{r}(\tfrac{1}{2}\boldsymbol{X} - \tfrac{1}{2}\boldsymbol{X}')\|_q && \text{(since } \tfrac{1}{2}\boldsymbol{X} - \tfrac{1}{2}\boldsymbol{X}' \text{ is symmetric)}\\
&= \|\tfrac{1}{2}a + \tfrac{1}{2}\boldsymbol{r}\boldsymbol{X} + \tfrac{1}{2}a - \tfrac{1}{2}\boldsymbol{r}\boldsymbol{X}'\|_q &&\\
&\leq \|\tfrac{1}{2}a + \tfrac{1}{2}\boldsymbol{r}\boldsymbol{X}\|_q + \|\tfrac{1}{2}a - \tfrac{1}{2}\boldsymbol{r}\boldsymbol{X}'\|_q && \text{(triangle inequality for } \|\cdot\|_q\text{)}\\
&= \|\tfrac{1}{2}a + \tfrac{1}{2}\boldsymbol{r}\boldsymbol{X}\|_q + \|\tfrac{1}{2}a + \tfrac{1}{2}\boldsymbol{r}\boldsymbol{X}'\|_q && (-\boldsymbol{r} \text{ distributed as } \boldsymbol{r})\\
&= \|a + \boldsymbol{r}\boldsymbol{X}\|_q. && \square
\end{aligned}$$

By employing these randomization/symmetrization techniques we obtain a $(2,q)$-hypercontractivity statement for all mean-zero random variables $\boldsymbol{X}$ with $\frac{\|\boldsymbol{X}\|_q}{\|\boldsymbol{X}\|_2}$ bounded, giving a good answer to Question 10.11:

Theorem 10.16. *Let $\boldsymbol{X}$ satisfy $\mathbf{E}[\boldsymbol{X}] = 0$, $\|\boldsymbol{X}\|_2 = 1$, $\|\boldsymbol{X}\|_q = C$, where $q > 2$. Then $\boldsymbol{X}$ is $(2,q,\frac{1}{2}\rho)$-hypercontractive for $\rho = \frac{1}{\sqrt{q-1}\|\boldsymbol{X}\|_q}$. (If $\boldsymbol{X}$ is symmetric, then the factor of $\frac{1}{2}$ may be omitted.)*

Proof. By Lemma 10.15 we have

$$\|a + \tfrac{1}{2}\rho\boldsymbol{X}\|_q^2 \leq \|a + \rho\boldsymbol{r}\boldsymbol{X}\|_q^2.$$

Since $\boldsymbol{r}\boldsymbol{X}$ is a symmetric random variable satisfying $\|\boldsymbol{r}\boldsymbol{X}\|_2 = 1$, $\|\boldsymbol{r}\boldsymbol{X}\|_q = C$, Theorem 10.13 implies

$$\|a + \rho\boldsymbol{r}\boldsymbol{X}\|_q^2 \leq \|a + \boldsymbol{r}\boldsymbol{X}\|_2^2 = a^2 + 1 = \|a + \boldsymbol{X}\|_2^2.$$

This completes the proof. $\square$

If $\boldsymbol{X}$ is a discrete random variable then instead of computing $\frac{\|\boldsymbol{X}\|_2}{\|\boldsymbol{X}\|_q}$ it can sometimes be convenient to use a bound based on the minimum value of $\boldsymbol{X}$'s probability mass function. The following is a simple generalization of Proposition 9.5, whose proof is left for Exercise 10.17:

Proposition 10.17. *Let $\boldsymbol{X}$ be a discrete random variable with probability mass function π. Write*

$$\lambda = \min(\pi) = \min_{x \in \mathrm{range}(\boldsymbol{X})} \{\mathbf{Pr}[\boldsymbol{X} = x]\}.$$

Then for any $q > 2$ we have $\|\boldsymbol{X}\|_q \le (1/\lambda)^{1/2-1/q} \cdot \|\boldsymbol{X}\|_2$.

As a consequence of Theorem 10.16, if in addition $\mathbf{E}[\boldsymbol{X}] = 0$ then $\boldsymbol{X}$ is $(2, q, \frac{1}{2}\rho)$-hypercontractive for $\rho = \frac{1}{\sqrt{q-1}} \cdot \lambda^{1/2-1/q}$, and also $(q', 2, \frac{1}{2}\rho)$-hypercontractive by Proposition 10.8. (If $\boldsymbol{X}$ is symmetric then the factor of $\frac{1}{2}$ may be omitted.)

For each $q > 2$, the value $\rho = \Theta_q(\lambda^{1/2-1/q})$ in Proposition 10.17 has the optimal dependence on λ, up to a constant. In fact, a perfectly sharp version of Proposition 10.17 is known. The most important case is when $\boldsymbol{X}$ is a λ-biased bit; more precisely, when $\boldsymbol{X} = \phi(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \sim \pi_\lambda$ in the notation of Definition 8.39. In that case, the below theorem (whose very technical proof is left to Exercises 10.19–10.21) is due to Latała and Oleszkiewicz [LO94]. The case of general discrete random variables is a reduction to the two-valued case due to Wolff [Wol07].

Theorem 10.18. *Let $\boldsymbol{X}$ be a mean-zero discrete random variable and let $\lambda < 1/2$ be the least value of its probability mass function, as in Proposition 10.17. Then for $q > 2$ it holds that $\boldsymbol{X}$ is $(2, q, \rho)$-hypercontractive and $(q', 2, \rho)$-hypercontractive for*

$$\rho = \sqrt{\frac{\exp(u/q) - \exp(-u/q)}{\exp(u/q') - \exp(-u/q')}} = \sqrt{\frac{\sinh(u/q)}{\sinh(u/q')}}, \text{ with } u \text{ defined by } \exp(-u) = \tfrac{\lambda}{1-\lambda}. \tag{10.5}$$

This value of ρ is optimal, even under the assumption that $\boldsymbol{X}$ is two-valued.

Remark 10.19. It's not hard to see that for $\lambda \to 1/2$ (hence $u \to 0$) we get $\rho \to \sqrt{\frac{1/q-(-1/q)}{1/q'-(-1/q')}} = \frac{1}{\sqrt{q-1}}$, consistent with the Two-Point Inequality from Section 10.1. Also, for $\lambda \to 0$ (hence $u \to \infty$) we get $\rho \sim \sqrt{\frac{\lambda^{-1/q}}{\lambda^{-1/q'}}} = \lambda^{1/2-1/q}$, showing that Proposition 10.17 is sharp up to a constant. Exercise 10.18 asks you to investigate the function defining ρ in (10.5) more carefully. In particular, you'll show that $\rho \ge \frac{1}{\sqrt{q-1}} \cdot \lambda^{1/2-1/q}$ holds for all λ. Hence we can omit the factor of $\frac{1}{2}$ from the simpler bound in Proposition 10.17 even for nonsymmetric random variables.

Corollary 10.20. *Let (Ω, π) be a finite probability space, $|\Omega| \ge 2$, in which every outcome has probability at least λ. Let $f \in L^2(\Omega, \pi)$. Then for any $q > 2$ and $0 \le \rho \le \frac{1}{\sqrt{q-1}} \cdot \lambda^{1/2-1/q}$,*

$$\|\mathrm{T}_\rho f\|_q \le \|f\|_2 \quad \textit{and} \quad \|\mathrm{T}_\rho f\|_2 \le \|f\|_{q'}.$$

Proof. Recalling Chapter 8.3, this follows from the decomposition $f(x) = f^{\emptyset} + f^{=\{1\}}$, under which $\mathrm{T}_\rho f = f^{\emptyset} + \rho f^{=\{1\}}$. Note that for $\boldsymbol{x} \sim \pi$ the random variable $f^{=\{1\}}(\boldsymbol{x})$ has mean zero, and the least value of its probability mass function is at least λ. □

The General Hypercontractivity Theorem stated at the beginning of the chapter now follows by applying the Hypercontractivity Induction Theorem from Section 10.1.

10.3. Applications of general hypercontractivity

In this section we will collect some applications of the General Hypercontractivity Theorem, including generalizations of the facts from Section 9.5. We begin by bounding the q-norms of low-degree functions. The proof is essentially the same as that of Theorem 9.21; see Exercise 10.28.

Theorem 10.21. *In the setting of the General Hypercontractivity Theorem, if f has degree at most k, then*

$$\|f\|_q \leq (\sqrt{q-1} \cdot \lambda^{1/q-1/2})^k \|f\|_2.$$

Next we turn to an analogue of Theorem 9.22, getting a relationship between the 2-norm and the 1-norm for low-degree functions. The proof (Exercise 10.31) needs $(2,q,\rho)$-hypercontractivity with q tending to 2, so to get the most elegant statement requires appealing to the sharp bound from Theorem 10.18:

Theorem 10.22. *In the setting of the General Hypercontractivity Theorem, if f has degree at most k, then*

$$\|f\|_2 \leq c(\lambda)^k \|f\|_1, \quad \textit{where } c(\lambda) = \sqrt{\tfrac{1-\lambda}{\lambda}}^{1/(1-2\lambda)}.$$

We have $c(\lambda) \sim 1/\sqrt{\lambda}$ as $\lambda \to 0$, $c(\lambda) \to e$ as $\lambda \to \frac{1}{2}$, and in general, $c(\lambda) \leq e/\sqrt{2\lambda}$.

Just as in Chapter 9.5 we obtain (Exercise 10.32) the following as a corollary:

Theorem 10.23. *In the setting of the General Hypercontractivity Theorem, if f is a nonconstant function of degree at most k, then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \pi^{\otimes n}}[f(\boldsymbol{x}) > \mathbf{E}[f]] \geq \tfrac{1}{4}(e^2/2\lambda)^{-k} \geq (15/\lambda)^{-k}.$$

Extending Theorem 9.23, the concentration bound for degree-k functions, is straightforward (see Exercise 10.33). We again get that the probability of exceeding t standard deviations decays like $\exp(-\Theta(t^{2/k}))$, though the constant in the $\Theta(\cdot)$ is linear in λ:

Theorem 10.24. *In the setting of the General Hypercontractivity Theorem, if f has degree at most k, then for any $t \geq \sqrt{2e/\lambda}^k$,*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi^{\otimes n}}[|f(\boldsymbol{x})| \geq t\|f\|_2] \leq \lambda^k \exp\left(-\tfrac{k}{2e}\lambda t^{2/k}\right).$$

Next, we give a generalization of the Small-Set Expansion Theorem, the proof being left for Exercise 10.34.

Theorem 10.25. *Let (Ω,π) be a finite probability space, $|\Omega| \geq 2$, in which every outcome has probability at least λ. Let $A \subseteq \Omega^n$ have "volume" α; i.e., suppose $\mathbf{Pr}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\boldsymbol{x} \in A] = \alpha$. Let $q \geq 2$. Then for any*

$$0 \leq \rho \leq \tfrac{1}{q-1} \cdot \lambda^{1-2/q}$$

(or even ρ as large as the square of the quantity in Theorem 10.18) we have

$$\mathbf{Stab}_\rho[1_A] = \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim\pi^{\otimes n}\\ \boldsymbol{y}\sim N_\rho(\boldsymbol{x})}}[\boldsymbol{x} \in A, \boldsymbol{y} \in A] \leq \alpha^{2-2/q}.$$

Similarly, we can generalize Corollary 9.25, bounding the stable influence of a coordinate by a power of the usual influence:

Theorem 10.26. *In the setting of Theorem 10.25, if $f : \Omega^n \to \{-1,1\}$, then*

$$\rho\mathbf{Inf}_i^{(\rho)}[f] \leq \mathbf{Inf}_i[f]^{2-2/q}.$$

for all $i \in [n]$. In particular, by selecting $q = 4$ we get

$$\sum_{S\ni i} (\sqrt{\lambda}/3)^{|S|}\|f^{=S}\|_2^2 \leq \mathbf{Inf}_i[f]^{3/2}. \tag{10.6}$$

Proof. Applying the General Hypercontractivity Theorem to $\mathrm{L}_i f$ and squaring we get

$$\|\mathrm{T}_{\sqrt{\rho}}\mathrm{L}_i f\|_2^2 \leq \|\mathrm{L}_i f\|_{q'}^2.$$

By definition, the left-hand side is $\rho\mathbf{Inf}_i^{(\rho)}[f]$. The right-hand side is $(\|\mathrm{L}_i f\|_{q'}^{q'})^{2-2/q}$, and $\|\mathrm{L}_i f\|_{q'}^{q'} \leq \mathbf{Inf}_i[f]$ by Exercise 8.10(*b*). □

The KKL Edge-Isoperimetic Theorem in this setting now follows by an almost verbatim repetition of the proof from Chapter 9.6.

KKL Isoperimetric Theorem for general product space domains. *In the setting of the General Hypercontractivity Theorem, suppose f has range $\{-1,1\}$ and is nonconstant. Let $\widetilde{\mathbf{I}}[f] = \mathbf{I}[f]/\mathbf{Var}[f] \geq 1$. Then*

$$\mathbf{MaxInf}[f] \geq \tfrac{1}{\widetilde{\mathbf{I}}[f]^2} \cdot (9/\lambda)^{-\widetilde{\mathbf{I}}[f]}.$$

As a consequence, $\mathbf{MaxInf}[f] \geq \Omega(\frac{1}{\log(1/\lambda)}) \cdot \mathbf{Var}[f] \cdot \frac{\log n}{n}$.

Proof. (Cf. Exercise 9.29.) The proof is essentially identical to the one in Chapter 9.6, but using (10.6) from Theorem 10.26. Summing this inequality over all $i \in [n]$ yields

$$\sum_{S \subseteq [n]} |S| (\sqrt{\lambda}/3)^{|S|} \|f^{=S}\|_2^2 \leq \sum_{i=1}^{n} \mathbf{Inf}_i[f]^{3/2} \leq \mathbf{MaxInf}[f]^{1/2} \cdot \mathbf{I}[f]. \tag{10.7}$$

On the left-hand side above we will drop the factor of $|S|$ for $|S| > 0$. We also introduce the set-valued random variable $\boldsymbol{S}$ defined by $\mathbf{Pr}[\boldsymbol{S} = S] = \|f^{=S}\|_2^2/\mathbf{Var}[f]$ for $S \neq \emptyset$. Note that $\mathbf{E}[|\boldsymbol{S}|] = \widetilde{\mathbf{I}}[f]$. Thus

$$\text{LHS(10.7)} \geq \mathbf{Var}[f] \cdot \mathop{\mathbf{E}}_{\boldsymbol{S}}[(\sqrt{\lambda}/3)^{|\boldsymbol{S}|}] \geq \mathbf{Var}[f] \cdot (\sqrt{\lambda}/3)^{\mathbf{E}[|\boldsymbol{S}|]} = \mathbf{Var}[f] \cdot (\sqrt{\lambda}/3)^{\widetilde{\mathbf{I}}[f]},$$

where we used that $s \mapsto (\sqrt{\lambda}/3)^s$ is convex. The first statement of the theorem now follows after rearrangement. As for the second statement, there is some universal $c > 0$ such that

$$\widetilde{\mathbf{I}}[f] \leq c \cdot \tfrac{1}{\log(1/\lambda)} \cdot \log n \quad \Longrightarrow \quad \tfrac{1}{\widetilde{\mathbf{I}}[f]^2} \cdot (9/\lambda)^{-\widetilde{\mathbf{I}}[f]} = O(1/\lambda)^{-\widetilde{\mathbf{I}}[f]} \geq \frac{1}{\sqrt{n}},$$

say, in which case our lower bound for $\mathbf{MaxInf}[f]$ is $\frac{1}{\sqrt{n}} \gg \frac{\log n}{n}$. On the other hand,

$$\widetilde{\mathbf{I}}[f] \geq c \cdot \tfrac{1}{\log(1/\lambda)} \cdot \log n \quad \Longrightarrow \quad \mathbf{I}[f] \geq \Omega(\tfrac{1}{\log(1/\lambda)}) \cdot \mathbf{Var}[f] \cdot \log n,$$

in which case even the average influence of f is $\Omega(\frac{1}{\log(1/\lambda)}) \cdot \mathbf{Var}[f] \cdot \frac{\log n}{n}$. □

Similarly, essentially no extra work is required to generalize Theorem 9.28 and Friedgut's Junta Theorem to general product space domains; see Exercise 10.35. For example, we have:

Friedgut's Junta Theorem for general product space domains. *In the setting of the General Hypercontractivity Theorem, if f has range $\{-1,1\}$ and $0 < \epsilon \leq 1$, then f is ϵ-close to a $(1/\lambda)^{O(\mathbf{I}[f]/\epsilon)}$-junta $h : \Omega^n \to \{-1,1\}$ (i.e., $\mathbf{Pr}_{\boldsymbol{x} \sim \pi^{\otimes n}}[f(\boldsymbol{x}) \neq h(\boldsymbol{x})] \leq \epsilon$).*

We conclude this section by establishing "sharp thresholds" – in the sense of Chapter 8.4 – for monotone transitive-symmetric functions with critical probability in the range $[1/n^{o(1)}, 1 - 1/n^{o(1)}]$. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a nonconstant monotone function and define the (strictly increasing) curve $F : [0,1] \to [0,1]$ by $F(p) = \mathbf{Pr}_{\boldsymbol{x} \sim \pi_p^{\otimes n}}[f(\boldsymbol{x}) = -1]$. Recall that the critical probability p_c is defined to be the value such that $F(p_c) = 1/2$; equivalently, such that $\mathbf{Var}[f^{(p_c)}] = 1$. Recall also the Margulis–Russo Formula, which says that

$$\frac{d}{dp} F(p) = \frac{1}{\sigma^2} \cdot \mathbf{I}[f^{(p)}],$$

where

$$\sigma^2 = \sigma^2(p) = \mathop{\mathbf{Var}}_{\pi_p}[\boldsymbol{x}_i] = 4p(1-p) = \Theta(\min(p, 1-p)).$$

Remark 10.27. Since we will not be concerned with constant factors, it's helpful in the following discussion to mentally replace σ^2 with $\min(p, 1-p)$. In fact it's even more helpful to always assume $p \leq 1/2$ and replace σ^2 with p.

Now suppose f is a transitive-symmetric function, e.g., a graph property. This means that all of its influences are the same, i.e.,

$$\mathbf{Inf}_i[f^{(p)}] = \mathbf{MaxInf}[f^{(p)}] = \frac{1}{n}\mathbf{I}[f^{(p)}]$$

for all $i \in [n]$. It thus follows from the KKL Theorem for general product spaces that

$$\mathbf{I}[f^{(p)}] \geq \Omega\big(\tfrac{1}{\log(1/\min(p,1-p))}\big) \cdot \mathbf{Var}[f^{(p)}] \cdot \log n;$$

hence

$$\frac{d}{dp}F(p) \geq \mathbf{Var}[f^{(p)}] \cdot \Omega\big(\tfrac{1}{\sigma^2 \ln(e/\sigma^2)}\big) \cdot \log n. \tag{10.8}$$

(As mentioned in Remark 10.27, assuming $p \leq 1/2$ you can read $\sigma^2 \ln(e/\sigma^2)$ as $p\log(1/p)$.)

If we take $p = p_c$ in inequality (10.8) we conclude that $F(p)$ has a large derivative at its critical probability: $F'(p_c) \geq \Omega(\frac{1}{p_c \log(1/p_c)}) \cdot \log n$, assuming $p_c \leq 1/2$. In particular if $\log(1/p_c) \ll \log n$ – that is, $p_c > 1/n^{o(1)}$ – then $F'(p_c) = \omega(\frac{1}{p_c})$. This suggests that f has a "sharp threshold"; i.e., $F(p)$ jumps from near 0 to near 1 in an interval of the form $p_c(1 \pm o(1))$. However, largeness of $F'(p_c)$ is not quite enough to establish a sharp threshold (see Exercise 8.30); we need to have $F'(p)$ large *throughout* the range of p near p_c where $\mathbf{Var}[f^{(p)}]$ is large. Happily, inequality (10.8) provides precisely this.

Remark 10.28. Even if we are only concerned about monotone functions f with $p_c = 1/2$, we still need the KKL Theorem for general product spaces to establish a sharp threshold. Though $F'(1/2) \geq \Omega(\log n)$ can be derived using just the uniform-distribution KKL Theorem from Chapter 9.6, we also need to know that $F'(p) \geq \Omega(\log n)$ continues to hold for $p = 1/2 \pm O(1/\log n)$.

Making the above ideas precise, we can establish the following result of Friedgut and Kalai [FK96] (cf. Exercises 8.28, 8.29):

Theorem 10.29. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a nonconstant, monotone, transitive-symmetric function and let $F : [0,1] \to [0,1]$ be the strictly increasing function defined by $F(p) = \mathbf{Pr}_{\boldsymbol{x} \sim \pi_p^{\otimes n}}[f(\boldsymbol{x}) = -1]$. Let p_c be the critical probability such that $F(p_c) = 1/2$ and assume without loss of generality that $p_c \leq 1/2$. Fix $0 < \epsilon < 1/4$ and let*

$$\eta = B\log(1/\epsilon) \cdot \frac{\log(1/p_c)}{\log n},$$

where $B > 0$ is a certain universal constant. Then assuming $\eta \leq 1/2$,

$$F(p_c \cdot (1-\eta)) \leq \epsilon, \qquad F(p_c \cdot (1+\eta)) \geq 1-\epsilon.$$

Proof. Let p be in the range $p_c \cdot (1 \pm \eta)$. By the assumption $\eta \leq 1/2$ we also have $\frac{1}{2}p_c \leq p \leq \frac{3}{2}p_c \leq \frac{3}{4}$. It follows that the quantity $\sigma^2 \ln(e/\sigma^2)$ in the KKL corollary (10.8) is within a universal constant factor of $p_c \log(1/p_c)$. Thus for all p in the range $p_c \cdot (1 \pm \eta)$ we obtain

$$F'(p) \geq \mathbf{Var}[f^{(p)}] \cdot \Omega\big(\tfrac{1}{p_c \log(1/p_c)}\big) \cdot \log n.$$

Using $\mathbf{Var}[f^{(p)}] = 4F(p)(1-F(p))$, the definition of η, and a suitable choice of B, this is equivalent to

$$F'(p) \geq \frac{2\ln(1/2\epsilon)}{\eta p_c} F(p)(1-F(p)). \tag{10.9}$$

We now show that (10.9) implies that $F(p_c - \eta p_c) \leq \epsilon$ and leave the implication $F(p_c + \eta p_c) \geq 1-\epsilon$ to Exercise 10.36. For $p \leq p_c$ we have $1 - F(p) \geq 1/2$ and hence

$$F'(p) \geq \frac{\ln(1/2\epsilon)}{\eta p_c} F(p) \quad \implies \quad \frac{d}{dp} \ln F(p) = \frac{F'(p)}{F(p)} \geq \frac{\ln(1/2\epsilon)}{\eta p_c}.$$

It follows that

$$\ln F(p_c - \eta p_c) \leq \ln F(p_c) - \ln(1/2\epsilon) = \ln(1/2) - \ln(1/2\epsilon) = \ln \epsilon;$$

i.e., $F(p_c - \eta p_c) \leq \epsilon$ as claimed. □

This proof establishes that every monotone transitive-symmetric function with critical probability at least $1/n^{o(1)}$ (and at most $1 - 1/n^{o(1)}$) has a sharp threshold. Unfortunately, the restriction on the critical probability can't be removed. The simplest example illustrating this is the logical OR function $\mathrm{OR}_n : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ (equivalently, the graph property of containing an edge), which has critical probability $p_c \sim \frac{\ln 2}{n}$. Even though OR_n is transitive-symmetric, it has constant total influence at its critical probability, $\mathbf{I}[\mathrm{OR}_n^{(p_c)}] \sim 2\ln 2$. Indeed, OR_n doesn't have a sharp threshold; i.e., it's not true that $\mathbf{Pr}_{\pi_p}[\mathrm{OR}_n(\boldsymbol{x}) = \mathsf{True}] = 1 - o(1)$ for $p = p_c(1 + o(1))$. For example, if $\boldsymbol{x}$ is drawn from the $(2p_c)$-biased distribution we still just have $\mathbf{Pr}[\mathrm{OR}_n(\boldsymbol{x}) = \mathsf{True}] \approx 3/4$. On the other hand, most "interesting" monotone transitive-symmetric functions *do* have a sharp threshold; in Section 10.5 we'll derive a more sophisticated method for establishing this.

10.4. More on randomization/symmetrization

In Section 10.3 we collected a number of consequences of the General Hypercontractivity Theorem for functions $f \in L^2(\Omega^n, \pi^{\otimes n})$. All of these had a dependence on "λ", the least probability of an outcome under π. This can sometimes be quite expensive; for example, the KKL Theorem and its consequence Theorem 10.29 are trivialized when $\lambda = 1/n^{\Theta(1)}$.

However, as mentioned in Section 10.2, when working with *symmetric* random variables $\boldsymbol{X}$, the "randomization" trick sometimes lets us reduce to the analysis of uniformly random ± 1 bits (which have $\lambda = 1/2$). Further, Lemma 10.15 suggests a way of "symmetrizing" general mean-zero random variables (at least if we don't mind applying $\mathrm{T}_{\frac{1}{2}}$). In this section we will develop the randomization/symmetrization technique more thoroughly and see an application: bounding the $L^p \to L^p$ norm of the "low-degree projection" operator.

Informally, applying the randomization/symmetrization technique to $f \in L^2(\Omega^n, \pi^{\otimes n})$ means introducing n independent uniformly random bits $\boldsymbol{r} = (\boldsymbol{r}_1, \dots, \boldsymbol{r}_n) \sim \{-1,1\}^n$ and then "multiplying the ith input to f by $\boldsymbol{r}_i$". Of course Ω is just an abstract set so this doesn't quite make sense. What we really mean is "multiplying $\mathrm{L}_i f$, the ith part of f's Fourier expansion (orthogonal decomposition), by $\boldsymbol{r}_i$". Let's see some examples:

Example 10.30. Let $f : \{-1,1\}^n \to \mathbb{R}$ be a usual Boolean function with Fourier expansion

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

Its randomization/symmetrization will be the function

$$\widetilde{f}(r, x) = \sum_{S \subseteq [n]} \widehat{f}(S) \prod_{i \in S} r_i x_i = \sum_{S \subseteq [n]} \widehat{f}(S) x^S r^S.$$

The key observation is that for random inputs $\boldsymbol{x}, \boldsymbol{r} \sim \{-1,1\}^n$, the random variables $f(\boldsymbol{x})$ and $\widetilde{f}(\boldsymbol{r}, \boldsymbol{x})$ are *identically distributed*. This is simply because $\boldsymbol{x}_i$ is a symmetric random variable, so it has the same distribution as $\boldsymbol{r}_i \boldsymbol{x}_i$.

Example 10.31. Let's return to Examples 8.10 and 8.15 from Chapter 8.1. Here we had $\Omega = \{a, b, c\}$ with π the uniform distribution, and we defined a certain Fourier basis $\{\phi_0 \equiv 1, \phi_1, \phi_2\}$. A typical $f : \Omega^3 \to \mathbb{R}$ here might look like

$$\begin{aligned} f(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = \tfrac{1}{3} - \tfrac{1}{4} \cdot \phi_1(\boldsymbol{x}_1) + \tfrac{3}{2} \cdot \phi_2(\boldsymbol{x}_1) + \cdot \phi_1(\boldsymbol{x}_2) + \tfrac{1}{2} \cdot \phi_2(\boldsymbol{x}_2) - \tfrac{2}{3} \cdot \phi_2(\boldsymbol{x}_3) \\ + \tfrac{1}{6} \cdot \phi_1(\boldsymbol{x}_1) \cdot \phi_2(\boldsymbol{x}_3) + \tfrac{1}{8} \cdot \phi_1(\boldsymbol{x}_2) \cdot \phi_1(\boldsymbol{x}_3) \\ - \tfrac{1}{10} \cdot \phi_1(\boldsymbol{x}_1) \cdot \phi_2(\boldsymbol{x}_2) \cdot \phi_3(\boldsymbol{x}_3) + \tfrac{1}{5} \cdot \phi_2(\boldsymbol{x}_1) \cdot \phi_2(\boldsymbol{x}_2) \cdot \phi_2(\boldsymbol{x}_3). \end{aligned}$$

The randomization/symmetrization of this function would be the following function $\widetilde{f} \in L^2(\{-1,1\}^3 \times \Omega^3, \pi_{1/2}^{\otimes 3} \otimes \pi^{\otimes 3})$:

$$\begin{aligned} \widetilde{f}(\boldsymbol{r}, \boldsymbol{x}) = \tfrac{1}{3} - \tfrac{1}{4}\phi_1(\boldsymbol{x}_1) \cdot \boldsymbol{r}_1 + \tfrac{3}{2}\phi_2(\boldsymbol{x}_1) \cdot \boldsymbol{r}_1 + \phi_1(\boldsymbol{x}_2) \cdot \boldsymbol{r}_2 + \tfrac{1}{2}\phi_2(\boldsymbol{x}_2) \cdot \boldsymbol{r}_2 - \tfrac{2}{3}\phi_2(\boldsymbol{x}_3) \cdot \boldsymbol{r}_3 \\ + \tfrac{1}{6}\phi_1(\boldsymbol{x}_1) \cdot \phi_2(\boldsymbol{x}_3) \cdot \boldsymbol{r}_1 \boldsymbol{r}_3 + \tfrac{1}{8}\phi_1(\boldsymbol{x}_2) \cdot \phi_1(\boldsymbol{x}_3) \cdot \boldsymbol{r}_2 \boldsymbol{r}_3 \\ - \tfrac{1}{10}\phi_1(\boldsymbol{x}_1) \cdot \phi_2(\boldsymbol{x}_2) \cdot \phi_3(\boldsymbol{x}_3) \cdot \boldsymbol{r}_1 \boldsymbol{r}_2 \boldsymbol{r}_3 + \tfrac{1}{5}\phi_2(\boldsymbol{x}_1) \cdot \phi_2(\boldsymbol{x}_2) \cdot \phi_2(\boldsymbol{x}_3) \cdot \boldsymbol{r}_1 \boldsymbol{r}_2 \boldsymbol{r}_3. \end{aligned}$$

There's no obvious way to compare the distributions of $f(\boldsymbol{x})$ and $\widetilde{f}(\boldsymbol{r}, \boldsymbol{x})$. However, looking carefully at Example 8.10 we see that the basis function ϕ_2 has the property that $\phi_2(\boldsymbol{x}_i)$ is a symmetric real random variable when $\boldsymbol{x}_i \sim \pi$.

In particular, $\boldsymbol{r}_i \cdot \phi_2(\boldsymbol{x}_i)$ has the same distribution as $\phi_2(\boldsymbol{x}_i)$. Therefore if $g \in L^2(\Omega^n, \pi^{\otimes n})$ has the lucky property that its Fourier expansion happens to only use ϕ_2 and never uses ϕ_1, then we *do* have that $g(\boldsymbol{x})$ and $\widetilde{g}(\boldsymbol{r}, \boldsymbol{x})$ are identically distributed.

Let's give a formal definition of randomization/symmetrization.

Definition 10.32. Let $f \in L^2(\Omega^n, \pi^{\otimes n})$. The *randomization / symmetrization* of f is the function $\widetilde{f} \in L^2(\{-1,1\}^n \times \Omega^n, \pi_{1/2}^{\otimes n} \otimes \pi^{\otimes n})$ defined by

$$\widetilde{f}(\boldsymbol{r}, \boldsymbol{x}) = \sum_{S \subseteq [n]} \boldsymbol{r}^S f^{=S}(\boldsymbol{x}), \tag{10.10}$$

where we recall the notation $r^S = \prod_{i \in S} r_i$.

Remark 10.33. Another way of defining $\widetilde{f}$ is to stipulate that for each $x \in \Omega^n$, the function $\widetilde{f}_{|x} : \{-1,1\}^n \to \mathbb{R}$ is defined to be the Boolean function whose Fourier coefficient on S is $f^{=S}(x)$. (This is more evident from (10.10) if you swap the positions of $\boldsymbol{r}^S$ and $f^{=S}(\boldsymbol{x})$.)

In light of this remark, the basic Parseval formula for Boolean functions implies that for all $x \in \Omega^n$,

$$\|\widetilde{f}_{|x}\|_{2,\boldsymbol{r}}^2 = \sum_{S \subseteq [n]} f^{=S}(x)^2.$$

(The notation $\|\cdot\|_{2,\boldsymbol{r}}$ emphasizes that the norm is computed with respect to the random inputs $\boldsymbol{r}$.) If we take the expectation of the above over $\boldsymbol{x} \sim \pi^{\otimes n}$, the left-hand side becomes $\|\widetilde{f}\|_{2,\boldsymbol{r},\boldsymbol{x}}^2$ and the right-hand side becomes $\|f\|_{2,\boldsymbol{x}}^2$, by Parseval's formula for $L^2(\Omega^n, \pi^{\otimes n})$. Thus:

Proposition 10.34. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then $\|\widetilde{f}\|_2 = \|f\|_2$.*

Thus randomization/symmetrization doesn't change 2-norms. What about q-norms for $q \neq 2$? As discussed in Examples 10.30 and 10.31, if f has the lucky property that its Fourier expansion only contains symmetric basis functions then $\widetilde{f}(\boldsymbol{r}, \boldsymbol{x})$ and $f(\boldsymbol{x})$ have identical distributions, so their q-norms are identical. The essential feature of the randomization/symmetrization technique is that even for general f the q-norms don't change much – if you are willing to apply T_ρ for some constant ρ:

Theorem 10.35. *For $f \in L^2(\Omega^n, \pi^{\otimes n})$ and $q > 1$,*

$$\|\mathrm{T}_{\frac{1}{2}} f\|_q \leq \|\widetilde{f}\|_q \leq \|\mathrm{T}_{c_q^{-1}} f\|_q. \tag{10.11}$$

Equivalently,

$$\|\widetilde{\mathrm{T}_{c_q} f}\|_q \leq \|f\|_q \leq \|\widetilde{\mathrm{T}_2 f}\|_q.$$

Here $0 < c_q \leq 1$ is a constant depending only on q; in particular, we may take $c_4 = c_{4/3} = \frac{2}{5}$.

The two inequalities in (10.11) are not too difficult to prove; for example, you might already correctly guess that the left-hand inequality follows from our first randomization/symmetrization Lemma 10.15 and an induction. We'll give the proofs at the end of this section. But first, let's illustrate how you might use them by solving the following basic problem concerning low-degree projections:

Question 10.36. *Let $k \in \mathbb{N}$, let $1 < q < \infty$, and let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Can $\|f^{\le k}\|_q$ be much larger than $\|f\|_q$? To put the question in reverse, suppose $g \in L^2(\Omega^n, \pi^{\otimes n})$ has degree at most k; is it possible to make the q-norm of g much smaller by adding terms of degree exceeding k to its Fourier expansion?*

The question has a simple answer if $q = 2$: in this case we have $\|f^{\le k}\|_2 \le \|f\|_2$ always. This follows from Paresval:

$$\|f^{\le k}\|_2^2 = \sum_{j=0}^{k} \mathbf{W}^j[f] \le \sum_{j=0}^{n} \mathbf{W}^j[f] = \|f\|_2^2. \tag{10.12}$$

When $q \neq 2$ things are not so simple, so let's first consider the most familiar setting of $\Omega = \{-1,1\}$, $\pi = \pi_{1/2}$. In this case we can relate the q-norm and the 2-norm via the Hypercontractivity Theorem:

Proposition 10.37. *Let $k \in \mathbb{N}$ and let $g : \{-1,1\}^n \to \mathbb{R}$. Then for $q \ge 2$ we have $\|g^{\le k}\|_q \le \sqrt{q-1}^k \|g\|_q$ and for $1 < q \le 2$ we have $\|g^{\le k}\|_q \le (1/\sqrt{q-1})^k \|g\|_q$.*

This proposition is an easy consequence of the Hypercontractivity Theorem and already appeared as Exercise 9.8. The simplest case, $q = 4$, follows from the Bonami Lemma alone:

$$\|g^{\le k}\|_4 \le \sqrt{3}^k \|g^{\le k}\|_2 \le \sqrt{3}^k \|g\|_2 \le \sqrt{3}^k \|g\|_4. \tag{10.13}$$

Now let's consider functions $f \in L^2(\Omega^n, \pi^{\otimes n})$ on general product spaces; for simplicity, we'll continue to focus on the case $q = 4$. One possibility is to repeat the above proof using the General Hypercontractivity Theorem (more specifically, Theorem 10.21). This would give us $\|f^{\le k}\|_4 \le \sqrt{3/\lambda}^k \|f\|_4$. However, we will see that it's possible to get a bound completely independent of λ – i.e., independent of (Ω, π) – using randomization/symmetrization.

First, suppose we are in the lucky case described in Example 10.31 in which f's Fourier spectrum only uses symmetric basis functions. In this case $f^{\le k}(\boldsymbol{x})$ and $\widetilde{f^{\le k}}(\boldsymbol{r}, \boldsymbol{x})$ have the same distribution for any k, and we can leverage the $L^2(\{-1,1\})$ bound (10.13) to get the same result for f. First,

$$\|f^{\le k}\|_4 = \|\widetilde{f^{\le k}}\|_4 = \Big\| \, \|\widetilde{f^{\le k}}_{|\boldsymbol{x}}(\boldsymbol{r})\|_{4,\boldsymbol{r}} \, \Big\|_{4,\boldsymbol{x}}.$$

For each outcome $\boldsymbol{x} = x$, the inner function $g(r) = \widetilde{f^{\le k}}_{|x}(r)$ is a degree-k function of $r \in \{-1,1\}^n$. Therefore we can apply (10.13) with this g to deduce

$$\Big\| \, \|\widetilde{f^{\le k}}_{|\boldsymbol{x}}(\boldsymbol{r})\|_{4,\boldsymbol{r}} \, \Big\|_{4,\boldsymbol{x}} \le \Big\| \, \sqrt{3}^k \, \|\widetilde{f}_{|\boldsymbol{x}}(\boldsymbol{r})\|_{4,\boldsymbol{r}} \, \Big\|_{4,\boldsymbol{x}} = \sqrt{3}^k \, \|\widetilde{f}\|_4 = \sqrt{3}^k \, \|f\|_4.$$

Thus we see that we can deduce (10.13) "automatically" for these luckily symmetric f, with no dependence on "λ". We'll now show that we can get something similar for a completely general f using the randomization/symmetrization Theorem 10.35. This will cause us to lose a factor of $(2 \cdot \frac{5}{2})^k$, due to application of T_2 and $\mathrm{T}_{\frac{5}{2}}$; to prepare for this, we first extend the calculation in (10.13) slightly.

Lemma 10.38. *Let $k \in \mathbb{N}$ and let $g : \{-1,1\}^n \to \mathbb{R}$. Then for any $0 < \rho \le 1$,*

$$\|g^{\le k}\|_4 \le (\sqrt{3}/\rho)^k \, \|\mathrm{T}_\rho g\|_4.$$

Proof. We have

$$\|g^{\le k}\|_4 \le \sqrt{3}^k \, \|g^{\le k}\|_2 \le (\sqrt{3}/\rho)^k \, \|\mathrm{T}_\rho g\|_2 \le (\sqrt{3}/\rho)^k \, \|\mathrm{T}_\rho g\|_4.$$

Here the first inequality is Bonami's Lemma and the second is because

$$\|g^{\le k}\|_2^2 = \sum_{j=0}^{k} \mathbf{W}^j[f] \le (1/\rho^2)^k \sum_{j=0}^{k} \rho^{2j} \mathbf{W}^j[f] \le (1/\rho^2)^k \sum_{j=0}^{n} \rho^{2j} \mathbf{W}^j[f] = (1/\rho^2)^k \, \|\mathrm{T}_\rho g\|_2^2.$$

□

We can now give a good answer to Question 10.36, showing that low-degree projection doesn't substantially increase any q-norm:

Theorem 10.39. *Let $k \in \mathbb{N}$ and let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then for $q > 1$ we have $\|f^{\le k}\|_q \le C_q^k \|f\|_q$. Here C_q is a constant depending only on q; in particular we may take $C_4, C_{4/3} = 5\sqrt{3} \le 9$.*

Proof. We will give the proof for $q = 4$; the other cases are left for Exercise 10.16. Using the randomization/symmetrization Theorem 10.35,

$$\|f^{\le k}\|_4 \le \|\widetilde{\mathrm{T}_2 f^{\le k}}\|_4 = \Big\| \, \|\widetilde{\mathrm{T}_2 f^{\le k}}_{|\boldsymbol{x}}(\boldsymbol{r})\|_{4,\boldsymbol{r}} \, \Big\|_{4,\boldsymbol{x}}.$$

For a given outcome $\boldsymbol{x} = x$, let's write $g = \widetilde{\mathrm{T}_2 f}_{|x} : \{-1,1\}^n \to \mathbb{R}$, so that we have $\|g^{\le k}(\boldsymbol{r})\|_4$ on the inside above. For clarity, we remark that g is the Boolean function whose Fourier coefficient on S is $2^{|S|} f^{=S}(x)$. We apply Lemma 10.38 to this g, with $\rho = \frac{1}{5}$. Note that $\mathrm{T}_\rho g$ is then the Boolean function whose Fourier coefficient on S is $(\frac{2}{5})^{|S|} f^{=S}(x)$; i.e., it is $\widetilde{\mathrm{T}_{\frac{2}{5}} f}_{|x}$. Thus we deduce

$$\Big\| \, \|\widetilde{\mathrm{T}_2 f^{\le k}}_{|\boldsymbol{x}}(\boldsymbol{r})\|_{4,\boldsymbol{r}} \, \Big\|_{4,\boldsymbol{x}} \le \Big\| \, (5\sqrt{3})^k \, \|\widetilde{\mathrm{T}_{\frac{1}{5}} f}_{|\boldsymbol{x}}(\boldsymbol{r})\|_{4,\boldsymbol{r}} \, \Big\|_{4,\boldsymbol{x}} = (5\sqrt{3})^k \, \|\widetilde{\mathrm{T}_{\frac{2}{5}} f}\|_4 \le (5\sqrt{3})^k \, \|f\|_4,$$

where the last step is the "un-randomization/symmetrization" inequality from Theorem 10.35. □

The remainder of this section is devoted to the proof of Theorem 10.35, which lets us compare norms of a function and its randomization/symmetrization. It will help to view randomization/symmetrization from an operator perspective. To do this, we need to slightly extend our T_ρ notation, allowing for “different noise rates on different coordinates”.

Definition 10.40. For $i \in [n]$ and $\rho \in \mathbb{R}$, let T^i_ρ be the operator on $L^2(\Omega^n, \pi^{\otimes n})$ defined by

$$\mathrm{T}^i_\rho f = \rho f + (1-\rho)\mathrm{E}_i f = \mathrm{E}_i f + \rho \mathrm{L}_i f = \sum_{S \not\ni i} f^{=S} + \rho \sum_{S \ni i} f^{=S}. \tag{10.14}$$

Furthermore, for $r = (r_1, \dots, r_n) \in \mathbb{R}^n$, let T_r be the operator on $L^2(\Omega^n, \pi^{\otimes n})$ defined by $\mathrm{T}_r = \mathrm{T}^1_{r_1}\mathrm{T}^2_{r_2}\cdots\mathrm{T}^n_{r_n}$. From the third formula in (10.14) we have

$$\mathrm{T}_r f = \sum_{S \subseteq [n]} r^S f^{=S}, \tag{10.15}$$

where we use the notation $r^S = \prod_{i \in S} r_i$. In particular, $\mathrm{T}_{(\rho,\dots,\rho)}$ is the usual T_ρ operator. We remark that when $r \in [0,1]^n$ we have

$$\mathrm{T}_r f(x) = \mathop{\mathbf{E}}_{\boldsymbol{y}_1 \sim N_{r_1}(x_1), \dots, \boldsymbol{y}_n \sim N_{r_n}(x_n)} [f(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)].$$

These generalizations of the noise operator behave the way you would expect; you are referred to Exercise 8.11 for some basic properties. Now comparing (10.15) and (10.10) reveals the connection to randomization/symmetrization:

Fact 10.41. *For $f \in L^2(\Omega^n, \pi^{\otimes n})$, $x \in \Omega^n$, and $r \in \{-1,1\}^n$,*

$$\widetilde{f}(r,x) = \mathrm{T}_r f(x).$$

In other words, randomization/symmetrization of f means applying $\mathrm{T}_{(\pm1,\pm1,\dots,\pm1)}$ to f for a random choice of signs. We use this viewpoint to prove Theorem 10.35, which we do in two steps:

Theorem 10.42. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then for any $q \geq 1$,*

$$\|\mathrm{T}_{\frac12} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \leq \|\mathrm{T}_{\boldsymbol{r}} f(\boldsymbol{x})\|_{q,\boldsymbol{r},\boldsymbol{x}} \tag{10.16}$$

for $\boldsymbol{x} \sim \pi^{\otimes n}$, $\boldsymbol{r} \sim \{-1,1\}^n$. In other words, $\|\mathrm{T}_{\frac12} f\|_q \leq \|\widetilde{f}\|_q$.

Proof. In brief, the result follows from our first randomization/symmetrization result, Lemma 10.15, and an induction. To fill in the details, we begin by showing that if $h \in L^2(\Omega, \pi)$ is any one-input function and $\boldsymbol{\omega} \sim \pi$, $\boldsymbol{b} \sim \{-1,1\}$, then

$$\|\mathrm{T}_{\frac12} h(\boldsymbol{\omega})\|_{q,\boldsymbol{\omega}} \leq \|\mathrm{T}_{\boldsymbol{b}} h(\boldsymbol{\omega})\|_{q,\boldsymbol{b},\boldsymbol{\omega}}. \tag{10.17}$$

This follows immediately from Lemma 10.15 because $h^{=\{1\}}(\boldsymbol{x})$ is a mean-zero random variable (cf. the proof of Corollary 10.20). Next, we show that for any $g \in L^2(\Omega^n, \pi^{\otimes n})$ and any $i \in [n]$,

$$\|\mathrm{T}^i_{\frac{1}{2}} g(\boldsymbol{x})\|_{q,\boldsymbol{x}} \le \|\mathrm{T}^i_{\boldsymbol{r}_i} g(\boldsymbol{x})\|_{q,\boldsymbol{r}_i,\boldsymbol{x}}. \tag{10.18}$$

Assuming $i = 1$ for notational simplicity, and writing $x = (x_1, x')$ where $x' = (x_2, \dots, x_n)$, we have

$$\|\mathrm{T}^i_{\frac{1}{2}} g(\boldsymbol{x})\|_{q,\boldsymbol{x}} = \Big\| \; \|\mathrm{T}^i_{\frac{1}{2}} g(\boldsymbol{x}_1, \boldsymbol{x}')\|_{q,\boldsymbol{x}_1} \; \Big\|_{q,\boldsymbol{x}'} = \Big\| \; \|(\mathrm{T}_{\frac{1}{2}} g_{|\boldsymbol{x}'})(\boldsymbol{x}_1)\|_{q,\boldsymbol{x}_1} \; \Big\|_{q,\boldsymbol{x}'}.$$

(You are asked to carefully justify the second equality here in Exercise 10.10.) Now for each outcome of $\boldsymbol{x}'$ we can apply (10.17) with $h = g_{|\boldsymbol{x}'}$ to deduce

$$\Big\| \; \|(\mathrm{T}_{\frac{1}{2}} g_{|\boldsymbol{x}'})(\boldsymbol{x}_1)\|_{q,\boldsymbol{x}_1} \; \Big\|_{q,\boldsymbol{x}'} \le \Big\| \; \|(\mathrm{T}_{\boldsymbol{r}_1} g_{|\boldsymbol{x}'})(\boldsymbol{x}_1)\|_{q,\boldsymbol{x}_1,\boldsymbol{r}_1} \; \Big\|_{q,\boldsymbol{x}'} = \|\mathrm{T}^i_{\boldsymbol{r}_i} g(\boldsymbol{x})\|_{q,\boldsymbol{r}_i,\boldsymbol{x}}.$$

Finally, we illustrate the first step of the induction. For distinct indices i, j,

$$\|\mathrm{T}^i_{\frac{1}{2}} \mathrm{T}^j_{\frac{1}{2}} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \le \|\mathrm{T}^i_{\boldsymbol{r}_i} \mathrm{T}^j_{\frac{1}{2}} f(\boldsymbol{x})\|_{q,\boldsymbol{r}_i,\boldsymbol{x}}$$

by applying (10.18) with $g = \mathrm{T}^j_{\frac{1}{2}} f$. Then

$$\|\mathrm{T}^i_{\boldsymbol{r}_i} \mathrm{T}^j_{\frac{1}{2}} f(\boldsymbol{x})\|_{q,\boldsymbol{r}_i,\boldsymbol{x}} = \Big\| \; \|\mathrm{T}^i_{\boldsymbol{r}_i} \mathrm{T}^j_{\frac{1}{2}} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \; \Big\|_{q,\boldsymbol{r}_i} = \Big\| \; \|\mathrm{T}^j_{\frac{1}{2}} \mathrm{T}^i_{\boldsymbol{r}_i} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \; \Big\|_{q,\boldsymbol{r}_i},$$

where we used that $\mathrm{T}^i_{\rho_i}$ and $\mathrm{T}^j_{\rho_j}$ commute. Now for each outcome of $\boldsymbol{r}_i$ we can apply (10.18) with $g = \mathrm{T}^i_{\boldsymbol{r}_i} f$ to get

$$\Big\| \; \|\mathrm{T}^j_{\frac{1}{2}} \mathrm{T}^i_{\boldsymbol{r}_i} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \; \Big\|_{q,\boldsymbol{r}_i} \le \Big\| \; \|\mathrm{T}^j_{\boldsymbol{r}_j} \mathrm{T}^i_{\boldsymbol{r}_i} f(\boldsymbol{x})\|_{q,\boldsymbol{r}_j,\boldsymbol{x}} \; \Big\|_{q,\boldsymbol{r}_i} = \|\mathrm{T}^i_{\boldsymbol{r}_i} \mathrm{T}^j_{\boldsymbol{r}_j} f(\boldsymbol{x})\|_{q,\boldsymbol{r}_i,\boldsymbol{r}_j,\boldsymbol{x}}.$$

Thus we have shown

$$\|\mathrm{T}^i_{\frac{1}{2}} \mathrm{T}^j_{\frac{1}{2}} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \le \|\mathrm{T}^i_{\boldsymbol{r}_i} \mathrm{T}^j_{\boldsymbol{r}_j} f(\boldsymbol{x})\|_{q,\boldsymbol{r}_i,\boldsymbol{r}_j,\boldsymbol{x}}.$$

Continuing the induction in the same way completes the proof. □

To prove the "un-randomization/symmetrization" inequality in Theorem 10.35, we first establish an elementary lemma about mean-zero random variables:

Lemma 10.43. *Let $q \ge 2$. Then there is a small enough $0 < c_q \le 1$ such that*

$$\|a - c_q \boldsymbol{X}\|_q \le \|a + \boldsymbol{X}\|_q$$

for any $a \in \mathbb{R}$ and any random variable $\boldsymbol{X}$ satisfying $\mathbf{E}[\boldsymbol{X}] = 0$ and $\|\boldsymbol{X}\|_q < \infty$. In particular we may take $c_4 = \frac{2}{5}$.

