The Dissertation Committee for Alexander Alexandrovich Sherstov certifies that this is the approved version of the following dissertation:

Lower Bounds in Communication Complexity and Learning Theory via Analytic Methods

Committee:

Adam R. Klivans, Supervisor

Anna Gál

C. Gregory Plaxton

Alexander A. Razborov

David Zuckerman

Lower Bounds in Communication Complexity and Learning Theory via Analytic Methods

by

Alexander Alexandrovich Sherstov, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2009

In loving memory of my grandmother, Lyubov Panteleyevna Shilova (1923–2004).

Acknowledgments

I am deeply indebted to my adviser Adam Klivans for his fundamental role in my doctoral work. Adam provided me with every bit of guidance, assistance, and expertise that I needed during my first few semesters; then, when I felt ready to venture into research on my own and branch out into new research areas, Adam gave me the freedom to do whatever I wanted, at the same time continuing to contribute valuable feedback, advice, and encouragement. In addition to our academic collaboration, I greatly value the close personal rapport that Adam and I have forged over the years. I quite simply cannot imagine a better adviser.

I would like to thank the theory faculty at the University of Texas at Austin, especially Anna Gál, Adam Klivans, and David Zuckerman, for the substantial influence that their courses have had on my research. I gratefully acknowledge the members of my Ph.D. committee for their time and valuable feedback on a preliminary version of this thesis. I would particularly like to acknowledge Alexander Razborov, who has long been an inspiring figure for me. Among many other things, I am thankful to Sasha for inviting me to the Institute for Advanced Study in the fall of 2007 and our fruitful collaboration there.

I am very thankful to Matthew Fike, my English professor at the American University in Bulgaria in 1999 and a good friend ever since, for all his help, advice, and encouragement. The computer science faculty at Hope College, my alma mater, also played a crucial role in my academic career. I am particularly thankful to Herb Dershem and Ryan McFall for being the outstanding educators that they are and for all their encouragement and support during my time at Hope and later on. I feel extremely privileged to have been their student. I am very thankful to the professors of French at Hope College and the University of Texas at Austin, particularly Brigitte Hamon-Porter, Anne Larsen, Carl Blyth, Robert Dawson (who sadly passed in June 2007), Thomas Vessely, as well as teaching assistant Julie Ouvrard and my tutor and close friend Emilie Destruel, for helping me realize my dream of mastering French and for teaching me so much about France and French culture. Je ne vous en remercierai jamais assez.

I owe a great debt of gratitude to my teachers at High School No. 1 in my hometown Karaganda, Kazakhstan. These men and women provided their students with a world-class high school education in mathematics, physics, biology, chemistry, and other subjects without being paid a salary for months on end, amid the disastrous economic conditions and political turmoil that resulted from the dissolution of the Soviet Union. I am particularly thankful to my mathematics teacher and mentor Vera Petrovna Onishchenko, whose pedagogical talent is greatly responsible for my love of mathematics. As time goes on, I realize more and more clearly the huge impact that she has had on my academic career.

I would like to thank my friends in the computer science department at Austin for all the great times that we have shared. I am particularly thankful to Mazda Ahmadi, a racquetball partner and a good friend, for his infinite patience with my command of the racquet and reaction speed.

I am deeply thankful to my family for their love, support, and sacrifices. Without them, this thesis would never have been written. I dedicate this thesis to the memory of my grandmother Lyubov Panteleyevna Shilova, whose role in my life was, and remains, immense. This last word of acknowledgment I have saved for my dear wife Lily Mihalkova, who has been with me all these years and has made them the best years of my life.

Alexander A. Sherstov

Austin, Texas
August 2009

Lower Bounds in Communication Complexity and Learning Theory via Analytic Methods

Publication No. _____

Alexander Alexandrovich Sherstov, Ph.D.
The University of Texas at Austin, 2009

Supervisor: Adam R. Klivans

A central goal of theoretical computer science is to characterize the limits of efficient computation in a variety of models. We pursue this research objective in the contexts of *communication complexity* and *computational learning theory*. In the former case, one seeks to understand which distributed computations require a significant amount of communication among the parties involved. In the latter case, one aims to rigorously explain why computers cannot master some prediction tasks or learn from past experience. While communication and learning may seem to have little in common, they turn out to be closely related, and much insight into both can be gained by studying them *jointly*. Such is the approach pursued in this thesis. We answer several fundamental questions in communication complexity and learning theory and in so doing discover new relations between the two topics. A consistent theme in our work is the use of analytic methods to solve the problems at hand, such as approximation theory, Fourier analysis, matrix analysis, and duality.

- We contribute a novel technique, the *pattern matrix method*, for proving lower bounds on communication. Using our method, we solve an open problem due to Krause and Pudlák (1997) on the comparative power of two well-studied circuit classes: majority circuits and constant-depth AND/OR/NOT circuits. Next, we prove that the pattern matrix method applies not only to classical communication but also to the more powerful quantum model. In particular, we contribute lower bounds for a new class of quantum communication

problems, broadly subsuming the celebrated work by Razborov (2002) who used different techniques. In addition, our method has enabled considerable progress by a number of researchers in the area of multiparty communication.

- Second, we study *unbounded-error* communication, a natural model with applications to matrix analysis, circuit complexity, and learning. We obtain essentially optimal lower bounds for all symmetric functions, giving the first strong results for unbounded-error communication in years. Next, we resolve a longstanding open problem due to Babai, Frankl, and Simon (1986) on the comparative power of unbounded-error communication and alternation, showing that $\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}$. The latter result also yields an unconditional, exponential lower bound for learning DNF formulas by a large class of algorithms, which explains why this central problem in computational learning theory remains open after more than 20 years of research.

- We establish the computational intractability of learning *intersections of halfspaces*, a major unresolved challenge in computational learning theory. Specifically, we obtain the first exponential, near-optimal lower bounds for the learning complexity of this problem in Kearns' statistical query model, Valiant's PAC model (under standard cryptographic assumptions), and various analytic models. We also prove that the intersection of even two halfspaces on $\{0,1\}^n$ cannot be sign-represented by a polynomial of degree less than $\Theta(\sqrt{n})$, which is an exponential improvement on previous lower bounds and solves an open problem due to Klivans (2002).

- We fully determine the relations and gaps among three key complexity measures of a communication problem: product discrepancy, sign-rank, and discrepancy. As an application, we solve an open problem due to Kushilevitz and Nisan (1997) on distributional complexity under product versus nonproduct distributions, as well as separate the communication classes $\mathsf{PP}^{\mathsf{cc}}$ and $\mathsf{UPP}^{\mathsf{cc}}$ due to Babai, Frankl, and Simon (1986). We give interpretations of our results in purely learning-theoretic terms.

Table of Contents

Chapter 1

Introduction

A broad goal of theoretical computer science is to characterize the limits of efficient computation in a variety of models. We pursue this research objective in the contexts of *communication* and *learning*. In the former case, one seeks to understand which distributed computations require a significant amount of communication among the parties involved. In the latter case, one aims to rigorously explain why computers cannot master some prediction tasks or learn from past experience. The complexity of communication and the complexity of learning are vibrant areas of theoretical computer science, studied for their intrinsic appeal as well as applications to other areas. While communication and learning may seem to have little in common, they turn out to be closely related, and much insight into both can be gained by studying them *jointly*. Such is the approach pursued in this thesis. Our work is based on the fact that communication and learning naturally lend themselves to the methods of mathematical analysis, such as approximation theory, matrix analysis, linear programming duality, and Fourier analysis. We use these analytic methods to answer several fundamental questions in communication complexity and learning theory, and in so doing discover new relations between the two topics. Before we go into further details, we need to describe the subject of our study somewhat more formally.

1.1 Communication complexity

Communication complexity theory, initiated in a seminal 1979 paper by Yao [223], studies of the amount of information exchange necessary in order to compute a given Boolean function when its arguments are distributed among several parties. More precisely, consider a Boolean function $f\colon X \times Y \to \{-1,+1\}$ for some finite sets X and Y. The canonical model features two parties, traditionally called Alice and Bob. Alice receives an argument $x \in X$, her counterpart Bob receives an argument $y \in Y$, and their objective is to determine the value $f(x,y)$. To this end, Alice and Bob can exchange binary messages, i.e., strings of 0 and 1, via a shared communication channel according to an agreed-upon protocol. For a given function f, the main research question is to determine whether a protocol exists that will allow Alice and Bob to evaluate $f(x,y)$ correctly for any x and y while exchanging only a small number of bits.

To illustrate, consider the *disjointness problem*, where Alice and Bob each receive a subset of $\{1, 2, \ldots, n\}$ and need to determine if these subsets intersect. A trivial solution would be for Alice to send her entire set to Bob, using the canonical n-bit encoding. This trivial protocol turns out to be essentially optimal [137]: one can prove a lower bound of $\Omega(n)$ bits on the communication required. This lower bound remains valid [102, 176] even if Alice and Bob simply want to predict the correct answer with probability 51%. On the other hand, consider the *equality problem*, in which Alice and Bob receive strings $x \in \{0,1\}^n$ and $y \in \{0,1\}^n$, respectively, and need to determine whether $x = y$. While an exact solution still requires $\Omega(n)$ bits of communication, it may surprise the reader that the parties can solve this problem with correctness probability 99% by exchanging only a constant number of bits [137].

Over the past thirty years, communication complexity has evolved into a challenging and active research area. First of all, it provides a natural model in which to study the limits of efficient computation. Furthermore, communication complexity turns out to be essential for understanding a host of other computational models. To illustrate how communication complexity sheds light on seemingly unrelated questions, consider the problem of laying out a computer chip. A typical such chip, shown schematically in Figure 1.1, features n Boolean inputs $x_1, \ldots, x_n$ that feed into a circuit of wires and gates. The chip computes a certain function $f : \{0,1\}^n \to \{0,1\}$ of the inputs, with the output appearing at the output port. The chip operates in cycles, each corresponding to signals propagating through one Boolean gate. The designer's challenge is to lay out the chip so as to minimize the number of compute cycles as well as the area of the chip. In a classic paper, Thompson [212] recasts this task as a communication problem induced by f and derives a lower bound on the chip area and compute time in terms of the communication complexity of f.

As another well-studied application, consider the design of *streaming algorithms*. In this setting, a stream of data is arriving so abundant that it is impractical to store it all in memory. Such situations arise frequently in network analysis, database management, and other large-scale computational tasks. In such cases one scans the data stream one item at a time, retaining only a small amount of information about the previously viewed items. In an influential paper, Alon, Matias, and Szegedy [14] modeled a streaming algorithm as a communication protocol and showed that several natural streaming problems require large amounts of memory.

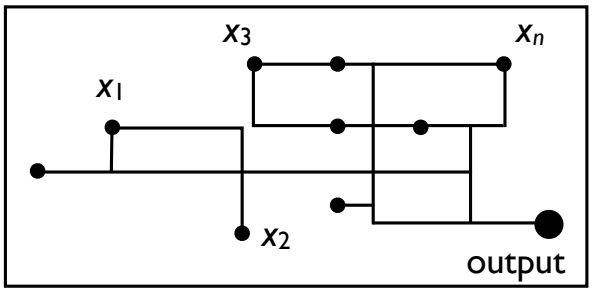

Figure 1.1: A chip layout problem.

Chip layout and streaming algorithms are just the tip of the iceberg as far as the uses of communication complexity are concerned. Communication complexity plays a vital role in theoretical computer science, with applications to circuit complexity, Turing machines, branching programs, data structures, derandomization, private information retrieval, communication networks, matrix analysis, and learning theory. To accommodate this array of applications, various models of communication are in use. Many of these models are treated in this thesis, including randomized communication, quantum communication, multiparty communication, small-bias communication, and unbounded-error communication.

1.2 Computational learning theory

As computing technology is put to use in more and more areas of human activity, there is an increasing need for algorithms that automatically improve their performance over time, or *learn* from past experience. Computational tasks that would benefit from this kind of learning include estimating the credit-worthiness of a loan applicant based on historical data, classifying satellite images based on previously labeled examples, and predicting stock market performance based on past trends. *Computational learning theory* is an area of theoretical computer science that seeks to design efficient and provably correct algorithms for natural learning tasks and to rigorously establish why some other learning tasks are inherently intractable.

A central computational abstraction of learning is the *probably approximately correct* model due to Valiant [213], commonly abbreviated PAC. This

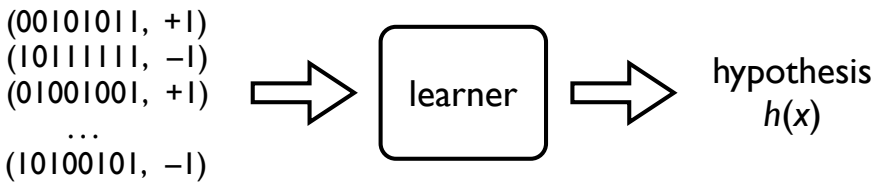

Figure 1.2: A computational view of learning.

framework, illustrated in Figure 1.2, models a learning task as a family $\mathscr{C}$ of Boolean functions $f: X \to \{-1, +1\}$ on a given set X, with the typical case being $X = \{0,1\}^n$. The functions in $\mathscr{C}$ are called *concepts* and the family $\mathscr{C}$ is called the *concept class*. Fix a probability distribution μ on X and choose function $f \in \mathscr{C}$, both unknown to the learner. The learner receives the evaluations of f on a small set of points $x^{(1)}, \ldots, x^{(m)} \in X$, each drawn independently according to μ. Based on this training data alone and the fact that $f \in \mathscr{C}$, the learner's challenge is to produce a hypothesis $h: X \to \{-1, +1\}$ that agrees with the unknown function almost perfectly, in the sense that $\mathbf{P}_\mu[f(x) = h(x)] > 0.99$. Given a concept class $\mathscr{C}$, the main research question is to design a computationally efficient PAC learning algorithm for $\mathscr{C}$ or to give a rigorous argument for the problem's computational intractability. Apart from the PAC model, several other learning models are considered in this thesis and earlier literature, distinguished by the kind of information that learner receives about the unknown function.

To illustrate these definitions, consider the concept class of polynomial-size formulas in disjunctive normal form (DNF), commonly called *DNF formulas* for short. These are functions $f: \{0,1\}^n \to \{-1, +1\}$ expressible in the form $f(x_1, \ldots, x_n) = \bigvee_{i=1}^m T_i(x_1, \ldots, x_n)$, where $m < n^c$ for some positive constant c and each T_i is a conjunction of the literals $x_1, \ldots, x_n$ or their negations. Learning DNF formulas efficiently amounts to giving a polynomial-time algorithm that takes as input a random sample of strings from $\{0,1\}^n$, each labeled according to the unknown function f, and produces an accurate approximation $h: \{0,1\}^n \to \{-1, +1\}$ to f. Note that the hypothesis h itself can be any polynomial-time computable function and not necessarily a DNF formula. The challenge of learning DNF formulas in

polynomial time, posed in Valiant's seminal 1984 paper on computational learning theory [213], remains unresolved to this day.

The study of computational learning is motivated on the one hand by practical applications such as computer vision, data mining, electronic commerce, and networking, and on the other hand by the connections between learning and other areas of theoretical computer science such as circuit complexity and cryptography. In particular, the work in this thesis exploits the close relationship between learning and communication complexity. As described above, a learning problem is given by a family $\mathscr{C}$ of functions $X \to \{-1, +1\}$, which has the following convenient representation in matrix form:

$$A = \begin{array}{c} x \in X \\ \left[\begin{array}{c} \\ \cdots\cdots\cdots\cdots\cdots f(x) \cdots\cdots\cdots\cdots\cdots \\ \\ \end{array}\right] \end{array} f \in \mathscr{C}$$

Similarly, a communication problem is given by a function $f : X \times Y \to \{-1, +1\}$ for some finite sets X and Y and has the representation

$$B = \begin{array}{c} x \in X \\ \left[\begin{array}{c} \\ \cdots\cdots\cdots\cdots\cdots f(x,y) \cdots\cdots\cdots\cdots\cdots \\ \\ \end{array}\right] \end{array} y \in Y$$

The matrices A and B above, one arising in learning theory and the other in communication complexity, are distinguished exclusively by the row and column indices.

Analytically speaking, A and B are the same object: a sign matrix. In this light, our approach is to focus on the main mathematical object, the sign matrix, without undue concern for its origin (learning or communication). This approach allows us to discover novel techniques for studying sign matrices, use them to solve challenging open problems in learning and communication, and reveal new relations between the two areas.

1.3 A historical overview

Problems in computational complexity frequently admit formulations in terms of multivariate real polynomials, rational functions, or matrices. This is particularly true of the work in this thesis, and a recurring theme in our research is the use of analytic methods such as approximation theory, Fourier analysis, matrix analysis, and linear programming duality. Analytic methods have a long history of use in computational complexity. We now pause to survey this literature, focusing on analytic work in communication complexity and learning theory.

Approximation theory

Modern approximation theory is an active area of mathematical analysis that originated in the mid-nineteenth century. The monograph by Cheney [60] offers a thorough treatment of the fundamentals of approximation theory, with a shorter and gentler introduction available due to Rivlin [185]. In computational complexity, one typically considers a Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$ and studies the least degree of a real polynomial that approximates f pointwise within a given error ε. More formally, for $0 < \varepsilon < 1$ one studies the *ε-approximate degree* of f, denoted $\deg_\varepsilon(f)$ and defined as the least degree of a polynomial p such that

$$\max_{x\in\{0,1\}^n} |f(x) - p(x)| \leqslant \varepsilon.$$

There is an extensive literature on the ε-approximate degree of Boolean functions [160, 165, 100, 52, 4, 15, 221], both for the canonical setting $\varepsilon = 1/3$ and various other settings. An important limiting case of the approximate degree is the

quantity

$$\deg_{\pm}(f) = \lim_{\varepsilon \nearrow 1} \deg_{\varepsilon}(f),$$

called the *threshold degree* of f. A moment's reflection reveals that the threshold degree is the least degree of a polynomial p that represents f in sign: $f(x) \equiv \operatorname{sgn} p(x)$. The threshold degree, too, has been extensively studied by complexity theorists [153, 21, 132, 133, 122, 121, 163].

Uniform approximation and sign-representation of Boolean functions have found various applications in the literature, both on the algorithmic front and on the complexity-theoretic front. To start with, the fastest known algorithms for PAC learning DNF formulas [122] and intersections of halfspaces [121] crucially exploit the fact that these concept classes admit efficient sign-representations by polynomials. Representation of Boolean functions by real polynomials have been used to prove lower bounds on various types of Boolean and threshold circuits and to relate computational complexity classes [33, 166, 210, 21, 132, 218, 133]. Remarkable relationships have been discovered [158, 160, 28] between the above analytic representations of Boolean functions and the more traditional, combinatorial models such as decision trees. In more recent literature, approximation theory has also been applied to prove communication lower bounds [177]. For detailed background on complexity-theoretic applications of polynomial approximation, we refer the reader to the excellent surveys in [55, 31, 190].

In addition to real polynomials, there is a considerable literature on representations of Boolean functions by polynomials with integer coefficients, or *perceptrons* as they are sometimes called [153, 155, 32, 92, 218, 124, 169, 170]. Finally, it is natural to study approximation by *rational functions*. This subject in approximation theory dates back to the classical work by Zolotarev [228] and Newman [156] and has seen much research, as surveyed in [168]. In particular, rational functions have found various applications in computational complexity [33, 166, 210, 121, 1] and play an important role in the concluding chapters of this thesis.

Fourier analysis

A counterpart to uniform approximation on the Boolean hypercube is Fourier analysis, which is motivated by least-squares approximation. Consider a Boolean function $f: \{-1, +1\}^n \to \{-1, +1\}$. By elementary linear algebra, f admits a unique representation of the form

$$f(x) = \sum_{S \subseteq \{1,2,\ldots,n\}} \hat{f}(S) \prod_{i \in S} x_i,$$

for some reals $\hat{f}(S)$ called the *Fourier coefficients* of f. A straightforward yet powerful observation, known as Parseval's identity, is that

$$\sum_{S \subseteq \{1,2,\ldots,n\}} \hat{f}(S)^2 = 1.$$

In this light, it is meaningful to ask how the "Fourier mass" is distributed among the 2^n Fourier coefficients. Of particular interest is the case when the high-order Fourier coefficients are negligible in magnitude, i.e., when $\sum_{|S|>k} \hat{f}(S)^2 < \varepsilon$ for some small ε and $k \ll n$. In this case, the Fourier mass of f is concentrated on the low-order Fourier coefficients, and one can show that the truncated, low-degree expression

$$\sum_{S \subseteq \{1,2,\ldots,n\},\ |S| \leqslant k} \hat{f}(S) \prod_{i \in S} x_i$$

is an accurate least-squares approximation to f.

Concentration results for the Fourier spectrum and other Fourier-based techniques have been used to design learning algorithms with respect to the uniform distribution for a variety of concept classes [141, 136, 150, 98, 121, 154, 49], including decision trees, DNF formulas, constant-depth circuits, intersections of halfspaces, and juntas. In computational complexity, the Fourier spectrum has been used to obtain communication lower bounds [171, 114] and to study various types of threshold

circuits [44, 207, 45, 92]. Other prominent uses of Fourier analysis in complexity theory include probabilistically checkable proofs and social choice theory, as surveyed in [66, 162].

Matrix analysis

Problems in communication complexity and learning theory have representations in the form of sign matrices, as described earlier in Section 1.2. Such representations allow for a fruitful use of matrix analysis. The singular values, for example, provide a wealth of information about a given matrix, including its rank, spectral norm, trace norm, and various other properties. Typically, it is necessary to study the spectral properties not only for a given sign matrix A but also for all real matrices in its ε-neighborhood: $\{B : \|A - B\|_\infty < \varepsilon\}$.

Among the first uses of matrix analysis in communication complexity is the influential paper of Mehlhorn and Schmidt [152], who showed that the rank of a communication matrix implies a lower bound on its deterministic communication complexity. The longstanding *log-rank conjecture*, due to Lovász and Saks [147], is also matrix-analytic in nature: it states that the logarithm of the rank of a sign matrix characterizes its deterministic communication complexity up to a polynomial. It is well-known [130, 226, 134, 54, 177] that the communication complexity of a sign matrix A in several classical and quantum models can be bounded from below by the minimum rank or minimum trace norm of the real matrices in the ε-neighborhood of A. The breakthrough lower bound, due to Forster [70], for the *unbounded-error* model of communication crucially uses matrix analysis and compactness arguments. Matrix-analytic methods have also been used to study threshold circuits [130, 71] and matrix rigidity [146, 103, 73]. In computational learning theory, matrix analysis has been used in the context of learning via *Euclidean embeddings* and *statistical queries* [70, 72, 222, 206, 73].

Linear programming duality

Linear programming duality is an analytic tool with applications in areas as diverse as game theory, economics, algorithm design, and complexity theory. The duality of linear programming admits several equivalent formulations. One such, known as Farkas' Lemma [194], states that the system of linear inequalities $Ax \leqslant b$, where A

is a real matrix and b a vector, has a solution $x \geqslant 0$ if and only if there does not exist a vector $y \geqslant 0$ such that $y^{\mathsf{T}}A \geqslant 0$ and $y^{\mathsf{T}}b < b$. The crux is that every infeasible system $Ax \leqslant b$, $x \geqslant 0$ has a corresponding vector y that witnesses the system's infeasibility. In computational complexity, this witness is an object of great interest in its own right because it can be used to construct a new witness, exhibiting the infeasibility of another system.

Let us consider an illustrative use of duality, due to O'Donnell and Servedio [163]. Recall that the threshold degree of a Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$, denoted $\deg_{\pm}(f)$, is the least degree of a real polynomial p with $f(x) \equiv \operatorname{sgn} p(x)$. Fix two Boolean functions $f, g\colon \{0,1\}^n \to \{-1,+1\}$ and let $f \oplus g$ denote their product: $(f \oplus g)(x,y) = f(x)g(x)$. Then $f \oplus g$ is a Boolean function itself, and it is meaningful to speak of its threshold degree. Observe that

$$\deg_{\pm}(f \oplus g) \leqslant \deg_{\pm}(f) + \deg_{\pm}(g),$$

since given any polynomials p, q with $f(x) \equiv \operatorname{sgn} p(x)$ and $g(x) \equiv q(x)$, we have $(f \oplus g)(x,y) \equiv \operatorname{sgn}\{p(x)q(y)\}$. Is this simple upper bound tight? It turns out that it is, and the proof is an elegant application of linear programming duality [163]. In more detail, the fact that f and g have threshold degree at least $\deg_{\pm}(f)$ and $\deg_{\pm}(g)$, respectively, can be certified by a certain witness in each case. The existence of these two witnesses is assured by Farkas' Lemma. Combining them yields the sought witness that $\deg_{\pm}(f \oplus g) \geqslant \deg_{\pm}(f) + \deg_{\pm}(g)$.

Apart from the above example, a well-known application of duality in the complexity literature is due to Linial and Nisan [143], who studied the problem of approximating the probability of a union of events $A_1 \cup \cdots \cup A_n$ by the probabilities of the intersections $\bigcap_{i \in S} A_i$ for small sets S. Tarui and Tsukiji [211] used this duality-based result to give an improved algorithm for learning DNF formulas. In communication complexity, Yao [224] used the famous minimax theorem for zero-sum games to relate randomized and deterministic communication. Kushilevitz and Nisan [137] used duality to study a different complexity measure of a communication problem. More recently, Linial and Shraibman [145] used linear programming duality to prove the equivalence of the *discrepancy* of a communication problem and a learning-theoretic notion, *margin complexity.* Linear programming duality is an essential tool in many chapters of this thesis.

1.4 Our contributions

We will describe the contributions of this thesis on a topic-by-topic basis, focusing first on communication complexity, then on computational learning theory, and finally on interdisciplinary results that span both areas.

The pattern matrix method

Recall that in communication complexity, one seeks to prove a lower bound on the amount of communication needed to compute a given Boolean function in a given model. This thesis contributes a novel technique for communication lower bounds, the *pattern matrix method.* Our technique takes any given Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$ and creates from it a natural two-party communication problem $F(x,y)$. The crux is that the communication complexity of $F(x,y)$ can be conveniently determined from standard analytic properties of f. To illustrate, if f cannot be approximated pointwise within $1/3$ by a real polynomial of low degree, then $F(x,y)$ will have high communication complexity in the randomized model. If, in addition, f cannot even be represented in sign by a low-degree real polynomial, then the resulting communication problem $F(x,y)$ will have high cost in almost all models, including computation with an exponentially small advantage over random guessing. Recall that real polynomials are among the most extensively studied objects in theoretical computer science, with the literature dating as far back as 1961. In other words, the pattern matrix method takes several decades' worth of work on representations of Boolean functions by real polynomials and puts them at the disposal of communication complexity. The key to our method is to discover a way to exploit the *dual* formulation of uniform approximation, an approach that departs entirely from previous work [54, 114, 177]. Other essential ingredients of the pattern matrix method are Fourier analysis and spectral techniques.

We demonstrate the power of the pattern matrix method by solving several problems in the area. First, we solve an open problem posed in 1997 by Krause and Pudlák [132] on the comparative power of two well-studied circuit classes: constant-depth AND/OR/NOT circuits and majority circuits. Specifically, we prove that constant-depth AND/OR/NOT circuits cannot be simulated by depth-2 majority circuits of subexponential size. This result is best possible since an efficient simulation always exists by depth-3 majority circuits [11].

Second, we prove that the pattern matrix method applies not only to classical communication but also to the more challenging *quantum* model. In particular, we contribute lower bounds for a new class of quantum communication problems, broadly subsuming the celebrated work by Razborov [177] on the disjointness function and other symmetric problems. Our techniques are completely different from Razborov's. As another application, we exhibit a large new class of communication problems for which quantum protocols are essentially no better than their classical counterparts, subsuming a previous such class due to Razborov [177]. Understanding the comparative power of quantum and classical communication is a fundamental goal in the area.

We are pleased to report that the pattern matrix method has also enabled important progress by a number of researchers in the area of *multiparty* communication. Lee and Shraibman [140] and Chattopadhyay and Ada [59] observed that our method readily generalizes to three and more players, thereby obtaining much improved lower bounds on the communication complexity of the disjointness function. David and Pitassi [64] combined these techniques with an ingenious use of the probabilistic method, thereby separating the multiparty classes $\mathsf{NP}^{\mathsf{cc}}$ and $\mathsf{BPP}^{\mathsf{cc}}$. Their probabilistic construction was made explicit in a follow-up paper by David, Pitassi, and Viola [65]. Continuing this line of work is a recent paper by Beame and Huynh-Ngoc [29], who give much improved multiparty communication lower bounds for constant-depth circuits. In joint work with D. Gavinsky, we contribute to this growing body of work a proof that $\mathsf{NP}^{\mathsf{cc}} \neq \mathsf{coNP}^{\mathsf{cc}}$ in multiparty communication, and other separations [81].

Unbounded-error communication and sign-rank

The *unbounded-error* model is the most powerful of the primary models of communication, with applications to learning theory, threshold circuit complexity, and matrix analysis. The unbounded-error communication complexity of a Boolean matrix $A \in \{-1,+1\}^{m \times n}$ is precisely determined the *sign-rank* of A, or equivalently the least rank of a real matrix R with $A_{ij} = \operatorname{sgn} R_{ij}$ for all i, j. We contribute essentially optimal lower bounds for every symmetric Boolean function in this model. These are the first strong results for unbounded-error communication in years, since Forster's breakthrough lower bounds for Hadamard matrices and their generaliza-

tions [70]. Our proof uses approximation theory, Fourier analysis, and random walks, in addition to the pattern matrix method.

In a follow-up result on the topic, proved jointly by A. A. Razborov and the author, we solve a longstanding open problem posed by Babai, Frankl, and Simon [23]. This problem asks whether unbounded-error communication protocols are more powerful than the polynomial hierarchy $\mathsf{PH}^{\mathsf{cc}}$, another model which corresponds to communication with a bounded number of $\exists$ and $\forall$ quantifiers. We give the strongest negative answer to this question, showing that even the function $f(x, y) = \bigwedge_{i=1}^{n} \bigvee_{j=1}^{n} (x_{ij} \wedge y_{ij})$, which corresponds to two quantifiers, does not have an efficient unbounded-error protocol. At the heart of our proof is a new method for analyzing multivariate polynomials p on $\mathbb{R}^n$, which works by projecting p in several ways to a univariate polynomial, analyzing these simpler objects using approximation theory, and recombining the results using Fourier-theoretic tools.

Our joint work with A. A. Razborov additionally shows that polynomial-size DNF formulas have exponentially high sign-rank. Put differently, we give an unconditional, exponential lower bound for learning DNF formulas by a large class of algorithms, which helps explain why this central problem in learning theory remains unsolved after more than 20 years of research [214, 106, 17, 18, 217, 19, 7, 89, 136, 6, 8, 39, 150, 46, 90, 138, 5, 189, 48, 47, 211, 122].

Learning intersections of halfspaces

A major unresolved challenge in computational learning theory is learning intersections of halfspaces, i.e., intersections of Boolean functions of the form $f(x) = \operatorname{sgn}(\sum a_i x_i - \theta)$ for some fixed reals $a_1, \ldots, a_n, \theta$. While efficient algorithms have long been known for learning a single halfspace, no polynomial-time algorithm has been found for learning the intersection of even two halfspaces. In joint work with A. R. Klivans, we account for this lack of progress by establishing the first exponential, near-optimal lower bounds for learning the intersection of n^{ε} halfspaces in Kearns' *statistical-query* model [104]. We obtain an analogous result for the PAC model, showing that an efficient learning algorithm for the intersection of n^{ε} halfspaces would violate standard cryptographic assumptions. Our techniques center around perceptrons and Fourier analysis.

These two results leave open the possibility of efficiently learning the intersection of k halfspaces for k small, such as a constant. The author addresses this challenge in the concluding part of the thesis. Specifically, we construct two halfspaces on $\{0,1\}^n$ whose intersection cannot be sign-represented by a polynomial of degree less than $\Theta(\sqrt{n})$. This lower bound is an exponential improvement on previous work. It solves an open problem due to Klivans [120] and rules out the use of perceptron-based algorithms for learning the intersection of even two halfspaces. Our techniques feature novel applications of linear programming duality, lower bounds for rational functions, and other analytic work. In particular, we contribute new techniques for studying compositions $F(f_1, \ldots, f_k)$ in terms of the analytic properties of the constituent functions $F, f_1, \ldots, f_k$.

In summary, our study of intersections of halfspaces paints a rather complete picture of the problem's intractability, including representation-independent cryptographic hardness results as well as various model-specific, exponential lower bounds.

Relations and separations in communication and learning

By treating communication problems abstractly as sign matrices, we are able to relate the computational power of several other models of communication. One open problem that we solve, due to Kushilevitz and Nisan [137], is as follows. When proving a lower bound on the communication requirements of a Boolean function $f(x, y)$, one must typically find a probability distribution μ with respect to which f is hard to compute. In practice, it is far easier to analyze *product distributions,* whereby the two arguments to the function $f(x, y)$ are distributed independently. Kushilevitz and Nisan asked in 1997 whether the exclusive use of product distributions can severely affect the quality of the solution. We answer this question, exhibiting a gap of $\Theta(1)$ versus $\Theta(n)$ for the product and nonproduct communication complexity of a certain function $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$. Only a quadratic separation between the two quantities was known prior to our work [23].

Next, we study three complexity measures of a communication problem: *product discrepancy*, *sign-rank*, and *discrepancy.* Remarkably, all three play a key role in learning theory [145, 198] and are known in the learning literature as *statistical query dimension*, *dimension complexity*, and *margin complexity*. All that

was known prior to our work is that the margin complexity gives an upper bound on the dimension complexity [35]. We completely determine all the other relations and gaps among the three quantities. As an application to communication complexity, we separate the communication classes $\mathsf{PP}^{\mathsf{cc}}$ and $\mathsf{UPP}^{\mathsf{cc}}$, whose equality or inequality was open since their introduction in 1986 by Babai, Frankl, and Simon [23]. These two classes were separated independently and at the same time by Buhrman, Vereshchagin, and de Wolf [53], using quite different techniques. As an application of our work to learning theory, we prove a near-tight upper bound of $2(n+1)^2$ on the statistical query dimension of halfspaces, considerably improving on the earlier estimate of $n^{O(1)}$ from the influential 1994 paper of Blum et al. [37].

1.5 Organization of this thesis

This thesis is organized in two parts. Part I focuses on communication complexity and consists of Chapters 3–11. Part II studies computational learning theory and consists of Chapters 12–17. Preceding both parts of the thesis is a chapter on general analytic and combinatorial background. In what follows, we give a more detailed account of the organization on a chapter-by-chapter basis. We will later complement this overview with a chart of chapter dependencies.

Chapter 2 sets the stage for the technical development to follow with notation, conventions, and required analytic background. In particular, we review key analytic properties of Boolean functions and matrices, which play a central role throughout this thesis. Also covered here are combinatorial complexity measures of Boolean functions, which turn out to be closely related to their analytic counterparts. To best organize the various definitions and notations, we conclude Chapter 2 with a table of technical symbols and their meanings. A comprehensive such table, including symbols introduced in later chapters, is provided in Appendix A for the reader's convenience.

Chapter 3 marks the beginning of Part I of this thesis on communication complexity. We introduce the canonical framework of two-party communication and the deterministic, randomized, and nondeterministic models. More advanced formalisms, such as the quantum, unbounded-error, and multiparty models, will be introduced as needed in later chapters. We close Chapter 3 by discussing an

elegant technique for communication lower bounds, the *discrepancy method*, and its generalizations.

Chapter 4 addresses the challenge of proving communication lower bounds in the randomized model, both for bounded-error protocols and for small-bias protocols. It is here that we develop our technique, the *pattern matrix method*, that converts standard analytic properties of Boolean functions into lower bounds for the associated communication problems. As an application, we establish the separations $\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$ and $\Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$ in communication complexity and solve an open problem in circuit complexity due to Krause and Pudlák [132]. Various other applications of the pattern matrix method are presented in later chapters as we further develop our technique.

Chapter 5 focuses on the *quantum* model of communication, which is a counterpart to the classical randomized model. We prove that the pattern matrix method applies unchanged to this model, yielding a new source of communication lower bounds. As an illustration of the quantum pattern matrix method, we give a new and simple proof of Razborov's breakthrough lower bounds for disjointness and the other symmetric functions [177]. Finally, we contrast the pattern matrix method with a different duality-based technique, the *block composition method* of Shi and Zhu [205].

Chapter 6 studies the comparative power of classical and quantum communication. It is a longstanding goal in computational complexity to prove that quantum bounded-error protocols cannot be superpolynomially more efficient than their classical counterparts. Here, we prove this conjecture for a new class of communication problems, subsuming previous results. In particular, we prove a polynomial relationship between the quantum and classical complexity of computing $f(x \wedge y)$ and $f(x \vee y)$ on input x, y, where f is any given Boolean function. We prove analogous results for other function compositions. Finally, the explore the implications of our techniques for the *log-rank conjecture.*

Chapter 7 examines the *unbounded-error* model. Our main result here is a near-tight lower bound on the unbounded-error communication complexity of every symmetric function, i.e., every function of the form $f(x, y) = D(\sum x_i y_i)$ for some predicate $D\colon \{0, 1, \ldots, n\} \to \{-1, +1\}$. The pattern matrix method of the previous chapters continues to play an important role in this chapter but needs to

be complemented with other results on random walks, matrix analysis, and approximation theory.

Chapter 8 continues our study of unbounded-error communication complexity, this time with a focus on the circuit class AC^0. The main result of this chapter is the first polynomial lower bound on the unbounded-error communication complexity of a function in AC^0. As a corollary, we establish the separations $\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}$ and $\Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}$ in communication complexity, thereby solving a longstanding open problem due to Babai et al. [23]. As another corollary, we obtain the first exponential, tight lower bound on the sign-rank of polynomial-size DNF formulas as well as the first exponential lower bound on the size of threshold-of-majority circuits for AC^0. This chapter is joint work with A. A. Razborov.

Chapter 9 explores generalizations of the pattern matrix method to multiparty communication. Recall that the pattern matrix method has been adapted to multiparty communication and has enabled substantial progress in the area. We give a detailed and integrated treatment of these developments, which we hope will serve as a self-contained reference and spur further progress. Covered here are the improved lower bounds for the disjointness function due to Lee and Shraibman [140] and Chattopadhyay and Ada [59], a separation of $\mathsf{NP}^{\mathsf{cc}}$ from $\mathsf{BPP}^{\mathsf{cc}}$ due to David, Pitassi, and Viola [65], and a separation of $\mathsf{NP}^{\mathsf{cc}}$ from $\mathsf{coNP}^{\mathsf{cc}}$ and $\mathsf{coMA}^{\mathsf{cc}}$ due to Gavinsky and the author [81].

Chapter 10 determines the relations and gaps among three complexity measures of a communication problem: product discrepancy, nonproduct discrepancy, and sign-rank. As a corollary, we prove that the containment $\mathsf{PP}^{\mathsf{cc}} \subseteq \mathsf{UPP}^{\mathsf{cc}}$ is proper. Next, we solve an open problem due to Kushilevitz and Nisan [137], exhibiting a gap of $\Theta(1)$ versus $\Theta(n)$ between the product and nonproduct distributional complexities of a function on n-bit strings. Finally, we prove that product discrepancy, originally defined in communication complexity, is equivalent to the learning-theoretic notion of statistical query dimension. Other connections between learning and communication are established here, involving in particular the notion of sign-rank. This chapter is placed strategically to provide a segue from Part I on communication complexity to Part II on learning theory.

Chapter 11 concludes Part I of this thesis with a summary of our contributions in communication complexity and a discussion of several open problems related to our work.

Chapter 12, our first chapter on learning theory, focuses on the problem of PAC learning intersections of halfspaces on the hypercube $\{0,1\}^n$. This problem has long resisted attack and remains a central challenge in the area. Our main result here shows that in fact, under a widely believed cryptographic assumption, no efficient algorithm exists for the task. We obtain analogous hardness results for learning other concept classes, such as majority circuits and arithmetic circuits. Analytic representations of Boolean functions play a key role in our proofs. This chapter is joint work with A. R. Klivans.