Proof. We will only prove the statement for $q = 4$; you are asked to establish the general case in Exercise 10.13. By homogeneity we may assume $a = 1$; then raising the inequality to the 4th power we need to show

$$\mathbf{E}[(1-c\boldsymbol{X})^4] \le \mathbf{E}[(1+\boldsymbol{X})^4]$$

for small enough c. Expanding both sides and using $\mathbf{E}[\boldsymbol{X}] = 0$, this is equivalent to

$$\mathbf{E}[(1-c^4)\boldsymbol{X}^4 + (4+4c^3)\boldsymbol{X}^3 + (6-6c^2)\boldsymbol{X}^2] \ge 0. \tag{10.19}$$

It suffices to find c such that

$$(1-c^4)x^2 + (4+4c^3)x + (6-6c^2) \ge 0 \quad \forall x \in \mathbb{R}; \tag{10.20}$$

then we can multiply this inequality by x^2 and take expectations to obtain (10.19). This last problem is elementary, and Exercise 10.14 asks you to find the largest c that works (the answer is $c \approx .435$). To see that $c = \frac{2}{5}$ suffices, we use the fact that $x \ge -\frac{2}{9}x^2 - \frac{9}{8}$ for all x (because the difference of the left- and right-hand sides is $\frac{1}{72}(4x+9)^2$). Putting this into (10.20), it remains to ensure

$$(\tfrac{1}{9} - \tfrac{8}{9}c^3 - c^4)x^2 + (\tfrac{3}{2} - 6c^2 - \tfrac{9}{2}c^3) \ge 0 \quad \forall x \in \mathbb{R},$$

and when $c = \frac{2}{5}$ this is the trivially true statement $\frac{161}{5625}x^2 + \frac{63}{250} \ge 0$. □

Theorem 10.44. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$. Then for any $q > 1$,*

$$\|\mathrm{T}_{c_q \boldsymbol{r}} f(\boldsymbol{x})\|_{q,\boldsymbol{r},\boldsymbol{x}} \le \|f(\boldsymbol{x})\|_{q,\boldsymbol{x}}$$

for $\boldsymbol{x} \sim \pi^{\otimes n}$, $\boldsymbol{r} \sim \{-1,1\}^n$. In other words, $\|\widetilde{\mathrm{T}_{c_q} f}\|_q \le \|f\|_q$. Here $0 < c_q \le 1$ is a constant depending only on q; in particular we may take $c_4, c_{4/3} = \frac{2}{5}$.

Proof. In fact, we can show that for *every* outcome $\boldsymbol{r} = r \in \{-1,1\}^n$ we have

$$\|\mathrm{T}_{c_q r} f(\boldsymbol{x})\|_{q,\boldsymbol{x}} \le \|f(\boldsymbol{x})\|_{q,\boldsymbol{x}}$$

for sufficiently small $c_q > 0$. Note that on the left-hand side we have

$$\|\mathrm{T}^1_{\pm c_q} \mathrm{T}^2_{\pm c_q} \cdots \mathrm{T}^n_{\pm c_q} f(\boldsymbol{x})\|_{q,\boldsymbol{x}}.$$

We know that T^i_ρ is a contraction in L^q for any $\rho \ge 0$ (Exercise 8.11). Hence it suffices to show that $\mathrm{T}^i_{-c_q}$ is a contraction in L^q, i.e., that

$$\|\mathrm{T}^i_{-c_q} g(\boldsymbol{x})\|_{q,\boldsymbol{x}} \le \|g(\boldsymbol{x})\|_{q,\boldsymbol{x}} \tag{10.21}$$

for all $g \in L^2(\Omega^n, \pi^{\otimes n})$. Similar to the proof of Theorem 10.42, it suffices to show

$$\|\mathrm{T}_{-c_q} h\|_q \le \|h\|_q \tag{10.22}$$

for all one-input functions $h \in L^2(\Omega, \pi)$, because then (10.21) holds pointwise for all outcomes of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \dots, \boldsymbol{x}_n$. By Proposition 9.19, if we prove (10.22) for some q, then the same constant c_q works for the conjugate

Hölder index q'; thus we may restrict attention to $q \geq 2$. Now the result follows from Lemma 10.43 by taking $a = h^{=\emptyset}$ and $\boldsymbol{X} = h^{=\{1\}}(\boldsymbol{x})$. □

10.5. Highlight: General sharp threshold theorems

In Chapter 8.4 we described the problem of "threshold phenomena" for monotone functions $f : \{-1,1\}^n \to \{-1,1\}$. As p increases from 0 to 1, we are interested in whether $\mathbf{Pr}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[f(\boldsymbol{x}) = -1]$ has a "sharp threshold", jumping quickly from near 0 to near 1 around the critical probability $p = p_c$. The "sharp threshold principle" tells us that this occurs (roughly speaking) if and only if the total influence of f under its critical distribution, $\mathbf{I}[f^{(p_c)}]$, is $\omega(1)$. (See Exercise 8.28 for more precise statements.) This motivates finding a characterization of functions with small total influence. Indeed, finding such a characterization is a perfectly natural question even for not-necessarily-monotone Boolean-valued functions $f \in L^2(\Omega^n, \pi^{\otimes n})$.

For the usual uniform distribution on $\{-1,1\}^n$, Friedgut's Junta Theorem from Chapter 9.6 provides a very good characterization: $f : \{-1,1\}^n \to \{-1,1\}$ can only have $O(1)$ total influence if it's (close to) an $O(1)$-junta. By the version of Friedgut's Junta Theorem for general product spaces (Section 10.3), the same holds for Boolean-valued $f \in L^2(\{-1,1\}^n, \pi_p^{\otimes n})$ so long as p is not too close to 0 or to 1. However, for p as small as $1/n^{\Theta(1)}$, the "junta"-size promised by Friedgut's Junta Theorem may be larger than n. (Cf. the breakdown of Friedgut and Kalai's sharp threshold result Theorem 10.29 for $p \leq 1/n^{\Theta(1)}$.) This is a shame, as many natural graph properties for which we'd like to show a sharp threshold – e.g., (non-)3-colorability – have $p = 1/n^{\Theta(1)}$. At a technical level, the reason for the breakdown for very small p is the dependence on the "λ" parameter in the General Hypercontractivity Theorem. But there's a more fundamental reason for its failure, as suggested by the example at the end of Section 10.3: Friedgut's Junta Theorem simply isn't true for such small p.

Example 10.45. Here are some examples of Friedgut's Junta Theorem failing for small p:

- The logical OR function $\mathrm{OR}_n : \{-1,1\}^n \to \{-1,1\}$ has critical probability $p_c \sim \frac{\ln 2}{n}$, and its total influence at this probability is $\mathbf{I}[\mathrm{OR}_n^{(p_c)}] \sim 2\ln 2$, a small constant. Yet it's easy to see that under the p_c-biased distribution, OR_n is not even, say, .1-close to any junta on $o(n)$ coordinates. (That is, for every $o(n)$-junta h, $\mathbf{Pr}_{\boldsymbol{x}\sim\pi_{p_c}^{\otimes n}}[f(\boldsymbol{x}) \neq h(\boldsymbol{x})] > .1$.)
- Consider the function $f : \{-1,1\}^n \to \{-1,1\}$ that is True (-1) if and only if there exists a "run" of three consecutive -1's in its input. (We allow runs to "wrap around", thus making f a transitive-symmetric function.) It's not hard to show that the critical probability for this f satisfies $p_c = \Theta(1/n^{1/3})$. Furthermore, since f is a computable by a DNF of width 3,

> Exercise 8.26(*b*) shows that $\mathbf{I}[f^{(p_c)}] \leq 12$, a small constant. But again, this f is not close to any $o(n)$-junta under the p_c-biased distribution. A similar example is $\mathrm{Clique}_3 : \{\mathsf{True}, \mathsf{False}\}^{\binom{v}{2}} \to \{\mathsf{True}, \mathsf{False}\}$, the graph property of containing a triangle.

We see from these examples that for p very small, we can't hope to show that low-influence functions are close to juntas. However, these counterexample functions still have low complexity in a weaker sense – they are computable by narrow DNFs. Indeed, Friedgut [**Fri99**] suggests this as a characterization:

Friedgut's Conjecture. *There is a function* $w : \mathbb{R}^+ \times (0,1) \to \mathbb{R}^+$ *such that the following holds: If* $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ *is a monotone function,* $0 < p \leq 1/2$*, and* $\mathbf{I}[f^{(p)}] \leq K$*, then* f *is* ϵ*-close under* $\pi_p^{\otimes n}$ *to a monotone DNF of width at most* $w(K, \epsilon)$*.*

The assumption of monotonicity is essential in this conjecture; see Exercise 10.38.

Short of proving his conjecture, Friedgut managed to show:

Friedgut's Sharp Threshold Theorem. *The above conjecture holds when* f *is a graph property.*

This gives a very good characterization of monotone graph properties with low total influence, one that works no matter how small p is. Friedgut also extended his result to monotone hypergraph properties; this was sufficient for him to show that several interesting hypergraph (or hypergraph-like) properties have sharp thresholds – for example, the property of a random 3-uniform hypergraph containing a perfect matching, or the property of a random width-3 DNF formula being a tautology. (Interestingly, for neither of these properties do we know precisely where the critical probability p_c is; nevertheless, we know there is a sharp threshold around it.) Roughly speaking one needs to show that at the critical probability, these properties can't be well-approximated by narrow DNFs because they are almost surely not determined just by "local" information about the (hyper)graph. This kind of deduction takes some effort in random graph theory and we won't discuss it further here beyond Exercise 10.42; for a survey, see Friedgut [**Fri05**].

Friedgut's proof is rather long and it relies heavily on the function being a graph or hypergraph property. Following Friedgut's work, Bourgain [**Bou99**] gave a shorter proof of an alternative characterization. Bourgain's characterization is not as strong as Friedgut's for monotone graph properties; however, it has the advantage that it works for low-influence functions on *any* product probability space. (In particular, there is no monotonicity assumption since the domain need not be $\{\mathsf{True}, \mathsf{False}\}^n$.) We first make a quick definition and then state Bourgain's theorem.

Definition 10.46. Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be $\{-1,1\}$-valued. For $T \subseteq [n]$, $y \in \Omega^T$, and $\tau > 0$, we say that the restriction y_T is a *τ-booster* if $f^{\subseteq T}(y) \ge \mathbf{E}[f] + \tau$. (Recall that $f^{\subseteq T}(y) = \mathbf{E}[f_{\overline{T}|y}]$.) In case $\tau < 0$ we say that y_T is a τ-booster if $f^{\subseteq T}(y) \le \mathbf{E}[f] - |\tau|$.

Bourgain's Sharp Threshold Theorem. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be $\{-1,1\}$-valued with $\mathbf{I}[f] \le K$. Assume $\mathbf{Var}[f] \ge .01$. Then there is some τ (either positive or negative) with $|\tau| \ge \exp(-O(K^2))$ such that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\exists T \subseteq [n], |T| \le O(K) \textit{ such that } \boldsymbol{x}_T \textit{ is a } \tau\textit{-booster}] \ge |\tau|.$$

Thinking of K as an absolute constant, the theorem says that for a typical input string x, there is a large chance that it contains a constant-sized substring that is an $\Omega(1)$-booster for f. In the particular case of monotone $f \in L^2(\{\mathsf{True}, \mathsf{False}\}^n, \pi_p^{\otimes n})$ with p small, it's not hard to deduce (Exercise 10.40) that in fact there exists a T with $|T| \le O(K)$ such that restricting all coordinates in T to be True increases $\mathbf{Pr}_{\pi_p^{\otimes n}}[f = \mathsf{True}]$ by $\exp(-O(K^2))$. This is a qualitatively weaker conclusion than what you get from Friedgut's Sharp Threshold Theorem when f is a graph property with $\mathbf{I}[f] \le O(1)$ – in that case, by taking T to be any of the width-$O(1)$ terms in the approximating DNF one can increase $\mathbf{Pr}_{\pi_p^{\otimes n}}[f = \mathsf{True}]$ not just by $\Omega(1)$ but up to almost 1. Nevertheless, Bourgain's theorem apparently suffices to deduce any of the sharp thresholds results obtainable from Friedgut's theorem [**Fri05**]. For a very high-level sketch of how Bourgain's theorem would apply in the case of 3-colorability of random graphs, see Exercise 10.42.

The last part of this section will be devoted to proving Bourgain's Sharp Threshold Theorem. Before doing this, we add one more remark. Hatami [**Hat12**] has significantly generalized Bourgain's work, establishing the following characterization of Boolean-valued functions with low total influence:

Hatami's Theorem. *Let $f \in L^2(\Omega^n, \pi^{\otimes n})$ be a $\{-1,1\}$-valued function with $\mathbf{I}[f] \le K$. Then for every $\epsilon > 0$, the function f is ϵ-close (under $\pi^{\otimes n}$) to an $\exp(O(K^3/\epsilon^3))$-"pseudo-junta" $h : \Omega^n \to \{-1,1\}$.*

The term "pseudo-junta" is defined in Exercise 10.39. A K-pseudo-junta h has the property that $\mathbf{I}[h] \le 4K$; thus Hatami's Theorem shows that having $O(1)$ total influence is essentially equivalent to being an $O(1)$-pseudo-junta. A downside of the result, however, is that being a K-pseudo-junta is not a "syntactic" property; it depends on the probability distribution $\pi^{\otimes n}$.

Let's now turn to proving Bourgain's Sharp Threshold Theorem. In fact, Bourgain proved the theorem as a corollary of the following main result:

Theorem 10.47. *Let (Ω, π) be a finite probability space and let $f : \Omega^n \to \{-1,1\}$. Let $0 < \epsilon < 1/2$ and write $k = \mathbf{I}[f]/\epsilon$. Then for each $x \in \Omega^n$ it's possible to define a set of "notable coordinates" $J_x \subseteq [n]$ satisfying $|J_x| \leq \exp(O(k))$ such that*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \pi^{\otimes n}} \left[\sum_{S \not\in \mathscr{F}_{\boldsymbol{x}}} f^{=S}(\boldsymbol{x})^2 \right] \leq 2\epsilon.$$

Here $\mathscr{F}_x = \{S : S \subseteq J_x, |S| \leq k\}$, a collection always satisfying $|\mathscr{F}_x| \leq \exp(O(k^2))$.

You may notice that this theorem looks extremely similar to Friedgut's Junta Theorem from Chapter 9.6 (and the $\exp(-O(\mathbf{I}[f]^2))$ quantity in Bourgain's Sharp Threshold Theorem looks similar to the Fourier coefficient lower bound in Corollary 9.32). Indeed, the only difference between Theorem 10.47 and Friedgut's Junta Theorem is that in the latter, the "notable coordinates" J can be "named in advance" – they're simply the coordinates j with $\mathbf{Inf}_j[f] = \sum_{S \ni j} \widehat{f}(S)^2$ large. By contrast, in Theorem 10.47 the notable coordinates depend on the input x. As we will see in the proof, they are precisely the coordinates j such that $\sum_{S \ni j} f^{=S}(x)^2$ is large. Of course, in the setting of $f : \{-1,1\}^n \to \{-1,1\}$ we have $f^{=S}(x)^2 = \widehat{f}(S)^2$ for all x, so the two definitions coincide. But in the general setting of $f \in L^2(\Omega^n, \pi^{\otimes n})$ it makes sense that we can't name the notable coordinates in advance and rather have to "wait until x is chosen". For example, for the OR_n function as in Example 10.45, there are no notable coordinates to be named in advance, but once x is chosen the few coordinates on which x takes the value True (if any exist) will be the notable ones.

The proof of Theorem 10.47 mainly consists of adding the randomization/symmetrization technique to the proof of Friedgut's Junta Theorem (more precisely, Theorem 9.28) to avoid dependence on the minimum probability of π. This randomization/symmetrization is applied to what are essentially the key inequalities in that proof:

$$\|\mathrm{T}_{\frac{1}{\sqrt{3}}} \mathrm{L}_i f\|_2^2 \leq \|\mathrm{L}_i f\|_{4/3}^2 = \|\mathrm{L}_i f\|_{4/3}^{2/3} \cdot \|\mathrm{L}_i f\|_{4/3}^{4/3} \leq \|\mathrm{L}_i f\|_{4/3}^{2/3} \cdot \mathbf{Inf}_i[f].$$

(The last inequality here is Exercise 8.10(*b*).) The overall proof needs one more minor twist: since we work on a "per-x" basis and not in expectation, it's possible that the set of notable coordinates can be improbably large. (Think again about the example of OR_n; for $\boldsymbol{x} \sim \pi_{1/n}^{\otimes n}$ we expect only a constant number of coordinates of $\boldsymbol{x}$ to be True, but it's not always uniformly bounded.) This is combated using the principle that low-degree functions are "reasonable" (together with randomization/symmetrization).

Proof of Theorem 10.47. By the simple "Markov argument" (see Proposition 3.2) we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\sum_{|S|>k} f^{=S}(\boldsymbol{x})^2\right] = \sum_{|S|>k} \|f^{=S}\|_2^2 \leq \mathbf{I}[f]/k = \epsilon.$$

Thus it suffices to define the sets J_x so that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\sum_{|S|\leq k,\ S\not\subseteq J_{\boldsymbol{x}}} f^{=S}(\boldsymbol{x})^2\right] \leq \epsilon. \tag{10.23}$$

We'll first define "notable coordinate" sets $J'_x \subseteq [n]$ which almost do the trick:

$$J'_x = \left\{ j \in [n] : \sum_{S\ni j} f^{=S}(x)^2 \geq \tau \right\}, \quad \tau = c^{-k}.$$

(where $c > 1$ is a universal constant). Using this definition, the main effort of the proof will be to show

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\sum_{|S|\leq k,\ S\not\subseteq J'_{\boldsymbol{x}}} f^{=S}(\boldsymbol{x})^2\right] \leq \epsilon/2. \tag{10.24}$$

This looks better than (10.23); the only problem is that the sets J'_x don't always satisfy $|J'_x| \leq \exp(O(k))$ as needed. However, "in expectation" $|J'_x|$ ought not be much larger than $1/\tau = c^k$. Thus we introduce the event

$$\text{“}J'_{\boldsymbol{x}} \text{ is too big”} \quad \Longleftrightarrow \quad |J'_{\boldsymbol{x}}| \geq C^k$$

(where $C > c$ is another universal constant) and define

$$J_x = \begin{cases} J'_x & \text{if } J'_x \text{ is not too big,} \\ \emptyset & \text{if } J'_x \text{ is too big.} \end{cases}$$

The last part of the proof will be to show that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\mathbf{1}[J'_{\boldsymbol{x}} \text{ is too big}] \cdot \sum_{0<|S|\leq k} f^{=S}(\boldsymbol{x})^2\right] \leq \epsilon/2. \tag{10.25}$$

Together, (10.25) and (10.24) establish (10.23). We will first prove (10.24) and then prove (10.25). As a small aside, we'll see that for both inequalities we could obtain a bound much less than $\epsilon/2$ if desired.

To prove (10.24), we mimic the proof of Theorem 9.28 but add in randomization/symmetrization. The key step is encapsulated in the following lemma. Note that the lemma also holds with the more natural definition $g = \mathrm{L}_i f$; the additional $\mathrm{T}_{\frac{2}{5}}$ is to facilitate future "un-randomization/symmetrization".

Lemma 10.48. *Fix $x \in \Omega^n$ and $i \in J'_x$. Then writing $g = \mathrm{T}_{\frac{2}{5}} \mathrm{L}_i f$ we have*

$$\|\mathrm{T}_{\frac{1}{\sqrt{3}}} \widetilde{g}_{|x}\|_2^2 \leq \tau^{1/3} \|\widetilde{g}_{|x}\|_{4/3}^{4/3}.$$

Proof. Here $\widetilde{g}$ is the randomization/symmetrization of g, so $\widetilde{g}_{|x} = \widetilde{g}_{|x}(r)$ is a function on the uniform-distribution hypercube. Applying the basic $(4/3,2)$-Hypercontractivity Theorem we have

$$\|\mathrm{T}_{\frac{1}{\sqrt{3}}}\widetilde{g}_{|x}\|_2^2 \le \|\widetilde{g}_{|x}\|_{4/3}^2 = (\|\widetilde{g}_{|x}\|_{4/3}^2)^{1/3}\cdot\|\widetilde{g}_{|x}\|_{4/3}^{4/3} \le (\|\widetilde{g}_{|x}\|_2^2)^{1/3}\cdot\|\widetilde{g}_{|x}\|_{4/3}^{4/3}.$$

But by the usual Parseval Theorem,

$$\|\widetilde{g}_{|x}\|_2^2 = \sum_{S\subseteq[n]} g^{=S}(x)^2 = \sum_{S\ni i}(2/5)^{2|S|}f^{=S}(x)^2 \le \sum_{S\ni i} f^{=S}(x)^2 \le \tau,$$

the last inequality due to the assumption that $i \in J'_x$. □

We now establish (10.24):

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{|S|\le k,\ S\not\subseteq J'_{\boldsymbol{x}}} f^{=S}(\boldsymbol{x})^2\right] &\le (5\sqrt{3}/2)^{2k}\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{S\not\subseteq J'_{\boldsymbol{x}}}(\mathrm{T}_{\frac{2}{5\sqrt{3}}}f^{=S})(\boldsymbol{x})^2\right]\\
&\le 20^k\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{i\not\in J'_{\boldsymbol{x}}}\sum_{S\ni i}(\mathrm{T}_{\frac{2}{5\sqrt{3}}}f^{=S})(\boldsymbol{x})^2\right]\\
&= 20^k\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{i\not\in J'_{\boldsymbol{x}}}\|\mathrm{T}_{\frac{1}{\sqrt{3}}}\widetilde{g^i}_{|\boldsymbol{x}}\|_2^2\right] && (\text{for } g^i = \mathrm{T}_{\frac{2}{5}}\mathrm{L}_i f)\\
&\le 20^k\tau^{1/3}\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\sum_{i\not\in J'_{\boldsymbol{x}}}\|\widetilde{g^i}_{|\boldsymbol{x}}\|_{4/3}^{4/3}\right] && (\text{Lemma 10.48})\\
&\le 20^k\tau^{1/3}\cdot\sum_{i=1}^n\|\mathrm{L}_i f\|_{4/3}^{4/3} && (\text{Theorem 10.35})\\
&\le 20^k\tau^{1/3}\cdot\sum_{i=1}^n\mathbf{Inf}_i[f] && (\text{Exercise 8.10}(b))\\
&= 20^k\tau^{1/3}\cdot\mathbf{I}[f] = (20c^{-1/3})^k k\epsilon \le \epsilon/2,
\end{aligned}$$

the last inequality because $(20c^{-1/3})^k k \le 1/2$ for all $k \ge 0$ once c is a large enough constant.

The last task in the proof is to establish (10.25). Using Cauchy–Schwarz,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\mathbf{1}[J'_{\boldsymbol{x}}\text{ is too big}]\cdot\sum_{0<|S|\le k} f^{=S}(\boldsymbol{x})^2\right]\\
&\qquad\le\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\mathbf{1}[J'_{\boldsymbol{x}}\text{ is too big}]^2\right]}\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\Big(\sum_{0<|S|\le k} f^{=S}(\boldsymbol{x})^2\Big)^2\right]}. \qquad (10.26)
\end{aligned}$$

For the first factor on the right of (10.26) we use Markov's inequality:

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\mathbf{1}[J'_{\boldsymbol{x}} \text{ is too big}]^2\right] &= \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[J'_{\boldsymbol{x}} \text{ is too big}] = \mathop{\mathbf{Pr}}_{\boldsymbol{x}}[|J'_{\boldsymbol{x}}| \geq C^k] \\ &\leq C^{-k}\mathop{\mathbf{E}}_{\boldsymbol{x}}[|J'_{\boldsymbol{x}}|] \leq C^{-k}\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\Big(\sum_{i=1}^{n}\sum_{S\ni i} f^{=S}(\boldsymbol{x})^2\Big)/\tau\right] = C^{-k}c^k\cdot\mathbf{I}[f]. \qquad (10.27)\end{aligned}$$

As for the second factor on the right of (10.26), let's write $h = \mathrm{T}_{\frac{2}{5}}(f - f^{=\emptyset})$. (We are being slightly finicky about $f^{=\emptyset}$ just in case it's very large.) Then

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\Big(\sum_{0<|S|\leq k} f^{=S}(\boldsymbol{x})^2\Big)^2\right] &\leq (5/2)^{4k}\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\Big(\sum_{S\neq\emptyset}(\mathrm{T}_{\frac{2}{5}}f^{=S})(\boldsymbol{x})^2\Big)^2\right] \\ &= (5/2)^{4k}\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\|\widetilde{h}_{|\boldsymbol{x}}\|_2^4\right] \\ &\leq 40^k\cdot\mathop{\mathbf{E}}_{\boldsymbol{x}}\left[\|\widetilde{h}_{|\boldsymbol{x}}\|_4^4\right] \\ &\leq 40^k\cdot\|f - f^{=\emptyset}\|_4^4 && \text{(Theorem 10.35)} \\ &\leq 40^k\cdot 2^2\mathop{\mathbf{E}}_{\boldsymbol{x}}[(f-f^{=\emptyset})^2] && \text{(since } |f - f^{=\emptyset}| \leq 2 \text{ always)} \\ &= 4\cdot 40^k\cdot\mathbf{Var}[f] \leq 4\cdot 40^k\cdot\mathbf{I}[f]. && (10.28)\end{aligned}$$

Substituting (10.27) and (10.28) into (10.26) gives

$$\begin{aligned}&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\mathbf{1}[J'_{\boldsymbol{x}} \text{ is too big}]\cdot\sum_{0<|S|\leq k} f^{=S}(\boldsymbol{x})^2\right] \\ &\qquad\qquad \leq \sqrt{C^{-k}c^k\cdot 4\cdot 40^k}\cdot\mathbf{I}[f] = 2(\tfrac{40c}{C})^{k/2}k\epsilon \leq \epsilon/2,\end{aligned}$$

the last inequality again holding for all $k \geq 0$ once C is chosen large enough compared to c. □

We end this chapter by deducing Bourgain's Sharp Threshold Theorem from Theorem 10.47.

Proof of Bourgain's Sharp Threshold Theorem. We take $\epsilon = .001$ in Theorem 10.47 and obtain the associated collections of subsets $\mathscr{F}_x$, where each $|\mathscr{F}_x| \leq \exp(O(K^2))$ and each $S \in \mathscr{F}_x$ satisfies $|S| \leq O(K)$. Using the fact that $f^{=\emptyset}(x)^2 = 1 - \mathbf{Var}[f] \leq .99$ for each x we get

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\sum_{S\in\mathscr{F}_{\boldsymbol{x}}\setminus\{\emptyset\}} f^{=S}(\boldsymbol{x})^2\right] \geq 1 - 2\epsilon - .99 = .008.$$

We always have $|\mathscr{F}_x \setminus \{\emptyset\}| \leq \exp(O(K^2))$, and there's also no harm in assuming $|\mathscr{F}_x \setminus \{\emptyset\}| > 0$. It follows that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[\max_{S\in\mathscr{F}_{\boldsymbol{x}}\setminus\{\emptyset\}}\{f^{=S}(\boldsymbol{x})^2\}\right] \geq \frac{.008}{\exp(O(K^2))} = \exp(-O(K^2)).$$

Thus for each x we can define a set S_x with $0 < |S_x| \le O(K)$ such that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[f^{=S_{\boldsymbol{x}}}(\boldsymbol{x})^2\right] \ge \exp(-O(K^2)). \tag{10.29}$$

By Exercise 8.19 we have $|f^{=S_x}(x)| \le 2^{|S_x|} \le 2^{O(K)}$ and hence $f^{=S_{\boldsymbol{x}}}(\boldsymbol{x})^2 \le \exp(O(K))$ always. It follows from (10.29) that we must have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi^{\otimes n}}\left[f^{=S_{\boldsymbol{x}}}(\boldsymbol{x})^2 \ge \exp(-O(K^2))\right] \ge \exp(-O(K^2)).$$

We will complete the proof by showing that whenever $f^{=S_{\boldsymbol{x}}}(\boldsymbol{x})^2 \ge \exp(-O(K^2))$ occurs, there exists $T \subseteq S_{\boldsymbol{x}}$ such that $\boldsymbol{x}_T$ is a $\pm\exp(-O(K^2))$-booster for f. Thus we either have a $+\exp(-O(K^2))$-booster with probability at least $\frac{1}{2}\exp(-O(K^2))$, or a $-\exp(-O(K^2))$ with probability at least $\frac{1}{2}\exp(-O(K^2))$; either way, the proof will be complete.

Assume then that $f^{=S_x}(x)^2 \ge \exp(-O(K^2))$; equivalently,

$$|f^{=S_x}(x)| \ge \exp(-O(K^2)).$$

Let's now work with $g = f - \mathbf{E}[f]$. Of course $g^{=T} = f^{=T}$ for all $T \ne \emptyset$; since $S_x \ne \emptyset$ the above inequality tells us that $|g^{=S_x}(x)| \ge \exp(-O(K^2))$. Recall the formula

$$g^{=S_x}(x) = \sum_{\emptyset \ne T \subseteq S_x} (-1)^{|S_x|-|T|} g^{\subseteq T}(x);$$

we dropped the $T = \emptyset$ term since it's 0. As there are only $2^{|S_x|} - 1 = \exp(O(K))$ terms in the above sum, we deduce there must exist some $T \subseteq S_x$ with $0 < |T| \le O(K)$ such that

$$|g^{\subseteq T}(x)| \ge \exp(-O(K^2))/\exp(O(K)) = \exp(-O(K^2)).$$

But $g^{\subseteq T} = f^{\subseteq T} - \mathbf{E}[f]$, so the above gives us $|f^{\subseteq T}(x) - \mathbf{E}[f]| \ge \exp(-O(K^2))$. This precisely says that x_T is a $\pm\exp(-O(K^2))$-booster, as desired. □

For a relaxation of the assumption $\mathbf{Var}[f] \ge .01$ in this theorem, see Exercise 10.41.

10.6. Exercises and notes

10.1 Let $\boldsymbol{X}$ be a random variable and let $1 \le r \le \infty$. Recall that the triangle (Minkowski) inequality implies that for real-valued functions f_1, f_2,

$$\|f_1(\boldsymbol{X}) + f_2(\boldsymbol{X})\|_r \le \|f_1(\boldsymbol{X})\|_r + \|f_2(\boldsymbol{X})\|_r.$$

More generally, if $w_1, \dots, w_m$ are nonnegative reals $f_1, \dots, f_m$ are real functions, then

$$\|w_1 f_1(\boldsymbol{X}) + \cdots + w_m f_m(\boldsymbol{X})\|_r \le w_1\|f_1(\boldsymbol{X})\|_r + \cdots + w_m\|f_m(\boldsymbol{X})\|_r.$$

Still more generally, if $\boldsymbol{Y}$ is a random variable independent of $\boldsymbol{X}$ and $f(\boldsymbol{X},\boldsymbol{Y})$ is a (measurable) real-valued function, then it holds that

$$\Big\|\mathop{\mathbf{E}}_{\boldsymbol{Y}}[f(\boldsymbol{X},\boldsymbol{Y})]\Big\|_{r,\boldsymbol{X}} \leq \mathop{\mathbf{E}}_{\boldsymbol{Y}}[\|f(\boldsymbol{X},\boldsymbol{Y})\|_{r,\boldsymbol{X}}].$$

Using this last fact, show that whenever $0 < p \leq q \leq \infty$,

$$\Big\| \ \|f(\boldsymbol{X},\boldsymbol{Y})\|_{p,\boldsymbol{Y}} \ \Big\|_{q,\boldsymbol{X}} \leq \Big\| \ \|f(\boldsymbol{X},\boldsymbol{Y})\|_{q,\boldsymbol{X}} \ \Big\|_{p,\boldsymbol{Y}}.$$

(Hint: Raise the inequality to the power of p and use $r = q/p$.)

10.2 The goal of this exercise is to prove Proposition 9.15: If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent (p,q,ρ)-hypercontractive random variables, then so is $\boldsymbol{X}+\boldsymbol{Y}$. Let $a,b \in \mathbb{R}$.

(*a*) First obtain

$$\|a+\rho b(\boldsymbol{X}+\boldsymbol{Y})\|_{q,\boldsymbol{X},\boldsymbol{Y}} \leq \Big\| \ \|a+\rho b\boldsymbol{X}+b\boldsymbol{Y}\|_{p,\boldsymbol{Y}} \ \Big\|_{q,\boldsymbol{X}}.$$

(*b*) Next, upper-bound this by

$$\Big\| \ \|a+b\boldsymbol{Y}+\rho b\boldsymbol{X}\|_{q,\boldsymbol{X}} \ \Big\|_{p,\boldsymbol{Y}}.$$

(Hint: Exercise 10.1.)

(*c*) Finally, upper-bound this by

$$\Big\| \ \|a+b\boldsymbol{Y}+b\boldsymbol{X}\|_{p,\boldsymbol{X}} \ \Big\|_{p,\boldsymbol{Y}} = \|a+b(\boldsymbol{X}+\boldsymbol{Y})\|_{p,\boldsymbol{X},\boldsymbol{Y}}.$$

10.3 Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be independent (p,q,ρ)-hypercontractive random variables. Let $F(x) = \sum_{S\subseteq[n]} \widehat{F}(S)x^S$ be an n-variate multilinear polynomial. Define formally the multilinear polynomial $\mathrm{T}_\rho F(x) = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{F}(S)x^S$. The goal of this exercise is to show

$$\|\mathrm{T}_\rho F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)\|_q \leq \|F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)\|_p. \tag{10.30}$$

Note that this result yields an alternative deduction of the Hypercontractivity Theorem for ±1 bits from the Two-Point Inequality. A (notationally intense) generalization of this exercise can also be used as an alternative inductive strategy for deducing the General Hypercontractivity Theorem from Proposition 10.17 or Theorem 10.18.

(*a*) Why is Exercise 10.2 a special case of (10.30)?

(*b*) Begin the inductive proof of (10.30) by showing that the base case $n = 0$ is trivial.

(*c*) For the case of general n, first establish

$$\|\mathrm{T}_\rho F(\boldsymbol{X})\|_q \leq \Big\| \ \|\mathrm{T}'_\rho E(\boldsymbol{X}') + \boldsymbol{X}_n \mathrm{T}'_\rho D(\boldsymbol{X}')\|_{p,\boldsymbol{X}_n} \ \Big\|_{q,\boldsymbol{X}'},$$

where we are using the notation $x' = (x_1,\dots,x_{n-1})$, $F(x) = E(x') + x_n D(x')$, and T'_ρ for the operator acting formally on $(n-1)$-variate multilinear polynomials.

(*d*) Complete the inductive step, using steps similar to Exercises 10.2(*b*),(*c*). (Hint: For X_n a real constant, why is $\mathrm{T}'_\rho E(\boldsymbol{X}') + X_n \mathrm{T}'_\rho D(\boldsymbol{X}') = \mathrm{T}'_\rho(E + X_n D)(\boldsymbol{X}')$?)

10.4 This exercise is concerned with the possibility of a converse for Proposition 10.8.

(*a*) In our proof of the Two-Point Inequality we used Proposition 9.19 to deduce that a uniform bit $\boldsymbol{x} \sim \{-1,1\}$ is (p,q,ρ)-hypercontractivity if it's (q',p',ρ)-hypercontractive. Why can't we use Proposition 9.19 to deduce this for a general random variable $\boldsymbol{X}$?

(*b*) For each $1 < p < 2$, exhibit a random variable $\boldsymbol{X}$ that is $(p,2,\rho)$-hypercontractive (for some ρ) but not $(2,p',\rho)$-hypercontractive.

10.5 (*a*) Regarding Remark 10.2, heuristically justify (in the manner of Exercise 9.24(*a*)) the following statement: If $A,B \subseteq \{-1,1\}^n$ are concentric Hamming balls with volumes $\exp(-\frac{a^2}{2})$ and $\exp(-\frac{b^2}{2})$ and $\rho a \le b \le a$ (where $0 < \rho < 1$), then

$$\mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] \gtrapprox \exp\left(-\frac{1}{2}\frac{a^2-2\rho ab+b^2}{1-\rho^2}\right);$$

and further, if $b < \rho a$, then $\mathbf{Pr}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] \sim \mathbf{Pr}[\boldsymbol{x} \in A]$. Here you should treat ρ as fixed and $a,b \to \infty$.

(*b*) Similarly, heuristically justify that the Reverse Small-Set Expansion Theorem is essentially sharp by considering diametrically opposed Hamming balls.

10.6 The goal of this exercise (and Exercise 10.7) is to prove the Reverse Hypercontractivity Theorem and its equivalent Two-Function version:

Reverse Hypercontractivity Theorem. *Let $f : \{-1,1\}^n \to \mathbb{R}^{\ge 0}$ be a nonnegative function and let $-\infty \le q < p \le 1$. Then $\|\mathrm{T}_\rho f\|_q \ge \|f\|_p$ for $0 \le \rho \le \sqrt{(1-p)/(1-q)}$.*

Reverse Two-Function Hypercontractivity Theorem. *Let $f,g : \{-1,1\}^n \to \mathbb{R}^{\ge 0}$ be nonnegative, let $r,s \le 0$, and assume $0 \le \rho \le \sqrt{rs} \le 1$. Then*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{y})\\ \rho\text{-correlated}}}[f(\boldsymbol{x})g(\boldsymbol{y})] \ge \|f\|_{1+r}\|g\|_{1+s}.$$

Recall that for $-\infty < p < 0$ and for positive functions $f \in L^2(\Omega,\pi)$ the "norm" $\|f\|_p$ retains the definition $\mathbf{E}[f^p]^{-1/p}$. (The cases of $p = -\infty$, $p = 0$, and nonnegative functions are defined by appropriate limits; in particular $\|f\|_{-\infty}$ is the minimum of f's values, $\|f\|_0$ is the geometric mean of f's values, and $\|f\|_p$ is 0 whenever f is not everywhere positive. We also define p' by $\frac{1}{p} + \frac{1}{p'} = 1$, with $0' = 0$.)

The Reverse Two-Function Hypercontractivity Theorem can be thought of as a generalization of the lesser known "reverse Hölder inequality" in the setting of $L^2(\{-1,1\}^n, \pi_{1/2}^{\otimes n})$:

Reverse Hölder inequality. *Let $f \in L^2(\Omega, \pi)$ be a positive function. Then for any $p < 1$,*

$$\|f\|_p = \inf\{\mathbf{E}[fg] : g > 0, \|g\|_{p'} = 1\}.$$

In particular, for $r < 0$ and $f, g > 0$ we have $\mathbf{E}[fg] \geq \|f\|_{1+r}\|g\|_{1+1/r}$.

(a) Show that to prove these two Reverse Hypercontractivity Theorems it suffices to consider the case of $f, g : \{-1,1\}^n \to \mathbb{R}^+$, i.e., strictly positive functions.

(b) Show that the Reverse Two-Function Hypercontractivity Theorem is equivalent (via the reverse Hölder inequality) to the Reverse Hypercontractivity Theorem.

(c) Reduce the Reverse Two-Function Hypercontractivity Theorem to the $n = 1$ case. (Hint: Virtually identical to the Two-Function Hypercontractivity Induction.) Further reduce to following:

Reverse Two-Point Inequality. *Let $-\infty \leq q < p \leq 1$ and let $0 \leq \rho \leq \sqrt{(1-p)/(1-q)}$. Then $\|\mathrm{T}_\rho f\|_q \geq \|f\|_p$ for any $f : \{-1,1\} \to \mathbb{R}^+$.*

10.7 The goal of this exercise is to prove the Reverse Two-Point Inequality.

(a) Similar to the non-reverse case, the main effort is proving the inequality assuming that $0 < q < p \leq 1$ and that $\rho = \sqrt{(1-p)/(1-q)}$. Do this by mimicking the proof of the Two-Point Inequality. (Hint: You will need the inequality $(1+t)^\theta \geq 1 + \theta t$ for $\theta \geq 1$, and you will need to show that $\frac{j-r}{\sqrt{1-r}}$ is an increasing function of r on $[0,1)$ for all $j \geq 2$.)

(b) Extend to the case of $0 \leq \rho \leq \sqrt{(1-p)/(1-q)}$. (Hint: Use the fact that for any $f : \{-1,1\}^n \to \mathbb{R}^{\geq 0}$ and $-\infty \leq p \leq q \leq \infty$ we have $\|f\|_p \leq \|f\|_q$. You can prove this generalization of Exercise 1.13 by reducing to the case of negative p and q to the case of positive p and q.)

(c) Establish the $q = -\infty$ case of the Reverse Two-Point Inequality.

(d) Show that the cases $-\infty < q < p < 0$ follow by "duality". (Hint: Like Proposition 9.19 but with the reverse Hölder inequality.)

(e) Show that the cases $q < 0 < p$ follow by the semigroup property of T_ρ.

(f) Finally, treat the cases of $p = 0$ or $q = 0$.

10.8 Give a simple proof of the $n = 1$ case of the Reverse Two-Function Hypercontractivity Theorem when $r = s = -1/2$. (Hint: Replace f and g by f^2 and g^2; then you don't even need to assume f and g are nonnegative.) Can you also give a simple proof when $r = s = -1 + 1/k$ for integers $k > 2$?

10.9 By selecting "r" $= -\rho\frac{\rho a+b}{a+\rho b}$ and "s" $= -\rho\frac{a+\rho b}{\rho a+b}$, prove the Reverse Small-Set Expansion Theorem mentioned in Remark 10.3. (Hint: The negative norm of a 0-1-indicator is 0, so be sure to verify no negative norms arise.)

10.10 Let $g \in L^2(\Omega^n, \pi^{\otimes n})$. Writing $x = (x_1, x')$, where $x' = (x_2, \dots, x_n)$, carefully justify the following identity of one-input functions: $(\mathrm{T}^1_\rho g)_{|x'} = \mathrm{T}_\rho(g_{|x'})$. (Hint: You may want to refer to Exercise 8.21.)

10.11 Prove Proposition 10.12.

10.12 Let $\boldsymbol{X}$ be a random variable and let $\boldsymbol{Y}$ denote its symmetrization $\boldsymbol{X} - \boldsymbol{X}'$, where $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$. Show for any $t, \theta \in \mathbb{R}$ that $\mathbf{Pr}[|\boldsymbol{Y}| \geq t] \leq 2\,\mathbf{Pr}[|\boldsymbol{X} - \theta| \geq t/2]$.

10.13 The goal of this exercise is to establish Lemma 10.43.

(*a*) Show that we may take $c_2 = 1$ (and that equality holds). Henceforth assume $q > 2$.

(*b*) By following the idea of our $q = 4$ proof, reduce to showing that there exists $0 < c_q < 1$ such that

$$|1 - c_q x|^q + c_q q x - 1 \leq |1 + x|^q - qx - 1 \quad \forall x \in \mathbb{R}.$$

(*c*) Further reduce to showing there exists $0 < c_q < 1$ such that

$$\frac{|1 - c_q x|^q + c_q q x - 1}{x^2} \leq \frac{|1 + x|^q - qx - 1}{x^2} \quad \forall x \in \mathbb{R}. \tag{10.31}$$

Here you should also establish that both sides are continuous functions of $x \in \mathbb{R}$ once the value at $x = 0$ is defined appropriately.

(*d*) Show that there exists $M > 0$ such that for *every* $0 < c_q < \frac{1}{2}$, inequality (10.31) holds once $|x| \geq M$. (Hint: Consider the limit of both sides as $|x| \to \infty$.)

(*e*) Argue that it suffices to show that

$$\frac{|1 + x|^q - qx - 1}{x^2} \geq \eta \tag{10.32}$$

for some universal positive constant $\eta > 0$. (Hint: A uniform continuity argument for $(x, c_q) \in [-M, M] \times [0, \frac{1}{2}]$.)

(*f*) Establish (10.32). (Hint: The best possible η is 1, but to just achieve some positive η, argue using Bernoulli's inequality that $\frac{|1+x|^q - qx - 1}{x^2}$ is everywhere positive and then observe that it tends to ∞ as $|x| \to \infty$.)

(*g*) Possibly using a different argument, what is the best asymptotic bound you can achieve for c_q? Is $c_q \geq \Omega(\frac{\log q}{q})$ possible?

10.14 Show that the largest c for which inequality (10.20) holds is the smaller real root of $c^4 - 2c^3 - 2c + 1 = 0$, namely, $c \approx .435$.

10.15 (*a*) Show that $1 + 6c^2x^2 + c^4x^4 \leq 1 + 6x^2 + 4x^3 + x^4$ holds for all $x \in \mathbb{R}$ when $c = 1/2$. (Can you also establish it for $c \approx .5269$?)

(*b*) Show that if $\boldsymbol{X}$ is a random variable satisfying $\mathbf{E}[\boldsymbol{X}] = 0$ and $\|\boldsymbol{X}\|_4 < \infty$, then $\|a + \frac{1}{2}\boldsymbol{r}\boldsymbol{X}\|_4 \leq \|a + \boldsymbol{X}\|_4$ for all $a \in \mathbb{R}$, where $\boldsymbol{r} \sim \{-1,1\}$ is a uniformly random bit independent of $\boldsymbol{X}$. (Cf. Lemma 10.15.)

(*c*) Establish the following improvement of Theorem 10.44 in the case of $q = 4$: for all $f \in L^2(\Omega^n, \pi^{\otimes n})$,

$$\|\mathrm{T}_{\frac{1}{2}\boldsymbol{r}} f(\boldsymbol{x})\|_{4,\boldsymbol{r},\boldsymbol{x}} \leq \|f(\boldsymbol{x})\|_{4,\boldsymbol{x}}$$

(where $\boldsymbol{x} \sim \pi^{\otimes n}$, $\boldsymbol{r} \sim \{-1,1\}^n$).

10.16 Complete the proof of Theorem 10.39. (Hint: You'll need to rework Exercise 9.8 as in Lemma 10.38.)

10.17 Prove Proposition 10.17.

10.18 Recall from (10.5) the function $\rho = \rho(\lambda)$ defined for $\lambda \in (0,1/2)$ (and fixed $q > 2$) by

$$\rho = \rho(\lambda) = \sqrt{\frac{\exp(u/q) - \exp(-u/q)}{\exp(u/q') - \exp(-u/q')}} = \sqrt{\frac{\sinh(u/q)}{\sinh(u/q')}},$$

where $u = u(\lambda)$ is defined by $\exp(-u) = \frac{\lambda}{1-\lambda}$.

(*a*) Show that ρ is an increasing function of λ. (Hint: One route is to reduce to showing that ρ^2 is a decreasing function of $u \in (0,\infty)$, reduce to showing that $q\tanh(u/q)$ is an increasing function of $q \in (1,\infty)$, reduce to showing $\frac{\tanh r}{r}$ is a decreasing function of $r \in (0,\infty)$, and reduce to showing $\sinh(2r) \geq 2r$.)

(*b*) Verify the following statements from Remark 10.19:

$$\text{for fixed } q \text{ and } \lambda \to 1/2, \quad \rho \to \frac{1}{\sqrt{q-1}};$$
$$\text{for fixed } q \text{ and } \lambda \to 0, \quad \rho \sim \lambda^{1/2-1/q}.$$

Also show:

$$\text{for fixed } \lambda \text{ and } q \to \infty, \quad \rho \sim \sqrt{\frac{u}{\sinh u}}\sqrt{\frac{1}{q}},$$

and $\sqrt{\frac{u}{\sinh u}} \sim 2\lambda \ln(1/\lambda)$ for $\lambda \to 0$.

(*c*) Show that $\rho \geq \frac{1}{\sqrt{q-1}}\lambda^{1/2-1/q}$ holds for all λ.

10.19 Let (Ω, π) be a finite probability space, $|\Omega| \geq 2$, in which every outcome has probability at least λ. Let $1 < p < 2$ and $0 < \rho < 1$. The goal of this exercise is to prove the result of Wolff [Wol07] that, subject to $\|\mathrm{T}_\rho f\|_2 = 1$, every $f \in L^2(\Omega, \pi)$ that minimizes $\|f\|_p$ takes on at most two values (and there is at least one minimizing f).

(*a*) We consider the equivalent problem of minimizing $F(f) = \|f\|_p^p$ subject to $G(f) = \|\mathrm{T}_\rho f\|_2^2 = 1$. Show that both $F(f)$ and $G(f)$ are $\mathscr{C}^1$ functionals (identifying functions f with points in $\mathbb{R}^\Omega$).

(*b*) Argue from continuity that the minimum value for $\|f\|_p^p$ subject to $\|\mathrm{T}_\rho f\|_2^2 = 1$ is attained. Henceforth write f_0 to denote any minimizer; the goal is to show that f_0 takes on at most two values.

(*c*) Show that f_0 is either everywhere nonnegative or everywhere nonpositive. (Hint: By homogeneity our problem is equivalent to maximizing $\|\mathrm{T}_\rho f\|_2$ subject to $\|f\|_p = 1$; now use Exercise 2.34.) Replacing f_0 by $|f_0|$ if necessary, henceforth assume f_0 is nonnegative.

(*d*) Show that $\nabla F(f_0) = \pi \cdot p f_0^{p-1}$ and $\nabla G(f_0) = \pi \cdot 2\mathrm{T}_{\rho^2} f_0$. Here $\pi \cdot g$ signifies the pointwise product of functions on Ω, with π thought of as a function $\Omega \to \mathbb{R}^{\geq 0}$. (Hint: For the latter, write $G(f) = \langle \mathrm{T}_{\rho^2} f, f\rangle$.)

(*e*) Use the method of Lagrange Multipliers to show that $c f_0^{p-1} = T_{\rho^2} f_0$ for some $c \in \mathbb{R}^+$. (Hint: You'll need to note that $\nabla G(f_0) \neq 0$.)

(*f*) Writing $\mu = \mathbf{E}[f_0]$, argue that each value $y = f(\omega)$ satisfies the equation

$$c y^{p-1} = \rho^2 y + (1-\rho^2)\mu. \tag{10.33}$$

(*g*) Show that (10.33) has at most two solutions for $y \in \mathbb{R}^+$, thereby completing the proof that f_0 takes on at most two values. (Hint: Strict concavity of y^{p-1}.)

(*h*) Suppose $q > 2$. By slightly modifying the above argument, show that subject to $\|g\|_2 = 1$, every $g \in L^2(\Omega, \pi)$ that maximizes $\|\mathrm{T}_\rho g\|_q$ takes on at most two values (and there is at least one maximizing g). (Hint: At some point you might want to make the substitution $g = \mathrm{T}_\rho f$; note that g is two-valued if f is.)

10.20 Fix $1 < p < 2$ and $0 < \lambda < 1/2$. Let $\Omega = \{-1, 1\}$ and $\pi = \pi_\lambda$, meaning $\pi(-1) = \lambda$, $\pi(1) = 1-\lambda$. The goal of this exercise is to show the result of Latała and Oleszkiewicz [LO94]: the largest value of ρ for which $\|\mathrm{T}_\rho f\|_2 \leq \|f\|_p$ holds for all $f \in L^2(\Omega, \pi)$ is as given in Theorem 10.18; i.e., it satisfies

$$\rho^2 = r^* = \frac{\exp(u/p') - \exp(-u/p')}{\exp(u/p) - \exp(-u/p)}, \tag{10.34}$$

where u is defined by $\exp(-u) = \frac{\lambda}{1-\lambda}$. (Here we are using $p = q'$ to facilitate the proof; we get the $(2,q)$-hypercontractivity statement by Proposition 9.19.)

(*a*) Let's introduce the notation $\alpha = \lambda^{1/p}$, $\beta = (1-\lambda)^{1/p}$. Show that

$$r^* = \frac{\alpha^p \beta^{2-p} - \alpha^{2-p}\beta^p}{\alpha^2 - \beta^2}.$$

(*b*) Let $f \in L^2(\Omega, \pi)$. Write $\mu = \mathbf{E}[f]$ and $\delta = \mathrm{D}_1 f = \hat{f}(1)$. Our goal will be to show

$$\mu^2 + \delta^2 r^* = \|T_{\sqrt{r^*}} f\|_2^2 \leq \|f\|_p^2. \tag{10.35}$$

In the course of doing this, we'll also exhibit a nonconstant function f that makes the above inequality sharp. Why does this establish that no larger value of ρ is possible?

(*c*) Show that without loss of generality we may assume

$$f(-1) = \frac{1+y}{\alpha}, \quad f(1) = \frac{1-y}{\beta}$$

for some $-1 < y < 1$. (Hint: First use Exercise 2.34 and a continuity argument to show that we may assume $f > 0$; then use homogeneity of (10.35).)

(*d*) The left-hand side of (10.35) is now a quadratic function of y. Show that our r^* is precisely such that

$$\text{LHS(10.35)} = Ay^2 + C$$

for some constants A, C; i.e., r^* makes the linear term in y drop out. (Hint: Work exclusively with the α, β notation and recall from Definition 8.44 that $\delta^2 = \lambda(1-\lambda)(f(1)-f(-1))^2 = \alpha^p\beta^p(f(1)-f(-1))^2$.)