Chapter 13 continues our work on lower bounds for learning intersections of halfspaces. Recall that in the previous chapter, we derive representation-independent, cryptographic hardness results for learning intersections of halfspaces on $\{0,1\}^n$. Here, we complement those results with *unconditional* lower bounds for learning intersections of halfspaces in Kearns' statistical query model [104]. In particular, we prove that any statistical-query algorithm for learning the intersection of $\sqrt{n}$ halfspaces on $\{0,1\}^n$ runs in time $\exp\{\Omega(\sqrt{n})\}$. This lower bound is an exponential improvement on previous work. In addition, we derive near-tight, exponential lower bounds on the *threshold density* of this concept class, placing it beyond the scope of Jackson's Harmonic sieve algorithm [98]. This chapter is joint work with A. R. Klivans.

Chapter 14 studies the *agnostic model*, an abstraction of learning from noisy training data. Both algorithmic results and lower bounds for this model have seen limited progress. We contribute several new lower bounds, ruling out the use of the current techniques for learning concept classes as simple as decision lists and disjunctions. Along the way we relate agnostic learning, via the pattern matrix method, to our work on communication complexity as well as to an algorithmic problem known as *approximate inclusion/exclusion.* Parts of this chapter (Sections 14.3–14.5) are joint work with A. R. Klivans.

Chapter 15 takes an in-depth look at the sign-representation of Boolean functions by real polynomials. We prove that, for any Boolean functions f and g, the intersection $f(x) \wedge g(y)$ has threshold degree $O(d)$ if and only if there exist rational functions F, G of degree $O(d)$ with $\|f - F\|_\infty + \|g - G\|_\infty < 1$. This characterization extends to conjunctions of three and more functions as well as various other compositions. This result is of interest because of the applications of the threshold degree in previous chapters and in earlier literature. As a concrete

application in the next chapter, we solve an open problem in learning theory, due to Klivans [120], on the threshold degree of the intersection of two halfspaces.

Chapter 16, as outlined in the previous paragraph, studies the structural complexity of the intersection of two halfspaces. Specifically, we construct two halfspaces on $\{0,1\}^n$ whose intersection has threshold degree $\Theta(\sqrt{n})$, an exponential improvement on previous lower bounds. This solves an open problem due to Klivans [120] and rules out the use of perceptron-based techniques for PAC-learning the intersection of even two halfspaces. We also prove that the intersection of two majority functions has threshold degree $\Omega(\log n)$, which is tight and settles a conjecture of O'Donnell and Servedio [163]. We obtain these results by a thorough study of the rational approximation of halfspaces, along with the relationship between rational approximation and sign-representation proved in the previous chapter.

Chapter 17 summarizes our contributions in computational learning theory and presents a number of open problems related to our work. This chapter concludes Part II on learning theory and the thesis itself.

1.6 Chapter dependencies

We have organized this thesis so as to make certain technical shortcuts available to readers with an interest in one particular result or application. We provide a chart of chapter dependencies in Figure 1.3, with Chapters 11 and 17 (summary and open problems) omitted. This chart only concerns the *main* results of each chapter. Observations, remarks, comparisons, and technical results on a lesser scale warrant additional dependencies.

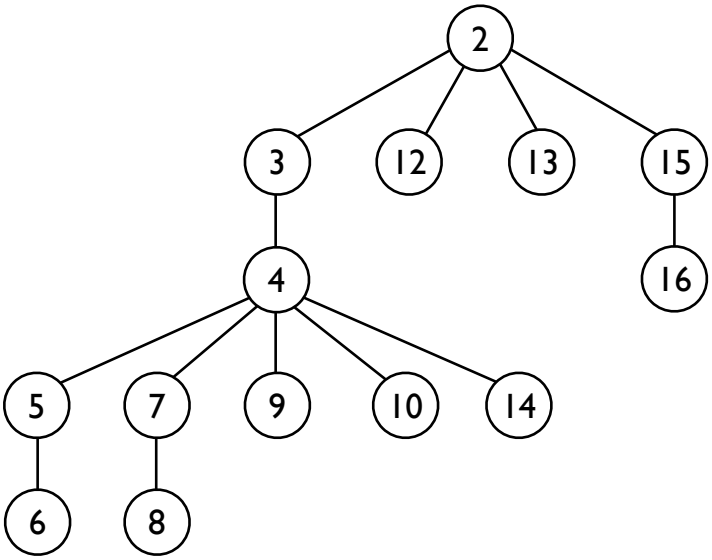

Figure 1.3: A chart of chapter dependencies.

Chapter 2

Notation and Technical Preliminaries

This chapter sets the stage with notation, conventions, and technical background. In particular, we will review key analytic properties of Boolean functions and matrices, which play a central role in this thesis. Also covered here are combinatorial complexity measures of Boolean functions, which ultimately turn out to be closely related to their analytic counterparts.

2.1 General conventions and notation

We view Boolean functions as mappings $X \to \{-1,+1\}$ for a finite set X, where -1 and $+1$ correspond to "true" and "false," respectively. Typically, the domain will be $X = \{0,1\}^n$ or $X = \{0,1\}^n \times \{0,1\}^n$. Given a function $f\colon X \to \{-1,+1\}$, its negation $\neg f\colon X \to \{-1,+1\}$ is given by $\neg f(x) \equiv -f(x)$. Sometimes we will write $\overline{f}$ in place of $\neg f$. Given a function $f\colon X \to \{-1,+1\}$ and a subset $A \subseteq X$, we let $f|_A$ denote the restriction of f to A. In other words, the function $f|_A\colon A \to \{-1,+1\}$ is given by $f_A(x) = f(x)$. The standard functions OR_n, AND_n, MAJ_n, PARITY_n, each a mapping $\{0,1\}^n \to \{-1,+1\}$, are given by

$$\begin{aligned}
\mathrm{OR}_n(x) = -1 &\quad\Leftrightarrow\quad \sum x_i > 0,\\
\mathrm{AND}_n(x) = -1 &\quad\Leftrightarrow\quad \sum x_i = n,\\
\mathrm{MAJ}_n(x) = -1 &\quad\Leftrightarrow\quad \sum x_i > n/2,\\
\mathrm{PARITY}_n(x) = -1 &\quad\Leftrightarrow\quad \sum x_i \text{ is odd.}
\end{aligned}$$

A *predicate* is a mapping $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$. The notation $[n]$ stands for the set $\{1,2,\ldots,n\}$. For a set $S \subseteq [n]$, its *characteristic vector* $\mathbf{1}_S \in \{0,1\}^n$ is defined by

$$(\mathbf{1}_S)_i = \begin{cases} 1 & \text{if } i \in S,\\ 0 & \text{otherwise.}\end{cases}$$

For a finite set X, the notation $\mathscr{P}(X)$ stands for the family of all $2^{|X|}$ subsets of X. For a string $x \in \{0,1\}^n$ and a subset $S \subseteq \{1,2,\ldots,n\}$, we define $x|_S =$

$(x_{i_1}, x_{i_2}, \ldots, x_{i_{|S|}}) \in \{0,1\}^{|S|}$, where $i_1 < i_2 < \cdots < i_{|S|}$ are the elements of S. For $b \in \{0,1\}$, we write $\neg b = 1 - b$. For $x \in \{0,1\}^n$, we adopt the shorthand $|x| = x_1 + \cdots + x_n$. For $x, y \in \{0,1\}^n$, the notation $x \wedge y \in \{0,1\}^n$ refers to the component-wise conjunction of x and y. Analogously, the string $x \vee y$ stands for the component-wise disjunction of x and y. In particular, the notation $|x \wedge y|$ stands for the number of positions in which the strings x and y both have a 1. Throughout this thesis, "log" refers to the logarithm to base 2. As usual, we denote the base of the natural logarithm by $\mathrm{e} = 2.718\ldots$. For any mapping $\phi\colon X \to \mathbb{R}$, where X is a finite set, we adopt the standard notation $\|\phi\|_\infty = \max_{x \in X} |\phi(x)|$ and $\|\phi\|_1 = \sum_{x \in X} |\phi(x)|$. We use the symbol P_d to refer to the family of univariate real polynomials of degree at most d. We adopt the standard definition of the sign function:

$$\operatorname{sgn} t = \begin{cases} -1, & t < 0, \\ 0, & t = 0, \\ 1, & t > 0. \end{cases}$$

Equations and inequalities involving vectors in $\mathbb{R}^n$, such as $x < y$ or $x \geqslant 0$, are to be interpreted component-wise, as usual.

A *halfspace*, also known as a *linear threshold function* or a *linear threshold gate*, is a function $f\colon \{0,1\}^n \to \{-1,+1\}$ of the form $f(x) \equiv \operatorname{sgn}(\sum a_i x_i - \theta)$ for some fixed reals $a_1, \ldots, a_n, \theta$. Observe that a linear threshold gate generalizes the majority function.

A *decision list* is a Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$ specified by a fixed permutation $\sigma\colon [n] \to [n]$, a fixed vector $a \in \{-1,+1\}^{n+1}$, and a fixed vector $b \in \{0,1\}^n$. The computation of f on input $x \in \{0,1\}^n$ proceeds as follows. If $x_{\sigma(i)} \neq b_i$ all $i = 1, 2, \ldots, n$, then one outputs a_{n+1}. Otherwise, one outputs a_i, where $i \in \{1, 2, \ldots, n\}$ is the least integer with $x_{\sigma(i)} = b_i$.

A *Boolean formula* is any Boolean circuit in which each gate, except for the output gate and the inputs $x_1, \ldots, x_n$, feeds into exactly one other gate. An AND/OR/NOT formula is a Boolean formula with gates AND, OR, NOT. A *read-once* formula is a Boolean formula in which the inputs $x_1, \ldots, x_n$ each feed into at most one gate. Note that it is meaningful to speak of read-once AND/OR/NOT formulas, for example.

We now recall the Fourier transform over $\mathbb{Z}_2^n$. Consider the vector space of functions $\{0,1\}^n \to \mathbb{R}$, equipped with the inner product

$$\langle f, g \rangle = 2^{-n} \sum_{x \in \{0,1\}^n} f(x)g(x).$$

For $S \subseteq [n]$, define $\chi_S : \{0,1\}^n \to \{-1,+1\}$ by $\chi_S(x) = (-1)^{\sum_{i \in S} x_i}$. Then $\{\chi_S\}_{S \subseteq [n]}$ is an orthonormal basis for the inner product space in question. As a result, every function $f : \{0,1\}^n \to \mathbb{R}$ has a unique representation of the form

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

where $\hat{f}(S) = \langle f, \chi_S \rangle$. The reals $\hat{f}(S)$ are called the *Fourier coefficients of* f. The *degree* of f, denoted $\deg(f)$, is the quantity $\max\{|S| : \hat{f}(S) \neq 0\}$. It is clear that the degree-d real polynomials on $\{0,1\}^n$ are precisely the functions $f : \{0,1\}^n \to \mathbb{R}$ with $\deg(f) = d$. The orthonormality of $\{\chi_S\}$ immediately yields *Parseval's identity*:

$$\sum_{S \subseteq [n]} \hat{f}(S)^2 = \langle f, f \rangle = \mathop{\mathbf{E}}_{x}[f(x)^2]. \tag{2.1}$$

Observe that the Fourier transform $\hat{f}$ is a real function on $\mathscr{P}(\{1,2,\ldots,n\})$. In particular, we have the shorthands

$$\|\hat{f}\|_\infty = \max_{S \subseteq [n]} |\hat{f}(S)|,$$
$$\|\hat{f}\|_1 = \sum_{S \subseteq [n]} |\hat{f}(S)|.$$

The following fact is immediate from the definition of the Fourier coefficients.

PROPOSITION 2.1. *Let $f: \{0,1\}^n \to \mathbb{R}$ be given. Then*

$$\|\hat{f}\|_\infty \leqslant 2^{-n} \sum_{x \in \{0,1\}^n} |f(x)|.$$

For given functions $f, g: \{0,1\}^n \to \mathbb{R}$, we let $fg: \{0,1\}^n \to \mathbb{R}$ stand for their pointwise multiplication: $(fg)(x) = f(x)g(x)$. In particular, the notation $\widehat{fg}$ stands for the Fourier transform of the function fg.

Let S_n stand for the symmetric group on n elements. For a string $x \in \{0,1\}^n$ and a permutation $\sigma \in S_n$, we write $\sigma x = (x_{\sigma(1)}, \ldots, x_{\sigma(n)})$. A function $\phi: \{0,1\}^n \to \mathbb{R}$ is called *symmetric* if $\phi(x) \equiv \phi(\sigma x)$ for all permutations $\sigma \in S_n$. Equivalently, ϕ is symmetric if the value $\phi(x)$ is uniquely determined by $\sum x_i$. Note that there is a one-to-one correspondence between predicates $D: \{0,1,\ldots,n\} \to \{-1,+1\}$ and symmetric Boolean functions $f: \{0,1\}^n \to \{-1,+1\}$. Namely, one associates a predicate D with the symmetric Boolean function $f(x) \equiv D(\sum x_i)$. Observe also that for every function $\phi: \{0,1\}^n \to \mathbb{R}$ (symmetric or not), the derived function

$$\phi_{\mathrm{symm}}(x) = \mathop{\mathbf{E}}_{\sigma \in S_n} [\phi(\sigma x)]$$

is symmetric. Symmetric functions on $\{0,1\}^n$ are intimately related to univariate polynomials, as demonstrated by Minsky and Papert's *symmetrization argument*.

PROPOSITION 2.2 (Minsky and Papert [153]). *Let $\phi: \{0,1\}^n \to \mathbb{R}$ be given, $d = \deg(\phi)$. Then there is a polynomial $p \in P_d$ with*

$$\mathop{\mathbf{E}}_{\sigma \in S_n} [\phi(\sigma x)] = p(|x|)$$

for all $x \in \{0,1\}^n$.

When speaking of communication complexity, we will abuse the terminology slightly and use the term *symmetric function* to refer to functions $f: \{0,1\}^n \times$

$\{0,1\}^n \to \{-1,+1\}$ of the form $f(x,y) = D(\sum x_i y_i)$ for a given predicate $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$.

2.2 Analytic properties of Boolean functions

Let $f\colon \{0,1\}^n \to \mathbb{R}$ be a given function. As we saw in the previous section, any such function f has an *exact* representation as a linear combination of the characters χ_S, where $|S| \leqslant \deg(f)$. A fundamental question to ask is how closely f can be *approximated* by a degree-d polynomial, where $d \ll \deg(f)$. More formally, for each function $f\colon \{0,1\}^n \to \mathbb{R}$, we define

$$E(f,d) = \min_p \|f - p\|_\infty,$$

where the minimum is over real polynomials of degree up to d. The *ε-approximate degree* of f, denoted $\deg_\varepsilon(f)$, is the least d with $E(f,d) \leqslant \varepsilon$. In words, the ε-approximate degree of f is the least degree of a polynomial that approximates f uniformly within ε. There is an extensive literature on the ε-approximate degree of Boolean functions [160, 165, 100, 52, 4, 15, 199, 221], for the canonical setting $\varepsilon = 1/3$ and various other settings. The choice of $\varepsilon = 1/3$ is a convention and can be replaced by any other constant in $(0,1)$, without affecting $\deg_\varepsilon(f)$ by more than a multiplicative constant:

PROPOSITION 2.3 (Folklore). *Let $f\colon \{0,1\}^n \to \{-1,+1\}$ be a given function, ε a constant with $0 < \varepsilon < 1$. Then*

$$\deg_\varepsilon(f) = \Theta(\deg_{1/3}(f)).$$

PROOF. We assume that $\varepsilon \leqslant 1/3$, the complementary case being analogous. Put $d = \deg_{1/3}(f)$ and fix $\phi \in \operatorname{span}\{\chi_S : |S| \leqslant d\}$ such that $\|f - \phi\|_\infty \leqslant 1/3$. By basic approximation theory [185, Cor. 1.4.1], there exists a univariate polynomial p of degree $O(1/\varepsilon)$ that sends $[-\frac{4}{3}, -\frac{2}{3}] \to [-1-\varepsilon, -1+\varepsilon]$ and $[\frac{2}{3}, \frac{4}{3}] \to [1-\varepsilon, 1+\varepsilon]$. Then $p(\phi(x))$ is the sought approximant of f. □

Another well-studied notion is the *threshold degree* $\deg_{\pm}(f)$, defined for a Boolean function $f:\{0,1\}^n \to \{-1,+1\}$ as the least degree of a real polynomial p with $f(x) \equiv \operatorname{sgn} p(x)$. In words, $\deg_{\pm}(f)$ is the least degree of a polynomial that represents f in sign. This notion has been investigated in numerous works [153, 21, 132, 133, 122, 121, 163] in complexity theory and learning theory. Several synonyms for threshold degree are in use, including "strong degree" [21], "voting polynomial degree" [132], "PTF degree" [164], and "sign degree" [53]. It is useful to keep in mind the following alternate characterization of the threshold degree, as a limit process:

$$\deg_{\pm}(f) = \lim_{\varepsilon \searrow 0} \deg_{1-\varepsilon}(f).$$

So far we have considered representations of Boolean functions by *real* polynomials. Restricting the polynomials to have *integer* coefficients yields another useful representation scheme. The main complexity measure here is the sum of the absolute values of the coefficients. Specifically, for a Boolean function $f:\{0,1\}^n \to \{-1,+1\}$, its *degree-$d$ threshold weight* $W(f,d)$ is defined to be the minimum $\sum_{|S| \leqslant d} |\lambda_S|$ over all integers λ_S such that

$$f(x) \equiv \operatorname{sgn}\left(\sum_{S \subseteq \{1,\ldots,n\},\, |S| \leqslant d} \lambda_S \chi_S(x) \right). \tag{2.2}$$

If no such integers λ_S can be found, we put $W(f,d) = \infty$. It is clear that the following three conditions are equivalent: $W(f,d) = \infty$; $E(f,d) = 1$; $d < \deg_{\pm}(f)$. In all expressions involving $W(f,d)$, we adopt the standard convention that $1/\infty = 0$ and $\min\{t,\infty\} = t$ for any real t. A closely related notion is the *degree-d threshold density* of f, denoted $\operatorname{dns}(f,d)$ and defined to be the minimum $|\{S : \lambda_S \neq 0\}|$ over all coefficients λ_S such that (2.2) holds. If no such coefficients can be found, we let $\operatorname{dns}(f,d) = \infty$. We define

$$W(f) = \min_{d=0,1,\ldots,n} W(f,d)$$

and

$$\mathrm{dns}(f) = \min_{d=0,1,\ldots,n} \mathrm{dns}(f,d).$$

There is a substantial body of work on threshold weight and density as well as their applications [153, 155, 44, 207, 45, 82, 32, 218, 133, 124, 131, 169, 170].

As one might expect, representations of Boolean functions by real and integer polynomials are closely related. In particular, we have the following relationship between $E(f,d)$ and $W(f,d)$.

THEOREM 2.4. *Let $f:\{0,1\}^n \to \{-1,+1\}$ be given. Then for $d = 0,1,\ldots,n$,*

$$\frac{1}{1-E(f,d)} \leqslant W(f,d) \leqslant \frac{2}{1-E(f,d)} \left\{ \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d} \right\}^{3/2},$$

with the convention that $1/0 = \infty$.

Similar statements have been noted earlier [133, 53]. We prove Theorem 2.4 by modifying a recent analysis due to Buhrman et al. [53, Cor. 1].

PROOF OF THEOREM 2.4. Recall that $W(f,d) = \infty$ if and only if $E(f,d) = 1$. In what follows, we focus on the complementary case when $W(f,d) < \infty$ and $E(f,d) < 1$.

To prove the lower bound on $W(f,d)$, fix integers λ_S with $\sum_{|S|\leqslant d} |\lambda_S| = W(f,d)$ such that the polynomial $p(x) = \sum_{|S|\leqslant d} \lambda_S \chi_S(x)$ satisfies $f(x) \equiv \operatorname{sgn} p(x)$. Then $1 \leqslant f(x)p(x) \leqslant W(f,d)$ and therefore

$$E(f,d) \leqslant \left\| f - \frac{1}{W(f,d)}\, p \right\|_\infty \leqslant 1 - \frac{1}{W(f,d)}.$$

To prove the upper bound on $W(f,d)$, fix any degree-d polynomial p such that $\|f-p\|_\infty = E(f,d)$. Define $\delta = 1 - E(f,d) > 0$ and $N = \sum_{i=0}^{d} \binom{n}{i}$. For a

real t, let $\operatorname{rnd} t$ be the result of rounding t to the closest integer, so that $|t - \operatorname{rnd} t| \leqslant 1/2$. We claim that the polynomial

$$q(x) = \sum_{|S| \leqslant d} \operatorname{rnd}(M\,\hat{p}(S))\chi_S(x),$$

where $M = 3N/(4\delta)$, satisfies $f(x) \equiv \operatorname{sgn} q(x)$. Indeed,

$$\begin{aligned}
\left| f(x) - \frac{1}{M} q(x) \right| &\leqslant |f(x) - p(x)| + \frac{1}{M} |Mp(x) - q(x)| \\
&\leqslant 1 - \delta + \frac{1}{M} \sum_{|S| \leqslant d} |M\,\hat{p}(S) - \operatorname{rnd}(M\,\hat{p}(S))| \\
&\leqslant 1 - \delta + \frac{N}{2M} \\
&< 1.
\end{aligned}$$

It remains to examine the sum of the coefficients of q. We have:

$$\begin{aligned}
\sum_{|S| \leqslant d} |\operatorname{rnd}(M\,\hat{p}(S))| &\leqslant \frac{1}{2} N + M \sum_{|S| \leqslant d} |\hat{p}(S)| \\
&\leqslant \frac{1}{2} N + M \left(N \mathop{\mathbf{E}}_{x} \left[p(x)^2 \right] \right)^{1/2} \\
&\leqslant \frac{2N\sqrt{N}}{\delta},
\end{aligned}$$

where the second step follows by an application of the Cauchy-Schwarz inequality and Parseval's identity (2.1). □

A *polynomial threshold function* (PTF) of degree d is any function $f\colon \{0,1\}^n \to \{-1,+1\}$ that can be expressed as

$$f(x) = \operatorname{sgn}\left(\sum_{|S|\leqslant d} \lambda_S \chi_S(x)\right) \tag{2.3}$$

for some integer coefficients λ_S. The *weight* of a polynomial threshold function f is the minimum $\sum |\lambda_S|$ over all choices of coefficients λ_S for which (2.3) holds. In this terminology, the threshold degree of a given Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$ is the least d for which f is a degree-d polynomial threshold function. Analogously, the degree-d threshold weight of a given Boolean function f is the weight of f as a degree-d polynomial threshold function. Observe that the terms "halfspace" and "linear threshold function" are equivalent.

For a fixed d, a sequence of degree-d polynomial threshold functions $f_1, f_2, \ldots, f_n, \ldots$, where $f_n\colon \{0,1\}^n \to \{-1,+1\}$, is called *light* if $W(f_n, d) \leqslant n^c$ for some constant and all n. More generally, a sequence $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_n, \ldots$, where $\mathscr{C}_n$ is a family of degree-d polynomial threshold functions on $\{0,1\}^n$, is called *light* if $\max_{f\in\mathscr{C}_n} W(f,d) \leqslant n^c$ for some constant $c > 1$ and all n. A common use of this terminology is to speak of light halfspaces, as in Chapter 12.

We now recall Paturi's tight estimate [165] of the approximate degree for each symmetric Boolean function.

THEOREM 2.5 (Paturi [165]). *Let $f\colon \{0,1\}^n \to \{-1,+1\}$ be a given function such that $f(x) \equiv D(\sum x_i)$ for some predicate $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$. Then*

$$\deg_{1/3}(f) = \Theta\left(\sqrt{n\ell_0(D)} + \sqrt{n\ell_1(D)}\right),$$

where $\ell_0(D) \in \{0,1,\ldots,\lfloor n/2\rfloor\}$ and $\ell_1(D) \in \{0,1,\ldots,\lceil n/2\rceil\}$ are the smallest integers such that D is constant in the range $[\ell_0(D), n-\ell_1(D)]$.

The notions of uniform approximation and sign-representation, defined above for Boolean functions on the hypercube, extend naturally to any finite domain

$X \subset \mathbb{R}^n$. Specifically, the ε-approximate degree of a function $f\colon X \to \{-1,+1\}$, denoted $\deg_\varepsilon(f)$, is the least degree of a real polynomial with $\|f-p\|_\infty \leqslant \varepsilon$. Analogously, the threshold degree of a function $f\colon X \to \{-1,+1\}$, denoted $\deg_\pm(f)$, is the least degree of a real polynomial with $f(x) \equiv \operatorname{sgn} p(x)$.

2.3 Combinatorial properties of Boolean functions

Every function $f\colon \{0,1\}^n \to \mathbb{R}$ has a unique representation of the form

$$f(x) = \sum_{S \subseteq \{1,\ldots,n\}} \alpha_S \prod_{i \in S} x_i$$

for some reals α_S. This representation is to be contrasted with the Fourier transform of f, discussed above. We define the *monomial count* or *number of monomials* in f by $\operatorname{mon}(f) = |\{S : \alpha_S \neq 0\}|$.

For $i = 1,2,\ldots,n$, we let $e_i \in \{0,1\}^n$ stand for the vector with 1 in the ith component and zeroes everywhere else. For a set $S \subseteq \{1,\ldots,n\}$, we define $e_S \in \{0,1\}^n$ by $e_S = \sum_{i \in S} e_i$. In particular, $e_\varnothing = 0$. Fix a Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$. For $\ell = 1,2,\ldots,n$, the *ℓ-block sensitivity* of f, denoted $\operatorname{bs}_\ell(f)$, is defined as the largest k for which there exist nonempty disjoint sets $S_1,\ldots,S_k \subseteq \{1,\ldots,n\}$, each containing no more than ℓ elements, such that

$$f(z \oplus e_{S_1}) = f(z \oplus e_{S_2}) = \cdots = f(z \oplus e_{S_k}) \neq f(z)$$

for some $z \in \{0,1\}^n$. One distinguishes two extremal cases. The *sensitivity* of f, denoted $\operatorname{s}(f)$, is defined by $\operatorname{s}(f) = \operatorname{bs}_1(f)$. The *block sensitivity* of f, denoted $\operatorname{bs}(f)$, is defined by $\operatorname{bs}(f) = \operatorname{bs}_n(f)$. In this context, the term *block* simply refers to a subset $S \subseteq \{1,2,\ldots,n\}$. We say that block $S \subseteq \{1,2,\ldots,n\}$ is *sensitive* for f on input z if $f(z) \neq f(z \oplus e_S)$. Sensitivity, block sensitivity, and ℓ-block sensitivity were introduced by Cook et al. [63], Nisan [158], and Kenyon and Kutin [109], respectively. Buhrman and de Wolf [54] define an additional variant of sensitivity: the *zero block sensitivity* of f, denoted $\operatorname{zbs}(f)$, is the largest k for which there exist

nonempty disjoint sets $S_1, \ldots, S_k \subseteq \{1, \ldots, n\}$ such that

$$f(z \oplus e_{S_1}) = f(z \oplus e_{S_2}) = \cdots = f(z \oplus e_{S_k}) \neq f(z)$$

for some $z \in \{0,1\}^n$ with $z|_{S_1 \cup \cdots \cup S_k} = (0, 0, \ldots, 0)$. As an illustrative example, the AND function on n bits satisfies $\mathrm{s}(\mathrm{AND}_n) = \mathrm{bs}(\mathrm{AND}_n) = n$ and $\mathrm{zbs}(\mathrm{AND}_n) = 1$. Similarly, we have $\mathrm{s}(\mathrm{OR}_n) = \mathrm{bs}(\mathrm{OR}_n) = \mathrm{zbs}(\mathrm{OR}_n) = n$.

A *decision tree* is a classical combinatorial model of computation. Given a function $f : \{0,1\}^n \to \{-1, +1\}$, consider the task of determining $f(x)$ by adaptively querying individual bits of $x = (x_1, \ldots, x_n)$. The complexity measure of interest is the maximum number of bits queried on any input x before the value $f(x)$ is known. Any valid algorithm for f in this model can be represented by a rooted binary tree whose internal nodes are labeled with integers from $\{1, 2, \ldots, n\}$, whose leafs are labeled -1 or $+1$, and whose edges are labeled 0 or 1. This representation, called a *decision tree*, has the obvious semantics: one starts at the root node, queries the variable that corresponds to the node's label i, branches to the right or left depending on the value of x_i, and continues the process until a leaf is reached and the value $f(x)$ is known. The *decision tree complexity* $\mathrm{dt}(f)$ is the least complexity of a query algorithm for f, or equivalently, the least depth of a decision tree for f.

It is well-known [158, 160] that the decision tree complexity $\mathrm{dt}(f)$, block sensitivity $\mathrm{bs}(f)$, degree $\deg(f)$, and approximate degree $\deg_{1/3}(f)$ are polynomially related for every Boolean function $f : \{0,1\}^n \to \{-1, +1\}$. For the purposes of this thesis, we note down the following relationships.

THEOREM 2.6 (Nisan and Smolensky [55, Thm. 12]). *For every Boolean function* $f : \{0,1\}^n \to \{-1, +1\}$,

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

THEOREM 2.7 (Nisan and Szegedy [160]). *For every* $f : \{0,1\}^n \to \{-1,+1\}$,

$$\mathrm{bs}(f) \leqslant O(\deg_{1/3}(f)^2).$$

THEOREM 2.8 (Beals et al. [28, §5]). *For every* $f : \{0,1\}^n \to \{-1,+1\}$,

$$\mathrm{dt}(f) \leqslant \mathrm{bs}(f)^3.$$

A consequence of Theorems 2.7 and 2.8 is the following result.

THEOREM 2.9 (Beals et al. [28, §5]). *For every* $f : \{0,1\}^n \to \{-1,+1\}$,

$$\mathrm{dt}(f) \leqslant O(\deg_{1/3}(f)^6).$$

It is unknown whether sensitivity is polynomially related to block sensitivity and the other quantities that figure in Theorems 2.6–2.9. An elegant result due to Kenyon and Kutin [109] shows, however, that the sensitivity and ℓ-block sensitivity of a Boolean function are polynomially related for all constant ℓ. For our purposes, the case $\ell = 2$ is all that is needed.

THEOREM 2.10 (Kenyon and Kutin [109, Cor. 3.1]). *For every* $f : \{0,1\}^n \to \{-1,+1\}$,

$$\Omega(\sqrt{\mathrm{bs}_2(f)}) \leqslant \mathrm{s}(f) \leqslant \mathrm{bs}(f).$$

The lower bound in Theorem 2.10 is asymptotically tight [187]. We refer the interested reader to the excellent survey by Buhrman and de Wolf [55] for further results on decision trees and related combinatorial complexity measures.

2.4 Matrix analysis

Matrix analysis plays a considerable role in this thesis. We assume familiarity with basic matrix-analytic notions and facts, viz., the singular value decomposition, positive semidefinite matrices, matrix similarity, matrix trace and its properties, the Kronecker product and its spectral properties, the relation between singular values and eigenvalues, and eigenvalue computation for matrices of simple form. This material is entirely classical, with a variety of excellent texts available. Standard references on the subject are the monographs by Golub and Van Loan [85] and Horn and Johnson [95]. The review below is limited to notation and the more substantial results.

The symbol $\mathbb{R}^{m\times n}$ refers to the family of all $m \times n$ matrices with real entries. A *sign matrix* is a matrix with entries $+1$ or -1. A *probability matrix* is a matrix whose entries are nonnegative and sum to 1. We specify matrices by their generic entry, e.g., $A = [F(i,j)]_{i,j}$. In most matrices that arise in this work, the exact ordering of the columns (and rows) is irrelevant. In such cases we describe a matrix by the notation $[F(i,j)]_{i\in I,\, j\in J}$, where I and J are some index sets. We denote the rank of $A \in \mathbb{R}^{m\times n}$ by $\operatorname{rk} A$. We also write

$$\|A\|_\infty = \max_{i,j} |A_{ij}|, \qquad \|A\|_1 = \sum_{i,j} |A_{ij}|.$$

We denote the singular values of A by $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \cdots \geqslant \sigma_{\min\{m,n\}}(A) \geqslant 0$. Recall that the spectral norm, trace norm, and Frobenius norm of A are given by

$$\begin{aligned} \|A\| &= \max_{x\in\mathbb{R}^n,\, \|x\|=1} \|Ax\| = \sigma_1(A), \\ \|A\|_\Sigma &= \sum \sigma_i(A), \\ \|A\|_{\mathrm{F}} &= \sqrt{\sum A_{ij}^2} = \sqrt{\sum \sigma_i(A)^2}. \end{aligned}$$

The vector norm $\|\cdot\|$ above and throughout this thesis is the Euclidean norm $\|\cdot\|_2$. The following relationship follows at once by the Cauchy-Schwarz inequality:

$$\|A\|_\Sigma \leqslant \|A\|_{\mathrm{F}} \sqrt{\operatorname{rk} A} \qquad (A \in \mathbb{R}^{m\times n}). \tag{2.4}$$

For a square matrix $A \in \mathbb{R}^{n\times n}$, its trace is given by $\operatorname{tr} A = \sum A_{ii}$.

Recall that every matrix $A \in \mathbb{R}^{m\times n}$ has a singular value decomposition $A = U\Sigma V^{\mathsf{T}}$, where U and V are orthogonal matrices and Σ is diagonal with entries $\sigma_1(A), \sigma_2(A), \ldots, \sigma_{\min\{m,n\}}(A)$. For $A, B \in \mathbb{R}^{m\times n}$, we write $\langle A, B\rangle = \sum A_{ij}B_{ij} = \operatorname{tr}(AB^{\mathsf{T}})$. The Hadamard product of matrices $A = [A_{ij}]$ and $B = [B_{ij}]$, also known as the Schur product, is given by $A \circ B = [A_{ij}B_{ij}]$. The Kronecker product of $A = [A_{ij}]$ and $B = [B_{kl}]$ is given by $A \otimes B = [A_{ij}B_{kl}]_{(i,k),(j,l)}$. A useful consequence of the singular value decomposition is:

$$\langle A, B\rangle \leqslant \|A\|\ \|B\|_\Sigma \qquad (A, B \in \mathbb{R}^{m\times n}). \tag{2.5}$$

Following Razborov [177], we define the *ε-approximate trace norm* of a matrix $F \in \mathbb{R}^{m\times n}$ by

$$\|F\|_{\Sigma,\varepsilon} = \min\{\|A\|_\Sigma : \|F - A\|_\infty \leqslant \varepsilon\}.$$

The next proposition is a trivial consequence of (2.5).

PROPOSITION 2.11. *Let $F \in \mathbb{R}^{m\times n}$ and $\varepsilon \geqslant 0$. Then*

$$\|F\|_{\Sigma,\varepsilon} \geqslant \sup_{\Psi\neq 0} \frac{\langle F, \Psi\rangle - \varepsilon\|\Psi\|_1}{\|\Psi\|}.$$

PROOF. Fix any $\Psi \neq 0$ and A such that $\|F - A\|_\infty \leqslant \varepsilon$. Then $\langle A, \Psi\rangle \leqslant \|A\|_\Sigma\|\Psi\|$ by (2.5). On the other hand, $\langle A, \Psi\rangle \geqslant \langle F, \Psi\rangle - \|A - F\|_\infty\|\Psi\|_1 \geqslant \langle F, \Psi\rangle - \varepsilon\|\Psi\|_1$. Comparing these two estimates gives the sought lower bound on $\|A\|_\Sigma$. □

Following Buhrman and de Wolf [54], we define the *ε-approximate rank* of a matrix $F \in \mathbb{R}^{m\times n}$ by

$$\operatorname{rk}_\varepsilon F = \min\{\operatorname{rk} A : \|F - A\|_\infty \leqslant \varepsilon\}.$$

The approximate rank and approximate trace norm are related by virtue of the singular value decomposition, as follows.

PROPOSITION 2.12. *Let $F \in \mathbb{R}^{m\times n}$ and $\varepsilon \geqslant 0$ be given. Then*

$$\operatorname{rk}_\varepsilon F \geqslant \frac{(\|F\|_{\Sigma,\varepsilon})^2}{\sum_{i,j}(|F_{ij}| + \varepsilon)^2}.$$

PROOF (adapted from [126]). Fix A with $\|F - A\|_\infty \leqslant \varepsilon$. Then

$$\begin{aligned}\|F\|_{\Sigma,\varepsilon} &\leqslant \|A\|_\Sigma \\ &\leqslant \|A\|_{\mathrm{F}}\sqrt{\operatorname{rk} A} \\ &\leqslant \left(\sum_{i,j}(|F_{ij}| + \varepsilon)^2\right)^{1/2}\sqrt{\operatorname{rk} A}.\end{aligned}$$

□

The *sign-rank* of a matrix $F \in \mathbb{R}^{m\times n}$, denoted $\operatorname{rk}_\pm F$, is the least rank of a real matrix A such that $F_{ij}A_{ij} > 0$ for all i, j with $F_{ij} \neq 0$. In words, the sign-rank of F is the least rank of a real matrix with the same sign pattern as F, except for any zero entries in F. For sign matrices $F \in \{-1, +1\}^{m\times n}$, an equivalent definition is

$$\operatorname{rk}_\pm F = \lim_{\varepsilon \searrow 0} \operatorname{rk}_{1-\varepsilon} F.$$

This fundamental notion has been studied in contexts as diverse as matrix analysis, communication complexity, circuit complexity, and computational learning

theory [167, 13, 35, 70, 71, 122, 142] and figures prominently in this thesis. In general, the sign-rank of a matrix can be vastly smaller than its rank. For example, consider the following nonsingular matrices of order $n \geqslant 3$, representing well-known problems GREATER-THAN and EQUALITY in communication complexity:

$$\begin{bmatrix} 1 & & & & & \\ & 1 & & 1 & & \\ & & 1 & & & \\ & & & \ddots & & \\ -1 & & & & 1 & \\ & & & & & 1 \end{bmatrix}, \qquad \begin{bmatrix} -1 & & & & & \\ & -1 & & & 1 & \\ & & -1 & & & \\ & & & \ddots & & \\ & 1 & & & -1 & \\ & & & & & -1 \end{bmatrix}.$$

These matrices have sign-rank at most 2 and 3, respectively. Indeed, the first matrix has the same sign pattern as the matrix $[2(j-i)+1]_{i,j}$. The second has the same sign pattern as the matrix $[(1-\varepsilon) - \langle v_i, v_j \rangle]_{i,j}$, where $v_1, v_2, \ldots, v_n \in \mathbb{R}^2$ are arbitrary pairwise distinct unit vectors and ε is a suitably small positive real, cf. Paturi and Simon [167, §5].

We close this section with a classical bound on the trace norm of a matrix product, which we state with a proof for the reader's convenience.

PROPOSITION 2.13. *For all real matrices A and B of compatible dimensions,*

$$\|AB\|_\Sigma \leqslant \|A\|_{\mathrm{F}} \, \|B\|_{\mathrm{F}}.$$

PROOF. Write the singular value decomposition $AB = U\Sigma V^{\mathsf{T}}$. Let $u_1, u_2, \ldots$ and $v_1, v_2, \ldots$ stand for the columns of U and V, respectively. By definition, $\|AB\|_\Sigma$

is the sum of the diagonal entries of Σ. We have:

$$\begin{aligned}\|AB\|_\Sigma &= \sum (U^\mathsf{T} ABV)_{ii} = \sum (u_i^\mathsf{T} A)(Bv_i) \\ &\leqslant \sum \|A^\mathsf{T} u_i\| \, \|Bv_i\| \\ &\leqslant \sqrt{\sum \|A^\mathsf{T} u_i\|^2} \sqrt{\sum \|Bv_i\|^2} = \|U^\mathsf{T} A\|_\mathrm{F} \, \|BV\|_\mathrm{F} \\ &= \|A\|_\mathrm{F} \, \|B\|_\mathrm{F}. \end{aligned}$$

□

Hadamard product and inner product generalize from matrices to arbitrary tensors, as follows. For tensors $A, B \colon X_1 \times \cdots \times X_k \to \mathbb{R}$, where X_i is a finite set, $i = 1, 2, \ldots, k$, define

$$\langle A, B \rangle = \sum_{x_1 \in X_1} \cdots \sum_{x_k \in X_k} A(x_1, \ldots, x_k) B(x_1, \ldots, x_k).$$

Define the *Hadamard product* of A and B to be the tensor $A \circ B \colon X_1 \times \cdots \times X_k \to \mathbb{R}$ given by $(A \circ B)(x_1, \ldots, x_k) = A(x_1, \ldots, x_k) B(x_1, \ldots, x_k)$.