(*e*) Compute that

$$A = 2\frac{\beta^{p-1}-\alpha^{p-1}}{\beta-\alpha}. \tag{10.36}$$

(Hint: You'll want to multiply the above expression by $\alpha^p + \beta^p = 1$.)

(*f*) Show that

$$\text{RHS(10.35)} = ((1+y)^p + (1-y)^p)^{2/p}.$$

Why does it now suffice to show (10.35) just for $0 \le y < 1$?

(*g*) Let $y^* = \frac{\beta-\alpha}{\beta+\alpha} > 0$. Show that if $y = -y^*$, then f is a constant function and both sides of (10.35) are equal to $\frac{4}{(\alpha+\beta)^2}$.

(*h*) Deduce that both sides of (10.35) are equal to $\frac{4}{(\alpha+\beta)^2}$ for $y = y^*$. Verify that after scaling, this yields the following nonconstant function for which (10.35) is sharp: $f(x) = \exp(-xu/p)$.

(*i*) Write $y = \sqrt{z}$ for $0 \le z < 1$. By now we have reduced to showing

$$Az + C \le ((1+\sqrt{z})^p + (1-\sqrt{z})^p)^{2/p},$$

knowing that both sides are equal when $\sqrt{z} = y^*$. Calling the expression on the right $\phi(z)$, show that

$$\frac{d}{dz}\phi(z)\Big|_{\sqrt{z}=y^*} = A.$$

(Hint: You'll need $\alpha^p + \beta^p = 1$, as well as the fact from part (*h*) that $\phi(z) = \frac{4}{(\alpha+\beta)^2}$ when $\sqrt{z} = y^*$.) Deduce that we can complete the proof by showing that $\phi(z)$ is convex for $z \in [0,1)$.

(*j*) Show that ϕ is indeed convex on $[0,1)$ by showing that its derivative is a nondecreasing function of z. (Hint: Use the Generalized Binomial Theorem as well as $1 < p < 2$ to show that $(1+\sqrt{z})^p + (1-\sqrt{z})^p$ is expressible as $\sum_{j=0}^{\infty} b_j z^j$ where each b_j is positive.)

10.21 Complete the proof of Theorem 10.18. (Hint: Besides Exercises 10.19 and 10.20, you'll also need Exercise 10.18(*a*).)

10.22 (*a*) Let $\Phi : [0,\infty) \to \mathbb{R}$ be defined by $\Phi(x) = x \ln x$, where we take $0 \ln 0 = 0$. Verify that Φ is a smooth, strictly convex function.

(*b*) Consider the following:

Definition 10.49. Let $g \in L^2(\Omega, \pi)$ be a nonnegative function. The *entropy* of g is defined by

$$\mathbf{Ent}[g] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi}[\Phi(g(\boldsymbol{x}))] - \Phi\Big(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi}[g(\boldsymbol{x})]\Big).$$

Verify that $\mathbf{Ent}[g] \geq 0$ always, that $\mathbf{Ent}[g] = 0$ if and only if g is constant, and that $\mathbf{Ent}[cg] = c\mathbf{Ent}[g]$ for any constant $c \geq 0$.

(*c*) Suppose φ is a probability density on $\{-1,1\}^n$ (recall Definition 1.20). Show that $\mathbf{Ent}[\varphi] = D_{\mathrm{KL}}(\varphi \parallel \pi_{1/2}^{\otimes n})$, the Kullback–Leibler divergence of the uniform distribution from φ (more precisely, the distribution with density φ).

10.23 The goal of this exercise is to establish:

The Log-Sobolev Inequality. *Let* $f : \{-1,1\}^n \to \mathbb{R}$. *Then* $\frac{1}{2}\mathbf{Ent}[f^2] \leq \mathbf{I}[f]$.

(*a*) Writing $\rho = e^{-t}$, the $(p,2)$-Hypercontractivity Theorem tells us that

$$\|\mathrm{T}_{e^{-t}} f\|_2^2 \leq \|f\|_{1+\exp(-2t)}^2$$

for all $t \geq 0$. Denote the left- and right-hand sides as $\mathrm{LHS}(t)$, $\mathrm{RHS}(t)$. Verify that these are smooth functions of $t \in [0,\infty)$ and that $\mathrm{LHS}(0) = \mathrm{RHS}(0)$. Deduce that $\mathrm{LHS}'(0) \leq \mathrm{RHS}'(0)$.

(*b*) Compute $\mathrm{LHS}'(0) = -2\mathbf{I}[f]$. (Hint: Pass through the Fourier representation; cf. Exercise 2.18.)

(*c*) Compute $\mathrm{RHS}'(0) = -\mathbf{Ent}[f^2]$, thereby deducing the Log-Sobolev Inequality. (Hint: As an intermediate step, define $F(t) = \mathbf{E}[|f|^{1+\exp(-2t)}]$ and show that $\mathrm{RHS}'(0) = F(0)\ln F(0) + F'(0)$.)

10.24 (*a*) Let $f : \{-1,1\}^n \to \mathbb{R}$. Show that $\mathbf{Ent}[(1+\epsilon f)^2] \sim 2\mathbf{Var}[f]\epsilon^2$ as $\epsilon \to 0$.

(*b*) Deduce the Poincaré Inequality for f from the Log-Sobolev Inequality.

10.25 (*a*) Deduce from the Log-Sobolev Inequality that for $f : \{-1,1\}^n \to \{-1,1\}$ with $\alpha = \min\{\mathbf{Pr}[f=1], \mathbf{Pr}[f=-1]\}$,

$$2\alpha \ln(1/\alpha) \leq \mathbf{I}[f]. \tag{10.37}$$

This is off by a factor of $\ln 2$ from the optimal edge-isoperimetric inequality Theorem 2.39. (Hint: Apply the inequality to either $\frac{1}{2}-\frac{1}{2}f$ or $\frac{1}{2}+\frac{1}{2}f$.)

(*b*) Give a more streamlined direct derivation of (10.37) by differentiating the Small-Set Expansion Theorem.

10.26 This exercise gives a direct proof of the Log-Sobolev Inequality.

(*a*) The first step is to establish the $n=1$ case. Toward this, show that we may assume $f : \{-1,1\} \to \mathbb{R}$ is nonnegative and has mean 1. (Hints: Exercise 2.14, Exercise 10.22(*b*).)

(*b*) Thus it remains to establish $\frac{1}{2}\mathbf{Ent}[(1+b\boldsymbol{x})^2] \le b^2$ for $b \in [-1,1]$. Show that $g(b) = b^2 - \frac{1}{2}\mathbf{Ent}[(1+b\boldsymbol{x})^2]$ is smooth on $[-1,1]$ and satisfies $g(0) = 0$, $g'(0) = 0$, and $g''(b) = \frac{2b^2}{1+b^2} + \ln\frac{1+b^2}{1-b^2} \ge 0$ for $b \in (-1,1)$. Explain why this completes the proof of the $n=1$ case of the Log-Sobolev Inequality.

(*c*) Show that for any two functions $f_+, f_- : \{-1,1\}^n \to \mathbb{R}$,

$$\left(\frac{\sqrt{\mathbf{E}[f_+^2]}-\sqrt{\mathbf{E}[f_-^2]}}{2}\right)^2 \le \mathbf{E}\left[\left(\frac{f_+-f_-}{2}\right)^2\right].$$

(Hint: The triangle inequality for $\|\cdot\|_2$.)

(*d*) Prove the Log-Sobolev Inequality via “induction by restrictions” (as described in Section 9.4). (Hint: For the right-hand side, establish $\mathbf{Inf}[f] = \mathbf{E}[(\frac{f_+-f_-}{2})^2] + \frac{1}{2}\mathbf{I}[f_+] + \frac{1}{2}\mathbf{I}[f_-]$. For the left-hand side, apply induction, then the $n=1$ base case, then part (*c*).)

10.27 (*a*) By following the strategy of Exercise 10.23, establish the following:

Log-Sobolev Inequality for general product space domains. *Let* $f \in L^2(\Omega^n, \pi^{\otimes n})$ *and write* $\lambda = \min(\pi)$, $\lambda' = 1-\lambda$, $\exp(-u) = \frac{\lambda}{\lambda'}$. *Then* $\frac{1}{2}\varrho\mathbf{Ent}[f^2] \le \mathbf{I}[f]$, *where*

$$\varrho = \varrho(\lambda) = \frac{\tanh(u/2)}{u/2} = 2\frac{\lambda'-\lambda}{\ln\lambda' - \ln\lambda}.$$

(*b*) Show that $\varrho(\lambda) \sim 2/\ln(1/\lambda))$ as $\lambda \to 0$.

(*c*) Let $f : \{-1,1\}^n \to \{-1,1\}$ and treat $\{-1,1\}^n$ as having the p-biased distribution $\pi_p^{\otimes n}$. Write $q = 1-p$. Show that if $\alpha = \min\{\mathbf{Pr}_{\pi_p}[f = 1], \mathbf{Pr}_{\pi_p}[f=-1]\}$, then

$$4\frac{q-p}{\ln q - \ln p}\alpha\ln(1/\alpha) \le \mathbf{I}[f^{(p)}]$$

and hence, for $p \to 0$,

$$\alpha\log_p\alpha \le (1+o_p(1))p \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[\mathrm{sens}_f(\boldsymbol{x})]. \tag{10.38}$$

We remark that (10.38) is known to hold without the $o_p(1)$ for all $p \le 1/2$.

10.28 Prove Theorem 10.21. (Hint: Recall Proposition 8.28.)

10.29 Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent $(2, q, \rho)$-hypercontractive random variables and let $F(x) = \sum_{|S| \leq k} \widehat{F}(S) x^S$ be an n-variate multilinear polynomial of degree at most k. Show that

$$\|F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\|_q \leq (1/\rho)^k \|F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\|_2.$$

(Hint: You'll need Exercise 10.3.)

10.30 Let $0 < \lambda \leq 1/2$ and let (Ω, π) be a finite probability space in which some outcome $\omega_0 \in \Omega$ has $\pi(\omega_0) = \lambda$. (For example, $\Omega = \{-1, 1\}$, $\pi = \pi_\lambda$.) Define $f \in L^2(\Omega, \pi)$ by setting $f(\omega_0) = 1$, $f(\omega) = 0$ for $\omega \neq \omega_0$. For $q \geq 2$, compute $\|f\|_q / \|f\|_2$ and deduce (in light of the proof of Theorem 10.21) that Corollary 10.20 cannot hold for $\rho > \lambda^{1/2 - 1/q}$.

10.31 Prove Theorem 10.22.

10.32 Prove Theorem 10.23.

10.33 Prove Theorem 10.24. (Hint: Immediately worsen $q-1$ to q so that finding the optimal choice of q is easier.)

10.34 Prove Theorem 10.25.

10.35 Prove Friedgut's Junta Theorem for general product spaces as stated in Section 10.3.

10.36 Show that (10.9) implies $F(p_c + \eta p_c) \geq 1 - \epsilon$ in the proof of Theorem 10.29. (Hint: Consider $\frac{d}{dp} \ln(1 - F(p))$.)

10.37 Justify the various calculations and observations in Example 10.45.

10.38 (*a*) Let $p = \frac{1}{n}$ and let $f \in L^2(\{-1, 1\}^n, \pi_p^{\otimes n})$ be any Boolean-valued function. Show that $\mathbf{I}[f] \leq 4$. (Hint: Proposition 8.45.)

(*b*) Let us specialize to the case $f = \chi_{[n]}$. Show that f is not .1-close to any width-$O(1)$ DNF (under the $\frac{1}{n}$-biased distribution, for n sufficiently large). This shows that the assumption of monotonicity can't be removed from Friedgut's Conjecture. (Hint: Show that fixing any constant number of coordinates cannot change the bias of $\chi_{[n]}$ very much.)

10.39 A function $h : \Omega^n \to \Sigma$ is said to expressed as a *pseudo-junta* if the following hold: There are "juntas" $f_1, \dots, f_m : \Omega^n \to \{\mathsf{True}, \mathsf{False}\}$ with domains $J_1, \dots, J_m \subseteq [n]$ respectively. Further, $g : (\Omega \cup \{*\})^n \to \Sigma$, where $*$ is a new symbol not in Ω. Finally, for each input $x \in \Omega^n$ we have $h(x) = g(y)$, where for $j \in [n]$,

$$y_j = \begin{cases} x_j & \text{if } j \in J_i \text{ for some } i \text{ with } f_i(x) = \mathsf{True}, \\ * & \text{else.} \end{cases}$$

An alternative explanation is that on input x, the junta f_i decides whether the coordinates in its domain are "notable"; then, $h(x)$ must be determined based only on the set of all notable coordinates. Finally, if π is a distribution on Ω, we say that the pseudo-junta has *width-k under $\pi^{\otimes n}$* if

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\#\{j : \boldsymbol{y}_j \neq *\}] \leq k;$$

in other words, the expected number of notable coordinates is at most k. For $h \in L^2(\Omega^n, \pi^{\otimes n})$ we simply say that h is a *k-pseudo-junta*. Show that if such a k-pseudo-junta h is $\{-1,1\}$-valued, then $\mathbf{I}[f] \leq 4k$. (Hint: Referring to the second statement in Proposition 8.24, consider the notable coordinates for both $\boldsymbol{x}$ and $\boldsymbol{x}' = (\boldsymbol{x}_i, \dots, \boldsymbol{x}_{i-1}, \boldsymbol{x}'_i, \boldsymbol{x}_{i+1}, \dots, \boldsymbol{x}_n)$.)

10.40 Establish the following further consequence of Bourgain's Sharp Threshold Theorem: Let $f : \{\mathsf{True}, \mathsf{False}\}^n \to \{\mathsf{True}, \mathsf{False}\}$ be a monotone function with $\mathbf{I}[f^{(p)}] \leq K$. Assume $\mathbf{Var}[f] \geq .01$ and $0 < p \leq \exp(-cK^2)$, where c is a large universal constant. Then there exists $T \subseteq [n]$ with $|T| \leq O(K)$ such that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[f(\boldsymbol{x}) = \mathsf{True} \mid \boldsymbol{x}_i = \mathsf{True} \text{ for all } i \in T] \geq \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi_p^{\otimes n}}[f(\boldsymbol{x}) = \mathsf{True}] + \exp(-O(K^2)).$$

(Hint: Bourgain's Sharp Threshold Theorem yields a booster either toward True or toward False. In the former case you're easily done; to rule out the latter case, use the fact that $p|T| \ll \exp(-O(K^2))$.)

10.41 Suppose that in Bourgain's Sharp Threshold Theorem we drop the assumption that $\mathbf{Var}[f] \geq .01$. (Assume at least that f is nonconstant.) Show that there is some τ with $|\tau| \geq \mathbf{stddev}[f] \cdot \exp(-O(\mathbf{I}[f]^2/\mathbf{Var}[f]^2))$ such that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\pi^{\otimes n}}[\exists T \subseteq [n], |T| \leq O(\mathbf{I}[f]/\mathbf{Var}[f]) \text{ such that } \boldsymbol{x}_T \text{ is a } \tau\text{-booster}] \geq |\tau|.$$

(Cf. Exercise 9.32.)

10.42 In this exercise we give the beginnings of the idea of how Bourgain's Sharp Threshold Theorem can be used to show sharp thresholds for interesting monotone properties. We will consider $\neg$3Col, the property of a random v-vertex graph $\boldsymbol{G} \sim \mathcal{G}(v,p)$ being non-3-colorable.

(*a*) Prove that the critical probability p_c satisfies $p_c \leq O(1/v)$; i.e., establish that there is a universal constant C such that $\mathbf{Pr}[\boldsymbol{G} \sim \mathcal{G}(v, C/v) \text{ is 3-colorable}] = o_n(1)$. (Hint: Union-bound over all potential 3-colorings.)

(*b*) Toward showing (non-)3-colorability has a sharp threshold, suppose the property had constant total influence at the critical probability. Bourgain's Sharp Threshold Theorem would imply that there is a τ of constant magnitude such that for $\boldsymbol{G} \sim \mathcal{G}(v, p_c)$, there is a $|\tau|$ chance that $\boldsymbol{G}$ contains a τ-boosting induced subgraph $\boldsymbol{G}_T$. There are two cases, depending on the sign of τ. It's easy to rule out that the boost is in favor of 3-colorability; the absence of a few edges shouldn't increase the probability of 3-colorability by much (cf. Exercise 10.41).

On the other hand, it might seem plausible that the *presence* of a certain constant number of edges should boost the probability of non-3-colorability by a lot. For example, the presence of a 4-clique immediately boosts the probability to 1. However, the point is that *at the critical probability* it is very unlikely that $\boldsymbol{G}$ contains a 4-clique (or indeed, any "local" witness to non-3-colorability). Short of showing this, prove at least that the expected number of 4-cliques in $\boldsymbol{G} \sim \mathscr{G}(v,p)$ is $o_v(1)$ unless $p = \Omega(v^{-2/3}) \gg p_c$.

Notes. As mentioned, the standard template introduced by Bonami [**Bon70**] for proving the Hypercontractivity Theorem for ±1 bits is to first prove the Two-Point Inequality, and then do the induction described in Exercise 10.3. Bonami's original proof of the Two-Point Inequality reduced to the $1 \leq p < q \leq 2$ case as we did, but then her calculus was a little more cumbersome. We followed the proof of the Two-Point Inequality appearing in Janson [**Jan97**]. Our use of two-function hypercontractivity theorems to facilitate induction and avoid the use of Exercise 10.1 is nontraditional; it was inspired by Mossel et al. [**MOR$^+$06**], Barak et al. [**BBH$^+$12**], and Kauers et al. [**KOTZ13**]. The other main approach for proving the Hypercontractivity Theorem is to derive it from the Log-Sobolev Inequality (see Exercise 10.23), as was done by Gross [**Gro75**].

We are not aware of the Generalized Small-Set Expansion Theorem appearing previously in the literature; however, in a sense it's almost identical to the Reverse Small-Set Expansion Theorem, which is due to Mossel et al. [**MOR$^+$06**]. The Reverse Hypercontractivity Inequality itself is due to Borell [**Bor82**]; the presentation in Exercises 10.6–10.9 follows Mossel et al. [**MOR$^+$06**]. For more on reverse hypercontractivity, including the very surprising fact that the Reverse Hypercontractivity Inequality holds with no change in constants for every product probability space, see Mossel, Oleszkiewicz, and Sen [**MOS12**].

As mentioned in Chapter 9 the definition of a hypercontractive random variable is due to Krakowiak and Szulga [**KS88**]. Many of the basic facts from Section 10.2 (and also Exercise 10.2) are from this work and the earlier work of Borell [**Bor84**]; see also various other works [**KW92, Jan97, Szu98, MOO10**]. As mentioned, the main part of Theorem 10.18 (the case of biased bits) is essentially from Latała and Oleszkiewicz [**LO94**]; see also Oleszkiewicz [**Ole03**]. Our Exercise 10.20 fleshes out (and slightly simplifies) their computations but introduces no new idea. Earlier works [**BKK$^+$92, Tal94, FK96, Fri98**] had established forms of the General Hypercontractivity Theorem for λ-biased bits, giving as applications KKL-type theorems in this setting with the correct asymptotic dependence on λ. We should also mention that the sharp Log-Sobolev Inequality for product space domains (mentioned in Exercise 10.27) was derived independently of the Latała–Oleszkiewicz work by Higuchi and Yoshida [**HY95**] (without proof), by Diaconis and Saloff-Coste [**DSC96**] (with proof), and possibly also by Oscar Rothaus (see [**BL98**]). Unlike in the case of uniform ±1 bits, it's not known how to derive Latała and Oleszkiewicz's optimal biased hypercontractive inequality from the optimal biased Log-Sobolev Inequality.

Kahane [**Kah68**] has been credited with pioneering the randomization/symmetrization trick for random variables. The entirety of Section 10.4 is due to Bourgain [**Bou79**], though our presentation was significantly informed by the expertise of Krzysztof Oleszkiewicz (and our proof of Lemma 10.43 is slightly different). Like Bourgain, we don't give any explicit dependence for the constant C_q in Theorem 10.39; however, Kwapień [**Kwa10**] has shown that one may take $C_{q'} = C_q = O(q/\log q)$ for $q \geq 2$. Our proof of Bourgain's Theorem 10.47 follows the original [**Bou99**] extremely closely, though we also valued the easier-to-read version of Bal [**Bal13**].

The biased edge-isoperimetric inequality (10.38) from Exercise 10.27 was proved by induction on n, without the additional $o_p(1)$ error, by Russo [**Rus82**] (and also independently by Kahn and Kalai [**KK07**]). We remark that this work and the earlier [**Rus81**] already contain the germ of the idea that monotone functions with small influences have sharp thresholds. Regarding the sharp threshold for 3-colorability discussed in Exercise 10.42, Alon and Spencer [**AS08**] contains a nice elementary proof of the fact that at the critical probability for 3-colorability, every subgraph on ϵv vertices is 3-colorable, for some universal $\epsilon > 0$. The existence of a sharp threshold for k-colorability was proven by Achlioptas and Friedgut [**AF99**], with Achlioptas and Naor [**AN05**] essentially determining the location.

Chapter 11

Gaussian space and Invariance Principles

The final destination of this chapter is a proof of the following theorem due to Mossel, O'Donnell, and Oleszkiewicz [**MOO05b, MOO10**], first mentioned in Chapter 5.2:

Majority Is Stablest Theorem. *Fix $\rho \in (0,1)$. Let $f : \{-1,1\}^n \to [-1,1]$ have* $\mathbf{E}[f] = 0$. *Then, assuming* $\mathbf{MaxInf}[f] \leq \epsilon$*, or more generally that f has no (ϵ,ϵ)-notable coordinates,*

$$\mathbf{Stab}_\rho[f] \leq 1 - \tfrac{2}{\pi}\arccos\rho + o_\epsilon(1).$$

This bound is tight; recalling Theorem 2.45, the bound $1 - \frac{2}{\pi}\arccos\rho$ is achieved by taking $f = \mathrm{Maj}_n$, the volume-$\frac{1}{2}$ Hamming ball indicator, for $n \to \infty$. More generally, in Section 11.7 we'll prove the General-Volume Majority Is Stablest Theorem, which shows that for *any* fixed volume, "Hamming ball indicators have maximal noise stability among small-influence functions".

There are two main ideas underlying this theorem. The first is that "functions on Gaussian space" are a special case of small-influence Boolean functions. In other words, a Boolean function may always be a "Gaussian function in disguise". This motivates *analysis of Gaussian functions*, the topic introduced in Sections 11.1 and 11.2. It also means that a prerequisite for proving the (General-Volume) Majority Is Stablest Theorem is proving its Gaussian special cases, namely, Borell's Isoperimetric Theorem (Section 11.3) and the Gaussian Isoperimetric Inequality (Section 11.4). In many ways, working in the Gaussian setting is nicer because tools like rotational symmetry and differentiation are available.

The second idea is the converse to the first: In Section 11.6 we prove the *Invariance Principle*, a generalization of the Berry–Esseen Central Limit Theorem, which shows that any low-degree (or uniformly noise-stable) Boolean function with small influences is approximable by a Gaussian function. In fact, the Invariance Principle roughly shows that given such a Boolean function, if you plug *any* independent mean-0, variance-1 random variables into its Fourier expansion, the distribution doesn't change much. In Section 11.7 we use the Invariance Principle to prove the Majority Is Stablest Theorem by reducing to its Gaussian special case, Borell's Isoperimetric Theorem.

11.1. Gaussian space and the Gaussian noise operator

We begin with a few definitions concerning Gaussian space.

Notation 11.1. Throughout this chapter we write φ for the pdf of a standard Gaussian random variable, $\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2)$. We also write Φ for its cdf, and $\overline{\Phi}$ for the complementary cdf $\overline{\Phi}(t) = 1 - \Phi(t) = \Phi(-t)$. We write $\boldsymbol{z} \sim \mathrm{N}(0,1)^n$ to denote that $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ is a random vector in $\mathbb{R}^n$ whose components $\boldsymbol{z}_i$ are independent Gaussians. Perhaps the most important property of this distribution is that it's rotationally symmetric; this follows because the pdf at z is $\frac{1}{(2\pi)^{n/2}}\exp(-\frac{1}{2}(z_1^2 + \cdots + z_n^2))$, which depends only on the length $\|z\|_2^2$ of z.

Definition 11.2. For $n \in \mathbb{N}^+$ and $1 \leq p \leq \infty$ we write $L^p(\mathbb{R}^n, \gamma)$ for the space of Borel functions $f : \mathbb{R}^n \to \mathbb{R}$ that have finite pth moment $\|f\|_p^p$ under the Gaussian measure (the "γ" stands for Gaussian). Here for a function f on Gaussian space we use the notation

$$\|f\|_p = \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \mathrm{N}(0,1)^n}[|f(\boldsymbol{z})|^p]^{1/p}.$$

All functions $f : \mathbb{R}^n \to \mathbb{R}$ and sets $A \subseteq \mathbb{R}^n$ are henceforth assumed to be Borel without further mention.

Notation 11.3. When it's clear from context that f is a function on Gaussian space we'll use shorthand notation like $\mathbf{E}[f] = \mathbf{E}_{\boldsymbol{z} \sim \mathrm{N}(0,1)^n}[f(\boldsymbol{z})]$. If $f = 1_A$ is the 0-1 indicator of a subset $A \subseteq \mathbb{R}^n$ we'll also write

$$\mathrm{vol}_\gamma(A) = \mathbf{E}[1_A] = \mathop{\mathbf{Pr}}_{\boldsymbol{z} \sim \mathrm{N}(0,1)^n}[\boldsymbol{z} \in A]$$

for the *Gaussian volume* of A.

Notation 11.4. For $f, g \in L^2(\mathbb{R}^n, \gamma)$ we use the inner product notation $\langle f, g\rangle = \mathbf{E}[fg]$, under which $L^2(\mathbb{R}^n, \gamma)$ is a separable Hilbert space.

If you're only interested in Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ you might wonder why it's necessary to study Gaussian space. As discussed at the beginning of the chapter, the reason is that functions on Gaussian space are *special cases* of Boolean functions. Conversely, even if you're only interested in studying functions of Gaussian random variables, sometimes the easiest proof technique involves "simulating" the Gaussians using sums of random bits. Let's discuss this in a little more detail. Recall that the Central Limit Theorem tells us that for $\boldsymbol{x} \sim \{-1,1\}^M$, the distribution of $\frac{1}{\sqrt{M}}(\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_M)$ approaches that of a standard Gaussian as $M \to \infty$. This is the sense in which a standard Gaussian random variable $\boldsymbol{z} \sim \mathrm{N}(0,1)$ can be "simulated" by random bits. If we want d independent Gaussians we can simulate them by summing up M independent d-dimensional vectors of random bits.

Definition 11.5. The function $\mathrm{BitsToGaussians}_M : \{-1,1\}^M \to \mathbb{R}$ is defined by

$$\mathrm{BitsToGaussians}_M(x) = \tfrac{1}{\sqrt{M}}(x_1 + \cdots + x_M).$$

More generally, the function $\mathrm{BitsToGaussians}_M^d : \{-1,1\}^{dM} \to \mathbb{R}^d$ is defined on an input $x \in \{-1,1\}^{d \times M}$, thought of as a matrix of column vectors $\vec{x}_1, \dots, \vec{x}_M \in \{-1,1\}^d$, by

$$\mathrm{BitsToGaussians}_M^d(x) = \tfrac{1}{\sqrt{M}}(\vec{x}_1 + \cdots + \vec{x}_M).$$

Although M needs to be large for this simulation to be accurate, many of the results we've developed in the analysis of Boolean functions $f : \{-1,1\}^M \to \mathbb{R}$ are independent of M. A further key point is that this simulation preserves polynomial degree: if $p(\boldsymbol{z}_1, \dots, \boldsymbol{z}_d)$ is a degree-k polynomial applied to d independent standard Gaussians, the "simulated version" $p \circ \mathrm{BitsToGaussians}_M^d : \{-1,1\}^{dM} \to \mathbb{R}$ is a degree-k Boolean function. These facts allow us to transfer many results from the analysis of Boolean functions to the analysis of Gaussian functions. On the other hand, it also means that to fully understand Boolean functions, we need to understand the "special case" of functions on Gaussian space: a Boolean function may essentially be a function on Gaussian space "in disguise". For example, as we saw in Chapter 5.3, there is a sense in which the majority function Maj_n "converges" as $n \to \infty$; what it's converging to is the sign function on 1-dimensional Gaussian space, $\mathrm{sgn} \in L^1(\mathbb{R}, \gamma)$.

We'll begin our study of Gaussian functions by developing the analogue of the most important operator on Boolean functions, namely the noise operator T_ρ. Suppose we take a pair of ρ-correlated M-bit strings $(\boldsymbol{x}, \boldsymbol{x}')$ and use them to form approximate Gaussians,

$$\boldsymbol{y} = \mathrm{BitsToGaussians}_M(\boldsymbol{x}), \qquad \boldsymbol{y}' = \mathrm{BitsToGaussians}_M(\boldsymbol{x}').$$

For each M it's easy to compute that $\mathbf{E}[\boldsymbol{y}] = \mathbf{E}[\boldsymbol{y}'] = 0$, $\mathbf{Var}[\boldsymbol{y}] = \mathbf{Var}[\boldsymbol{y}'] = 1$, and $\mathbf{E}[\boldsymbol{y}\boldsymbol{y}'] = \rho$. As noted in Chapter 5.2, a multidimensional version of the Central Limit Theorem (see, e.g., Exercises 5.33, 11.46) tells us that the joint distribution of $(\boldsymbol{y}, \boldsymbol{y}')$ converges to a pair of Gaussian random variables with the same properties. We call these ρ-correlated Gaussians.

Definition 11.6. For $-1 \le \rho \le 1$, we say that the random variables $(\boldsymbol{z}, \boldsymbol{z}')$ are *ρ-correlated (standard) Gaussians* if they are jointly Gaussian and satisfy $\mathbf{E}[\boldsymbol{z}] = \mathbf{E}[\boldsymbol{z}'] = 0$, $\mathbf{Var}[\boldsymbol{z}] = \mathbf{Var}[\boldsymbol{z}'] = 1$, and $\mathbf{E}[\boldsymbol{z}\boldsymbol{z}'] = \rho$. In other words, if

$$(\boldsymbol{z}, \boldsymbol{z}') \sim \mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

Note that the definition is symmetric in $\boldsymbol{z}$, $\boldsymbol{z}'$ and that each is individually distributed as $\mathrm{N}(0,1)$.

Fact 11.7. *An equivalent definition is to say that $\boldsymbol{z} = \langle \vec{u}, \vec{\boldsymbol{g}} \rangle$ and $\boldsymbol{z}' = \langle \vec{v}, \vec{\boldsymbol{g}} \rangle$, where $\vec{\boldsymbol{g}} \sim \mathrm{N}(0,1)^d$ and $\vec{u}, \vec{v} \in \mathbb{R}^d$ are any two unit vectors satisfying $\langle \vec{u}, \vec{v} \rangle = \rho$. In particular we may choose $d = 2$, $\vec{u} = (1,0)$, and $\vec{v} = (\rho, \sqrt{1-\rho^2})$, thereby defining $\boldsymbol{z} = \boldsymbol{g}_1$ and $\boldsymbol{z}' = \rho \boldsymbol{g}_1 + \sqrt{1-\rho^2}\boldsymbol{g}_2$.*

Remark 11.8. In Fact 11.7 it's often convenient to write $\rho = \cos\theta$ for some $\theta \in \mathbb{R}$, in which case we may define the ρ-correlated Gaussians as $\boldsymbol{z} = \langle \vec{u}, \vec{\boldsymbol{g}} \rangle$ and $\boldsymbol{z}' = \langle \vec{v}, \vec{\boldsymbol{g}} \rangle$ for any unit vectors $\vec{u}, \vec{v}$ making an angle of θ; e.g., $\vec{u} = (1,0)$, $\vec{v} = (\cos\theta, \sin\theta)$.

Definition 11.9. For a fixed $z \in \mathbb{R}$ we say random variable $\boldsymbol{z}'$ is a *Gaussian ρ-correlated to z*, written $\boldsymbol{z}' \sim N_\rho(z)$, if $\boldsymbol{z}'$ is distributed as $\rho z + \sqrt{1-\rho^2}\boldsymbol{g}$ where $\boldsymbol{g} \sim \mathrm{N}(0,1)$. By Fact 11.7, if we draw $\boldsymbol{z} \sim \mathrm{N}(0,1)$ and then form $\boldsymbol{z}' \sim N_\rho(\boldsymbol{z})$, we obtain a ρ-correlated pair of Gaussians $(\boldsymbol{z}, \boldsymbol{z}')$.

Definition 11.10. For $-1 \le \rho \le 1$ and $n \in \mathbb{N}^+$ we say that the $\mathbb{R}^n$-valued random variables $(\boldsymbol{z}, \boldsymbol{z}')$ are *ρ-correlated n-dimensional Gaussian random vectors* if each component pair $(\boldsymbol{z}_1, \boldsymbol{z}'_1)$, …, $(\boldsymbol{z}_n, \boldsymbol{z}'_n)$ is a ρ-correlated pair of Gaussians, and the n pairs are mutually independent. We also naturally extend the definition of $\boldsymbol{z}' \sim N_\rho(z)$ to the case of $z \in \mathbb{R}^n$; this means $\boldsymbol{z}' = \rho z + \sqrt{1-\rho^2}\boldsymbol{g}$ for $\boldsymbol{g} \sim \mathrm{N}(0,1)^n$.

Remark 11.11. Thus, if $\boldsymbol{z} \sim \mathrm{N}(0,1)^n$ and then $\boldsymbol{z}' \sim N_\rho(\boldsymbol{z})$ we obtain a ρ-correlated n-dimensional pair $(\boldsymbol{z}, \boldsymbol{z}')$. It follows from this that the joint distribution of such a pair is rotationally symmetric (since the distribution of a single n-dimensional Gaussian is).

Now we can introduce the Gaussian analogue of the noise operator.

Definition 11.12. For $\rho \in [-1,1]$, the *Gaussian noise operator* U_ρ is the linear operator defined on the space of functions $f \in L^1(\mathbb{R}^n, \gamma)$ by

$$\mathrm{U}_\rho f(z) = \mathop{\mathbf{E}}_{\boldsymbol{z}' \sim N_\rho(z)}[f(\boldsymbol{z}')] = \mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[f(\rho z + \sqrt{1-\rho^2}\boldsymbol{g})].$$

Fact 11.13. *(Exercise 11.3.) If $f \in L^1(\mathbb{R}^n, \gamma)$ is an n-variate multilinear polynomial, then $\mathrm{U}_\rho f(z) = f(\rho z)$.*

Remark 11.14. Our terminology is nonstandard. The Gaussian noise operators are usually collectively referred to as the *Ornstein–Uhlenbeck semigroup* (or sometimes as the *Mehler transforms*). They are typically defined for $\rho = e^{-t} \in [0,1]$ (i.e., for $t \in [0,\infty]$) by

$$\mathrm{P}_t f(z) = \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)^n}[f(e^{-t}z + \sqrt{1-e^{-2t}}\boldsymbol{g})] = \mathrm{U}_{e^{-t}} f(z).$$

The term "semigroup" refers to the fact that the operators satisfy $\mathrm{P}_{t_1}\mathrm{P}_{t_2} = \mathrm{P}_{t_1+t_2}$, i.e., $\mathrm{U}_{\rho_1}\mathrm{U}_{\rho_2} = \mathrm{U}_{\rho_1\rho_2}$ (which holds for all $\rho_1, \rho_2 \in [-1,1]$; see Exercise 11.4).

Before going further let's check that U_ρ is a bounded operator on all of $L^p(\mathbb{R}^n,\gamma)$ for $p \geq 1$; in fact, it's a contraction (cf. Exercise 2.33):

Proposition 11.15. *For each $\rho \in [-1,1]$ and $1 \leq p \leq \infty$ the operator U_ρ is a contraction on $L^p(\mathbb{R}^n,\gamma)$; i.e., $\|\mathrm{U}_\rho f\|_p \leq \|f\|_p$.*

Proof. The proof for $p = \infty$ is easy; otherwise, the result follows from Jensen's inequality, using that $t \mapsto |t|^p$ is convex:

$$\begin{aligned}\|\mathrm{U}_\rho f\|_p^p = \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\mathrm{N}(0,1)^n}[|\mathrm{U}_\rho f(\boldsymbol{z})|^p] &= \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\mathrm{N}(0,1)^n}\left[\left|\mathop{\mathbf{E}}_{\boldsymbol{z}'\sim N_\rho(\boldsymbol{z})}[f(\boldsymbol{z}')]\right|^p\right] \\ &\leq \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\mathrm{N}(0,1)^n}\left[\mathop{\mathbf{E}}_{\boldsymbol{z}'\sim N_\rho(\boldsymbol{z})}[|f(\boldsymbol{z}')|^p]\right] = \|f\|_p^p. \qquad \square\end{aligned}$$

As in the Boolean case, you should think of the Gaussian noise operator as having a "smoothing" effect on functions. As ρ goes from 1 down to 0, $\mathrm{U}_\rho f$ involves averaging f's values over larger and larger neighborhoods. In particular U_1 is the identity operator, $\mathrm{U}_1 f = f$, and $\mathrm{U}_0 f = \mathbf{E}[f]$, the constant function. In Exercises 11.5, 11.6 you are asked to verify the following facts, which say that for any f, as $\rho \to 1^-$ we get a sequence of smooth (i.e., $\mathscr{C}^\infty$) functions $\mathrm{U}_\rho f$ that tend to f.

Proposition 11.16. *Let $f \in L^1(\mathbb{R}^n,\gamma)$ and let $-1 < \rho < 1$. Then $\mathrm{U}_\rho f$ is a smooth function.*

Proposition 11.17. *Let $f \in L^1(\mathbb{R}^n,\gamma)$. As $\rho \to 1^-$ we have $\|\mathrm{U}_\rho f - f\|_1 \to 0$.*

Having defined the Gaussian noise operator, we can also make the natural definition of Gaussian noise stability (for which we'll use the same notation as in the Boolean case):

Definition 11.18. For $f \in L^2(\mathbb{R}^n,\gamma)$ and $\rho \in [-1,1]$, the *Gaussian noise stability of f at ρ* is defined to be

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ n\text{-dimensional} \\ \rho\text{-correlated Gaussians}}}[f(\boldsymbol{z})f(\boldsymbol{z}')] = \langle f, \mathrm{U}_\rho f\rangle = \langle \mathrm{U}_\rho f, f\rangle.$$

(Here we used that $(\boldsymbol{z}',\boldsymbol{z})$ has the same distribution as $(\boldsymbol{z},\boldsymbol{z}')$ and hence U_ρ is self-adjoint.)

Example 11.19. Let $f : \mathbb{R} \to \{0,1\}$ be the 0-1 indicator of the nonpositive halfline: $f = 1_{(-\infty,0]}$. Then

$$\mathbf{Stab}_\rho[f] = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated} \\ \text{standard Gaussians}}}[f(\boldsymbol{z})f(\boldsymbol{z}')] = \mathbf{Pr}[\boldsymbol{z} \le 0, \boldsymbol{z}' \le 0] = \frac{1}{2} - \frac{1}{2}\frac{\arccos\rho}{\pi}, \tag{11.1}$$

with the last equality being *Sheppard's Formula*, which we stated in Section 5.2 and now prove.

Proof of Sheppard's Formula. Since $(-\boldsymbol{z},-\boldsymbol{z}')$ has the same distribution as $(\boldsymbol{z},\boldsymbol{z}')$, proving (11.1) is equivalent to proving

$$\mathbf{Pr}[\boldsymbol{z} \le 0, \boldsymbol{z}' \le 0 \text{ or } \boldsymbol{z} > 0, \boldsymbol{z}' > 0] = 1 - \frac{\arccos\rho}{\pi}.$$

The complement of the above event is the event that $f(\boldsymbol{z}) \neq f(\boldsymbol{z}')$ (up to measure 0); thus it's further equivalent to prove

$$\mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \\ \cos\theta\text{-correlated}}}[f(\boldsymbol{z}) \neq f(\boldsymbol{z}')] = \tfrac{\theta}{\pi} \tag{11.2}$$

for all $\theta \in [0,\pi]$. As in Remark 11.8, this suggests defining $\boldsymbol{z} = \langle \vec{u}, \vec{\boldsymbol{g}} \rangle$, $\boldsymbol{z}' = \langle \vec{v}, \vec{\boldsymbol{g}} \rangle$, where $\vec{u},\vec{v} \in \mathbb{R}^2$ is some fixed pair of unit vectors making an angle of θ, and $\vec{\boldsymbol{g}} \sim \mathrm{N}(0,1)^2$. Thus we want to show

$$\mathop{\mathbf{Pr}}_{\vec{\boldsymbol{g}} \sim \mathrm{N}(0,1)^2}[\langle \vec{u}, \vec{\boldsymbol{g}} \rangle \le 0 \ \&\ \langle \vec{v}, \vec{\boldsymbol{g}} \rangle > 0 \text{ or vice versa}] = \tfrac{\theta}{\pi}.$$

But this last identity is easy: If we look at the diameter of the unit circle that is perpendicular to $\vec{\boldsymbol{g}}$, then the event above is equivalent (up to measure 0) to the event that this diameter "splits" $\vec{u}$ and $\vec{v}$. By the rotational symmetry of $\vec{\boldsymbol{g}}$, the probability is evidently θ (the angle between $\vec{u},\vec{v}$) divided by π (the range of angles for the diameter). □

Corollary 11.20. *Let $H \subset \mathbb{R}^n$ be any halfspace (open or closed) with boundary hyperplane containing the origin. Let $h = \pm 1_H$. Then* $\mathbf{Stab}_\rho[h] = 1 - \frac{2}{\pi}\arccos\rho$.

Proof. We may assume H is open (since its boundary has measure 0). By the rotational symmetry of correlated Gaussians (Remark 11.11), we may rotate H to the form $H = \{z \in \mathbb{R}^n : z_1 > 0\}$. Then it's clear that the noise stability of $h = \pm 1_H$ doesn't depend on n, i.e., we may assume $n = 1$. Thus $h = \mathrm{sgn} = 1 - 2f$, where $f = 1_{(-\infty,0]}$ as in Example 11.19. Now if $(\boldsymbol{z},\boldsymbol{z}')$ denote ρ-correlated standard Gaussians, it follows from (11.1) that

$$\begin{aligned}\mathbf{Stab}_\rho[h] = \mathbf{E}[h(\boldsymbol{z})h(\boldsymbol{z}')] &= \mathbf{E}[(1-2f(\boldsymbol{z}))(1-2f(\boldsymbol{z}'))] \\ &= 1 - 4\mathbf{E}[f] + 4\mathbf{Stab}_\rho[f] = 1 - \tfrac{2}{\pi}\arccos\rho. \qquad \square\end{aligned}$$

Remark 11.21. The quantity $\mathbf{Stab}_\rho[\mathrm{sgn}] = 1 - \frac{2}{\pi}\arccos\rho$ is also precisely the limiting noise stability of Maj_n, as stated in Theorem 2.45 and justified in Chapter 5.2.

We've defined the key Gaussian noise operator U_ρ and seen (Proposition 11.15) that it's a contraction on all $L^p(\mathbb{R}^n, \gamma)$. Is it also hypercontractive? In fact, we'll now show that the Hypercontractivity Theorem for uniform ±1 bits holds identically in the Gaussian setting. The proof is simply a reduction to the Boolean case, and it will use the following standard fact (see Janson [**Jan97**, Theorem 2.6] or Teuwen [**Teu12**, Section 1.3] for the proof in case of L^2; to extend to other L^p you can use Exercise 11.1):

Theorem 11.22. *For each $n \in \mathbb{N}^+$, the set of multivariate polynomials is dense in $L^p(\mathbb{R}^n, \gamma)$ for all $1 \le p < \infty$.*

Gaussian Hypercontractivity Theorem. *Let $f, g \in L^1(\mathbb{R}^n, \gamma)$, let $r, s \ge 0$, and assume $0 \le \rho \le \sqrt{rs} \le 1$. Then*

$$\langle f, \mathrm{U}_\rho g\rangle = \langle \mathrm{U}_\rho f, g\rangle = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated} \\ n\text{-dimensional Gaussians}}}[f(\boldsymbol{z})g(\boldsymbol{z}')] \le \|f\|_{1+r}\|g\|_{1+s}.$$

Proof. (We give a sketch; you are asked to fill in the details in Exercise 11.2.) We may assume that $f \in L^{1+r}(\mathbb{R}^n, \gamma)$ and $g \in L^{1+s}(\mathbb{R}^n, \gamma)$. We may also assume $f, g \in L^2(\mathbb{R}^n, \gamma)$ by a truncation and monotone convergence argument; thus the left-hand side is finite by Cauchy–Schwarz. Finally, we may assume that f and g are multivariate polynomials, using Theorem 11.22. For fixed $M \in \mathbb{N}^+$ we consider "simulating" $(\boldsymbol{z}, \boldsymbol{z}')$ using bits. More specifically, let $(\boldsymbol{x}, \boldsymbol{x}') \in \{-1,1\}^{nM} \times \{-1,1\}^{nM}$ be a pair ρ-correlated random strings and define the joint $\mathbb{R}^n$-valued random variables $\boldsymbol{y}, \boldsymbol{y}'$ by

$$\boldsymbol{y} = \mathrm{BitsToGaussians}_M^n(\boldsymbol{x}), \qquad \boldsymbol{y}' = \mathrm{BitsToGaussians}_M^n(\boldsymbol{x}').$$

By a multidimensional Central Limit Theorem we have that

$$\mathbf{E}[f(\boldsymbol{y})g(\boldsymbol{y}')] \xrightarrow{M\to\infty} \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \\ \rho\text{-correlated}}}[f(\boldsymbol{z})g(\boldsymbol{z}')].$$

(Since f and g are polynomials, we can even reduce to a Central Limit Theorem for bivariate monomials.) We further have

$$\mathbf{E}[|f(\boldsymbol{y})|^{1+r}]^{1/(1+r)} \xrightarrow{M\to\infty} \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\mathrm{N}(0,1)^n}[|f(\boldsymbol{z})|^{1+r}]^{1/(1+r)}$$

and similarly for g. (This can also be proven by the multidimensional Central Limit Theorem, or by the one-dimensional Central Limit Theorem together with some tricks.) Thus it suffices to show

$$\mathbf{E}[f(\boldsymbol{y})g(\boldsymbol{y}')] \le \mathbf{E}[|f(\boldsymbol{y})|^{1+r}]^{1/(1+r)}\,\mathbf{E}[|g(\boldsymbol{y}')|^{1+s}]^{1/(1+s)}$$

for any fixed M. But we can express $f(\boldsymbol{y}) = F(\boldsymbol{x})$ and $g(\boldsymbol{y}') = G(\boldsymbol{x}')$ for some $F, G : \{-1,1\}^{nM} \to \mathbb{R}$ and so the above inequality holds by the Two-Function Hypercontractivity Theorem (for ±1 bits). □

An immediate corollary, using the proof of Proposition 10.4, is the standard one-function form of hypercontractivity:

Theorem 11.23. *Let $1 \le p \le q \le \infty$ and let $f \in L^p(\mathbb{R}^n, \gamma)$. Then $\|\mathrm{U}_\rho f\|_q \le \|f\|_p$ for $0 \le \rho \le \sqrt{\frac{p-1}{q-1}}$.*

We conclude this section by discussing the Gaussian space analogue of the discrete Laplacian operator. Taking our cue from Exercise 2.18 we make the following definition:

Definition 11.24. The *Ornstein–Uhlenbeck operator* L (also called the *infinitesimal generator* of the Ornstein–Uhlenbeck semigroup, or the *number operator*) is the linear operator acting on functions $f \in L^2(\mathbb{R}^n, \gamma)$ by

$$\mathrm{L}f = \frac{d}{d\rho}\mathrm{U}_\rho f\Big|_{\rho=1} = -\frac{d}{dt}\mathrm{U}_{e^{-t}} f\Big|_{t=0}$$

(provided Lf exists in $L^2(\mathbb{R}^n, \gamma)$). Notational warning: It is common to see this as the definition of −L.

Remark 11.25. We will not be completely careful about the domain of the operator L in this section; for precise details, see Exercise 11.18.

Proposition 11.26. *Let $f \in L^2(\mathbb{R}^n, \gamma)$ be in the domain of* L*, and further assume for simplicity that f is $\mathscr{C}^3$. Then we have the formula*

$$\mathrm{L}f(x) = x \cdot \nabla f(x) - \Delta f(x),$$

where Δ denotes the usual Laplacian differential operator, $\cdot$ denotes the dot product, and ∇ denotes the gradient.

Proof. We give the proof in the case $n = 1$, leaving the general case to Exercise 11.7. We have

$$\mathrm{L}f(x) = -\lim_{t \to 0^+} \frac{\mathbf{E}_{\boldsymbol{z}\sim \mathrm{N}(0,1)}[f(e^{-t}x + \sqrt{1-e^{-2t}}\boldsymbol{z})] - f(x)}{t}. \tag{11.3}$$

Applying Taylor's theorem to f we have

$$f(e^{-t}x + \sqrt{1-e^{-2t}}\boldsymbol{z}) \approx f(e^{-t}x) + f'(e^{-t}x)\sqrt{1-e^{-2t}}\boldsymbol{z} + \tfrac{1}{2}f''(e^{-t}x)(1-e^{-2t})\boldsymbol{z}^2,$$

where the ≈ denotes that the two quantities differ by at most $C(1-e^{-2t})^{3/2}|\boldsymbol{z}|^3$ in absolute value, for some constant C depending on f and x. Substituting this into (11.3) and using $\mathbf{E}[\boldsymbol{z}] = 0$, $\mathbf{E}[\boldsymbol{z}^2] = 1$, and that $\mathbf{E}[|\boldsymbol{z}|^3]$ is an absolute constant, we get

$$\mathrm{L}f(x) = -\lim_{t\to 0^+}\left(\frac{f(e^{-t}x)-f(x)}{t} + \frac{\frac{1}{2}f''(e^{-t}x)(1-e^{-2t})}{t}\right),$$

using the fact that $\frac{(1-e^{-2t})^{3/2}}{t} \to 0$. But this is easily seen to be $xf'(x) - f''(x)$, as claimed. □

An easy consequence of the semigroup property is the following:

Proposition 11.27. *The following equivalent identities hold:*

$$\frac{d}{d\rho}\mathrm{U}_\rho f = \rho^{-1}\mathrm{L}\mathrm{U}_\rho f = \rho^{-1}\mathrm{U}_\rho \mathrm{L}f,$$
$$\frac{d}{dt}\mathrm{U}_{e^{-t}} f = -\mathrm{L}\mathrm{U}_{e^{-t}} f = -\mathrm{U}_{e^{-t}}\mathrm{L}f.$$

Proof. This follows from

$$\begin{aligned}\frac{d}{dt}\mathrm{U}_{e^{-t}}f(x) &= \lim_{\delta\to 0}\frac{\mathrm{U}_{e^{-t-\delta}}f(x) - \mathrm{U}_{e^{-t}}f(x)}{\delta}\\ &= \lim_{\delta\to 0}\frac{\mathrm{U}_{e^{-\delta}}\mathrm{U}_{e^{-t}}f(x) - \mathrm{U}_{e^{-t}}f(x)}{\delta} = \lim_{\delta\to 0}\frac{\mathrm{U}_{e^{-t}}\mathrm{U}_{e^{-\delta}}f(x) - \mathrm{U}_{e^{-t}}f(x)}{\delta}.\end{aligned}$$ □

We also have the following formula:

Proposition 11.28. *Let $f, g \in L^2(\mathbb{R}^n,\gamma)$ be in the domain of* L*, and further assume for simplicity that they are $\mathscr{C}^3$. Then*

$$\langle f, \mathrm{L}g\rangle = \langle \mathrm{L}f, g\rangle = \langle \nabla f, \nabla g\rangle. \tag{11.4}$$

Proof. It suffices to prove the inequality on the right of (11.4). We again treat only the case of $n = 1$, leaving the general case to Exercise 11.8. Using Proposition 11.26,

$$\begin{aligned}\langle \mathrm{L}f, g\rangle &= \int_{\mathbb{R}}(xf'(x) - f''(x))g(x)\varphi(x)\,dx\\ &= \int_{\mathbb{R}} xf'(x)g(x)\varphi(x)\,dx + \int_{\mathbb{R}} f'(x)(g\varphi)'(x)\,dx \qquad \text{(integration by parts)}\\ &= \int_{\mathbb{R}} xf'(x)g(x)\varphi(x)\,dx + \int_{\mathbb{R}} f'(x)(g'(x)\varphi(x) + g(x)\varphi'(x))\,dx\\ &= \int_{\mathbb{R}} f'(x)g'(x)\varphi(x)\,dx,\end{aligned}$$

using the fact that $\varphi'(x) = -x\varphi(x)$. □

Finally, by differentiating the Gaussian Hypercontractivity Inequality we obtain the Gaussian Log-Sobolev Inequality (see Exercise 10.23; the proof is the same as in the Boolean case):

Gaussian Log-Sobolev Inequality. *Let $f \in L^2(\mathbb{R}^n, \gamma)$ be in the domain of* L. *Then*

$$\tfrac{1}{2}\mathbf{Ent}[f^2] \le \mathbf{E}[\|\nabla f\|^2].$$

It's tempting to use the notation $\mathbf{I}[f]$ for $\mathbf{E}[\|\nabla f\|^2]$; however, you have to be careful because this quantity is not equal to $\sum_{i=1}^n \mathbf{E}[\mathbf{Var}_{\boldsymbol{z}_i}[f]]$ unless f is a multilinear polynomial. See Exercise 11.13.

11.2. Hermite polynomials

Having defined the basic operators of importance for functions on Gaussian space, it's useful to also develop the analogue of the Fourier expansion. To do this we'll proceed as in Chapter 8.1, looking for a complete orthonormal "Fourier basis" for $L^2(\mathbb{R}, \gamma)$, which we can extend to $L^2(\mathbb{R}^n, \gamma)$ by taking products. It's natural to start with polynomials; by Theorem 11.22 we know that the collection $(\phi_j)_{j\in\mathbb{N}}$, $\phi_j(z) = z^j$ is a complete basis for $L^2(\mathbb{R}, \gamma)$. To get an orthonormal ("Fourier") basis we can simply perform the Gram–Schmidt process. Calling the resulting basis $(h_j)_{j\in\mathbb{N}}$ (with "h" standing for "Hermite"), we get

$$h_0(z) = 1, \quad h_1(z) = z, \quad h_2(z) = \frac{z^2-1}{\sqrt{2}}, \quad h_3(z) = \frac{z^3-3z}{\sqrt{6}}, \quad \dots \tag{11.5}$$

Here, e.g., we obtained $h_3(z)$ in two steps. First, we made $\phi_3(z) = z^3$ orthogonal to $h_0, \dots, h_2$ as

$$z^3 - \langle \boldsymbol{z}^3, 1\rangle \cdot 1 - \langle \boldsymbol{z}^3, \boldsymbol{z}\rangle \cdot z - \langle \boldsymbol{z}^3, \tfrac{\boldsymbol{z}^2-1}{\sqrt{2}}\rangle \cdot \tfrac{z^2-1}{\sqrt{2}} = z^3 - 3z,$$

where $\boldsymbol{z} \sim \mathrm{N}(0,1)$ and we used the fact that $\boldsymbol{z}^3$ and $\boldsymbol{z}^3 \cdot \frac{\boldsymbol{z}^2-1}{\sqrt{2}}$ are odd functions and hence have Gaussian expectation 0. Then we defined $h_3(z) = \frac{z^3-3z}{\sqrt{6}}$ after determining that $\mathbf{E}[(\boldsymbol{z}^3 - 3\boldsymbol{z})^2] = 6$.

Let's develop a more explicit definition of these Hermite polynomials. The computations involved in the Gram–Schmidt process require knowledge of the moments of a Gaussian random variable $\boldsymbol{z} \sim \mathrm{N}(0,1)$. It's most convenient to understand these moments through the moment generating function of $\boldsymbol{z}$, namely

$$\mathbf{E}[\exp(t\boldsymbol{z})] = \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tz} e^{-\frac{1}{2}z^2}\, dz = e^{\frac{1}{2}t^2} \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}(z-t)^2}\, dz = \exp(\tfrac{1}{2}t^2). \tag{11.6}$$

In light of our interest in the U_ρ operators, and the fact that orthonormality involves pairs of basis functions, we'll in fact study the moment generating function of a pair $(\boldsymbol{z}, \boldsymbol{z}')$ of ρ-correlated standard Gaussians. To compute it, assume $(\boldsymbol{z},\boldsymbol{z}')$ are generated as in Fact 11.7 with $\vec{u},\vec{v}$ unit vectors in $\mathbb{R}^2$. Then

$$\begin{aligned}\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \rho\text{-correlated}}}[\exp(s\boldsymbol{z}+t\boldsymbol{z}')] &= \mathop{\mathbf{E}}_{\substack{\boldsymbol{g}_1,\boldsymbol{g}_2\sim\mathrm{N}(0,1)\\ \text{independent}}}[\exp(s(u_1\boldsymbol{g}_1+u_2\boldsymbol{g}_2)+t(v_1\boldsymbol{g}_1+v_2\boldsymbol{g}_2))]\\ &= \mathop{\mathbf{E}}_{\boldsymbol{g}_1\sim\mathrm{N}(0,1)}[\exp((su_1+tv_1)\boldsymbol{g}_1)]\mathop{\mathbf{E}}_{\boldsymbol{g}_2\sim\mathrm{N}(0,1)}[\exp((su_2+tv_2)\boldsymbol{g}_2)]\\ &= \exp(\tfrac{1}{2}(su_1+tv_1)^2)\exp(\tfrac{1}{2}(su_2+tv_2)^2)\\ &= \exp(\tfrac{1}{2}\|\vec{u}\|_2^2s^2+\langle\vec{u},\vec{v}\rangle st+\tfrac{1}{2}\|\vec{v}\|_2^2t^2)\\ &= \exp(\tfrac{1}{2}(s^2+2\rho st+t^2)),\end{aligned}$$

where the third equality used (11.6). Dividing by $\exp(\frac{1}{2}(s^2+t^2))$ it follows that

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \rho\text{-correlated}}}[\exp(s\boldsymbol{z}-\tfrac{1}{2}s^2)\exp(t\boldsymbol{z}'-\tfrac{1}{2}t^2)] = \exp(\rho st) = \sum_{j=0}^{\infty}\frac{\rho^j}{j!}s^jt^j. \qquad (11.7)$$

Inside the expectation above we essentially have the expression $\exp(tz-\frac{1}{2}t^2)$ appearing twice. It's easy to see that if we take the power series in t for this expression, the coefficient on t^j will be a polynomial in z with leading term $\frac{1}{j!}z^j$. Let's therefore write

$$\exp(tz-\tfrac{1}{2}t^2) = \sum_{j=0}^{\infty}\frac{1}{j!}H_j(z)t^j, \qquad (11.8)$$

where $H_j(z)$ is a monic polynomial of degree j. Now substituting this into (11.7) yields

$$\sum_{j,k=0}^{\infty}\frac{1}{j!k!}\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \rho\text{-correlated}}}[H_j(\boldsymbol{z})H_k(\boldsymbol{z}')]s^jt^k = \sum_{j=0}^{\infty}\frac{\rho^j}{j!}s^jt^j.$$

Equating coefficients, it follows that we must have

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \rho\text{-correlated}}}[H_j(\boldsymbol{z})H_k(\boldsymbol{z}')] = \begin{cases} j!\rho^j & \text{if } j=k,\\ 0 & \text{if } j\neq k.\end{cases}$$

In particular (taking $\rho = 1$),

$$\langle H_j,H_k\rangle = \begin{cases} j! & \text{if } j=k,\\ 0 & \text{if } j\neq k;\end{cases} \qquad (11.9)$$

i.e., the polynomials $(H_j)_{j\in\mathbb{N}}$ are orthogonal. Furthermore, since H_j is monic and of degree j, it follows that the H_j's are precisely the polynomials that arise in the Gram–Schmidt orthogonalization of $\{1,z,z^2,\dots\}$. We also see from (11.9) that the orthonormalized polynomials $(h_j)_{j\in\mathbb{N}}$ are obtained by setting $h_j = \frac{1}{\sqrt{j!}}H_j$.

Let's summarize and introduce the terminology for what we've deduced.

Definition 11.29. The *probabilists' Hermite polynomials* $(H_j)_{j\in\mathbb{N}}$ are the univariate polynomials defined by the identity (11.8). An equivalent definition (Exercise 11.9) is

$$H_j(z) = \frac{(-1)^j}{\varphi(z)} \cdot \frac{d^j}{dz^j}\varphi(z). \tag{11.10}$$

The *normalized Hermite polynomials* $(h_j)_{j\in\mathbb{N}}$ are defined by $h_j = \frac{1}{\sqrt{j!}}H_j$; the first four are given explicitly in (11.5). For brevity we'll simply refer to the h_j's as the "Hermite polynomials", though this is not standard terminology.

Proposition 11.30. *The Hermite polynomials $(h_j)_{j\in\mathbb{N}}$ form a complete orthonormal basis for $L^2(\mathbb{R},\gamma)$. They are also a "Fourier basis", since $h_0 = 1$.*

Proposition 11.31. *For any $\rho \in [-1,1]$ we have*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \rho\text{-correlated}}}[h_j(\boldsymbol{z})h_k(\boldsymbol{z}')] = \langle h_j, \mathrm{U}_\rho h_k\rangle = \langle \mathrm{U}_\rho h_j, h_k\rangle = \begin{cases}\rho^j & \text{if } j = k,\\ 0 & \text{if } j \neq k.\end{cases}$$

From this "Fourier basis" for $L^2(\mathbb{R},\gamma)$ we can construct a "Fourier basis" for $L^2(\mathbb{R}^n,\gamma)$ just by taking products, as in Proposition 8.13.

Definition 11.32. For a multi-index $\alpha \in \mathbb{N}^n$ we define the *(normalized multivariate) Hermite polynomial* $h_\alpha : \mathbb{R}^n \to \mathbb{R}$ by

$$h_\alpha(z) = \prod_{j=1}^n h_{\alpha_j}(z_j).$$

Note that the total degree of h_α is $|\alpha| = \sum_j \alpha_j$. We also identify a subset $S \subseteq [n]$ with its indicator α defined by $\alpha_j = 1_{j\in S}$; thus $h_S(z)$ denotes $z^S = \prod_{j\in S} z_j$.

Proposition 11.33. *The Hermite polynomials $(h_\alpha)_{\alpha\in\mathbb{N}^n}$ form a complete orthonormal (Fourier) basis for $L^2(\mathbb{R}^n,\gamma)$. Further, for any $\rho\in[-1,1]$ we have*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \rho\text{-correlated}}}[h_\alpha(\boldsymbol{z})h_\beta(\boldsymbol{z}')] = \langle h_\alpha, \mathrm{U}_\rho h_\beta\rangle = \langle \mathrm{U}_\rho h_\alpha, h_\beta\rangle = \begin{cases}\rho^{|\alpha|} & \text{if } \alpha = \beta,\\ 0 & \text{if } \alpha \neq \beta.\end{cases}$$

We can now define the "Hermite expansion" of Gaussian functions.

Definition 11.34. Every $f \in L^2(\mathbb{R}^n,\gamma)$ is uniquely expressible as

$$f = \sum_{\alpha\in\mathbb{N}^n} \widehat{f}(\alpha)h_\alpha,$$

where the real numbers $\widehat{f}(\alpha)$ are called the *Hermite coefficients of f* and the convergence is in $L^2(\mathbb{R}^n,\gamma)$; i.e.,

$$\left\| f - \sum_{|\alpha|\le k} \widehat{f}(\alpha)h_\alpha \right\|_2 \to 0 \quad \text{as } k\to\infty.$$

This is called the *Hermite expansion* of f.

Remark 11.35. If $f : \mathbb{R}^n \to \mathbb{R}$ is a multilinear polynomial, then it "is its own Hermite expansion":

$$f(z) = \sum_{S \subseteq [n]} \widehat{f}(S) z^S = \sum_{S \subseteq [n]} \widehat{f}(S) h_S(z) = \sum_{\alpha_1, \dots, \alpha_n \leq 1} \widehat{f}(\alpha) h_\alpha(z).$$

Proposition 11.36. *The Hermite coefficients of $f \in L^2(\mathbb{R}^n, \gamma)$ satisfy the formula*

$$\widehat{f}(\alpha) = \langle f, h_\alpha \rangle,$$

and for $f, g \in L^2(\mathbb{R}^n, \gamma)$ we have the Plancherel formula

$$\langle f, g \rangle = \sum_{\alpha \in \mathbb{N}^n} \widehat{f}(\alpha) \widehat{g}(\alpha).$$

From this we may deduce:

Proposition 11.37. *For $f \in L^2(\mathbb{R}^n, \gamma)$, the function $\mathrm{U}_\rho f$ has Hermite expansion*

$$\mathrm{U}_\rho f = \sum_{\alpha \in \mathbb{N}^n} \rho^{|\alpha|} \widehat{f}(\alpha) h_\alpha$$

and hence

$$\mathbf{Stab}_\rho[f] = \sum_{\alpha \in \mathbb{N}^n} \rho^{|\alpha|} \widehat{f}(\alpha)^2.$$

Proof. Both statements follow from Proposition 11.36, with the first using

$$\widehat{\mathrm{U}_\rho f}(\alpha) = \langle \mathrm{U}_\rho f, h_\alpha \rangle = \langle \sum_\beta \mathrm{U}_\rho \widehat{f}(\beta) h_\beta, h_\alpha \rangle = \sum_\beta \widehat{f}(\beta) \langle \mathrm{U}_\rho h_\beta, h_\alpha \rangle = \rho^{|\alpha|} \widehat{f}(\alpha);$$

we also used Proposition 11.33 and the fact that U_ρ is a contraction in $L^2(\mathbb{R}^n, \gamma)$. □

Remark 11.38. When $f : \mathbb{R}^n \to \mathbb{R}$ is a multilinear polynomial, this formula for $\mathrm{U}_\rho f$ agrees with the formula $f(\rho z)$ given in Fact 11.13.

Remark 11.39. In a sense it's not very important to know the explicit formulas for the Hermite polynomials, (11.5), (11.8); it's usually enough just to know that the formula for $\mathrm{U}_\rho f$ from Proposition 11.37 holds.

Finally, by differentiating the formula in Proposition 11.37 at $\rho = 1$ we deduce the following formula for the Ornstein–Uhlenbeck operator (explaining why it's sometimes called the number operator):

Proposition 11.40. *For $f \in L^2(\mathbb{R}^n, \gamma)$ in the domain of L we have*

$$\mathrm{L} f = \sum_{\alpha \in \mathbb{N}^n} |\alpha| \widehat{f}(\alpha) h_\alpha.$$

(Actually, Exercise 11.18 asks you to formally justify this and the fact that f is in the domain of L if and only if $\sum_\alpha |\alpha|^2 \widehat{f}(\alpha)^2 < \infty$.) For additional facts about Hermite polynomials, see Exercises 11.9–11.14.

11.3. Borell's Isoperimetric Theorem

If we believe that the Majority Is Stablest Theorem should be true, then we also have to believe in its "Gaussian special case". Let's see what this Gaussian special case is. Suppose $f : \mathbb{R}^n \to [-1,1]$ is a "nice" function (smooth, say, with all derivatives bounded) having $\mathbf{E}[f] = 0$. You're encouraged to think of f as (a smooth approximation to) the indicator $\pm 1_A$ of some set $A \subseteq \mathbb{R}^n$ of Gaussian volume $\mathrm{vol}_\gamma(A) = \frac{1}{2}$. Now consider the Boolean function $g : \{-1,1\}^{nM} \to \{-1,1\}$ defined by

$$g = f \circ \mathrm{BitsToGaussians}_M^n.$$

Using the multidimensional Central Limit Theorem, for any $\rho \in (0,1)$ we should have

$$\mathbf{Stab}_\rho[g] \xrightarrow{M\to\infty} \mathbf{Stab}_\rho[f],$$

where on the left we have Boolean noise stability and on the right we have Gaussian noise stability. Using $\mathbf{E}[g] \to \mathbf{E}[f] = 0$, the Majority Is Stablest Theorem would tell us that

$$\mathbf{Stab}_\rho[g] \le 1 - \tfrac{2}{\pi} \arccos \rho + o_\epsilon(1),$$

where $\epsilon = \mathbf{MaxInf}[g]$. But $\epsilon = \epsilon(M) \to 0$ as $M \to \infty$. Thus we should simply have the Gaussian noise stability bound

$$\mathbf{Stab}_\rho[f] \le 1 - \tfrac{2}{\pi} \arccos \rho. \tag{11.11}$$

(By a standard approximation argument this extends from "nice" $f : \mathbb{R}^n \to [-1,1]$ with $\mathbf{E}[f] = 0$ to any measurable $f : \mathbb{R}^n \to [-1,1]$ with $\mathbf{E}[f] = 0$.) Note that the upper bound (11.11) is achieved when f is the ± 1-indicator of any halfspace through the origin; see Corollary 11.20. (Note also that if $n = 1$ and $f = \mathrm{sgn}$, then the function g is simply Maj_M.)

The "isoperimetric inequality" (11.11) is indeed true, and is a special case of a theorem first proved by Borell [Bor85].

Borell's Isoperimetric Theorem (volume-$\frac{1}{2}$ case). *Fix $\rho \in (0,1)$. Then for any $f \in L^2(\mathbb{R}^n, \gamma)$ with range $[-1,1]$ and $\mathbf{E}[f] = 0$,*

$$\mathbf{Stab}_\rho[f] \le 1 - \tfrac{2}{\pi} \arccos \rho,$$

with equality if f is the ± 1-indicator of any halfspace through the origin.

Remark 11.41. In Borell's Isoperimetric Theorem, nothing is lost by restricting attention to functions with range $\{-1,1\}$, i.e., by considering only $f = \pm 1_A$ for $A \subseteq \mathbb{R}^n$. This is because the case of range $[-1,1]$ follows straightforwardly from the case of range $\{-1,1\}$, essentially because $\sqrt{\mathbf{Stab}_\rho[f]} = \|\mathrm{U}_{\sqrt{\rho}} f\|_2$ is a convex functional of f; see Exercise 11.25.

More generally, Borell showed that for any fixed volume $\alpha \in [0,1]$, the maximum Gaussian noise stability of a set of volume α is no greater than that of a halfspace of volume α. We state here the more general theorem, using range $\{0,1\}$ rather than range $\{-1,1\}$ for future notational convenience (and with Remark 11.41 applying equally):

Borell's Isoperimetric Theorem. *Fix $\rho \in (0,1)$. Then for any $f \in L^2(\mathbb{R}^n,\gamma)$ with range $[0,1]$ and $\mathbf{E}[f] = \alpha$,*

$$\mathbf{Stab}_\rho[f] \le \Lambda_\rho(\alpha).$$

Here $\Lambda_\rho(\alpha)$ is the Gaussian quadrant probability *function, discussed in Exercises 5.32 and 11.19, and equal to $\mathbf{Stab}_\rho[1_H]$ for any (every) halfspace $H \subseteq \mathbb{R}^n$ having Gaussian volume $\mathrm{vol}_\gamma(H) = \alpha$.*

We've seen that the volume-$\frac{1}{2}$ case of Borell's Isoperimetric Theorem is a special case of the Majority Is Stablest Theorem, and similarly, the general version of Borell's theorem is a special case of the General-Volume Majority Is Stablest Theorem mentioned at the beginning of the chapter. As a consequence, proving Borell's Isoperimetric Theorem is a *prerequisite* for proving the General-Volume Majority Is Stablest Theorem. In fact, our proof in Section 11.7 of the latter will be a reduction to the former.

The proof of Borell's Isoperimetric Theorem itself is not too hard; one of five known proofs, the one due to Mossel and Neeman [**MN12**], is outlined in Exercises 11.26–11.29. If our main goal is just to prove the basic Majority Is Stablest Theorem, then we only need the volume-$\frac{1}{2}$ case of Borell's Isoperimetric Inequality. Luckily, there's a very simple proof of this volume-$\frac{1}{2}$ case for "many" values of ρ, as we will now explain.

Let's first slightly rephrase the statement of Borell's Isoperimetric Theorem in the volume-$\frac{1}{2}$ case. By Remark 11.41 we can restrict attention to sets; then the theorem asserts that among sets of Gaussian volume $\frac{1}{2}$, halfspaces through the origin have maximal noise stability, for each positive value of ρ. Equivalently, halfspaces through the origin have minimal noise *sensitivity* under correlation $\cos\theta$, for $\theta \in (0,\frac{\pi}{2})$. The formula for this minimal noise sensitivity was given as (11.2) in our proof of Sheppard's Formula. Thus we have:

Equivalent statement of the volume-$\frac{1}{2}$ Borell Isoperimetric Theorem. *Fix $\theta \in (0,\frac{\pi}{2})$. Then for any $A \subset \mathbb{R}^n$ with $\mathrm{vol}_\gamma(A) = \frac{1}{2}$,*

$$\mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \cos\theta\text{-correlated}}}[1_A(\boldsymbol{z}) \neq 1_A(\boldsymbol{z}')] \ge \frac{\theta}{\pi},$$

with equality if A is any halfspace through the origin.

In the remainder of this section we'll show how to prove this formulation of the theorem whenever $\theta = \frac{\pi}{2\ell}$, where ℓ is a positive integer. This gives the volume-$\frac{1}{2}$ case of Borell's Isoperimetric Inequality for all ρ of the form $\arccos \frac{\pi}{2\ell}$, $\ell \in \mathbb{N}^+$; in particular, for an infinite sequence of ρ's tending to 1. To prove the theorem for these values of θ, it's convenient to introduce notation for the following noise sensitivity variant:

Definition 11.42. For $A \subseteq \mathbb{R}^n$ and $\delta \in \mathbb{R}$ (usually $\delta \in [0,\pi]$) we write $\mathbf{RS}_A(\delta)$ for the *rotation sensitivity of A at δ*, defined by

$$\mathbf{RS}_A(\delta) = \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}') \\ \cos\delta\text{-correlated}}}[1_A(\boldsymbol{z}) \neq 1_A(\boldsymbol{z}')].$$

The key property of this definition is the following:

Theorem 11.43. *For any $A \subseteq \mathbb{R}^n$ the function $\mathbf{RS}_A(\delta)$ is subadditive; i.e.,*

$$\mathbf{RS}_A(\delta_1 + \cdots + \delta_\ell) \leq \mathbf{RS}_A(\delta_1) + \cdots + \mathbf{RS}_A(\delta_\ell).$$

In particular, for any $\delta \in \mathbb{R}$ and $\ell \in \mathbb{N}^+$,

$$\mathbf{RS}_A(\delta) \leq \ell \cdot \mathbf{RS}_A(\delta/\ell).$$

Proof. Let $\boldsymbol{g}, \boldsymbol{g}' \sim \mathrm{N}(0,1)^n$ be drawn independently and define $\boldsymbol{z}(\theta) = (\cos\theta)\boldsymbol{g} + (\sin\theta)\boldsymbol{g}'$. Geometrically, as θ goes from 0 to $\frac{\pi}{2}$ the random vectors $\boldsymbol{z}(\theta)$ trace from $\boldsymbol{g}$ to $\boldsymbol{g}'$ along the origin-centered ellipse passing through these two points. The random vectors $\boldsymbol{z}(\theta)$ are jointly normal, with each individually distributed as $\mathrm{N}(0,1)^n$. Further, for each fixed $\theta, \theta' \in \mathbb{R}$ the pair $(\boldsymbol{z}(\theta), \boldsymbol{z}(\theta'))$ constitute ρ-correlated Gaussians with

$$\rho = \cos\theta\cos\theta' + \sin\theta\sin\theta' = \cos(\theta' - \theta).$$

Now consider the sequence $\theta_0, \dots, \theta_\ell$ defined by the partial sums of the δ_i's, i.e., $\theta_j = \sum_{i=1}^{j} \delta_i$. We get that $\boldsymbol{z}(\theta_0)$ and $\boldsymbol{z}(\theta_\ell)$ are $\cos(\delta_1 + \cdots + \delta_\ell)$-correlated, and that $\boldsymbol{z}(\theta_{j-1})$ and $\boldsymbol{z}(\theta_j)$ are $\cos\delta_j$-correlated for each $j \in [\ell]$. Thus

$$\begin{aligned}\mathbf{RS}_A(\delta_1 + \cdots + \delta_\ell) &= \mathbf{Pr}[1_A(\boldsymbol{z}(\theta_0)) \neq 1_A(\boldsymbol{z}(\theta_\ell))] \\ &\leq \sum_{j=1}^{\ell} \mathbf{Pr}[1_A(\boldsymbol{z}(\theta_j)) \neq 1_A(\boldsymbol{z}(\theta_{j-1}))] = \sum_{j=1}^{\ell} \mathbf{RS}_A(\delta_j), \qquad (11.12)\end{aligned}$$

where the inequality is the union bound. □

With this subadditivity result in hand, it's indeed easy to prove the equivalent statement of the volume-$\frac{1}{2}$ Borell Isoperimetric Theorem for any $\theta \in \{\frac{\pi}{4}, \frac{\pi}{6}, \frac{\pi}{8}, \frac{\pi}{10}, \dots\}$. As we'll see in Section 11.7, the case of $\theta = \frac{\pi}{4}$ can be used to give an excellent UG-hardness result for the Max-Cut CSP.

Corollary 11.44. *The equivalent statement of the volume-$\frac{1}{2}$ Borell Isoperimetric Theorem holds whenever $\theta = \frac{\pi}{2\ell}$ for $\ell \in \mathbb{N}^+$.*

Proof. The exact statement we need to show is $\mathbf{RS}_A(\frac{\pi}{2\ell}) \geq \frac{1}{2\ell}$. This follows by taking $\delta = \frac{\pi}{2}$ in Theorem 11.43 because

$$\mathbf{RS}_A(\tfrac{\pi}{2}) = \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\\ \text{0-correlated}}}[1_A(\boldsymbol{z}) \neq 1_A(\boldsymbol{z}')] = \tfrac{1}{2},$$

using that 0-correlated Gaussians are independent and that $\mathrm{vol}_\gamma(A) = \frac{1}{2}$. □

Remark 11.45. Although Sheppard's Formula already tells us that equality holds in this corollary when A is a halfspace through the origin, it's also not hard to derive this directly from the proof. The only inequality in the proof, (11.12), is an equality when A is a halfspace through the origin, because the elliptical arc can only cross such a halfspace 0 or 1 times.

Remark 11.46. Suppose that $A \subseteq \mathbb{R}^n$ not only has volume $\frac{1}{2}$, it has the property that $x \in A$ if and only if $-x \notin A$; in other words, the ±1-indicator of A is an odd function. (In both statements, we allow a set of measure 0 to be ignored.) An example set with this property is any halfspace through the origin. Then $\mathbf{RS}_A(\pi) = 1$, and hence we can establish Corollary 11.44 more generally for any $\theta \in \{\frac{\pi}{1}, \frac{\pi}{2}, \frac{\pi}{3}, \frac{\pi}{4}, \frac{\pi}{5}, \dots\}$ by taking $\delta = \pi$ in the proof.

11.4. Gaussian surface area and Bobkov's Inequality

This section is devoted to studying the *Gaussian Isoperimetric Inequality*. This inequality is a special case of the Borell Isoperimetric Inequality (and hence also a special case of the General-Volume Majority Is Stablest Theorem); in particular, it's the special case arising from the limit $\rho \to 1^-$.

Restating Borell's theorem using rotation sensitivity we have that for any $A \subseteq \mathbb{R}^n$, if $H \subseteq \mathbb{R}^n$ is a halfspace with the same Gaussian volume as A then for all ϵ,

$$\mathbf{RS}_A(\epsilon) \geq \mathbf{RS}_H(\epsilon).$$

Since $\mathbf{RS}_A(0) = \mathbf{RS}_H(0) = 0$, it follows that

$$\mathbf{RS}'_A(0^+) \geq \mathbf{RS}'_H(0^+).$$

(Here we are considering the one-sided derivatives at 0, which can be shown to exist, though $\mathbf{RS}'_A(0^+)$ may equal $+\infty$; see the notes at the end of this chapter.) As will be explained shortly, $\mathbf{RS}'_A(0^+)$ is precisely $\sqrt{2/\pi} \cdot \mathrm{surf}_\gamma(A)$, where $\mathrm{surf}_\gamma(A)$ denotes the "Gaussian surface area" of A. Therefore the above inequality is equivalent to the following:

Gaussian Isoperimetric Inequality. *Let* $A \subseteq \mathbb{R}^n$ *have* $\mathrm{vol}_\gamma(A) = \alpha$ *and let* $H \subseteq \mathbb{R}^n$ *be any halfspace with* $\mathrm{vol}_\gamma(H) = \alpha$. *Then* $\mathrm{surf}_\gamma(A) \geq \mathrm{surf}_\gamma(H)$.

Remark 11.47. As shown in Proposition 11.49 below, the right-hand side in this inequality is equal to $\mathcal{U}(\alpha)$, where $\mathcal{U}$ is the *Gaussian isoperimetric function*, encountered earlier in Definition 5.26 and defined by $\mathcal{U} = \varphi \circ \Phi^{-1}$.

Let's now discuss the somewhat technical question of how to properly define $\mathrm{surf}_\gamma(A)$, the Gaussian surface area of a set A. Perhaps the most natural definition would be to equate it with the *Gaussian Minkowski content* of the boundary ∂A of A,

$$\gamma^+(\partial A) = \liminf_{\epsilon\to 0^+} \frac{\mathrm{vol}_\gamma(\{z : \mathrm{dist}(z, \partial A) < \epsilon/2\})}{\epsilon}. \tag{11.13}$$

(Relatedly, one might also consider the surface integral over ∂A of the Gaussian pdf φ.) Under the "official" definition of $\mathrm{surf}_\gamma(A)$ we give below in Definition 11.48, we'll indeed have $\mathrm{surf}_\gamma(A) = \gamma^+(\partial A)$ whenever A is sufficiently nice – say, a disjoint union of closed, full-dimensional, convex sets. However, the Minkowski content definition is not a good one in general because it's possible to have $\gamma^+(\partial A_1) \neq \gamma^+(\partial A_2)$ for some sets A_1 and A_2 that are equivalent up to measure 0. (For more information, see Exercise 11.15 and the notes at the end of this chapter.)

As mentioned above, one "correct" definition is $\mathrm{surf}_\gamma(A) = \sqrt{\pi/2}\cdot \mathbf{RS}'_A(0^+)$. This definition has the advantage of being insensitive to measure-0 changes to A. To connect this unusual-looking definition with Minkowski content, let's heuristically interpret $\mathbf{RS}'_A(0^+)$. We start by thinking of it as $\frac{\mathbf{RS}_A(\epsilon)}{\epsilon}$ for "infinitesimal ϵ". Now $\mathbf{RS}_A(\epsilon)$ can be thought of as the probability that the line segment $\boldsymbol{\ell}$ joining two $\cos\epsilon$-correlated Gaussians crosses ∂A. Since $\sin\epsilon \approx \epsilon$, $\cos\epsilon \approx 1$ up to $O(\epsilon^2)$, we can think of these correlated Gaussians as $\boldsymbol{g}$ and $\boldsymbol{g} + \epsilon\boldsymbol{g}'$ for independent $\boldsymbol{g}, \boldsymbol{g}' \sim \mathrm{N}(0,1)^n$. When $\boldsymbol{g}$ lands near ∂A, the length of $\boldsymbol{\ell}$ in the direction perpendicular to ∂A will, in expectation, be $\epsilon\,\mathbf{E}[|\mathrm{N}(0,1)|] = \sqrt{2/\pi}\epsilon$. Thus $\mathbf{RS}_A(\epsilon)$ should essentially be $\sqrt{2/\pi}\epsilon\cdot \mathrm{vol}_\gamma(\{z : \mathrm{dist}(z,\partial A) < \epsilon/2\})$ and we have heuristically justified

$$\sqrt{\pi/2}\cdot \mathbf{RS}'_A(0^+) = \sqrt{\pi/2}\cdot \lim_{\epsilon\to 0^+} \frac{\mathbf{RS}_A(\epsilon)}{\epsilon} \overset{?}{=} \gamma^+(\partial A). \tag{11.14}$$

One more standard idea for the definition of $\mathrm{surf}_\gamma(A)$ is "$\mathbf{E}[\|\nabla 1_A\|]$". This doesn't quite make sense since $1_A \in L^1(\mathbb{R}^n, \gamma)$ is not actually differentiable. However, we might consider replacing it with the limit of $\mathbf{E}[\|\nabla f_m\|]$ for a sequence (f_m) of smooth functions approximating 1_A. To see why this notion should agree with the Gaussian Minkowski content $\gamma^+(\partial A)$ for nice enough A, let's suppose we have a smooth approximator f to 1_A that agrees with 1_A on $\{z : \mathrm{dist}(z, \partial A) \geq \epsilon/2\}$ and is (essentially) a linear function on $\{z : \mathrm{dist}(z, \partial A) < \epsilon/2\}$. Then $\|\nabla f\|$ will be 0 on the former set and (essentially) constantly $1/\epsilon$ on the latter (since it must climb from 0 to 1 over a distance of ϵ). Thus we indeed have

$$\mathbf{E}[\|\nabla f\|] \approx \frac{\mathrm{vol}_\gamma(\{z : \mathrm{dist}(z, \partial A) < \epsilon/2\})}{\epsilon} \approx \gamma^+(\partial A),$$

as desired. We summarize the above technical discussion with the following definition/theorem, which is discussed further in the notes at the end of this chapter:

Definition 11.48. For any $A \subseteq \mathbb{R}^n$, we define its *Gaussian surface area* to be

$$\mathrm{surf}_\gamma(A) = \sqrt{\pi/2} \cdot \mathbf{RS}'_A(0^+) \in [0, \infty].$$

An equivalent definition is

$$\mathrm{surf}_\gamma(A) = \inf \left\{ \liminf_{m \to \infty} \mathop{\mathbf{E}}_{\boldsymbol{z} \sim \mathrm{N}(0,1)^n} [\|\nabla f_m(\boldsymbol{z})\|] \right\},$$

where the infimum is over all sequences $(f_m)_{m \in \mathbb{N}}$ of smooth $f_m : \mathbb{R}^n \to [0,1]$ with first partial derivatives in $L^2(\mathbb{R}^n, \gamma)$ such that $\|f_m - 1_A\|_1 \to 0$. Furthermore, this infimum is actually achieved by taking $f_m = \mathrm{U}_{\rho_m} f$ for any sequence $\rho_m \to 1^-$. Finally, the equality $\mathrm{surf}_\gamma(A) = \gamma^+(\partial A)$ with Gaussian Minkowski content holds if A is a disjoint union of closed, full-dimensional, convex sets.

To get further acquainted with this definition, let's describe the Gaussian surface area of some basic sets. We start with halfspaces, which as mentioned in Remark 11.47 have Gaussian surface area given by the Gaussian isoperimetric function.

Proposition 11.49. *Let $H \subseteq \mathbb{R}^n$ be any halfspace (open or closed) with $\mathrm{vol}_\gamma(H) = \alpha \in (0,1)$. Then $\mathrm{surf}_\gamma(H) = \mathcal{U}(\alpha) = \varphi(\Phi^{-1}(\alpha))$. In particular, if $\alpha = 1/2$ – i.e., H's boundary contains the origin – then $\mathrm{surf}_\gamma(H) = \frac{1}{\sqrt{2\pi}}$.*

Proof. Just as in the proof of Corollary 11.20, by rotational symmetry we may assume H is a 1-dimensional halfline, $H = (-\infty, t]$. Since $\mathrm{vol}_\gamma(H) = \alpha$, we have $t = \Phi^{-1}(\alpha)$. Then $\mathrm{surf}_\gamma(H)$ is equal to

$$\gamma^+(\partial H) = \lim_{\epsilon \to 0^+} \frac{\mathrm{vol}_\gamma(\{z \in \mathbb{R} : \mathrm{dist}(z, \partial H) < \frac{\epsilon}{2}\})}{\epsilon} = \lim_{\epsilon \to 0^+} \frac{\int_{t-\epsilon/2}^{t+\epsilon/2} \varphi(s)\,ds}{\epsilon} = \varphi(t) = \mathcal{U}(\alpha).$$

□

Here are some more Gaussian surface area bounds:

Example 11.50. In Exercise 11.16 you are asked to generalize the above computation and show that if $A \subseteq \mathbb{R}$ is the union of disjoint nondegenerate intervals $[t_1, t_2], [t_3, t_4], \dots, [t_{2m-1}, t_{2m}]$ then $\mathrm{surf}_\gamma(A) = \sum_{i=1}^{2m} \varphi(t_i)$. Perhaps the next easiest example is when $A \subseteq \mathbb{R}^n$ is an origin-centered ball; Ball [**Bal93**] gave an explicit formula for $\mathrm{surf}_\gamma(A)$ in terms of the dimension and radius, one which is always less than $\sqrt{\frac{2}{\pi}}$ (see Exercise 11.17). This upper bound was extended to non-origin-centered balls in Klivans et al. [**KOS08**]. Ball also showed that every convex set $A \subseteq \mathbb{R}^n$ satisfies $\mathrm{surf}_\gamma(A) \leq O(n^{1/4})$; Nazarov [**Naz03**] showed that this bound is tight up to the constant, using a construction highly reminiscent of Talagrand's Exercise 4.18. As noted in

Klivans et al. [KOS08], Nazarov's work also immediately implies that an intersection of k halfspaces has Gaussian surface area at most $O(\sqrt{\log k})$ (tight for appropriately sized cubes in $\mathbb{R}^k$), and that any cone in $\mathbb{R}^n$ with apex at the origin has Gaussian surface area at most 1. Finally, by proving the "Gaussian special case" of the Gotsman–Linial Conjecture, Kane [Kan11] established that if $A \subseteq \mathbb{R}^n$ is a degree-k "polynomial threshold function" – i.e., $A = \{z : p(z) > 0\}$ for p an n-variate degree-k polynomial – then $\mathrm{surf}_\gamma(A) \leq \frac{k}{\sqrt{2\pi}}$. This is tight for every k (even when $n = 1$).

Though we've shown that the Gaussian Isoperimetric Inequality follows from Borell's Isoperimetric Theorem, we now discuss some alternative proofs. In the special case of sets of Gaussian volume $\frac{1}{2}$, we can again get a very simple proof using the subadditivity property of Gaussian rotation sensitivity, Theorem 11.43. That result easily yields the following kind of "concavity property" concerning Gaussian surface area:

Theorem 11.51. *Let $A \subseteq \mathbb{R}^n$. Then for any $\delta > 0$,*

$$\sqrt{\pi/2} \cdot \frac{\mathbf{RS}_A(\delta)}{\delta} \leq \mathrm{surf}_\gamma(A).$$

Proof. For $\delta > 0$ and $\epsilon = \delta/\ell$, $\ell \in \mathbb{N}^+$, Theorem 11.43 is equivalent to

$$\frac{\mathbf{RS}_A(\delta)}{\delta} \leq \frac{\mathbf{RS}_A(\epsilon)}{\epsilon}.$$

Taking $\ell \to \infty$ hence $\epsilon \to 0^+$, the right-hand side becomes $\mathbf{RS}'_A(0^+) = \sqrt{2/\pi} \cdot \mathrm{surf}_\gamma(A)$. □

If we take $\delta = \pi/2$ in this theorem, the left-hand side becomes

$$\sqrt{2/\pi} \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{z},\boldsymbol{z}' \sim \mathrm{N}(0,1)^n \\ \text{independent}}} [1_A(\boldsymbol{z}) \neq 1_A(\boldsymbol{z}')] = 2\sqrt{2/\pi} \cdot \mathrm{vol}_\gamma(A)(1 - \mathrm{vol}_\gamma(A)).$$

Thus we obtain a simple proof of the following result, which includes the Gaussian Isoperimetric Inequality in the volume-$\frac{1}{2}$ case:

Theorem 11.52. *Let $A \subseteq \mathbb{R}^n$. Then*

$$2\sqrt{2/\pi} \cdot \mathrm{vol}_\gamma(A)(1 - \mathrm{vol}_\gamma(A)) \leq \mathrm{surf}_\gamma(A).$$

In particular, if $\mathrm{vol}_\gamma(A) = \frac{1}{2}$, then we get the tight Gaussian Isoperimetric Inequality statement $\mathrm{surf}_\gamma(A) \geq \frac{1}{\sqrt{2\pi}} = \mathcal{U}(\frac{1}{2})$.

As for the full Gaussian Isoperimetric Inequality, it's a pleasing fact that it can be derived by pure analysis of Boolean functions. This was shown by Bobkov [Bob97], who proved the following very interesting isoperimetric inequality about Boolean functions:

Bobkov's Inequality. *Let $f : \{-1,1\}^n \to [0,1]$. Then*

$$\mathcal{U}(\mathbf{E}[f]) \leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\|(\mathcal{U}(f(\boldsymbol{x})), \nabla f(\boldsymbol{x}))\|]. \tag{11.15}$$

Here ∇f is the discrete gradient (as in Definition 2.34) and $\|\cdot\|$ is the usual Euclidean norm (in $\mathbb{R}^{n+1}$). Thus to restate the inequality,

$$\mathcal{U}(\mathbf{E}[f]) \leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[\sqrt{\mathcal{U}(f(\boldsymbol{x}))^2 + \sum_{i=1}^{n} \mathrm{D}_i f(\boldsymbol{x})^2}\right].$$

In particular, suppose $f = 1_A$ is the 0-1 indicator of a subset $A \subseteq \{-1,1\}^n$. Then since $\mathcal{U}(0) = \mathcal{U}(1) = 0$ we obtain $\mathcal{U}(\mathbf{E}[1_A]) \leq \mathbf{E}[\|\nabla 1_A\|]$.

As Bobkov noted, by the usual Central Limit Theorem argument one can straightforwardly obtain inequality (11.15) in the setting of functions $f \in L^2(\mathbb{R}^n, \gamma)$ with range $[0,1]$, provided f is sufficiently smooth (for example, if f is in the domain of L; see Exercise 11.18). Then given $A \subseteq \mathbb{R}^n$, by taking a sequence of smooth approximations to 1_A as in Definition 11.48, the Gaussian Isoperimetric Inequality $\mathcal{U}(\mathbf{E}[1_A]) \leq \mathrm{surf}_\gamma(A)$ is recovered.

Given $A \subseteq \{-1,1\}^n$ we can write the quantity $\mathbf{E}[\|\nabla 1_A\|]$ appearing in Bobkov's Inequality as

$$\mathbf{E}[\|\nabla 1_A\|] = \tfrac{1}{2} \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[\sqrt{\mathrm{sens}_A(\boldsymbol{x})}\right], \tag{11.16}$$

using the fact that for $1_A : \{-1,1\}^n \to \{0,1\}$ we have

$$\mathrm{D}_i 1_A(\boldsymbol{x})^2 = \tfrac{1}{4} \cdot \mathbf{1}[\text{coordinate } i \text{ is pivotal for } 1_A \text{ on } x].$$

The quantity in (11.16) – (half of) the expected square-root of the number of pivotal coordinates – is an interesting possible notion of "Boolean surface area" for sets $A \subseteq \{-1,1\}^n$. It was first essentially proposed by Talagrand [**Tal93**]. By Cauchy–Schwarz it's upper-bounded by (half of) the square-root of our usual notion of boundary size, average sensitivity:

$$\mathbf{E}[\|\nabla 1_A\|] \leq \sqrt{\mathbf{E}[\|\nabla 1_A\|^2]} = \sqrt{\mathbf{I}[1_A]}. \tag{11.17}$$

(Note that $\mathbf{I}[1_A]$ here is actually one quarter of the average sensitivity of A, because we're using 0-1 indicators as opposed to ±1). But the inequality in (11.17) is often far from sharp. For example, while the majority function has average sensitivity $\Theta(\sqrt{n})$, the expected square-root of its sensitivity is $\Theta(1)$ because a $\Theta(1/\sqrt{n})$-fraction of strings have sensitivity $\lceil n/2 \rceil$ and the remainder have sensitivity 0.

Let's turn to the proof of Bobkov's Inequality. As you are asked to show in Exercise 11.20, the general-n case of Bobkov's Inequality follows from the $n = 1$ case by a straightforward "induction by restrictions". Thus just as in the proof of the Hypercontractivity Theorem, it suffices to prove the $n = 1$ "two-point inequality", an elementary inequality about two real numbers:

Bobkov's Two-Point Inequality. *Let $f:\{-1,1\}\to[0,1]$. Then*

$$\mathcal{U}(\mathbf{E}[f])\le\mathbf{E}[\|(\mathcal{U}(f),\nabla f)\|].$$

Writing $f(x)=a+bx$, this is equivalent to saying that provided $a\pm b\in[0,1]$,

$$\mathcal{U}(a)\le\tfrac{1}{2}\|(\mathcal{U}(a+b),b)\|+\tfrac{1}{2}\|(\mathcal{U}(a-b),b)\|.$$

Remark 11.53. The only property of $\mathcal{U}$ used in proving this inequality is that it satisfies (Exercise 5.43) the differential equation $\mathcal{U}\mathcal{U}''=-1$ on $(0,1)$.

Bobkov's proof of the two-point inequality was elementary but somewhat long and hard to motivate. In contrast, Barthe and Maurey [BM00] gave a fairly short proof of the inequality, but it used methods from stochastic calculus, namely Itô's Formula. We present here an elementary discretization of the Barthe–Maurey proof.

Proof of Bobkov's Two-Point Inequality. By symmetry and continuity we may assume $\delta\le a-b<a+b\le 1-\delta$ for some $\delta>0$. Let $\tau=\tau(\delta)>0$ be a small quantity to be chosen later such that b/τ is an integer. Let $\boldsymbol{y}_0,\boldsymbol{y}_1,\boldsymbol{y}_2,\dots$ be a random walk within $[a-b,a+b]$ that starts at $\boldsymbol{y}_0=a$, takes independent equally likely steps of $\pm\tau$, and is absorbed at the endpoints $a\pm b$. Finally, for $t\in\mathbb{N}$, define $\boldsymbol{z}_t=\|(\mathcal{U}(\boldsymbol{y}_t),\tau\sqrt{t})\|$. The key claim for the proof is:

Claim 11.54. *Assuming $\tau=\tau(\delta)>0$ is small enough, $(\boldsymbol{z}_t)_t$ is a submartingale with respect to $(\boldsymbol{y}_t)_t$, i.e., $\mathbf{E}[\boldsymbol{z}_{t+1}\mid\boldsymbol{y}_0,\dots,\boldsymbol{y}_t]=\mathbf{E}[\boldsymbol{z}_{t+1}\mid\boldsymbol{y}_t]\ge\boldsymbol{z}_t$.*

Let's complete the proof given the claim. Let $\boldsymbol{T}$ be the stopping time at which $\boldsymbol{y}_t$ first reaches $a\pm b$. By the Optional Stopping Theorem we have $\mathbf{E}[\boldsymbol{z}_0]\le\mathbf{E}[\boldsymbol{z}_{\boldsymbol{T}}]$; i.e.,

$$\mathcal{U}(a)\le\mathbf{E}[\|(\mathcal{U}(\boldsymbol{z}_{\boldsymbol{T}}),\tau\sqrt{\boldsymbol{T}})\|].\tag{11.18}$$

In the expectation above we can condition on whether the walk stopped at $a+b$ or $a-b$. By symmetry, both events occur with probability 1/2 and neither changes the conditional distribution of $\boldsymbol{T}$. Thus we get

$$\begin{aligned}\mathcal{U}(a)&\le\tfrac{1}{2}\mathbf{E}[\|(\mathcal{U}(a+b),\tau\sqrt{\boldsymbol{T}})\|]+\tfrac{1}{2}\mathbf{E}[\|(\mathcal{U}(a-b),\tau\sqrt{\boldsymbol{T}})\|]\\&\le\tfrac{1}{2}\|(\mathcal{U}(a+b),\sqrt{\mathbf{E}[\tau^2\boldsymbol{T}]})\|+\tfrac{1}{2}\|(\mathcal{U}(a-b),\sqrt{\mathbf{E}[\tau^2\boldsymbol{T}]})\|,\end{aligned}$$

with the second inequality using concavity of $v\mapsto\sqrt{u^2+v}$. But it's a well-known fact (following immediately from Exercise 11.22) that $\mathbf{E}[\boldsymbol{T}]=(b/\tau)^2$. Substituting this into the above completes the proof.

It remains to verify Claim 11.54. Actually, although the claim is true as stated (see Exercise 11.23) it will be more natural to prove the following slightly weaker claim:

$$\mathbf{E}[\boldsymbol{z}_{t+1}\mid\boldsymbol{y}_t]\ge\boldsymbol{z}_t-C_\delta\tau^3\tag{11.19}$$

for some constant C_δ depending only on δ. This is still enough to complete the proof: Applying the Optional Stopping Theorem to the submartingale $(\boldsymbol{z}_t + C_\delta \tau^3 t)_t$ we get that (11.18) holds up to an additive $C_\delta \tau^3 \mathbf{E}[\boldsymbol{T}] = C_\delta b^2 \tau$. Then continuing with the above we deduce Bobkov's Inequality up to $C_\delta b^2 \tau$, and we can make τ arbitrarily small.

Even though we only need to prove (11.19), let's begin a proof of the original Claim 11.54 anyway. Fix $t \in \mathbb{N}^+$ and condition on $\boldsymbol{y}_t = y$. If y is $a \pm b$, then the walk is stopped and the claim is clear. Otherwise, $\boldsymbol{y}_{t+1}$ is $y \pm \tau$ with equal probability, and we want to verify the following inequality (assuming $\tau > 0$ is sufficiently small as a function of δ, independent of y):

$$\begin{aligned}\|(\mathscr{U}(y), \tau\sqrt{t})\| &\leq \tfrac{1}{2}\|(\mathscr{U}(y+\tau), \tau\sqrt{t+1})\| + \tfrac{1}{2}\|(\mathscr{U}(y-\tau), \tau\sqrt{t+1})\| \qquad (11.20)\\ &= \tfrac{1}{2}\left\|\left(\sqrt{\mathscr{U}(y+\tau)^2+\tau^2}, \tau\sqrt{t}\right)\right\| + \tfrac{1}{2}\left\|\left(\sqrt{\mathscr{U}(y-\tau)^2+\tau^2}, \tau\sqrt{t}\right)\right\|.\end{aligned}$$

By the triangle inequality, it's sufficient to show

$$\mathscr{U}(y) \leq \tfrac{1}{2}\sqrt{\mathscr{U}(y+\tau)^2+\tau^2} + \tfrac{1}{2}\sqrt{\mathscr{U}(y-\tau)^2+\tau^2},$$

and this is actually necessary too, being the $t = 0$ case of (11.20). (In fact, this is identical to Bobkov's Two-Point Inequality itself, except now we may assume τ is sufficiently small.) Finally, since we actually only need the weakened submartingale statement (11.19), we'll instead establish

$$\mathscr{U}(y) - C_\delta \tau^3 \leq \tfrac{1}{2}\sqrt{\mathscr{U}(y+\tau)^2+\tau^2} + \tfrac{1}{2}\sqrt{\mathscr{U}(y-\tau)^2+\tau^2} \qquad (11.21)$$

for some constant C_δ depending only on δ and for every $\tau \leq \frac{\delta}{2}$. We do this using Taylor's theorem. Write $V_y(\tau)$ for the function of τ on the right-hand side of (11.21). For any $y \in [a-b, a+b]$ the function V_y is smooth on $[0, \frac{\delta}{2}]$ because $\mathscr{U}$ is a smooth, positive function on $[\frac{\delta}{2}, 1-\frac{\delta}{2}]$. Thus

$$V_y(\tau) = V_y(0) + V_y'(0)\tau + \tfrac{1}{2}V_y''(0)\tau^2 + \tfrac{1}{6}V_y'''(\xi)\tau^3$$

for some ξ between 0 and τ. The magnitude of $V_y'''(\xi)$ is indeed bounded by some C_δ depending only on δ, using the fact that $\mathscr{U}$ is smooth and positive on $[\frac{\delta}{2}, 1-\frac{\delta}{2}]$. But $V_y(0) = \mathscr{U}(y)$, and it's straightforward to calculate that

$$V_y'(0) = 0, \qquad V_y''(0) = \mathscr{U}''(y) + 1/\mathscr{U}(y) = 0,$$

the last identity used the key property $\mathscr{U}'' = -1/\mathscr{U}$ mentioned in Remark 11.53. Thus we conclude $V_y(\tau) \geq \mathscr{U}(y) - C_\delta \tau^3$, verifying (11.21) and completing the proof. □

As a matter of fact, by a minor adjustment (Exercise 11.24) to this random walk argument we can establish the following generalization of Bobkov's Inequality:

Theorem 11.55. *Let $f : \{-1,1\}^n \to [0,1]$. Then $\mathbf{E}[\|(\mathscr{U}(\mathrm{T}_\rho f), \nabla \mathrm{T}_\rho f)\|]$ is an increasing function of $\rho \in [0,1]$. We recover Bobkov's Inequality by considering $\rho = 0,1$.*

We end this section by remarking that De, Mossel, and Neeman [DMN13] have given a "Bobkov-style" Boolean inductive proof that yields both Borell's Isoperimetric Theorem and also the Majority Is Stablest Theorem (albeit with some aspects of the Invariance Principle-based proof appearing in the latter case); see Exercise 11.30 and the notes at the end of this chapter.