2.5 Learning-theoretic complexity measures

Let X be a finite set, such as $X = \{0,1\}^n$. Let $\mathscr{C}$ be a given set of functions $X \to \{-1, +1\}$. Computational learning theory is concerned with the task of efficiently constructing an approximation to an unknown function f based only on the membership $f \in \mathscr{C}$ and on the values of f on a small sample of points from X. The given set $\mathscr{C}$ of functions is called a *concept class.* We identify $\mathscr{C}$ with its *characteristic matrix* $M_\mathscr{C}$, given by $M_\mathscr{C} = [f(x)]_{f \in \mathscr{C}, x \in X}$. In what follows, we use $\mathscr{C}$ and its characteristic matrix interchangeably.

Let μ be a probability distribution over X. Then the following is a natural notion of distance between functions $f, g: X \to \{-1, +1\}$:

$$\Delta_\mu(f, g) = \mathop{\mathbf{P}}_{x \sim \mu}\Big[f(x) \neq g(x)\Big].$$

A central model in learning theory is Valiant's *probably approximately correct* model [213], commonly abbreviated PAC. A concept class $\mathscr{C}$ is *PAC-learnable* to accuracy ε and confidence δ under distribution μ from m examples if there is an algorithm L that, for every unknown $f \in \mathscr{C}$, takes as input i.i.d. examples $x^{(1)}, x^{(2)}, \ldots, x^{(m)} \sim \mu$ and their labels $f(x^{(1)}), f(x^{(2)}), \ldots, f(x^{(m)})$, and with probability at least $1 - \delta$ produces a hypothesis $h: X \to \{-1, +1\}$ with $\Delta_\mu(h, f) \leqslant \varepsilon$. The probability is over the random choice of examples and any internal randomization in L.

For a sign matrix A and its corresponding concept class, define its *Vapnik-Chervonenkis dimension* $\mathrm{vc}(A)$ to be the largest d such that A features a $2^d \times d$ submatrix whose rows are the distinct elements of $\{-1, +1\}^d$. The Vapnik-Chervonenkis dimension is a combinatorial quantity that exactly captures the learning complexity of a concept class. This is borne out by the following classical theorem due to Vapnik and Chervonenkis [215] and Blumer et al. [42].

THEOREM 2.14 (Vapnik-Chervonenkis Theorem [215, 42]). *Let $\mathscr{C}$ be a concept class and μ a distribution. Then $\mathscr{C}$ is learnable to accuracy ε and confidence δ under μ from*

$$m = O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta} + \frac{\mathrm{vc}(\mathscr{C})}{\varepsilon}\log\frac{1}{\varepsilon}\right)$$

examples. Moreover, any algorithm that outputs as a hypothesis some member of $\mathscr{C}$ consistent with the given m examples will successfully learn $\mathscr{C}$.

The precise formulation above can be found, along with a proof, in the textbook by Kearns and Vazirani [108, Thm. 3.3]. Theorem 2.14 almost matches the information-theoretic lower bounds on the number of examples necessary. These

lower bounds come in many different flavors; for example, see [108, Thm. 3.5]. We will need the following specialized version.

PROPOSITION 2.15 (Sherstov [197]). *Let μ be a probability distribution and $\mathscr{C}$ a concept class such that $\Delta_\mu(f, f') > \varepsilon$ for every two distinct $f, f' \in \mathscr{C}$. Then learning $\mathscr{C}$ to accuracy $\varepsilon/2$ and confidence δ under μ requires $\log\{|\mathscr{C}|(1-\delta)\}$ examples.*

PROOF. Let L be a learner for $\mathscr{C}$ that uses m examples, achieving accuracy $\varepsilon/2$ and confidence δ. View L as a function $L(x^{(1)}, y_1, \ldots, x^{(m)}, y_m, r)$ that takes labeled training examples and a random string as input and outputs a hypothesis. With this notation, we have:

$$\mathop{\mathbf{E}}_{f\in\mathscr{C}}\left[\mathop{\mathbf{P}}_{x^{(1)},\ldots,x^{(m)},r}\left[\Delta_\mu\Big(f,\ L\Big(x^{(1)}, f(x^{(1)}), \ldots, x^{(m)}, f(x^{(m)}), r\Big)\Big) \leqslant \frac{\varepsilon}{2}\right]\right] \geqslant 1-\delta.$$

Reordering the expectation and probability operators yields

$$\mathop{\mathbf{E}}_{x^{(1)},\ldots,x^{(m)},r}\left[\mathop{\mathbf{P}}_{f\in\mathscr{C}}\left[\Delta_\mu\Big(f,\ L\Big(x^{(1)}, f(x^{(1)}), \ldots, x^{(m)}, f(x^{(m)}), r\Big)\Big) \leqslant \frac{\varepsilon}{2}\right]\right] \geqslant 1-\delta.$$

Thus, there is a fixed choice of $x^{(1)}, \ldots, x^{(m)}, r$ for which

$$\mathop{\mathbf{P}}_{f\in\mathscr{C}}\left[\Delta_\mu\Big(f,\ L\Big(x^{(1)}, f(x^{(1)}), \ldots, x^{(m)}, f(x^{(m)}), r\Big)\Big) \leqslant \frac{\varepsilon}{2}\right] \geqslant 1-\delta. \tag{2.6}$$

With $x^{(1)}, \ldots, x^{(m)}, r$ fixed in this way, algorithm L becomes a deterministic mapping from $\{-1,+1\}^m$ to the hypothesis space. In particular, L can output at most 2^m different hypotheses. Equation (2.6) says that L produces $(\varepsilon/2)$-approximates for at least $(1-\delta)\,|\mathscr{C}|$ functions in $\mathscr{C}$. Since no hypothesis can be an $(\varepsilon/2)$-approximator for two different functions in $\mathscr{C}$, we have $2^m \geqslant (1-\delta)\,|\mathscr{C}|$. □

Of considerable algorithmic importance in learning theory is the notion of a Euclidean embedding. Let $A \in \{-1,+1\}^{M\times N}$ be a given sign matrix, correspond-

ing to some concept class. A *Euclidean embedding* of A is any system of vectors $u_1, \ldots, u_M \in \mathbb{R}^k$ and $v_1, \ldots, v_N \in \mathbb{R}^k$ (for some finite k) such that $\langle u_i, v_j \rangle A_{ij} > 0$ for all i, j. The integer k is called the *dimension* of the embedding. The quantity

$$\gamma = \min_{i,j} \frac{|\langle u_i, v_j \rangle|}{\|u_i\| \, \|v_j\|}$$

is called the *margin* of the embedding. Observe that in the terminology of Section 2.4, the smallest dimension of a Euclidean embedding of A is precisely $\mathrm{rk}_{\pm} A$. The *margin complexity* $\mathrm{mc}(A)$ is the minimum $1/\gamma$ over all embeddings of A.

Recall that e_i denotes the vector with 1 in the ith component and zeroes elsewhere. The following is a trivial embedding of a sign matrix $A = [a_1 \mid \ldots \mid a_N] \in \{-1, +1\}^{M \times N}$: label the rows by vectors $e_1, \ldots, e_M \in \mathbb{R}^M$ and the columns by vectors $a_1, \ldots, a_N$. It is clear that this embedding has dimension M and margin $1/\sqrt{M}$. By interchanging the roles of the rows and columns, we obtain the following well-known fact:

PROPOSITION 2.16. *Let $A \in \{-1, +1\}^{M \times N}$. Then*

$$1 \leqslant \mathrm{rk}_{\pm} A \leqslant \min\{M, N\},$$
$$1 \leqslant \mathrm{mc}(A) \leqslant \min\left\{\sqrt{M}, \sqrt{N}\right\}.$$

The final learning-theoretic complexity measure in this section pertains to learning by statistical queries, a model due to Kearns [104] that will receive thorough treatment in Part II of this thesis. Let X be a finite set. For a family $\mathscr{C}$ of functions $X \to \{-1, +1\}$ and a distribution μ on X, the *statistical query (SQ) dimension* of $\mathscr{C}$ under μ, denoted $\mathrm{sq}_{\mu}(\mathscr{C})$, is defined as the largest integer d for which there are functions $f_1, f_2, \ldots, f_d \in \mathscr{C}$ such that

$$\left| \mathop{\mathbf{E}}_{x \sim \mu} \left[f_i(x) f_j(x) \right] \right| \leqslant \frac{1}{d}$$

for all $i \neq j$. We put $\mathrm{sq}(\mathscr{C}) = \max_\mu \mathrm{sq}_\mu(\mathscr{C})$. For a sign matrix $A \in \{-1,+1\}^{M\times N}$, we define $\mathrm{sq}_\mu(A) = \mathrm{sq}_\mu(\mathscr{C})$ and $\mathrm{sq}(A) = \mathrm{sq}(\mathscr{C})$, where $\mathscr{C}$ is the concept class with characteristic matrix A. To illustrate, the Hadamard matrix $H = [(-1)^{\sum x_i y_i}]_{x,y\in\{0,1\}^n}$ satisfies $H^\mathsf{T} H = 2^n I$ and thus has SQ dimension 2^n with respect to the uniform distribution on the columns.

An excellent reference for further background on computational learning is the textbook by Kearns and Vazirani [108].

2.6 Summary

This chapter introduced a variety of definitions and background results on the analytic and combinatorial theory of Boolean functions and matrices. For the reader's convenience, we provide a table of the key quantities introduced here, along with a brief definition and a page reference for each. A comprehensive table of all symbols and notation in this thesis is available in Appendix A.

Symbol	*Meaning*	*Pages*
$[n]$	the set $\{1,2,\ldots,n\}$	23
$\mathbf{1}_S,\ e_S$	the characteristic vector of $S \subseteq \{1,2,\ldots,n\}$	23, 32
e_i	the characteristic vector of $\{i\}$	32
$\lvert x\rvert$	the Hamming weight $\sum x_i$	24
$x\lvert_S$	projection of $x \in \{0,1\}^n$ onto the set $S \subset \{1,2,\ldots,n\}$	23
P_d	the family of univariate polynomials of degree up to d	24
S_n	the symmetric group on n elements	26
σx	the string $(x_{\sigma(1)},\ldots,x_{\sigma(n)})$	26
$\hat{f}(S)$	Fourier transform of a function $f\colon\{0,1\}^n \to \mathbb{R}$	25
fg	pointwise product of functions $f, g\colon\{0,1\}^n \to \mathbb{R}$	26
χ_S	character of the Fourier transform on $\mathbb{Z}_2^n$	25
$E(f,d)$	least error in a degree-d uniform approximation of f	27
$W(f,d)$	degree-d threshold weight of f	28
$\mathrm{dns}(f,d)$	degree-d threshold density of f	28
$W(f)$	threshold weight of f	28
$\mathrm{dns}(f)$	threshold density of f	29
$\deg_\varepsilon(f)$	ε-approximate degree of f	32

Symbol	*Meaning*	*Pages*
$\deg_{\pm}(f)$	threshold degree of f	32
$\mathrm{mon}(f)$	monomial count of f	32
$\mathrm{s}(f)$	sensitivity of f	32
$\mathrm{bs}(f)$	block sensitivity of f	32
$\mathrm{bs}_{\ell}(f)$	ℓ-block sensitivity of f	32
$\mathrm{zbs}(f)$	zero block sensitivity of f	32
$\mathrm{dt}(f)$	decision tree complexity of f	33
$\mathrm{rk}\, A$	rank of a real matrix A	35
$\mathrm{rk}_{\varepsilon}\, A$	ε-approximate rank of a real matrix A	36
$\mathrm{rk}_{\pm}\, A$	sign-rank of a real matrix A	37
$\mathrm{tr}\, A$	trace of a square real matrix A	36
$\sigma_i(A)$	the ith largest singular value of a real matrix A	35
$\langle A, B \rangle$	inner product of real matrices or tensors A and B	36
$A \circ B$	Hadamard product of real matrices or tensors A and B	36
$A \otimes B$	Kronecker product of real matrices A and B	36
$\lVert \cdot \rVert_{\infty}$	ℓ_{∞} norm on real functions and matrices	24, 35
$\lVert \cdot \rVert_{1}$	ℓ_{1} norm on real functions and matrices	24, 35
$\lVert \cdot \rVert$	Euclidean norm on vectors or spectral norm on matrices	36
$\lVert \cdot \rVert_{\mathrm{F}}$	Frobenius norm on matrices	35
$\lVert \cdot \rVert_{\Sigma}$	trace norm on matrices	35
$\lVert \cdot \rVert_{\Sigma,\varepsilon}$	ε-approximate trace norm on matrices	36
$\mathrm{vc}(A)$	Vapnik-Chervonenkis dimension of the sign matrix A	40
$\mathrm{mc}(A)$	margin complexity of the sign matrix A	42
$\mathrm{sq}(A)$	statistical query (SQ) dimension of the sign matrix A	42

Part I

Communication Complexity

Chapter 3

Fundamentals of Communication Complexity

This chapter sets the stage for our results on communication complexity. We describe the canonical framework of communication complexity and review the deterministic, randomized, and nondeterministic models. More advanced formalisms, such as the quantum, unbounded-error, and multiparty models, will be introduced as needed in later chapters. We close this chapter by discussing an important technique for communication lower bounds, the *discrepancy method*, and its generalizations.

3.1 Deterministic, nondeterministic, and randomized models

Communication complexity studies the amount of communication necessary in order to compute a given Boolean function when its arguments are distributed among several parties. Initiated in a seminal paper by Yao [223] over three decades ago, communication complexity has evolved in a central topic of complexity theory, studied for its intrinsic appeal as well as numerous applications to circuit complexity, computational learning theory, streaming algorithms, data structures, quantum computing, and other fundamental topics in theoretical computer science.

The simplest models in communication complexity are the deterministic and nondeterministic models. Let $f: X \times Y \to \{-1, +1\}$ be a given function, where X and Y are finite sets. There are two parties, traditionally called Alice and Bob. Alice receives an input $x \in X$, Bob receives $y \in Y$, and their objective is to compute $f(x, y)$. The exact meaning of *compute* will depend on the communication model in question. Alice and Bob communicate by exchanging bits 0 and 1 via a shared communication channel, according to a *protocol* agreed upon in advance. Formally, a protocol is a fixed agreement between Alice and Bob that specifies:

(1) for each history of previously transmitted bits, an output value -1 or $+1$ if the communication is over, and an indication of who is to speak next if the communication is to continue;

(2) for the party to speak next, a value 0 or 1 that is to be transmitted, based on the history of previously transmitted bits and the party's own input (x for Alice, y for Bob).

The *cost* of a protocol is the maximum number of bits exchanged on any input (x, y). A protocol is said to compute f *deterministically* if the output of the pro-

tocol on input (x, y) is always $f(x, y)$. A protocol is said to compute f *nondeterministically* if the protocol always outputs $+1$ on inputs $(x, y) \in f^{-1}(+1)$ and outputs -1 at least on some executions for every input $(x, y) \in f^{-1}(-1)$. The *deterministic* (respectively, *nondeterministic*) communication complexity of f is the least cost of a protocol that computes f deterministically (respectively, nondeterministically). The deterministic and nondeterministic communication complexities of f are denoted $D(f)$ and $N(f)$, respectively. The *co-nondeterministic* communication complexity of f is the quantity $N(-f)$.

In the *randomized* model, the parties further have access to an unlimited supply of shared random bits. The cost of a randomized protocol is still the maximum number of bits exchanged between the parties on any input. A randomized protocol is said to *compute f with error ε* if on every input (x, y), the protocol produces the correct output $f(x, y)$ with probability at least $1-\varepsilon$. The *ε-error randomized* communication complexity of f, denoted $R_\varepsilon(f)$, is the least cost of a randomized protocol that computes f with error ε. The canonical setting is $\varepsilon = 1/3$, corresponding to *bounded-error* randomized communication complexity, but any other parameter $\varepsilon \in (0, 1/2)$ can be considered. In particular, it is of considerable importance to understand *small-bias* randomized communication complexity, which corresponds to $\varepsilon = \frac{1}{2} - o(1)$. It is useful to keep in mind that the error probability of a randomized protocol can be reduced from $1/3$ to any desired constant $\varepsilon > 0$ by executing the protocol $\Theta\left(\log \frac{1}{\varepsilon}\right)$ times and outputting the majority answer. In other words, one has

$$R_\varepsilon(f) = O\left(R_{1/3}(f) \log \frac{1}{\varepsilon}\right)$$

by basic probability, and thus the setting $\varepsilon = 1/3$ entails no loss of generality in the study of bounded-error communication complexity.

Applications of communication complexity have motivated numerous other models of communication, including the unbounded-error model, multiparty models, and quantum models. We will take a close look these models in later chapters of this thesis. In the meantime, the interested reader may wish to consult the book by Kushilevitz and Nisan [137], a treasure trove of information on communication

complexity. An introductory chapter in a thesis, no matter how detailed, cannot rival the encyclopedic treatment given to the subject by Kushilevitz and Nisan.

3.2 Discrepancy method

Crucial to the study of communication complexity is the notion of *discrepancy*. This concept figures prominently in the study of bounded-error and small-bias communication as well as various applications, such as learning theory and circuit complexity. Formally, for a Boolean function $f: X \times Y \to \{-1, +1\}$ and a probability distribution μ on $X \times Y$, the *discrepancy* of f under μ is defined by

$$\operatorname{disc}_\mu(f) = \max_{\substack{S \subseteq X, \\ T \subseteq Y}} \left| \sum_{x \in S} \sum_{y \in T} \mu(x, y) f(x, y) \right|. \tag{3.1}$$

We put

$$\operatorname{disc}(f) = \min_\mu \{\operatorname{disc}_\mu(f)\}.$$

A *product distribution* μ on $X \times Y$ is a distribution that has the representation $\mu(x, y) = \mu_X(x)\mu_Y(y)$ for some distributions μ_X and μ_Y on X and Y, respectively. We let $\operatorname{disc}^\times(f)$ stand for the minimum discrepancy of f under product distributions:

$$\operatorname{disc}^\times(f) = \min_{\mu \text{ product}} \{\operatorname{disc}_\mu(f)\}.$$

Throughout this thesis, we will identify a function $f: X \times Y \to \{-1, +1\}$ with its communication matrix $F = [f(x, y)]_{x,y}$. For example, we will use the conventions $\operatorname{disc}_\mu(F) = \operatorname{disc}_\mu(f)$ and $\operatorname{disc}(F) = \operatorname{disc}(f)$. Furthermore, it is natural to identify a probability matrix $P = [P_{xy}]$ with the probability distribution μ on $X \times Y$ given by $\mu(x, y) = P_{xy}$. In such cases we write $\operatorname{disc}_P(F)$ to mean $\operatorname{disc}_\mu(f)$.

We will shortly see that, in a precise and meaningful sense, the discrepancy $\mathrm{disc}(F)$ can be regarded as a combinatorial complexity measure of the sign matrix F.

Intimately related to randomized communication complexity is what is called the *distributional* communication complexity. For a function $f: X \times Y \to \{-1, +1\}$ and a distribution μ on $X \times Y$, the *μ-distributional* communication complexity $D^{\mu}_{\varepsilon}(f)$ is the least cost of a deterministic protocol $\Pi: X \times Y \to \{-1, +1\}$ such that $\mathbf{P}_{\mu}[\Pi(x, y) = f(x, y)] \geqslant 1 - \varepsilon$. Yao observed [224] the following relationship in light of the famous minimax theorem for zero-sum games.

THEOREM 3.1 (Yao [224]). *Let $f: X \times Y \to \{-1, +1\}$ be a given function, for finite sets X and Y. Then*

$$R_{\varepsilon}(f) = \max_{\mu} \{D^{\mu}_{\varepsilon}(f)\}.$$

A detailed derivation of this result is also available in the monograph of Kushilevitz and Nisan [137, Thm. 3.20]. Yao's observation has been the basis for most lower bounds on randomized communication complexity: one defines a probability distribution μ on $X \times Y$ and argues that the cost of the best deterministic protocol with error at most ε over μ must be high. Discrepancy provides a powerful means of proving lower bounds on distributional communication complexity, as shown in the following result (see Kushilevitz and Nisan [137], pp. 36–38).

PROPOSITION 3.2 (Discrepancy method). *For every function $f: X \times Y \to \{-1, +1\}$, every $\gamma \in (0, 1)$, and every distribution μ on $X \times Y$,*

$$R_{1/2-\gamma/2}(f) \geqslant D^{\mu}_{1/2-\gamma/2}(f) \geqslant \log\left(\frac{\gamma}{\mathrm{disc}_{\mu}(f)}\right).$$

Proposition 3.2 is frequently used to derive communication lower bounds not only for bounded-error protocols but also protocols with error $\frac{1}{2} - o(1)$. In this latter context, one says that a protocol has *advantage* γ if the protocol has error probability $\frac{1}{2} - \frac{1}{2}\gamma$. Proposition 3.2 bears out our earlier remark that discrepancy

can be viewed as a combinatorial complexity measure of a sign matrix, with small discrepancy corresponding to high complexity.

Now that we have seen that discrepancy is an important quantity, we discuss techniques for estimating it. The following result gives a bound on the discrepancy that does not feature a bothersome maximization operator as in (3.1).

LEMMA 3.3 (Discrepancy estimate [25, 62, 173, 69]). *Let $f: X \times Y \to \{-1, +1\}$ be a given function, where X and Y are finite sets. Let μ be a probability distribution over $X \times Y$. Then*

$$\operatorname{disc}_\mu(f)^2 \leqslant |X| \sum_{y,y' \in Y} \left| \sum_{x \in X} \mu(x,y)\mu(x,y') f(x,y) f(x,y') \right|.$$

PROOF (adapted from [173]). Let $S \subseteq X$ and $T \subseteq Y$ be sets for which the maximum is achieved in (3.1), and let $R = S \times T$. Define $\alpha_x = 1$ for all $x \in S$, and likewise $\beta_y = 1$ for all $y \in T$. For all remaining x and y, let α_x and β_y be independent random variables distributed uniformly over $\{-1, +1\}$. Passing to expectations,

$$\begin{aligned}
&\left| \mathbf{E}\left[\sum_{x,y} \alpha_x \beta_y \mu(x,y) f(x,y) \right] \right| \\
&\quad = \left| \sum_{(x,y) \in R} \mathbf{E}[\underbrace{\alpha_x \beta_y}_{=1}] \mu(x,y) f(x,y) + \sum_{(x,y) \notin R} \underbrace{\mathbf{E}[\alpha_x \beta_y]}_{=0} \mu(x,y) f(x,y) \right| \\
&\quad = \operatorname{disc}_\mu(M).
\end{aligned}$$

In particular, there exists a fixed assignment $\alpha_x, \beta_y \in \{-1, +1\}$ for all x, y such that

$$\operatorname{disc}_\mu(f) \leqslant \left| \sum_{x,y} \alpha_x \beta_y \mu(x,y) f(x,y) \right|.$$

Squaring both sides and applying the Cauchy-Schwarz inequality,

$$\begin{aligned}\operatorname{disc}_\mu(f)^2 &\leqslant |X| \sum_x \left(\alpha_x \sum_y \beta_y \mu(x,y) f(x,y) \right)^2 \\ &= |X| \sum_{y,y'} \beta_y \beta_{y'} \sum_x \mu(x,y)\mu(x,y') f(x,y) f(x,y') \\ &\leqslant |X| \sum_{y,y'} \left| \sum_x \mu(x,y)\mu(x,y') f(x,y) f(x,y') \right|,\end{aligned}$$

as desired. □

We close this section with a matrix-analytic reformulation of discrepancy, following Kushilevitz and Nisan [137, Ex. 3.29].

PROPOSITION 3.4. *Let X, Y be finite sets, $f : X \times Y \to \{-1, +1\}$ a given function. Then*

$$\operatorname{disc}_P(f) \leqslant \sqrt{|X|\,|Y|}\,\|P \circ F\|,$$

where $F = [f(x,y)]_{x \in X,\, y \in Y}$ and P is any probability matrix. In particular,

$$\operatorname{disc}(f) \leqslant \sqrt{|X|\,|Y|} \min_P \|P \circ F\|,$$

where the minimum is over probability matrices P.

PROOF. We have:

$$\begin{aligned}\operatorname{disc}_P(f) &= \max_{S,T} \left|\mathbf{1}_S^{\mathsf{T}} (P \circ F)\, \mathbf{1}_T\right| \\ &\leqslant \max_{S,T} \left\{\|\mathbf{1}_S\| \cdot \|P \circ F\| \cdot \|\mathbf{1}_T\|\right\} \\ &= \|P \circ F\| \sqrt{|X|\,|Y|}.\end{aligned}$$

□

To illustrate Proposition 3.4, consider the well-studied inner product function $\mathrm{IP}_n \colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$, given by $\mathrm{IP}_n(x,y) = (-1)^{\sum x_i y_i}$.

PROPOSITION 3.5 (Discrepancy of inner product [61, 23, 137]). *Let $\mathscr{U}$ stand for the uniform distribution on $\{0,1\}^n \times \{0,1\}^n$. Then*

$$\operatorname{disc}_{\mathscr{U}}(\mathrm{IP}_n) \leqslant 2^{-n/2}.$$

In particular,

$$R_{1/3}(\mathrm{IP}_n) \geqslant D^{\mathscr{U}}_{1/3}(\mathrm{IP}_n) = \Theta(n). \tag{3.2}$$

PROOF (adapted from [137]). Let $H = [(-1)^{\sum x_i y_i}]_{x,y \in \{0,1\}^n}$ be the communication matrix of IP_n. By Proposition 3.4,

$$\operatorname{disc}_{\mathscr{U}}(\mathrm{IP}_n) \leqslant 2^n \|4^{-n} H\| = 2^{-n/2},$$

where the last equality exploits the fact that $H^{\mathsf{T}} H = 2^n I$. This upper bound on the discrepancy, along with Proposition 3.2, forces (3.2). □

3.3 Generalized discrepancy method

As one can see from Proposition 3.2, the discrepancy method is particularly strong in that it gives communication lower bounds for protocols with error probability vanishingly close to trivial, $\frac{1}{2} - o(1)$. This strength of the discrepancy method is at once a weakness when it comes to proving lower bounds for bounded-error communication. Let us consider a classical example of this phenomenon, the disjointness function.

EXAMPLE 3.6. The well-studied disjointness function $\mathrm{DISJ}_n \colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ is given by $\mathrm{DISJ}_n(x,y) = \bigvee_{i=1}^n (x_i \wedge y_i)$. We claim that this function has an efficient randomized protocol with nonnegligible probability of correctness. Specifically, Alice and Bob randomly pick an index $i \in \{1,2,\ldots,n\}$ and exchange the bits x_i and y_i. If $x_i = y_i = 1$, they output -1 ("true"); otherwise, they output -1 with probability $\frac{1}{2} - \frac{1}{4n}$ and $+1$ with the complementary probability. Calculations reveal that this protocol has constant cost and error probability at most $\frac{1}{2} - \frac{1}{4n}$, whence $R_{1/2-1/4n}(\mathrm{DISJ}_n) = O(1)$. In view of Proposition 3.2, we conclude that $\mathrm{disc}(\mathrm{DISJ}_n) = \Omega(1/n)$. In particular, the discrepancy method cannot yield a lower bound better than $\Omega(\log n)$ on the bounded-error communication complexity of DISJ_n. Yet it is well-known [102, 176] that $R_{1/3}(\mathrm{DISJ}_n) = \Theta(n)$.

The *generalized* discrepancy method is a clever extension of the traditional discrepancy method that avoids the difficulty just cited. To the best of our knowledge, this idea originated in a paper by Klauck [114, Thm. 4] and was reformulated more broadly by Razborov [177]. The development in [114] and [177] takes place in the quantum model of communication. However, the basic mathematical technique is in no way restricted to the quantum model, and we will focus here on a model-independent version of the generalized discrepancy method from [203]. Let $f \colon X \times Y \to \{-1,+1\}$ be a given function whose communication complexity we wish to estimate. The underlying communication model is *irrelevant* at this point. Suppose we can find a function $h \colon X \times Y \to \{-1,+1\}$ and a distribution μ on $X \times Y$ that satisfy the following two properties. First, the functions f and h are well correlated under μ:

$$\mathop{\mathbf{E}}_{(x,y)\sim\mu} [f(x,y)h(x,y)] \geqslant \varepsilon, \tag{3.3}$$

where $\varepsilon > 0$ is a given constant. Second, no low-cost protocol Π in the given model of communication can compute h to a substantial advantage under μ. Formally, if Π is a protocol in the given model with cost C bits (with output values ± 1), then

$$\mathop{\mathbf{E}}_{(x,y)\sim\mu}[h(x,y)\,\mathbf{E}\,[\Pi(x,y)]] \leqslant 2^{O(C)}\gamma, \tag{3.4}$$

where $\gamma = o(1)$. The inner expectation in (3.4) is over the internal operation of the protocol on the fixed input (x, y).

If the conditions (3.3) and (3.4) hold, we claim that any protocol in the given model that computes f with error at most $\varepsilon/3$ on each input must have cost $\Omega(\log\{\varepsilon/\gamma\})$. Indeed, let Π be a protocol with $\mathbf{P}[\Pi(x,y) \neq f(x,y)] \leqslant \varepsilon/3$ for all x, y. Then standard manipulations reveal:

$$\mathop{\mathbf{E}}_{\mu}[h(x,y)\,\mathbf{E}\,[\Pi(x,y)]] \geqslant \mathop{\mathbf{E}}_{\mu}[f(x,y)h(x,y)] - 2\cdot\frac{\varepsilon}{3} \geqslant \frac{\varepsilon}{3},$$

where the last step uses (3.3). In view of (3.4), this shows that Π must have cost $\Omega(\log\{\varepsilon/\gamma\})$.

We attach the term *generalized discrepancy method* to this abstract framework. Observe that original discrepancy method, Proposition 3.2, corresponds to the case when $f = h$ and the communication takes place in the two-party randomized model. The purpose of our abstract discussion was to expose the fundamental mathematical technique in question, which is independent of the communication model. Indeed, the communication model enters the picture only in the proof of (3.4). It is here that the analysis must exploit the particularities of the model. We will now specialize our discussion to the two-party model of bounded-error communication. It will be convenient to state this result in matrix-analytic notation.

THEOREM 3.7 (Sherstov [203], implicit). *Let $F = [F_{xy}]_{x\in X,\, y\in Y}$ be a given sign matrix, for some finite sets X and Y. Then for all sign matrices $H = [H_{xy}]$ and all*

probability matrices $P = [P_{xy}]$,

$$2^{R_\varepsilon(F)} \geqslant \frac{\langle F, H \circ P\rangle - 2\varepsilon}{\operatorname{disc}_P(H)}. \tag{3.5}$$

In particular,

$$2^{R_\varepsilon(F)} \geqslant \sup_{\Psi \neq 0} \frac{\langle F, \Psi\rangle - 2\varepsilon \|\Psi\|_1}{\|\Psi\| \sqrt{|X|\,|Y|}}. \tag{3.6}$$

Note that one recovers Proposition 3.2, the ordinary discrepancy method, by setting $H = F$ in (3.5).

PROOF OF THEOREM 3.7. We first show that (3.6) follows from (3.5). By linearity, one may assume that $\|\Psi\|_1 = 1$, in which case $\Psi = H \circ P$ for some sign matrix H and some probability matrix P. But then (3.6) is immediate from (3.5) by the method of Proposition 3.4.

It remains to prove (3.5). Put $c = R_\varepsilon(F)$. Theorem 3.1 immediately shows that there exists a deterministic protocol $\Pi\colon X \times Y \to \{-1, +1\}$ with communication cost at most c and $\mathbf{P}_P[F_{xy} \neq \Pi(x, y)] \leqslant \varepsilon$. Viewing the protocol as the sign matrix $\Pi = [\Pi(x, y)]_{x,y}$, we obtain

$$\langle \Pi, H \circ P\rangle \geqslant \langle F, H \circ P\rangle - 2\varepsilon.$$

On the other hand, the ordinary discrepancy method (Proposition 3.2) states that

$$\langle \Pi, H \circ P\rangle \leqslant 2^c \operatorname{disc}_P(H).$$

Comparing the last two inequalities completes the proof. □

In subsequent chapters, we will see that Theorem 3.7 generalizes in a straightforward way to quantum communication and multiparty communication.

3.4 Communication complexity classes

Analogous to computational complexity, it is natural to define complexity classes for communication. Throughout this section, the symbol $\{f_n\}$ shall stand for a family of functions $f_1, f_2, \ldots, f_n, \ldots$, where $f_n : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$.

As one might expect by analogy with computational complexity, the classes P^{cc} and $\mathsf{BPP}^{\mathsf{cc}}$ correspond to communication problems with efficient deterministic and randomized protocols, respectively. Formally, a function family $\{f_n\}$ is in P^{cc} if and only if for some constant $c > 1$ and all $n > c$, one has $D(f_n) \leqslant \log^c n$. Similarly, a function family $\{f_n\}$ is in $\mathsf{BPP}^{\mathsf{cc}}$ if and only if for some constant $c > 1$ and all $n > c$, one has $R_{1/3}(f_n) \leqslant \log^c n$.

The next two complexity classes that we consider correspond to small-bias communication and are analogues of the class PP in computational complexity. The first of these, $\mathsf{PP}^{\mathsf{cc}}$, was originally defined in [23] as the class of communication problems that have an efficient protocol with nonnegligible bias. More precisely, a family $\{f_n\}$ is in $\mathsf{PP}^{\mathsf{cc}}$ if and only if

$$R_{\frac{1}{2}-\exp(-\log^c n)}(f_n) \leqslant \log^c n$$

for some constant $c > 1$ and all $n > c$. For our purposes, it will be more convenient to use an equivalent characterization of $\mathsf{PP}^{\mathsf{cc}}$ in terms of discrepancy:

THEOREM 3.8 (Klauck [116]). *A family $\{f_n\}$ is in $\mathsf{PP}^{\mathsf{cc}}$ if and only if*

$$\operatorname{disc}(f_n) \geqslant 2^{-\log^c n}$$

for some constant $c > 1$ and all $n > c$.

The other class that corresponds to small-bias communication is $\mathsf{UPP}^{\mathsf{cc}}$. The definition of this class requires *private-coin* randomized protocols, in which Alice and Bob each have an unlimited private source of random bits. This is in contrast to the *public-coin* randomized protocols considered earlier, in which Alice and Bob share an unlimited source of random bits. Unless specified otherwise, all randomized protocols in this thesis are public-coin, and we will use the term

"randomized protocol" as a shorthand for "public-coin randomized protocol." Now, a family $\{f_n\}$ is in the class $\mathsf{UPP}^{\mathsf{CC}}$ if and only if for some constant $c > 1$ and all $n > c$, there is a private-coin randomized protocol with cost at most $\log^c n$ that computes f_n with error probability strictly less than $1/2$ on every input. This original definition from [23] admits an elegant matrix-analytic characterization:

THEOREM 3.9 (Paturi and Simon [167]). *A family $\{f_n\}$ is in the class $\mathsf{UPP}^{\mathsf{CC}}$ if and only if*

$$\mathrm{rk}_{\pm}\, F_n \leqslant 2^{\log^c n}$$

for some constant $c > 1$ and all $n > c$, where $F_n = [f(x,y)]_{x,y \in \{0,1\}^n}$.

The reader may wonder in what way this purely matrix-analytic definition is related to communication. We will discuss this relationship in detail in Chapter 7, dedicated to *unbounded-error* communication complexity. For now, we record the containment

$$\mathsf{PP}^{\mathsf{CC}} \subseteq \mathsf{UPP}^{\mathsf{CC}}. \tag{3.7}$$

due to Babai et al. [23]. This containment will be formally proved in Section 10.5, where we will also show that it is proper.

We now turn to the communication-complexity analogue of the polynomial hierarchy PH. A function $f_n \colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ is called a *rectangle* if there exist subsets $A, B \subseteq \{0,1\}^n$ such that

$$f_n(x,y) = -1 \quad \Leftrightarrow \quad x \in A,\ y \in B.$$

We call f_n the *complement of a rectangle* if the negated function $\neg f_n = -f_n$ is a rectangle.

DEFINITION 3.10 (Babai et al. [23]).

(1) A family $\{f_n\}$ is in Π_0^{cc} if and only if each f_n is a rectangle. A family $\{f_n\}$ is in Σ_0^{cc} if and only if $\{\neg f_n\}$ is in Π_0^{cc}.

(2) Fix an integer $k = 1, 2, 3, 4, \ldots$. A family $\{f_n\}$ is in Σ_k^{cc} if and only if for some constant $c > 1$ and all $n > c$,

$$f_n = \bigvee_{i_1=1}^{2^{\log^c n}} \bigwedge_{i_2=1}^{2^{\log^c n}} \bigvee_{i_3=1}^{2^{\log^c n}} \cdots \bigodot_{i_k=1}^{2^{\log^c n}} g_n^{i_1,i_2,\ldots,i_k},$$

where $\bigodot = \bigvee$ (resp., $\bigodot = \bigwedge$) for k odd (resp., even); and each $g_n^{i_1,i_2,\ldots,i_k}$ is a rectangle (resp., the complement of a rectangle) for k odd (resp., even). A family $\{f_n\}$ is in Π_k^{cc} if and only if $\{\neg f_n\}$ is in Σ_k^{cc}.

(3) The polynomial hierarchy is given by $\mathsf{PH}^{\mathsf{cc}} = \bigcup_k \Sigma_k^{\mathsf{cc}} = \bigcup_k \Pi_k^{\mathsf{cc}}$, where $k = 0, 1, 2, 3, \ldots$ ranges over all constants.

(4) A family $\{f_n\}$ is in $\mathsf{PSPACE}^{\mathsf{cc}}$ if and only if for some constant $c > 1$ and all $n > c$,

$$f_n = \bigvee_{i_1=1}^{2^{\log^c n}} \bigwedge_{i_2=1}^{2^{\log^c n}} \bigvee_{i_3=1}^{2^{\log^c n}} \cdots \bigvee_{i_k=1}^{2^{\log^c n}} g_n^{i_1,i_2,\ldots,i_k},$$

where $k < \log^c n$ is odd and each $g_n^{i_1,i_2,\ldots,i_k}$ is a rectangle.

Thus, the zeroth level (Σ_0^{cc} and Π_0^{cc}) of the polynomial hierarchy consists of rectangles and complements of rectangles, the simplest functions in communication complexity. It is also straightforward to check that $\{f_n\} \in \Sigma_1^{\mathsf{cc}}$ if and only if $N(f_n) \leqslant \log^c n$ for some constant $c > 1$ and all $n > c$. Likewise, $\{f_n\} \in \Pi_1^{\mathsf{cc}}$ if and only if $N(-f_n) \leqslant \log^c n$ for some constant $c > 1$ and all $n > c$. In summary, the first level of the polynomial hierarchy corresponds to functions with efficient nondeterministic and co-nondeterministic protocols. For this reason, one uses the following equivalent notation: $\mathsf{NP}^{\mathsf{cc}} = \Sigma_1^{\mathsf{cc}}$, $\mathsf{coNP}^{\mathsf{cc}} = \Pi_1^{\mathsf{cc}}$.

We will revisit the above complexity classes several times in this thesis. In the meantime, we refer the interested reader to the original article by Babai et al. [23] for further results.

3.5 Summary

For the reader's convenience, we summarize in tabular form the communication-related notation introduced in this chapter.

Symbol	*Meaning*	*Pages*
$D(f)$	deterministic communication complexity of f	48
$N(f)$	nondeterministic communication complexity of f	48
$R_\varepsilon(f)$	ε-error randomized communication complexity of f	48
$D^\mu_\varepsilon(f)$	ε-error μ-distributional communication complexity of f	50
$\mathrm{disc}_\mu(f)$	discrepancy of f with respect to μ	49
$\mathrm{disc}(f)$	minimum discrepancy of f under any distribution	49
$\mathrm{disc}^\times(f)$	minimum discrepancy of f under a product distribution	49
P^{cc}	sign matrices with low deterministic complexity	57
$\mathsf{NP}^{\mathsf{cc}}$	sign matrices with low nondeterministic complexity	59
$\mathsf{coNP}^{\mathsf{cc}}$	sign matrices with low co-nondeterministic complexity	59
$\mathsf{BPP}^{\mathsf{cc}}$	sign matrices with low randomized complexity	57
$\mathsf{PP}^{\mathsf{cc}}$	sign matrices with nonnegligible discrepancy	57
$\mathsf{UPP}^{\mathsf{cc}}$	sign matrices with low sign-rank	58
$\mathsf{PH}^{\mathsf{cc}}$	polynomial hierarchy in communication	59
Σ_k^{cc}, Π_k^{cc}	kth level of the polynomial hierarchy	59
$\mathsf{PSPACE}^{\mathsf{cc}}$	polynomial space in communication	59

Chapter 4

Bounded-Error Communication and Discrepancy

In this chapter, we address the challenge of proving communication lower bounds in the randomized model, both for bounded-error protocols and for small-bias protocols. Specifically, we develop a novel technique, the *pattern matrix method*, that converts standard analytic properties of Boolean functions into lower bounds for the associated communication problems. As an application, we establish the separations $\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$ and $\Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$ in communication complexity and solve an open problem in circuit complexity due to Krause and Pudlák [132]. Various other applications will be presented in later chapters as we further develop our technique.