11.5. The Berry–Esseen Theorem

Now that we've built up some results concerning Gaussian space, we're motivated to try reducing problems involving Boolean functions to problems involving Gaussian functions. The key tool for this is the Invariance Principle, discussed at the beginning of the chapter. As a warmup, this section is devoted to proving (a form of) the Berry–Esseen Theorem. As discussed in Chapter 5.2, the Berry–Esseen Theorem is a quantitative form of the Central Limit Theorem for finite sums of independent random variables. We restate it here:

Berry–Esseen Theorem. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent random variables with $\mathbf{E}[\boldsymbol{X}_i] = 0$ and $\mathbf{Var}[\boldsymbol{X}_i] = \sigma_i^2$, and assume $\sum_{i=1}^n \sigma_i^2 = 1$. Let $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{X}_i$ and let $\boldsymbol{Z} \sim \mathrm{N}(0,1)$ be a standard Gaussian. Then for all $u \in \mathbb{R}$,*

$$|\mathbf{Pr}[\boldsymbol{S} \le u] - \mathbf{Pr}[\boldsymbol{Z} \le u]| \le c\gamma,$$

where

$$\gamma = \sum_{i=1}^n \|\boldsymbol{X}_i\|_3^3$$

and c is a universal constant. (For definiteness, $c = .56$ is acceptable.)

In this traditional statement of Berry–Esseen, the error term γ is a little opaque. To say that γ is small is to simultaneously say two things: the random variables $\boldsymbol{X}_i$ are all "reasonable" (as in Chapter 9.1); and, none is too dominant in terms of variance. In Chapter 9.1 we discussed several related notions of "reasonableness" for a random variable $\boldsymbol{X}$. It was convenient there to use the definition that $\|\boldsymbol{X}\|_4^4$ is not much larger than $\|\boldsymbol{X}\|_2^4$. For the Berry–Esseen Theorem it's more convenient (and slightly stronger) to use the analogous condition for the 3rd moment. (For the Invariance Principle it will be more convenient to use $(2,3,\rho)$- or $(2,4,\rho)$-hypercontractivity.) The implication for Berry–Esseen is the following:

Remark 11.56. In the Berry–Esseen Theorem, if all of the $\boldsymbol{X}_i$'s are "reasonable" in the sense that $\|\boldsymbol{X}_i\|_3^3 \leq B\|\boldsymbol{X}_i\|_2^3 = B\sigma_i^3$, then we can use the bound

$$\gamma \leq B \cdot \max_i\{\sigma_i\}, \tag{11.22}$$

as this is a consequence of

$$\gamma = \sum_{i=1}^n \|\boldsymbol{X}_i\|_3^3 \leq B \sum_{i=1}^n \sigma_i^3 \leq B \cdot \max_i\{\sigma_i\} \cdot \sum_{i=1}^n \sigma_i^2 = B \cdot \max_i\{\sigma_i\}.$$

(Cf. Remark 5.15.) Note that some "reasonableness" condition must hold if $\boldsymbol{S} = \sum_i \boldsymbol{X}_i$ is to behave like a Gaussian. For example, if each $\boldsymbol{X}_i$ is the "unreasonable" random variable which is $\pm\sqrt{n}$ with probability $\frac{1}{2n^2}$ each and 0 otherwise, then $\boldsymbol{S} = 0$ except with probability at most $\frac{1}{n}$ – quite unlike a Gaussian. Further, even assuming reasonableness we still need a condition like (11.22) ensuring that no $\boldsymbol{X}_i$ is too dominant ("influential") in terms of variance. For example, if $\boldsymbol{X}_1 \sim \{-1,1\}$ is a uniformly random bit and $\boldsymbol{X}_2, \dots, \boldsymbol{X}_n \equiv 0$, then $\boldsymbol{S} \equiv \boldsymbol{X}_1$, which is again quite unlike a Gaussian.

There are several known ways to prove the Berry–Esseen Theorem; for example, using characteristic functions (i.e., "real" Fourier analysis), or Stein's Method. We'll use the "Replacement Method" (also known as the Lindeberg Method, and similar to the "Hybrid Method" in theoretical cryptography). Although it doesn't always give the sharpest results, it's a very flexible technique which generalizes easily to higher-degree polynomials of random variables (as in the Invariance Principle) and random vectors. The Replacement Method suggests itself as soon as the Berry–Esseen Theorem is written in a slightly different form: Instead of trying to show

$$\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n \approx \boldsymbol{Z}, \tag{11.23}$$

where $\boldsymbol{Z} \sim \mathrm{N}(0,1)$, we'll instead try to show the equivalent statement

$$\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n \approx \boldsymbol{Z}_1 + \boldsymbol{Z}_2 + \cdots + \boldsymbol{Z}_n, \tag{11.24}$$

where the $\boldsymbol{Z}_i$'s are independent Gaussians with $\boldsymbol{Z}_i \sim \mathrm{N}(0,\sigma_i^2)$. The statements (11.23) and (11.24) really are identical, since the sum of independent Gaussians is Gaussian, with the variances adding. The Replacement Method proves (11.24) by replacing the $\boldsymbol{X}_i$'s with $\boldsymbol{Z}_i$'s one by one. Roughly speaking, we introduce the "hybrid" random variables

$$\boldsymbol{H}_t = \boldsymbol{Z}_1 + \cdots + \boldsymbol{Z}_t + \boldsymbol{X}_{t+1} + \cdots + \boldsymbol{X}_n,$$

show that $\boldsymbol{H}_{t-1} \approx \boldsymbol{H}_t$ for each $t \in [n]$, and then simply add up the n errors.

As a matter of fact, the Replacement Method doesn't really have anything to do with Gaussian random variables. It actually seeks to show that

$$\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n \approx \boldsymbol{Y}_1 + \boldsymbol{Y}_2 + \cdots + \boldsymbol{Y}_n,$$

whenever $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n,\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ are independent random variables with "matching first and second moments", meaning $\mathbf{E}[\boldsymbol{X}_i]=\mathbf{E}[\boldsymbol{Y}_i]$ and $\mathbf{E}[\boldsymbol{X}_i^2]=\mathbf{E}[\boldsymbol{Y}_i^2]$ for each $i\in[n]$. (The error will be proportional to $\sum_i(\|\boldsymbol{X}_i\|^3+\|\boldsymbol{Y}_i\|_3^3)$.) Another way of putting it (roughly speaking) is that the linear form $x_1+\cdots+x_n$ is *invariant* to what independent random variables you substitute in for $x_1,\dots,x_n$, so long as you always use the same first and second moments. The fact that we can take the $\boldsymbol{Y}_i$'s to be Gaussians (with $\boldsymbol{Y}_i\sim\mathrm{N}(\mathbf{E}[\boldsymbol{X}_i],\mathbf{Var}[\boldsymbol{X}_i])$) and then in the end use the fact that the sum of Gaussians is Gaussians to derive the simpler-looking

$$\boldsymbol{S}=\sum_{i=1}^n\boldsymbol{X}_i\approx\mathrm{N}(\mathbf{E}[\boldsymbol{S}],\mathbf{Var}[S])$$

is just a pleasant bonus (and one that we'll no longer get once we look at *nonlinear* polynomials of random variables in Section 11.6). Indeed, the remainder of this section will be devoted to showing that

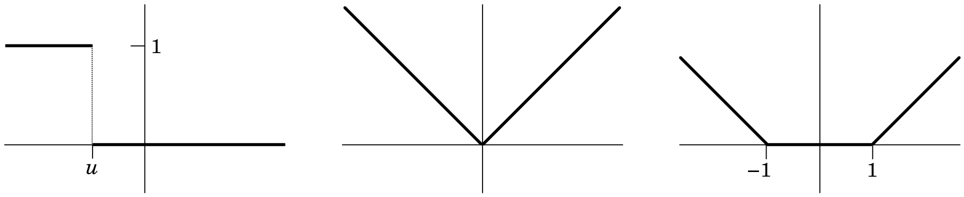
$$\boldsymbol{S}_X=\boldsymbol{X}_1+\cdots+\boldsymbol{X}_n\quad\text{is "close" to}\quad\boldsymbol{S}_Y=\boldsymbol{Y}_1+\cdots+\boldsymbol{Y}_n$$

whenever the $\boldsymbol{X}_i$'s and $\boldsymbol{Y}_i$'s are independent, "reasonable" random variables with matching first and second moments.

To do this, we'll first have to discuss in more detail what it means for two random variables to be "close". A traditional measure of closeness between two random variables $\boldsymbol{S}_X$ and $\boldsymbol{S}_Y$ is the "cdf-distance" used in the Berry–Esseen Theorem: $\mathbf{Pr}[\boldsymbol{S}_X\le u]\approx\mathbf{Pr}[\boldsymbol{S}_Y\le u]$ for every $u\in\mathbb{R}$. But there are other natural measures of closeness too. We might want to know that the absolute moments of $\boldsymbol{S}_X$ and $\boldsymbol{S}_Y$ are close; for example, that $\|\boldsymbol{S}_X\|_1\approx\|\boldsymbol{S}_Y\|_1$. Or, we might like to know that $\boldsymbol{S}_X$ and $\boldsymbol{S}_Y$ stray from the interval $[-1,1]$ by about the same amount: $\mathbf{E}[\mathrm{dist}_{[-1,1]}(\boldsymbol{S}_X)]\approx\mathbf{E}[\mathrm{dist}_{[-1,1]}(\boldsymbol{S}_Y)]$. Here we are using:

Definition 11.57. For any interval $\emptyset\neq I\subsetneq\mathbb{R}$ the function $\mathrm{dist}_I:\mathbb{R}\to\mathbb{R}^{\ge0}$ is defined to measure the distance of a point from I; i.e., $\mathrm{dist}_I(s)=\inf_{u\in I}\{|s-u|\}$.

All of the closeness measures just described can be put in a common framework: they are requiring $\mathbf{E}[\psi(\boldsymbol{S}_X)]\approx\mathbf{E}[\psi(\boldsymbol{S}_Y)]$ for various "test functions" (or "distinguishers") $\psi:\mathbb{R}\to\mathbb{R}$.



$\psi(s) = 1_{s \le u}$ $\psi(s) = |s|$ $\psi(s) = \mathrm{dist}_{[-1,1]}(s)$

Figure 11.1. The test functions ψ used for judging $\mathbf{Pr}[\boldsymbol{S}_X \le u] \approx \mathbf{Pr}[\boldsymbol{S}_Y \le u]$, $\|\boldsymbol{S}_X\|_1 \approx \|\boldsymbol{S}_Y\|_1$, and $\mathbf{E}[\mathrm{dist}_{[-1,1]}(\boldsymbol{S}_X)] \approx \mathbf{E}[\mathrm{dist}_{[-1,1]}(\boldsymbol{S}_Y)]$, respectively

It would be nice to prove a version of the Berry–Esseen Theorem that showed closeness for all the test functions ψ depicted in Figure 11.1, and more. What class of tests might we able to handle? On one hand, we can't be *too* ambitious. For example, suppose each $\boldsymbol{X}_i \sim \{-1,1\}$, each $\boldsymbol{Y}_i \sim \mathrm{N}(0,1)$, and $\psi(s) = 1_{s \in \mathbb{Z}}$. Then $\mathbf{E}[\psi(\boldsymbol{S}_X)] = 1$ because $\boldsymbol{S}_X$ is supported on the integers, but $\mathbf{E}[\psi(\boldsymbol{S}_Y)] = 0$ because $\boldsymbol{S}_Y \sim \mathrm{N}(0,n)$ is a continuous random variable. On the other hand, there are some simple kinds of tests ψ for which we have exact equality. For example, if $\psi(s) = s$, then $\mathbf{E}[\psi(\boldsymbol{S}_X)] = \mathbf{E}[\psi(\boldsymbol{S}_Y)]$; this is by the assumption of matching first moments, $\mathbf{E}[\boldsymbol{X}_i] = \mathbf{E}[\boldsymbol{Y}_i]$ for all i. Similarly, if $\psi(s) = s^2$, then

$$\begin{aligned}\mathbf{E}[\psi(\boldsymbol{S}_X)] = \mathbf{E}\Big[\big(\sum_i \boldsymbol{X}_i\big)^2\Big] &= \sum_i \mathbf{E}[\boldsymbol{X}_i^2] + \sum_{i \neq j} \mathbf{E}[\boldsymbol{X}_i \boldsymbol{X}_j] \\ &= \sum_i \mathbf{E}[\boldsymbol{X}_i^2] + \sum_{i \neq j} \mathbf{E}[\boldsymbol{X}_i]\mathbf{E}[\boldsymbol{X}_j] \qquad (11.25)\end{aligned}$$

(using independence of the $\boldsymbol{X}_i$'s); and also

$$\mathbf{E}[\psi(\boldsymbol{S}_Y)] = \sum_i \mathbf{E}[\boldsymbol{Y}_i^2] + \sum_{i \neq j} \mathbf{E}[\boldsymbol{Y}_i]\mathbf{E}[\boldsymbol{Y}_j]. \qquad (11.26)$$

The quantities (11.25) and (11.26) are equal because of the matching first and second moment conditions.

As a consequence of these observations we have $\mathbf{E}[\psi(\boldsymbol{S}_X)] = \mathbf{E}[\psi(\boldsymbol{S}_Y)]$ for any quadratic polynomial $\psi(s) = a + bs + cs^2$. This suggests that to handle a general test ψ we try to approximate it by a quadratic polynomial up to some error; in other words, consider its 2nd-order Taylor expansion. For this to make sense the function ψ must have a continuous 3rd derivative, and the error we incur will involve the magnitude of this derivative. Indeed, we will now prove a variant of the Berry–Esseen Theorem for the class of $\mathscr{C}^3$ test functions ψ with ψ''' uniformly bounded. You might be concerned that this class doesn't contain *any* of the interesting test functions depicted in Figure 11.1. But we'll be able to handle even those test functions with some loss in the parameters by using a simple "hack" – approximating them by smooth functions, as suggested in Figure 11.2.

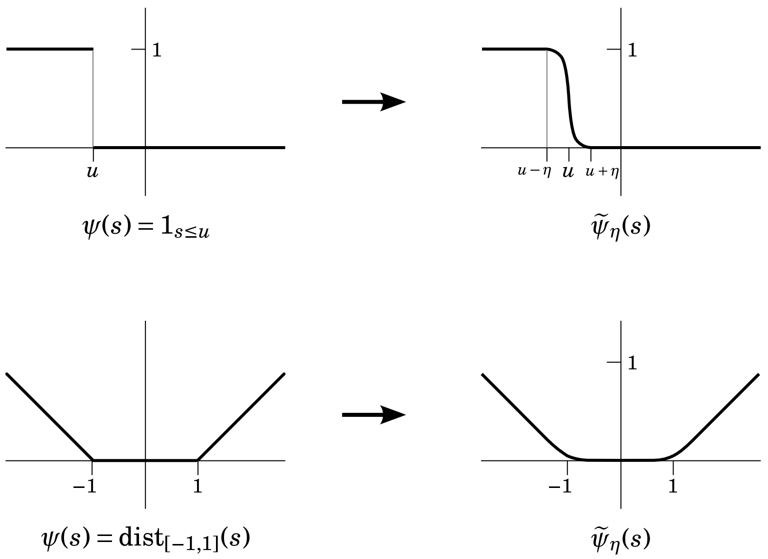


Figure 11.2. The step function $\psi(s) = 1_{s \le u}$ can be smoothed out on the interval $[u-\eta, u+\eta]$ so that the resulting function $\widetilde{\psi}_\eta$ satisfies $\|\widetilde{\psi}_\eta'''\|_\infty \le O(1/\eta^3)$. Similarly, we can smooth out $\psi(s) = \mathrm{dist}_{[-1,1]}(s)$ to a function $\widetilde{\psi}_\eta$ satisfying $\|\psi - \widetilde{\psi}\|_\infty \le \eta$ and $\|\widetilde{\psi}_\eta'''\|_\infty \le O(1/\eta^2)$.

Invariance Principle for Sums of Random Variables. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ be independent random variables with matching* 1*st and* 2*nd moments; i.e., $\mathbf{E}[\boldsymbol{X}_i^k] = \mathbf{E}[\boldsymbol{Y}_i^k]$ for $i \in [n]$, $k \in \{1,2\}$. Write $\boldsymbol{S}_X = \sum_i \boldsymbol{X}_i$ and $\boldsymbol{S}_Y = \sum_i \boldsymbol{Y}_i$. Then for any $\psi : \mathbb{R} \to \mathbb{R}$ with continuous third derivative,*

$$\left|\mathbf{E}[\psi(\boldsymbol{S}_X)] - \mathbf{E}[\psi(\boldsymbol{S}_Y)]\right| \le \tfrac{1}{6}\|\psi'''\|_\infty \cdot \gamma_{XY},$$

where $\gamma_{XY} = \sum_i (\|\boldsymbol{X}_i\|_3^3 + \|\boldsymbol{Y}_i\|_3^3)$.

Proof. The proof is by the Replacement Method. For $0 \le t \le n$, define the "hybrid" random variable

$$\boldsymbol{H}_t = \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_t + \boldsymbol{X}_{t+1} + \cdots + \boldsymbol{X}_n,$$

so $\boldsymbol{S}_X = \boldsymbol{H}_0$ and $\boldsymbol{S}_Y = \boldsymbol{H}_n$. Thus by the triangle inequality,

$$\left|\mathbf{E}[\psi(\boldsymbol{S}_X)] - \mathbf{E}[\psi(\boldsymbol{S}_Y)]\right| \le \sum_{t=1}^{n} \left|\mathbf{E}[\psi(\boldsymbol{H}_{t-1})] - \mathbf{E}[\psi(\boldsymbol{H}_t)]\right|.$$

Given the definition of γ_{XY}, we can complete the proof by showing that for each $t \in [n]$,

$$\begin{aligned}\tfrac{1}{6}\|\psi'''\|_\infty \cdot (\mathbf{E}[|\boldsymbol{X}_t|^3] + \mathbf{E}[|\boldsymbol{Y}_t|^3]) &\ge \left|\mathbf{E}[\psi(\boldsymbol{H}_{t-1})] - \mathbf{E}[\psi(\boldsymbol{H}_t)]\right| \\ &= \left|\mathbf{E}[\psi(\boldsymbol{H}_{t-1}) - \psi(\boldsymbol{H}_t)]\right| \\ &= \left|\mathbf{E}[\psi(\boldsymbol{U}_t + \boldsymbol{X}_t) - \psi(\boldsymbol{U}_t + \boldsymbol{Y}_t)]\right|, \qquad (11.27)\end{aligned}$$

where

$$\boldsymbol{U}_t = \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_{t-1} + \boldsymbol{X}_{t+1} + \cdots + \boldsymbol{X}_n.$$

Note that $\boldsymbol{U}_t$ is independent of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. We are now comparing ψ's values at $\boldsymbol{U}_t + \boldsymbol{X}_t$ and $\boldsymbol{U}_t + \boldsymbol{Y}_t$, with the presumption that $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are rather small compared to $\boldsymbol{U}_t$. This clearly suggests the use of Taylor's theorem: For all $u, \delta \in \mathbb{R}$,

$$\psi(u+\delta) = \psi(u) + \psi'(u)\delta + \tfrac{1}{2}\psi''(u)\delta^2 + \tfrac{1}{6}\psi'''(u^*)\delta^3,$$

for some $u^* = u^*(u,\delta)$ between u and $u+\delta$. Applying this pointwise with $u = \boldsymbol{U}_t$, $\delta = \boldsymbol{X}_t, \boldsymbol{Y}_t$ yields

$$\psi(\boldsymbol{U}_t + \boldsymbol{X}_t) = \psi(\boldsymbol{U}_t) + \psi'(\boldsymbol{U}_t)\boldsymbol{X}_t + \tfrac{1}{2}\psi''(\boldsymbol{U}_t)\boldsymbol{X}_t^2 + \tfrac{1}{6}\psi'''(\boldsymbol{U}_t^*)\boldsymbol{X}_t^3$$
$$\psi(\boldsymbol{U}_t + \boldsymbol{Y}_t) = \psi(\boldsymbol{U}_t) + \psi'(\boldsymbol{U}_t)\boldsymbol{Y}_t + \tfrac{1}{2}\psi''(\boldsymbol{U}_t)\boldsymbol{Y}_t^2 + \tfrac{1}{6}\psi'''(\boldsymbol{U}_t^{**})\boldsymbol{Y}_t^3$$

for some random variables $\boldsymbol{U}_t^*, \boldsymbol{U}_t^{**}$. Referring back to our goal of (11.27), what happens when we subtract these two identities and take expectations? The $\psi(\boldsymbol{U}_t)$ terms cancel. The next difference is

$$\mathbf{E}[\psi'(\boldsymbol{U}_t)(\boldsymbol{X}_t - \boldsymbol{Y}_t)] = \mathbf{E}[\psi'(\boldsymbol{U}_t)] \cdot \mathbf{E}[\boldsymbol{X}_t - \boldsymbol{Y}_t] = \mathbf{E}[\psi'(\boldsymbol{U}_t)] \cdot 0 = 0,$$

where the first equality used that $\boldsymbol{U}_t$ is independent of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, and the second equality used the matching 1st moments of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. An identical argument, using matching 2nd moments, shows that the shows that the difference of the quadratic terms disappears in expectation. Thus we're left only with the "error term":

$$\begin{aligned}\left|\mathbf{E}[\psi(\boldsymbol{U}_t + \boldsymbol{X}_t) - \psi(\boldsymbol{U}_t + \boldsymbol{Y}_t)]\right| &= \tfrac{1}{6}\left|\mathbf{E}[\psi'''(\boldsymbol{U}_t^*)\boldsymbol{X}_t^3 - \psi'''(\boldsymbol{U}_t^{**})\boldsymbol{Y}_t^3]\right| \\ &\leq \tfrac{1}{6}\|\psi'''\|_\infty \cdot (\mathbf{E}[|\boldsymbol{X}_t|^3] + \mathbf{E}[|\boldsymbol{Y}_t|^3]),\end{aligned}$$

where the last step used the triangle inequality. This confirms (11.27) and completes the proof. □

We can now give a Berry–Esseen-type corollary by taking the $\boldsymbol{Y}_i$'s to be Gaussians:

Variant Berry–Esseen Theorem. *In the setting of the Berry–Esseen Theorem, for all $\mathscr{C}^3$ functions $\psi : \mathbb{R} \to \mathbb{R}$,*

$$\left|\mathbf{E}[\psi(\boldsymbol{S})] - \mathbf{E}[\psi(\boldsymbol{Z})]\right| \leq \tfrac{1}{6}(1 + 2\sqrt{\tfrac{2}{\pi}})\|\psi'''\|_\infty \cdot \gamma \leq .433\|\psi'''\|_\infty \cdot \gamma.$$

Proof. Applying the preceding theorem with $\boldsymbol{Y}_i \sim \mathrm{N}(0, \sigma_i^2)$ (and hence $\boldsymbol{S}_Y \sim \mathrm{N}(0,1)$), it suffices to show that

$$\gamma_{XY} = \sum_{i=1}^n (\|\boldsymbol{X}_i\|_3^3 + \|\boldsymbol{Y}_i\|_3^3) \leq (1 + 2\sqrt{\tfrac{2}{\pi}}) \cdot \gamma = (1 + 2\sqrt{\tfrac{2}{\pi}}) \cdot \sum_{i=1}^n \|\boldsymbol{X}_i\|_3^3. \qquad (11.28)$$

In particular, we just need to show that $\|\boldsymbol{Y}_i\|_3^3 \leq 2\sqrt{\frac{2}{\pi}}\|\boldsymbol{X}_i\|_3^3$ for each i. This holds because Gaussians are extremely reasonable; by explicitly computing 3rd absolute moments we indeed obtain

$$\|\boldsymbol{Y}_i\|_3^3 = \sigma_i^3\|\mathrm{N}(0,1)\|_3^3 = 2\sqrt{\tfrac{2}{\pi}}\sigma_i^3 = 2\sqrt{\tfrac{2}{\pi}}\|\boldsymbol{X}_i\|_2^3 \leq 2\sqrt{\tfrac{2}{\pi}}\|\boldsymbol{X}_i\|_3^3. \qquad \Box$$

This version of the Berry–Esseen Theorem is incomparable with the standard version. Sometimes it can be stronger; for example, if for some reason we wanted to show $\mathbf{E}[\cos \boldsymbol{S}] \approx \mathbf{E}[\cos \boldsymbol{Z}]$ then the Variant Berry–Esseen Theorem gives this with error $.433\gamma$, whereas it can't be directly deduced from the standard Berry–Esseen at all. On the other hand, as we'll see shortly, we can only obtain the standard Berry–Esseen conclusion from the Variant version with an error bound of $O(\gamma^{1/4})$ rather than $O(\gamma)$.

We end this section by describing the "hacks" which let us extend the Variant Berry–Esseen Theorem to cover certain non-$\mathscr{C}^3$ tests ψ. As mentioned the idea is to smooth them out, or "mollify" them:

Proposition 11.58. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be c-Lipschitz. Then for any $\eta > 0$ there exists $\widetilde{\psi}_\eta : \mathbb{R} \to \mathbb{R}$ satisfying $\|\psi - \widetilde{\psi}_\eta\|_\infty \leq c\eta$ and $\|\widetilde{\psi}_\eta^{(k)}\|_\infty \leq C_k c/\eta^{k-1}$ for each $k \in \mathbb{N}^+$. Here C_k is a constant depending only on k, and $\widetilde{\psi}_\eta^{(k)}$ denotes the kth derivative of $\widetilde{\psi}_\eta$.*

The proof is straightforward, taking $\widetilde{\psi}_\eta(s) = \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)}[\psi(s+\eta\boldsymbol{g})]$; see Exercise 11.38.

As $\eta \to 0$ this gives a better and better smooth approximation to ψ, but also a larger and larger value of $\|\widetilde{\psi}_\eta'''\|_\infty$. Trading these off gives the following:

Corollary 11.59. *In the setting of the Invariance Principle for Sums of Random Variables, if we merely have that $\psi : \mathbb{R} \to \mathbb{R}$ is c-Lipschitz, then*

$$\left|\mathbf{E}[\psi(\boldsymbol{S}_X)] - \mathbf{E}[\psi(\boldsymbol{S}_Y)]\right| \leq O(c) \cdot \gamma_{XY}^{1/3}.$$

Proof. Applying the Invariance Principle for Sums of Random Variables with the test $\widetilde{\psi}_\eta$ from Proposition 11.58 we get

$$\left|\mathbf{E}[\widetilde{\psi}_\eta(\boldsymbol{S}_X)] - \mathbf{E}[\widetilde{\psi}_\eta(\boldsymbol{S}_Y)]\right| \leq O(c/\eta^2) \cdot \gamma_{XY}.$$

But $\|\widetilde{\psi}_\eta - \psi\|_\infty \leq c\eta$ implies

$$\left|\mathbf{E}[\widetilde{\psi}_\eta(\boldsymbol{S}_X)] - \mathbf{E}[\psi(\boldsymbol{S}_X)]\right| \leq \mathbf{E}[|\widetilde{\psi}_\eta(\boldsymbol{S}_X) - \psi(\boldsymbol{S}_X)|] \leq c\eta$$

and similarly for $\boldsymbol{S}_Y$. Thus we get

$$\left|\mathbf{E}[\psi(\boldsymbol{S}_X)] - \mathbf{E}[\psi(\boldsymbol{S}_Y)]\right| \leq O(c) \cdot (\eta + \gamma_{XY}/\eta^2)$$

which yields the desired bound by taking $\eta = \gamma_{XY}^{1/3}$. $\Box$

Remark 11.60. It's obvious that the dependence on c in this theorem should be linear in c; in fact, since we can always divide ψ by c it would have sufficed to prove the theorem assuming $c = 1$.

This corollary covers all Lipschitz tests, which suffices for the functions $\psi(s) = |s|$ and $\psi(s) = \mathrm{dist}_{[-1,1]}(s)$ from Figure 11.1. However, it still isn't enough for the test $\psi(s) = 1_{s \le u}$ – i.e., for establishing cdf-closeness as in the usual Berry–Esseen Theorem. Of course, we can't hope for a smooth approximator $\widetilde{\psi}_\eta$ satisfying $|\widetilde{\psi}_\eta(s) - 1_{s\le u}| \le \eta$ for all s because of the discontinuity at u. However, as suggested in Figure 11.2, if we're willing to exclude $s \in [u-\eta, u+\eta]$ we can get an approximator with third derivative bound $O(1/\eta^3)$, and thereby obtain (Exercises 11.41, 11.42):

Corollary 11.61. *In the setting of the Invariance Principle for Sums of Random Variables, for all $u \in \mathbb{R}$ we have*

$$\mathbf{Pr}[\boldsymbol{S}_Y \le u - \epsilon] - \epsilon \le \mathbf{Pr}[\boldsymbol{S}_X \le u] \le \mathbf{Pr}[\boldsymbol{S}_Y \le u + \epsilon] + \epsilon$$

for $\epsilon = O(\gamma_{XY}^{1/4})$; i.e., $\boldsymbol{S}_X$ and $\boldsymbol{S}_Y$ have Lévy distance $d_L(\boldsymbol{S}_X, \boldsymbol{S}_Y) \le O(\gamma_{XY}^{1/4})$.

Finally, in the Berry–Esseen setting where $\boldsymbol{S}_Y \sim \mathrm{N}(0,1)$, we can appeal to the "anticoncentration" of Gaussians:

$$\mathbf{Pr}[\mathrm{N}(0,1) \le u+\epsilon] = \mathbf{Pr}[\mathrm{N}(0,1) \le u] + \mathbf{Pr}[u < \mathrm{N}(0,1) \le u+\epsilon] \le \mathbf{Pr}[\mathrm{N}(0,1) \le u] + \tfrac{\epsilon}{\sqrt{2\pi}},$$

and similarly for $\mathbf{Pr}[\mathrm{N}(0,1) \le u - \epsilon]$. This lets us convert the Lévy distance bound into a cdf-distance bound. Recalling (11.28), we immediately deduce the following weaker version of the classical Berry–Esseen Theorem:

Corollary 11.62. *In the setting of the Berry–Esseen Theorem, for all $u \in \mathbb{R}$,*

$$|\mathbf{Pr}[\boldsymbol{S} \le u] - \mathbf{Pr}[\boldsymbol{Z} \le u| \le O(\gamma^{1/4}),$$

where the $O(\cdot)$ hides a universal constant.

Although the error bound here is weaker than necessary by a power of 1/4, this weakness will be more than made up for by the ease with which the Replacement Method generalizes to other settings. In the next section we'll see it applied to nonlinear polynomials of independent random variables. Exercise 11.46 outlines how to use it to give a Berry–Esseen theorem for sums of independent random vectors; as you'll see, other than replacing Taylor's theorem with its multivariate form, hardly a symbol in the proof changes.

11.6. The Invariance Principle

Let's summarize the Variant Berry–Esseen Theorem and proof from the preceding section, using slightly different notation. (Specifically, we'll rewrite $\boldsymbol{X}_i = a_i \boldsymbol{x}_i$ where $\mathbf{Var}[\boldsymbol{x}_i] = 1$, so $a_i = \pm\sigma_i$.) We showed that if $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{y}_1, \dots, \boldsymbol{y}_n$

are independent mean-0, variance-1 random variables, reasonable in the sense of having third absolute moment at most B, and if $a_1,\dots,a_n$ are real constants assumed for normalization to satisfy $\sum_i a_i^2 = 1$, then

$$a_1\boldsymbol{x}_1 + \cdots + a_n\boldsymbol{x}_n \approx a_1\boldsymbol{y}_1 + \cdots + a_n\boldsymbol{y}_n,$$
$$\text{with error bound proportional to } B\max\{|a_i|\}.$$

We think of this as saying that the linear form $a_1x_1 + \cdots + a_nx_n$ is (roughly) *invariant* to what independent mean-0, variance-1, reasonable random variables are substituted for the x_i's, so long as all $|a_i|$'s are "small" (compared to the overall variance). In this section we generalize this statement to degree-k multilinear polynomial forms, $\sum_{|S|\leq k} a_S x^S$. The appropriate generalization of the condition that "all $|a_i|$'s are small" is the condition that all "influences" $\sum_{S\ni i} a_S^2$ are small. We refer to these nonlinear generalizations of Berry–Esseen as *Invariance Principles*.

In this section we'll develop the most basic Invariance Principle, which involves replacing bits by Gaussians for a single Boolean function f. We'll show that this doesn't change the distribution of f much provided f has small influences and provided that f is of "constant degree" – or at least, provided f is uniformly noise-stable so that it's "close to having constant degree". Invariance Principles in much more general settings are possible – for example Exercises 11.48 and 11.49 describe variants which handle several functions applied to correlated inputs, and functions on general product spaces. Here we'll just focus on the simplest possible Invariance Principle, which is already sufficient for the proof of the Majority Is Stablest Theorem in Section 11.7.

Let's begin with some notation.

Definition 11.63. Let F be a formal multilinear polynomial over the sequence of indeterminates $x = (x_1,\dots,x_n)$:

$$F(x) = \sum_{S\subseteq[n]} \widehat{F}(S) \prod_{i\in S} x_i,$$

where the coefficients $\widehat{F}(S)$ are real numbers. We introduce the notation

$$\mathbf{Var}[F] = \sum_{S\neq\emptyset} \widehat{F}(S)^2, \qquad \mathbf{Inf}_i[F] = \sum_{S\ni i} \widehat{F}(S)^2.$$

Remark 11.64. To justify this notation, we remark that we'll always consider F applied to a sequence $\boldsymbol{z} = (\boldsymbol{z}_1,\dots,\boldsymbol{z}_n)$ independent random variables satisfying $\mathbf{E}[\boldsymbol{z}_i] = 0$, $\mathbf{E}[\boldsymbol{z}_i^2] = 1$. Under these circumstances the collection of monomial random variables $\prod_{i\in S}\boldsymbol{z}_i$ is orthonormal and so it's easy to see (cf. Chapter 8.2) that

$$\mathbf{E}[F(\boldsymbol{z})] = \widehat{F}(\emptyset), \quad \mathbf{E}[F(\boldsymbol{z})^2] = \sum_{S\subseteq[n]} \widehat{F}(S)^2, \quad \mathbf{Var}[F(\boldsymbol{z})] = \mathbf{Var}[F] = \sum_{S\neq\emptyset} \widehat{F}(S)^2.$$

We also have $\mathbf{E}[\mathbf{Var}_{\boldsymbol{z}_i}[F(\boldsymbol{z})]] = \mathbf{Inf}_i[F] = \sum_{S\ni i}\widehat{F}(S)^2$, though we won't use this.

As in the Berry–Esseen Theorem, to get good error bounds we'll need our random variables $\boldsymbol{z}_i$ to be "reasonable". Sacrificing generality for simplicity in this section, we'll take the bounded 4th-moment notion from Definition 9.1 which will allow us to use the basic Bonami Lemma (more precisely, Corollary 9.6):

Hypothesis 11.65. *The random variable $\boldsymbol{z}_i$ satisfies $\mathbf{E}[\boldsymbol{z}_i] = 0$, $\mathbf{E}[\boldsymbol{z}_i^2] = 1$, $\mathbf{E}[\boldsymbol{z}_i^3] = 0$, and is "9-reasonable" in the sense of Definition 9.1; i.e., $\mathbf{E}[\boldsymbol{z}_i^4] \leq 9$.*

The main examples we have in mind are that each $\boldsymbol{z}_i$ is either a uniform ±1 random bit or a standard Gaussian. (There are other possibilities, though; e.g., $\boldsymbol{z}_i$ could be uniform on the interval $[-\sqrt{3}, \sqrt{3}]$.)

We can now prove the most basic Invariance Principle, for low-degree multilinear polynomials of random variables:

Basic Invariance Principle. *Let F be a formal n-variate multilinear polynomial of degree at most $k \in \mathbb{N}$,*

$$F(x) = \sum_{S \subseteq [n], |S| \leq k} \widehat{F}(S) \prod_{i \in S} x_i.$$

Let $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ be sequences of independent random variables, each satisfying Hypothesis 11.65. Assume $\psi : \mathbb{R} \to \mathbb{R}$ is $\mathscr{C}^4$ with $\|\psi''''\|_\infty \leq C$. Then

$$\left|\mathbf{E}[\psi(F(\boldsymbol{x}))] - \mathbf{E}[\psi(F(\boldsymbol{y}))]\right| \leq \tfrac{C}{12} \cdot 9^k \cdot \sum_{t=1}^{n} \mathbf{Inf}_t[F]^2. \tag{11.29}$$

Remark 11.66. The proof will be very similar to the one we used for Berry–Esseen except that we'll take a 3rd-order Taylor expansion rather than a 2nd-order one (so that we can use the easy Bonami Lemma). As you are asked to show in Exercise 11.47, had we only required that ψ be $\mathscr{C}^3$ and that the $\boldsymbol{x}_i$'s and $\boldsymbol{y}_i$'s be $(2,3,\rho)$-hypercontractive with 2nd moment equal to 1, then we could obtain

$$\left|\mathbf{E}[\psi(F(\boldsymbol{x}))] - \mathbf{E}[\psi(F(\boldsymbol{y}))]\right| \leq \tfrac{\|\psi'''\|_\infty}{3} \cdot (1/\rho)^{3k} \cdot \sum_{t=1}^{n} \mathbf{Inf}_t[F]^{3/2}.$$

Proof. The proof uses the Replacement Method. For $0 \leq t \leq n$ we define

$$\boldsymbol{H}_t = F(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t, \boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_n),$$

so $F(\boldsymbol{x}) = \boldsymbol{H}_0$ and $F(\boldsymbol{y}) = \boldsymbol{H}_n$. We will show that

$$\left|\mathbf{E}[\psi(\boldsymbol{H}_{t-1}) - \psi(\boldsymbol{H}_t)]\right| \leq \tfrac{C}{12} \cdot 9^k \cdot \mathbf{Inf}_t[F]^2; \tag{11.30}$$

as in our proof of the Berry–Esseen Theorem, this will complete the proof after summing over t and using the triangle inequality. To analyze (11.30) we separate out the part of $F(x)$ that depends on x_t; i.e., we write $F(x) = \mathrm{E}_tF(x) + x_t\mathrm{D}_tF(x)$, where the formal polynomials E_tF and D_tF are defined by

$$\mathrm{E}_tF(x) = \sum_{S \not\ni t} \widehat{F}(S) \prod_{i \in S} x_i, \qquad \mathrm{D}_tF(x) = \sum_{S \ni t} \widehat{F}(S) \prod_{i \in S \setminus \{t\}} x_i.$$

Note that neither E_tF nor D_tF depends on the indeterminate x_t; thus we can define

$$\begin{aligned} \boldsymbol{U}_t &= \mathrm{E}_tF(\boldsymbol{y}_1, \dots, \boldsymbol{y}_{t-1}, \cdot, \boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_n), \\ \boldsymbol{\Delta}_t &= \mathrm{D}_tF(\boldsymbol{y}_1, \dots, \boldsymbol{y}_{t-1}, \cdot, \boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_n), \end{aligned}$$

so that

$$\boldsymbol{H}_{t-1} = \boldsymbol{U}_t + \boldsymbol{\Delta}_t \boldsymbol{x}_t, \qquad \boldsymbol{H}_t = \boldsymbol{U}_t + \boldsymbol{\Delta}_t \boldsymbol{y}_t.$$

We now use a 3rd-order Taylor expansion to bound (11.30):

$$\psi(\boldsymbol{H}_{t-1}) = \psi(\boldsymbol{U}_t) + \psi'(\boldsymbol{U}_t)\boldsymbol{\Delta}_t\boldsymbol{x}_t + \tfrac{1}{2}\psi''(\boldsymbol{U}_t)\boldsymbol{\Delta}_t^2\boldsymbol{x}_t^2 + \tfrac{1}{6}\psi'''(\boldsymbol{U}_t)\boldsymbol{\Delta}_t^3\boldsymbol{x}_t^3 + \tfrac{1}{24}\psi''''(\boldsymbol{U}_t^*)\boldsymbol{\Delta}_t^4\boldsymbol{x}_t^4$$

$$\psi(\boldsymbol{H}_t) = \psi(\boldsymbol{U}_t) + \psi'(\boldsymbol{U}_t)\boldsymbol{\Delta}_t\boldsymbol{y}_t + \tfrac{1}{2}\psi''(\boldsymbol{U}_t)\boldsymbol{\Delta}_t^2\boldsymbol{y}_t^2 + \tfrac{1}{6}\psi'''(\boldsymbol{U}_t)\boldsymbol{\Delta}_t^3\boldsymbol{y}_t^3 + \tfrac{1}{24}\psi''''(\boldsymbol{U}_t^{**})\boldsymbol{\Delta}_t^4\boldsymbol{y}_t^4$$

for some random variables $\boldsymbol{U}_t^*$ and $\boldsymbol{U}_t^{**}$. As in the proof of the Berry–Esseen Theorem, when we subtract these and take the expectation there are significant simplifications. The 0th-order terms cancel. As for the 1st-order terms,

$$\mathbf{E}[\psi'(\boldsymbol{U}_t)\boldsymbol{\Delta}_t\boldsymbol{x}_t - \psi'(\boldsymbol{U}_t)\boldsymbol{\Delta}_t\boldsymbol{y}_t] = \mathbf{E}[\psi'(\boldsymbol{U}_t)\boldsymbol{\Delta}_t \cdot (\boldsymbol{x}_t - \boldsymbol{y}_t)] = \mathbf{E}(\psi'(\boldsymbol{U}_t)\boldsymbol{\Delta}_t] \cdot \mathbf{E}[\boldsymbol{x}_t - \boldsymbol{y}_t] = 0.$$

The second equality here crucially uses the fact that $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ are independent of $\boldsymbol{U}_t$, $\boldsymbol{\Delta}_t$. The final equality only uses the fact that $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ have matching 1st moments (and not the stronger assumption that both of these 1st moments are 0). The 2nd- and 3rd-order terms will similarly cancel, using the fact that $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ have matching 2nd and 3rd moments. Finally, for the "error" term we'll just use $|\psi''''(\boldsymbol{U}_t^*)|, |\psi''''(\boldsymbol{U}_t^{**})| \leq C$ and the triangle inequality; we thus obtain

$$\left|\mathbf{E}[\psi(\boldsymbol{H}_{t-1}) - \psi(\boldsymbol{H}_t)]\right| \leq \tfrac{C}{24} \cdot (\mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{x}_t)^4] + \mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{y}_t)^4]).$$

To complete the proof of (11.30) we now just need to bound

$$\mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{x}_t)^4],\ \mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{y}_t)^4] \leq 9^k \cdot \mathbf{Inf}_t[F]^2,$$

which we'll do using the Bonami Lemma. We'll give the proof for $\mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{x}_t)^4]$, the case of $\mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{y}_t)^4]$ being identical. We have

$$\boldsymbol{\Delta}_t\boldsymbol{x}_t = \mathrm{L}_tF(\boldsymbol{y}_1, \dots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_n),$$

where

$$\mathrm{L}_tF(x) = x_t\mathrm{D}_tF(x) = \sum_{S \ni t} \widehat{F}(S) \prod_{i \in S} x_i.$$

Since L_tF has degree at most k we can apply the Bonami Lemma (more precisely, Corollary 9.6) to obtain

$$\mathbf{E}[(\boldsymbol{\Delta}_t\boldsymbol{x}_t)^4] \leq 9^k\, \mathbf{E}[\mathrm{L}_tF(\boldsymbol{y}_1, \dots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_n)^2]^2.$$

But since $\boldsymbol{y}_1,\dots,\boldsymbol{y}_{t-1},\boldsymbol{x}_t,\dots,\boldsymbol{x}_n$ are independent with mean 0 and 2nd moment 1, we have (see Remark 11.64)

$$\mathbf{E}[\mathrm{L}_t F(\boldsymbol{y}_1,\dots,\boldsymbol{y}_{t-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1},\dots,\boldsymbol{x}_n)^2] = \sum_{S\subseteq[n]} \widehat{\mathrm{L}_t F}(S)^2 = \sum_{S\ni t} \widehat{F}(S)^2 = \mathbf{Inf}_t[F].$$

Thus we indeed have $\mathbf{E}[(\Delta_t \boldsymbol{x}_t)^4] \le 9^k \cdot \mathbf{Inf}_t[F]^2$, and the proof is complete. □

Corollary 11.67. *In the setting of the preceding theorem, if we furthermore have* $\mathbf{Var}[F] \le 1$ *and* $\mathbf{Inf}_t[F] \le \epsilon$ *for all* $t \in [n]$, *then*

$$\left|\mathbf{E}[\psi(F(\boldsymbol{x}))] - \mathbf{E}[\psi(F(\boldsymbol{y}))]\right| \le \tfrac{C}{12} \cdot k9^k \cdot \epsilon.$$

Proof. We have $\sum_t \mathbf{Inf}_t[F]^2 \le \epsilon \sum_t \mathbf{Inf}_t[F] \le \epsilon \sum_S |S| \widehat{F}(S)^2 \le \epsilon k \,\mathbf{Var}[F]$. □

Corollary 11.68. *In the setting of the preceding corollary, if we merely have that* $\psi : \mathbb{R} \to \mathbb{R}$ *is c-Lipschitz (rather than* $\mathscr{C}^4$*), then*

$$\left|\mathbf{E}[\psi(F(\boldsymbol{x}))] - \mathbf{E}[\psi(F(\boldsymbol{y}))]\right| \le O(c) \cdot 2^k \epsilon^{1/4}.$$

Proof. Just as in the proof of Corollary 11.59, by using $\widetilde{\psi}_\eta$ from Proposition 11.58 (which has $\|\widetilde{\psi}_\eta''''\|_\infty \le O(c/\eta^3)$) we obtain

$$\left|\mathbf{E}[\psi(F(\boldsymbol{x}))] - \mathbf{E}[\psi(F(\boldsymbol{y}))]\right| \le O(c) \cdot (\eta + k9^k \epsilon/\eta^3).$$

The proof is completed by taking $\eta = \sqrt[4]{k9^k\epsilon} \le 2^k \epsilon^{1/4}$. □

Let's connect this last corollary back to the study of Boolean functions. Suppose $f : \{-1,1\}^n \to \mathbb{R}$ has ϵ-small influences (in the sense of Definition 6.9) and degree at most k. Letting $\boldsymbol{g} = (\boldsymbol{g}_1,\dots,\boldsymbol{g}_n)$ be a sequence of independent standard Gaussians, Corollary 11.68 tells us that for any Lipschitz ψ we have

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\psi(f(\boldsymbol{x}))] - \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)^n}[\psi(f(\boldsymbol{g}))] \right| \le O(2^k \epsilon^{1/4}). \tag{11.31}$$

Here the expression "$f(\boldsymbol{g})$" is an abuse of notation indicating that the real numbers $\boldsymbol{g}_1,\dots,\boldsymbol{g}_n$ are substituted into f's Fourier expansion (multilinear polynomial representation).

At first it may seem peculiar to substitute arbitrary real numbers into the Fourier expansion of a Boolean function. Actually, if all the numbers being substituted are in the range $[-1,1]$ then there's a natural interpretation: as you were asked to show in Exercise 1.4, if $\mu \in [-1,1]^n$, then $f(\mu) = \mathbf{E}[f(\boldsymbol{y})]$ where $\boldsymbol{y} \sim \{-1,1\}^n$ is drawn from the product distribution in which $\mathbf{E}[\boldsymbol{y}_i] = \mu_i$. On the other hand, there doesn't seem to be any obvious meaning when real numbers outside the range $[-1,1]$ are substituted into f's Fourier expansion, as may certainly occur when we consider $f(\boldsymbol{g})$.

Nevertheless, (11.31) says that when f is a low-degree, small-influence function, the distribution of the random variable $f(\boldsymbol{g})$ will be close to that of $f(\boldsymbol{x})$. Now suppose $f : \{-1,1\}^n \to \{-1,1\}$ is Boolean-valued and unbiased.

Then (11.31) might seem impossible; how could the continuous random variable $f(\boldsymbol{g})$ essentially be -1 with probability $1/2$ and $+1$ with probability $1/2$? The solution to this mystery is that there *are no* low-degree, small-influence, unbiased Boolean-valued functions. This is a consequence of the OSSS Inequality – more precisely, Exercise 8.44(*b*) – which shows that in this setting we will always have $\epsilon \geq 1/k^3$ in (11.31), rendering the bound very weak. If the Aaronson–Ambainis Conjecture holds (see the notes in Chapter 8.7), a similar statement is true even for functions with range $[-1,1]$.

The reason (11.31) is still useful is that we can apply it to small-influence, low-degree functions which are *almost* $\{-1,1\}$-valued, or $[-1,1]$-valued. Such functions can arise from truncating a very noise-stable Boolean-valued function to a large but constant degree. For example, we might profitably apply (11.31) to $f = \mathrm{Maj}_n^{\leq k}$ and then deduce some consequences for $\mathrm{Maj}_n(\boldsymbol{x})$ using the fact that $\mathbf{E}[(\mathrm{Maj}_n^{\leq k}(\boldsymbol{x}) - \mathrm{Maj}_n(\boldsymbol{x}))^2] = \mathbf{W}^{>k}[\mathrm{Maj}_n] \leq O(1/\sqrt{k})$ (Corollary 5.23). Let's consider this sort of idea more generally:

Corollary 11.69. *Let $f : \{-1,1\}^n \to \mathbb{R}$ have $\mathbf{Var}[f] \leq 1$. Let $k \geq 0$ and suppose $f^{\leq k}$ has ϵ-small influences. Then for any c-Lipschitz $\psi : \mathbb{R} \to \mathbb{R}$ we have*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\psi(f(\boldsymbol{x}))] - \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)^n}[\psi(f(\boldsymbol{g}))] \right| \leq O(c) \cdot \left(2^k \epsilon^{1/4} + \|f^{>k}\|_2\right). \tag{11.32}$$

In particular, suppose $h : \{-1,1\}^n \to \mathbb{R}$ has $\mathbf{Var}[h] \leq 1$ and no (ϵ,δ)-notable coordinates (we assume $\epsilon \leq 1$, $\delta \leq \frac{1}{20}$). Then

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}[\psi(\mathrm{T}_{1-\delta} h(\boldsymbol{x}))] - \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)^n}[\psi(\mathrm{T}_{1-\delta} h(\boldsymbol{g}))] \right| \leq O(c) \cdot \epsilon^{\delta/3}.$$

Proof. For the first statement we simply decompose $f = f^{\leq k} + f^{>k}$. Then the left-hand side of (11.32) can be written as

$$\begin{aligned} &\left| \mathbf{E}[\psi(f^{\leq k}(\boldsymbol{x}) + f^{>k}(\boldsymbol{x}))] - \mathbf{E}[\psi(f^{\leq k}(\boldsymbol{g}) + f^{>k}(\boldsymbol{g}))] \right| \\ &\qquad \leq \left| \mathbf{E}[\psi(f^{\leq k}(\boldsymbol{x}))] - \mathbf{E}[\psi(f^{\leq k}(\boldsymbol{g}))] \right| + c\,\mathbf{E}[|f^{>k}(\boldsymbol{x})|] + c\,\mathbf{E}[|f^{>k}(\boldsymbol{g})|], \end{aligned}$$

using the fact that ψ is c-Lipschitz. The first quantity is at most $O(c) \cdot 2^k \epsilon^{1/4}$, by Corollary 11.68 (even if k is not an integer). As for the other two quantities, Cauchy–Schwarz implies

$$\mathbf{E}[|f^{>k}(\boldsymbol{x})|] \leq \sqrt{\mathbf{E}[f^{>k}(\boldsymbol{x})^2]} = \sqrt{\sum_{|S|>k} \widehat{f}(S)^2} = \|f^{>k}\|_2,$$

and the same bound also holds for $\mathbf{E}[|f^{>k}(\boldsymbol{g})|]$; this uses the fact that $\mathbf{E}[f^{>k}(\boldsymbol{g})^2] = \sum_{|S|>k} \widehat{f}(S)^2$ just as in Remark 11.64. This completes the proof of (11.32).