4.1 Introduction

Randomized protocols have been the focus of much research in communication complexity since the introduction of the area by Yao three decades ago [223]. A variety of techniques have been developed for proving communication lower bounds, e.g., [102, 176, 82, 171, 56, 144, 77]. The main contribution of this chapter is a strong new technique for lower bounds on communication complexity, the *pattern matrix method*. The method converts analytic properties of Boolean functions into lower bounds for the corresponding communication problems. The analytic properties in question, discussed in Section 2.2, pertain to the approximation and sign-representation of a given Boolean function by real polynomials of low degree, which are among the oldest and most studied objects in theoretical computer science. In other words, the pattern matrix method takes the wealth of results available on the representations of Boolean functions by real polynomials and puts them at the disposal of communication complexity.

As the method's name suggests, the central concept in this chapter is what we call a *pattern matrix*. We introduce the communication problem of computing

$$f(x|_V),$$

where $f\colon \{0,1\}^t \to \{-1,+1\}$ is a fixed Boolean function; the string $x \in \{0,1\}^n$ is Alice's input (n is a multiple of t); and the set $V \subset \{1,2,\ldots,n\}$ with $|V| = t$ is Bob's input. In words, this communication problem corresponds to a situation

when the function f depends on only t of the inputs $x_1, \ldots, x_n$. Alice knows the values of all the inputs $x_1, \ldots, x_n$ but does not know which t of them are relevant. Bob, on the other hand, knows which t inputs are relevant but does not know their values. For the purposes of this introduction, one can think of the (n, t, f)-*pattern matrix* as the matrix $[f(x|_V)]_{x,V}$, where V ranges over the $(n/t)^t$ sets that have exactly one element from each block of the following partition:

$$\{1, \ldots, n\} = \left\{1, 2, \ldots, \frac{n}{t}\right\} \cup \left\{\frac{n}{t} + 1, \ldots, \frac{2n}{t}\right\} \cup \cdots \cup \left\{\frac{(t-1)n}{t} + 1, \ldots, n\right\}.$$

We defer the precise definition to Section 4.2. Observe that restricting V to be of special form only makes our results stronger.

As a first result, we prove in Section 4.4 that the randomized communication complexity of the (n, t, f)-pattern matrix F obeys

$$R_\delta(F) \geqslant \frac{1}{2} \deg_\varepsilon(f) \log\left(\frac{n}{t}\right) - \log\left(\frac{1}{\varepsilon - 2\delta}\right) \tag{4.1}$$

for any $\varepsilon \in [0, 1)$ and any $\delta < \varepsilon/2$. This equation gives lower bounds for both bounded-error protocols and small-bias protocols. For example, it follows that

$$R_{1/3}(F) \geqslant \Omega\Big(\deg_{1/3}(f) \log \frac{n}{t}\Big). \tag{4.2}$$

This lower bound on bounded-error communication is tight up to a polynomial factor, even for deterministic protocols. The lower bounds (4.1) and (4.2) are of interest because pattern matrices arise as submatrices in natural communication problems. For example, (4.2) shows that for every function $f : \{0, 1\}^t \to \{-1, +1\}$, the composition

$$F(x, y) = f(\ldots, (x_{i,1} y_{i,1} \vee x_{i,2} y_{i,2} \vee x_{i,3} y_{i,3} \vee x_{i,4} y_{i,4}), \ldots)$$

has bounded-error communication complexity $\Omega(\deg_{1/3}(f))$. We will see a number of other such examples in later chapters.

In Section 4.5, we prove an additional lower bound for small-bias communication in terms of threshold weight $W(f,d)$. In particular, we are able to characterize the discrepancy of every pattern matrix:

$$\operatorname{disc}(F) \leqslant \min_{d=1,\ldots,t} \max \left\{ \left(\frac{2t}{W(f,d-1)} \right)^{1/2}, \left(\frac{t}{n} \right)^{d/2} \right\},$$

which is essentially tight. In Section 4.6, we apply this characterization to the well-studied class AC^0, the class of polynomial-size constant-depth circuits of AND, OR, and NOT gates. We construct the first function $f \in \mathsf{AC}^0$ with exponentially small discrepancy, thereby establishing the separations

$$\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}, \qquad \Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$$

in communication complexity. These separations are best possible in that $\mathsf{PP}^{\mathsf{cc}}$ trivially contains the first two levels of the polynomial hierarchy: Σ_0^{cc}, Σ_1^{cc}, Π_0^{cc}, Π_1^{cc}. Independently of the author, Buhrman et al. [53] exhibited another AC^0 function with exponentially small discrepancy, with a much different proof. An advantage of our technique is that it works not only for AC^0 but for an arbitrary base function f with high threshold weight. In particular, we are able to give a simple alternate proof of the result by Buhrman et al. [53] using pattern matrices.

As another application of our lower bounds for small-bias communication, we prove in Section 4.7 that the AC^0 function

$$f(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij})$$

cannot be computed by a depth-2 majority circuit of size less than $2^{\Theta(m)}$. This solves an open problem due to Krause and Pudlák [132] and matches Allender's classical

result [11] that every function in AC^0 can be efficiently computed by a depth-3 majority circuit.

An overview of the proofs. The setting in which to best describe our proofs is the *generalized discrepancy method,* discussed in Section 3.3. Let $F(x, y)$ be a Boolean function whose bounded-error communication complexity is of interest. Recall that the generalized discrepancy method asks for a Boolean function $H(x, y)$ and a distribution μ on (x, y)-pairs such that:

(1) the functions F and H have correlation $\Omega(1)$ under μ; and

(2) all low-cost protocols have negligible advantage in computing H under μ.

If such H and μ indeed exist, it follows that no low-cost protocol can compute F to high accuracy (otherwise it would be a good predictor for the hard function H as well). This method generalizes Yao's original discrepancy method [137], in which $H = F$. The advantage of the generalized version is that it makes it possible, in theory, to prove lower bounds for functions such as DISJOINTNESS, to which the traditional method does not apply.

The hard part is, of course, finding H and μ with the desired properties. Except in rather restricted cases [114, Thm. 4], it was not known how to do it. As a result, the generalized discrepancy method was of limited practical use prior to this work. Here we overcome this difficulty, obtaining H and μ for a broad range of problems, namely, the communication problems of computing $f(x|_V)$.

Pattern matrices are a crucial ingredient of our solution. In Section 4.2, we derive a closed-form expression for the singular values of a pattern matrix and their multiplicities. This spectral information reduces our search from H and μ to a much smaller and simpler object, namely, a function $\psi\colon \{0,1\}^t \to \mathbb{R}$ with certain properties. On the one hand, ψ must be well-correlated with the base function f. On the other hand, ψ must be orthogonal to all low-degree polynomials. We establish the existence of such ψ in Section 4.3 by passing to the *linear programming dual* of the approximate degree of f. Although the approximate degree and its dual are classical notions, we are not aware of any previous use of this duality to prove communication lower bounds. For the results that feature threshold weight, we combine the above program with the dual characterization of threshold weight.

Looking ahead, we will see in the next chapter that these results apply to the *quantum* model, regardless of prior entanglement. In that context, we will contrast our approach with the work of Shi and Zhu [205], who independently used the dual characterization of the approximate degree in a rather different way.

4.2 Pattern matrices and their spectrum

In this section we study the first component of our proof, a certain family of real matrices that we introduce. Our goal here is to explicitly calculate their singular values. As we shall see later, this provides a convenient means to generate hard communication problems.

Let t and n be positive integers, where $t < n$ and $t \mid n$. Partition $[n]$ into t contiguous blocks, each with n/t elements:

$$[n] = \left\{1, 2, \ldots, \frac{n}{t}\right\} \cup \left\{\frac{n}{t} + 1, \ldots, \frac{2n}{t}\right\} \cup \cdots \cup \left\{\frac{(t-1)n}{t} + 1, \ldots, n\right\}.$$

Let $\mathscr{V}(n,t)$ denote the family of subsets $V \subseteq [n]$ that have exactly one element in each of these blocks (in particular, $|V| = t$). Clearly, $|\mathscr{V}(n,t)| = (n/t)^t$. Recall that for a bit string $x \in \{0,1\}^n$ and a set $V \in \mathscr{V}(n,t)$, the projection of x onto V is given by

$$x|_V = (x_{i_1}, x_{i_2}, \ldots, x_{i_t}) \in \{0,1\}^t,$$

where $i_1 < i_2 < \cdots < i_t$ are the elements of V. We are ready for a formal definition of our matrix family.

DEFINITION 4.1 (Sherstov [203]). For $\phi\colon \{0,1\}^t \to \mathbb{R}$, the (n,t,ϕ)-*pattern matrix* is the real matrix A given by

$$A = \Big[\phi(x|_V \oplus w)\Big]_{x \in \{0,1\}^n,\, (V,w) \in \mathscr{V}(n,t) \times \{0,1\}^t}.$$

In words, A is the matrix of size 2^n by $(n/t)^t 2^t$ whose rows are indexed by strings $x \in \{0,1\}^n$, whose columns are indexed by pairs $(V, w) \in \mathscr{V}(n,t) \times \{0,1\}^t$, and whose entries are given by $A_{x,(V,w)} = \phi(x|_V \oplus w)$.

The logic behind the term "pattern matrix" is as follows: a mosaic arises from repetitions of a pattern in the same way that A arises from applications of ϕ to various subsets of the variables. Our approach to analyzing the singular values of a pattern matrix A will be to represent it as the sum of simpler matrices and analyze them instead. For this to work, we should be able to reconstruct the singular values of A from those of the simpler matrices. Just when this can be done is the subject of the following lemma.

LEMMA 4.2 (Sherstov [203]). *Let A, B be real matrices with $AB^\mathsf{T} = 0$ and $A^\mathsf{T}B = 0$. Then the nonzero singular values of $A + B$, counting multiplicities, are $\sigma_1(A), \ldots, \sigma_{\mathrm{rk}\,A}(A), \sigma_1(B), \ldots, \sigma_{\mathrm{rk}\,B}(B)$.*

PROOF. The claim is trivial when $A = 0$ or $B = 0$, so assume otherwise. Since the singular values of $A + B$ are precisely the square roots of the eigenvalues of $(A+B)(A+B)^\mathsf{T}$, it suffices to compute the spectrum of the latter matrix. Now,

$$\begin{aligned}(A+B)(A+B)^\mathsf{T} &= AA^\mathsf{T} + BB^\mathsf{T} + \underbrace{AB^\mathsf{T}}_{=0} + \underbrace{BA^\mathsf{T}}_{=0} \\ &= AA^\mathsf{T} + BB^\mathsf{T}. \end{aligned} \tag{4.3}$$

Fix spectral decompositions

$$AA^\mathsf{T} = \sum_{i=1}^{\mathrm{rk}\,A} \sigma_i(A)^2 u_i u_i^\mathsf{T}, \qquad BB^\mathsf{T} = \sum_{j=1}^{\mathrm{rk}\,B} \sigma_j(B)^2 v_j v_j^\mathsf{T}.$$

Then

$$\begin{aligned}\sum_{i=1}^{\operatorname{rk} A}\sum_{j=1}^{\operatorname{rk} B}\sigma_i(A)^2\sigma_j(B)^2\langle u_i, v_j\rangle^2 &= \left\langle \sum_{i=1}^{\operatorname{rk} A}\sigma_i(A)^2 u_i u_i^\mathsf{T}, \sum_{j=1}^{\operatorname{rk} B}\sigma_j(B)^2 v_j v_j^\mathsf{T}\right\rangle \\ &= \langle AA^\mathsf{T}, BB^\mathsf{T}\rangle \\ &= \operatorname{tr}(AA^\mathsf{T}BB^\mathsf{T}) \\ &= \operatorname{tr}(A\cdot 0\cdot B^\mathsf{T}) \\ &= 0. \end{aligned} \tag{4.4}$$

Since $\sigma_i(A)\,\sigma_j(B) > 0$ for all i, j, it follows from (4.4) that $\langle u_i, v_j\rangle = 0$ for all i, j. Put differently, the vectors $u_1, \ldots, u_{\operatorname{rk} A}, v_1, \ldots, v_{\operatorname{rk} B}$ form an orthonormal set. Recalling (4.3), we conclude that the spectral decomposition of $(A+B)(A+B)^\mathsf{T}$ is

$$\sum_{i=1}^{\operatorname{rk} A}\sigma_i(A)^2 u_i u_i^\mathsf{T} + \sum_{j=1}^{\operatorname{rk} B}\sigma_j(B)^2 v_j v_j^\mathsf{T},$$

and thus the nonzero eigenvalues of $(A+B)(A+B)^\mathsf{T}$ are as claimed. □

We are ready for the main result of this section.

THEOREM 4.3 (Sherstov [203]). *Let $\phi\colon\{0,1\}^t \to \mathbb{R}$ be given. Let A be the (n, t, ϕ)-pattern matrix. Then the nonzero singular values of A, counting multiplicities, are:*

$$\bigcup_{S:\hat{\phi}(S)\neq 0}\left\{\sqrt{2^{n+t}\binom{n}{t}^t}\cdot|\hat{\phi}(S)|\left(\frac{t}{n}\right)^{|S|/2}, \qquad \textit{repeated } \binom{n}{t}^{|S|} \textit{ times}\right\}.$$

In particular,

$$\|A\| \;=\; \sqrt{2^{n+t}\left(\frac{n}{t}\right)^{t}}\;\max_{S\subseteq[t]}\left\{|\hat{\phi}(S)|\left(\frac{t}{n}\right)^{|S|/2}\right\}.$$

PROOF. For each $S \subseteq [t]$, let A_S be the (n, t, χ_S)-pattern matrix. Thus,

$$A = \sum_{S\subseteq[t]} \hat{\phi}(S) A_S. \tag{4.5}$$

Fix arbitrary $S, T \subseteq [t]$ with $S \neq T$. Then

$$\begin{aligned} A_S A_T^{\mathsf{T}} &= \left[\sum_{V\in\mathscr{V}(n,t)}\;\sum_{w\in\{0,1\}^t} \chi_S(x|_V \;\oplus\; w)\;\chi_T(y|_V \;\oplus\; w)\right]_{x,y} \\ &= \left[\sum_{V\in\mathscr{V}(n,t)} \chi_S(x|_V)\;\chi_T(y|_V)\underbrace{\sum_{w\in\{0,1\}^t}\chi_S(w)\;\chi_T(w)}_{=0}\right]_{x,y} \\ &= 0. \end{aligned} \tag{4.6}$$

Similarly,

$$A_S^{\mathsf{T}} A_T = \left[\chi_S(w)\;\chi_T(w')\underbrace{\sum_{x\in\{0,1\}^n}\chi_S(x|_V)\;\chi_T(x|_{V'})}_{=0}\right]_{(V,w),(V',w')} = 0. \tag{4.7}$$

By (4.5)–(4.7) and Lemma 4.2, the nonzero singular values of A are the union of the nonzero singular values of all $\hat{\phi}(S)A_S$, counting multiplicities. Therefore, the proof will be complete once we show that the only nonzero singular value of $A_S^{\mathsf{T}} A_S$

is $2^{n+t}(n/t)^{t-|S|}$, with multiplicity $(n/t)^{|S|}$. It is convenient to write this matrix as the Kronecker product

$$A_S^{\mathsf{T}} A_S \;=\; [\chi_S(w)\chi_S(w')]_{w,w'} \;\otimes\; \left[\sum_{x\in\{0,1\}^n} \chi_S(x|_V)\;\chi_S(x|_{V'})\right]_{V,V'}.$$

The first matrix in this factorization has rank 1 and entries ± 1, which means that its only nonzero singular value is 2^t with multiplicity 1. The other matrix, call it M, is permutation-similar to

$$2^n \begin{bmatrix} J & & & \\ & J & & \\ & & \ddots & \\ & & & J \end{bmatrix},$$

where J is the all-ones square matrix of order $(n/t)^{t-|S|}$. This means that the only nonzero singular value of M is $2^n(n/t)^{t-|S|}$ with multiplicity $(n/t)^{|S|}$. It follows from elementary properties of the Kronecker product that the spectrum of $A_S^{\mathsf{T}} A_S$ is as claimed. □

4.3 Duals of approximation and sign-representation

We now develop the second ingredient of our proof, the dual characterizations of the uniform approximation and sign-representation of Boolean functions by real polynomials. As a starting point, we recall a classical result from approximation theory due to Ioffe and Tikhomirov [97] on the duality of norms. A more recent treatment is available in the textbook of DeVore and Lorentz [67], p. 61, Thm. 1.3. We provide a short and elementary proof of this result in Euclidean space, which will suffice for our purposes. We let $\mathbb{R}^X$ stand for the linear space of real functions on the set X.

THEOREM 4.4 (Ioffe and Tikhomirov [97]). *Let X be a finite set. Fix $\Phi \subseteq \mathbb{R}^X$ and a function $f : X \to \mathbb{R}$. Then*

$$\min_{\phi \in \operatorname{span}(\Phi)} \|f - \phi\|_\infty = \max_{\psi} \left\{ \sum_{x \in X} f(x)\psi(x) \right\}, \tag{4.8}$$

where the maximum is over all functions $\psi : X \to \mathbb{R}$ such that

$$\sum_{x \in X} |\psi(x)| \leqslant 1$$

and, for each $\phi \in \Phi$,

$$\sum_{x \in X} \phi(x)\psi(x) = 0.$$

PROOF. The theorem holds trivially when $\operatorname{span}(\Phi) = \{0\}$. Otherwise, let $\phi_1, \ldots, \phi_k$ be a basis for $\operatorname{span}(\Phi)$. Observe that the left member of (4.8) is the optimum of the following linear program in the variables $\varepsilon, \alpha_1, \ldots, \alpha_k$:

minimize:	ε	
subject to:	$\left\lvert f(x) - \sum_{i=1}^{k} \alpha_i \phi_i(x) \right\rvert \leqslant \varepsilon$	for each $x \in X$,
	$\alpha_i \in \mathbb{R}$	for each i,
	$\varepsilon \geqslant 0.$	

Standard manipulations reveal the dual:

maximize:	$\sum_{x \in X} \psi_x f(x)$	
subject to:	$\sum_{x \in X} \lvert\psi_x\rvert \leqslant 1,$	
	$\sum_{x \in X} \psi_x \phi_i(x) = 0$	for each i,
	$\psi_x \in \mathbb{R}$	for each $x \in X$.

Both programs are clearly feasible and thus have the same finite optimum. We have already observed that the optimum of first program is the left-hand side of (4.8). Since $\phi_1, \ldots, \phi_k$ form a basis for $\operatorname{span}(\Phi)$, the optimum of the second program is by definition the right-hand side of (4.8). □

As a corollary to Theorem 4.4, we obtain a dual characterization of the approximate degree.

THEOREM 4.5. *Fix $\varepsilon \geqslant 0$. Let $f : \{0,1\}^n \to \mathbb{R}$ be given, $d = \deg_\varepsilon(f) \geqslant 1$. Then there is a function $\psi : \{0,1\}^n \to \mathbb{R}$ such that*

$$\begin{aligned} &\hat{\psi}(S) = 0 && (|S| < d), \\ &\sum_{x \in \{0,1\}^n} |\psi(x)| = 1, \\ &\sum_{x \in \{0,1\}^n} \psi(x) f(x) > \varepsilon. \end{aligned}$$

PROOF. Set $X = \{0,1\}^n$ and $\Phi = \{\chi_S : |S| < d\} \subset \mathbb{R}^X$. Since $\deg_\varepsilon(f) = d$, we conclude that

$$\min_{\phi \in \operatorname{span}(\Phi)} \|f - \phi\|_\infty > \varepsilon.$$

In view of Theorem 4.4, we can take ψ to be any function for which the maximum is achieved in (4.8). □

We now state the dual characterization of the threshold degree, which is better known as Gordan's Transposition Theorem.

THEOREM 4.6 (Gordan [86]). *Let* $f : \{0,1\}^n \to \{-1,+1\}$ *be given. Then* $d < \deg_{\pm}(f)$ *if and only if there is a distribution* μ *over* $\{0,1\}^n$ *with*

$$\mathop{\mathbf{E}}_{x\sim\mu}[f(x)\chi_S(x)] = 0 \qquad\qquad (|S| \leqslant d).$$

Theorem 4.6 has a short proof using linear programming duality, as explained in [194, §7.8] and [202, §2.2]. We close this section with one final dual characterization, corresponding to sign-representation by integer polynomials.

THEOREM 4.7 (Freund [74], Hajnal et al. [88]). *Fix a function* $f : \{0,1\}^n \to \{-1,+1\}$ *and an integer* $d \geqslant \deg_{\pm}(f)$. *Then for every distribution* μ *on* $\{0,1\}^n$,

$$\max_{|S|\leqslant d} \left| \mathop{\mathbf{E}}_{x\sim\mu}[f(x)\chi_S(x)] \right| \geqslant \frac{1}{W(f,d)}. \tag{4.9}$$

Furthermore, there exists a distribution μ *such that*

$$\max_{|S|\leqslant d} \left| \mathop{\mathbf{E}}_{x\sim\mu}[f(x)\chi_S(x)] \right| \leqslant \left(\frac{2n}{W(f,d)}\right)^{1/2}. \tag{4.10}$$

Inequalities (4.9) and (4.10) are originally due to Hajnal et al. [88] and Freund [74], respectively. An integrated and simplified treatment of both results is available in the work of Goldmann et al. [82], Lem. 4 and Thm. 10.

4.4 Lower bounds for bounded-error communication

The previous two sections examined relevant dual representations and the spectrum of pattern matrices. Having studied these notions in their pure and basic form, we now apply our findings to communication complexity. Specifically, we establish the *pattern matrix method* for bounded-error communication complexity, which gives strong lower bounds for every pattern matrix generated by a Boolean function with high approximate degree.

THEOREM 4.8 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, for a given function $f\colon \{0,1\}^t \to \{-1,+1\}$. Then for every $\varepsilon \in [0,1)$ and every $\delta < \varepsilon/2$,*

$$R_\delta(F) \geqslant \frac{1}{2}\deg_\varepsilon(f)\log\left(\frac{n}{t}\right) - \log\left(\frac{1}{\varepsilon - 2\delta}\right). \tag{4.11}$$

In particular,

$$R_{1/7}(F) > \frac{1}{2}\deg_{1/3}(f)\log\left(\frac{n}{t}\right) - 5. \tag{4.12}$$

PROOF. Since (4.11) immediately implies (4.12), we will focus on the former in the remainder of the proof. Let $d = \deg_\varepsilon(f) \geqslant 1$. By Theorem 4.5, there is a function $\psi\colon \{0,1\}^t \to \mathbb{R}$ such that:

$$\hat{\psi}(S) = 0 \qquad (|S| < d), \tag{4.13}$$

$$\sum_{z\in\{0,1\}^t} |\psi(z)| = 1, \tag{4.14}$$

$$\sum_{z\in\{0,1\}^t} \psi(z) f(z) > \varepsilon. \tag{4.15}$$

Let Ψ be the $(n, t, 2^{-n}(n/t)^{-t}\psi)$-pattern matrix. Then (4.14) and (4.15) show that

$$\|\Psi\|_1 = 1, \qquad \langle F, \Psi\rangle > \varepsilon. \tag{4.16}$$

Our last task is to calculate $\|\Psi\|$. By (4.14) and Proposition 2.1,

$$\max_{S \subseteq [t]} |\hat{\psi}(S)| \leqslant 2^{-t}. \tag{4.17}$$

Theorem 4.3 yields, in view of (4.13) and (4.17):

$$\|\Psi\| \leqslant \left(\frac{t}{n}\right)^{d/2} \left(2^{n+t} \left(\frac{n}{t}\right)^{t}\right)^{-1/2}. \tag{4.18}$$

Now (4.11) follows from (4.16), (4.18), and the generalized discrepancy method (Theorem 3.7). □

Theorem 4.8 gives lower bounds not only for bounded-error communication but also for communication protocols with error probability $\frac{1}{2} - o(1)$. For example, if a function $f : \{0,1\}^t \to \{-1,+1\}$ requires a polynomial of degree d for approximation within $1 - o(1)$, equation (4.11) gives a lower bound for small-bias communication. We will complement and refine that estimate in the next section, which is dedicated to small-bias communication.

Pattern matrices are of interest because they occur as submatrices in natural communication problems. We will illustrate this point throughout this thesis. At present, we record a corollary of Theorem 4.8 in terms of function composition.

COROLLARY 4.9 (Sherstov [203]). *Let $f : \{0,1\}^t \to \{-1,+1\}$ be given. Define $F : \{0,1\}^{4t} \times \{0,1\}^{4t} \to \{-1,+1\}$ by*

$$F(x,y) = f(\ldots, (x_{i,1}y_{i,1} \vee x_{i,2}y_{i,2} \vee x_{i,3}y_{i,3} \vee x_{i,4}y_{i,4}), \ldots).$$

Then

$$R_{1/7}(F) > \frac{1}{2} \deg_{1/3}(f) - 5.$$

PROOF. The $(2t, t, f)$-pattern matrix is as a submatrix of $[F(x, y)]_{x,y\in\{0,1\}^{4t}}$. □

Finally, we show that the lower bound in Theorem 4.8 for bounded-error communication complexity is tight up to a polynomial factor, even for deterministic protocols. The proof follows a well-known argument in the literature [51, 28] and was pointed out to us by R. de Wolf [220].

PROPOSITION 4.10 (R. de Wolf [220], personal communication). *Let F be the (n, t, f)-pattern matrix, where $f : \{0,1\}^t \to \{-1,+1\}$ is given. Then*

$$D(F) \leqslant O\Big(\mathrm{dt}(f) \log \frac{n}{t}\Big) \leqslant O\Big(\deg_{1/3}(f)^6 \log \frac{n}{t}\Big). \tag{4.19}$$

In particular, (4.12) is tight up to a polynomial factor.

PROOF. The second inequality in (4.19) follows by Theorem 2.9. Therefore, it suffices to prove an upper bound of $O(d \log(n/t))$ on the deterministic communication complexity of F, where $d = \mathrm{dt}(f)$ is the decision tree complexity of f. The needed deterministic protocol is well-known. Fix a depth-d decision tree for f. Let $(x, (V, w))$ be a given input. Alice and Bob start at the root of the decision tree, labeled by some variable $i \in \{1, \ldots, t\}$. By exchanging $\lceil \log(n/t) \rceil + 2$ bits, Alice and Bob determine $(x|_V)_i \oplus w_i \in \{0,1\}$ and take the corresponding branch of the tree. The process repeats until a leaf is reached, at which point both parties learn $f(x|_V \oplus w)$. □

4.5 Lower bounds for small-bias communication

As we have already mentioned, Theorem 4.8 of the previous section can be used to obtain lower bounds not only for bounded-error communication but also small-bias communication. In the latter case, one first needs to show that the base function $f : \{0,1\}^t \to \{-1,+1\}$ cannot be approximated pointwise within $1 - o(1)$ by a real polynomial of a given degree d. In this section, we derive a different lower bound for small-bias communication, this time using the assumption that the threshold weight $W(f, d)$ is high. We will see that this new lower bound is nearly optimal and closely related to the lower bound in Theorem 4.8.

THEOREM 4.11 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, for a given function $f\colon \{0,1\}^t \to \{-1,+1\}$. Then for every integer $d \geqslant 1$ and real $\gamma \in (0,1)$,*

$$R_{1/2-\gamma/2}(F) \geqslant \frac{1}{2} \min\left\{ d \log \frac{n}{t},\ \log \frac{W(f, d-1)}{2t} \right\} - \log \frac{1}{\gamma}. \tag{4.20}$$

In particular,

$$R_{1/2-\gamma/2}(F) \geqslant \frac{1}{2} \deg_{\pm}(f) \log\left(\frac{n}{t}\right) - \log \frac{1}{\gamma}. \tag{4.21}$$

PROOF. Letting $d = \deg_{\pm}(f)$ in (4.20) yields (4.21), since $W(f, d-1) = \infty$ in that case. In the remainder of the proof, we focus on (4.20) alone.

We claim that there exists a distribution μ on $\{0,1\}^t$ such that

$$\max_{|S|<d} \left| \mathop{\mathbf{E}}_{z \sim \mu} [f(z)\chi_S(z)] \right| \leqslant \left(\frac{2t}{W(f, d-1)} \right)^{1/2}. \tag{4.22}$$

For $d \leqslant \deg_{\pm}(f)$, the claim holds by Theorem 4.6 since $W(f, d-1) = \infty$ in that case. For $d > \deg_{\pm}(f)$, the claim holds by Theorem 4.7. Letting $\psi\colon \{0,1\}^t \to \mathbb{R}$ be given by $\psi(z) = f(z)\mu(z)$, we have from (4.22) that

$$|\hat{\psi}(S)| \leqslant 2^{-t} \left(\frac{2t}{W(f, d-1)} \right)^{1/2} \qquad (|S| < d), \tag{4.23}$$

$$\sum_{z \in \{0,1\}^t} |\psi(z)| = 1, \tag{4.24}$$

$$\sum_{z \in \{0,1\}^t} \psi(z) f(z) = 1. \tag{4.25}$$

Let Ψ be the $(n, t, 2^{-n}(n/t)^{-t}\psi)$-pattern matrix. Then (4.24) and (4.25) show that

$$\|\Psi\|_1 = 1, \qquad \langle F, \Psi \rangle = 1. \tag{4.26}$$

It remains to calculate $\|\Psi\|$. By (4.24) and Proposition 2.1,

$$\max_{S \subseteq [t]} |\hat{\psi}(S)| \leqslant 2^{-t}. \tag{4.27}$$

Theorem 4.3 yields, in view of (4.23) and (4.27):

$$\|\Psi\| \leqslant \max\left\{ \left(\frac{t}{n}\right)^{d/2}, \left(\frac{2t}{W(f, d-1)}\right)^{1/2} \right\} \left(2^{n+t}\left(\frac{n}{t}\right)^{t}\right)^{-1/2}. \tag{4.28}$$

Now (4.20) follows from (4.26), (4.28), and the generalized discrepancy method (Theorem 3.7). □

Recall from Theorem 2.4 that the quantities $E(f, d)$ and $W(f, d)$ are related for all f and d. In particular, the lower bounds for small-bias communication in Theorems 4.8 and 4.11 are quite close, and either one can be approximately deduced from the other. In deriving both results from scratch, as we did, our motivation was to obtain the tightest bounds and to illustrate the pattern matrix method in different contexts. We will now see that the lower bound in Theorem 4.11 is close to optimal.

THEOREM 4.12 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, for a given function $f : \{0,1\}^t \to \{-1,+1\}$. Then for every integer $d \geqslant \deg_{\pm}(f)$,*

$$R_{1/2-\gamma/2}(F) \leqslant d \log\left(\frac{n}{t}\right) + 3,$$

where $\gamma = 1/W(f, d)$.

PROOF. The communication protocol that we will describe is standard and has been used in one form or another in several works, e.g., [167, 82, 198, 202]. Put

$W = W(f, d)$ and fix a representation

$$f(z) \equiv \operatorname{sgn}\left(\sum_{S \subseteq [t],\, |S| \leqslant d} \lambda_S \chi_S(z)\right),$$

where the integers λ_S satisfy $\sum |\lambda_S| = W$. On input $(x, (V, w))$, the protocol proceeds as follows. Let $i_1 < i_2 < \cdots < i_t$ be the elements of V. Alice and Bob use their shared randomness to pick a set $S \subseteq [t]$ with $|S| \leqslant d$, according to the probability distribution $|\lambda_S|/W$. Next, Bob sends Alice the indices $\{i_j : j \in S\}$ as well as the bit $\chi_S(w)$. With this information, Alice computes the product $\operatorname{sgn}(\lambda_S)\chi_S(x|_V)\chi_S(w) = \operatorname{sgn}(\lambda_S)\chi_S(x|_V \oplus w)$ and announces the result as the output of the protocol.

Assuming an optimal encoding of the messages, the communication cost of this protocol is bounded by

$$\left\lceil \log \binom{n}{t}^d \right\rceil + 2 \leqslant d \log \left(\frac{n}{t}\right) + 3,$$

as desired. On each input x, V, w, the output of the protocol is a random variable $P(x, V, w) \in \{-1, +1\}$ that obeys

$$\begin{aligned} f(x|_V \oplus w)\, \mathbf{E}[P(x, V, w)] &= f(x|_V \oplus w) \sum_{|S| \leqslant d} \frac{|\lambda_S|}{W} \operatorname{sgn}(\lambda_S)\chi_S(x|_V \oplus w) \\ &= \frac{1}{W} \left| \sum_{|S| \leqslant d} \lambda_S \chi_S(x|_V \oplus w) \right| \\ &\geqslant \frac{1}{W}, \end{aligned}$$

which means that the protocol produces the correct answer with probability no smaller than $\frac{1}{2} + \frac{1}{2W}$. □

We close this section by characterizing the discrepancy of pattern matrices in terms of threshold weight.

THEOREM 4.13 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, for a given function $f : \{0,1\}^t \to \{-1,+1\}$. Then for every integer $d \geqslant 0$,*

$$\operatorname{disc}(F) \geqslant \frac{1}{8W(f,d)} \left(\frac{t}{n}\right)^d \tag{4.29}$$

and

$$\operatorname{disc}(F)^2 \leqslant \max\left\{\frac{2t}{W(f,d-1)}, \left(\frac{t}{n}\right)^d\right\}. \tag{4.30}$$

In particular,

$$\operatorname{disc}(F) \leqslant \left(\frac{t}{n}\right)^{\deg_\pm(f)/2}. \tag{4.31}$$

PROOF. The lower bound (4.29) follows from Theorem 4.12 and Proposition 3.2. For the upper bound (4.30), construct the matrix Ψ as in the proof of Theorem 4.11. Then (4.26) shows that $\Psi = F \circ P$ for a probability matrix P. As a result, (4.30) follows from (4.28) and Proposition 3.4. Finally, (4.31) follows by taking $d = \deg_\pm(f)$ in (4.30), since $W(f, d-1) = \infty$ in that case. □

Threshold weight is typically easier to analyze than the approximate degree. For completeness, however, we will now supplement Theorem 4.13 with an alternate bound on the discrepancy of a pattern matrix in terms of the approximate degree.

THEOREM 4.14 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, for a given function $f : \{0,1\}^t \to \{-1,+1\}$. Then for every $\gamma > 0$,*

$$\operatorname{disc}(F) \leqslant \gamma + \left(\frac{t}{n}\right)^{\deg_{1-\gamma}(f)/2}.$$

PROOF. Let $d = \deg_{1-\gamma}(f) \geqslant 1$. Define $\varepsilon = 1 - \gamma$ and construct the matrix Ψ as in the proof of Theorem 4.8. Then (4.16) shows that $\Psi = H \circ P$ for a sign matrix H and a probability matrix P. By (4.18) and Proposition 3.4,

$$\operatorname{disc}_P(H) \leqslant \left(\frac{t}{n}\right)^{d/2}. \tag{4.32}$$

Moreover,

$$\begin{aligned} \operatorname{disc}_P(F) &\leqslant \operatorname{disc}_P(H) + \|(F - H) \circ P\|_1 \\ &= \operatorname{disc}_P(H) + 1 - \langle F, H \circ P \rangle \\ &\leqslant \operatorname{disc}_P(H) + \gamma, \end{aligned} \tag{4.33}$$

where the last step follows because $\langle F, \Psi \rangle > \varepsilon = 1 - \gamma$ by (4.16). The proof is complete in view of (4.32) and (4.33). □

4.6 Separation of the polynomial hierarchy from $\mathsf{PP}^{\mathsf{cc}}$

As an application of our results on small-bias communication, we will now examine the discrepancy of AC^0, the class of polynomial-size constant-depth circuits with AND, OR, NOT gates. We will prove the first exponentially small upper bound on the discrepancy of a function in AC^0, thereby separating the polynomial hierarchy in communication complexity from $\mathsf{PP}^{\mathsf{cc}}$.

Consider the function $\mathrm{MP}_m\colon \{0,1\}^{4m^3} \to \{-1,+1\}$ given by

$$\mathrm{MP}_m(x) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{4m^2} x_{ij}.$$

This function was originally defined and studied by Minsky and Papert [153] in their seminal monograph on perceptrons. We have:

THEOREM 4.15 (Sherstov [202, 203]). *Let* $f(x,y) = \mathrm{MP}_m(x \wedge y)$. *Then*

$$\mathrm{disc}(f) = \exp\{-\Omega(m)\}.$$

PROOF. Put $d = \lfloor m/2 \rfloor$. A well-known result of Minsky and Papert [153] states that $\deg_{\pm}(\mathrm{MP}_d) \geqslant d$. Since the $(8d^3, 4d^3, \mathrm{MP}_d)$-pattern matrix is a submatrix of $[f(x,y)]_{x,y}$, the proof is complete in view of equation (4.31) of Theorem 4.13. □

Theorem 4.15 gives a function in AC^0 with exponentially small discrepancy. Independently of the author, another such function was exhibited by Buhrman et al. [53] with very different techniques. The key building block of that construction is the ODD-MAX-BIT function $\mathrm{OMB}_n\colon \{0,1\}^n \to \{-1,+1\}$ due to Beigel [32], which is given by

$$\mathrm{OMB}_n(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n} (-2)^i x_i\right). \tag{4.34}$$

Since

$$\mathrm{OMB}_n(x) = \bigvee_{i \text{ odd}} (x_i \wedge \overline{x}_{i+1} \wedge \cdots \wedge \overline{x}_n),$$

it is clear that OMB_n belongs to the class AC^0. Buhrman et al. [53, §3.2] proved the following result.

THEOREM 4.16 (Buhrman et al. [53]). *Let* $f(x, y) = \mathrm{OMB}_n(x \wedge y)$. *Then*

$$\mathrm{disc}(f) = \exp\{-\Omega(n^{1/3})\}.$$

Using the results of this chapter, we are able to give a simple alternate proof of this theorem.

PROOF (Sherstov [203]). Put $m = \lfloor n/4 \rfloor$. A well-known result due to Beigel [32] shows that $W(\mathrm{OMB}_m, cm^{1/3}) \geqslant \exp(cm^{1/3})$ for some absolute constant $c > 0$. Since the $(2m, m, \mathrm{OMB}_m)$-pattern matrix is a submatrix of $[f(x, y)]_{x,y}$, the proof is complete by Theorem 4.13. □

The work in this section establishes new separations for the communication complexity classes Σ_2^{cc}, Π_2^{cc}, and $\mathsf{PP}^{\mathsf{cc}}$, reviewed in detail in Section 3.4. If $f_n \colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$, $n = 1, 2, 3, 4, \ldots$, is an AC^0 circuit family of depth k with an OR gate at the top (resp., AND gate), then by definition $\{f_n\} \in \Sigma_{k-1}^{\mathsf{cc}}$ (resp., $\{f_n\} \in \Pi_{k-1}^{\mathsf{cc}}$). In particular, the depth-3 circuit family $\{f_n\}$ in Theorem 4.15 is in Π_2^{cc}, whereas $\{\neg f_n\}$ is in Σ_2^{cc}. Since the discrepancy of a function remains unchanged under negation, Theorem 4.15 has the following corollary.

COROLLARY 4.17. $\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$, $\Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}$.

By the same argument, Corollary 4.17 is immediate from Theorem 4.16. This corollary is best possible in that $\mathsf{PP}^{\mathsf{cc}}$ trivially contains the zeroth and first levels of the polynomial hierarchy:

PROPOSITION 4.18. *Let* $f \colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ *be a function such that* $f(x, y) = \bigvee_{i=1}^{s} g_i(x, y)$, *where* $g_1, \ldots, g_s \in \Pi_0^{\mathsf{cc}}$ *are each a rectangle. Then*

$$\mathrm{disc}(f) \geqslant \Omega\left(\frac{1}{s}\right). \tag{4.35}$$

In particular,

$$\Sigma_0^{\mathsf{cc}}, \Sigma_1^{\mathsf{cc}}, \Pi_0^{\mathsf{cc}}, \Pi_1^{\mathsf{cc}} \subseteq \mathsf{PP}^{\mathsf{cc}}.$$

PROOF. The containments $\Sigma_0^{\mathsf{cc}}, \Sigma_1^{\mathsf{cc}} \subseteq \mathsf{PP}^{\mathsf{cc}}$ are immediate from (4.35). They in turn imply the containments $\Pi_0^{\mathsf{cc}}, \Pi_1^{\mathsf{cc}} \subseteq \mathsf{PP}^{\mathsf{cc}}$ since the discrepancy of a function is invariant under negation. To prove (4.35), observe that

$$f = \text{MAJORITY}(g_1, \ldots, g_s, g_{s+1}, \ldots, g_{2s-1}),$$

where we define $g_{s+1} \equiv g_{s+2} \equiv \cdots \equiv g_{2s-1} \equiv -1$ (identically true). Consider the randomized protocol in which the parties pick $i \in \{1, 2, \ldots, 2s-1\}$ uniformly at random, evaluate g_i using constant communication, and output the result. This protocol evaluates f correctly with probability at least $\frac{1}{2} + \Omega\left(\frac{1}{s}\right)$. Thus,

$$R_{1/2-\Omega(1/s)}(f) = O(1).$$

In view of Proposition 3.2, this completes the proof. □

Proposition 4.18 also shows that the discrepancy of every AC^0 circuit of depth 2 is at least $n^{-O(1)}$. In particular, the exponentially small upper bounds on discrepancy, obtained in Theorems 4.15 and 4.16 for circuits of depth 3, are optimal with respect to depth.

4.7 Separation of AC^0 from majority circuits

A natural and important computational model is that of a polynomial-size circuit of majority gates. This model has been extensively studied for the past two decades [153, 207, 88, 82, 159, 208, 209, 196]. Research has shown that majority circuits of depth 2 and 3 already possess surprising computational power. Indeed, it is a longstanding open problem [132] to exhibit a Boolean function that *cannot* be computed by a depth-3 majority circuit of polynomial size. In particular, the

arithmetic operations of powering, multiplication, and division on n-bit integer arguments can all be computed by depth-3 majority circuits of polynomial size [209]. An even more striking example is the addition of n n-bit integers, which is computable by a depth-2 majority circuit of polynomial size [209]. Depth-2 majority circuits of polynomial size can also compute every symmetric function (such as PARITY) and every DNF and CNF formula of polynomial size.

This goal of this section is to relate the computational power of majority circuits to that of AC^0. A well-known result due to Allender [11] states that every function in AC^0 can be computed by a depth-3 majority circuit of quasipolynomial size. It was an open problem to determine whether Allender's simulation is optimal. Specifically, Krause and Pudlák [132, §6] asked whether every function in AC^0 can be computed by a depth-2 majority circuit of quasipolynomial size. We solve this problem:

THEOREM 4.19 (Sherstov [202, 203]). *The function*

$$f(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij})$$

cannot be computed by a majority vote of fewer than $\exp\{\Omega(m)\}$ *linear threshold gates.*

In other words, Allender's simulation is optimal in a strong sense. The lower bound in Theorem 4.19 is an exponential improvement on previous work. The best previous lower bound [88, 159] was quasipolynomial.

To prove Theorem 4.19, we first recall a well-known relationship between the discrepancy of a function and its circuit complexity.

THEOREM 4.20 (Nisan [159]). *Let* $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ *be a given function. Then* f *cannot be computed by a majority vote of fewer than*

$$\frac{1}{2n \operatorname{disc}(f)^{\alpha}}$$

linear threshold gates, for some absolute constant $\alpha > 0$.

PROOF. Proposition 3.2 states that for every $\gamma > 0$,

$$R_{1/2-\gamma/2}(f) \geqslant \log\left(\frac{\gamma}{\operatorname{disc}(f)}\right). \tag{4.36}$$

On the other hand, fix a representation $f \equiv \mathrm{MAJ}(h_1, h_2, \ldots, h_s)$, where each h_i is a linear threshold function. Nisan [159] proved that every linear threshold gate $h\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ satisfies $R_\varepsilon(h) = O(\log n + \log \frac{1}{\varepsilon})$. This gives a nontrivial protocol for f, whereby the parties randomly pick $i \in \{1, 2, \ldots, s\}$, evaluate h_i correctly with probability $1 - \frac{1}{4s}$ using $O(\log n + \log s)$ communication, and output the result. Since this protocol computes $f(x, y)$ correctly with probability at least $\left(\frac{1}{2} + \frac{1}{2s}\right) - \frac{1}{4s} = \frac{1}{2} + \frac{1}{4s}$ on every input, we have

$$R_{1/2-1/4s}(f) = O(\log n + \log s). \tag{4.37}$$

The theorem follows from (4.36) and (4.37). □

We can now settle the main result of this section.

PROOF OF THEOREM 4.19. Immediate from Theorems 4.15 and 4.20. □

An analogous argument reveals another AC^0 function that requires a depth-2 majority circuit of exponential size:

THEOREM 4.21. *The function $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ given by*

$$f(x, y) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n} (-2)^i x_i y_i\right)$$

cannot be computed by a majority vote of fewer than $\exp\{\Omega(n^{1/3})\}$ linear threshold gates.

PROOF. Immediate from Theorems 4.16 and 4.20. □

4.8 A combinatorial analysis of pattern matrices

In Sections 4.2–4.5, we used matrix-analytic methods to determine the discrepancy and randomized communication complexity of pattern matrices. We revisit those derivations here using simpler, combinatorial methods. The resulting bounds will be weaker than before. We will see in Chapter 9, however, that this combinatorial approach readily applies to the multiparty model of communication, where matrix-analytic techniques do not work.

We will need the following technical observation.

PROPOSITION 4.22. *Let $\nu(x)$ be a probability distribution on $\{0,1\}^r$. Fix $i_1, \ldots, i_r \in \{1,2,\ldots,r\}$. Then*

$$\sum_{x \in \{0,1\}^r} \nu(x_{i_1}, \ldots, x_{i_r}) \leqslant 2^{r-|\{i_1,\ldots,i_r\}|},$$

where $|\{i_1,\ldots,i_r\}|$ is the number of distinct integers among $i_1,\ldots,i_r$.

PROOF. Immediate from the fact that $\sum_{x \in \{0,1\}^r} \nu(x_1,\ldots,x_r) \leqslant 1$. □

The main technical tool of this section is the following theorem.

THEOREM 4.23 (Sherstov [202, 195]). *Fix a function $h\colon \{0,1\}^t \to \{-1,+1\}$ and a probability distribution μ on $\{0,1\}^t$. Let n be a given integer, $t \mid n$. Define matrices*

$$H = [h(x|_V)]_{x,V}, \qquad P = 2^{-n+t}\binom{n}{t}^{-t}[\mu(x|_V)]_{x,V},$$

where the rows and columns are indexed as follows: $x \in \{0,1\}^n$, $V \in \mathscr{V}(n,t)$. *Let* d *be an integer such that* $\widehat{\mu h}(S) = 0$ *whenever* $|S| < d$. *Then*

$$\mathrm{disc}_P(H) \leqslant \left(\frac{4\mathrm{e}t^2}{nd}\right)^{d/2}.$$

PROOF. Throughout the proof, the symbol $\mathscr{U}$ shall stand for the uniform distribution over the relevant domain. By Lemma 3.3,

$$\mathrm{disc}_P(H)^2 \leqslant 4^t \mathop{\mathbf{E}}_{(S,T)\sim\mathscr{U}} |\Gamma(S,T)|, \tag{4.38}$$

where we let

$$\Gamma(S,T) = \mathop{\mathbf{E}}_{x\sim\mathscr{U}}\Big[\mu(x|_S)\mu(x|_T)h(x|_S)h(x|_T)\Big].$$

To analyze this expression, we prove two key claims.

CLAIM 4.24. *Assume that* $|S \cap T| \leqslant d-1$. *Then* $\Gamma(S,T) = 0$.

PROOF. The claim is immediate from the fact that the Fourier spectrum of the function μh is supported on characters of order d and higher. For completeness, we include a more detailed derivation. By renumbering the variables, we may assume that $S = \{1,2,\ldots,t\}$. Then

$$\begin{aligned}\Gamma(S,T) &= \mathop{\mathbf{E}}_{x\sim\mathscr{U}}\Big[\mu(x_1,\ldots,x_t)\mu(x|_T)h(x_1,\ldots,x_t)h(x|_T)\Big] \\ &= 2^{-n}\sum_{x_1,\ldots,x_t}\mu(x_1,\ldots,x_t)h(x_1,\ldots,x_t)\underbrace{\sum_{x_{t+1},\ldots,x_n}\mu(x|_T)h(x|_T)}\,.\end{aligned}$$

Since $|S \cap T| \leqslant d-1$, the underbraced expression is a real function of fewer than d variables. The claim follows by the assumption on the Fourier spectrum of μh. □

CLAIM 4.25. *Assume that* $|S \cap T| = k$. *Then* $|\Gamma(S,T)| \leqslant 2^{k-2t}$.

PROOF. The claim is immediate from Proposition 4.22. For completeness, we include a more detailed explanation. By renumbering the variables, we may assume that

$$\begin{aligned} S &= \{1, 2, \ldots, t\}, \\ T &= \{1, 2, \ldots, k\} \cup \{t+1, t+2, \ldots, t+(t-k)\}. \end{aligned}$$

We have:

$$\begin{aligned} |\Gamma(S,T)| &\leqslant \mathop{\mathbf{E}}_{x \sim \mathscr{U}} \left| \mu(x|_S)\mu(x|_T)h(x|_S)h(x|_T) \right| \\ &= \mathop{\mathbf{E}}_{x_1,\ldots,x_{2t-k}} [\mu(x_1,\ldots,x_t)\mu(x_1,\ldots,x_k,x_{t+1},\ldots,x_{2t-k})] \\ &\leqslant \underbrace{\mathop{\mathbf{E}}_{x_1,\ldots,x_t} [\mu(x_1,\ldots,x_t)]}_{=2^{-t}} \cdot \max_{x_1,\ldots,x_k} \underbrace{\mathop{\mathbf{E}}_{x_{t+1},\ldots,x_{2t-k}} [\mu(x_1,\ldots,x_k,x_{t+1},\ldots,x_{2t-k})]}_{\leqslant 2^{-(t-k)}}. \end{aligned}$$

The bounds 2^{-t} and $2^{-(t-k)}$ follow because μ is a probability distribution. □

In view of Claims 4.24 and 4.25, inequality (4.38) simplifies to

$$\mathrm{disc}_P(H)^2 \leqslant \sum_{k=d}^{t} 2^k \, \mathbf{P}[|S \cap T| = k].$$

Since

$$\mathbf{P}[|S \cap T| = k] \leqslant \binom{t}{k} \left(\frac{t}{n}\right)^k \leqslant \left(\frac{\mathrm{e}t^2}{nk}\right)^k,$$

and since the discrepancy cannot exceed 1, we conclude that

$$\operatorname{disc}_P(H)^2 \leqslant \min\left\{1,\ \sum_{k=d}^{t}\left(\frac{2et^2}{nk}\right)^k\right\} \leqslant \left(\frac{4et^2}{nd}\right)^d.$$

This completes the proof of Theorem 4.23. □

With Theorem 4.23 in hand, we now obtain a weaker form of Theorem 4.8 on the randomized communication complexity of pattern matrices.

THEOREM 4.26 (Sherstov [195, 203]). *Let $f\colon\{0,1\}^t \to \{-1,+1\}$ be a given function. Let n be a given integer, $t \mid n$. Put $F = [f(x|_V)]_{x,V}$, where the indices range as follows: $x \in \{0,1\}^n$, $V \in \mathscr{V}(n,t)$. Then for every parameter $\varepsilon \in [0,1)$ and $\delta < \varepsilon/2$,*

$$R_\delta(F) \geqslant \frac{1}{2}\deg_\varepsilon(f)\log\left(\frac{n\deg_\varepsilon(f)}{4et^2}\right) - \log\left(\frac{1}{\varepsilon - 2\delta}\right). \tag{4.39}$$

In particular,

$$R_{1/7}(F) > \frac{1}{2}\deg_{1/3}(f)\log\left(\frac{n\deg_{1/3}(f)}{4et^2}\right) - 5. \tag{4.40}$$

PROOF. Since (4.39) immediately implies (4.40), we will focus on the former in the remainder of the proof. Let $d = \deg_\varepsilon(f) \geqslant 1$. By Theorem 4.5, there exists a function $h\colon\{0,1\}^t \to \{-1,+1\}$ and a probability distribution μ on $\{0,1\}^t$ such that

$$\widehat{\mu h}(S) = 0, \qquad |S| < d, \tag{4.41}$$

$$\sum_{z\in\{0,1\}^t} f(z)\mu(z)h(z) > \varepsilon. \tag{4.42}$$

Letting $H = [h(x|_V)]_{x,V}$ and $P = 2^{-n+t}\binom{n}{t}^{-t}[\mu(x|_V)]_{x,V}$, we obtain from (4.41) and Theorem 4.23 that

$$\operatorname{disc}_P(H) \leqslant \left(\frac{4et^2}{nd}\right)^{d/2}. \tag{4.43}$$

At the same time, one sees from (4.42) that

$$\langle F, H \circ P\rangle > \varepsilon. \tag{4.44}$$

The theorem now follows from (4.43), (4.44), and the generalized discrepancy method (Theorem 3.7). □

We close this section with a weaker form of Theorem 4.14 on the discrepancy of a pattern matrix.

THEOREM 4.27 (Sherstov [202, 203]). *Let $f\colon\{0,1\}^t \to \{-1,+1\}$ be a given non-constant function. Let n be a given integer, $t \mid n$. Put $F = [f(x|_V)]_{x,V}$, where the indices range as follows: $x \in \{0,1\}^n$, $V \in \mathscr{V}(n,t)$. Then for all $\gamma > 0$,*

$$\operatorname{disc}(F) \leqslant \gamma + \left(\frac{4et^2}{n\deg_{1-\gamma}(f)}\right)^{\deg_{1-\gamma}(f)/2}. \tag{4.45}$$

In particular,

$$\operatorname{disc}(F) \leqslant \left(\frac{4et^2}{n\deg_{\pm}(f)}\right)^{\deg_{\pm}(f)/2}. \tag{4.46}$$

PROOF. Letting $\gamma \searrow 0$ in (4.45) yields (4.46). In the remainder of the proof, we will focus on the former bound. Let $d = \deg_{1-\gamma}(f) \geqslant 1$. By Theorem 4.5, there exists a function $h\colon\{0,1\}^t \to \{-1,+1\}$ and a probability distribution μ on $\{0,1\}^t$

such that

$$\widehat{\mu h}(S) = 0, \qquad |S| < d, \tag{4.47}$$

$$\sum_{z \in \{0,1\}^t} f(z)\mu(z)h(z) > 1 - \gamma. \tag{4.48}$$

Letting $H = [h(x|_V)]_{x,V}$ and $P = 2^{-n+t}\left(\frac{n}{t}\right)^{-t}[\mu(x|_V)]_{x,V}$, we obtain from (4.47) and Theorem 4.23 that

$$\mathrm{disc}_P(H) \leqslant \left(\frac{4\mathrm{e}t^2}{nd}\right)^{d/2}. \tag{4.49}$$

Moreover,

$$\begin{aligned} \mathrm{disc}_P(F) &\leqslant \mathrm{disc}_P(H) + \|(F - H) \circ P\|_1 \\ &= \mathrm{disc}_P(H) + 1 - \langle F \circ H, P \rangle \\ &\leqslant \mathrm{disc}_P(H) + \gamma, \end{aligned} \tag{4.50}$$

where the last step uses (4.48). The theorem follows from (4.49) and (4.50). □

Chapter 5

Quantum Communication

A counterpart to the classical randomized model of communication is the more powerful *quantum* model. We prove that the pattern matrix method of the previous chapter applies unchanged to this more difficult model, yielding a new source of communication lower bounds. As an illustration of the quantum pattern matrix method, we give a new and simple proof of Razborov's breakthrough lower bounds for disjointness and the other symmetric functions [177]. Finally, we contrast the pattern matrix method with another duality-based technique, the *block composition method* of Shi and Zhu [205].

5.1 Introduction

Quantum communication complexity, introduced by Yao [226], studies the amount of quantum communication necessary to compute a Boolean function F whose arguments are distributed among several parties. As before, one considers a function $F\colon X \times Y \to \{-1,+1\}$, where X and Y are some finite sets. One of the parties, Alice, receives an input $x \in X$, and the other party, Bob, receives an input $y \in Y$. Their objective is to evaluate $F(x,y)$. To this end, Alice and Bob can exchange messages back and forth through a shared communication channel. This time, however, they can exchange *quantum* bits in addition to classical information. Furthermore, Alice and Bob can take advantage of arbitrary *prior entanglement,* in the sense of Einstein, Podolsky, and Rosen. The ε-error quantum communication complexity of F with prior entanglement, denoted $Q^*_\varepsilon(F)$, is the least cost of a protocol that computes F correctly with probability at least $1-\varepsilon$ on every input. We defer a more detailed and rigorous definition of the quantum model to Section 5.2. Quantum communication is of course the counterpart of the classical randomized model, studied in the previous chapter, in which the parties exchange classical bits 0, 1 and additionally share an unlimited supply of unbiased random bits.

Quantum computing has drawn considerable interest, both as a natural and conceivably practical model and as a valuable source of new problems and insights in physics, information theory, computer science, and other disciplines. The theory of quantum communication complexity in particular has seen steady progress over the past two decades [226, 16, 54, 117, 114, 177, 144], although quantum protocols remain less understood than classical ones. Our main result is that the pattern matrix

method, developed in the previous chapter for classical protocols, extends readily to the quantum model, yielding a powerful new technique for communication lower bounds. In particular, we prove in Section 5.4 that the (n, t, f)-pattern matrix F obeys

$$Q^*_{1/7}(F) > \frac{1}{4} \deg_{1/3}(f) \log \left(\frac{n}{t}\right) - 3, \tag{5.1}$$

for every Boolean function $f: \{0,1\}^t \to \{-1,+1\}$. Analogous to the previous chapter, we also obtain lower bounds for quantum protocols with small bias.

Recall from Proposition 4.10 that the lower bound (5.1) for bounded-error communication is within a polynomial of optimal, even for *classical deterministic* protocols. In particular, equation (5.1) exhibits a new class of communication problems (namely, the family of all pattern matrices) with polynomially related classical and quantum complexity. Prior to our work, the largest class of problems with polynomially related quantum and classical bounded-error complexities was the class of symmetric functions [177], which is broadly subsumed by pattern matrices. Exhibiting a polynomial relationship between the quantum and classical bounded-error complexities for *all* functions $F: X \times Y \to \{-1,+1\}$ is a longstanding open problem.

As illustration of the pattern matrix method in the quantum model, we revisit the quantum communication complexity of symmetric functions. In this framework Alice has a string $x \in \{0,1\}^n$, Bob has a string $y \in \{0,1\}^n$, and their objective is to compute $D(\sum x_i y_i)$ for some predicate $D: \{0,1,\ldots,n\} \to \{-1,+1\}$ fixed in advance. This setting encompasses several familiar functions, such as DISJOINTNESS (determining if x and y intersect) and INNER PRODUCT MODULO 2 (determining if x and y intersect in an odd number of positions). In a celebrated result, Razborov [177] established optimal lower bounds on the quantum communication complexity of every function of such form:

THEOREM (Razborov [177]). *Let $D: \{0,1,\ldots,n\} \to \{-1,+1\}$ be a given predicate. Put $f(x,y) = D(\sum x_i y_i)$. Then*

$$Q^*_{1/3}(f) \geqslant \Omega(\sqrt{n \ell_0(D)} + \ell_1(D)),$$

where $\ell_0(D) \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ *and* $\ell_1(D) \in \{0, 1, \ldots, \lceil n/2 \rceil\}$ *are the smallest integers such that* D *is constant in the range* $[\ell_0(D), n - \ell_1(D)]$.

Using the pattern matrix method, we give a new and simple proof of Razborov's result in Section 5.5. No alternate proof was available prior to our work, despite a long line of research on this problem [16, 54, 117, 114, 96, 144].

In the concluding part of this chapter, we contrast the pattern matrix method with a different technique for lower bounds on bounded-error communication, the *block composition method* of Shi and Zhu [205]. Discovered independently of the pattern matrix method, the technique of Shi and Zhu also exploits the dual characterization of the approximate degree but in a rather different way. We offer a full technical exposition of the block composition method in Section 5.6, followed by a comparison of the two methods in Section 5.7.

5.2 Quantum model of communication

There are several equivalent ways to describe a quantum communication protocol. Our description closely follows Razborov [177]. Let $\mathscr{A}$ and $\mathscr{B}$ be complex finite-dimensional Hilbert spaces. Let $\mathscr{C}$ be a Hilbert space of dimension 2, whose orthonormal basis we denote by $|0\rangle$, $|1\rangle$. Consider the tensor product $\mathscr{A} \otimes \mathscr{C} \otimes \mathscr{B}$, which is itself a Hilbert space with an inner product inherited from $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{C}$. The *state* of a quantum system is a unit vector in $\mathscr{A} \otimes \mathscr{C} \otimes \mathscr{B}$, and conversely any such unit vector corresponds to a distinct quantum state. The quantum system starts in a given state and traverses a sequence of states, each obtained from the previous one via a unitary transformation chosen according to the protocol. Formally, a *quantum communication protocol* is a finite sequence of unitary transformations

$$U_1 \otimes I_{\mathscr{B}},\ I_{\mathscr{A}} \otimes U_2,\ U_3 \otimes I_{\mathscr{B}},\ I_{\mathscr{A}} \otimes U_4,\ \ldots,\ U_{2k-1} \otimes I_{\mathscr{B}},\ I_{\mathscr{A}} \otimes U_{2k},$$

where: $I_{\mathscr{A}}$ and $I_{\mathscr{B}}$ are the identity transformations in $\mathscr{A}$ and $\mathscr{B}$, respectively; $U_1, U_3, \ldots, U_{2k-1}$ are unitary transformations in $\mathscr{A} \otimes \mathscr{C}$; and $U_2, U_4, \ldots, U_{2k}$ are unitary transformations in $\mathscr{C} \otimes \mathscr{B}$. The *cost* of the protocol is the length of this

sequence, namely, $2k$. On Alice's input $x \in X$ and Bob's input $y \in Y$ (where X, Y are given finite sets), the computation proceeds as follows.

1. The quantum system starts out in an initial state $\mathsf{Initial}(x, y)$.
2. Through successive applications of the above unitary transformations, the system reaches the state

$$\begin{aligned}&\mathsf{Final}(x, y)\\ &\quad = (I_{\mathscr{A}} \otimes U_{2k})(U_{2k-1} \otimes I_{\mathscr{B}}) \cdots (I_{\mathscr{A}} \otimes U_2)(U_1 \otimes I_{\mathscr{B}})\, \mathsf{Initial}(x, y).\end{aligned}$$

3. Let v denote the projection of $\mathsf{Final}(x, y)$ onto $\mathscr{A} \otimes \mathrm{span}(|1\rangle) \otimes \mathscr{B}$. The output of the protocol is 1 with probability $\langle v, v \rangle$, and 0 with the complementary probability $1 - \langle v, v \rangle$.

All that remains is to specify how the initial state $\mathsf{Initial}(x, y) \in \mathscr{A} \otimes \mathscr{C} \otimes \mathscr{B}$ is constructed from x, y. It is here that the model with prior entanglement differs from the model without prior entanglement. In the model without prior entanglement, $\mathscr{A}$ and $\mathscr{B}$ have orthonormal bases $\{|x, w\rangle : x \in X,\ w \in W\}$ and $\{|y, w\rangle : y \in Y,\ w \in W\}$, respectively, where W is a finite set corresponding to the private workspace of each of the parties. The initial state is the pure state

$$\mathsf{Initial}(x, y) = |x, 0\rangle\, |0\rangle\, |y, 0\rangle,$$

where $0 \in W$ is a certain fixed element. In the model with prior entanglement, the spaces $\mathscr{A}$ and $\mathscr{B}$ have orthonormal bases $\{|x, w, e\rangle : x \in X,\ w \in W,\ e \in E\}$ and $\{|y, w, e\rangle : y \in Y,\ w \in W,\ e \in E\}$, respectively, where W is as before and E is a finite set corresponding to the prior entanglement. The initial state is now the entangled state

$$\mathsf{Initial}(x, y) = \frac{1}{\sqrt{|E|}} \sum_{e \in E} |x, 0, e\rangle\, |0\rangle\, |y, 0, e\rangle.$$

Apart from finite size, no assumptions are made about W or E. In particular, the model with prior entanglement allows for an unlimited supply of entangled qubits. This mirrors the unlimited supply of shared random bits in the classical public-coin randomized model.

Let $f: X \times Y \to \{-1, +1\}$ be a given function. A quantum protocol Π is said to compute f with error ε if

$$\mathbf{P}\Big[f(x, y) = (-1)^{\Pi(x,y)}\Big] \geqslant 1 - \varepsilon$$

for all x, y, where the random variable $\Pi(x, y) \in \{0, 1\}$ is the output of the protocol on input (x, y). Let $Q_\varepsilon(f)$ denote the least cost of a quantum protocol without prior entanglement that computes f with error ε. Define $Q^*_\varepsilon(f)$ analogously for protocols with prior entanglement. The precise choice of a constant $0 < \varepsilon < 1/2$ affects $Q_\varepsilon(f)$ and $Q^*_\varepsilon(f)$ by at most a constant factor, and thus the setting $\varepsilon = 1/3$ entails no loss of generality.

Let $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a predicate. We associate with D the function $f: \{0, 1\}^n \times \{0, 1\}^n \to \{-1, +1\}$ defined by $f(x, y) = D(\sum x_i y_i)$. We let $Q_\varepsilon(D) = Q_\varepsilon(f)$ and $Q^*_\varepsilon(D) = Q^*_\varepsilon(f)$. More generally, by computing D in the quantum model we mean computing the associated function f.

5.3 Quantum generalized discrepancy

The generalized discrepancy method, discussed in Section 3.3, readily extends to the quantum model of communication. A starting point in our discussion is the following fact due to Linial and Shraibman [144, Lem. 10], with closely analogous statements established earlier by Yao [226], Kremer [134], and Razborov [177].

THEOREM 5.1 (Protocol factorization [226, 134, 177, 144]). *Let X, Y be finite sets. Let Π be a quantum protocol (with or without prior entanglement) with cost C qubits and input sets X and Y. Then*

$$\Big[\mathbf{E}[\Pi(x, y)]\Big]_{x,y} = AB$$

for some real matrices A, B *with* $\|A\|_{\mathrm{F}} \leqslant 2^C \sqrt{|X|}$ *and* $\|B\|_{\mathrm{F}} \leqslant 2^C \sqrt{|Y|}$.

Theorem 5.1 states that the matrix of acceptance probabilities of every low-cost protocol Π has a nontrivial factorization. This transition from quantum protocols to matrix factorization is now a standard technique and has been used by various authors in various contexts. The generalized discrepancy method was first applied in the quantum setting by Klauck [114, Thm. 4] and reformulated more broadly by Razborov [177]. The treatment in [177] being informal, we now propose a precise formulation of the quantum generalized discrepancy method and supply a proof.

THEOREM 5.2 (Quantum generalized discrepancy). *Let* X, Y *be finite sets and* $f : X \times Y \to \{-1, +1\}$ *a given function. Let* $\Psi = [\Psi_{xy}]_{x \in X,\, y \in Y}$ *be any real matrix with* $\|\Psi\|_1 = 1$. *Then for each* $\varepsilon > 0$,

$$4^{Q_\varepsilon(f)} \geqslant 4^{Q^*_\varepsilon(f)} \geqslant \frac{\langle \Psi, F \rangle - 2\varepsilon}{3\,\|\Psi\| \sqrt{|X|\,|Y|}},$$

where $F = [f(x, y)]_{x \in X,\, y \in Y}$.

PROOF. Fix any quantum protocol that computes f with error ε and cost C. Let the random variable $\Pi(x, y) \in \{0, 1\}$ denote the output of the protocol on input $x \in X$, $y \in Y$. Define the matrix of acceptance probabilities

$$\Pi = \Big[\, \mathbf{E}[\Pi(x, y)] \Big]_{x \in X,\, y \in Y}.$$

Then we can write $F = (J - 2\Pi) + 2E$, where J is the all-ones matrix and E is some matrix with $\|E\|_\infty \leqslant \varepsilon$. As a result,

$$\begin{aligned} \langle \Psi, J - 2\Pi \rangle &= \langle \Psi, F \rangle - 2 \langle \Psi, E \rangle \\ &\geqslant \langle \Psi, F \rangle - 2\varepsilon \, \|\Psi\|_1 \\ &= \langle \Psi, F \rangle - 2\varepsilon. \end{aligned} \tag{5.2}$$

On the other hand, Theorem 5.1 guarantees the existence of matrices A and B with $AB = \Pi$ and $\|A\|_{\mathrm{F}} \|B\|_{\mathrm{F}} \leqslant 4^C \sqrt{|X|\,|Y|}$. Therefore,

$$\begin{aligned}
\langle \Psi, J - 2\Pi \rangle &\leqslant \|\Psi\| \, \|J - 2\Pi\|_\Sigma && \text{by (2.5)} \\
&\leqslant \|\Psi\| \, \left(\sqrt{|X|\,|Y|} + 2\,\|\Pi\|_\Sigma \right) && \text{since } \|J\|_\Sigma = \sqrt{|X|\,|Y|} \\
&\leqslant \|\Psi\| \, \left(\sqrt{|X|\,|Y|} + 2\,\|A\|_{\mathrm{F}} \, \|B\|_{\mathrm{F}} \right) && \text{by Prop. 2.13} \\
&\leqslant \|\Psi\| \, \left(2 \cdot 4^C + 1 \right) \sqrt{|X|\,|Y|}. && (5.3)
\end{aligned}$$

The theorem follows by comparing (5.2) and (5.3). □

5.4 Lower bounds using the pattern matrix method

Now that we have a quantum analogue of generalized discrepancy, we will see that the pattern matrix method of Chapter 4 applies equally well to quantum communication. Indeed, the reader will note that our proofs are identical in the randomized and quantum cases.

THEOREM 5.3 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, for a given function $f : \{0,1\}^t \to \{-1,+1\}$. Then for every parameter $\varepsilon \in [0,1)$ and every $\delta < \varepsilon/2$,*

$$Q^*_\delta(F) \geqslant \frac{1}{4} \deg_\varepsilon(f) \log\left(\frac{n}{t}\right) - \frac{1}{2} \log\left(\frac{3}{\varepsilon - 2\delta}\right). \tag{5.4}$$

In particular,

$$Q^*_{1/7}(F) > \frac{1}{4} \deg_{1/3}(f) \log\left(\frac{n}{t}\right) - 3. \tag{5.5}$$

PROOF. Let $d = \deg_\varepsilon(f) \geqslant 1$ and define Ψ as in the proof of Theorem 4.8. Then (5.4) follows from (4.16), (4.18), and the quantum generalized discrepancy method (Theorem 5.2). Finally, (5.5) follows immediately from (5.4). □

Recall from Proposition 4.10 that the lower bound (5.5) derived above for bounded-error quantum communication complexity is tight up to a polynomial factor, even for classical deterministic protocols. Analogous to the classical case, we have the following corollary of Theorem 5.3 on function composition.

COROLLARY 5.4 (Sherstov [203]). *Let $f\colon \{0,1\}^t \to \{-1,+1\}$ be given. Define $F\colon \{0,1\}^{4t} \times \{0,1\}^{4t} \to \{-1,+1\}$ by*

$$F(x,y) = f(\ldots, (x_{i,1}y_{i,1} \vee x_{i,2}y_{i,2} \vee x_{i,3}y_{i,3} \vee x_{i,4}y_{i,4}), \ldots).$$

Then

$$Q^*_{1/7}(F) > \frac{1}{4}\deg_{1/3}(f) - 3.$$

PROOF. The $(2t, t, f)$-pattern matrix is as a submatrix of $[F(x,y)]_{x,y\in\{0,1\}^{4t}}$. □

Note that Theorem 5.3 yields lower bounds not only for bounded-error communication but also small-bias communication. In the latter case, one first needs to show that the base function $f\colon \{0,1\}^t \to \{-1,+1\}$ cannot be approximated pointwise within $1 - o(1)$ by a real polynomial of a given degree d. In our next result, we derive a different lower bound for small-bias quantum communication, this time using the assumption that the threshold weight $W(f,d)$ is high.

THEOREM 5.5 (Sherstov [203]). *Let F be the (n,t,f)-pattern matrix, where $f\colon \{0,1\}^t \to \{-1,+1\}$ is a given function. Then for every integer $d \geqslant 1$ and real $\gamma \in (0,1)$,*

$$Q^*_{1/2-\gamma/2}(F) \geqslant \frac{1}{4}\min\left\{d\log\frac{n}{t},\ \log\frac{W(f,d-1)}{2t}\right\} - \frac{1}{2}\log\frac{3}{\gamma}. \tag{5.6}$$

In particular,

$$Q^*_{1/2-\gamma/2}(F) \geqslant \frac{1}{4}\deg_{\pm}(f)\log\left(\frac{n}{t}\right) - \frac{1}{2}\log\frac{3}{\gamma}. \tag{5.7}$$

PROOF. Define the matrix Ψ as in the proof of Theorem 4.11. Then (5.6) follows immediately from (4.26), (4.28), and the quantum generalized discrepancy method (Theorem 5.2). Letting $d = \deg_{\pm}(f)$ in (5.6) yields (5.7), since $W(f, d-1) = \infty$ in that case. □

Recall from Theorem 2.4 that the quantities $E(f,d)$ and $W(f,d)$ are related for all f and d. In particular, the lower bounds for small-bias communication in Theorems 5.3 and 5.5 are quite close, as their classical counterparts in Chapter 4. In addition, these lower bounds on quantum small-bias communication are close to optimal even for classical protocols, as we showed in Theorem 4.12.

5.5 Tight lower bounds for symmetric functions

As another illustration of the pattern matrix method, we now revisit Razborov's optimal lower bounds on the quantum communication complexity of symmetric functions:

THEOREM 5.6 (Razborov [177]). *Let $D\colon\{0,1,\ldots,n\} \to \{-1,+1\}$ be a given predicate. Then*

$$Q^*_{1/3}(D) \geqslant \Omega(\sqrt{n\ell_0(D)} + \ell_1(D)),$$

where $\ell_0(D) \in \{0,1,\ldots,\lfloor n/2\rfloor\}$ and $\ell_1(D) \in \{0,1,\ldots,\lceil n/2\rceil\}$ are the smallest integers such that D is constant in the range $[\ell_0(D), n-\ell_1(D)]$.

Using the quantum version of the pattern matrix method, we give a new and simple proof of this theorem. No alternate proof was available prior to this work, despite the fact that this problem has drawn the attention of various re-

searchers [16, 54, 117, 114, 96, 144]. Moreover, the next-best lower bounds for general predicates were nowhere close to Theorem 5.6. To illustrate, consider the disjointness predicate D, given by $D(t) = 1 \Leftrightarrow t = 0$. Theorem 5.6 shows that it has communication complexity $\Omega(\sqrt{n})$, while the next-best lower bound [16, 54] was only $\Omega(\log n)$.

We first solve the problem for all predicates D that change value close to 0. Extension to the general case will require an additional step.

THEOREM 5.7. *Let $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a given predicate. Suppose that $D(\ell) \neq D(\ell - 1)$ for some $\ell \leqslant \frac{1}{8}n$. Then*

$$Q^*_{1/3}(D) \geqslant \Omega(\sqrt{n\ell}).$$

PROOF (Sherstov [203]). It suffices to show that $Q^*_{1/7}(D) \geqslant \Omega(\sqrt{n\ell})$. Define $f: \{0,1\}^{\lfloor n/4 \rfloor} \to \{-1, +1\}$ by $f(z) = D(|z|)$. Then $\deg_{1/3}(f) \geqslant \Omega(\sqrt{n\ell})$ by Theorem 2.5. Theorem 5.3 implies that

$$Q^*_{1/7}(F) \geqslant \Omega(\sqrt{n\ell}),$$

where F is the $(2\lfloor n/4 \rfloor, \lfloor n/4 \rfloor, f)$-pattern matrix. Since F occurs as a submatrix of $[D(|x \wedge y|)]_{x,y}$, the proof is complete. □

The remainder of this section is a simple if tedious exercise in shifting and padding. We note that Razborov's proof concludes in a similar way (see [177], beginning of Section 5).

THEOREM 5.8. *Let $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a given predicate. Suppose that $D(\ell) \neq D(\ell - 1)$ for some $\ell > \frac{1}{8}n$. Then*

$$Q^*_{1/3}(D) \geqslant c(n - \ell) \tag{5.8}$$

for some absolute constant $c > 0$.

PROOF (Sherstov [203]). Consider the communication problem of computing $D(|x \wedge y|)$ when the last k bits in x and y are fixed to 1. In other words, the new problem is to compute $D_k(|x' \wedge y'|)$, where $x', y' \in \{0,1\}^{n-k}$ and the predicate $D_k\colon \{0,1,\ldots,n-k\} \to \{-1,+1\}$ is given by $D_k(i) \equiv D(k+i)$. Since the new problem is a restricted version of the original, we have

$$Q^*_{1/3}(D) \geqslant Q^*_{1/3}(D_k). \tag{5.9}$$

We complete the proof by placing a lower bound on $Q^*_{1/3}(D_k)$ for

$$k = \ell - \left\lfloor \frac{\alpha}{1-\alpha} \cdot (n-\ell) \right\rfloor,$$

where $\alpha = \frac{1}{8}$. Note that k is an integer between 1 and ℓ (because $\ell > \alpha n$). The equality $k = \ell$ occurs if and only if $\left\lfloor \frac{\alpha}{1-\alpha}(n-\ell) \right\rfloor = 0$, in which case (5.8) holds trivially for c suitably small. Thus, we can assume that $1 \leqslant k \leqslant \ell - 1$, in which case $D_k(\ell - k) \neq D_k(\ell - k - 1)$ and $\ell - k \leqslant \alpha(n-k)$. Therefore, Theorem 5.7 is applicable to D_k and yields:

$$Q^*_{1/3}(D_k) \geqslant C\sqrt{(n-k)(\ell-k)}, \tag{5.10}$$

where $C > 0$ is an absolute constant. Calculations reveal:

$$n - k = \left\lfloor \frac{1}{1-\alpha} \cdot (n-\ell) \right\rfloor, \qquad \ell - k = \left\lfloor \frac{\alpha}{1-\alpha} \cdot (n-\ell) \right\rfloor. \tag{5.11}$$

The theorem is now immediate from (5.9)–(5.11). □

Together, Theorems 5.7 and 5.8 give the main result of this section:

THEOREM (Razborov [177]). *Let* $D\colon\{0,1,\ldots,n\}\to\{-1,+1\}$. *Then*

$$Q^*_{1/3}(D)\geqslant\Omega(\sqrt{n\ell_0(D)}+\ell_1(D)),$$

where $\ell_0(D)\in\{0,1,\ldots,\lfloor n/2\rfloor\}$ *and* $\ell_1(D)\in\{0,1,\ldots,\lceil n/2\rceil\}$ *are the smallest integers such that* D *is constant in the range* $[\ell_0(D),n-\ell_1(D)]$.

PROOF (Sherstov [203]). If $\ell_0(D)\neq 0$, set $\ell=\ell_0(D)$ and note that $D(\ell)\neq D(\ell-1)$ by definition. One of Theorems 5.7 and 5.8 must be applicable, and therefore $Q^*_{1/3}(D)\geqslant\min\{\Omega(\sqrt{n\ell}),\ \Omega(n-\ell)\}$. Since $\ell\leqslant n/2$, this simplifies to

$$Q^*_{1/3}(D)\geqslant\Omega(\sqrt{n\ell_0(D)}). \tag{5.12}$$

If $\ell_1(D)\neq 0$, set $\ell=n-\ell_1(D)+1\geqslant n/2$ and note that $D(\ell)\neq D(\ell-1)$ as before. By Theorem 5.8,

$$Q^*_{1/3}(D)\geqslant\Omega(\ell_1(D)). \tag{5.13}$$

The theorem follows from (5.12) and (5.13). □

5.6 Lower bounds using the block composition method

We now present a different technique for lower bounds on bounded-error classical and quantum communication, the *block composition method* of Shi and Zhu [205]. Discovered independently of the author's pattern matrix method [203], the technique of Shi and Zhu also exploits the dual characterization of the approximate degree (Theorem 4.5) but in a rather different way. The pattern matrix method is based on the idea of applying the same function to distinct sets of variables, whereas the block composition method is based on the idea of hardness amplification by composition.