As for the second statement of the corollary, let $f = \mathrm{T}_{1-\delta} h$. The assumptions on h imply that $\mathbf{Var}[f] \leq 1$ and that $f^{\leq k}$ has ϵ-small influences for any k;

the latter is true because

$$\mathbf{Inf}_i[f^{\le k}] = \sum_{|S|\le k, S\ni i} (1-\delta)^{2|S|}\widehat{h}(S)^2 \le \sum_{S\ni i}(1-\delta)^{|S|-1}\widehat{h}(S)^2 = \mathbf{Inf}_i^{(1-\delta)}[h] \le \epsilon$$

since h has no (ϵ,δ)-notable coordinate. Furthermore,

$$\|f^{>k}\|_2^2 = \sum_{|S|>k}(1-\delta)^{2|S|}\widehat{h}(S)^2 \le (1-\delta)^{2k}\,\mathbf{Var}[h] \le (1-\delta)^{2k} \le \exp(-2k\delta)$$

for any $k \ge 1$; i.e., $\|f^{>k}\|_2 \le \exp(-k\delta)$. So applying the first part of the corollary gives

$$\big|\mathbf{E}[\psi(f(\boldsymbol{x}))] - \mathbf{E}[\psi(f(\boldsymbol{g}))]\big| \le O(c)\cdot\big(2^k\epsilon^{1/4} + \exp(-k\delta)\big) \tag{11.33}$$

for any $k \ge 0$. Choosing $k = \frac{1}{3}\ln(1/\epsilon)$, the right-hand side of (11.33) becomes

$$O(c)\cdot\left(\epsilon^{-(1/3)\ln 2}\epsilon^{1/4} + \epsilon^{\delta/3}\right) \le O(c)\cdot\epsilon^{\delta/3},$$

where the inequality uses the assumption $\delta \le \frac{1}{20}$ (numerically, $\frac{1}{4} - \frac{1}{3}\ln 2 \approx \frac{1}{53}$). This completes the proof of the second statement of the corollary. □

Finally, if we think of the Basic Invariance Principle as the nonlinear analogue of our Variant Berry–Esseen Theorem, it's natural to ask for the nonlinear analogue of the Berry–Esseen Theorem itself, i.e., a statement showing cdf-closeness of $F(\boldsymbol{x})$ and $F(\boldsymbol{g})$. It's straightforward to obtain a Lévy distance bound just as in the degree-1 case, Corollary 11.61; Exercise 11.44 asks you to show the following:

Corollary 11.70. *In the setting of Corollary 11.67 we have the Lévy distance bound $d_L(F(\boldsymbol{x}),F(\boldsymbol{y})) \le O(2^k\epsilon^{1/5})$. In the setting of Remark 11.66 we have the bound $d_L(F(\boldsymbol{x}),F(\boldsymbol{y})) \le (1/\rho)^{O(k)}\epsilon^{1/8}$.*

Suppose we now want actual cdf-closeness in the case that $\boldsymbol{y} \sim \mathrm{N}(0,1)^n$. In the degree-1 (Berry–Esseen) case we used the fact that degree-1 polynomials of independent Gaussians have good anticoncentration. The analogous statement for higher-degree polynomials of Gaussians is not so easy to prove; however, Carbery and Wright [**CW01**, Theorem 8] have obtained the following essentially optimal result:

Carbery–Wright Theorem. *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a polynomial (not necessarily multilinear) of degree at most k, let $\boldsymbol{g} \sim \mathrm{N}(0,1)^n$, and assume $\mathbf{E}[p(\boldsymbol{g})^2] = 1$. Then for all $\epsilon > 0$,*

$$\mathbf{Pr}[|p(\boldsymbol{g})| \le \epsilon] \le O(k\epsilon^{1/k}),$$

where the $O(\cdot)$ hides a universal constant.

Using this theorem it's not hard (see Exercise 11.45) to obtain:

Theorem 11.71. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be of degree at most k, with ϵ-small influences and* $\mathbf{Var}[f] = 1$. *Then for all $u \in \mathbb{R}$,*

$$|\mathbf{Pr}[f(\boldsymbol{x}) \leq u] - \mathbf{Pr}[f(\boldsymbol{g}) \leq u]| \leq O(k) \cdot \epsilon^{1/(4k+1)},$$

where the $O(\cdot)$ hides a universal constant.

11.7. Highlight: Majority Is Stablest Theorem

The Majority Is Stablest Theorem (to be proved at the end of this section) was originally conjectured in 2004 [**KKMO04, KKMO07**]. The motivation came from studying the approximability of the Max-Cut CSP. Recall that Max-Cut is perhaps the simplest possible constraint satisfaction problem: the domain of the variables is $\Omega = \{-1,1\}$ and the only constraint allowed is the binary non-equality predicate, $\neq : \{-1,1\}^2 \to \{0,1\}$. As we mentioned briefly in Chapter 7.3, Goemans and Williamson [**GW95**] gave a very sophisticated efficient algorithm using "semidefinite programming" which $(c_{\mathrm{GW}}\beta, \beta)$-approximates Max-Cut for every β, where $c_{\mathrm{GW}} \approx .8786$ is a certain trigonometric constant.

Turning to hardness of approximation, we know from Theorem 7.40 (developed in [**KKMO04**]) that to prove UG-hardness of $(\alpha+\delta, \beta-\delta)$-approximating Max-Cut, it suffices to construct an (α,β)-Dictator-vs.-No-Notables test which uses the predicate $\neq$. As we'll see in this section, the quality of the most natural such test can be easily inferred from the Majority Is Stablest Theorem. Assuming that theorem (as Khot et al. [**KKMO04**] did), we get a surprising conclusion: It's UG-hard to approximate the Max-Cut CSP any better than the Goemans–Williamson Algorithm does. In other words, the peculiar approximation guarantee of Goemans and Williamson on the very simple Max-Cut problem is optimal (assuming the Unique Games Conjecture).

Let's demystify this somewhat, starting with a description of the Goemans–Williamson Algorithm. Let $G = (V,E)$ be an n-vertex input graph for the algorithm; we'll write $(\boldsymbol{v},\boldsymbol{w}) \sim E$ to denote that $(\boldsymbol{v},\boldsymbol{w})$ is a uniformly random edge (i.e., $\neq$-constraint) in the graph. The first step of the Goemans–Williamson Algorithm is to solve following optimization problem:

$$\begin{array}{ll} \text{maximize} & \mathop{\mathbf{E}}_{(\boldsymbol{v},\boldsymbol{w})\sim E}\left[\frac{1}{2} - \frac{1}{2}\langle \vec{U}(\boldsymbol{v}), \vec{U}(\boldsymbol{w})\rangle\right] \\ \text{subject to} & \vec{U} : V \to S^{n-1}. \end{array} \tag{SDP}$$

Here S^{n-1} denotes the set of all unit vectors in $\mathbb{R}^n$. Somewhat surprisingly, since this optimization problem is a "semidefinite program" it can be solved in polynomial time using the Ellipsoid Algorithm. (Technically, it can only be solved up to any desired additive tolerance $\epsilon > 0$, but we'll ignore this point.) Let's write SDPOpt(G) for the optimum value of (SDP), and Opt(G) for the optimum Max-Cut value for G. We claim that (SDP) is a *relaxation* of the Max-Cut CSP on input G, and therefore

$$\mathrm{SDPOpt}(G) \geq \mathrm{Opt}(G).$$

To see this, simply note that if $F^* : V \to \{-1,1\}$ is an optimal assignment ("cut") for G then we can define $\vec{U}(v) = (F^*(v), 0, \dots, 0) \in S^{n-1}$ for each $v \in V$ and achieve the optimal cut value $\mathrm{Val}_G(F^*)$ in (SDP).

The second step of the Goemans–Williamson Algorithm might look familiar from Fact 11.7 and Remark 11.8. Let $\vec{U}^* : V \to S^{n-1}$ be the optimal solution for (SDP), achieving $\mathrm{SDPOpt}(G)$; abusing notation we'll write $\vec{U}^*(v) = \vec{v}$. The algorithm now chooses $\vec{\boldsymbol{g}} \sim \mathrm{N}(0,1)^n$ at random and outputs the assignment (cut) $\boldsymbol{F} : V \to \{-1,1\}$ defined by $\boldsymbol{F}(v) = \mathrm{sgn}(\langle \vec{v}, \vec{\boldsymbol{g}} \rangle)$. Let's analyze the (expected) quality of this assignment. The probability the algorithm's assignment $\boldsymbol{F}$ cuts a particular edge $(v,w) \in E$ is

$$\mathop{\mathbf{Pr}}_{\vec{\boldsymbol{g}} \sim \mathrm{N}(0,1)^n}[\mathrm{sgn}(\langle \vec{v}, \vec{\boldsymbol{g}} \rangle) \neq \mathrm{sgn}(\langle \vec{w}, \vec{\boldsymbol{g}} \rangle)].$$

This is precisely the probability that $\mathrm{sgn}(\boldsymbol{z}) \neq \mathrm{sgn}(\boldsymbol{z}')$ when $(\boldsymbol{z}, \boldsymbol{z}')$ is a pair of $\langle \vec{v}, \vec{w} \rangle$-correlated 1-dimensional Gaussians. Writing $\angle(\vec{v}, \vec{w}) \in [0, \pi]$ for the angle between the unit vectors $\vec{v}, \vec{w}$, we conclude from Sheppard's Formula (see (11.2)) that

$$\mathop{\mathbf{Pr}}_{\vec{\boldsymbol{g}}}[\boldsymbol{F} \text{ cuts edge } (v,w)] = \frac{\angle(\vec{v}, \vec{w})}{\pi}.$$

By linearity of expectation we can compute the expected value of the algorithm's assignment $\boldsymbol{F}$:

$$\mathop{\mathbf{E}}_{\vec{\boldsymbol{g}}}[\mathrm{Val}_G(\boldsymbol{F})] = \mathop{\mathbf{E}}_{(\boldsymbol{v},\boldsymbol{w}) \sim E}\left[\angle(\vec{\boldsymbol{v}}, \vec{\boldsymbol{w}})/\pi\right]. \tag{11.34}$$

On the other hand, by definition we have

$$\mathrm{SDPOpt}(G) = \mathop{\mathbf{E}}_{(\boldsymbol{v},\boldsymbol{w}) \sim E}\left[\tfrac{1}{2} - \tfrac{1}{2}\cos\angle(\vec{\boldsymbol{v}}, \vec{\boldsymbol{w}})\right]. \tag{11.35}$$

It remains to compare (11.34) and (11.35). Define

$$c_{\mathrm{GW}} = \min_{\theta \in [0,\pi]} \left\{ \frac{\theta/\pi}{\frac{1}{2} - \frac{1}{2}\cos\theta} \right\} \approx .8786. \tag{11.36}$$

Then from (11.34) and (11.35) we immediately get

$$\mathop{\mathbf{E}}_{\vec{\boldsymbol{g}}}[\mathrm{Val}_G(\boldsymbol{F})] \geq c_{\mathrm{GW}} \cdot \mathrm{SDPOpt}(G) \geq c_{\mathrm{GW}} \cdot \mathrm{Opt}(G);$$

i.e., in expectation the Goemans–Williamson Algorithm delivers a cut of value at least c_{GW} times the Max-Cut. In other words, it's a $(c_{\mathrm{GW}}\beta, \beta)$-approximation algorithm, as claimed. By being a little bit more careful about this analysis (Exercise 11.33) you can show following additional result:

Theorem 11.72. [**GW95**]. *Let $\theta \in [\theta^*, \pi]$, where $\theta^* \approx .74\pi$ is the minimizing θ in (11.36) (also definable as the positive solution of $\tan(\theta/2) = \theta$). Then on any graph G with $\mathrm{SDPOpt}(G) \geq \frac{1}{2} - \frac{1}{2}\cos\theta$, the Goemans–Williamson Algorithm produces a cut of (expected) value at least θ/π. In particular, the algorithm is a $(\theta/\pi, \frac{1}{2} - \frac{1}{2}\cos\theta)$-approximation algorithm for Max-Cut.*

Example 11.73. Consider the Max-Cut problem on the 5-vertex cycle graph $\mathbb{Z}_5$. The best bipartition of this graph cuts 4 out of the 5 edges; hence $\mathrm{Opt}(\mathbb{Z}_5) = \frac{4}{5}$. Exercise 11.32 asks you to show that taking

$$\vec{U}(v) = (\cos\tfrac{4\pi v}{5}, \sin\tfrac{4\pi v}{5}), \quad v \in \mathbb{Z}_5,$$

in the semidefinite program (SDP) establishes that $\mathrm{SDPOpt}(\mathbb{Z}_5) \geq \frac{1}{2} - \frac{1}{2}\cos\frac{4\pi}{5}$. (These are actually unit vectors in $\mathbb{R}^2$ rather than in $\mathbb{R}^5$ as (SDP) requires, but we can pad out the last three coordinates with zeroes.) This example shows that the Goemans–Williamson analysis in Theorem 11.72 lower-bounding $\mathrm{Opt}(G)$ in terms of $\mathrm{SDPOpt}(G)$ cannot be improved (at least when $\mathrm{SDPOpt}(G) = \frac{4}{5}$). This is termed an optimal *integrality gap*. In fact, Theorem 11.72 also implies that $\mathrm{SDPOpt}(\mathbb{Z}_5)$ must equal $\frac{1}{2} - \frac{1}{2}\cos\frac{4\pi}{5}$, for if it were greater, the theorem would falsely imply that $\mathrm{Opt}(\mathbb{Z}_5) > \frac{4}{5}$. Note that the Goemans–Williamson Algorithm actually finds the maximum cut when run on the cycle graph $\mathbb{Z}_5$. For a related example, see Exercise 11.35.

Now we explain the result of Khot et al. [**KKMO04**], that the Majority Is Stablest Theorem implies it's UG-hard to approximate Max-Cut better than the Goemans–Williamson Algorithm does:

Theorem 11.74. [**KKMO04**]. *Let $\theta \in (\frac{\pi}{2}, \pi)$. Then for any $\delta > 0$ it's UG-hard to $(\theta/\pi + \delta, \frac{1}{2} - \frac{1}{2}\cos\theta)$-approximate Max-Cut.*

Proof. It follows from Theorem 7.40 that we just need to construct a $(\theta/\pi, \frac{1}{2} - \frac{1}{2}\cos\theta)$-Dictator-vs.-No-Notables test using the predicate $\neq$. (See Exercise 11.36 for an extremely minor technical point.) It's very natural to try the following, with $\beta = \frac{1}{2} - \frac{1}{2}\cos\theta \in (\frac{1}{2}, 1)$:

***β*-Noise Sensitivity Test.** *Given query access to $f : \{-1,1\}^n \to \{-1,1\}$:*

- *Choose $\boldsymbol{x} \sim \{-1,1\}^n$ and form $\boldsymbol{x}'$ by reversing each bit of $\boldsymbol{x}$ independently with probability $\beta = \frac{1}{2} - \frac{1}{2}\cos\theta$. In other words let $(\boldsymbol{x}, \boldsymbol{x}')$ be a pair of $\cos\theta$-correlated strings. (Note that $\cos\theta < 0$.)*
- *Query f at $\boldsymbol{x}$, $\boldsymbol{x}'$.*
- *Accept if $f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$.*

By design,

$$\mathbf{Pr}[\text{the test accepts } f] = \mathbf{NS}_\beta[f] = \tfrac{1}{2} - \tfrac{1}{2}\mathbf{Stab}_{\cos\theta}[f]. \tag{11.37}$$

(We might also express this as "$\mathbf{RS}_f(\theta)$".) In particular, if f is a dictator, it's accepted with probability exactly $\beta = \frac{1}{2} - \frac{1}{2}\cos\theta$. To complete the proof that this is a $(\theta/\pi, \frac{1}{2} - \frac{1}{2}\cos\theta)$-Dictator-vs.-No-Notables test, let's suppose $f : \{-1,1\}^n \to [-1,1]$ has no (ϵ,ϵ)-notable coordinates and show that (11.37) is at most $\theta/\pi + o_\epsilon(1)$. (Regarding f having range $[-1,1]$, recall Remark 7.38.)

At first it might look like we can immediately apply the Majority Is Stablest Theorem; however, the theorem's inequality goes the "wrong way" and the correlation parameter $\rho = \cos\theta$ is negative. These two difficulties actually cancel each other out. Note that

$$\begin{aligned}
\mathbf{Pr}[\text{the test accepts } f] &= \tfrac{1}{2} - \tfrac{1}{2}\mathbf{Stab}_{\cos\theta}[f] \\
&= \tfrac{1}{2} - \tfrac{1}{2}\sum_{k=0}^{n}(\cos\theta)^k\mathbf{W}^k[f] \\
&\leq \tfrac{1}{2} + \tfrac{1}{2}\sum_{k \text{ odd}}(-\cos\theta)^k\mathbf{W}^k[f] \qquad (\text{since } \cos\theta < 0) \\
&= \tfrac{1}{2} + \tfrac{1}{2}\mathbf{Stab}_{-\cos\theta}[f^{\text{odd}}], \qquad (11.38)
\end{aligned}$$

where $f^{\text{odd}} : \{-1,1\}^n \to [-1,1]$ is the odd part of f (see Exercise 1.8) defined by

$$f^{\text{odd}}(x) = \tfrac{1}{2}(f(x) - f(-x)) = \sum_{|S| \text{ odd}} \widehat{f}(S)x^S.$$

Now we're really in a position to apply the Majority Is Stablest Theorem to f^{odd}, because $-\cos\theta \in (0,1)$, $\mathbf{E}[f^{\text{odd}}] = 0$, and f^{odd} has no (ϵ,ϵ)-notable coordinates (since it's formed from f by just dropping some terms in the Fourier expansion). Using $-\cos\theta = \cos(\pi - \theta)$, the result is that

$$\mathbf{Stab}_{-\cos\theta}[f^{\text{odd}}] \leq 1 - \tfrac{2}{\pi}\arccos(\cos(\pi-\theta)) + o_\epsilon(1) = 2\theta/\pi - 1 + o_\epsilon(1).$$

Putting this into (11.38) yields

$$\mathbf{Pr}[\text{the test accepts } f] \leq \tfrac{1}{2} + \tfrac{1}{2}(2\theta/\pi - 1 + o_\epsilon(1)) = \theta/\pi + o_\epsilon(1),$$

as needed. □

Remark 11.75. There's actually still a mismatch between the algorithmic guarantee of Theorem 11.72 and the UG-hardness result Theorem 11.74, concerning the case of $\theta \in (\frac{\pi}{2}, \theta^*)$. In fact, for these values of θ – i.e., $\frac{1}{2} \leq \beta \lessapprox .8446$ – *neither* result is sharp; see O'Donnell and Wu [**OW08**].

Remark 11.76. If we want to prove UG-hardness of $(\theta'/\pi + \delta, \frac{1}{2} - \frac{1}{2}\cos\theta')$-approximating Max-Cut, we don't need the full version of Borell's Isoperimetric Theorem; we only need the volume-$\frac{1}{2}$ case with parameter $\theta = \pi - \theta'$. Corollary 11.44 gave a simple proof of this result for $\theta = \frac{\pi}{4}$, hence $\theta' = \frac{3}{4}\pi$. This yields UG-hardness of $(\frac{3}{4} + \delta, \frac{1}{2} + \frac{1}{2\sqrt{2}})$-approximating Max-Cut. The ratio between α and β here is approximately .8787, very close to the Goemans–Williamson constant $c_{\text{GW}} \approx .8786$.

Finally, we will prove the General-Volume Majority Is Stablest Theorem, by using the Invariance Principle to reduce it to Borell's Isoperimetric Theorem.

General-Volume Majority Is Stablest Theorem. *Let* $f : \{-1,1\}^n \to [0,1]$. *Suppose that* $\mathbf{MaxInf}[f] \le \epsilon$, *or more generally, that* f *has no* $(\epsilon, \frac{1}{\log(1/\epsilon)})$*-notable coordinates. Then for any* $0 \le \rho < 1$,

$$\mathbf{Stab}_\rho[f] \le \Lambda_\rho(\mathbf{E}[f]) + O\big(\tfrac{\log\log(1/\epsilon)}{\log(1/\epsilon)}\big) \cdot \tfrac{1}{1-\rho}. \tag{11.39}$$

(Here the $O(\cdot)$ *bound has no dependence on* ρ*.)*

Proof. The proof involves using the Basic Invariance Principle twice (in the form of Corollary 11.69). To facilitate this we introduce $f' = \mathrm{T}_{1-\delta} f$, where (with foresight) we choose

$$\delta = 3\tfrac{\log\log(1/\epsilon)}{\log(1/\epsilon)}.$$

(We may assume ϵ is sufficiently small so that $0 < \delta \le \frac{1}{20}$.) Note that $\mathbf{E}[f'] = \mathbf{E}[f]$ and that

$$\mathbf{Stab}_\rho[f'] = \sum_{S \subseteq [n]} \rho^{|S|}(1-\delta)^{2|S|}\widehat{f}(S)^2 = \mathbf{Stab}_{\rho(1-\delta)^2}[f].$$

But

$$\big|\mathbf{Stab}_{\rho(1-\delta)^2}[f] - \mathbf{Stab}_\rho[f]\big| \le (\rho - \rho(1-\delta)^2) \cdot \tfrac{1}{1-\rho} \cdot \mathbf{Var}[f] \le 2\delta \cdot \tfrac{1}{1-\rho} \tag{11.40}$$

by Exercise 2.46, and with our choice of δ this can be absorbed into the error of (11.39). Thus it suffices to prove (11.39) with f' in place of f.

Let $\mathrm{Sq} : \mathbb{R} \to \mathbb{R}$ be the continuous function which agrees with $t \mapsto t^2$ for $t \in [0,1]$ and is constant outside $[0,1]$. Note that Sq is 2-Lipschitz. We will apply the second part of Corollary 11.69 with "h" set to $\mathrm{T}_{\sqrt{\rho}} f$ (and thus $\mathrm{T}_{1-\delta} h = \mathrm{T}_{\sqrt{\rho}} f'$). This is valid since the variance and $(1-\delta)$-stable influences of h are only smaller than those of f. Thus

$$\Big|\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[\mathrm{Sq}(\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{x}))] - \mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{Sq}(\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{g}))]\Big| \le O(\epsilon^{\delta/3}) = O\big(\tfrac{1}{\log(1/\epsilon)}\big), \tag{11.41}$$

using our choice of δ. (In fact, it's trading off this error with (11.40) that led to our choice of δ.) Now $\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{x}) = \mathrm{T}_{(1-\delta)\sqrt{\rho}} f(\boldsymbol{x})$ is always bounded in $[0,1]$, so

$$\mathrm{Sq}(\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{x})) = (\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{x}))^2 \quad \Longrightarrow \quad \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[\mathrm{Sq}(\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{x}))] = \mathbf{Stab}_\rho[f'].$$

Furthermore, $\mathrm{T}_{\sqrt{\rho}} f'(\boldsymbol{g})$ is the same as $\mathrm{U}_{\sqrt{\rho}} f'(\boldsymbol{g})$ because f' is a multilinear polynomial. (Both are equal to $f'(\rho \boldsymbol{g})$; see Fact 11.13.) Thus in light of (11.41), to complete the proof of (11.39) it suffices to show

$$\Big|\mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} f'(\boldsymbol{g}))] - \Lambda_\rho(\mathbf{E}[f'])\Big| \le O\big(\tfrac{1}{\log(1/\epsilon)}\big). \tag{11.42}$$

Define the function $F : \mathbb{R}^n \to [0,1]$ by

$$F(g) = \mathrm{trunc}_{[0,1]}(f'(g)) = \begin{cases} 0 & \text{if } f'(g) < 0, \\ f'(g) & \text{if } f'(g) \in [0,1], \\ 1 & \text{if } f'(g) > 1. \end{cases}$$

We will establish the following two inequalities, which together imply (11.42):

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} f'(\boldsymbol{g}))] - \mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} F(\boldsymbol{g}))] \right| \le O\left(\tfrac{1}{\log(1/\epsilon)}\right), \tag{11.43}$$

$$\mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} F(\boldsymbol{g}))] \le \Lambda_\rho(\mathbf{E}[f']) + O\left(\tfrac{1}{\log(1/\epsilon)}\right). \tag{11.44}$$

Both of these inequalities will in turn follow from

$$\mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[|f'(\boldsymbol{g}) - F(\boldsymbol{g})|] = \mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{dist}_{[0,1]}(f'(\boldsymbol{g}))] \le O\left(\tfrac{1}{\log(1/\epsilon)}\right). \tag{11.45}$$

Let's show how (11.43) and (11.44) follow from (11.45), leaving the proof of (11.45) to the end. For (11.43),

$$\begin{aligned} \left| \mathbf{E}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} f'(\boldsymbol{g}))] - \mathbf{E}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} F(\boldsymbol{g}))] \right| &\le 2\,\mathbf{E}[|\mathrm{U}_{\sqrt{\rho}} f'(\boldsymbol{g}) - \mathrm{U}_{\sqrt{\rho}} F(\boldsymbol{g})|] \\ &\le 2\,\mathbf{E}[|f'(\boldsymbol{g}) - F(\boldsymbol{g})|] \le O\left(\tfrac{1}{\log(1/\epsilon)}\right), \end{aligned}$$

where the first inequality used that Sq is 2-Lipschitz, the second inequality used the fact that $\mathrm{U}_{\sqrt{\rho}}$ is a contraction on $L^1(\mathbb{R}^n, \gamma)$, and the third inequality was (11.45). As for (11.44), $\mathrm{U}_{\sqrt{\rho}} F$ is bounded in $[0,1]$ since F is. Thus

$$\mathbf{E}[\mathrm{Sq}(\mathrm{U}_{\sqrt{\rho}} F(\boldsymbol{g}))] = \mathbf{E}[(\mathrm{U}_{\sqrt{\rho}} F(\boldsymbol{g}))^2] = \mathbf{Stab}_\rho[F] \le \Lambda_\rho(\mathbf{E}[F(\boldsymbol{g})]),$$

where we used Borell's Isoperimetric Theorem. But $|\mathbf{E}[F(\boldsymbol{g})] - \mathbf{E}[f'(\boldsymbol{g})]| \le O\left(\frac{1}{\log(1/\epsilon)}\right)$ by (11.45), and Λ_ρ is easily shown to be 2-Lipschitz (Exercise 11.19(*e*)). This establishes (11.44).

It therefore remains to show (11.45), which we do by applying the Invariance Principle one more time. Taking ψ to be the 1-Lipschitz function $\mathrm{dist}_{[0,1]}$ in Corollary 11.69 we deduce

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{g} \sim \mathrm{N}(0,1)^n}[\mathrm{dist}_{[0,1]}(f'(\boldsymbol{g}))] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[\mathrm{dist}_{[0,1]}(f'(\boldsymbol{x}))] \right| \le O(\epsilon^{\delta/3}) = O\left(\tfrac{1}{\log(1/\epsilon)}\right).$$

But $\mathbf{E}[\mathrm{dist}_{[0,1]} f'(\boldsymbol{x})] = 0$ since $f'(\boldsymbol{x}) = \mathrm{T}_{1-\delta} f(\boldsymbol{x}) \in [0,1]$ always. This establishes (11.45) and completes the proof. □

We conclude with one more application of the Majority Is Stablest Theorem. Recall Kalai's version of Arrow's Theorem from Chapter 2.5, i.e., Theorem 2.56. It states that in a 3-candidate Condorcet election using the voting rule $f : \{-1,1\}^n \to \{-1,1\}$, the probability of having a Condorcet winner – often called a *rational outcome* – is precisely $\frac{3}{4} - \frac{3}{4}\mathbf{Stab}_{-1/3}[f]$. As we saw in the proof of Theorem 11.74 near (11.38), this is in turn at most $\frac{3}{4} + \frac{3}{4}\mathbf{Stab}_{1/3}[f^{\mathrm{odd}}]$, with equality if f is already odd. It follows from the Majority Is Stablest

Theorem that among all voting rules with ϵ-small influences (a condition all reasonable voting rules should satisfy), majority rule is the "most rational". Thus we see that the principle of representative democracy can be derived using analysis of Boolean functions.

11.8. Exercises and notes

11.1 Let $\mathscr{A}$ be the set of all functions $f : \mathbb{R}^n \to \mathbb{R}$ which are finite linear combinations of indicator functions of boxes. Prove that $\mathscr{A}$ is dense in $L^1(\mathbb{R}^n, \gamma)$.

11.2 Fill in proof details for the Gaussian Hypercontractivity Theorem.

11.3 Prove Fact 11.13. (Cf. Exercise 2.25.)

11.4 Show that $\mathrm{U}_{\rho_1}\mathrm{U}_{\rho_2} = \mathrm{U}_{\rho_1\rho_2}$ for all $\rho_1, \rho_2 \in [-1,1]$. (Cf. Exercise 2.32.)

11.5 Prove Proposition 11.16. (Hint: For $\rho \neq 0$, write $g(z) = \mathrm{U}_\rho f(z)$ and show that $g(z/\rho)$ is a smooth function using the relationship between convolution and derivatives.)

11.6 (*a*) Prove Proposition 11.17. (Hint: First prove it for bounded continuous f; then make an approximation and use Proposition 11.15.)

(*b*) Deduce more generally that for $f \in L^1(\mathbb{R}^n, \gamma)$ the map $\rho \mapsto \mathrm{U}_\rho f$ is "strongly continuous" on $[0,1]$, meaning that for any $\rho \in [0,1]$ we have $\|\mathrm{U}_{\rho'} f - \mathrm{U}_\rho f\|_1 \to 0$ as $\rho' \to \rho$. (Hint: Use Exercise 11.4.)

11.7 Complete the proof of Proposition 11.26 by establishing the case of general n.

11.8 Complete the proof of Proposition 11.28 by establishing the case of general n.

11.9 (*a*) Establish the alternative formula (11.10) for the probabilists' Hermite polynomials $H_j(z)$ given in Definition 11.29; equivalently, establish the formula

$$H_j(z) = (-1)^j \exp(\tfrac{1}{2}z^2) \cdot \left(\frac{d}{dz}\right)^j \exp(-\tfrac{1}{2}z^2).$$

(Hint: Complete the square on the left-hand side of (11.8); then differentiate j times with respect to t and evaluate at 0.)

(*b*) Establish the recursion

$$H_j(z) = (z - \tfrac{d}{dz})H_{j-1}(z) \qquad \Longleftrightarrow \qquad h_j(z) = \tfrac{1}{\sqrt{j}} \cdot (z - \tfrac{d}{dz})h_{j-1}(z)$$

for $j \in \mathbb{N}^+$, and hence the formula $H_j(z) = (z - \frac{d}{dz})^j 1$.

(*c*) Show that $h_j(z)$ is an odd function of z if j is odd and an even function of z if j is even.

11.10 (*a*) Establish the derivative formula for Hermite polynomials:

$$H_j'(z) = j \cdot H_{j-1}(z) \qquad \Longleftrightarrow \qquad h_j'(z) = \sqrt{j} \cdot h_{j-1}(z).$$

(*b*) By combining this with the other formula for $H'_j(z)$ implicit in Exercise 11.9(*b*), deduce the recursion

$$H_{j+1}(z) = zH_j(z) - jH_{j-1}(z).$$

(*c*) Show that $H_j(z)$ satisfies the second-order differential equation

$$jH_j(z) = zH'_j(z) - H''_j(z).$$

(It's equivalent to say that $h_j(z)$ satisfies it.) Observe that this is consistent with Propositions 11.26 and 11.40 and says that H_j (equivalently, h_j) is an eigenfunction of the Ornstein–Uhlenbeck operator L, with eigenvalue j.

11.11 Prove that

$$H_j(x+y) = \sum_{k=0}^{j} \binom{j}{k} x^{j-k} H_k(y).$$

11.12 (*a*) By equating both sides of (11.8) with

$$\mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)}[\exp(t(z+i\boldsymbol{g}))]$$

(where $i = \sqrt{-1}$), show that

$$H_j(z) = \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)}[(z+i\boldsymbol{g})^j].$$

(*b*) Establish the explicit formulas

$$H_j(z) = \sum_{k=0}^{\lfloor j/2 \rfloor} (-1)^k \binom{j}{2k} \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)}[\boldsymbol{g}^{2k}] z^{j-2k}$$
$$= j! \cdot \left(\frac{z^j}{0!! \cdot j!} - \frac{z^{j-2}}{2!! \cdot (j-2)!} + \frac{z^{j-4}}{4!! \cdot (j-4)!} - \frac{z^{j-6}}{6!! \cdot (j-6)!} + \cdots \right).$$

11.13 (*a*) Establish the formula

$$\mathbf{E}[\|\nabla f\|^2] = \sum_{\alpha\in\mathbb{N}^n} |\alpha| \widehat{f}(\alpha)^2$$

for all $f \in L^2(\mathbb{R}^n, \gamma)$ (or at least for all n-variate polynomials f).

(*b*) For $f \in L^2(\mathbb{R}^n, \gamma)$, establish the formula

$$\sum_{i=1}^{n} \mathbf{E}[\mathop{\mathbf{Var}}_{\boldsymbol{z}_i}[f]] = \sum_{\alpha\in\mathbb{N}^n} (\#\alpha) \widehat{f}(\alpha)^2.$$

11.14 Show that for all $j \in \mathbb{N}$ and all $z \in \mathbb{R}$ we have

$$\tbinom{n}{j}^{-1/2} \cdot K_j^{(n)}\left(\frac{n}{2} - z\frac{\sqrt{n}}{2}\right) \xrightarrow{n\to\infty} h_j(z),$$

where $K_j^{(n)}$ is the Kravchuk polynomial of degree j from Exercise 5.28 (with its dependence on n indicated in the superscript).

11.15 Recall the definition (11.13) of the Gaussian Minkowski content of the boundary ∂A of a set $A \subseteq \mathbb{R}^n$. Sometimes the following very similar definition is also proposed for the Gaussian surface area of A:

$$M(A) = \liminf_{\epsilon \to 0^+} \frac{\mathrm{vol}_\gamma(\{z : \mathrm{dist}(z, A) < \epsilon\}) - \mathrm{vol}_\gamma(A)}{\epsilon}.$$

Consider the following subsets of $\mathbb{R}$:

$$A_1 = \emptyset, \quad A_2 = \{0\}, \quad A_3 = (-\infty, 0), \quad A_4 = (-\infty, 0], \quad A_5 = \mathbb{R} \setminus \{0\}, \quad A_6 = \mathbb{R}.$$

(*a*) Show that

$$\begin{array}{lll}
\gamma^+(A_1) = 0 & M(A_1) = 0 & \mathrm{surf}_\gamma(A_1) = 0 \\
\gamma^+(A_2) = \frac{1}{\sqrt{2\pi}} & M(A_2) = \sqrt{\frac{2}{\pi}} & \mathrm{surf}_\gamma(A_2) = 0 \\
\gamma^+(A_3) = \frac{1}{\sqrt{2\pi}} & M(A_3) = \frac{1}{\sqrt{2\pi}} & \mathrm{surf}_\gamma(A_3) = \frac{1}{\sqrt{2\pi}} \\
\gamma^+(A_4) = \frac{1}{\sqrt{2\pi}} & M(A_4) = \frac{1}{\sqrt{2\pi}} & \mathrm{surf}_\gamma(A_4) = \frac{1}{\sqrt{2\pi}} \\
\gamma^+(A_5) = \frac{1}{\sqrt{2\pi}} & M(A_5) = 0 & \mathrm{surf}_\gamma(A_5) = 0 \\
\gamma^+(A_6) = 0 & M(A_6) = 0 & \mathrm{surf}_\gamma(A_6) = 0.
\end{array}$$

(*b*) For $A \subseteq \mathbb{R}^n$, the *essential boundary* (or *measure-theoretic boundary*) of A is defined to be

$$\partial_* A = \left\{ x \in \mathbb{R}^n : \lim_{\delta \to 0^+} \frac{\mathrm{vol}_\gamma(A \cap B_\delta(x))}{\mathrm{vol}_\gamma(B_\delta(x))} \neq 0, 1 \right\},$$

where $B_\delta(x)$ denotes the ball of radius δ centered at x. In other words, $\partial_* A$ is the set of points where the "local density of A" is strictly between 0 and 1. Show that if we replace ∂A with $\partial_* A$ in the definition (11.13) of the Gaussian Minkowski content of the boundary of A, then we have the identity $\gamma^+(\partial_* A_i) = \mathrm{surf}_\gamma(A_i)$ for all $1 \leq i \leq 6$. Remark: In fact, the equality $\gamma^+(\partial_* A) = \mathrm{surf}_\gamma(A)$ is known to hold for every set A such that $\partial_* A$ is "rectifiable".

11.16 Justify the formula for the Gaussian surface area of unions of intervals stated in Example 11.50.

11.17 (*a*) Let $B_r \subset \mathbb{R}^n$ denote the ball of radius $r > 0$ centered at the origin. Show that

$$\mathrm{surf}_\gamma(B_r) = \frac{n}{2^{n/2}(n/2)!} r^{n-1} e^{-r^2/2}. \tag{11.46}$$

(*b*) Show that (11.46) is maximized when $r = \sqrt{n-1}$. (In case $n = 1$, this should be interpreted as $r \to 0^+$.)

(*c*) Let $S(n)$ denote this maximizing value, i.e., the value of (11.46) with $r = \sqrt{n-1}$. Show that $S(n)$ decreases from $\sqrt{\frac{2}{\pi}}$ to a limit of $\frac{1}{\sqrt{\pi}}$ as n increases from 1 to ∞.

11.18 (*a*) For $f \in L^2(\mathbb{R}^n,\gamma)$, show that $\mathrm{L}f$ is defined, i.e.,

$$\lim_{t\to 0} \frac{f - \mathrm{U}_{e^{-t}} f}{t}$$

exists in $L^2(\mathbb{R}^n,\gamma)$, if and only if $\sum_{\alpha\in\mathbb{N}^n} |\alpha|^2 \widehat{f}(\alpha)^2 < \infty$. (Hint: Proposition 11.37.)

(*b*) Formally justify Proposition 11.40.

(*c*) Let $f \in L^2(\mathbb{R}^n,\gamma)$. Show that $\mathrm{U}_\rho f$ is in the domain of L for any $\rho \in (-1,1)$.

Remark: It can be shown that the $\mathscr{C}^3$ hypothesis in Propositions 11.26 and 11.28 is not necessary (provided the derivatives are interpreted in the distributional sense); see, e.g., Bogachev [**Bog98**, Chapter 1] for more details.

11.19 This exercise is concerned with (a generalization of) the function appearing in Borell's Isoperimetric Theorem.

Definition 11.77. For $\rho \in [-1,1]$ we define the *Gaussian quadrant probability* function $\Lambda_\rho : [0,1]^2 \to [0,1]$ by

$$\Lambda_\rho(\alpha,\beta) = \mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated} \\ \text{standard Gaussians}}}[\boldsymbol{z} \le t, \boldsymbol{z}' \le t'],$$

where t and t' are defined by $\Phi(t) = \alpha$, $\Phi(t') = \beta$. This is a slight reparametrization of the bivariate Gaussian cdf. We also use the shorthand notation

$$\Lambda_\rho(\alpha) = \Lambda_\rho(\alpha,\alpha),$$

which we encountered in Borell's Isoperimetric Theorem (and also in Exercises 5.32 and 9.24, with a different, but equivalent, definition).

(*a*) Confirm the statement from Borell's Isoperimetric Theorem, that for every $H \subseteq \mathbb{R}^n$ with $\mathrm{vol}_\gamma(H) = \alpha$ we have $\mathbf{Stab}_\rho[1_H] = \Lambda_\rho(\alpha)$.

(*b*) Verify the following formulas:

$$\begin{aligned}
\Lambda_\rho(\alpha,\beta) &= \Lambda_\rho(\beta,\alpha),\\
\Lambda_0(\alpha,\beta) &= \alpha\beta,\\
\Lambda_1(\alpha,\beta) &= \min(\alpha,\beta),\\
\Lambda_{-1}(\alpha,\beta) &= \max(\alpha+\beta-1,0),\\
\Lambda_\rho(\alpha,0) &= \Lambda_\rho(0,\alpha) = 0,\\
\Lambda_\rho(\alpha,1) &= \Lambda_\rho(1,\alpha) = \alpha,\\
\Lambda_{-\rho}(\alpha,\beta) &= \alpha - \Lambda_\rho(\alpha,1-\beta) = \beta - \Lambda_\rho(1-\alpha,\beta),\\
\Lambda_\rho(\tfrac12,\tfrac12) &= \tfrac12 - \tfrac12\frac{\arccos\rho}{\pi}.
\end{aligned}$$

(*c*) Prove that $\Lambda_\rho(\alpha,\beta) \gtrless \alpha\beta$ according as $\rho \gtrless 0$, for all $0 < \alpha,\beta < 1$.

(*d*) Establish

$$\frac{d}{d\alpha}\Lambda_\rho(\alpha,\beta) = \Phi\left(\frac{t'-\rho t}{\sqrt{1-\rho^2}}\right), \quad \frac{d}{d\beta}\Lambda_\rho(\alpha,\beta) = \Phi\left(\frac{t-\rho t'}{\sqrt{1-\rho^2}}\right),$$

where $t = \Phi^{-1}(\alpha)$, $t' = \Phi^{-1}(\beta)$ as usual.

(*e*) Show that

$$|\Lambda_\rho(\alpha,\beta) - \Lambda_\rho(\alpha',\beta')| \le |\alpha-\alpha'| + |\beta-\beta'|,$$

and hence $\Lambda_\rho(\alpha)$ is a 2-Lipschitz function of α.

11.20 Show that the general-n case of Bobkov's Inequality follows by induction from the $n = 1$ case.

11.21 Let $f : \{-1,1\}^n \to \{-1,1\}$ and let $\alpha = \min\{\mathbf{Pr}[f = 1], \mathbf{Pr}[f = -1]\}$. Deduce $\mathbf{I}[f] \ge 4\mathscr{U}(\alpha)^2$ from Bobkov's Inequality. Show that this recovers the edge-isoperimetric inequality for the Boolean cube (Theorem 2.39) up to a constant factor. (Hint: For the latter problem, use Proposition 5.27.)

11.22 Let $d_1, d_2 \in \mathbb{N}$. Suppose we take a simple random walk on $\mathbb{Z}$, starting from the origin and moving by ±1 at each step with equal probability. Show that the expected time it takes to first reach either $-d_1$ or $+d_2$ is $d_1 d_2$.

11.23 Prove Claim 11.54. (Hint: For the function $V_y(\tau)$ appearing in the proof of Bobkov's Two-Point Inequality, you'll want to establish that $V_y'''(0) = 0$ and that $V_y''''(0) = \frac{2+10\mathscr{U}'(y)^2}{\mathscr{U}(y)^3} > 0$.)

11.24 Prove Theorem 11.55. (Hint: Have the random walk start at $\boldsymbol{y}_0 = a \pm \rho b$ with equal probability, and define $\boldsymbol{z}_t = \|(\mathscr{U}(\boldsymbol{y}_t), \rho b, \tau\sqrt{t})\|$. You'll need the full generality of Exercise 11.22.)

11.25 Justify Remark 11.41 (in the general-volume context) by showing that Borell's Isoperimetric Theorem for all functions in $K = \{f : \mathbb{R}^n \to [0,1] \mid \mathbf{E}[f] = \alpha\}$ can be deduced from the case of functions in $\partial K = \{f : \mathbb{R}^n \to \{0,1\} \mid \mathbf{E}[f] = \alpha\}$. (Hint: As stated in the remark, the intuition is that $\sqrt{\mathbf{Stab}_\rho[f]}$ is a norm and that K is a convex set whose extreme points are ∂K. To make this precise, you may want to use Exercise 11.1.)

11.26 The goal of this exercise and Exercises 11.27–11.29 is to give the proof of Borell's Isoperimetric Theorem due to Mossel and Neeman [**MN12**]. In fact, their proof gives the following natural "two-set" generalization of the theorem (Borell's original work [**Bor85**] proved something even more general):

Two-Set Borell Isoperimetric Theorem. *Fix $\rho \in (0,1)$ and $\alpha, \beta \in [0,1]$. Then for any $A, B \subseteq \mathbb{R}^n$ with $\mathrm{vol}_\gamma(A) = \alpha$, $\mathrm{vol}_\gamma(B) = \beta$,*

$$\mathop{\mathbf{Pr}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated} \\ n\text{-dimensional Gaussians}}}[\boldsymbol{z} \in A, \boldsymbol{z}' \in B] \le \Lambda_\rho(\alpha,\beta). \tag{11.47}$$

By definition of $\Lambda_\rho(\alpha,\beta)$, equality holds if A and B are parallel halfspaces. Taking $\beta=\alpha$ and $B=A$ in this theorem gives Borell's Isoperimetric Theorem as stated in Section 11.3 (in the case of range $\{0,1\}$, at least, which is equivalent by Exercise 11.25). It's quite natural to guess that parallel halfspaces should maximize the "joint Gaussian noise stability" quantity on the left of (11.47), especially in light of Remark 10.2 from Chapter 10.1 concerning the analogous Generalized Small-Set Expansion Theorem. Just as our proof of the Small-Set Expansion Theorem passed through the Two-Function Hypercontracitivity Theorem to facilitate induction, so too does the Mossel–Neeman proof pass through the following "two-function version" of Borell's Isoperimetric Theorem:

Two-Function Borell Isoperimetric Theorem. *Fix* $\rho\in(0,1)$ *and let* $f,g\in L^2(\mathbb{R}^n,\gamma)$ *have range* $[0,1]$. *Then*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated}\\ n\text{-dimensional Gaussians}}}[\Lambda_\rho(f(\boldsymbol{z}),g(\boldsymbol{z}'))]\leq\Lambda_\rho(\mathbf{E}[f],\mathbf{E}[g]).$$

(*a*) Show that the Two-Function Borell Isoperimetric Theorem implies the Two-Set Borell Isoperimetric Theorem and the Borell Isoperimetric Theorem (for functions with range $[0,1]$). (Hint: You may want to use facts from Exercise 11.19.)

(*b*) Show conversely that the Two-Function Borell Isoperimetric Theorem (in dimension n) is implied by the Two-Set Borell Isoperimetric Theorem (in dimension $n+1$). (Hint: Given $f:\mathbb{R}^n\to[0,1]$, define $A\subseteq\mathbb{R}^{n+1}$ by $(z,t)\in A \iff f(z)\geq\Phi(t)$.)

(*c*) Let $\ell_1,\ell_2:\mathbb{R}^n\to\mathbb{R}$ be defined by $\ell_i(z)=\langle a,z\rangle+b_i$ for some $a\in\mathbb{R}^n$, $b_1,b_2\in\mathbb{R}$. Show that equality occurs in the Two-Function Borell Isoperimetric Theorem if $f(z)=1_{\ell_1(z)\geq 0}$, $g(z)=1_{\ell_2(z)\geq 0}$ or if $f(z)=\Phi(\ell_1(z))$, $g(z)=\Phi(\ell_2(z))$.

11.27 Show that the inequality in the Two-Function Borell Isoperimetric Theorem "tensorizes" in the sense that if it holds for $n=1$, then it holds for all n. Your proof should not use any property of the function Λ_ρ, nor any property of the ρ-correlated n-dimensional Gaussian distribution besides the fact that it's a product distribution. (Hint: Induction by restrictions as in the proof of the Two-Function Hypercontractivity Induction Theorem from Chapter 9.4.)

11.28 Let $I_1,I_2\subseteq\mathbb{R}$ be open intervals and let $\mathscr{F}:I_1\times I_2\to\mathbb{R}$ be $\mathscr{C}^2$. For $\rho\in\mathbb{R}$, define the matrix

$$H_\rho\mathscr{F}=(H\mathscr{F})\circ\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix},$$

where $H\mathscr{F}$ denotes the Hessian of $\mathscr{F}$ and $\circ$ is the entrywise (Hadamard) product. We say that $\mathscr{F}$ is *ρ-concave* (terminology introduced by Ledoux [Led13]) if $H_\rho \mathscr{F}$ is everywhere negative semidefinite. Note that the $\rho = 1$ case corresponds to the usual notion of concavity, and the $\rho = 0$ case corresponds to concavity separately along the two coordinates. The goal of this exercise is to show that the Gaussian quadrant probability Λ_ρ function is ρ-concave for all $\rho \in (0,1)$.

(*a*) Extending Exercise 11.19(*d*), show that for any $\rho \in (-1,1)$,

$$\frac{d^2}{d\alpha^2}\Lambda_\rho(\alpha,\beta) = -\frac{\rho}{\sqrt{1-\rho^2}} \cdot \frac{1}{\phi(t)} \cdot \phi\left(\frac{t' - \rho t}{\sqrt{1-\rho^2}}\right),$$

and deduce a similar formula for $\frac{d^2}{d\beta^2}\Lambda_\rho(\alpha,\beta)$.

(*b*) Show that

$$\frac{d^2}{d\alpha\, d\beta}\Lambda_\rho(\alpha,\beta) = \frac{1}{\sqrt{1-\rho^2}} \cdot \frac{1}{\phi(t')} \cdot \phi\left(\frac{t' - \rho t}{\sqrt{1-\rho^2}}\right),$$

and deduce a similar (in fact, equal) formula for $\frac{d^2}{d\beta\, d\alpha}\Lambda_\rho(\alpha,\beta)$.

(*c*) Show that $\det(H_\rho \Lambda_\rho) = 0$ on all of $(0,1)^2$.

(*d*) Show that if $\rho \in (0,1)$, then $\frac{d^2}{d\alpha^2}\Lambda_\rho$, $\frac{d^2}{d\beta^2}\Lambda_\rho < 0$ on $(0,1)^2$. Deduce that Λ_ρ is ρ-concave.

11.29 This exercise is devoted to Mossel and Neeman's proof [MN12] of the Two-Function Borell Isoperimetric Theorem in the case $n = 1$. For another approach, see Exercise 11.30. By Exercise 11.27, this is sufficient to establish the case of general n. (Actually, the proof in this exercise works essentially verbatim in the general n case, but we stick to $n = 1$ for simplicity.)

(*a*) More generally, we intend to prove that for $f, g : \mathbb{R} \to [0,1]$,

$$\lambda(\rho) = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated} \\ \text{standard Gaussians}}}[\Lambda_\rho(\mathrm{U}_\rho f(\boldsymbol{z}), \mathrm{U}_\rho g(\boldsymbol{z}'))]$$

is a nonincreasing function of $0 < \rho < 1$ (cf. Theorem 11.55). Obtain the desired conclusion by taking $\rho \to 0^+, 1^-$. (Hint: You'll need Exercises 11.6 and 11.19(*e*).)

(*b*) Write $f_\rho = \mathrm{U}_\rho f$, $g_\rho = \mathrm{U}_\rho g$ for brevity, and write $\partial_i \Lambda_\rho$ ($i = 1,2$) for the partial derivatives of Λ_ρ. Also let $\boldsymbol{h}_1, \boldsymbol{h}_2$ denote independent standard Gaussians. Use the Chain Rule and Proposition 11.27 to establish

$$\begin{aligned} \lambda'(\rho) = &\ \mathbf{E}[(\partial_1\Lambda_\rho)(f_\rho(\boldsymbol{h}_1), g_\rho(\rho\boldsymbol{h}_1 + \sqrt{1-\rho^2}\boldsymbol{h}_2)) \cdot \mathrm{L}f_\rho(\boldsymbol{h}_1)] && (11.48) \\ &+ \mathbf{E}[(\partial_2\Lambda_\rho)(f_\rho(\rho\boldsymbol{h}_2 + \sqrt{1-\rho^2}\boldsymbol{h}_1), g_\rho(\boldsymbol{h}_2)) \cdot \mathrm{L}g_\rho(\boldsymbol{h}_2)]. && (11.49) \end{aligned}$$

(*c*) Use Proposition 11.28 to show that the first expectation (11.48) equals

$$\mathbf{E}[(\partial_{11}\Lambda_\rho f)(f_\rho, g_\rho) \cdot (f'_\rho)^2 + \rho \cdot (\partial_{21}\Lambda_\rho f)(f_\rho, g_\rho) \cdot f'_\rho \cdot g'_\rho],$$

where f_ρ, f'_ρ are evaluated at $\boldsymbol{h}_1$ and g_ρ, g'_ρ are evaluated at $\rho \boldsymbol{h}_1 + \sqrt{1-\rho^2}\boldsymbol{h}_2$. Give a similar formula for (11.49).

(*d*) Deduce that

$$\lambda'(\rho) = \mathop{\mathbf{E}}_{\substack{(\boldsymbol{z},\boldsymbol{z}')\ \rho\text{-correlated} \\ \text{standard Gaussians}}}\left[\begin{bmatrix} f'_\rho(\boldsymbol{z}) & g'_\rho(\boldsymbol{z}')\end{bmatrix} \cdot (H_\rho \Lambda_\rho)(f_\rho(\boldsymbol{z}), g_\rho(\boldsymbol{z}')) \cdot \begin{bmatrix} f'_\rho(\boldsymbol{z}) \\ g'_\rho(\boldsymbol{z}')\end{bmatrix}\right],$$

where H_ρ is as in Exercise 11.28, and that indeed λ is a nonincreasing function.

11.30 (*a*) Suppose the Two-Function Borell Isoperimetric Theorem were to hold for 1-bit functions, i.e., for $f, g : \{-1,1\} \to [0,1]$. Then the easy induction of Exercise 11.27 would extend the result to n-bit functions $f, g : \{-1,1\}^n \to [0,1]$; in turn, this would yield the Two-Function Borell Isoperimetric Theorem for 1-dimensional Gaussian functions (i.e., Exercise 11.29), by the usual Central Limit Theorem argument. Show, however, that dictator functions provide a counterexample to a potential "1-bit Two-Function Borell Isoperimetric Theorem".

(*b*) Nevertheless, the idea can be salvaged by proving a weakened version of the inequality for 1-bit functions that has an "error term" that is a *superlinear* function of f and g's "influences". Fix $\rho \in (0,1)$ and some small $\epsilon > 0$. Let $f, g : \{-1,1\} \to [\epsilon, 1-\epsilon]$. Show that

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{x}') \\ \rho\text{-correlated}}}[\Lambda_\rho(f(\boldsymbol{x}), g(\boldsymbol{x}'))] \leq \Lambda_\rho(\mathbf{E}[f], \mathbf{E}[g]) + C_{\rho,\epsilon} \cdot (\mathbf{E}[|\mathrm{D}_1 f|^3] + \mathbf{E}[|\mathrm{D}_1 g|^3]),$$

where $C_{\rho,\epsilon}$ is a constant depending only on ρ and ϵ. (Hint: Perform a 2nd-order Taylor expansion of Λ_ρ around $(\mathbf{E}[f], \mathbf{E}[g])$; in expectation, the quadratic term should be

$$\begin{bmatrix} \mathrm{D}_1 f & \mathrm{D}_1 g\end{bmatrix} \cdot (H_\rho \Lambda_\rho)(\mathbf{E}[f], \mathbf{E}[g]) \cdot \begin{bmatrix} \mathrm{D}_1 f \\ \mathrm{D}_1 g\end{bmatrix}.$$

As in Exercise 11.29, show this quantity is nonpositive.)

(*c*) Extend the previous result by induction to obtain the following theorem of De, Mossel, and Neeman [DMN13]:

Theorem 11.78. *For each $\rho \in (0,1)$ and $\epsilon > 0$, there exists a constant $C_{\rho,\epsilon}$ such that the following holds: If $f, g : \{-1,1\}^n \to [\epsilon, 1-\epsilon]$, then*

$$\mathop{\mathbf{E}}_{\substack{(\boldsymbol{x},\boldsymbol{x}') \\ \rho\text{-correlated}}}[\Lambda_\rho(f(\boldsymbol{x}), g(\boldsymbol{x}'))] \leq \Lambda_\rho(\mathbf{E}[f], \mathbf{E}[g]) + C_{\rho,\epsilon} \cdot (\Delta_n[f] + \Delta_n[g]).$$

Here we using the following inductive notation: $\Delta_1[f] = \mathbf{E}[|f - \mathbf{E}[f]|^3]$, and

$$\Delta_n[f] = \mathop{\mathbf{E}}_{\boldsymbol{x}_n \sim \{-1,1\}}\left[\Delta_{n-1}[f_{|\boldsymbol{x}_n}]\right] + \Delta_1[f^{\subseteq\{n\}}].$$

(*d*) Prove by induction that $\Delta_n[f] \le 8\sum_{i=1}^{n} \|\mathrm{D}_i f\|_3^3$.

(*e*) Suppose that $f, g \in L^2(\mathbb{R}, \gamma)$ have range $[\epsilon, 1-\epsilon]$ and are c-Lipschitz. Show that for any $M \in \mathbb{N}^+$, the Two-Function Borell Isoperimetric Theorem holds for f, g with an additional additive error of $O(M^{-1/2})$, where the constant in the $O(\cdot)$ depends only on ρ, ϵ, and c. (Hint: Use $\mathrm{BitsToGaussians}_M$.)

(*f*) By an approximation argument, deduce the Two-Function Borell Isoperimetric Theorem for general $f, g \in L^2(\mathbb{R}, \gamma)$ with range $[0,1]$; i.e., prove Exercise 11.29.

11.31 Fix $0 < \rho < 1$ and suppose $f \in L^1(\mathbb{R}, \gamma)$ is nonnegative and satisfies $\mathbf{E}[f] = 1$. Note that $\mathbf{E}[\mathrm{U}_\rho f] = 1$ as well. The goal of this problem is to show that $\mathrm{U}_\rho f$ satisfies an improved Markov inequality: $\mathbf{Pr}[\mathrm{U}_\rho f > t] = O(\frac{1}{t\sqrt{\ln t}}) = o(\frac{1}{t})$ as $t \to \infty$. This gives a quantitative sense in which U_ρ is a "smoothing operator": $\mathrm{U}_\rho f$ can never look too much like like a step function (the tight example for Markov's inequality).

(*a*) For simplicity, let's first assume $\rho = 1/\sqrt{2}$. Given $t > \sqrt{2}$, select $h > 0$ such that $\varphi(h) = t/\sqrt{\pi}$. Show that $h \sim \sqrt{2\ln t}$.

(*b*) Let $H = \{z : \mathrm{U}_\rho f(z) > t\}$. Show that if $H \subseteq (-\infty, -h] \cup [h, \infty)$, then we have $\mathbf{Pr}[\mathrm{U}_\rho f > t] \lesssim \frac{\sqrt{2/\pi}}{t\sqrt{\ln t}}$, as desired. (Hint: You'll need $\overline{\Phi}(u) < \varphi(u)/u$.)

(*c*) Otherwise, we wish to get a contradiction. First, show that there exists $y \in (-h, h)$ and $\delta_0 > 0$ such that $\mathrm{U}_\rho f(z) > t$ for all $t \in (y - \delta_0, y + \delta_0)$. (Hint: You'll need that $\mathrm{U}_\rho f$ is continuous; see Exercise 11.5.)

(*d*) For $0 < \delta < \delta_0$, define $g \in L^1(\mathbb{R}, \gamma)$ by $g(z) = \frac{1}{2\delta} 1_{(y-\delta, y+\delta)}$. Show that $0 \le \mathrm{U}_\rho g \le \frac{1}{\sqrt{\pi}}$ pointwise. (Hint: Why is $\mathrm{U}_\rho g(z)$ maximized at $\sqrt{2}y$?)

(*e*) Show that $\frac{1}{\sqrt{\pi}} \ge \langle f, \mathrm{U}_\rho g\rangle > t\,\mathbf{E}[g]$.

(*f*) Derive a contradiction by taking $\delta \to 0$, thereby showing that indeed $\mathbf{Pr}[\mathrm{U}_\rho f > t] \lesssim \frac{\sqrt{2/\pi}}{t\sqrt{\ln t}}$.

(*g*) Show that this result is tight by constructing an appropriate f.

(*h*) Generalize the above to show that for any fixed $0 < \rho < 1$ we have $\mathbf{Pr}[\mathrm{U}_\rho f > t] \lesssim \frac{1}{\sqrt{\pi(1-\rho^2)}} \frac{1}{t\sqrt{\ln t}}$.

11.32 As described in Example 11.73, show that $\mathrm{SDPOpt}(\mathbb{Z}_5) \ge \frac{1}{2} - \frac{1}{2}\cos\frac{4\pi}{5} = \frac{5}{8} + \frac{\sqrt{5}}{8}$.

11.33 Prove Theorem 11.72.

11.34 Consider the generalization of the Max-Cut CSP in which the variable set is V, the domain is $\{-1, 1\}$, and each constraint is an equality of two literals, i.e., it's of the form $bF(v) = b'F(v')$ for for some $v, v' \in V$ and $b, b' \in \{-1, 1\}$. This CSP is traditionally called Max-E2-Lin. Given an instance $\mathscr{P}$, write $(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{b}, \boldsymbol{b}') \sim \mathscr{P}$ to denote a uniformly chosen constraint.

The natural SDP relaxation (which can also be solved efficiently) is the following:

$$\begin{aligned}\text{maximize}\quad & \mathop{\mathbf{E}}_{(\boldsymbol{v},\boldsymbol{v}',\boldsymbol{b},\boldsymbol{b}')\sim\mathscr{P}}\left[\tfrac{1}{2}+\tfrac{1}{2}\langle \boldsymbol{b}\vec{U}(\boldsymbol{v}),\boldsymbol{b}'\vec{U}(\boldsymbol{v}')\rangle\right]\\ \text{subject to}\quad & \vec{U}:V\to S^{n-1}.\end{aligned}$$

Show that the Goemans–Williamson algorithm, when using this SDP, is a $(c_{\mathrm{GW}}\beta,\beta)$-approximation algorithm for Max-E2Lin, and that it also has the same refined guarantee as in Theorem 11.72.

11.35 This exercise builds on Exercise 11.34. Consider the following instance $\mathscr{P}$ of Max-E2-Lin: The variable set is $\mathbb{Z}_4$ and the constraints are

$$F(0)=F(1),\quad F(1)=F(2),\quad F(2)=F(3),\quad F(3)=-F(0).$$

(*a*) Show that $\mathrm{Opt}(\mathscr{P})=\frac{3}{4}$.

(*b*) Show that $\mathrm{SDPOpt}(\mathscr{P})\ge\frac{1}{2}+\frac{1}{2\sqrt{2}}$. (Hint: Very similar to Exercise 11.32; you can use four unit vectors at 45° angles in $\mathbb{R}^2$.)

(*c*) Deduce that $\mathrm{SDPOpt}(\mathscr{P})=\frac{1}{2}+\frac{1}{2\sqrt{2}}$ and that this is an optimal SDP integrality gap for Max-E2Lin. (Cf. Remark 11.76.)

11.36 In our proof of Theorem 11.74 it's stated that showing the β-Noise Sensitivity Test is a $(\theta/\pi,\frac{1}{2}-\frac{1}{2}\cos\theta)$-Dictator-vs.-No-Notables test implies the desired UG-hardness of $(\theta/\pi+\delta,\frac{1}{2}-\frac{1}{2}\cos\theta)$-approximating Max-Cut (for any constant $\delta>0$). There are two minor technical problems with this: First, the test can only actually be implemented when β is a rational number. Second, even ignoring this, Theorem 7.40 only directly yields hardness of $(\theta/\pi+\delta,\frac{1}{2}-\frac{1}{2}\cos\theta-\delta)$-approximation. Show how to overcome both technicalities. (Hint: Continuity.)

11.37 Use Corollary 11.59 (and (11.28)) to show that in the setting of the Berry–Esseen Theorem, $|\|\boldsymbol{S}\|_1-\sqrt{2/\pi}|\le O(\gamma^{1/3})$. (Cf. Exercise 5.31.)

11.38 The goal of this exercise is to prove Proposition 11.58.

(*a*) Reduce to the case $c=1$.

(*b*) Reduce to the case $\eta=1$. (Hint: Dilate the input by a factor of η.)

(*c*) Assuming henceforth that $c=\eta=1$, we define $\widetilde{\psi}(s)=\mathbf{E}[\psi(s+\boldsymbol{g})]$ for $\boldsymbol{g}\sim\mathrm{N}(0,1)$ as suggested; i.e., $\widetilde{\psi}=\psi*\varphi$, where φ is the Gaussian pdf. Show that indeed $\|\widetilde{\psi}-\psi\|_\infty\le\sqrt{2/\pi}\le 1$.

(*d*) To complete the proof we need to show that for all $s\in\mathbb{R}$ and $k\in\mathbb{N}^+$ we have $|\widetilde{\psi}^{(k)}(s)|\le C_k$. Explain why, in proving this, we may assume $\psi(s)=0$. (Hint: This requires $k\ge 1$.)

(*e*) Assuming $\psi(s)=0$, show $|\widetilde{\psi}^{(k)}(s)|=|\psi*\varphi^{(k)}(s)|\le C_k$. (Hint: Show that $\varphi^{(k)}(s)=p(s)\varphi(s)$ for some polynomial $p(s)$ and use the fact that Gaussians have finite absolute moments.)

11.39 Establish the following multidimensional generalization of Proposition 11.58:

Proposition 11.79. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be c-Lipschitz. Then for any $\eta > 0$ there exists $\widetilde{\psi}_\eta : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|\psi - \widetilde{\psi}_\eta\|_\infty \leq c\sqrt{d}\eta$ and $\|\partial^\beta \widetilde{\psi}_\eta\|_\infty \leq C_{|\beta|} c\sqrt{d}/\eta^{|\beta|-1}$ for each multi-index $\beta \in \mathbb{N}^d$ with $|\beta| = \sum_i \beta_i \geq 1$, where C_k is a constant depending only on k.*

11.40 In Exercise 11.38 we "mollified" a function ψ by convolving it with the (smooth) pdf of a Gaussian random variable. It's sometimes helpful to instead use a random variable with bounded support (but still with a smooth pdf on all of $\mathbb{R}$). Here we construct such a random variable. Define $b : \mathbb{R} \to \mathbb{R}$ by

$$b(x) = \begin{cases} \exp\left(-\frac{1}{1-x^2}\right) & \text{if } -1 < x < 1, \\ 0 & \text{else.} \end{cases}$$

(*a*) Verify that $b(x) \geq 0$ for all x and that $b(-x) = b(x)$.
(*b*) Prove the following statement by induction on $k \in \mathbb{N}$: On $(-1,1)$, the kth derivative of b at x is of the form $p(x)(1-x^2)^{-2k} \cdot b(x)$, where $p(x)$ is a polynomial.
(*c*) Deduce that b is a smooth ($\mathscr{C}^\infty$) function on $\mathbb{R}$.
(*d*) Verify that $C = \int_{-1}^{1} b(x)\,dx$ satisfies $0 < C < \infty$ and that we can therefore define a real random variable $\boldsymbol{y}$, symmetric and supported on $(-1,1)$, with the smooth pdf $\widetilde{b}(y) = b(y)/C$. Show also that for $k \in \mathbb{N}$, the numbers $c_k = \|\widetilde{b}^{(k)}\|_\infty$ are finite and positive, where $\widetilde{b}^{(k)}$ denotes the kth derivative of $\widetilde{b}$.
(*e*) Give an alternate proof of Exercise 11.38 using $\boldsymbol{y}$ in place of $\boldsymbol{g}$.

11.41 Fix $u \in \mathbb{R}$, $\psi(s) = 1_{s \leq u}$, and $0 < \eta < 1/2$.
(*a*) Suppose we approximate ψ by a smooth function $\widetilde{\psi}_\eta$ as in Exercise 11.38, i.e., we define $\widetilde{\psi}_\eta(s) = \mathbf{E}[\psi(s + \eta\boldsymbol{g})]$ for $\boldsymbol{g} \sim \mathrm{N}(0,1)$. Show that $\widetilde{\psi}_\eta$ satisfies the following properties:
- $\widetilde{\psi}_\eta$ is a decreasing function with $\widetilde{\psi}_\eta(s) < \psi(s)$ for $s < u$ and $\widetilde{\psi}_\eta(s) > \psi(s)$ for $s > u$.
- $|\widetilde{\psi}_\eta(s) - \psi(s)| \leq \eta$ provided $|s - u| \geq O(\eta\sqrt{\log(1/\eta)})$.
- $\|\widetilde{\psi}_\eta^{(k)}\|_\infty \leq C_k/\eta^k$ for each $k \in \mathbb{N}$, where C_k depends only on k.

(*b*) Suppose we instead approximate ψ by the function $\widetilde{\psi}_\eta(s) = \mathbf{E}[\psi(s + \eta\boldsymbol{y})]$, where $\boldsymbol{y}$ is the random variable from Exercise 11.40. Show that $\widetilde{\psi}_\eta$ satisfies the following slightly nicer properties:
- $\widetilde{\psi}_\eta$ is a nonincreasing function which agrees with ψ on $(\infty, u-\eta]$ and on $[u+\eta, \infty)$.
- $\widetilde{\psi}_\eta$ is smooth and satisfies $\|\widetilde{\psi}_\eta^{(k)}\|_\infty \leq C_k/\eta^k$ for each $k \in \mathbb{N}$, where C_k depends only on k.

11.42 Prove Corollary 11.61 by first proving

$$\mathbf{Pr}[\boldsymbol{S}_Y \leq u - 2\eta] - O(\eta^{-3})\gamma_{XY} \leq \mathbf{Pr}[\boldsymbol{S}_X \leq u] \leq \mathbf{Pr}[\boldsymbol{S}_Y \leq u + 2\eta] + O(\eta^{-3})\gamma_{XY}.$$

(Hint: Obtain $\mathbf{Pr}[\boldsymbol{S}_X \leq u-\eta] \leq \mathbf{E}[\widetilde{\psi}_\eta(\boldsymbol{S}_X)] \approx \mathbf{E}[\widetilde{\psi}_\eta(\boldsymbol{S}_Y)] \leq \mathbf{Pr}[\boldsymbol{S}_Y \leq u+\eta]$ using properties from Exercise 11.41. Then replace u with $u+\eta$ and also interchange $\boldsymbol{S}_X$ and $\boldsymbol{S}_Y$.)

11.43 (*a*) Fix $q \in \mathbb{N}$. Establish the existence of a smooth function $f_q : \mathbb{R} \to \mathbb{R}$ that is 0 on $(-\infty, -\frac{1}{2}]$ and that agrees with some polynomial of degree exactly q on $[\frac{1}{2}, \infty)$. (Hint: Induction on q; the base case $q = 0$ is essentially Exercise 11.41, and the induction step can be achieved by integration.)

(*b*) Deduce that for any prescribed sequence $a_0, a_1, a_2, \dots$ that is eventually constantly 0, there is a smooth function $g : \mathbb{R} \to \mathbb{R}$ that is 0 on $(-\infty, -\frac{1}{2}]$ and has $g^{(k)}(\frac{1}{2}) = a_k$ for all $k \in \mathbb{N}$.

(*c*) Fix a univariate polynomial $p : \mathbb{R} \to \mathbb{R}$. Show that there is a smooth function $\widetilde{\psi} : \mathbb{R} \to \mathbb{R}$ that agrees with p on $[-1,1]$ and is identically 0 on $(-\infty, -2] \cup [2, \infty)$.

11.44 Establish Corollary 11.70.

11.45 Prove Theorem 11.71.

11.46 (*a*) By following our proof of the $d = 1$ case and using the multivariate Taylor theorem, establish the following:

Invariance Principle for Sums of Random Vectors. *Let $\vec{\boldsymbol{X}}_1, \dots, \vec{\boldsymbol{X}}_n$, $\vec{\boldsymbol{Y}}_1, \dots, \vec{\boldsymbol{Y}}_n$ be independent $\mathbb{R}^d$-valued random variables with matching means and covariance matrices; i.e.,* $\mathbf{E}[\vec{\boldsymbol{X}}_t] = \mathbf{E}[\vec{\boldsymbol{Y}}_t]$ *and* $\mathbf{Cov}[\vec{\boldsymbol{X}}_t] = \mathbf{Cov}[\vec{\boldsymbol{Y}}_t]$ *for all $t \in [n]$. (Note that the d individual components of a particular $\vec{\boldsymbol{X}}_t$ or $\vec{\boldsymbol{Y}}_t$ are not required to be independent.) Write $\vec{\boldsymbol{S}}_X = \sum_{t=1}^n \vec{\boldsymbol{X}}_t$ and $\vec{\boldsymbol{S}}_Y = \sum_{t=1}^n \vec{\boldsymbol{Y}}_t$. Then for any $\mathscr{C}^3$ function $\psi : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|\partial^\beta \psi\|_\infty \leq C$ for all $|\beta| = 3$,*

$$\left|\mathbf{E}[\psi(\vec{\boldsymbol{S}}_X)] - \mathbf{E}[\psi(\vec{\boldsymbol{S}}_Y)]\right| \leq C\gamma_{\vec{X}\vec{Y}},$$

where

$$\gamma_{\vec{X}\vec{Y}} = \sum_{\substack{\beta \in \mathbb{N}^d \\ |\beta|=3}} \frac{1}{\beta!} \sum_{t=1}^n \left(\mathbf{E}[|\vec{\boldsymbol{X}}_t^\beta|] + \mathbf{E}[|\vec{\boldsymbol{Y}}_t^\beta|]\right).$$

(*b*) Show that $\gamma_{\vec{X}\vec{Y}}$ satisfies

$$\gamma_{\vec{X}\vec{Y}} \leq \frac{d^2}{6} \sum_{t=1}^n \sum_{i=1}^d \left(\mathbf{E}[|\vec{\boldsymbol{X}}_t^{3e_i}|] + \mathbf{E}[|\vec{\boldsymbol{Y}}_t^{3e_i}|]\right).$$

Here $\vec{\boldsymbol{X}}_t^{3e_i}$ denotes the cube of the ith component of vector $\vec{\boldsymbol{X}}_t$, and similarly for $\vec{\boldsymbol{Y}}_t$. (Hint: $abc \leq \frac{1}{3}(a^3 + b^3 + c^3)$ for $a, b, c \geq 0$.)

(*c*) Deduce multivariate analogues of the Variant Berry–Esseen Theorem, Remark 11.56, and Corollary 11.59 (using Proposition 11.79).

11.47 Justify Remark 11.66. (Hint: You'll need Exercise 10.29.)

11.48 (*a*) Prove the following:

Multifunction Invariance Principle. *Let $F^{(1)}, \dots, F^{(d)}$ be formal n-variate multilinear polynomials each of degree at most $k \in \mathbb{N}$. Let $\vec{\boldsymbol{x}}_1, \dots, \vec{\boldsymbol{x}}_n$ and $\vec{\boldsymbol{y}}_1, \dots, \vec{\boldsymbol{y}}_n$ be independent $\mathbb{R}^d$-valued random variables such that $\mathbf{E}[\vec{\boldsymbol{x}}_t] = \mathbf{E}[\vec{\boldsymbol{y}}_t] = 0$ and $M_t = \mathbf{Cov}[\vec{\boldsymbol{x}}_t] = \mathbf{Cov}[\vec{\boldsymbol{y}}_t]$ for each $t \in [n]$. Assume each M_t has all its diagonal entries equal to 1 (i.e., each of the d components of $\vec{\boldsymbol{x}}_t$ has variance 1, and similarly for $\vec{\boldsymbol{y}}_t$). Further assume each component random variable $\vec{\boldsymbol{x}}_t^{(j)}$ and $\vec{\boldsymbol{y}}_t^{(j)}$ is $(2,3,\rho)$-hypercontractive ($t \in [n]$, $j \in [d]$). Then for any $\mathscr{C}^3$ function $\psi : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|\partial^\beta \psi\|_\infty \leq C$ for all $|\beta| = 3$,*

$$\left|\mathbf{E}[\psi(\vec{F}(\vec{\boldsymbol{x}}))] - \mathbf{E}[\psi(\vec{F}(\vec{\boldsymbol{y}}))]\right| \leq \tfrac{Cd^2}{3} \cdot (1/\rho)^{3k} \cdot \sum_{t=1}^{n} \sum_{j=1}^{d} \mathbf{Inf}_t[F^{(j)}]^{3/2}.$$

Here we are using the following notation: If $\vec{\boldsymbol{z}} = (\vec{\boldsymbol{z}}_1, \dots, \vec{\boldsymbol{z}}_n)$ is a sequence of $\mathbb{R}^d$-valued random variables, $\vec{F}(\vec{\boldsymbol{z}})$ denotes the vector in $\mathbb{R}^d$ whose jth component is $F^{(j)}(\vec{\boldsymbol{z}}_1^{(j)}, \dots, \vec{\boldsymbol{z}}_n^{(j)})$.

(Hint: Combine the proofs of the Basic Invariance Principle and the Invariance Principle for Sums of Random Vectors, Exercise 11.46. The only challenging part should be notation.)

(*b*) Show that if we further have $\mathbf{Var}[F^{(j)}] \leq 1$ and $\mathbf{Inf}_t[F^{(j)}] \leq \epsilon$ for all $j \in [d]$, $t \in [n]$, then

$$\left|\mathbf{E}[\psi(\vec{F}(\vec{\boldsymbol{x}}))] - \mathbf{E}[\psi(\vec{F}(\vec{\boldsymbol{y}}))]\right| \leq \tfrac{Cd^3}{3} \cdot k(1/\rho)^{3k} \cdot \epsilon^{1/2}.$$

11.49 (*a*) Prove the following:

Invariance Principle in general product spaces. *Let (Ω, π) be a finite probability space, $|\Omega| = m \geq 2$, in which every outcome has probability at least λ. Suppose $f \in L^2(\Omega^n, \pi^{\otimes n})$ has degree at most k; thus, fixing some Fourier basis $\phi_0, \dots, \phi_{m-1}$ for $L^2(\Omega, \pi)$, we have*

$$f = \sum_{\substack{\alpha \in \mathbb{N}^n_{<m} \\ \#\alpha \leq k}} \widehat{f}(\alpha)\phi_\alpha.$$

Introduce indeterminates $x = (x_{i,j})_{i \in [n], j \in [m-1]}$ and let F be the formal $(m-1)n$-variate polynomial of degree at most k defined by

$$F(x) = \sum_{\#\alpha \leq k} \widehat{f}(\alpha) \prod_{i \in \mathrm{supp}(\alpha)} x_{i,\alpha_i}.$$

Then for any $\psi : \mathbb{R} \to \mathbb{R}$ that is $\mathscr{C}^3$ and satisfies $\|\psi'''\|_\infty \leq C$ we have

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^{(m-1)n}}[\psi(F(\boldsymbol{x}))] - \mathop{\mathbf{E}}_{\boldsymbol{\omega} \sim \pi^{\otimes n}}[\psi(f(\boldsymbol{\omega}))]\right| \leq \tfrac{C}{3} \cdot (2\sqrt{2/\lambda})^k \cdot \sum_{i=1}^{n} \mathbf{Inf}_i[f]^{3/2}.$$

(Hint: For $0 \le t \le n$, define the function $h_t \in L^2(\Omega^t \times \{-1,1\}^{(m-1)(n-t)}, \pi^{\otimes t} \otimes \pi_{1/2}^{\otimes(m-1)(n-t)})$ via

$$h_t(\boldsymbol{\omega}_1,\dots,\boldsymbol{\omega}_t,\boldsymbol{x}_{t+1,1},\dots,\boldsymbol{x}_{n,m-1}) = \sum_{\#\alpha\le k} \widehat{f}(\alpha) \prod_{\substack{i\in\mathrm{supp}(\alpha)\\ i\le t}} \phi_{\alpha_i}(\boldsymbol{\omega}_i) \prod_{\substack{i\in\mathrm{supp}(\alpha)\\ i>t}} \boldsymbol{x}_{i,\alpha_i}.$$

Express

$$h_t = \mathrm{E}_t h_t + \mathrm{L}_t h_t = \mathrm{E}_t h_t + \sum_{j=1}^m D_j \cdot \phi_j(\boldsymbol{\omega}_t)$$

where

$$D_j = \sum_{\alpha:\alpha_t=j} \widehat{f}(\alpha) \prod_{\substack{i\in\mathrm{supp}(\alpha)\\ i<t}} \phi_{\alpha_i}(\boldsymbol{\omega}_i) \prod_{\substack{i\in\mathrm{supp}(\alpha)\\ i>t}} \boldsymbol{x}_{i,\alpha_i},$$

and note that $h_{t-1} = \mathrm{E}_t h_t + \sum_{j=1}^m D_j \cdot \boldsymbol{x}_{t,j}$.)

(*b*) In the setting of the previous theorem, show also that

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{g}\sim\mathrm{N}(0,1)^{(m-1)n}}[\psi(F(\boldsymbol{g}))] - \mathop{\mathbf{E}}_{\boldsymbol{\omega}\sim\pi^{\otimes n}}[\psi(f(\boldsymbol{\omega}))] \right| \le \tfrac{2C}{3} \cdot (2\sqrt{2/\lambda})^k \cdot \sum_{i=1}^n \mathbf{Inf}_i[f]^{3/2}.$$

(Hint: Apply the Basic Invariance Principle in the form of Exercise 11.47. How can you bound the $(m-1)n$ influences of F in terms of the n influences of f?)

11.50 Prove the following version of the General-Volume Majority Is Stablest Theorem in the setting of general product spaces:

Theorem 11.80. *Let (Ω,π) be a finite probability space in which each outcome has probability at least λ. Let $f \in L^2(\Omega^n,\pi^{\otimes n})$ have range $[0,1]$. Suppose that f has no $(\epsilon, \frac{1}{\log(1/\epsilon)})$-notable coordinates. Then for any $0 \le \rho < 1$,*

$$\mathbf{Stab}_\rho[f] \le \Lambda_\rho(\mathbf{E}[f]) + O\big(\tfrac{\log\log(1/\epsilon)}{\log(1/\epsilon)}\big) \cdot \tfrac{\log(1/\lambda)}{1-\rho}.$$

(Hint: Naturally, you'll need Exercise 11.49(*b*).)

Notes. The subject of Gaussian space is too enormous to be surveyed here; some recommended texts include Janson [**Jan97**] and Bogachev [**Bog98**], the latter having an extremely thorough bibliography. The Ornstein–Uhlenbeck semigroup dates back to the work of Uhlenbeck and Ornstein [**UO30**] whose motivation was to refine Einstein's theory of Brownian motion [**Ein05**] to take into account the inertia of the particle. The relationship between the action of U_ρ on functions and on Hermite expansions (i.e., Proposition 11.31) dates back even further, to Mehler [**Meh66**]. Hermite polynomials were first defined by Laplace [**Lap11**], and then studied by Chebyshev [**Che60**] and Hermite [**Her64**]. See Lebedev [**Leb72**, Chapter 4.15] for a proof of the pointwise convergence of a piecewise-$\mathscr{C}^1$ function's Hermite expansion.

As mentioned in Chapter 9.7, the Gaussian Hypercontractivity Theorem is originally due to Nelson [**Nel66**] and now has many known proofs. The idea behind the proof we presented – first proving the Boolean hypercontractivity result and then deducing the Gaussian case by the Central Limit Theorem – is due to Gross [**Gro75**] (see also Trotter [**Tro58**]). Gross actually used the idea to prove his Gaussian Log-Sobolev Inequality, and thereby deduced the Gaussian Hypercontractivity Theorem. Direct proofs of the Gaussian Hypercontractivity Theorem have been given by Neveu [**Nev76**] (using stochastic calculus), Brascamp and Lieb [**BL76**] (using rearrangement [**BL76**]), and Ledoux [**Led13**] (using a variation on Exercises 11.26–11.29); direct proofs of the Gaussian Log-Sobolev Inequality have been given by Adams and Clarke [**AC79**], by Bakry and Emery [**BÉ85**], and by Ledoux [**Led92**], the latter two using semigroup techniques. Bakry's survey [**Bak94**] on these topics is also recommended.

The Gaussian Isoperimetric Inequality was first proved independently by Borell [**Bor75**] and by Sudakov and Tsirel'son [**ST78**]. Both works derived the result by taking the isoperimetric inequality on the sphere (due to Lévy [**Lév22**] and Schmidt [**Sch48**], see also Figiel, Lindenstrauss, and Milman [**FLM77**]) and then taking "Poincaré's limit" – i.e., viewing Gaussian space as a projection of the sphere of radius $\sqrt{n}$ in n dimensions, with $n \to \infty$ (see Lévy [**Lév22**], McKean [**McK73**], and Diaconis and Freedman [**DF87**]). Ehrhard [**Ehr83**] gave a different proof using a symmetrization argument intrinsic to Gaussian space. This may be compared to the alternate proof of the spherical isoperimetric inequality [**Ben84**] based on the "two-point symmetrization" of Baernstein and Taylor [**BT76**] (analogous to Riesz rearrangement in Euclidean space and to the polarization operation from Exercise 2.52).

To carefully define Gaussian surface area for a broad class of sets requires venturing into the study of geometric measure theory and functions of bounded variation. For a clear and comprehensive development in the Euclidean setting (including the remark in Exercise 11.15(*b*)), see the book by Ambrosio, Fusco, and Pallara [**AFP00**]. There's not much difference between the Euclidean and finite-dimensional Gaussian settings; research on Gaussian perimeter tends to focus on the trickier infinite-dimensional case. For a thorough development of surface area in this latter setting (which of course includes finite-dimensional Gaussian space as a special case) see the work of Ambrosio, Miranda, Maniglia, and Pallara [**AMMP10**]; in particular, Theorem 4.1 in that work gives several additional equivalent definitions for surf_γ besides those in Definition 11.48. Regarding the fact that $\mathbf{RS}'_A(0^+)$ is an equivalent definition, the Euclidean analogue of this statement was proven in Miranda et al. [**MPPP07**] and the statement itself follows similarly [**Mir13**] using Ambrosio et al. [**AFR13**]. (Our heuristic justification of (11.14) is similar to the one given by Kane [**Kan11**].) Additional related results can be found in Hino [Hin10] (which includes the remark about convex sets at the end of Definition 11.48), Ambrosio and Figalli [AF11], Miranda et al. [MNP12], and Ambrosio et al. [AFR13].

The inequality of Theorem 11.51 is explicit in Ledoux [Led94] (see also the excellent survey [Led96]); he used it to deduce the Gaussian Isoperimetric Inequality. He also noted that it's essentially deducible from an earlier inequality of Pisier and Maurey [Pis86, Theorem 2.2]. Theorem 11.43, which expresses the subadditivity of rotation sensitivity, can be viewed as a discretization of the Pisier–Maurey inequality. This theorem appeared in work of Kindler and O'Donnell [KO12], which also made the observations about the volume-$\frac{1}{2}$ case of Borell's Isoperimetric Theorem at the end of Section 11.3 and in Remark 11.76.

Bobkov's Inequality [Bob97] in the special case of Gaussian space had already been implicitly established by Ehrhard [Ehr84]; the striking novelty of Bobkov's work (partially inspired by Talagrand [Tal93]) was his reduction to the two-point Boolean inequality. The proof of this inequality which we presented is, as mentioned a discretization of the stochastic calculus proof of Barthe and Maurey [BM00]. (In turn, they were extending the stochastic calculus proof of Bobkov's Inequality in the Gaussian setting due to Capitaine, Hsu, and Ledoux [CHL97].) The idea that it's enough to show that Claim 11.54 is "nearly true" by computing two derivatives – as opposed to showing it's exactly true by computing four derivatives – was communicated to the author by Yuval Peres. Following Bobkov's paper, Bakry and Ledoux [BL96] established Theorem 11.55 in very general infinite-dimensional settings including Gaussian space; Ledoux [Led98] further pointed out that the Gaussian version of Bobkov's Inequality has a very short and direct semigroup-based proof. See also Bobkov and Götze [BG99] and Tillich and Zémor [TZ00] for results similar to Bobkov's Inequality in other discrete settings.

Borell's Isoperimetric Theorem is from Borell [Bor85]. Borell's proof used "Ehrhard symmetrization" and actually gave much stronger results – e.g., that if $f, g \in L^2(\mathbb{R}^n, \gamma)$ are nonnegative and $q \geq 1$, then $\langle (\mathrm{U}_\rho f)^q, g\rangle$ can only increase under simultaneous Ehrhard symmetrization of f and g. There are at least four other known proofs of the basic Borell Isoperimetric Theorem. Beckner [Bec92] observed that the analogous isoperimetric theorem on the sphere follows from two-point symmetrization; this yields the Gaussian result via Poincaré's limit (for details, see Carlen and Loss [CL90]). (This proof is perhaps the conceptually simplest one, though carrying out all the technical details is a chore.) Mossel and Neeman [MN12] gave the proof based on semigroup methods outlined in Exercises 11.26–11.29, and later together with De [**DMN12**] gave a "Bobkov-style" Boolean proof (see Exercise 11.30). Finally, Eldan [**Eld13**] gave a proof using stochastic calculus.

As mentioned in Section 11.5 there are several known ways to prove the Berry–Esseen Theorem. Aside from the original method (characteristic functions), there is also Stein's Method [**Ste72, Ste86b**]; see also, e.g., [**Bol84, BH84, CGS11**]. The Replacement Method approach we presented originates in the work of Lindeberg [**Lin22**]. The mollification techniques used (e.g., those in Exercise 11.40) are standard. The Invariance Principle as presented in Section 11.6 is from Mossel, O'Donnell, and Oleszkiewicz [**MOO10**]. Further extensions (e.g., Exercise 11.48) appear in the work of Mossel [**Mos10**]. In fact the Invariance Principle dates back to the 1971 work of Rotar' [**Rot73, Rot74**]; therein he essentially proved the Invariance Principle for degree-2 multilinear polynomials (even employing the term "influence" as we do for the quantity in Definition 11.63). Earlier work on extending the Central Limit Theorem to higher-degree polynomials had focused on obtaining sufficient conditions for polynomials (especially quadratics) to have a Gaussian limit distribution; this is the subject of *U-statistics*. Rotar' emphasized the idea of invariance and of allowing any (quadratic) polynomial with low influences. Rotar' also credited Girko [**Gir73**] with related results in the case of positive definite quadratic forms. In 1975, Rotar' [**Rot75**] generalized his results to handle multilinear polynomials of any constant degree, and also random vectors (as in Exercise 11.48). (Rotar' also gave further refinements in 1979 [**Rot79**].)

The difference between the results of Rotar' [**Rot75**] and Mossel et al. [**MOO10**] comes in the treatment of the error bounds. It's somewhat difficult to extract simple-to-state error bounds from Rotar' [**Rot75**], as the error there is presented as a sum over $i \in [n]$ of expressions $\mathbf{E}[F(\boldsymbol{x})\mathbf{1}_{|F(\boldsymbol{x})|>u_i}]$, where u_i involves $\mathbf{Inf}_i[F]$. (Partly this is so as to generalize the statement of the Lindeberg CLT.) Nevertheless, the work of Rotar' implies a Lévy distance bound as in Corollary 11.70, with some inexplicit function $o_\epsilon(1)$ in place of $(1/\rho)^{O(k)}\epsilon^{1/8}$. By contrast, the work of Mossel et al. [**MOO10**] shows that a straightforward combination of the Replacement Method and hypercontractivity yields good, explicit error bounds. Regarding the Carbery–Wright Theorem [**CW01**], an alternative exposition appears in Nazarov, Sodin, and Vol'berg [**NSV02**].

Regarding the Majority Is Stablest Theorem (conjectured in Khot, Kindler, Mossel, and O'Donnell [**KKMO04**] and proved originally in Mossel, O'Donnell, and Oleszkiewicz [**MOO05b**]), it can be added that additional motivation for the conjecture came from Kalai [**Kal02**]. The fact that (SDP) is an efficiently computable relaxation for the Max-Cut problem dates back to the 1990 work of Delorme and Poljak [**DP93**]; however, they were unable to give an analysis relating its value to the optimum cut value. In fact, they conjectured that the case of the 5-cycle from Example 11.73 had the worst ratio of Opt(G) to SDPOpt(G). Goemans and Williamson [**GW94**] were the first to give a sharp analysis of the SDP (Theorem 11.72), at least for $\theta \geq \theta^*$. Feige and Schechtman [**FS02**] showed an optimal integrality gap for the SDP for all values $\theta \geq \theta^*$ (in particular, showing an integrality gap ratio of c_{GW}); interestingly, their construction essentially involved proving Borell's Isoperimetric Inequality (though they did it on the sphere rather than in Gaussian space). Both before and after the Khot et al. [**KKMO04**] UG-hardness result for Max-Cut there was a long line of work [**Kar99, Zwi99, AS00, ASZ02, CW04, KV05, FL06, KO06**] devoted to improving the known approximation algorithms and UG-hardness results, in particular for $\theta < \theta^*$. This culminated in the results from O'Donnell and Wu [**OW08**] (mentioned in Remark 11.75), which showed explicit matching (α, β)-approximation algorithms, integrality gaps, and UG-hardness results for all $\frac{1}{2} < \beta < 1$. The fact that the best integrality gaps matched the best UG-hardness results proved not to be a coincidence; in contemporaneous work, Raghavendra [**Rag08**] showed that for *any* CSP, *any* SDP integrality gap could be turned into a matching Dictator-vs.-No-Notables test. This implies the existence of matching efficient (α, β)-approximation algorithms and UG-hardness results for every CSP and every β. See Raghavendra's thesis [**Rag09**] for full details of his earlier publication [**Rag08**] (including some Invariance Principle extensions building further on Mossel [**Mos10**]); see also Austrin's work [**Aus07, Aus10**] for precursors to the Raghavendra theory.

Exercise 11.31 concerns a problem introduced by Talagrand [**Tal89**]. Talagrand offers a \$1,000 prize [**Tal06**] for a solution to the following Boolean version of the problem: Show that for any fixed $0 < \rho < 1$ and for $f : \{-1,1\}^n \to \mathbb{R}^{\geq 0}$ with $\mathbf{E}[f] = 1$ it holds that $\mathbf{Pr}[\mathrm{T}_\rho f > t] = o(1/t)$ as $t \to \infty$. (The rate of decay may depend on ρ but not, of course, on n; in fact, a bound of the form $O(\frac{1}{t\sqrt{\log t}})$ is expected.) The result outlined in Exercise 11.31 (obtained together with James Lee) is for the very special case of 1-dimensional Gaussian space; Ball, Barthe, Bednorz, Oleszkiewicz, and Wolff [**BBB$^+$13**] obtained the same result and also showed a bound of $O(\frac{\log\log t}{t\sqrt{\log t}})$ for d-dimensional Gaussian space (with the constant in the $O(\cdot)$ depending on d).

The Multifunction Invariance Principle (Exercise 11.48 and its special case Exercise 11.46) are from Mossel [**Mos10**]; the version for general product spaces (Exercise 11.49) is from Mossel, O'Donnell, and Oleszkiewicz [**MOO10**].

Some tips

- You might try using analysis of Boolean functions whenever you're faced with a problems involving Boolean strings in which both the uniform probability distribution and the Hamming graph structure play a role. More generally, the tools may still apply when studying functions on (or subsets of) product probability spaces.
- If you're mainly interested in unbiased functions, or subsets of volume $\frac{1}{2}$, use the representation $f : \{-1,1\}^n \to \{-1,1\}$. If you're mainly interested in subsets of small volume, use the representation $f : \{-1,1\}^n \to \{0,1\}$.
- As for the domain, if you're interested in the operation of adding two strings (modulo 2), use $\mathbb{F}_2^n$. Otherwise use $\{-1,1\}^n$.
- If you have a conjecture about Boolean functions:
  - Test it on dictators, majority, parity, tribes (and maybe recursive majority of 3). If it's true for these functions, it's probably true.
  - Try to prove it by induction on n.
  - Try to prove it in the special case of functions on Gaussian space.
- Try not to prove any bound on Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ that involves the parameter n.
- Analytically, the only multivariate polynomials we really know how to control are degree-1 polynomials. Try to reduce to this case if you can.
- Hypercontractivity is useful in two ways: (i) It lets you show that low-degree functions of independent random variables behave "reasonably". (ii) It implies that the noisy hypercube graph is a small-set expander.
- Almost any result about functions on the hypercube extends to the case of the p-biased cube, and more generally, to the case of functions on products of discrete probability spaces in which every outcome has probability at least p – possibly with a dependence on p, though.
- Every Boolean function consists of a junta part and Gaussian part.

Bibliography

[AA11] Scott Aaronson and Andris Ambainis. The need for structure in quantum speedups. In *Proceedings of the 2nd Annual Innovations in Theoretical Computer Science conference*, pages 338–352, 2011.

[Aar08] Scott Aaronson. How to solve longstanding open problems in quantum computing using only Fourier Analysis. Lecture at Banff International Research Station, 2008. http://www.scottaaronson.com/talks/openqc.ppt.

[ABH$^+$05] Sanjeev Arora, Eli Berger, Elad Hazan, Guy Kindler, and Muli Safra. On non-approximability for quadratic programs. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 206–215, 2005.

[ABI85] Noga Alon, László Babai, and Alon Itai. A fast and simple randomized algorithm for the maximal independent set problem. *Journal of Algorithms*, 7(4):567–583, 1985.

[AC79] Robert Adams and Frank Clarke. Gross's logarithmic Sobolev inequality: a simple proof. *American Journal of Mathematics*, 101(6):1265–1269, 1979.

[AF99] Dimitris Achlioptas and Ehud Friedgut. A sharp threshold for k-colorability. *Random Structures & Algorithms*, 14(1):63–70, 1999.

[AF11] Luigi Ambrosio and Alessio Figalli. Surface measures and convergence of the Ornstein–Uhlenbeck semigroup in Wiener spaces. *Annales de la faculté des sciences de Toulouse Mathématiques (série 6)*, 20(2):407–438, 2011.

[AFP00] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford University Press, 2000.

[AFR13] Luigi Ambrosio, Alessio Figalli, and Eris Runa. On sets of finite perimeter in Wiener spaces: reduced boundary and convergence to halfspaces. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Serie IX. Matematica e Applicazioni*, 24(1):111–122, 2013.

[AGHP92] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures & Algorithms*, 3(3):289–304, 1992.

[Ajt83] Miklós Ajtai. Σ^1_1-formulae on finite structures. *Annals of Pure and Applied Logic*, 24(1):1–48, 1983.

[AL93] Miklós Ajtai and Nathal Linial. The influence of large coalitions. *Combinatorica*, 13(2):129–145, 1993.

[ALM$^+$98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998.

[Ama11] Kazuyuki Amano. Tight bounds on the average sensitivity of k-CNF. *Theory of Computing*, 7(1):45–48, 2011.

[Amb03] Andris Ambainis. Polynomial degree vs. quantum query complexity. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 230–239, 2003.

[AMMP10] Luigi Ambrosio, Michele Miranda Jr., Stefania Maniglia, and Diego Pallara. BV functions in abstract Wiener spaces. *Journal of Functional Analysis*, 258(3):785–813, 2010.

[AN05] Dimitris Achlioptas and Assaf Naor. The two possible values of the chromatic number of a random graph. *Annals of Mathematics*, 162(3):1335–1351, 2005.

[Arr50] Kenneth Arrow. A difficulty in the concept of social welfare. *The Journal of Political Economy*, 58(4):328–346, 1950.

[Arr63] Kenneth Arrow. *Social choice and individual values*. Cowles Foundation, 1963.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the ACM*, 45(1):70–122, 1998.

[AS00] Noga Alon and Benjamin Sudakov. Bipartite subgraphs and the smallest eigenvalue. *Combinatorics, Probability and Computing*, 9(1):1–12, 2000.

[AS08] Noga Alon and Joel Spencer. *The Probabilistic Method*. Wiley–Interscience, third edition, 2008.

[ASZ02] Noga Alon, Benny Sudakov, and Uri Zwick. Constructing worst case instances for semidefinite programming based approximation algorithms. *SIAM Journal on Discrete Mathematics*, 15(1):58–72, 2002.

[Aus07] Per Austrin. Balanced Max-2Sat might not be hardest. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 189–197, 2007.

[Aus08] Per Austrin. *Conditional Inapproximability and Limited Independence*. PhD thesis, KTH Royal Institute of Technology, 2008.

[Aus10] Per Austrin. Towards sharp inapproximability for any 2-CSP. *SIAM Journal On Computing*, 39(6):2430–2463, 2010.

[Bak94] Dominique Bakry. L'hypercontractivité et son utilisation en théorie des semigroupes. In *Lectures on probability theory (Saint-Flour, 1992)*, volume 1581 of *Lecture Notes in Mathematics*, pages 1–114. Springer, Berlin, 1994.

[Bal93] Keith Ball. The reverse isoperimetric problem for Gaussian measure. *Discrete and Computational Geometry*, 10(4):411–420, 1993.

[Bal13] Deepak Bal. On sharp thresholds of monotone properties: Bourgain's proof revisited. Technical Report 1302.1162, arXiv, 2013.

[Ban65] John Banzhaf. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

[BBB$^+$13] Keith Ball, Franck Barthe, Witold Bednorz, Krzysztof Oleszkiewicz, and Paweł Wolff. L^1-smoothing for the Ornstein–Uhlenbeck semigroup. *Mathematika*, 59(1):160–168, 2013.

[BBH$^+$12] Boaz Barak, Fernando Brandão, Aram Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 307–326, 2012.

[BC99] Anna Bernasconi and Bruno Codenotti. Spectral analysis of Boolean functions as a graph eigenvalue problem. *IEEE Transactions on Computers*, 48(3):345–351, 1999.

[BCH$^+$96] Mihir Bellare, Don Coppersmith, Johan Håstad, Marcos Kiwi, and Madhu Sudan. Linearity testing in characteristic two. *IEEE Transactions on Information Theory*, 42(6):1781–1795, 1996.

[BÉ85] Dominiques Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités, XIX*, volume 1123 of *Lecture Notes in Mathematics*, pages 177–206. Springer, Berlin, 1985.

[Bea94] Paul Beame. A switching lemma primer. Technical Report UW-CSE-95-07-01, University of Washington, 1994.

[Bec75] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, 102:159–182, 1975.

[Bec92] William Beckner. Sobolev inequalities, the Poisson semigroup, and analysis on the sphere S^n. *Proceedings of the National Academy of Sciences*, 89(11):4816–4819, 1992.

[BEHW87] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.

[Ben84] Yoav Benyamini. Two-point symmetrization, the isoperimetric inequality on the sphere and some applications. In *Texas functional analysis seminar, 1983–1984*, volume 1984, pages 53–76, 1984.