Given functions $f\colon\{0,1\}^n\to\{-1,+1\}$ and $g\colon\{0,1\}^k\times\{0,1\}^k\to\{0,1\}$, let $f\circ g^n$ denote the composition of f with n independent copies of g. More

formally, the function $f \circ g^n : \{0,1\}^{nk} \times \{0,1\}^{nk} \to \{-1,+1\}$ is given by

$$(f \circ g^n)(x,y) = f(\ldots, g(x^{(i)}, y^{(i)}), \ldots),$$

where $x = (\ldots, x^{(i)}, \ldots) \in \{0,1\}^{nk}$ and $y = (\ldots, y^{(i)}, \ldots) \in \{0,1\}^{nk}$. The block composition method gives a lower bound on the communication complexity of $f \circ g^n$ in terms of certain properties of f and g. The relevant property of f is simply its approximate degree. The relevant property of g is its *spectral discrepancy*, formalized next.

DEFINITION 5.9 (Shi and Zhu [205]). Given $g : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$, its *spectral discrepancy* $\rho(g)$ is the least $\rho \geqslant 0$ for which there exist sets $A, B \subseteq \{0,1\}^k$ and a distribution μ on $A \times B$ such that

$$\left\| \left[\mu(x,y)(-1)^{g(x,y)} \right]_{x \in A, y \in B} \right\| \leqslant \frac{\rho}{\sqrt{|A|\,|B|}}, \tag{5.14}$$

$$\left\| \left[\mu(x,y) \right]_{x \in A, y \in B} \right\| \leqslant \frac{1+\rho}{\sqrt{|A|\,|B|}}, \tag{5.15}$$

and

$$\sum_{(x,y) \in A \times B} \mu(x,y)(-1)^{g(x,y)} = 0. \tag{5.16}$$

In view of (5.14) alone, the spectral discrepancy $\rho(g)$ is an upper bound on the discrepancy $\mathrm{disc}(g)$. The key additional requirement (5.15) is satisfied, for example, by doubly stochastic matrices [95, §8.7]: if $A = B$ and all row and column sums in $[\mu(x,y)]_{x \in A, y \in A}$ are $1/|A|$, then $\|[\mu(x,y)]_{x \in A, y \in A}\| = 1/|A|$.

As an illustration, consider the well-studied inner product function, given by $\mathrm{IP}_k(x,y) = \bigoplus_{i=1}^{k} (x_i \wedge y_i)$.

PROPOSITION 5.10 (Shi and Zhu [205]). *The function* IP_k *satisfies*

$$\rho(\mathrm{IP}_k) \leqslant \frac{1}{\sqrt{2^k - 1}}.$$

PROOF (Shi and Zhu [205]). Take μ to be the uniform distribution over $A \times B$, where $A = \{0,1\}^k \setminus \{0^k\}$ and $B = \{0,1\}^k$. □

We are prepared to state the block composition method.

THEOREM 5.11 (Shi and Zhu [205]). *Fix* $f : \{0,1\}^n \to \{-1,+1\}$ *and* $g : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$. *Put* $d = \deg_{1/3}(f)$ *and* $\rho = \rho(g)$. *If* $\rho \leqslant d/(2en)$, *then*

$$Q^*_{1/3}(f \circ g^n) = \Omega(d).$$

PROOF (adapted from [205]). Fix sets $A, B \subseteq \{0,1\}^k$ and a distribution μ on $A \times B$ with respect to which $\rho = \rho(g)$ is achieved. By Theorem 4.5, there exists $\psi : \{0,1\}^n \to \mathbb{R}$ such that

$$\hat{\psi}(S) = 0 \qquad \text{for } |S| < d, \tag{5.17}$$

$$\sum_{z \in \{0,1\}^n} |\psi(z)| = 1, \tag{5.18}$$

$$\sum_{z \in \{0,1\}^n} \psi(z) f(z) > \frac{1}{3}. \tag{5.19}$$

Define matrices

$$F = \Big[f(\dots, g(x^{(i)}, y^{(i)}), \dots) \Big]_{x,y},$$

$$\Psi = \left[2^n \psi(\dots, g(x^{(i)}, y^{(i)}), \dots) \prod_{i=1}^{n} \mu(x^{(i)}, y^{(i)}) \right]_{x,y},$$

where in both cases the row index $x = (\dots, x^{(i)}, \dots)$ ranges over A^n and the column index $y = (\dots, y^{(i)}, \dots)$ ranges over B^n. In view of (5.16) and (5.19),

$$\|\Psi\|_1 = 1, \qquad \langle F, \Psi \rangle > \frac{1}{3}. \tag{5.20}$$

We proceed to bound $\|\Psi\|$. Put

$$M_S = \left[\prod_{i \in S} (-1)^{g(x^{(i)}, y^{(i)})} \cdot \prod_{i=1}^{n} \mu(x^{(i)}, y^{(i)}) \right]_{x,y}, \qquad S \subseteq \{1, 2, \dots, n\}.$$

Then (5.14) and (5.15) imply, in view of the tensor structure of M_S, that

$$\|M_S\| \leqslant |A|^{-n/2} \, |B|^{-n/2} \, \rho^{|S|} (1 + \rho)^{n - |S|}. \tag{5.21}$$

On the other hand,

$$\begin{aligned}
\|\Psi\| &\leqslant \sum_{S \subseteq [n]} 2^n |\hat{\psi}(S)| \, \|M_S\| \\
&= \sum_{|S| \geqslant d} 2^n |\hat{\psi}(S)| \, \|M_S\| && \text{by (5.17)} \\
&\leqslant \sum_{|S| \geqslant d} \|M_S\| && \text{by (5.18) and Prop. 2.1} \\
&\leqslant |A|^{-n/2} \, |B|^{-n/2} \sum_{i=d}^{n} \binom{n}{i} \rho^i (1 + \rho)^{n-i} && \text{by (5.21).}
\end{aligned}$$

Since $\rho \leqslant d/(2en)$, we further have

$$\|\Psi\| \leqslant |A|^{-n/2} \, |B|^{-n/2} \, 2^{-\Theta(d)}. \tag{5.22}$$

In view of (5.20) and (5.22), the desired lower bound on $Q^*_{1/3}(F)$ now follows by the generalized discrepancy method (Theorem 5.2). □

Proposition 5.10 and Theorem 5.11 have the following consequence:

THEOREM 5.12 (Shi and Zhu [205]). *Fix a function* $f\colon \{0,1\}^n \to \{-1,+1\}$. *Then for* $k \geqslant 2\log n + 5$,

$$Q^*_{1/3}(f \circ \mathrm{IP}^n_k) \geqslant \Omega(\deg_{1/3}(f)).$$

For the disjointness function $\mathrm{DISJ}_k(x,y) = \bigvee_{i=1}^k (x_i \wedge y_i)$, Shi and Zhu obtain $\rho(\mathrm{DISJ}_k) = O(1/k)$. Unlike Proposition 5.10, this fact requires a nontrivial proof using Knuth's calculation of the eigenvalues of certain combinatorial matrices. In conjunction with Theorem 5.11, this upper bound on $\rho(\mathrm{DISJ}_k)$ leads with some work to the following result.

THEOREM 5.13 (Shi and Zhu [205]). *Let* $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$ *be a given predicate. Then*

$$Q^*_{1/3}(f) \geqslant \Omega(n^{1/3}\ell_0(D)^{2/3} + \ell_1(D)).$$

where $\ell_0(D) \in \{0,1,\ldots,\lfloor n/2 \rfloor\}$ *and* $\ell_1(D) \in \{0,1,\ldots,\lceil n/2 \rceil\}$ *are the smallest integers such that* D *is constant in the range* $[\ell_0(D), n - \ell_1(D)]$.

Thus, Shi and Zhu rederive a polynomially weaker form of Razborov's quantum lower bounds for symmetric predicates (Theorem 5.6).

5.7 Pattern matrix method vs. block composition method

Having described the two methods for quantum communication lower bounds, we now compare them in detail. To restate the block composition method,

$$Q^*_{1/3}(f \circ g^n) \geqslant \Omega(\deg_{1/3}(f))$$

provided that

$$\rho(g) \leqslant \frac{\deg_{1/3}(f)}{2en}.$$

The key player in this method is the quantity $\rho(g)$, which needs to be small. This poses two complications. First, the function g will generally need to depend on many variables, from $k = \Theta(\log n)$ to $k = n^{\Theta(1)}$, which weakens the final lower bounds on communication (recall that $\rho(g) \geqslant 2^{-k}$ always). For example, the lower bounds obtained in [205] for symmetric functions are polynomially weaker than Razborov's optimal lower bounds (see Theorems 5.13 and 5.6, respectively).

A second complication, as Shi and Zhu note, is that "estimating the quantity $\rho(g)$ is unfortunately difficult in general" [205]. For example, re-proving Razborov's lower bounds reduces to estimating $\rho(g)$ with g being the disjointness function. Shi and Zhu accomplish this using Hahn matrices, an advanced tool that is also the centerpiece of Razborov's own proof (Razborov's use of Hahn matrices is somewhat more demanding).

These complications do not arise in the pattern matrix method. For example, Theorem 5.3 shows that

$$Q^*_{1/3}(f \circ g^n) \geqslant \Omega(\deg_{1/3}(f))$$

for any function $g\colon \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$ such that the matrix $[g(x,y)]_{x,y}$ contains the following submatrix, up to permutations of rows and columns:

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \tag{5.23}$$

To illustrate, one can take

$$g(x,y) \;=\; x_1 y_1 \;\vee\; x_2 y_2 \;\vee\; x_3 y_3 \;\vee\; x_4 y_4,$$

or

$$g(x,y) \;=\; x_1 y_1 y_2 \;\vee\; \overline{x_1}\, y_1 \overline{y_2} \;\vee\; x_2 \overline{y_1}\, y_2 \;\vee\; \overline{x_2}\,\overline{y_1}\,\overline{y_2}.$$

(In particular, the pattern matrix method subsumes Theorem 5.12.) To summarize, there is a simple function g on only $k = 2$ variables that works universally for all f. This means no technical conditions to check, such as $\rho(g)$, and no blow-up in the number of variables. As a result, in Section 5.5 of this thesis we were able to re-prove Razborov's optimal lower bounds exactly. Moreover, the technical machinery involved was self-contained and disjoint from Razborov's proof.

We have just seen that the pattern matrix method gives strong lower bounds for many functions to which the block composition method does not apply. However, this does not settle the exact relationship between the scopes of applicability of the two methods. Several natural questions arise. If a function $g\colon \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$ has spectral discrepancy $\rho(g) \leqslant \frac{1}{2e}$, does the matrix $[g(x,y)]_{x,y}$ contain (5.23) as a submatrix, up to permutations of rows and columns? An affirmative answer would mean that the pattern matrix method has a strictly greater scope of applicability; a negative answer would mean that the block composition method works in some situations where the pattern matrix method does not apply. If the answer is negative, what can be said for $\rho(g) = o(1)$ or $\rho(g) = n^{-\Theta(1)}$?

Another issue concerns multiparty communication. As we will discuss in Chapter 9, the pattern matrix method extends readily to the multiparty model. This extension depends on the fact that the rows of a pattern matrix are applications of the same function to different subsets of the variables. In the general context of block composition (Section 5.6), it is unclear how to carry out this extension.

Chapter 6

Quantum vs. Classical Communication

A longstanding goal in computational complexity is to prove that quantum bounded-error protocols cannot be superpolynomially more efficient than their classical counterparts. Here, we prove this conjecture for a new class of communication problems, broadly subsuming previous results. In particular, we prove a polynomial relationship between the quantum and classical complexity of computing $f(x \wedge y)$ and $f(x \vee y)$ on input x, y, where f is any given Boolean function. We prove analogous results for other function compositions. Finally, the explore the implications of our techniques for the *log-rank conjecture.*

6.1 Introduction

A major open question in complexity theory and quantum computing is whether quantum communication can be significantly more powerful than classical communication, i.e., whether a superpolynomial gap exists between the quantities $Q^*_{1/3}(F)$ and $R_{1/3}(F)$ for some function $F: X \times Y \to \{-1, +1\}$. Exponential separations between quantum and classical complexity are known in several alternate models of communication [16, 172, 50, 26, 78, 79, 110, 77, 76, 80], such as one-way communication, simultaneous message passing, sampling, and computing a partial function or relation. However, these results do not apply to the original question about $Q^*_{1/3}(F)$ and $R_{1/3}(F)$, and the largest known separation between the two quantities is the quadratic gap for the disjointness function [177, 3].

It is conjectured that $Q^*_{1/3}(F)$ and $R_{1/3}(F)$ are polynomially related for all $F: X \times Y \to \{-1, +1\}$. Despite consistent research efforts, this conjecture appears to be beyond the reach of the current techniques. An intermediate goal, proposed by several authors [54, 116, 205] and still unattained, is to prove the conjecture for the class of communication problems $F: \{0,1\}^n \times \{0,1\}^n \to \{-1, +1\}$ of the form

$$F(x, y) = f(x \wedge y)$$

for a given function $f: \{0,1\}^n \to \{-1, +1\}$. There has been encouraging progress on this problem. To start with, Theorem 5.6 of Razborov solves this problem for the special case of symmetric f, as noted in [205]. Second, recall from Proposition 4.10 and Theorem 5.3 that the bounded-error quantum communication complexity of ev-

ery pattern matrix is within a polynomial of its classical deterministic complexity. In particular, the pattern matrix method is a new source of communication problems with polynomially related classical and quantum complexities, including all functions of the form

$$f(\ldots,\ (x_{i,1}y_{i,1} \vee x_{i,2}y_{i,2} \vee x_{i,3}y_{i,3} \vee x_{i,4}y_{i,4}),\ \ldots),$$

for arbitrary $f\colon\{0,1\}^n \to \{-1,+1\}$. The block composition method of Shi and Zhu, presented earlier in Section 5.6, is another important step toward quantum-classical equivalence. In particular, Shi and Zhu [205] prove a polynomial relationship between quantum and classical communication complexity for the family of functions

$$f(\ldots, g(x_{i,1}, y_{i,1}, \ldots, x_{i,k}, y_{i,k}), \ldots),$$

where $f\colon\{0,1\}^n \to \{-1,+1\}$ is arbitrary and g is any gadget on $2k \geqslant \Omega(\log n)$ variables that has certain pseudorandom analytic properties.

While the above results give further evidence that quantum and classical communication complexities are polynomially related, it remains open to prove this conjecture for all functions of the form $F(x,y) = f(x \wedge y)$. In this chapter, we will solve a variant of the $f(x \wedge y)$ question. Specifically, we will consider the communication problem of computing, on input $x, y \in \{0,1\}^n$, *both* of the quantities $f(x \wedge y)$ and $f(x \vee y)$. Our main result here is a polynomial relationship between the quantum and classical complexity of any such problem, regardless of f. We further show that the quantum complexity of any such problem is polynomially related to its deterministic classical complexity $D(F)$ and to the block sensitivity $\mathrm{bs}(f)$:

THEOREM 6.1 (Sherstov [201]). *Let $f\colon\{0,1\}^n \to \{-1,+1\}$ be arbitrary. Let F denote the communication problem of computing, on input $x, y \in \{0,1\}^n$, both of the quantities $f(x \wedge y)$ and $f(x \vee y)$. Then*

$$D(F) \geqslant R_{1/3}(F) \geqslant Q^*_{1/3}(F) \geqslant \Omega(\mathrm{bs}(f)^{1/4}) \geqslant \Omega(D(F)^{1/12}).$$

Theorem 6.1 and its generalizations in this chapter broadly subsume the quantum-classical equivalence given by Razborov's Theorem 5.6 and the pattern matrix method. A corollary of Theorem 6.1 is that given any f, a polynomial relationship between the classical and quantum complexities is assured for at least one of the communication problems $f(x \wedge y)$, $f(x \vee y)$. More precisely, we have:

COROLLARY 6.2 (Sherstov [201]). *Let $f : \{0,1\}^n \to \{-1,+1\}$ be arbitrary. Let F_1 and F_2 denote the communication problems of computing $f(x \wedge y)$ and $f(x \vee y)$, respectively. Then either*

$$D(F_1) \geqslant R_{1/3}(F_1) \geqslant Q^*_{1/3}(F_1) \geqslant \Omega(\mathrm{bs}(f)^{1/4}) \geqslant \Omega(D(F_1)^{1/12}) \tag{6.1}$$

or

$$D(F_2) \geqslant R_{1/3}(F_2) \geqslant Q^*_{1/3}(F_2) \geqslant \Omega(\mathrm{bs}(f)^{1/4}) \geqslant \Omega(D(F_2)^{1/12}) \tag{6.2}$$

or both.

PROOF. Theorem 6.1 immediately implies (6.1) if $Q^*_{1/3}(F_1) \geqslant Q^*_{1/3}(F_2)$ and implies (6.2) otherwise. □

Apart from giving a polynomial relationship between the quantum and classical complexity of our functions, Theorem 6.1 shows that prior entanglement does not affect their quantum complexity by more than a polynomial. It is an open problem [54] to prove a polynomial relationship for quantum communication complexity with and without prior entanglement, up to an additive logarithmic term. The current largest gap is an additive $\Theta(\log n)$ for the equality function.

In the concluding part of this chapter, we investigate applications of our techniques to related problems. In Section 6.6, we prove that the communication problem of computing $f(x \wedge y)$ and $f(x \vee y)$ satisfies another well-known conjecture, the *log-rank conjecture* of Lovász and Saks [147]. In Section 6.7, we note generalizations of our results on quantum-classical equivalence and the log-rank conjecture to arbitrary compositions of the form $f(g, \ldots, g)$.

6.2 Overview of the proofs

We obtain Theorem 6.1 by bringing together *analytic* and *combinatorial* views of the uniform approximation of Boolean functions. The analytic approach and combinatorial approach have each found important applications in isolation, e.g., [160, 28, 54, 177, 203, 205]. The key to our work is to find a way to combine them.

On the analytic side, a key ingredient in our solution is the pattern matrix method. Essential to the matrix-analytic version of this technique is a closed-form expression for the singular values of every pattern matrix

$$\Psi = \Big[\psi(x|_V \oplus w)\Big]_{x,(V,w)}, \tag{6.3}$$

in terms of the Fourier spectrum of $\psi\colon \{0,1\}^n \to \mathbb{R}$. In particular, Theorem 4.3 critically exploits the fact that the rows of Ψ are applications of the *same* function ψ to various subsets of the variables or their negations. In the communication problems of this chapter, this assumption is severely violated: as Bob's input y ranges over $\{0,1\}^n$, the induced functions $f_y(x) = f(x \wedge y)$ may have nothing to do with each other. This obstacle is fundamental: allowing a distinct function ψ in each row of (6.3) disrupts the spectral structure of Ψ and makes it impossible to force the desired spectral bounds.

We overcome this obstacle by exploiting the additional *combinatorial* structure of the base function $f\colon \{0,1\}^n \to \{-1,+1\}$, which never figured in the lower bounds of the previous chapters. Specifically, we consider the sensitivity of f, the block sensitivity of f, and their polynomial equivalence in our restricted setting, as given by Kenyon and Kutin's elegant Theorem 2.10. We use this combinatorial structure to identify a large submatrix inside $[f(x \wedge y)]_{x,y}$ or $[f(x \vee y)]_{x,y}$ which, albeit not directly representable in the form (6.3), has a certain *dual* matrix that can be represented precisely in this way. Since the pattern matrix method relies only on the spectral structure of this dual matrix, we are able to achieve our goal and place a strong lower bound on the quantum communication complexity. The corresponding upper bound for classical protocols has a short proof, analogous to de Wolf's Proposition 4.10.

The program of Theorem 6.1 can be equivalently described in terms of polynomials rather than functions. Let $\mathscr{F}$ be a subset of Boolean functions $\{0,1\}^n \to \{-1,+1\}$ none of which can be approximated within ε in the ℓ_∞ norm by a polynomial of degree less than d. For each $f \in \mathscr{F}$, our work in Section 4.3 implies the existence of a function $\psi\colon \{0,1\}^n \to \mathbb{R}$ such that $\sum f(x)\psi(x) > \varepsilon \sum |\psi(x)|$ and ψ has zero Fourier mass on the characters of order less than d. This dual object, the polynomial ψ, witnesses the fact that f has no low-degree approximant. Now, there is no reason to believe that a *single* witness polynomial ψ can be found that works for every function in $\mathscr{F}$. A key technical challenge in this work is to show that, under suitable combinatorial constraints that hold in our setting, the family $\mathscr{F}$ will indeed have a common witness polynomial ψ. In conjunction with the pattern matrix method, we are then able to solve the original problem.

Before reading on, the reader may wish to review the notation and results of Section 2.3, where we discussed the combinatorial complexity measures of Boolean functions. Given function $f\colon \{0,1\}^n \to \mathbb{R}$ and a string $z \in \{0,1\}^n$, we will use the symbol f_z to denote the function defined by $f_z(x) \equiv f(x \oplus z)$.

6.3 Combinatorial preliminaries

In this section, we develop the combinatorial component of our solution. A key combinatorial fact in our analysis is the following consequence of Kenyon and Kutin's Theorem 2.10 on the sensitivity of Boolean functions.

LEMMA 6.3 (Sherstov [201]). *Let $f\colon \{0,1\}^n \to \{-1,+1\}$ be a given function. Then there exists $g\colon \{0,1\}^n \to \{-1,+1\}$ such that*

$$\mathrm{s}(g) \geqslant \alpha \sqrt{\mathrm{bs}(f)} \tag{6.4}$$

for some absolute constant $\alpha > 0$ and

$$g(x) \equiv f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \tag{6.5}$$

for some $i_1, i_2, \ldots, i_n \in \{1,2,\ldots,n\}$.

PROOF. Put $k = \mathrm{bs}(f)$ and fix disjoint sets $S_1, \ldots, S_k \subseteq \{1, 2, \ldots, n\}$ such that one has $f(z \oplus e_{S_1}) = f(z \oplus e_{S_2}) = \cdots = f(z \oplus e_{S_k}) \neq f(z)$ for some $z \in \{0,1\}^n$. Let I be the set of all indices i such the string $z|_{S_i}$ features both zeroes and ones. Put $|I| = r$. For convenience of notation, we will assume that $I = \{1, 2, \ldots, r\}$. For $i = 1, 2, \ldots, r$, form the partition $S_i = A_i \cup B_i$, where

$$A_i = \{j \in S_i : z_j = 0\}, \qquad B_i = \{j \in S_i : z_j = 1\}.$$

Now let

$$g(x) = f\left(\bigoplus_{i=1}^{r} x_{\min A_i} e_{A_i} \oplus \bigoplus_{i=1}^{r} x_{\min B_i} e_{B_i} \oplus \bigoplus_{i=r+1}^{k} x_{\min S_i} e_{S_i} \oplus \bigoplus_{i \notin S_1 \cup \cdots \cup S_k} x_i e_i\right).$$

Then (6.5) is immediate. By the properties of f, we have $\mathrm{bs}_2(g) \geqslant k$, with the blocks $\{\min A_1, \min B_1\}, \ldots, \{\min A_r, \min B_r\}$ and $\{\min S_{r+1}\}, \ldots, \{\min S_k\}$ being sensitive for g on input $x = z$. As a result, Theorem 2.10 implies (6.4). □

6.4 Analytic preliminaries

We now turn to the analytic component of our solution. The main results of this section can all be derived by modifying Razborov's proof of the quantum lower bound for the disjointness function [177]. The alternate derivation presented here seems to have some advantages, however, as we discuss in Remark 6.6. We start by exhibiting a large family of Boolean functions whose inapproximability by low-degree polynomials in the uniform norm can be witnessed by a single, common dual object.

THEOREM 6.4 (Sherstov [201]). *Let $\mathscr{F}$ denote the set of all functions $f : \{0,1\}^n \to \{-1,+1\}$ such that $f(e_1) = f(e_2) = \cdots = f(e_n) \neq f(0) = 1$. Let $\delta > 0$ be a*

suitable absolute constant. Then there exists a function $\psi\colon \{0,1\}^n \to \mathbb{R}$ *such that:*

$$\hat{\psi}(S) = 0, \qquad |S| < \delta\sqrt{n}, \tag{6.6}$$

$$\sum_{x\in\{0,1\}^n} |\psi(x)| = 1, \tag{6.7}$$

$$\sum_{x\in\{0,1\}^n} \psi(x) f(x) > \frac{1}{3}, \qquad f \in \mathscr{F}. \tag{6.8}$$

PROOF. Let p be a univariate real polynomial that satisfies

$$\begin{aligned} &p(0) \in [2/3, 4/3], \\ &p(1) \in [-4/3, -2/3], \\ &p(i) \in [-4/3, 4/3], \qquad\qquad i = 2, 3, \ldots, n. \end{aligned}$$

It follows from basic approximation theory (viz., the inequalities of A. A. Markov and S. N. Bernstein) that any such polynomial p has degree at least $\delta\sqrt{n}$ for an absolute constant $\delta > 0$. See Nisan and Szegedy [160], pp. 308–309, for a short derivation.

By the symmetrization argument (Proposition 2.2), there does not exist a multivariate polynomial $\phi(x_1, \ldots, x_n)$ of degree less than $\delta\sqrt{n}$ such that

$$\begin{aligned} &\phi(0) \in [2/3, 4/3], \\ &\phi(e_i) \in [-4/3, -2/3], \qquad i = 1, 2, \ldots, n, \\ &\phi(x) \in [-4/3, 4/3], \qquad x \in \{0,1\}^n \setminus \{0, e_1, e_2, \ldots, e_n\}. \end{aligned}$$

Linear programming duality now implies the existence of ψ that obeys (6.6), (6.7), and additionally satisfies

$$\psi(0) - \sum_{i=1}^{n} \psi(e_i) - \sum_{\substack{x \in \{0,1\}^n \\ |x| \geqslant 2}} |\psi(x)| > \frac{1}{3},$$

which forces (6.8). □

We are now in a position to prove our main technical criterion for high quantum communication complexity. Our proof is based on the pattern matrix method of Chapters 4 and 5. The novelty of the development below resides in allowing the rows of the given Boolean matrix to derive from distinct Boolean functions, which considerably disrupts the spectral structure. We are able to force the same quantitative conclusion by using the fact that these Boolean functions, albeit distinct, share the relevant dual object. By identifying a set $S \subseteq \{1, 2, \ldots, N\}$ with its characteristic vector $\mathbf{1}_S \in \{0,1\}^N$, it will be convenient in what follows to view the set family $\mathscr{V}(N, n)$ from Section 4.2 as a family of strings in $\{0,1\}^N$.

THEOREM 6.5 (Sherstov [201]; cf. Razborov [177]). *Let $g\colon \{0,1\}^n \to \{-1,+1\}$ be a function such that $g(z \oplus e_1) = g(z \oplus e_2) = \cdots = g(z \oplus e_k) \neq g(z)$ for some $z \in \{0,1\}^n$ with $z_1 = \cdots = z_k = 0$. Then the matrix $G = [g(x \wedge y)]_{x,y \in \{0,1\}^n}$ satisfies*

$$Q^*_{1/3}(G) \geqslant \Omega(\sqrt{k}).$$

REMARK 6.6. As formulated above, Theorem 6.5 can be readily derived by modifying Razborov's proof of the $\Omega(\sqrt{n})$ quantum lower bound for the disjointness function [177, §5.3]. The derivation that we are about to give appears to offer some advantages. First, it is simpler and in particular does not require tools such as Hahn matrices in [177]. Second, it generalizes to any family $\mathscr{F}$ of functions with a common dual polynomial, whereas the method in [177] is restricted to symmetrizable families. Finally, the proof below generalizes to three and more communicating parties, as we will see in Chapter 9.

PROOF OF THEOREM 6.5. Without loss of generality, we assume that k is divisible by 4. Let $\mathscr{F}$ denote the system of all functions $f\colon \{0,1\}^{k/4} \to \{-1,+1\}$ such that $f(e_1) = f(e_2) = \cdots = f(e_{k/4}) \neq f(0) = 1$. By Theorem 6.4, there exists $\psi\colon \{0,1\}^{k/4} \to \mathbb{R}$ such that

$$\hat{\psi}(S) = 0, \qquad |S| < \delta\sqrt{k}, \tag{6.9}$$

$$\sum_{x\in\{0,1\}^{k/4}} |\psi(x)| = 1, \tag{6.10}$$

$$\sum_{x\in\{0,1\}^{k/4}} \psi(x) f(x) > \frac{1}{3}, \qquad f \in \mathscr{F}, \tag{6.11}$$

where $\delta > 0$ is an absolute constant. Now, let Ψ be the $(k/2, k/4, 2^{-3k/4}\psi)$-pattern matrix. It follows from (6.10) that

$$\|\Psi\|_1 = 1. \tag{6.12}$$

By (6.10) and Proposition 2.1,

$$\max_S |\hat{\psi}(S)| \leqslant 2^{-k/4}. \tag{6.13}$$

In view of (6.9) and (6.13), Theorem 4.3 yields

$$\|\Psi\| \leqslant 2^{-\delta\sqrt{k}/2}\, 2^{-k/2}. \tag{6.14}$$

Now, put

$$M = g(z)\left[g\left(z \oplus \bigoplus_{i=1}^{k/2} \{x_i y_{2i-1} e_{2i-1} \oplus \overline{x_i} y_{2i} e_{2i}\} \right) \right]_{x\in\{0,1\}^{k/2},\, y\in\mathscr{V}(k,k/4)},$$

where we identify $\mathscr{V}(k, k/4)$ in the natural way with a subset of $\{0,1\}^k$. Observe that

$$M = \Big[f_{V,w}(x|_V \oplus w) \Big]_{x \in \{0,1\}^{k/2}, (V,w) \in \mathscr{V}(k/2,k/4) \times \{0,1\}^{k/4}}$$

for some functions $f_{V,w} \in \mathscr{F}$. This representation makes it clear, in view of (6.11), that

$$\langle \Psi, M \rangle > \frac{1}{3}. \tag{6.15}$$

By (6.12), (6.14), (6.15) and the generalized discrepancy method (Theorem 5.2), we have $Q^*_{1/3}(M) \geqslant \Omega(\sqrt{k})$. It remains to note that M is a submatrix of $g(z)G$, so that $Q^*_{1/3}(G) \geqslant Q^*_{1/3}(M)$. □

We will also need the following equivalent formulation of Theorem 6.5, for disjunctions instead of conjunctions.

COROLLARY 6.7. *Let $g\colon \{0,1\}^n \to \{-1,+1\}$ be a function such that $g(z \oplus e_1) = g(z \oplus e_2) = \cdots = g(z \oplus e_k) \neq g(z)$ for some $z \in \{0,1\}^n$ with $z_1 = \cdots = z_k = 1$. Then the matrix $G = [g(x \vee y)]_{x,y \in \{0,1\}^n}$ satisfies*

$$Q^*_{1/3}(G) \geqslant \Omega(\sqrt{k}).$$

PROOF. Put $\tilde{g} = g_{(1,\ldots,1)}$ and $\tilde{z} = (1,\ldots,1) \oplus z$. Then $\tilde{z}_1 = \cdots = \tilde{z}_k = 0$ and $\tilde{g}(\tilde{z} \oplus e_1) = \tilde{g}(\tilde{z} \oplus e_2) = \cdots = \tilde{g}(\tilde{z} \oplus e_k) \neq \tilde{g}(\tilde{z})$. By Theorem 6.5, the matrix $\tilde{G} = [\tilde{g}(x \wedge y)]_{x,y \in \{0,1\}^n}$ satisfies $Q^*_{1/3}(\tilde{G}) \geqslant \Omega(\sqrt{k})$. It remains to note that G and $\tilde{G}$ are identical up to a permutation of rows and columns. □

We point out another simple corollary to Theorem 6.5.

COROLLARY 6.8. *Let $f:\{0,1\}^n \to \{-1,+1\}$ be given. Then for some $z \in \{0,1\}^n$, the matrix $F = [f_z(x \wedge y)]_{x,y} = [f(\ldots,(x_i \wedge y_i) \oplus z_i,\ldots)]_{x,y}$ obeys*

$$Q^*_{1/3}(F) = \Omega(\sqrt{\mathrm{bs}(f)}).$$

PROOF. Put $k = \mathrm{bs}(f)$ and fix $z \in \{0,1\}^n$ such that $\mathrm{zbs}(f_z) = k$. By an argument analogous to Lemma 6.3, one obtains a function $g:\{0,1\}^n \to \{-1,+1\}$ such that $g(e_1) = g(e_2) = \cdots = g(e_k) \neq g(0)$ and $g(x) \equiv f_z(\xi_1,\xi_2,\ldots,\xi_n)$ for some symbols $\xi_1,\xi_2,\ldots,\xi_n \in \{x_1,x_2,\ldots,x_n,0,1\}$. Then Theorem 6.5 implies that the matrix $G = [g(x \wedge y)]_{x,y\in\{0,1\}^n}$ satisfies $Q^*_{1/3}(G) \geqslant \Omega(\sqrt{k})$. On the other hand, $Q^*_{1/3}(F) \geqslant Q^*_{1/3}(G)$ by construction. □

6.5 Results on quantum-classical equivalence

We now combine the combinatorial and analytic development of the previous sections to obtain our main results. We start by proving relevant lower bounds against quantum protocols.

THEOREM 6.9 (Sherstov [201]). *Let $f:\{0,1\}^n \to \{-1,+1\}$ be a given function. Put $F_1 = [f(x \wedge y)]_{x,y}$ and $F_2 = [f(x \vee y)]_{x,y}$, where the row and column indices range over $\{0,1\}^n$. Then*

$$\max\{Q^*_{1/3}(F_1), Q^*_{1/3}(F_2)\} = \Omega(\mathrm{bs}(f)^{1/4}).$$

PROOF. By Lemma 6.3, there exists a function $g:\{0,1\}^n \to \{-1,+1\}$ such that

$$\mathrm{s}(g) \geqslant \Omega(\sqrt{\mathrm{bs}(f)}) \tag{6.16}$$

and

$$g(x) \equiv f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \tag{6.17}$$

for some $i_1, i_2, \ldots, i_n \in \{1, 2, \ldots, n\}$. By renumbering the variables if necessary, we see that at least one of the following statements must hold:

(1) $g(z \oplus e_1) = g(z \oplus e_2) = \cdots = g(z \oplus e_{\lceil s(g)/2 \rceil}) \neq g(z)$ for some $z \in \{0,1\}^n$ with $z_1 = z_2 = \cdots = z_{\lceil s(g)/2 \rceil} = 0$;

(2) $g(z \oplus e_1) = g(z \oplus e_2) = \cdots = g(z \oplus e_{\lceil s(g)/2 \rceil}) \neq g(z)$ for some $z \in \{0,1\}^n$ with $z_1 = z_2 = \cdots = z_{\lceil s(g)/2 \rceil} = 1$.

In the former case, Theorem 6.5 implies that the matrix $G_1 = [g(x \wedge y)]_{x,y \in \{0,1\}^n}$ satisfies $Q^*_{1/3}(G_1) \geqslant \Omega(\sqrt{s(g)})$, whence $Q^*_{1/3}(F_1) \geqslant Q^*_{1/3}(G_1) \geqslant \Omega(\mathrm{bs}(f)^{1/4})$ in view of (6.16) and (6.17).

In the latter case, Corollary 6.7 implies that $G_2 = [g(x \vee y)]_{x,y \in \{0,1\}^n}$ satisfies $Q^*_{1/3}(G_2) \geqslant \Omega(\sqrt{s(g)})$, whence $Q^*_{1/3}(F_2) \geqslant Q^*_{1/3}(G_2) \geqslant \Omega(\mathrm{bs}(f)^{1/4})$ in view of (6.16) and (6.17). □

Having obtained the desired lower bounds on quantum communication, we now turn to classical protocols. The argument here is closely analogous to that of Proposition 4.10.

THEOREM 6.10. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be given. Put $F_1 = [f(x \wedge y)]_{x,y}$ and $F_2 = [f(x \vee y)]_{x,y}$, where the row and column indices range over $\{0,1\}^n$. Then*

$$\max\{D(F_1), D(F_2)\} \leqslant 2\,\mathrm{dt}(f) \leqslant 2\,\mathrm{bs}(f)^3.$$

PROOF (adapted from [51, 28]). The second inequality follows immediately by Theorem 2.8, so we will focus on the first. Fix an optimal-depth decision tree for f. The protocol for F_1 is as follows. On input x and y, Alice and Bob start at the top node of the tree, read its label i, and exchange the two bits x_i and y_i. This allows them to compute $x_i \wedge y_i$ and to determine which branch to take next. The process repeats at the new node and so on, until the parties have reached a leaf node. Since the longest root-to-leaf path has length $\mathrm{dt}(f)$, the claim follows. The proof for F_2 is entirely analogous. □

Theorems 6.9 and 6.10 immediately imply our main result on quantum-classical equivalence, stated above as Theorem 6.1.

6.6 Applications to the log-rank conjecture

An important result due to Mehlhorn and Schmidt [152] shows that the rank of a communication matrix gives a lower bound on its deterministic communication complexity:

THEOREM 6.11 (Mehlhorn and Schmidt [152]). *Let $f: X \times Y \to \{-1, +1\}$ be a given function, where X, Y are finite sets. Put $F = [f(x, y)]_{x \in X,\, y \in Y}$. Then*

$$D(f) \geqslant \log_2 \operatorname{rk} F.$$

The well-known *log-rank conjecture* of Lovász and Saks [147, 148] states that the lower bound in Theorem 6.11 is tight up to a polynomial factor, i.e., the deterministic communication complexity $D(F)$ of every sign matrix satisfies $D(F) \leqslant (\log_2 \operatorname{rk} F)^c + c$ for some absolute constant $c > 0$. In this section we will prove that, in addition to having polynomially related quantum and classical communication complexities, the problem of computing $f(x \wedge y)$ and $f(x \vee y)$ satisfies the log-rank conjecture.

As an illustrative starting point, we first settle the log-rank conjecture for every pattern matrix.

THEOREM 6.12 (Sherstov [203]). *Let $f: \{0,1\}^t \to \{-1, +1\}$ be a given function, $d = \deg(f)$. Let F be the (n, t, f)-pattern matrix. Then*

$$\operatorname{rk} F \geqslant \left(\frac{n}{t}\right)^d \geqslant \exp\{\Omega(D(F)^{1/4})\}. \tag{6.18}$$

In particular, F satisfies the log-rank conjecture.

PROOF. Since $\hat{f}(S) \neq 0$ for some set S with $|S| = d$, Theorem 4.3 implies that F has at least $(n/t)^d$ nonzero singular values. This settles the first inequality in (6.18).

Proposition 4.10 implies that $D(F) \leqslant O(\operatorname{dt}(f) \log(n/t))$. Since $\operatorname{dt}(f) \leqslant 2 \deg(f)^4$ by Theorem 2.6, we obtain the second inequality in (6.18). □

The remainder of this section is based on the following result of Buhrman and de Wolf [54], who studied the special case of symmetric functions f in the same context.

THEOREM 6.13 (Buhrman and de Wolf [54]). *Let $f\colon\{0,1\}^n \to \mathbb{R}$ be given. Put $M = [f(x \wedge y)]_{x,y}$, where the row and column indices range over $\{0,1\}^n$. Then*

$$\operatorname{rk} M = \operatorname{mon}(f).$$

Our first observation is as follows.

LEMMA 6.14 (Sherstov [201]). *Let $f\colon\{0,1\}^n \to \mathbb{R}$ be given, where $f \not\equiv 0$ and $d = \deg(f)$. Then for some $z \in \{0,1\}^n$,*

$$\operatorname{mon}(f_z) \geqslant \left(\frac{3}{2}\right)^d.$$

PROOF. The proof is by induction on d. The base case $d = 0$ holds since $f \not\equiv 0$. Assume that the claim holds for all f of degree $d-1$. By renumbering the variables if necessary, we have $f(x) = x_1 p(x_2,\ldots,x_n) + q(x_2,\ldots,x_n)$ for some polynomial p of degree $d-1$. The inductive assumption guarantees the existence of $u \in \{0,1\}^{n-1}$ such that $\operatorname{mon}(p_u) \geqslant (3/2)^{d-1}$. Note that $\operatorname{mon}(f_{(0,u)}) = \operatorname{mon}(p_u) + \operatorname{mon}(q_u)$ and $\operatorname{mon}(f_{(1,u)}) \geqslant \operatorname{mon}(p_u) + |\operatorname{mon}(q_u) - \operatorname{mon}(p_u)|$. Thus,

$$\max\{\operatorname{mon}(f_{(0,u)}), \operatorname{mon}(f_{(1,u)})\} \geqslant \frac{3}{2}\operatorname{mon}(p_u) \geqslant \left(\frac{3}{2}\right)^d. \qquad \square$$

We will also need the following technical lemma.

LEMMA 6.15 (Sherstov [201]). *Let* $f:\{0,1\}^n \to \mathbb{R}$ *be given. Fix an index* $i = 1,2,\ldots,n$. *Define*

$$\tilde{f}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = f(x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_n).$$

Then

$$\max\{\mathrm{mon}(\tilde{f}),\ \mathrm{mon}(f_{e_i})\} \geqslant \frac{1}{2}\,\mathrm{mon}(f).$$

PROOF. Write

$$f(x) = x_i\, p(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) + \tilde{f}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n).$$

Then $\mathrm{mon}(\tilde{f}) + \mathrm{mon}(f_{e_i}) \geqslant \mathrm{mon}(\tilde{f}) + \mathrm{mon}(p) + |\mathrm{mon}(p) - \mathrm{mon}(\tilde{f})| \geqslant \mathrm{mon}(f)$, as desired. □

At last, we arrive at the main result of this section.

THEOREM 6.16 (Sherstov [201]). *Let* $f:\{0,1\}^n \to \{-1,+1\}$ *be given,* $d = \deg(f)$. *Put* $F_1 = [f(x \wedge y)]_{x,y}$ *and* $F_2 = [f(x \vee y)]_{x,y}$, *where the row and column indices range over* $\{0,1\}^n$. *Then*

$$\max\{\mathrm{rk}\,F_1,\ \mathrm{rk}\,F_2\} \geqslant \left(\frac{3}{2\sqrt{2}}\right)^d \geqslant 1.06^d. \tag{6.19}$$

In particular, the communication problem of computing, on input $x, y \in \{0,1\}^n$, *both of the quantities* $f(x \wedge y)$ *and* $f(x \vee y)$, *satisfies the log-rank conjecture.*

PROOF. To see how the last statement follows from the lower bound (6.19), note that $\max\{D(F_1), D(F_2)\} \leqslant 2\,\mathrm{dt}(f)$ by Theorem 6.10 and recall that $\mathrm{dt}(f) \leqslant 2\deg(f)^4$ by Theorem 2.6. In the remainder of the proof, we focus on (6.19) alone.

We assume that $d \geqslant 1$, the claim being trivial otherwise. By renumbering the variables if necessary, we may write

$$f(x) = \alpha x_1 x_2 \cdots x_d + \sum_{S \neq \{1,\ldots,d\}} \alpha_S \prod_{i \in S} x_i,$$

where $\alpha \neq 0$. Define $g(x_1, \ldots, x_d) = f(x_1, \ldots, x_d, 0, \ldots, 0)$. Then g is a nonzero polynomial of degree d, and Lemma 6.14 yields a vector $z \in \{0,1\}^d$ such that

$$\operatorname{mon}(g_z) \geqslant \left(\frac{3}{2}\right)^d.$$

By renumbering the variables if necessary, we may assume that $z = 0^t 1^{d-t}$. We complete the proof by analyzing the cases $t \leqslant d/2$ and $t > d/2$.

We will first consider the case when $t \leqslant d/2$. Let $\mathscr{F}$ stand for the set whose elements are the identity function on $\{0,1\}$ and the constant-one function on $\{0,1\}$. Lemma 6.15 provides functions $\phi_1, \ldots, \phi_t \in \mathscr{F}$ such that the polynomial $h(x_1, \ldots, x_d) = g_{1^d}(\phi_1(x_1), \ldots, \phi_t(x_t), x_{t+1}, \ldots, x_d)$ features at least $2^{-t} \operatorname{mon}(g_z) \geqslant (3/\{2\sqrt{2}\})^d$ monomials. By Theorem 6.13, the matrix $H = [h(x \wedge y)]_{x,y \in \{0,1\}^d}$ has rank at least $(3/\{2\sqrt{2}\})^d$. Since H is a submatrix of F_2, the theorem holds in this case.

The case $t > d/2$ is entirely symmetric, with F_1 playing the role of F_2. □

REMARK. By the results of Buhrman and de Wolf [54], Theorem 6.16 alone would suffice to obtain a polynomial relationship between classical and quantum communication complexity in the *exact* model. However, for our main result we need a polynomial relationship in the *bounded-error* model, which requires the full development of Sections 6.3–6.5.

6.7 Generalizations for arbitrary composed functions

Up to this point, we have focused on the communication problem of computing $f(x \wedge y)$ and $f(x \vee y)$. Here we point out that our results on quantum-classical

equivalence and the log-rank conjecture immediately apply to a broader class of communication problems. Specifically, we will consider compositions of the form $f(\dots, g_i(x^{(i)}, y^{(i)}), \dots)$, where one has a combining function $f\colon \{0,1\}^n \to \{-1,+1\}$ that receives input from intermediate functions $g_i\colon X_i \times Y_i \to \{0,1\}$, $i = 1, 2, \dots, n$. We will show that under natural assumptions on $g_1, \dots, g_n$, this composed function will have polynomially related quantum and classical bounded-error complexities and will satisfy the log-rank conjecture.

THEOREM 6.17 (Sherstov [201]). *Let $f\colon \{0,1\}^n \to \{-1,+1\}$ be a given function. Fix functions $g_i\colon X_i \times Y_i \to \{0,1\}$, for $i = 1, 2, \dots, n$. Assume that for each i, the matrix $[g_i(x^{(i)}, y^{(i)})]_{x^{(i)} \in X_i, y^{(i)} \in Y_i}$ contains the following submatrices*

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \tag{6.20}$$

up to a permutation of rows and columns. Put $F = [f(\dots, g_i(x^{(i)}, y^{(i)}), \dots)]$. Assume that for some constant $\alpha > 0$,

$$Q^*_{1/3}(g_i) \geqslant R_{1/3}(g_i)^\alpha, \qquad i = 1, 2, \dots, n. \tag{6.21}$$

Then for some constant $\beta = \beta(\alpha) > 0$,

$$R_{1/3}(F) \geqslant Q^*_{1/3}(F) \geqslant R_{1/3}(F)^\beta.$$

PROOF. Without loss of generality, we may assume that f depends on all of its n inputs (otherwise, disregard any irrelevant inputs from among $g_1, \dots, g_n$ in the analysis below). In particular, we have

$$Q^*_{1/3}(F) \geqslant Q^*_{1/3}(g_i), \qquad i = 1, 2, \dots, n. \tag{6.22}$$

Since each g_i contains the two-variable functions AND and OR as subfunctions, Corollary 6.8 shows that

$$Q^*_{1/3}(F) \geqslant \Omega(\sqrt{\mathrm{bs}(f)}). \tag{6.23}$$

Letting $d = \mathrm{dt}(f)$, we claim that

$$R_{1/3}(F) \leqslant O(d \log d) \max_{i=1,\dots,n} \{R_{1/3}(g_i)\}. \tag{6.24}$$

The proof of this bound is closely analogous to that of Theorem 6.10. Namely, Alice and Bob evaluate a depth-d decision tree for f. When a tree node calls for the ith variable, the parties run an optimal randomized protocol for g_i with error probability $\frac{1}{3d}$, which requires at most $O(R_{1/3}(g_i) \log d)$ bits of communication. Since all root-to-leaf paths have length at most d, the final answer will be correct with probability at least $2/3$.

In view of Theorem 2.8, the sought polynomial relationship between $R_{1/3}(F)$ and $Q^*_{1/3}(F)$ follows from (6.21)–(6.24). □

We now record an analogous result for the log-rank conjecture.

THEOREM 6.18 (Sherstov [201]). *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a given function. Fix functions $g_i : X_i \times Y_i \to \{0,1\}$, for $i = 1,2,\dots,n$. Assume that for each i, the matrix $[g_i(x^{(i)}, y^{(i)})]_{x^{(i)} \in X_i, y^{(i)} \in Y_i}$ contains (6.20) as submatrices, up to a permutation of rows and columns. Put $F = [f(\dots, g_i(x^{(i)}, y^{(i)}), \dots)]$. Assume that for some constant $c > 0$,*

$$D(g_i) \leqslant (\log_2 \mathrm{rk}\, G_i)^c, \qquad i = 1,2,\dots,n, \tag{6.25}$$

where $G_i = [(-1)^{g_i(x^{(i)}, y^{(i)})}]_{x^{(i)} \in X_i, y^{(i)} \in Y_i}$. Then for some constant $C = C(c) > 0$,

$$D(F) \leqslant (\log_2 \mathrm{rk}\, F)^C.$$

In particular, F satisfies the log-rank conjecture.

PROOF. Without loss of generality, we may assume that f depends on all of its n inputs (otherwise, disregard any irrelevant inputs from among $g_1, \ldots, g_n$ in the analysis below). In particular, we have

$$\operatorname{rk} F \geqslant \operatorname{rk} G_i, \qquad i = 1, 2, \ldots, n. \tag{6.26}$$

Since each g_i contains the two-variable functions AND and OR as subfunctions, Theorem 6.16 shows that

$$\operatorname{rk} F \geqslant \left(\frac{3}{2\sqrt{2}}\right)^{\deg(f)}. \tag{6.27}$$

Finally, we claim that

$$D(F) \leqslant 2 \operatorname{dt}(f) \max_{i=1,\ldots,n} \{D(g_i)\}. \tag{6.28}$$

The proof of this bound is closely analogous to that of Theorem 6.10. Namely, Alice and Bob evaluate an optimal-depth decision tree for f. When a tree node calls for the ith variable, the parties run an optimal deterministic protocol for g_i.

In view of (6.25)–(6.28) and Theorem 2.6, the proof is complete. □

The key property of $g_1, \ldots, g_n$ that we have used in this section is that their communication matrices contain (6.20) as submatrices. We close this section by observing that this property almost always holds. More precisely, we show that matrices that do not contain the submatrices (6.20) have a very restricted structure.

THEOREM 6.19 (Sherstov [201]). *A matrix $G \in \{0,1\}^{N \times M}$ does not contain*

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

as a submatrix if and only if $G = 0$, $G = J$, *or*

$$G' \sim \begin{bmatrix} J_1 & & & & \\ & J_2 & & 0 & \\ & & J_3 & & \\ & & & \ddots & \\ & 0 & & & J_k \end{bmatrix}, \tag{6.29}$$

where: G' *is the result of deleting any columns and rows in* G *that consist entirely of zeroes;* $J, J_1, J_2, \ldots, J_k$ *are all-*1 *matrices of appropriate dimensions; and* $\sim$ *denotes equality up to a permutation of rows and columns.*

PROOF. The "if" part is clear. We will prove the other direction by induction on the number of columns, M. The base case is trivial. For the inductive step, let $G \neq 0$ be a given matrix. Let J_1 be a maximal submatrix of G with all entries equal to 1. Then

$$G \sim \begin{bmatrix} J_1 & Z_1 \\ Z_2 & H \end{bmatrix}$$

for suitable matrices Z_1, Z_2, and H, possibly empty. By the maximality of J_1 and the fact that G does not contain A as a submatrix, it follows that either Z_1 is empty or $Z_1 = 0$. Likewise for Z_2. By the inductive hypothesis for H, the proof is complete. □

By reversing the roles of 0 and 1, one obtains from Theorem 6.19 an analogous characterization of all matrices $G = \{0,1\}^{N \times M}$ that do not contain

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

as a submatrix.

REMARK 6.20. The communication complexity of a Boolean matrix remains unaffected if one modifies it to retain only one copy of each column, removing any

duplicates. An analogous statement holds for the rows. In light of Theorem 6.19, this means that there are only four types of intermediate functions g for which our composition results (Theorem 6.17 and 6.18) fail. These are the functions g with matrix representations

$$I, \qquad \begin{bmatrix} I & \\ & 0 \end{bmatrix}, \qquad \begin{bmatrix} I \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} I & 0 \end{bmatrix}, \tag{6.30}$$

and their negations, where I is the identity matrix. The reason that Theorems 6.17 and 6.18 fail for such g is that the underlying quantum lower bound in terms of block sensitivity of the combining function f is no longer valid. For example, the first matrix type, I, corresponds to letting g be the equality function. Now, the conjunction of n equality functions is still an equality function, and its communication complexity is $O(1)$ both in the randomized and quantum models [137], which is much less than a hypothetical lower bound of $\Omega(\sqrt{n})$ that one would expect from the block sensitivity of $f = \mathrm{AND}_n$. The same $O(1)$ upper bound holds for a conjunction of arbitrarily many functions g of the second, third, and fourth type.

Chapter 7

Unbounded-Error Communication

This chapter focuses on the *unbounded-error* model, a deep and fascinating model in communication complexity with applications to circuit complexity, matrix analysis, and learning theory. Our main result here is a near-tight lower bound on the unbounded-error communication complexity of every symmetric function, i.e., every function of the form $f(x, y) = D(\sum x_i y_i)$ for some predicate $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$. The pattern matrix method of the previous chapters will continue to play an important role but will need to be complemented with other results on random walks, matrix analysis, and approximation theory.

7.1 Introduction and statement of results

The *unbounded-error* model, due to Paturi and Simon [167], is a rich and elegant model of communication. Fix a function $f: X \times Y \to \{-1, +1\}$, where X and Y are some finite sets. Alice receives an input $x \in X$, Bob receives $y \in Y$, and their objective is to compute $f(x, y)$. To this end, they exchange classical bits 0 and 1 through a shared communication channel according to a protocol established in advance. Alice and Bob each have an unlimited *private* source of random bits which they can use in deciding what messages to send. Eventually, Bob concludes this process by sending Alice a single bit, which is taken to be the output of their joint computation. Let the random variable $\Pi(x, y) \in \{-1, +1\}$ denote the protocol output when the parties receive inputs $x \in X$ and $y \in Y$. Alice and Bob's protocol is said to *compute* f if

$$\mathbf{P}[\Pi(x, y) = f(x, y)] > \frac{1}{2}$$

for all $x \in X$, $y \in Y$. The *cost* of a given protocol is the worst-case number of bits exchanged on any input (x, y). The unbounded-error communication complexity of f, denoted $U(f)$, is the least cost of a protocol that computes f.

The unbounded-error model occupies a special place in communication complexity because it is more powerful than any of the usual models, including deterministic, nondeterministic, bounded-error randomized, and bounded-error quantum with or without entanglement. Furthermore, the unbounded-error model has applications to matrix analysis, circuit complexity, and learning theory that the other

models cannot have. We defer a thorough discussion of the unbounded-error model, its applications, and quantitative comparisons with other models to Section 7.2.

Despite the many applications and intrinsic appeal of the unbounded-error model, progress in understanding it has been slow and difficult. Indeed, we are aware of only a few nontrivial results on this subject. Alon et al. [13] obtained strong lower bounds for random functions. No nontrivial lower bounds were available for any explicit functions until the breakthrough work of Forster [70], who proved strong lower bounds for the inner product function and, more generally, any function whose communication matrix has low spectral norm. Several extensions and refinements of Forster's method were proposed in subsequent work [71, 73].

This chapter focuses on symmetric functions, i.e., functions $f\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ of the form $f(x,y) = D(\sum x_i y_i)$ for a given predicate $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$. Symmetric functions are a well-studied class in communication complexity, as the reader may recall from Chapter 5 and earlier literature [102, 176, 177, 171, 27, 118]. Our main result is to settle the unbounded-error communication complexity of every such function, to within logarithmic factors. Since the unbounded-error model more powerful than the other models, earlier lower bounds for symmetric functions [102, 176, 177] are irrelevant to this project. The only symmetric function whose unbounded-error complexity was known prior to our work was inner product function $\mathrm{IP}_n(x,y) = \bigoplus_{i=1}^n (x_i \wedge y_i)$, for which Forster [70] proved a tight lower bound of $\Omega(n)$. The general result that we prove is in terms of the *degree* $\deg(D)$ of a given predicate D, defined as the number of times D changes value in $\{0,1,\ldots,n\}$. In other words, $\deg(D) = |\{i : D(i) \neq D(i-1)\}|$.

THEOREM 7.1 (Sherstov [204]). *Let* $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$ *be given,* $k = \deg(D)$. *Define* $f(x,y) = D(\sum x_i y_i)$. *Then*

$$\Theta(k/\log^5 n) \leqslant U(f) \leqslant \Theta(k \log n).$$

The upper bound in this result has a short and elementary demonstration. This chapter is devoted, then, almost entirely to the proof of the lower bound. We will give an intuitive overview of our proof in Section 7.3, after a more comprehensive discussion of the unbounded-error model in Section 7.2.

7.2 Unbounded-error model of communication

The unbounded-error model differs from the bounded-error randomized model of Chapter 4 in two vital ways. First, a bounded-error protocol must provide the correct output with probability $\frac{1}{2} + \Omega(1)$ on each input, whereas the correctness probability of an unbounded-error protocol need merely exceed $\frac{1}{2}$, possibly by an exponentially small amount. Second, the parties in the unbounded-error model no longer have access to a shared source of random bits. Indeed, allowing such a source would result in a trivial model, since every function $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ can be represented in the form

$$f(x,y) \equiv \operatorname{sgn}\left(\sum_{S,T \subseteq \{1,2,\ldots,n\}} \alpha_{S,T} \chi_S(x) \chi_T(y)\right)$$

for some reals $\alpha_{S,T}$ and therefore admits a trivial public-coin protocol with cost $O(1)$ and error probability strictly smaller than $\frac{1}{2}$. By contrast, a well-known result due to Newman [157] shows that shared randomness has essentially no effect on the bounded-error randomized complexity of any given function.

The unbounded-error model occupies a special place in the study of communication because it is more powerful than any of the usual models, including deterministic, nondeterministic, bounded-error randomized, and bounded-error quantum with or without entanglement. Frequently, the unbounded-error communication complexity of a function is exponentially smaller than its complexity in other models. For example, the disjointness function $\mathrm{DISJ}_n(x,y) = \bigvee_{i=1}^n (x_i \wedge y_i)$ has cost $\Theta(n)$ in the bounded-error randomized model [102, 176], cost $\Theta(\sqrt{n})$ in the bounded-error quantum model [177, 3], and cost $\Theta(\log n)$ in the unbounded-error model (Proposition 7.23 below). Furthermore, in Chapter 10 we will construct functions that have unbounded-error complexity $O(\log n)$ but require $\Omega(\sqrt{n})$ communication in the randomized and quantum models to even achieve advantage $2^{-\sqrt{n}/5}$ over random guessing.

The additional power of the unbounded-error model has a consequence that proving communication lower bounds in it requires richer mathematical machinery. Furthermore, the resulting lower bounds will have applications that other communication models could not have. We will now examine several such applications.

Sign-rank and rigidity. A compelling aspect of the unbounded-error model is that it has an exact matrix-analytic formulation. Recall from Chapter 2 that the sign-rank of a sign matrix $F = [F_{ij}]$ is the least rank of a real matrix $A = [A_{ij}]$ with $F_{ij} A_{ij} > 0$ for all i, j. In other words, sign-rank measures the sensitivity of the rank of F when its entries undergo sign-preserving perturbations. Sensitivity of the rank is an important and deep subject in complexity theory. For example, much work has focused on the closely related concept of *matrix rigidity* [103, 146]. Surprisingly, the notion of sign-rank is equivalent to unbounded-error communication complexity.

THEOREM 7.2 (Paturi and Simon [167, Thm. 2]). *Fix finite sets X, Y and a function $f: X \times Y \to \{-1, +1\}$. Put $F = [f(x, y)]_{x \in X, y \in Y}$. Then*

$$U(f) \;=\; \log \operatorname{rk}_{\pm} F \;\pm\; O(1).$$

In words, the unbounded-error complexity of a function essentially equals the logarithm of the sign-rank of its communication matrix. Thus, unbounded-error communication complexity embodies a fundamental question from matrix analysis, with close ties to computational complexity.

Circuit complexity. Recall from Chapter 2 that a linear threshold gate g with Boolean inputs $x_1, \ldots, x_n$ is a function of the form $g(x) = \operatorname{sgn}(\sum a_i x_i - \theta)$ for some fixed weights $a_1, \ldots, a_n, \theta$. Thus, a threshold gate generalizes the familiar majority gate. A major unsolved problem [132] in computational complexity is to exhibit a Boolean function that requires a depth-2 threshold circuit of superpolynomial size. Communication complexity has been crucial to the progress on this problem. Via randomized communication complexity and discrepancy, many explicit functions have been found that require *majority-of-threshold* circuits of exponential size, as the reader may recall from Chapter 4 and earlier papers [88, 82, 159, 196]. This solves a special case of the general problem. The unbounded-error model, or equivalently sign-rank, solves another case [71]: it supplies exponential lower bounds against *threshold-of-majority circuits,* i.e., circuits with a linear threshold gate at the top that receives inputs from majority gates. More precisely, Forster et al. [71] proved the following result.

THEOREM 7.3 (Forster et al. [71, Lem. 5]). *Let* $f\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ *be a given function. Suppose that* $f(x,y) \equiv \operatorname{sgn}(\sum_{i=1}^{s} \lambda_i g_i(x,y))$, *where each* g_i *is a linear threshold gate with integer weights bounded by* W *in absolute value. Then for* $F = [f(x,y)]_{x,y\in\{0,1\}^n}$,

$$s \geqslant \Omega\left(\frac{\operatorname{rk}_{\pm} F}{nW}\right).$$

Computational learning theory. We close this overview with an application of sign-rank and unbounded-error communication complexity to computational learning theory, a subject treated in Part II of this thesis. Computational learning theory seeks to approximately reconstruct an unknown function $f\colon \{0,1\}^n \to \{-1,+1\}$ based on its membership in a given set (the *concept class*) as well as its values on a small sample of points on the hypercube $\{0,1\}^n$.

In a seminal paper [213], Valiant formulated the *probably approximately correct* (PAC) model of learning, now a central model in computational learning theory. Research has shown that PAC learning is surprisingly difficult. (By "PAC learning," we shall always mean PAC learning under arbitrary distributions.) Indeed, the learning problem remains unsolved for such natural concept classes as DNF formulas of polynomial size and intersections of two halfspaces, whereas hardness results and lower bounds are abundant [105, 112, 127, 68, 128, 126]. One concept class for which efficient PAC learning algorithms are available is the class of halfspaces, i.e., functions $f\colon \mathbb{R}^n \to \{-1,+1\}$ representable as $f(x) \equiv \operatorname{sgn}(\sum a_i x_i - \theta)$ for some reals $a_1,\dots,a_n,\theta$. Halfspaces constitute one of the most studied classes in computational learning theory [186, 161, 153, 36] and a major success story of the field. Indeed, a significant part of computational learning theory attempts to learn rich concept classes by reducing them to halfspaces. The reduction works as follows. Let $\mathscr{C}$ be a given concept class, i.e., a set of Boolean functions $\{0,1\}^n \to \{-1,+1\}$. One seeks functions $\phi_1,\dots,\phi_r\colon \{0,1\}^n \to \mathbb{R}$ such that every $f \in \mathscr{C}$ has a representation

$$f(x) \equiv \operatorname{sgn}(a_1\phi_1(x) + \cdots + a_r\phi_r(x))$$

for some reals $a_1, \ldots, a_r$. This process is technically described as *embedding* $\mathscr{C}$ *in halfspaces of dimension* r. Once this is accomplished, $\mathscr{C}$ can be learned in time polynomial in n and r by any halfspace-learning algorithm.

For this approach to be practical, the number r of real functions needs to be reasonable (ideally, polynomial in n). It is thus of interest to determine what natural concept classes can be embedded in halfspaces of low dimension [35, 126]. The smallest dimension of such a representation is called the *sign-rank* of a given class. Formally, the sign-rank $\mathrm{rk}_{\pm}\,\mathscr{C}$ of a given class $\mathscr{C}$ of functions $\{0,1\}^n \to \{-1,+1\}$ is the least r for which there exist real functions $\phi_1, \ldots, \phi_r \colon \{0,1\}^n \to \mathbb{R}$ such that every $f \in \mathscr{C}$ is expressible as $f(x) \equiv \mathrm{sgn}(\sum a_i \phi_i(x))$ for some reals $a_1, \ldots, a_r$. To relate this discussion to the sign-rank of matrices, let $M_{\mathscr{C}} = [f(x)]_{f \in \mathscr{C},\, x \in \{0,1\}^n}$ be the characteristic matrix of $\mathscr{C}$. A moment's reflection reveals that

$$\mathrm{rk}_{\pm}\,\mathscr{C} = \mathrm{rk}_{\pm}\,M_{\mathscr{C}},$$

i.e., the sign-rank of a concept class is precisely the sign-rank of its characteristic matrix.

In summary, the study of sign-rank, or equivalently unbounded-error communication complexity, yields nontrivial PAC learning algorithms. In particular, the current fastest algorithm for learning polynomial-size DNF formulas, due to Klivans and Servedio [122], was obtained precisely by placing an upper bound of $2^{\tilde{O}(n^{1/3})}$ on the sign-rank of that concept class, with the functions ϕ_i corresponding to the monomials of degree up to $\tilde{O}(n^{1/3})$.

7.3 Overview of the proof

Our proof of Theorem 7.1 consists of two independent parts. First, we reduce the original problem to analyzing what we call *dense* predicates. We attach this term to the predicates $D \colon \{0,1,\ldots,n\} \to \{-1,+1\}$ that change value frequently and at roughly regular intervals. Dense predicates are highly structured and amenable to direct analysis, unlike general predicates. With this reduction in hand, we complete the proof by solving the problem for every dense predicate. We now describe the two technical components in greater detail.

Let $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a given predicate that is not dense. Any communication protocol that computes D can clearly compute the restriction of D to a given subinterval $\{i, i+1, \ldots, j\} \subseteq \{0, 1, \ldots, n\}$. Now, let $\mathscr{D}$ denote the set of all restrictions of D to subintervals of a given length. Using a probabilistic argument, we show that a dense predicate arises as the XOR of a small number T of predicates from $\mathscr{D}$, where T depends on the degree of D. As a result, if the original predicate D has unbounded-error complexity $\ll \deg(D)$, then some dense predicate will have disproportionately small unbounded-error complexity. This is the desired reduction. The technical challenge here is to show that a dense predicate can be obtained as the XOR of a *small* number of predicates from $\mathscr{D}$. To this end, we model the probabilistic argument as a random walk on $\mathbb{Z}_2^n$ and place a strong upper bound on its mixing time. Our analysis uses a known bound, due to Razborov [174], on the rate of convergence in terms of the probability of a basis for $\mathbb{Z}_2^n$.

It remains to describe our solution for dense predicates. Using Chebyshev polynomials and the Markov-Bernstein inequalities, Paturi [165] determined the least degree of a polynomial that approximates any given Boolean predicate on $\{0, 1, \ldots, n\}$ pointwise to within $1/3$. A starting point in our analysis of dense predicates is a related approximation problem, in which the nodes are no longer $\{0, 1, \ldots, n\}$ but are some arbitrary reals $\{\xi_1, \xi_2, \ldots, \xi_n\} \subset [0, n]$. Provided that the nodes are not too clumped together, we are able to prove strong lower bounds on the degree for a relevant class of approximation problems $f: \{\xi_1, \xi_2, \ldots, \xi_n\} \to \{0, 1\}$. Paturi's proof technique does not apply in this more general setting, and we give a direct analysis using fundamentals of approximation theory.

The crucial next step is to show that computation of dense predicates corresponds to the approximation problem just described, where the real nodes $\xi_1, \xi_2, \ldots, \xi_n$ are allowed to form clusters but must still cover much of the interval $[0, n]$. Linear-programming duality now tells us that, in a well-defined technical sense, a dense predicate behaves much like the parity function with respect to a smooth distribution on the inputs. This enables us to bound the spectral norm of relevant matrices using the pattern matrix method of Chapter 4. In a final step, we invoke Forster's generalized theorem [71] to obtain our main result.

7.4 Technical preliminaries

At several places in this chapter, it will be important to distinguish between addition over the reals and addition over GF(2). To avoid any confusion, we reserve the operator $+$ for the former and $\oplus$ for the latter. We will need the following fact about random walks on $\mathbb{Z}_2^n$.

PROPOSITION 7.4 (Folklore). *For an integer $T \geqslant 1$, let $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_T \in \{0,1\}$ be independent random variables, each taking on 1 with probability p. Then*

$$\mathbf{E}\Big[\mathbf{b}_1 \oplus \mathbf{b}_2 \oplus \cdots \oplus \mathbf{b}_T\Big] = \frac{1}{2} - \frac{1}{2}(1-2p)^T.$$

PROOF. Straightforward by induction on T. □

Fix a predicate $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$. We say that a *value change* occurs at index $t \in \{1,2,\ldots,n\}$ if $D(t) \neq D(t-1)$. The *degree* of D is defined by $\deg(D) = |\{t : D(t) \neq D(t-1)\}|$. To illustrate, the familiar predicate $\mathrm{PARITY}_n(t) = (-1)^t$ has degree n, whereas a constant predicate has degree 0. It is clear that $\deg(D)$ is the least degree of a real univariate polynomial p such that $D(t) = \operatorname{sgn} p(t)$, $t = 0,1,\ldots,n$, hence the term *degree*. Given two predicates $D_1, D_2\colon \{0,1,\ldots,n\} \to \{-1,+1\}$, recall that their XOR is the predicate $D_1 \oplus D_2\colon \{0,1,\ldots,n\} \to \{-1,+1\}$ defined by $(D_1 \oplus D_2)(t) = D_1(t)D_2(t)$.

In a breakthrough result, Forster [70] proved the first strong lower bound in the unbounded-error model for the inner product function and more generally any function whose communication matrix has low spectral norm. In view of Theorem 7.2, Forster's result admits equivalent formulations in terms of sign-rank and unbounded-error communication complexity.

THEOREM 7.5 (Forster [70]). *Let X, Y be finite sets and $M = [M_{xy}]_{x\in X, y\in Y}$ a sign matrix. Then*

$$\mathrm{rk}_{\pm} A \geqslant \frac{\sqrt{|X|\,|Y|}}{\|M\|}.$$

In particular, the inner product matrix $H = [(-1)^{\sum x_i y_i}]_{x,y \in \{0,1\}^n}$ *satisfies*

$$U(H) = \log \operatorname{rk}_{\pm} H \pm O(1) \geqslant \frac{1}{2}n - O(1).$$

Forster's proof generalizes to yield the following result, which serves as a crucial starting point for our work.

THEOREM 7.6 (Forster et al. [71, Thm. 3]). *Let* X, Y *be finite sets and* $M = [M_{xy}]_{x \in X, y \in Y}$ *a real matrix without zero entries. Then*

$$\operatorname{rk}_{\pm} M \geqslant \frac{\sqrt{|X|\,|Y|}}{\|M\|} \min_{x,y} |M_{xy}|.$$

Given functions $f, g \colon X \times Y \to \{-1, +1\}$, recall that their XOR is the function $f \oplus g \colon X \times Y \to \{-1, +1\}$ defined by $(f \oplus g)(x, y) = f(x, y)g(x, y)$. We have:

PROPOSITION 7.7 (Folklore). *Let* $f, g \colon X \times Y \to \{-1, +1\}$ *be arbitrary. Then*

$$U(f \oplus g) \leqslant U(f) + U(g).$$

PROOF. Alice and Bob can run a separate protocol for f and g and output the XOR of the two answers. It is straightforward to verify that this strategy is correct with probability greater than $1/2$. □

In this chapter, we will be primarily interested in the communication complexity of predicates $D \colon \{0, 1, \ldots, n\} \to \{-1, +1\}$. Specifically, we define $U(D)$ to be the unbounded-error communication complexity of the function $f \colon \{0,1\}^n \times \{0,1\}^n \to \{-1, +1\}$ given by $f(x, y) = D(\sum x_i y_i)$. In specifying matrices, we will use the symbol $*$ for entries whose values are irrelevant, as in the proofs of Lemmas 7.9 and 7.12. As a final convention, we will use two distinct versions of the XOR operator depending on the domain of the arguments, namely,

$\oplus\colon\{0,1\}\to\{0,1\}$ and $\oplus\colon\{-1,+1\}\to\{-1,+1\}$. In other words, we assume that $\oplus$ ranges in $\{0,1\}$ whenever its arguments take values in $\{0,1\}$, and likewise we assume that $\oplus$ ranges in $\{-1,+1\}$ whenever its arguments take values in $\{-1,+1\}$.

7.5 Reduction to high-degree predicates

Let $U(n,k)$ denote the minimum $U(D)$ over all predicates $D\colon\{0,1,\ldots,n\}\to\{-1,+1\}$ with $\deg(D)=k$. In this notation, our ultimate goal will be to bound $U(n,k)$ from below. This section takes a step in that direction. Specifically, we reduce the task of analyzing $U(n,k)$ to that of analyzing $U(n,\lceil\alpha n\rceil)$, where $\alpha\geqslant 1/4$. This focuses our efforts on high-degree predicates. In the next section, we will further reduce the problem to *dense* predicates, i.e., high-degree predicates that change value at more or less even intervals in $\{0,1,\ldots,n\}$. These reductions are essential because dense predicates behave more predictably and are much easier to analyze than arbitrary predicates. Dense predicates will be the focus of all later sections.

For a predicate $D\colon\{0,1,\ldots,n\}\to\{-1,+1\}$, we define its *flip vector* $v=(v_0,v_1,\ldots,v_n)\in\{0,1\}^{n+1}$ by

$$v_i=\begin{cases}1 & \text{if } D(i)\neq D(i-1),\\ 0 & \text{otherwise,}\end{cases}$$

where we adopt the convention that $D(-1)\equiv 1$. Note that $\deg(D)=v_1+v_2+\cdots+v_n$. Also, if D_1 and D_2 are predicates with flip vectors $v^{(1)}$ and $v^{(2)}$, then $D_1\oplus D_2$ has flip vector $v^{(1)}\oplus v^{(2)}$. Finally, given a predicate $D\colon\{0,1,\ldots,n\}\to\{-1,+1\}$, consider a derived predicate $D'\colon\{0,1,\ldots,m\}\to\{-1,+1\}$ given by $D'(t)\equiv D(t+\Delta)$, where $m\geqslant 1$ and $\Delta\geqslant 0$ are integers with $m+\Delta\leqslant n$. Then the flip vectors v and v' of D and D', respectively, are related as follows: $v'=(v_0\oplus\cdots\oplus v_\Delta,v_{\Delta+1},\ldots,v_{\Delta+m})$. From the standpoint of communication

complexity, D' can be computed by hardwiring some inputs to a protocol for D:

$$D'\Big(\Big|x_1x_2\ldots x_m \bigwedge y_1y_2\ldots y_m\Big|\Big) \\ = D\Big(\Big|x_1x_2\ldots x_m 1^{\Delta}0^{n-m-\Delta} \bigwedge y_1y_2\ldots y_m 1^{\Delta}0^{n-m-\Delta}\Big|\Big).$$

Therefore, $U(D') \leqslant U(D)$.

We begin with a technical lemma. Consider a Boolean vector $v = (v_1, v_2, \ldots, v_n)$. We show that there is a subvector $(v_i, v_{i+1}, \ldots, v_j)$ that is reasonably far from both endpoints of v and yet contains many of the "1" bits present in v.

LEMMA 7.8 (Sherstov [204]). *Let $v \in \{0,1\}^n$, $v \neq 0^n$. Put $k = v_1 + \cdots + v_n$. Then there are indices i, j with $i \leqslant j$ such that*

$$v_i + \cdots + v_j \geqslant \frac{1}{14} \frac{k}{1 + \log(n/k)} \tag{7.1}$$

and

$$\min\{i-1, n-j\} \geqslant j - i. \tag{7.2}$$

PROOF. By symmetry, we can assume that $v_1 + v_2 + \cdots + v_m \geqslant \frac{1}{2}k$ for some index $m \leqslant \lceil n/2 \rceil$. Let $\alpha \in (0, \frac{1}{2})$ be a parameter to be fixed later. Let $T \geqslant 0$ be the smallest integer such that

$$v_1 + v_2 + \cdots + v_{\lfloor m/2^T \rfloor} < (1-\alpha)^T (v_1 + v_2 + \cdots + v_m).$$

Clearly, $T \geqslant 1$. Since $v_1 + v_2 + \cdots + v_{\lfloor m/2^T \rfloor} \leqslant m/2^T$, we further obtain

$$1 \leqslant T \leqslant 1 + \frac{1 + \log(n/k)}{\log(2 - 2\alpha)}.$$

Now,

$$\begin{aligned} v_{\lfloor m/2^T \rfloor + 1} + \cdots + v_{\lfloor m/2^{T-1} \rfloor} &= \underbrace{(v_1 + \cdots + v_{\lfloor m/2^{T-1} \rfloor})}_{\geqslant (1-\alpha)^{T-1}(v_1 + v_2 + \cdots + v_m)} - \underbrace{(v_1 + \cdots + v_{\lfloor m/2^T \rfloor})}_{< (1-\alpha)^T (v_1 + v_2 + \cdots + v_m)} \\ &> \frac{1}{2}\alpha(1-\alpha)^{T-1}k \\ &\geqslant \frac{1}{2}\alpha(1 - \alpha(T-1))k \\ &\geqslant \frac{1}{2}\alpha\left(1 - \alpha \cdot \frac{1 + \log(n/k)}{\log(2-2\alpha)}\right)k. \end{aligned} \tag{7.3}$$

Set $\alpha = 0.23/(1 + \log(n/k))$, $i = \lfloor m/2^T \rfloor + 1$, and $j = \lfloor m/2^{T-1} \rfloor$. Then one easily verifies (7.2), while (7.1) is immediate from (7.3). □

We are now ready to prove the desired reduction to high-degree predicates. Throughout this proof, we will freely use the opening remarks of this section, often without mention.

LEMMA 7.9 (Sherstov [204]). *For all integers n, k with $1 \leqslant k \leqslant n$,*

$$U(n,k) \geqslant \frac{5}{6}\, K \min_{\substack{m = K, \ldots, n, \\ 1/4 \leqslant \alpha \leqslant 1}} \left\{ \frac{1}{m}\, U(m, \lceil \alpha m \rceil) \right\},$$

where

$$K = \left\lceil \frac{1}{14}\, \frac{k}{1 + \log(n/k)} \right\rceil.$$

PROOF. Let $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$ be any predicate with $\deg(D) = k$. As outlined before, the intuition is to express some complicated (i.e., high-degree) predicate as the XOR of a small number of predicates derived from D. The details follow.