[Ben04] Vidmantas Bentkus. A Lyapunov type bound in $\mathbf{R}^d$. *Rossiĭskaya Akademiya Nauk. Teoriya Veroyatnosteĭ i ee Primeneniya*, 49(2):400–410, 2004.

[Ber41] Andrew Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–139, 1941.

[BG99] Sergey Bobkov and Friedrich Götze. Discrete isoperimetric and Poincaré-type inequalities. *Probability Theory and Related Fields*, 114(2):245–277, 1999.

[BGR09] Steven Brams, William Gehrlein, and Fred Roberts, editors. *The Mathematics of Preference, Choice and Order*. Springer, 2009.

[BGS95] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free bits, PCPs, and non-approximability – towards tight results. Technical Report TR95-024, Electronic Colloquium on Computational Complexity, 1995.

[BGS98] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free bits, PCPs, and non-approximability – towards tight results. *SIAM Journal of Computing*, 27(3):804–915, 1998.

[BH57] Simon Broadbent and John Hammersley. Percolation processes I. Crystals and mazes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(3):629–641, 1957.

[BH84] Andrew Barbour and Peter Hall. Stein's method and the Berry–Esseen theorem. *Australian Journal of Statistics*, 26(1):8–15, 1984.

[BI87] Manuel Blum and Russell Impagliazzo. Generic oracles and oracle classes. In *Proceedings of the 28th Annual IEEE Symposium on Foundations of Computer Science*, pages 118–126, 1987.

[Bik66] Algimantas Bikelis. Estimates of the remainder in a combinatorial central limit theorem. *Litovskii Matematicheskii Sbornik*, 6(3):323–346, 1966.

[BKK$^+$92] Jean Bourgain, Jeff Kahn, Gil Kalai, Yitzhak Katznelson, and Nathan Linial. The influence of variables in product spaces. *Israel Journal of Mathematics*, 77(1):55–64, 1992.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Publications Mathématiques de l'IHÉS*, 90(1):5–43, 1999.

[BL76] Herm Brascamp and Elliott Lieb. Best constants in Young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics*, 20(2):151–173, 1976.

[BL85] Michael Ben-Or and Nathan Linial. Collective coin flipping, robust voting schemes and minima of Banzhaf values. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pages 408–416, 1985.

[BL90] Michael Ben-Or and Nathan Linial. Collective coin flipping. In Silvio Micali and Franco Preparata, editors, *Randomness and Computation*, volume 5 of *Advances in Computing Research: A research annual*, pages 91–115. JAI Press, 1990.

[BL96] Dominique Bakry and Michel Ledoux. Lévy–Gromov's isoperimetric inequality for an infinite dimensional diffusion generator. *Inventiones mathematicae*, 123(1):259–281, 1996.

[BL98] Sergey Bobkov and Michel Ledoux. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *Journal of Functional Analysis*, 156(2):347–365, 1998.

[Bla57] Julian Blau. The existence of social welfare functions. *Econometrica*, 25(2):302–313, 1957.

[BLR90] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pages 73–83, 1990.

[Blu03] Avrim Blum. Learning a function of r relevant variables. In Bernhard Schölkopf and Manfred Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory*, volume 2777 of *Lecture Notes in Computer Science*, pages 731–733. Springer, 2003.

[BM00] Franck Barthe and Bernard Maurey. Some remarks on isoperimetry of Gaussian type. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, 36(4):419–434, 2000.

[Bob97] Sergey Bobkov. An isoperimetric inequality on the discrete cube and an elementary proof of the isoperimetric inequality in Gauss space. *Annals of Probability*, 25(1):206–214, 1997.

[Bog98] Vladimir Bogachev. *Gaussian Measures*. Mathematical Series and Monographs. American Mathematical Society, 1998.

[BOH90] Yigal Brandman, Alon Orlitsky, and John Hennessy. A spectral lower bound technique for the size of decision trees and two-level AND/OR circuits. *IEEE Transactions on Computers*, 39(2):282–287, 1990.

[Bol84] Erwin Bolthausen. An estimate of the remainder in a combinatorial central limit theorem. *Probability Theory and Related Fields*, 66(3):379–386, 1984.

[Bol01] Béla Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2001.

[Bon68] Aline Bonami. Ensembles $\Lambda(p)$ dans le dual de D^∞. *Annales de l'Institut Fourier*, 18(2):193–204, 1968.

[Bon70] Aline Bonami. Étude des coefficients Fourier des fonctions de $L^p(G)$. *Annales de l'Institut Fourier*, 20(2):335–402, 1970.

[Bop97] Ravi Boppana. The average sensitivity of bounded-depth circuits. *Information Processing Letters*, 63(5):257–261, 1997.

[Bor75] Christer Borell. The Brunn–Minkowski inequality in Gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975.

[Bor79] Christer Borell. On the integrability of Banach space valued Walsh polynomials. In *Séminaire de Probabilités, XIII*, volume 721 of *Lecture Notes in Mathematics*, pages 1–3. Springer, Berlin, 1979.

[Bor82] Christer Borell. Positivity improving operators and hypercontractivity. *Mathematische Zeitschrift*, 180(2):225–234, 1982.

[Bor84] Christer Borell. On polynomial chaos and integrability. *Probabability and Mathematical Statistics*, 3(2):191–203, 1984.

[Bor85] Christer Borell. Geometric bounds on the Ornstein–Uhlenbeck velocity process. *Probability Theory and Related Fields*, 70(1):1–13, 1985.

[Bou79] Jean Bourgain. Walsh subspaces of l^p product spaces. In *Séminaire D'Analyse Fonctionnelle*, pages IV.1–IV.9. École Polytechnique, Centre De Mathématiques, 1979.

[Bou99] Jean Bourgain. On sharp thresholds of monotone properties. *Journal of the American Mathematical Society*, 12(4):1046–1053, 1999. Appendix to the main paper, *Sharp thresholds of graph properties, and the k-sat problem* by Ehud Friedgut.

[BOW10] Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. *Machine Learning*, 80(2):273–294, 2010.

[BR73] Leonid Balashov and Aleksandr Rubinshtein. Series with respect to the Walsh system and their generalizations. *Journal of Soviet Mathematics*, 1(6):727–763, 1973.

[BR08] Béla Bollobás and Oliver Riordan. Random graphs and branching processes. In Béla Bollobás, Robert Kozma, and Dezső Miklós, editors, *Handbook of large-scale random networks*, pages 15–116. Springer, 2008.

[Bra87] Yigal Brandman. *Spectral lower-bound techniques for logic circuits*. PhD thesis, Stanford University, 1987.

[Bru90] Jehoshua Bruck. Harmonic analysis of polynomial threshold functions. *SIAM Journal on Discrete Mathematics*, 3(2):168–177, 1990.

[BS92] Jehoshua Bruck and Roman Smolensky. Polynomial threshold functions, AC^0 functions and spectral norms. *SIAM Journal on Computing*, 21(1):33–42, 1992.

[BS08] Eli Ben-Sasson and Madhu Sudan. Short PCPs with polylog query complexity. *SIAM Journal on Computing*, 38(2):551–607, 2008.

[BSGH$^+$04] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs and applications to coding. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 1–10, 2004.

[BSSVW03] Eli Ben-Sasson, Madhu Sudan, Salil Vadhan, and Avi Wigderson. Randomness-efficient low degree tests and short PCPs via epsilon-biased sets. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 612–621, 2003.

[BSW05] Itai Benjamini, Oded Schramm, and David Wilson. Balanced Boolean functions that can be evaluated so that every input bit is unlikely to be read. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 244–250, 2005.

[BT76] Albert Baernstein and Bert Taylor. Spherical rearrangements, subharmonic functions, and $*$-functions in n-space. *Duke Mathematical Journal*, 43(2):245–268, 1976.

[BT87] Béla Bollobás and Andrew Thomason. Threshold functions. *Combinatorica*, 7(1):35–38, 1987.

[BT96] Nader Bshouty and Christino Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

[BT09] Avraham Ben-Aroya and Amnon Ta-Shma. Constructing small-bias sets from algebraic-geometric codes. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 191–197, 2009.

[BV07] Andrej Bogdanov and Emanuele Viola. Pseudorandom bits for polynomials. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 41–51, 2007.

[Car10] Claude Carlet. Boolean functions for cryptography and error-correcting codes. In Yves Crama and Peter Hammer, editors, *Boolean models and methods in mathematics, computer science, and engineering*, pages 257–397. Cambridge University Press, 2010.

[CFG$^+$85] Benny Chor, Joel Friedmann, Oded Goldreich, Johan Håstad, Steven Rudich, and Roman Smolensky. The bit extraction problem or t-resilient functions. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pages 396–407, 1985.

[CG92] Fan Chung and Ronald Graham. Quasi-random subsets of $\mathbb{Z}_n$. *Journal of Combinatorial Theory, Series A*, 61:64–86, 1992.

[CGG87] Benny Chor and Mihály Geréb-Graus. On the influence of single participant in coin flipping schemes. Technical report, Harvard University, 1987.

[CGG88] Benny Chor and Mihály Geréb-Graus. On the influence of single participant in coin flipping schemes. *SIAM Journal on Discrete Mathematics*, 1(4):411–415, 1988.

[CGS11] Louis Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer, 2011.

[CGW89] Fan Chung, Ronald Graham, and Richard Wilson. Quasi-random graphs. *Combinatorica*, 9(4):345–362, 1989.

[Che60] Pafnuty Chebyshev. Sur le développement des fonctions à une seule variable. *Bulletin de l'Académie impériale des sciences de St.-Pétersbourg*, 1:193–200, 1860.

[CHL97] Mireille Capitaine, Elton Hsu, and Michel Ledoux. Martingale representation and a simple proof of logarithmic Sobolev inequalities on path spaces. *Electronic Communications in Probability*, 2:71–81, 1997.

[Cho61] Chao-Kong Chow. On the characterization of threshold functions. In *Proceedings of the 2nd Annual Symposium on Switching Circuit Theory and Logical Design (FOCS)*, pages 34–38, 1961.

[CKS01] Nadia Creignou, Sanjeev Khanna, and Madhu Sudan. *Complexity classifications of Boolean constraint satisfaction problems*. Society for Industrial and Applied Mathematics, 2001.

[CL90] Eric Carlen and Michael Loss. Extremals of functionals with competing symmetries. *Journal of Functional Analysis*, 88(2):437–456, 1990.

[Col71] John Coleman. Control of collectivities and the power of a collectivity to act. In Bernhardt Lieberman, editor, *Social Choice*. Gordon and Breach, 1971.

[CW01] Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

[CW04] Moses Charikar and Anthony Wirth. Maximizing quadratic programs: extending Grothendieck's Inequality. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 54–60, 2004.

[dC85] Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris, de l'imprimerie royale, 1785.

[DF87] Persi Diaconis and David Freedman. A dozen de Finetti-style results in search of a theorem. *Annales de l'Institut Henri Poincaré (B)*, 23(S2):397–423, 1987.

[DFKO07] Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell. On the Fourier tails of bounded functions over the discrete cube. *Israel Journal of Mathematics*, 160(1):389–412, 2007.

[DHK$^+$10] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco Servedio, and Li-Yang Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 533–542, 2010.

[Dic01] Leonard Dickson. *Linear groups with an exposition of Galois field theory*. B. G. Teubner, 1901.

[Dil72] John Dillon. A survey of bent functions. *NSA Technical Journal*, pages 191–215, 1972.

[Din07] Irit Dinur. The PCP Theorem by gap amplification. *Journal of the ACM*, 54(3):1–44, 2007.

[dKPW04] Etienne de Klerk, Dmitrii Pasechnik, and Johannes Warners. On approximate graph colouring and MAX-k-CUT algorithms based on the ϑ-function. *Journal of Combinatorial Optimization*, 8(3):267–294, 2004.

[DMN12] Anindya De, Elchanan Mossel, and Joe Neeman. Majority is Stablest : discrete and SoS. Technical Report 1211.1001, arXiv, 2012.

[DMN13] Anindya De, Elchanan Mossel, and Joe Neeman. Majority is Stablest : Discrete and SoS. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, 2013.

[DP93] Charles Delorme and Svatopluk Poljak. Laplacian eigenvalues and the maximum cut problem. *Mathematical Programming*, 62(1–3):557–574, 1993.

[DR04] Irit Dinur and Omer Reingold. Assignment testers: Towards a combinatorial proof of the PCP Theorem. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 155–164, 2004.

[DS09] Ilias Diakonikolas and Rocco Servedio. Improved approximation of linear threshold functions. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity*, pages 161–172, 2009.

[DSC96] Persi Diaconis and Laurent Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Annals of Applied Probability*, 6(3):695–750, 1996.

[Ehr83] Antoine Ehrhard. Symétrisation dans l'espace de gauss. *Mathematica Scandinavica*, 53:281–301, 1983.

[Ehr84] Antoine Ehrhard. Inégalités isopérimétriques et intégrales de Dirichlet gaussiennes. *Annales Scientifiques de l'École Normale Supérieure. Quatrième Série*, 17(2):317–332, 1984.

[Ein05] Albert Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der physik*, 322(8):549–560, 1905.

[EKR99] Funda Ergün, Ravi Kumar, and Ronitt Rubinfeld. Fast approximate PCPs. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 41–50, 1999.

[Eld13] Ronen Eldan. A two-sided estimate for the Gaussian noise stability deficit. Technical Report 1307.2781, arXiv, 2013.

[Elg61] Calvin Elgot. Truth functions realizable by single threshold organs. In *Proceedings of the 2nd Annual Symposium on Switching Circuit Theory and Logical Design (FOCS)*, pages 225–245, 1961.

[Enf70] Per Enflo. On the nonexistence of uniform homeomorphisms between l_p-spaces. *Arkiv för matematik*, 8(2):103–105, 1970.

[Epp89] Jay Epperson. The hypercontractive approach to exactly bounding an operator with complex Gaussian kernel. *Journal of Functional Analysis*, 87(1):1–30, 1989.

[ER59] Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[ES81] Bradley Efron and Charles Stein. The jackknife estimate of variance. *Annals of Statistics*, 9(3):586–596, 1981.

[Ess42] Carl-Gustav Esseen. On the Liapounoff limit of error in the theory of probability. *Arkiv för matematik, astronomi och fysik*, 28(9):1–19, 1942.

[Fed69] Paul Federbush. Partially alternate derivation of a result of Nelson. *Journal of Mathematical Physics*, 10:50–52, 1969.

[FGL$^+$96] Uriel Feige, Shafi Goldwasser, László Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM*, 43(2):268–292, 1996.

[Fin49] Nathan Fine. On the Walsh functions. *Transactions of the American Mathematical Society*, 65(3):372–414, 1949.

[FJS91] Merrick Furst, Jeffrey Jackson, and Sean Smith. Improved learning of AC^0 functions. In *Proceedings of the 4th Annual Conference on Learning Theory*, pages 317–325, 1991.

[FK96] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the American Mathematical Society*, 124(10):2993–3002, 1996.

[FKN02] Ehud Friedgut, Gil Kalai, and Assaf Naor. Boolean functions whose Fourier transform is concentrated on the first two levels and neutral social choice. *Advances in Applied Mathematics*, 29(3):427–437, 2002.

[FL92] Uriel Feige and Lászlo Lovász. Two-prover one-round proof systems, their power and their problems. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pages 733–744, 1992.

[FL06] Uriel Feige and Michael Langberg. The RPR^2 rounding technique for semidefinite programs. *Journal of Algorithms*, 60(1):1–23, 2006.

[FLM77] Tadeusz Figiel, Joram Lindenstrauss, and Vitali Milman. The dimension of almost spherical sections of convex bodies. *Acta Mathematica*, 139(1-2):53–94, 1977.

[Fre79] Rūsiņš Freivalds. Fast probabilistic algorithms. In *Proceedings of the 4th Annual International Symposium on Mathematical Foundations of Computer Science*, pages 57–69, 1979.

[Fri98] Ehud Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–36, 1998.

[Fri99] Ehud Friedgut. Sharp thresholds of graph properties, and the k-SAT problem. *Journal of the American Mathematical Society*, 12(4):1017–1054, 1999.

[Fri05] Ehud Friedgut. Hunting for sharp thresholds. *Random Structures & Algorithms*, 26(1-2):37–51, 2005.

[FS95] Katalin Friedl and Madhu Sudan. Some improvements to total degree tests. In *Proceedings of the 3rd Annual Israel Symposium on Theory of Computing Systems*, pages 190–198, 1995.

[FS02] Uriel Feige and Gideon Schechtman. On the optimality of the random hyperplane rounding technique for Max-Cut. *Randoom Structures and Algorithms*, 20(3):403–440, 2002.

[FS07] Dvir Falik and Alex Samorodnitsky. Edge-isoperimetric inequalities and influences. *Combinatorics, Probability and Computing*, 16(5):693–712, 2007.

[FSS84] Merrick Furst, James Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17(1):13–27, 1984.

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connections to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[Gir73] Vyacheslav Girko. Limit theorems for random quadratic forms. *Izdat. Naukova Dumka*, pages 14–30, 1973.

[GK68] Mark Garman and Morton Kamien. The paradox of voting: probability calculations. *Behavioral Science*, 13(4):306–316, 1968.

[GKK08] Parikshit Gopalan, Adam Kalai, and Adam Klivans. Agnostically learning decision trees. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 527–536, 2008.

[GL89] Oded Goldreich and Leonid Levin. A hard-core predicate for all one-way functions. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, pages 25–32, 1989.

[GL94] Craig Gotsman and Nathan Linial. Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994.

[Gli68] James Glimm. Boson fields with nonlinear selfinteraction in two dimensions. *Communications in Mathematical Physics*, 8(1):12–25, 1968.

[GMR12] Parikshit Gopalan, Raghu Meka, and Omer Reingold. DNF sparsification and a faster deterministic counting algorithm. In *Proceedings of the 26th Annual IEEE Conference on Computational Complexity*, pages 126–135, 2012.

[Gol59] Solomon Golomb. On the classification of Boolean functions. *IRE Transactions on Circuit Theory*, 6(5):176–186, 1959.

[GOS^{+}11] Parikshit Gopalan, Ryan O'Donnell, Rocco Servedio, Amir Shpilka, and Karl Wimmer. Testing Fourier dimensionality and sparsity. *SIAM Journal on Computing*, 40(4):1075–1100, 2011.

[Gow01] W. Timothy Gowers. A new proof of Szemerédi's theorem. *Geometric and Functional Analysis*, 11(3):465–588, 2001.

[GOWZ10] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity*, pages 223–234, 2010.

[GR00] Mikael Goldmann and Alexander Russell. Spectral bounds on general hard core predicates. In *Proceedings of the 17th Annual Symposium on Theoretical Aspects of Computer Science*, pages 614–625, 2000.

[Gro72] Leonard Gross. Existence and uniqueness of physical ground states. *Journal of Functional Analysis*, 10:52–109, 1972.

[Gro75] Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[GS08] Ben Green and Tom Sanders. Boolean functions with small spectral norm. *Geometric and Functional Analysis*, 18(1):144–162, 2008.

[Gui52] George-Théodule Guilbaud. Les théories de l'intérêt général et le problème logique de l'agrégation. *Economie appliquée*, V(4):501–551, 1952.

[GW94] Michel Goemans and David Williamson. A 0.878 approximation algorithm for MAX-2SAT and MAX-CUT. In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, pages 422–431, 1994.

[GW95] Michel Goemans and David Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

[Haa10] Alfréd Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

[Haa82] Uffe Haagerup. The best constants in the Khinchine inequality. *Studia Mathematica*, 70(3):231–283, 1982.

[Háj68] Jaroslav Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *Annals of Mathematical Statistics*, 39(2):325–346, 1968.

[Har64] Lawrence Harper. Optimal assignments of numbers to vertices. *Journal of the Society for Industrial and Applied Mathematics*, 12(1):131–135, 1964.

[Hås87] Johan Håstad. *Computational Limitations for Small Depth Circuits*. MIT Press, 1987.

[Hås96] Johan Håstad. Testing of the long code and hardness for clique. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*, pages 11–19, 1996.

[Hås97] Johan Håstad. Some optimal inapproximability results. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pages 1–10, 1997.

[Hås99] Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182(1):105–142, 1999.

[Hås01a] Johan Håstad. A slight sharpening of LMN. *Journal of Computer and System Sciences*, 63(3):498–508, 2001.

[Hås01b] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001.

[Hås12] Johan Håstad. On the correlation of parity and small-depth circuits. Technical Report TR12-137, Electronic Colloquium on Computational Complexity, 2012.

[Hat12] Hamed Hatami. A structure theorem for Boolean functions with small total influences. *Annals of Mathematics*, 176(1):509–533, 2012.

[Her64] Charles Hermite. Sur un nouveau développement en série des fonctions. *Comptes rendus de l'Académie des sciences*, 58(2):93–100, 266–273, 1864.

[Hin10] Masanori Hino. Sets of finite perimeter and the Hausdorff-Gauss measure on the Wiener space. *Journal of Functional Analysis*, 258(5):1656–1681, 2010.

[HKM10] Prahladh Harsha, Adam Klivans, and Raghu Meka. Bounding the sensitivity of polynomial threshold functions. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 533–542, 2010.

[HMM82] Stanley Hurst, D. Michael Miller, and Jon Muzio. Spectral method of Boolean function complexity. *Electronics Letters*, 18(13):572–574, 1982.

[Hoe48] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[HY95] Yasunari Higuchi and Nobuo Yoshida. Analytic conditions and phase transition for Ising models. Unpublished lecture notes (in Japanese), 1995.

[IMP12] Russell Impagliazzo, William Matthews, and Ramamohan Paturi. A satisfiability algorithm for AC^0. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 961–972, 2012.

[Jac95] Jeffrey Jackson. *The Harmonic Sieve: a novel application of Fourier analysis to machine learning theory and practice*. PhD thesis, Carnegie Mellon University, 1995.

[Jac97] Jeffrey Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, 1997.

[Jan97] Svante Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, 1997.

[JKS03] T. S. Jayram, Ravi Kumar, and D. Sivakumar. Two applications of information complexity. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 673–682, 2003.

[Joh74] David Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, 1974.

[JOW12] Jacek Jendrej, Krzysztof Oleszkiewicz, and Jakub Wojtaszczyk. On some extensions of the FKN theorem. Manuscript, 2012.

[JZ11] Rahul Jain and Shengyu Zhang. The influence lower bound via query elimination. *Theory of Computing*, 7(1):147–153, 2011.

[Kah68] Jean-Pierre Kahane. *Some random series of functions*. D. C. Heath & Co., 1968.

[Kal02] Gil Kalai. A Fourier-theoretic perspective on the Condorcet paradox and Arrow's theorem. *Advances in Applied Mathematics*, 29(3):412–426, 2002.

[Kan11] Daniel Kane. *On Elliptic Curves, the ABC Conjecture, and Polynomial Threshold Functions*. PhD thesis, Harvard University, 2011.

[Kan12] Daniel Kane. The correct exponent for the Gotsman–Linial conjecture. Technical Report 1210.1283, arXiv, 2012.

[Kar76] Mark Karpovsky. *Finite orthogonal series in the design of digital devices: analysis, synthesis, and optimization*. Wiley, 1976.

[Kar99] Howard Karloff. How good is the Goemans–Williamson MAX CUT algorithm? *SIAM Journal of Computing*, 29(1):336–350, 1999.

[Kha93] Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, pages 372–381, 1993.

[Kho02] Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 767–775, 2002.

[Kho05] Subhash Khot. Inapproximability results via Long Code based PCPs. *ACM SIGACT News*, 36(2):25–42, 2005.

[Kho10a] Subhash Khot. Inapproximability of NP-complete problems, discrete Fourier analysis, and geometry. In *Proceedings of the International Congress of Mathematicians*, volume 901, pages 2676–2697, 2010.

[Kho10b] Subhash Khot. On the Unique Games Conjecture. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity*, pages 99–121, 2010.

[Kie69] Konrad Kiener. *Über Produkte von quadratisch integrierbaren Funktionen endlicher Vielfalt*. PhD thesis, Universität Innsbruck, 1969.

[Kin02] Guy Kindler. *Property Testing, PCP, and juntas*. PhD thesis, Tel Aviv University, 2002.

[KK07] Jeff Kahn and Gil Kalai. Thresholds and expectation thresholds. *Combinatorics, Probability and Computing*, 16(3):495–502, 2007.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on Boolean functions. In *Proceedings of the 29th Annual IEEE Symposium on Foundations of Computer Science*, pages 68–80, 1988.

[KKMO04] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 146–154, 2004.

[KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for Max-Cut and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.

[Kle66] Daniel Kleitman. Families of non-disjoint subsets. *Journal of Combinatorial Theory*, 1(1):153–155, 1966.

[KLX10] Tali Kaufman, Simon Litsyn, and Ning Xie. Breaking the ϵ-soundness bound of the linearity test over GF(2). *SIAM Journal on Computing*, 39(5):1988–2003, 2010.

[KM93] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.

[KO06] Subhash Khot and Ryan O'Donnell. SDP gaps and UGC-hardness for Max-Cut-Gain. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 217–226, 2006.

[KO12] Guy Kindler and Ryan O'Donnell. Gaussian noise sensitivity and Fourier tails. In *Proceedings of the 26th Annual IEEE Conference on Computational Complexity*, pages 137–147, 2012.

[KOS04] Adam Klivans, Ryan O'Donnell, and Rocco Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004.

[KOS08] Adam Klivans, Ryan O'Donnell, and Rocco Servedio. Learning geometric concepts via Gaussian surface area. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550, 2008.

[KOTZ13] Manuel Kauers, Ryan O'Donnell, Li-Yang Tan, and Yuan Zhou. Hypercontractive inequalities via SOS, with an application to Vertex-Cover. Manuscript, 2013.

[KP97] Matthias Krause and Pavel Pudlák. On the computational power of depth-2 circuits with threshold and modulo gates. *Theoretical Computer Science*, 174(1–2):137–156, 1997.

[KR82] Samuel Karlin and Yosef Rinott. Applications of ANOVA type decompositions for comparisons of conditional variance statistics including jackknife estimates. *Annals of Statistics*, 10(2):485–501, 1982.

[KR08] Subhash Khot and Oded Regev. Vertex Cover might be hard to approximate to within $2-\epsilon$. *Journal of Computer and System Sciences*, 74(3):335–349, 2008.

[Kra29] Mikahil (Krawtchouk) Kravchuk. Sur une généralisation des polynomes d'Hermite. *Comptes rendus de l'Académie des sciences*, 189:620–622, 1929.

[KS88] Wiesław Krakowiak and Jerzy Szulga. Hypercontraction principle and random multilinear forms. *Probability Theory and Related Fields*, 77(3):325–342, 1988.

[KS02] Guy Kindler and Shmuel Safra. Noise-resistant Boolean functions are juntas. Manuscript, 2002.

[KSTW01] Sanjeev Khanna, Madhu Sudan, Luca Trevisan, and David Williamson. The approximability of constraint satisfaction problems. *SIAM Journal on Computing*, 30(6):1863–1920, 2001.

[KV05] Subhash Khot and Nisheeth Vishnoi. The Unique Games Conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 53–62, 2005.

[KW92] Stanisław Kwapień and Wojbor Woyczyński. *Random series and stochastic integrals: Single and multiple*. Probability and Its Applications. Birkhäuser, 1992.

[Kwa10] Stanisław Kwapień. On Hoeffding decomposition in l_p. *Illinois Journal of Mathematics*, 54(3):1205–1211, 2010.

[KZ97] Howard Karloff and Uri Zwick. A 7/8-approximation algorithm for MAX 3SAT? In *Proceedings of the 38th Annual IEEE Symposium on Foundations of Computer Science*, pages 406–415, 1997.

[Lap11] Pierre-Simon Laplace. Mémoire sur les intégrales définies et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations. *Mémoires de la Classe des Sciences Mathématiques et Physiques de l'Institut Impérial de France, Année 1810*, 58:279–347, 1811.

[Leb72] Nikolaĭ Lebedev. *Special functions & their applications*. Dover Publications, 1972.

[Lec63] Robert Lechner. *Affine equivalence of switching functions*. PhD thesis, Harvard University, 1963.

[Lec71] Robert Lechner. Harmonic analysis of switching functions. In Amar Mukhophadhay, editor, *Recent developments in switching theory*, pages 121–228. Academic Press, 1971.

[Led92] Michel Ledoux. On an integral criterion for hypercontractivity of diffusion semigroups and extremal functions. *Journal of Functional Analysis*, 105(2):444–465, 1992.

[Led94] Michel Ledoux. Semigroup proofs of the isoperimetric inequality in Euclidean and Gauss space. *Bulletin des Sciences Mathématiques*, 118(6):485–510, 1994.

[Led96] Michel Ledoux. Isoperimetry and Gaussian analysis. In Pierre Bernard, editor, *Lectures on Probability Theory and Statistics*, volume XXIV of *Lecture Notes in Mathematics 1648*, pages 165–294. Springer, 1996.

[Led98] Michel Ledoux. A short proof of the Gaussian isoperimetric inequality. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progress in Probability*, pages 229–232. Birkhäuser, Basel, 1998.

[Led13] Michel Ledoux. Remarks on noise sensitivity, Brascamp–Lieb and Slepian inequalities. http://perso.math.univ-toulouse.fr/ledoux/files/2013/11/noise.pdf, 2013.

[Lee10] Homin Lee. Decision trees and influence: an inductive proof of the OSSS Inequality. *Theory of Computing*, 6(1):81–84, 2010.

[Leo12] Nikos Leonardos. An improved lower bound for the randomized decision tree complexity of recursive majority. Technical Report TR12-099, Electronic Colloquium on Computational Complexity, 2012.

[Lév22] Paul Lévy. *Leçons d'Analyse Fonctionnelle*. Gauthier-Villars, 1922.

[LG14] François Le Gall. Powers of tensors and fast matrix multiplication. Technical Report 1401.7714, arXiv, 2014.

[Lin22] Jarl Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922.

[LLS06] Sophie Laplante, Troy Lee, and Mario Szegedy. The quantum adversary method and classical formula size lower bounds. *Computational Complexity*, 15(2):163–196, 2006.

[LMN89] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform and Learnability. In *Proceedings of the 30th Annual IEEE Symposium on Foundations of Computer Science*, pages 574–579, 1989.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[LO94] Rafał Latała and Krzysztof Oleszkiewicz. On the best constant in the Khintchine–Kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994.

[Lov08] Shachar Lovett. Unconditional pseudorandom generators for low degree polynomials. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 557–562, 2008.

[LSP82] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

[LT09] Shachar Lovett and Yoav Tzur. Explicit lower bound for fooling polynomials by the sum of small-bias generators. In *Electronic Colloquium on Computational Complexity TR09-088*, 2009.

[LVW93] Michael Luby, Boban Veličković, and Avi Wigderson. Deterministic approximate counting of depth-2 circuits. In *Proceedings of the 2nd Annual Israel Symposium on Theory of Computing Systems*, pages 18–24, 1993.

[Man94] Yishay Mansour. Learning Boolean functions via the Fourier Transform. In Vwani Roychowdhury, Kai-Yeung Siu, and Alon Orlitsky, editors, *Theoretical Advances in Neural Computation and Learning*, chapter 11, pages 391–424. Kluwer Academic Publishers, 1994.

[Man95] Yishay Mansour. An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *Journal of Computer and System Sciences*, 50(3):543–550, 1995.

[Mar74] Grigory Margulis. Probabilistic characteristics of graphs with large connectivity. *Problemy Peredači Informacii*, 10(2):101–108, 1974.

[May52] Kenneth May. A set of independent necessary and sufficient conditions for simple majority decisions. *Econometrica*, 20(4):680–684, 1952.

[McK73] Henry McKean. Geometry of differential space. *Annals of Probability*, 1(2):197–206, 1973.

[Meh66] F. Gustav Mehler. Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer ordnung. *Journal für die reine und angewandte Mathematik*, 66:161–176, 1866.

[Mid04] Gatis Midrijānis. Exact quantum query complexity for total Boolean functions. arXiv:quant-ph/0403168, 2004.

[Mir13] Michele Miranda Jr. Personal communication to the author, October 2013.

[MN12] Elchanan Mossel and Joe Neeman. Robust optimality of Gaussian noise stability. Technical Report 1210.4126, arXiv, 2012.

[MNP12] Michele Miranda Jr., Matteo Novaga, and Diego Pallara. An introduction to BV functions in Wiener spaces. Technical Report 1212.5926, arXiv, 2012.

[MNSX11] Frédéric Magniez, Ashwin Nayak, Miklos Santha, and David Xiao. Improved bounds for the randomized decision tree complexity of recursive majority. In *Proceedings of the 38th Annual International Colloquium on Automata, Languages and Programming*, pages 317–329, 2011.

[MO05] Elchanan Mossel and Ryan O'Donnell. Coin flipping from a cosmic source: On error correction of truly random bits. *Random Structures & Algorithms*, 26(4):418–436, 2005.

[MOO05a] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 21–30, 2005.

[MOO05b] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. Technical Report math/0503503, arXiv, 2005.

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Annals of Mathematics*, 171(1):295–341, 2010.

[MOR$^+$06] Elchanan Mossel, Ryan O'Donnell, Oded Regev, Jeffrey Steif, and Benjamin Sudakov. Non-interactive correlation distillation, inhomogeneous Markov chains, and the reverse Bonami–Beckner inequality. *Israel Journal of Mathematics*, 154:299–336, 2006.

[MORS10] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco Servedio. Testing halfspaces. *SIAM Journal on Computing*, 39(5):2004–2047, 2010.

[MOS04] Elchanan Mossel, Ryan O'Donnell, and Rocco Servedio. Learning functions of k relevant variables. *Journal of Computer and System Sciences*, 69(3):421–434, 2004.

[Mos10] Elchanan Mossel. Gaussian bounds for noise correlation of functions. *Geometric and Functional Analysis*, 19(6):1713–1756, 2010.

[MOS12] Elchanan Mossel, Krzysztof Oleszkiewicz, and Arnab Sen. On reverse hypercontractivity. Technical Report 1108.1210, arXiv, 2012.

[MP11] Gretchen Matthews and Justin Peachey. Small-bias sets from extended norm-trace codes. Manuscript, 2011.

[MPPP07] Michele Miranda Jr., Diego Pallara, Fabio Paronetto, and Marc Preunkert. Short-time heat flow and functions of bounded variation in $\mathbf{R}^N$. *Annales de la Faculté des Sciences de Toulouse. Mathématiques. Série 6*, 16(1):125–145, 2007.

[MRRW77] Robert McEliece, Eugene Rodemich, Howard Rumsey, and Lloyd Welch. New upper bounds on the rate of a code via the Delsarte–MacWilliams inequalities. *IEEE Transactions on Information Theory*, 23(2):157–166, 1977.

[MS73] Tamás Matolcsi and József Szücs. Intersection des mesures spectrales conjuguées. *Comptes rendus de l'Académie des sciences*, 277:841–843, 1973.

[MS77] F. Jessie MacWilliams and Neil Sloane. *The theory of error-correcting codes*. North-Holland, 1977.

[Mul54a] David Muller. Application of Boolean algebra to switching circuit design and to error detection. *IRE Transactions on Electronic Computers*, 3(6):6–12, 1954.

[Mul54b] David Muller. Boolean algebras in electric circuit design. *The American Mathematical Monthly*, 61(7):27–28, 1954.

[Mül05] Paul Müller. *Isomorphisms between H^1 spaces*, volume 66 of *Monografie Matematyczne*. Birkhäuser Verlag, 2005.

[Nak35] Akira Nakashima. The theory of relay circuit composition. *The Journal of the Institute of Telegraph and Telephone Engineers of Japan*, 150:731–752, September 1935.

[Naz03] Fedor Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*, volume 1807, pages 169–187. Israel Seminar, 2003.

[Nel66] Edward Nelson. A quartic interaction in two dimensions. In *Mathematical Theory of Elementary Particles*, pages 69–73. MIT Press, 1966.

[Nel73] Edward Nelson. The free Markoff field. *Journal of Functional Analysis*, 12:211–227, 1973.

[Nev76] Jacques Neveu. Sur l'espérance conditionnelle par rapport à un mouvement brownien. *Annales de l'Institut Henri Poincaré (B)*, 12(2):105–109, 1976.

[Nin58] Ichizo Ninomiya. A theory of the coordinate representations of switching functions. *Memoirs of the Faculty of Engineering, Nagoya University*, 10:175–190, 1958.

[NN93] Joseph Naor and Moni Naor. Small-bias probability spaces: efficient constructions and applications. *SIAM Journal on Computing*, 22(4):838–856, 1993.

[NP00] Fedor Nazarov and Anatoliy Podkorytov. Ball, Haagerup, and distribution functions. *Complex Analysis, Operators, and Related Topics. Operator Theory: Advances and Applications*, 113:247–267, 2000.

[NS94] Noam Nisan and Mario Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994.

[NSV02] Fedor Nazarov, Mikhail Sodin, and Alexander Vol'berg. The geometric Kannan–Lovász–Simonovits lemma, dimension-free estimates for volumes of sublevel sets of polynomials, and distribution of zeros of random analytic functions. *Algebra i Analiz*, 14(2):214–234, 2002.

[NW95] Noam Nisan and Avi Wigderson. On rank vs. communication complexity. *Combinatorica*, 15(4):557–565, 1995.

[O'D03] Ryan O'Donnell. *Computational applications of noise sensitivity*. PhD thesis, Massachusetts Institute of Technology, 2003.

[O'D04] Ryan O'Donnell. Hardness amplification within NP. *Journal of Computer and System Sciences*, 69(1):68–94, 2004.

[Ole03] Krzysztof Oleszkiewicz. On a nonsymmetric version of the Khinchine–Kahane inequality. In Evariste Giné, Christian Houdré, and David Nualart, editors, *Stochastic inequalities and applications*, volume 56, pages 157–168. Birkhäuser, 2003.

[OS06] Ryan O'Donnell and Rocco Servedio. Learning monotone decision trees in polynomial time. In *Proceedings of the 21st Annual IEEE Conference on Computational Complexity*, pages 213–225, 2006.

[OS07] Ryan O'Donnell and Rocco Servedio. Learning monotone decision trees in polynomial time. *SIAM Journal on Computing*, 37(3):827–844, 2007.

[OS08] Ryan O'Donnell and Rocco Servedio. Extremal properties of polynomial threshold functions. *Journal of Computer and System Sciences*, 74(3):298–312, 2008.

[OSSS05] Ryan O'Donnell, Michael Saks, Oded Schramm, and Rocco Servedio. Every decision tree has an influential variable. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 31–39, 2005.

[OW08] Ryan O'Donnell and Yi Wu. An optimal SDP algorithm for Max-Cut, and equally optimal Long Code tests. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 335–344, 2008.

[OW09] Ryan O'Donnell and Yi Wu. 3-Bit dictator testing: 1 vs. 5/8. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 365–373, 2009.

[OW12] Ryan O'Donnell and John Wright. A new point of NP-hardness for Unique-Games. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 289–306, 2012.

[OW13] Ryan O'Donnell and Karl Wimmer. Sharpness of KKL on Schreier graphs. *Electronic Communications in Probability*, 18:1–12, 2013.

[Pal32] Raymond Paley. A remarkable series of orthogonal functions (I). *Proceedings of the London Mathematical Society*, 2(1):241–264, 1932.

[Pen46] Lionel Penrose. The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57, 1946.

[Per90] René Peralta. On the randomness complexity of algorithms. Technical Report 90-1, University of Wisconsin, Milwaukee, 1990.

[Per04] Yuval Peres. Noise stability of weighted majority. arXiv:math/0412377, 2004.

[Pis86] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. In *Probability and analysis (Varenna, 1985)*, volume 1206 of *Lecture Notes in Mathematics*, pages 167–241. Springer, Berlin, 1986.

[PRS01] Michael Parnas, Dana Ron, and Alex Samorodnitsky. Proclaiming dictators and juntas or testing Boolean formulae. In *Proceedings of the 5th Annual International Workshop on Randomized Techniques in Computation*, pages 273–284, 2001.

[PRS02] Michal Parnas, Dana Ron, and Alex Samorodnitsky. Testing basic Boolean formulae. *SIAM Journal on Discrete Mathematics*, 16(1):20–46, 2002.

[PZ78] Gilles Pisier and Joel Zinn. On the limit theorems for random variables with values in the spaces l_p $(2 \le p < \infty)$. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 41(4):289–304, 1978.

[Rag08] Prasad Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 245–254, 2008.

[Rag09] Prasad Raghavendra. *Approximating NP-hard problems: efficient algorithms and their limits*. PhD thesis, University of Washington, 2009.

[Rao47] Calyampudi Rao. Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of the Royal Statistical Society*, 9(1):128–139, 1947.

[Raz93] Alexander Razborov. Bounded arithmetic and lower bounds in boolean complexity. In Peter Clote and Jeffrey Remmel, editors, *Feasible Mathematics II*, pages 344–386. Birkhäuser, 1993.

[Rik61] William Riker. Voting and the summation of preferences: An interpretive bibliographic review of selected developments during the last decade. *American Political Science Review*, 55(4):900–911, 1961.

[Ros76] Haskell Rosenthal. Convolution by a biased coin. In *The Altgeld Book 1975/1976*, pages II.1–II.17. University of Illinois, 1976.

[Ros06] Raphaël Rossignol. Threshold for monotone symmetric properties through a logarithmic Sobolev inequality. *Annals of Probability*, 34(5):1707–1725, 2006.

[Rot53] Klaus Roth. On certain sets of integers. *Journal of the London Mathematical Society*, 28(1):104–109, 1953.

[Rot73] Vladimir Rotar'. Some limit theorems for polynomials of second order. *Teoriya Veroyatnostei i ee Primeneniya*, 18(3):527–534, 1973.

[Rot74] Vladimir Rotar'. Some limit theorems for polynomials of second degree. *Theory of Probability and its Applications*, 18(3):499–507, 1974.

[Rot75] Vladimir Rotar'. Limit theorems for multilinear forms and quasipolynomial functions. *Teoriya Veroyatnostei i ee Primeneniya*, 20(3):527–546, 1975.

[Rot76] Oscar Rothaus. On "bent" functions. *Journal of Combinatorial Theory, Series A*, 20(3):300–305, 1976.

[Rot79] Vladimir Rotar'. Limit theorems for polylinear forms. *Journal of Multivariate Analysis*, 9(4):511–530, 1979.

[Rot88] Alvin Roth, editor. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[Rou62] Jean-Jacques Rousseau. *Du Contrat Social*. Marc-Michel Rey, 1762.

[RR01] Yosef Rinott and Vladimir Rotar'. A remark on quadrant normal probabilities in high dimensions. *Statistics & Probability Letters*, 51(1):47–51, 2001.

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

[Rud62] Walter Rudin. *Fourier analysis on groups*. John Wiley & Sons, 1962.

[Rus81] Lucio Russo. On the critical percolation probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56(2):229–237, 1981.

[Rus82] Lucio Russo. An approximate zero-one law. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(1):129–139, 1982.

[RV80] Herman Rubin and Richard Vitale. Asymptotic distribution of symmetric statistics. *Annals of Statistics*, 8(1):165–170, 1980.

[Sae68] Sadahiro Saeki. On norms of idempotent measures. *Proceedings of the American Mathematical Society*, 19(3):600–602, 1968.

[SB91] Kai-Yeung Siu and Jehoshua Bruck. On the power of threshold circuits with small weights. *SIAM Journal on Discrete Mathematics*, 4(3):423–435, 1991.

[Sch48] Erhard Schmidt. Die Brunn-Minkowskische Ungleichung und ihr Spiegelbild sowie die isoperimetrische Eigenschaft der Kugel in der euklidischen und nichteuklidischen Geometrie. I. *Mathematische Nachrichten*, 1:81–157, 1948.

[Sch67] Michel Schreiber. Fermeture en probabilité des chaos de Wiener. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Séries A*, 265:859–861, 1967.

[Sch69] Michel Schreiber. Fermeture en probabilité de certains sous-espaces d'un espace L^2. Application aux chaos de Wiener. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14:36–48, 1969.

[Seg70] Irving Segal. Construction of non-linear local quantum processes: I. *Annals of Mathematics*, 92:462–481, 1970.

[Sha37] Claude Shannon. A symbolic analysis of relay and switching circuits. Master's thesis, Massachusetts Institute of Technology, 1937.

[Sha53] Lloyd Shapley. A value for n-person games. In Harold Kuhn and Albert Tucker, editors, *Contributions in the Theory of Games, volume II*, pages 307–317. Princeton University Press, 1953.

[She99] William Sheppard. On the application of the theory of error to cases of normal distribution and normal correlation. *Philosophical Transactions of the Royal Society of London, Series A*, 192:101–167, 531, 1899.

[She38] Victor Shestakov. *Some Mathematical Methods for the Construction and Simplification of Two-Terminal Electrical Networks of Class A*. PhD thesis, Lomonosov State University, 1938.

[She08] Jonah Sherman. The randomized decision tree complexity of the recursive majority of three function on 3^n inputs is at least 2.5^n. Unpublished, 2008.

[She13] Irina Shevtsova. On the absolute constants in the Berry–Esseen inequality and its structural and nonuniform improvements. *Informatika i Ee Primeneniya*, 7(1):124–125, 2013.

[SHK72] Barry Simon and Raphael Høegh-Krohn. Hypercontractive semigroups and two dimensional self-coupled Bose fields. *Journal of Functional Analysis*, 9:121–180, 1972.

[Sie84] Thomas Siegenthaler. Correlation-immunity of nonlinear combining functions for cryptographic applications. *IEEE Transactions on Information Theory*, 30(5):776–780, 1984.

[SS10] Oded Schramm and Jeffrey Steif. Quantitative noise sensitivity and exceptional times for percolation. *Annals of Mathematics*, 171(2):619–672, 2010.

[ST78] Vladimir Sudakov and Boris Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978. Originally published in *Zap. Nauchn. Sem. Leningrad. Otdel. Math. Inst. Steklova.*, 41:14–21, 1974.

[Ste72] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602. University of California Press, 1972.

[Ste86a] J. Michael Steele. An Efron–Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14(2):753–758, 1986.

[Ste86b] Charles Stein. *Approximate computation of expectations*. Institute of Mathematical Statistics Lecture Notes. Institute of Mathematical Statistics, Hayward, CA, 1986.

[Sub61] Bella Subbotovskaya. Realizations of linear functions by formulas using $\vee$, &, -. *Doklady Akademii Nauk SSSR*, 136(3):553–555, 1961.

[SW86] Michael Saks and Avi Wigderson. Probabilistic Boolean decision trees and the complexity of evaluating game trees. In *Proceedings of the 27th Annual IEEE Symposium on Foundations of Computer Science*, pages 29–38, 1986.

[Szu98] Jerzy Szulga. *Introduction to random chaos*. Chapman & Hall, 1998.

[Tak83] Akimichi Takemura. Tensor analysis of ANOVA decomposition. *Journal of the American Statistical Association*, 78(384):894–900, 1983.

[Tal89] Michel Talagrand. A conjecture on convolution operators, and a non-Dunford–Pettis operator on L^1. *Israel Journal of Mathematics*, 68(1):82–88, 1989.

[Tal93] Michel Talagrand. Isoperimetry, logarithmic Sobolev inequalities on the discrete cube and Margulis' graph connectivity theorem. *Geometric And Functional Analysis*, 3(3):298–314, 1993.

[Tal94] Michel Talagrand. On Russo's approximate zero-one law. *Annals of Probability*, 22(3):1576–1587, 1994.

[Tal96] Michel Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[Tal06] Michel Talagrand. Regularization from l^1 by convolution. http://www.math.jussieu.fr/~talagran/prizes/convolution.pdf, 2006.

[Tan61] Meyer Tannenbaum. The establishment of a unique representation for a linearly separable function. Technical report, Lockheed Missiles and Space Company, 1961. Threshold Switching Techniques, 20:1–5.

[Tar89] Gábor Tardos. Query complexity, or why is it difficult to separate $\mathsf{NP}^A \cap \mathsf{coNP}^A$ from P^A by random oracles? *Combinatorica*, 9(4):385–392, 1989.

[Ter99] Audrey Terras. *Fourier Analysis on Finite Groups and Applications*. Cambridge University Press, 1999.

[Teu12] Jonas Teuwen. A cornucopia of Hermite polynomials. http://fa.its.tudelft.nl/~teuwen/Writings/Proof-of-competency.pdf, 2012.

[Tho87] Andrew Thomason. Pseudo-random graphs. *Annals of Discrete Mathematics*, 144:307–331, 1987.

[Tit62] Robert Titsworth. *Correlation properties of cyclic sequences*. PhD thesis, California Institute of Technology, 1962.

[Tit63] Robert Titsworth. Optimal ranging codes. Technical Report 32-411, Jet Propulsion Laboratory, 1963.

[Tro58] Hale Trotter. Approximation of semi-groups of operators. *Pacific Journal of Mathematics*, 8:887–919, 1958.

[TSSW00] Luca Trevisan, Gregory Sorkin, Madhu Sudan, and David Williamson. Gadgets, approximation, and linear programming. *SIAM Journal on Computing*, 29(6):2074–2097, 2000.

[TZ00] Jean-Pierre Tillich and Gilles Zémor. Discrete isoperimetric inequalities and the probability of a decoding error. *Combinatorics, Probability and Computing*, 9(5):465–479, 2000.

[UO30] George Uhlenbeck and Leonard Ornstein. On the theory of the Brownian motion. *Physical Review*, 36(5):823–841, 1930.

[Val84] Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Val12] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas with noise. Technical Report TR12-006, Electronic Colloquium on Computational Complexity, 2012.

[Vil47] Naum Vilenkin. On a class of complete orthonormal systems. *Izvestiya Rossiiskoi Akademii Nauk, Seriya Matematicheskaya*, 11(4):363–400, 1947.

[Vio09] Emanuele Viola. Correlation bounds for polynomials over $\{0,1\}$. *SIGACT News*, 40(1):27–44, 2009.

[Vit84] Richard Vitale. An expansion for symmetric statistics and the Efron–Stein inequality. In *Inequalities in Statistics and Probability*, volume 5 of *Lecture Notes—Monograph Series*, pages 112–114. Institute of Mathematical Statistics, 1984.

[vM47] Richard von Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18(3):309–348, 1947.

[Wal23] Joseph Walsh. A closed set of normal orthogonal functions. *American Journal of Mathematics*, 45(1):5–24, 1923.

[Wei79] Fred Weissler. Two-point inequalities, the Hermite semigroup, and the Gauss–Weierstrass semigroup. *Journal of Functional Analysis*, 32(1):102–121, 1979.

[Wei80] Fred Weissler. Logarithmic Sobolev inequalities and hypercontractive estimates on the circle. *Journal of Functional Analysis*, 37(2):218–234, 1980.

[Wol07] Paweł Wolff. Hypercontractivity of simple random variables. *Studia Mathematica*, 180(3):219–236, 2007.

[XM88] Guozhen Xiao and James Massey. A spectral characterization of correlation-immune combining functions. *IEEE Transactions on Information Theory*, 34(3):569–571, 1988.

[Yan04] Ke Yang. On the (im)possibility of non-interactive correlation distillation. In *Proceedings of the 6th Annual Latin American Informatics Symposium*, pages 222–231, 2004.

[Yao77] Andrew Yao. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the 9th Annual ACM Symposium on Theory of Computing*, pages 222–227, 1977.

[Yao85] Andrew Yao. Separating the polynomial time hierarchy by oracles. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pages 1–10, 1985.

[Zhe27] Ivan Zhegalkin. On a technique of calculating propositions in symbolic logic. *Matematicheskii Sbornik*, 43:9–28, 1927.

[Zue89] Yuri Zuev. Asymptotics of the logarithm of the number of threshold functions of the algebra of logic. *Doklady Akademii Nauk SSSR*, 39(3):512–513, 1989.

[Zwi99] Uri Zwick. Outward rotations: A tool for rounding solutions of semidefinite programming relaxations, with applications to MAX CUT and other problems. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 679–687, 1999.

[Zwi02] Uri Zwick. Computer assisted proof of optimal approximability results. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 496–505, 2002.

Index

Revision history

Revision	Date	Description
v1.00	Apr. 8, 2014	Version published by Cambridge University Press
v1.01	Oct. 4, 2014	About 40 typos fixed