Let $v = (v_0, v_1, \ldots, v_n)$ be the flip vector of D. Apply Lemma 7.8 to $(v_1, \ldots, v_n)$ and let i, j be the resulting indices, $i \leqslant j$. Put $m = j - i + 1$. Since $v_i + \cdots + v_j \geqslant K$, we have

$$K \leqslant m \leqslant n. \tag{7.4}$$

Define predicates $D^{-(m-1)}, \ldots, D^0, \ldots, D^{m-1}$, each a mapping $\{0, 1, \ldots, m\} \to \{-1, +1\}$, by $D^r(t) \equiv D(t + i - 1 + r)$. Then (7.2) shows that each of these predicates can be computed by taking a protocol for D and fixing all but the first m variables to appropriate values. Thus,

$$U(D) \geqslant U(D^r), \qquad r = -(m-1), \ldots, (m-1). \tag{7.5}$$

The flip vector of D^0 is $(*, v_i, \ldots, v_j)$ for some $* \in \{0, 1\}$, which means that $\deg(D^0) = v_i + \cdots + v_j$. If $\deg(D^0) > m/2$, then the theorem is true for D in view of (7.4) and (7.5). Thus, we can assume the contrary:

$$K \leqslant v_i + \cdots + v_j \leqslant \frac{1}{2}m. \tag{7.6}$$

If we write the flip vectors of $D^{-(m-1)}, \ldots, D^{m-1}$ one after another as row vectors, we obtain the following matrix A:

$$A = \begin{bmatrix} * & * & * & * & * & \cdots & * & * & * & v_i \\ * & * & * & * & * & \cdots & * & * & v_i & v_{i+1} \\ * & * & * & * & * & \cdots & * & v_i & v_{i+1} & v_{i+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ * & v_i & v_{i+1} & v_{i+2} & v_{i+3} & \cdots & v_{j-3} & v_{j-2} & v_{j-1} & v_j \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ * & v_{j-2} & v_{j-1} & v_j & * & \cdots & * & * & * & * \\ * & v_{j-1} & v_j & * & * & \cdots & * & * & * & * \\ * & v_j & * & * & * & \cdots & * & * & * & * \end{bmatrix}.$$

Let T be a suitably large integer to be named later, and let $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(T)}$ be independent random vectors, each selected uniformly from among the rows of A. Put $\mathbf{u} = \mathbf{u}^{(1)} \oplus \mathbf{u}^{(2)} \oplus \cdots \oplus \mathbf{u}^{(T)}$. We will index the columns of A and the components of all these vectors by $0, 1, \ldots, m$ (left to right). Let p_r stand for the fraction of 1s in the rth column of A. Every column of A, except the zeroth, contains $v_i, \ldots, v_j$ and some $m - 1$ additional values. One infers from (7.6) that

$$\frac{K}{2m} \leqslant p_r \leqslant \frac{3}{4}, \qquad r = 1, 2, \ldots, m. \tag{7.7}$$

Therefore,

$$\begin{aligned} \mathbf{E}\Big[(\mathbf{u})_1 + \cdots + (\mathbf{u})_m\Big] &= \sum_{r=1}^{m} \mathbf{E}\Big[(\mathbf{u}^{(1)})_r \ \oplus \ \cdots \ \oplus \ (\mathbf{u}^{(T)})_r\Big] \\ &= \sum_{r=1}^{m} \left(\frac{1}{2} - \frac{1}{2}(1 - 2p_r)^T\right) && \text{by Proposition 7.4} \\ &\geqslant \frac{1}{2}m\left(1 - \frac{1}{\mathrm{e}^{TK/m}}\right) && \text{by (7.6), (7.7).} \end{aligned}$$

Fix $T = \lceil (\ln 2)m/K \rceil$. Then by the last calculation, there is a vector $u = (u_0, u_1, \ldots, u_m)$ that satisfies $u_1 + \cdots + u_m \geqslant m/4$ and is the XOR of some T rows of A. In other words, there is a predicate $D^{\oplus}: \{0, 1, \ldots, m\} \to \{-1, +1\}$ that satisfies $\deg(D^{\oplus}) \geqslant m/4$ and is the XOR of some $T \leqslant \frac{6m}{5K}$ predicates from among $D^{-(m-1)}, \ldots, D^{m-1}$. This completes the proof in view of (7.5) and Proposition 7.7. □

7.6 Reduction to dense predicates

The proof in this section uses the same setup as Lemma 7.9, except the argument is now more involved. The reason is that the previous averaging argument is not strong enough to yield a dense predicate, which is a highly structured object. To overcome this, we recast the previous argument as a random walk on $\mathbb{Z}_2^n$ and show that it mixes rapidly. In particular, we will need the following lemma that bounds the mixing time of a random walk.

LEMMA 7.10 (Razborov [174, Lem. 1]). *Fix a probability distribution μ on $\{0,1\}^n$. Let $\{v^{(1)}, v^{(2)}, \ldots, v^{(n)}\}$ be a basis for $\{0,1\}^n$ as a vector space over GF(2). Put*

$$p = \min \left\{ \mu(0^n), \mu(v^{(1)}), \mu(v^{(2)}), \ldots, \mu(v^{(n)}) \right\}.$$

Let $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(T)}$ be independent random vectors, each distributed according to μ. Then for every $v \in \{0,1\}^n$,

$$\left| \mathbf{P}\left[\mathbf{u}^{(1)} \oplus \cdots \oplus \mathbf{u}^{(T)} = v \right] - 2^{-n} \right| \leqslant \mathrm{e}^{-2Tp}.$$

REMARK. Razborov's article [174] is in Russian. For an English translation, see Jukna [99, Lem. 24.3].

We are ready to formally define dense predicates and give the promised reduction.

DEFINITION 7.11 (Sherstov [204]). Let n, b be positive integers and $d \geqslant 0$ a real number. A predicate D is called (n, b, d)-*dense* if D is a predicate $\{0, 1, \ldots, n\} \to \{-1, +1\}$ with flip vector $(v_0, v_1, \ldots, v_n)$ satisfying

$$v_{rb+1} + v_{rb+2} + \cdots + v_{(r+1)b} \geqslant d, \qquad r = 0, 1, 2, \ldots, \left\lfloor \frac{n}{b} \right\rfloor - 1.$$

LEMMA 7.12 (Sherstov [204]). *Let $D\colon \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a predicate with $\deg(D) \geqslant \frac{1}{4}n$. Let b be any integer with $1 \leqslant b \leqslant \frac{1}{350}n$. Then*

$$U(D) \geqslant \frac{b}{n \log n}\, U(D'),$$

where D' is a certain $(m, \lceil \log n \rceil b, \frac{1}{700}b)$-dense predicate and $\frac{1}{350}n \leqslant m \leqslant n$.

PROOF. Let $(v_0, v_1, \ldots, v_n)$ be the flip vector of D. Apply Lemma 7.8 to $(v_1, \ldots, v_n)$ and let i, ℓ be the resulting indices ($i \leqslant \ell$). It will be convenient to work with a somewhat smaller subvector $v = (v_i, \ldots, v_j)$, where we define $j \in \{i, \ldots, \ell\}$ to be the largest integer so that $b \mid (j - i + 1)$. Since $b \leqslant \frac{1}{350}n$ and $v_i + \cdots + v_\ell \geqslant \frac{1}{168}n$, this gives:

$$v_i + \cdots + v_j \geqslant \frac{1}{350}n. \tag{7.8}$$

Defining $m = j - i + 1$, we infer that $\frac{1}{350}n \leqslant m \leqslant n$, as desired. We view $v = (v_i, \ldots, v_j)$ as composed of consecutive blocks, each b bits long:

$$v = \left(\underbrace{\boxed{v_i, \ldots, v_{i+b-1}}}_{\text{block } 1}, \underbrace{\boxed{v_{i+b}, \ldots, v_{i+2b-1}}}_{\text{block } 2}, \cdots\cdots, \underbrace{\boxed{v_{j-b+1}, \ldots, v_j}}_{\text{block } m/b} \right). \tag{7.9}$$

For $r = 1, 2, \ldots, b$, define the *rth layer of* v, denoted $z^{(r)}$, to be the vector obtained by taking the rth component from each of the above blocks:

$$z^{(r)} = (v_{i-1+r}, v_{i-1+b+r}, \ldots, v_{j-b+r}) \in \{0,1\}^{m/b}.$$

We say of a layer z that it is *perfect* if it does not have $\lceil \log n \rceil$ consecutive components equal to 0. If more than $\frac{1}{700}b$ of the layers are perfect, take D' to be the predicate with flip vector $(v_0 \oplus \cdots \oplus v_{i-1}, v_i, \ldots, v_j)$. Clearly, D' is $(m, \lceil \log n \rceil b, \frac{1}{700}b)$-dense. Furthermore, $U(D') \leqslant U(D)$, by the same argument as in Lemma 7.9. As a result, the theorem holds in this case.

Thus, we may assume that at least $(1 - \frac{1}{700})b$ of the layers are not perfect. In view of (7.8), at most $(1 - \frac{1}{350})b$ layers can be zero vectors. Therefore, $\frac{1}{700}b$ or more layers are nonzero *and* not perfect. These are the only layers we will consider in the remainder of the proof.

Define predicates $D^{-(m-b)}, D^{-(m-2b)}, \ldots, D^{-b}, D^0, D^b, \ldots, D^{m-2b}, D^{m-b}$, each a mapping $\{0, 1, \ldots, m\} \to \{-1, +1\}$, by $D^r(t) \equiv D(t + i - 1 + r)$. These are a subset of the predicates from the proof of Lemma 7.9, and again

$$U(D) \geqslant U(D^r) \qquad \text{for each } r. \tag{7.10}$$

Writing the flip vectors of these predicates one after another as row vectors yields the following matrix B:

$B =$

$$\begin{bmatrix}
* & * & * & * & \cdots & * & * & \text{block } 1 \\
* & * & * & * & \cdots & * & \text{block } 1 & \text{block } 2 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
* & \text{block } 1 & \text{block } 2 & \text{block } 3 & \cdots & \text{block } \frac{m}{b}-2 & \text{block } \frac{m}{b}-1 & \text{block } \frac{m}{b} \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
* & \text{block } \frac{m}{b}-1 & \text{block } \frac{m}{b} & * & \cdots & * & * & * \\
* & \text{block } \frac{m}{b} & * & * & \cdots & * & * & *
\end{bmatrix},$$

where the blocks refer to the partition in (7.9). Let T be a suitably large integer to be named later, and let $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \dots, \mathbf{u}^{(T)}$ be independent random vectors, each selected uniformly from among the rows of B. Put $\mathbf{u} = \mathbf{u}^{(1)} \oplus \mathbf{u}^{(2)} \oplus \cdots \oplus \mathbf{u}^{(T)}$. We will index the columns of B and the components of $\mathbf{u}$ by $0, 1, \dots, m$ (left to right). Key to analyzing the distribution of $\mathbf{u}$ is the following claim.

CLAIM 7.13 (Sherstov [204]). *Let $T \geqslant (m/b)\ln n$. Let $\Delta \in \{1, 2, \dots, b\}$ be such that the layer $z^{(\Delta)}$ is nonzero and not perfect. Let $s \in \{0, b, 2b, 3b, \dots\}$ be such that $s + \lceil \log n \rceil b \leqslant m$. Then*

$$\mathbf{P}\Big[(\mathbf{u})_{s+\Delta} = (\mathbf{u})_{s+b+\Delta} = \cdots = (\mathbf{u})_{s+(\lceil \log n \rceil - 1)b+\Delta} = 0\Big] \leqslant \frac{2}{n}.$$

PROOF. Let B' be the matrix whose columns are the following columns of B: $s+\Delta, s+b+\Delta, \dots, s+(\lceil \log n \rceil - 1)b + \Delta$, in that order. Since $z^{(\Delta)}$ is nonzero and not perfect, $z^{(\Delta)}$ has $\lceil \log n \rceil + 1$ consecutive components with values either $0, 0, \dots, 0, 1$ or $1, 0, 0, \dots, 0$. Consequently, B' must contain one of the following submatrices, each of size $(\lceil \log n \rceil + 1) \times \lceil \log n \rceil$:

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ & & & & & 1 \\ & 0 & & & 1 & \\ & & & \iddots & & \\ & & 1 & & & \\ & 1 & & & * & \\ 1 & & & & & \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} & & & & & 1 \\ & * & & & 1 & \\ & & & \iddots & & \\ & & 1 & & & \\ & 1 & & & 0 & \\ 1 & & & & & \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

The claim now follows from Lemma 7.10, since $2^{-\lceil \log n \rceil} + \mathrm{e}^{-2T \cdot \frac{b}{2m}} \leqslant 2/n$. □

We return to the proof of the lemma. Fix $T = \lceil (m/b)\ln n \rceil$. Let $s = 0$ and apply Claim 7.13 with every $\Delta \in \{1, 2, \dots, b\}$ for which the layer $z^{(\Delta)}$ is nonzero and not perfect. Since there are at least $\frac{1}{700}b$ such choices for Δ, we conclude by

the union bound that

$$\mathbf{P}\left[(\mathbf{u})_1 + (\mathbf{u})_2 + \cdots + (\mathbf{u})_{\lceil \log n \rceil b} < \frac{1}{700} b\right] \leqslant b \cdot \frac{2}{n}.$$

The same calculation applies to the next set of $\lceil \log n \rceil b$ components of $\mathbf{u}$ (i.e., $s = \lceil \log n \rceil b$), and so on. Applying a union bound across all these $m/(\lceil \log n \rceil b)$ calculations, we find that with probability

$$1 - \frac{m}{\lceil \log n \rceil b} \left(b \cdot \frac{2}{n} \right) > 0,$$

the predicate whose flip vector is $\mathbf{u}$ is $(m, \lceil \log n \rceil b, \frac{1}{700} b)$-dense. Fix any such predicate D'. Since D' is the XOR of $T \leqslant (n \log n)/b$ predicates from among $D^{-(m-b)}, \ldots, D^{m-b}$, the lemma follows by (7.10) and Proposition 7.7. □

7.7 Univariate approximation with clusters of nodes

Crucial to our study of dense predicates are certain approximation problems to which they give rise. Roughly speaking, the hardness of such an approximation problem for low-degree polynomials translates into the communication hardness of the associated predicate. This section carries out the first part of the program, namely, showing that the approximation task at hand is hard for low-degree polynomials. We examine this question in its basic mathematical form, with no extraneous considerations to obscure our view. How communication fits in this picture will become clear in the next two sections.

For a finite set $X \subset \mathbb{R}$, a function $f : X \to \mathbb{R}$, and an integer $r \geqslant 0$, define

$$\varepsilon^*(f, X, r) = \min_{p \in P_r} \max_{x \in X} |p(x) - f(x)|.$$

In words, $\varepsilon^*(f, X, r)$ is the least error (in the uniform sense) to which a degree-r polynomial can approximate f on X. The following fact from approximation theory is useful in estimating this error; see, for example, Rivlin [185, Thm. 1.15].

FACT 7.14. *Let $X = \{x_1, x_2, \ldots, x_{r+2}\}$ be a set of $r+2$ distinct reals. Let $f : X \to \mathbb{R}$ be given. Put $\omega(x) = (x - x_1)(x - x_2) \cdots (x - x_{r+2})$. Then*

$$\varepsilon^*(f, X, r) = \frac{\left|\sum_{i=1}^{r+2} [f(x_i)/\omega'(x_i)]\right|}{\sum_{i=1}^{r+2} [1/|\omega'(x_i)|]}.$$

To develop some intuition for the work in this section, consider the following approximation problem. Let $f : \{0, 1, \ldots, n\} \to \{-1, +1\}$ be defined by

$$f(x) = \begin{cases} -1 & \text{if } x = \lfloor n/2 \rfloor, \\ +1 & \text{otherwise.} \end{cases}$$

It is well-known that any polynomial that approximates f within $1/3$ has degree $\Omega(n)$. For example, this follows from work by Paturi [165]. The approximation problem of interest to us is similar, except that our points need not be as evenly spaced as $0, 1, \ldots, n$ but rather may form clusters. As a result, Paturi's results and methods do not apply, and we approach this question differently, using the first-principles formula of Fact 7.14. Specifically, our main result in this section is as follows.

LEMMA 7.15 (Sherstov [204]). *Let positive integers L, d and a real number $B \geqslant d$ be given. Let $\{x_{ij} : i = 1, \ldots, L;\ j = 1, \ldots, d\}$ be a set of Ld distinct reals, where $x_{ij} \in [(i-1)B, iB]$ and*

$$|x_{ij} - x_{i'j'}| \geqslant 1 \qquad \text{for } (i, j) \neq (i', j'). \tag{7.11}$$

Let $x_0 \in [\frac{1}{4}LB, \frac{3}{4}LB]$. *Then any polynomial* p *with*

$$p(x_0) = 1, \qquad |p(x_{ij})| < \frac{1}{2}\left(\frac{1}{LB}\right)^{4d+1} \qquad \textit{for all } i, j$$

has degree at least $\left(\frac{1}{2}L - 1\right) d$.

PROOF. Define $f(x)$ by

$$f(x) = \begin{cases} 1 & \text{if } x = x_0, \\ 0 & \text{if } x = x_{ij} \text{ for some } i, j. \end{cases}$$

By symmetry, we can assume that $x_0 \in [\frac{1}{4}LB, \frac{1}{2}LB]$. Fix an integer $\ell \leqslant \lceil \frac{1}{2}L \rceil$ so that $x_0 \in [(\ell - 1)B, \ell B]$. Put

$$X = \{x_0\} \cup \{x_{ij} : i = 1, \ldots, 2\ell - 1; \;\; j = 1, \ldots, d\}.$$

With $\omega(x) = \prod_{y \in X}(x - y)$, Fact 7.14 implies that

$$\varepsilon^*(f, X, |X| - 2) \geqslant \frac{1}{|X|} \, \frac{\min_{x \in X} |\omega'(x)|}{|\omega'(x_0)|}. \tag{7.12}$$

We proceed to estimate the denominator and numerator of (7.12). Since x_0 is distinct from each x_{ij}, the quantity

$$\delta = \min_{\substack{i=1,\ldots,2\ell-1, \\ j=1,\ldots,d}} |x_0 - x_{ij}|$$

satisfies $\delta > 0$. We have:

$$|\omega'(x_0)| \;=\; \prod_{j=1}^{d}\prod_{i=1}^{2\ell-1}|x_0 - x_{ij}| \;\leqslant\; \delta \prod_{j=1}^{d}\prod_{i=1}^{2\ell-1} B \underbrace{\left\lceil \frac{|x_0 - x_{ij}|}{B} \right\rceil}_{\leqslant |i-\ell|+1}$$

$$\leqslant\; \delta \cdot \Big(\ell!\,\ell!\,B^{2\ell-1}\Big)^{d}. \tag{7.13}$$

On the other hand, every $x_{i'j'} \in X$ satisfies:

$$\begin{aligned}
|\omega'(x_{i'j'})| \;&=\; \prod_{x\in X\setminus\{x_{i'j'}\}} |x - x_{i'j'}| \\
&\geqslant\; \delta \prod_{j=1}^{d} \prod_{\substack{i=1,\dots,2\ell-1 \\ i\notin\{i'-1,i',i'+1\}}} |x_{ij} - x_{i'j'}| \qquad \text{by (7.11)} \\
&\geqslant\; \delta \prod_{j=1}^{d} \prod_{\substack{i=1,\dots,2\ell-1 \\ i\notin\{i'-1,i',i'+1\}}} B \underbrace{\left\lfloor \frac{|x_{ij} - x_{i'j'}|}{B} \right\rfloor}_{\geqslant |i-i'|-1} \\
&\geqslant\; \delta \cdot \left(\frac{\ell!\,\ell!\,B^{2\ell-4}}{\ell^4}\right)^{d}.
\end{aligned} \tag{7.14}$$

Now (7.12) yields, in view of (7.13) and (7.14):

$$\varepsilon^*(f, X, |X|-2) \geqslant \frac{1}{2}\left(\frac{1}{LB}\right)^{4d+1},$$

which concludes the proof since $|X| \geqslant \left(\frac{1}{2}L - 1\right) d + 1$. ∎

7.8 Key analytic property of dense predicates

We now transition to the final ingredient of our proof, *smooth orthogonalizing distributions* for a given predicate D. This informal term refers to a distribution on $\{0,1\}^n$ that does not put too little weight on any point (the *smooth* part) and under which $D(\sum x_i)$ is approximately orthogonal to all low-degree characters χ_S (the *orthogonalizing* part). Our task is to establish the existence of such distributions for every dense predicate. Once this is accomplished, we will be able to treat a dense predicate as if it were the familiar parity function, whose defining analytic property is precisely its orthogonality to the lower-order characters under the uniform distribution. Crucial to the development below will be the inapproximability result that we proved in Section 7.7.

For a polynomial p, a predicate $D\colon\{0,1,\ldots,n\}\to\{-1,+1\}$, and a number $N>0$, define the *advantage* of p in computing D by

$$\mathrm{adv}(p,N,D)=N\min_{t=0,\ldots,n}\{D(t)p(t)\}+\sum_{t=0}^{n}\frac{\binom{n}{t}}{2^n}D(t)p(t).$$

This quantity is conceptually close to the correlation of p and D with respect the binomial distribution. There is a substantial difference, however: if p and D differ in sign at some point, this causes a penalty term to be subtracted. We will be interested in values $N\gg 1$, when even a single error of p results in a large penalty. Define

$$\mathrm{adv}_r(N,D)=\max_p\ \mathrm{adv}(p,N,D),$$

where the maximization is over $p\in P_r$ with $|p(t)|\leqslant 1$ for $t=0,1,\ldots,n$. As we now show, this quantity is closely related to smooth orthogonalizing distributions for D.

THEOREM 7.16 (Sherstov [204]). *Fix a predicate $D\colon\{0,1,\ldots,n\}\to\{-1,+1\}$ and an integer $r\geqslant 0$. Then for every $N>1$, there is a distribution μ on $\{0,1\}^n$*

such that $\mu(x) \geqslant \frac{1}{2^n N}$ *for each* x *and*

$$\left|\mathop{\mathbf{E}}_{x}\left[D(\textstyle\sum x_i)\mu(x)\ \chi_S(x)\right]\right| \leqslant \frac{1}{2^n N}\,\mathrm{adv}_r(N-1, D), \qquad |S| \leqslant r.$$

PROOF. Abbreviate $f(x) = D(\sum x_i)$ and consider the following linear program:

$$\begin{array}{lll}
\text{variables:} & \mu(x) \text{ for all } x; \quad \varepsilon & \\
\text{minimize:} & \varepsilon & \\
\text{subject to:} & \left|\displaystyle\sum_{x\in\{0,1\}^n} \mu(x)\, f(x)\, \chi_S(x)\right| \leqslant \varepsilon & \text{for } |S| \leqslant r, \\
& \displaystyle\sum_{x\in\{0,1\}^n} \mu(x) = 1, & \\
& \mu(x) \geqslant \dfrac{1}{2^n N} & \text{for each } x.
\end{array} \tag{LP1}$$

It suffices to show that the optimum of this program is at most $\frac{1}{N}\,\mathrm{adv}_r(N-1, D)$. For this, we pass to the dual:

$$\begin{array}{lll}
\text{variables:} & \alpha_S \text{ (for } |S| \leqslant r); \quad \xi_x \text{ (for all } x); \quad \Delta & \\
\text{maximize:} & \dfrac{1}{N}\left((N-1)\Delta + \dfrac{1}{2^n}\displaystyle\sum_{x\in\{0,1\}^n} (\Delta + \xi_x)\right) & \\
\text{subject to:} & f(x)\displaystyle\sum_{|S|\leqslant r} \alpha_S\, \chi_S(x) \geqslant \Delta + \xi_x & \text{for all } x, \\
& \displaystyle\sum_{|S|\leqslant r} |\alpha_S| \leqslant 1, & \\
& \alpha_S \in \mathbb{R} & \text{for } |S| \leqslant r, \\
& \xi_x \geqslant 0 & \text{for all } x, \\
& \Delta \in \mathbb{R}. &
\end{array} \tag{LP2}$$

The dual programs (LP1) and (LP2) are both feasible and thus have the same finite optimum. Therefore, our task reduces to proving that the optimum of (LP2) is at most $\frac{1}{N}\,\mathrm{adv}_r(N-1, D)$. Fix an optimal solution to (LP2). Then

$$f(x) \sum_{|S| \leqslant r} \alpha_S \, \chi_S(x) = \Delta + \xi_x \qquad \text{for all } x, \tag{7.15}$$

since in case of a strict inequality ($>$) we could increase the corresponding variable ξ_x by a small amount to obtain a feasible solution with greater value. Furthermore, we claim that

$$\Delta = \min_{x \in \{0,1\}^n} \left\{ f(x) \sum_{|S| \leqslant r} \alpha_S \, \chi_S(x) \right\}. \tag{7.16}$$

Indeed, let m stand for the right-hand side of (7.16). Then $\Delta \leqslant m$ because each ξ_x is nonnegative. It remains to show that $\Delta \geqslant m$. If we had $\Delta < m$, then (7.15) would imply that $\xi_x \geqslant m - \Delta$ for all x. As a result, we could obtain a new feasible solution $\xi'_x = \xi_x + (\Delta - m)$ and $\Delta' = m$. This new solution satisfies $\Delta' + \xi'_x = \Delta + \xi_x$ for all x. Moreover, $\Delta' > \Delta$, which results in a greater objective value and yields the desired contradiction. In summary, $\Delta = m$.

In view of (7.15) and (7.16), the optimum of (LP2) is

$$\frac{1}{N} \max_{\phi} \left\{ (N-1) \min_x \left\{ f(x)\phi(x) \right\} \;+\; \frac{1}{2^n} \sum_x f(x)\phi(x) \right\}, \tag{7.17}$$

where the maximization is over functions ϕ of the form

$$\phi(x) = \sum_{|S| \leqslant r} \alpha_S \, \chi_S(x), \qquad \text{where } \sum_{|S| \leqslant r} |\alpha_S| \leqslant 1. \tag{7.18}$$

Fix ϕ that optimizes (7.17). By (7.18),

$$\max_{x\in\{0,1\}^n} |\phi(x)| \leqslant 1.$$

Put

$$\phi_{\text{symm}}(x) = \frac{1}{n!}\sum_{\sigma\in S_n} \phi(x_{\sigma(1)},\ldots,x_{\sigma(n)}).$$

Since f is symmetric, ϕ and ϕ_{symm} have the same objective value in (7.17). By the symmetrization argument (Proposition 2.2), there is a univariate polynomial $p \in P_r$ with

$$\phi_{\text{symm}}(x) = p(x_1+\cdots+x_n) \qquad \text{for all } x\in\{0,1\}^n.$$

For $t = 0,1,\ldots,n$,

$$\begin{aligned} |p(t)| &= |p(\overbrace{1+\cdots+1}^{t \text{ times}}+0+\cdots+0)| \\ &\leqslant \max_{x\in\{0,1\}^n} |\phi_{\text{symm}}(x)| \\ &\leqslant \max_{x\in\{0,1\}^n} |\phi(x)| \\ &\leqslant 1. \end{aligned}$$

Replacing $\phi(x)$ by $p(x_1+\cdots+x_n)$ in (7.17), we see that the optimum of (LP2) is at most

$$\frac{1}{N}\max_p \left\{ (N-1)\min_{t=0,\ldots,n}\{D(t)p(t)\} \;+\; \frac{1}{2^n}\sum_{t=0}^{n}\binom{n}{t}D(t)p(t)\right\},$$

where the maximization is over $p \in P_r$ with $|p(t)| \leqslant 1$ for $t = 0, 1, \ldots, n$. This latter quantity is $\frac{1}{N}\,\mathrm{adv}_r(N-1, D)$, by definition. □

Theorem 7.16 states that a smooth orthogonalizing distribution for D exists whenever low-degree polynomials have negligible advantage in computing D. Accordingly, we proceed to examine the advantage achievable by low-degree polynomials.

LEMMA 7.17 (Sherstov [204]). *Let D be an $(n, B, 2d+1)$-dense predicate, where n, B, d are positive integers. Assume that $\mathrm{adv}_r(N, D) \geqslant n2^{-n/6}$, where $r < \deg(D)$ and $N > 0$ are given. Then there are $\lfloor \frac{n}{B} \rfloor d$ distinct reals $\{x_{ij} : i = 1, \ldots, \lfloor \frac{n}{B} \rfloor;\ j = 1, \ldots, d\}$ and a polynomial $p \in P_r$ such that:*

$$
\begin{array}{ll}
x_{ij} \in [(i-1)B, iB] & \text{for all } i, j, \\
|x_{ij} - x_{i'j'}| \geqslant 1 & \text{for all } (i, j) \neq (i', j'), \\
|p(x_{ij})| \leqslant \sqrt{n}/N & \text{for all } i, j, \\
p(x_0) = 1 & \text{for some } x_0 \in [\tfrac{1}{4}n, \tfrac{3}{4}n].
\end{array}
$$

PROOF. Fix $q \in P_r$ with $|q(t)| \leqslant 1$ for $t = 0, 1, \ldots, n$ and $\mathrm{adv}(q, N, D) = \mathrm{adv}_r(N, D)$. Fix $k \in \{0, 1, \ldots, n\}$ with

$$
\binom{n}{k} D(k)q(k) = \max_{t=0,\ldots,n} \left\{ \binom{n}{t} D(t)q(t) \right\}.
$$

Since $\deg(q) < \deg(D)$, the quantity $\binom{n}{t} D(t)q(t)$ is positive for at most n values of $t = 0, 1, \ldots, n$. Therefore,

$$
\mathrm{adv}(q, N, D) \leqslant n \cdot \frac{\binom{n}{k}}{2^n} D(k)q(k) \leqslant n \cdot \frac{\binom{n}{k}}{2^n}.
$$

Recalling that $\mathrm{adv}(q, N, D) \geqslant n2^{-n/6}$, we infer that $\frac{1}{4}n \leqslant k \leqslant \frac{3}{4}n$. Put

$$p(t) = \frac{1}{q(k)}\, q(t).$$

Taking $x_0 = k$, we have $\frac{1}{4}n \leqslant x_0 \leqslant \frac{3}{4}n$ and $p(x_0) = 1$, as desired. It remains to find the points x_{ij}. For this, we need the following claim.

CLAIM 7.18 (Sherstov [204]). *Let a, b be integers with $a < b$ and $D(a) \neq D(b)$. Then $|p(\xi)| \leqslant \sqrt{n}/N$ for some $\xi \in [a, b]$.*

PROOF. If q vanishes at some point in $[a, b]$, we are done. In the contrary case, q is nonzero and has the same sign at every point of $[a, b]$, which means that either $q(a)D(a) < 0$ or $q(b)D(b) < 0$. Since $\mathrm{adv}(q, N, D) \geqslant 0$, we have:

$$\begin{aligned}\min\{|q(a)|, |q(b)|\} &\leqslant \frac{n}{N} \max_{t=0,\ldots,n} \left\{ \frac{\binom{n}{t}}{2^n} D(t)q(t) \right\} = \frac{n}{N} \cdot \frac{\binom{n}{k}}{2^n} \cdot |q(k)| \\ &\leqslant \frac{\sqrt{n}}{N}\, |q(k)|,\end{aligned}$$

and hence $\min\{|p(a)|, |p(b)|\} \leqslant \sqrt{n}/N$. □

Fix an integer $i = 1, 2, \ldots, \lfloor \frac{n}{B} \rfloor$. Since D is $(n, B, 2d + 1)$-dense, D changes value at least $2d$ times in $[(i-1)B + 1, iB]$. As a result, there are at least d pairs of integers $(a_1, b_1), \ldots, (a_d, b_d)$ with

$$D(a_1) \neq D(b_1), \quad D(a_2) \neq D(b_2), \quad \ldots, \quad D(a_d) \neq D(b_d)$$

and

$$(i-1)B + 1 \leqslant a_1 < b_1 < a_2 < b_2 < \cdots < a_d < b_d \leqslant iB.$$

In view of Claim 7.18, this provides the desired d points in $[(i-1)B+1, iB]$. □

We have reached the main result of this section.

THEOREM 7.19 (Sherstov [204]). *Let D be an $(n, B, 2d+1)$-dense predicate, where n, B, d are positive integers with $B \mid n$ and $n \geqslant 3B$. Then there is a distribution μ on $\{0,1\}^n$ such that:*

$$\begin{aligned} &\mu(x) \geqslant \frac{1}{2^n} \frac{1}{3n^{4d+1.5}} && \text{for each } x, \\ &\left| \mathop{\mathbf{E}}_{x} \left[D(|x|)\mu(x)\chi_S(x) \right] \right| \leqslant 2^{-7n/6} && \text{for } |S| < \frac{nd}{6B}. \end{aligned}$$

PROOF. Put $N = 3n^{4d+1.5}$. In view of Theorem 7.16, it is sufficient to show that $\mathrm{adv}_r(N-1, D) < n2^{-n/6}$ for all $r < \frac{nd}{6B}$. So assume, for the sake of contradiction, that $\mathrm{adv}_r(N-1, D) \geqslant n2^{-n/6}$ for some $r < \frac{nd}{6B}$. Since $\deg(D) \geqslant \frac{n}{B}(2d+1)$, we have $r < \deg(D)$. Thus, Lemma 7.17 is applicable and yields $\frac{nd}{B}$ distinct reals $\{x_{ij} : i = 1, \ldots, \frac{n}{B};\ j = 1, \ldots, d\}$ and a polynomial $p \in P_r$ such that:

$$\begin{aligned} &x_{ij} \in [(i-1)B, iB] && \text{for all } i, j, \\ &|x_{ij} - x_{i'j'}| \geqslant 1 && \text{for all } (i,j) \neq (i',j'), \\ &|p(x_{ij})| < \tfrac{1}{2}\left(\tfrac{1}{n}\right)^{4d+1} && \text{for all } i, j, \\ &p(x_0) = 1 && \text{for some } x_0 \in [\tfrac{1}{4}n, \tfrac{3}{4}n]. \end{aligned}$$

Applying Lemma 7.15 with $L = \frac{n}{B}$, we infer that $r \geqslant \left(\frac{1}{2}\frac{n}{B} - 1\right) d$, which yields $r \geqslant \frac{nd}{6B}$ since $\frac{n}{B} \geqslant 3$. We have reached the desired contradiction to $r < \frac{nd}{6B}$. □

7.9 Unbounded-error complexity of symmetric functions

This section consolidates the preceding developments into our main result, a near-optimal lower bound on the unbounded-error communication complexity of every symmetric function. As outlined earlier, we will first solve this problem for dense

predicates and then extend our work to the general case via the reductions of Sections 7.5 and 7.6.

THEOREM 7.20 (Sherstov [204]). *Let $\alpha > 0$ be a sufficiently small absolute constant. Let D be an $(m, b\lceil \log n \rceil, \frac{1}{700}b)$-dense predicate, where $\frac{1}{350}n \leqslant m \leqslant n$ and $b = \lfloor \alpha n / \log^2 n \rfloor$. Then*

$$U(D) \geqslant \Omega\left(\frac{n}{\log n}\right).$$

PROOF. Throughout the proof we will, without mention, use the assumption that n is large enough. This will simplify the setting of parameters, the manipulation of floors and ceilings, and generally make the proof easier to follow.

Fix an integer $v \in [\frac{1}{8}m, \frac{1}{4}m]$ with $b\lceil \log n \rceil \mid v$. It is clear that $v \gg 3b\lceil \log n \rceil$. Define $D' : \{0, 1, \ldots, v\} \to \{-1, +1\}$ by $D'(t) \equiv D(t)$. Since D' is $(v, b\lceil \log n \rceil, \frac{1}{700}b)$-dense, Theorem 7.19 provides a distribution μ on $\{0,1\}^v$ with

$$\mu(z) \geqslant 2^{-v} \; 2^{-\alpha n / 350 \log n}, \qquad z \in \{0,1\}^v, \tag{7.19}$$

and

$$\left| \mathop{\mathbf{E}}_{z} \left[D(|z|)\mu(z)\chi_S(z) \right] \right| \leqslant 2^{-7v/6}, \qquad |S| < \frac{v}{6 \cdot 1401 \lceil \log n \rceil}. \tag{7.20}$$

Define $\phi : \{0,1\}^v \to \mathbb{R}$ by $\phi(z) = D(|z|)\mu(z)$. Restating (7.20),

$$|\hat{\phi}(S)| \leqslant 2^{-7v/6}, \qquad |S| < \frac{v}{6 \cdot 1401 \lceil \log n \rceil}. \tag{7.21}$$

Furthermore, Proposition 2.1 reveals that

$$\max_{S \subseteq [v]} |\hat{\phi}(S)| \leqslant 2^{-v}. \tag{7.22}$$

Let A be the $(2v, v, 8^{-v}\phi)$-pattern matrix. By (7.21), (7.22), and Theorem 4.3,

$$\|A\| \leqslant 4^{-v} \; 2^{-v/12 \cdot 1401 \lceil \log n \rceil}. \tag{7.23}$$

By (7.19), every entry of A has absolute value at least $16^{-v} \, 2^{-\alpha n / 350 \log n}$. Combining this observation with (7.23) and Theorem 7.6,

$$\operatorname{rk}_{\pm} A \geqslant 2^{v/12 \cdot 1401 \lceil \log n \rceil} \; 2^{-\alpha n / 350 \log n}.$$

Recall that $v \geqslant \frac{1}{8} m \geqslant \frac{1}{8 \cdot 350} n$. Hence, for a suitably small constant $\alpha > 0$,

$$\operatorname{rk}_{\pm} A \geqslant 2^{\Omega(n/\log n)}.$$

It remains to relate the sign-rank of A to the communication complexity of D. Let F be the $(2v, v, f)$-pattern matrix, where $f(z) = D(|z|)$. Then $\operatorname{rk}_{\pm} A = \operatorname{rk}_{\pm} F$ because A and F have the same sign pattern. But F is a submatrix of the communication matrix of D, namely,

$$M = \Big[D(|x \wedge y|) \Big]_{x \in \{0,1\}^m, y \in \{0,1\}^m}.$$

Thus,

$$\operatorname{rk}_{\pm} M \geqslant \operatorname{rk}_{\pm} F = \operatorname{rk}_{\pm} A \geqslant 2^{\Omega(n/\log n)}.$$

In view of Theorem 7.2, the proof is complete. □

COROLLARY 7.21 (Sherstov [204]). *Let* $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$ *be a predicate with* $\deg(D) \geqslant \frac{1}{4}n$. *Then*

$$U(D) \geqslant \Omega\left(\frac{n}{\log^4 n}\right).$$

PROOF. Immediate from Lemma 7.12 and Theorem 7.20. □

At last, we arrive at the main result of this chapter, cf. Theorem 7.1 above.

THEOREM 7.22 (Sherstov [204]). *Let* $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$ *be a nonconstant predicate,* $k = \deg(D)$. *Then*

$$\Theta\left(\frac{k}{\{1+\log(n/k)\}\log^4 n}\right) \leqslant U(D) \leqslant \Theta\left(k \log \frac{2n}{k}\right).$$

PROOF. The lower bound is immediate from Lemma 7.9 and Corollary 7.21. To prove the upper bound, fix $p \in P_k$ with $\operatorname{sgn} p(t) = D(t)$ for $t = 0,1,\ldots,n$. Put

$$M = \Big[D(|x \wedge y|)\Big]_{x,y}, \qquad R = \Big[p(x_1 y_1 + \cdots + x_n y_n)\Big]_{x,y},$$

where the indices run as usual: $x, y \in \{0,1\}^n$. Then $M_{xy} R_{xy} > 0$ for all x and y. Thus, the sign-rank of M does not exceed $\sum_{i=0}^{k} \binom{n}{i}$. In view of Theorem 7.2, this completes the proof. □

7.10 Concluding remarks

It is natural to wonder whether the logarithmic factors in Theorem 7.22 can be eliminated. The answer varies from one predicate to another. There are indeed predicates $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$ for which $U(D) = \Theta(\deg(D))$. For example, the conjunction predicate, given by $\mathrm{AND}_n(t) = -1 \Leftrightarrow t = n$, has degree 1 and unbounded-error complexity $\Theta(1)$, as one can verify from the repre-

sentation $\mathrm{AND}_n(\sum x_i y_i) = 1 - 2\prod x_i \cdot \prod y_i$. Similarly, the familiar predicate $\mathrm{PARITY}_n(t) = (-1)^t$ has degree n and unbounded-error complexity $\Theta(n)$ by Forster's result [70]. At the same time, there are predicates D for which a logarithmic gap exists between $\deg(D)$ and $U(D)$. One such predicate is disjunction, given by $\mathrm{OR}_n(t) = 1 \Leftrightarrow t = 0$, which has degree 1 and unbounded-error complexity $\Theta(\log n)$:

PROPOSITION 7.23 (Sherstov [204]). $U(\mathrm{OR}_n) = \Theta(\log n)$.

PROOF. The upper bound is immediate from Theorem 7.22. For the lower bound, note that

$$\bigoplus_{i=1}^{t} x_i \wedge y_i = \bigvee_{i=1}^{4^t} f_i(x_1, \ldots, x_t) \wedge g_i(y_1, \ldots, y_t),$$

where f_i, g_i are suitable Boolean functions (in fact, conjunctions of literals). This yields the inequality $U(\mathrm{PARITY}_t) \leqslant U(\mathrm{OR}_{4^t})$, which completes the proof since $U(\mathrm{PARITY}_t) = \Theta(t)$ by Forster's result [70]. □

The lower bound of Proposition 7.23 is of course valid for any predicate D that contains disjunction or its negation as a subfunction. More precisely:

PROPOSITION 7.24 (Sherstov [204]). *Let $D\colon \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a predicate with flip vector v. If v contains the subvector $(1, \underbrace{0, 0, \ldots, 0}_{m})$, then*

$$U(D) \geqslant \Omega(\log m).$$

To illustrate, Proposition 7.24 shows that the majority predicate $\mathrm{MAJ}_n(t) = -1 \Leftrightarrow t > n/2$ has degree 1 and unbounded-error complexity $\Theta(\log n)$. Other threshold predicates can be handled analogously.

Chapter 8

Alternation vs. Unbounded-Error Communication

We continue our study of unbounded-error communication complexity, focusing this time on the circuit class AC^0. The main result of this chapter is the first polynomial lower bound on the unbounded-error communication complexity of a function in AC^0. As a corollary, we establish the separations $\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}$ and $\Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}$ in communication complexity, thereby solving a longstanding open problem due to Babai et al. [23]. As another corollary, we obtain the first exponential, tight lower bound on the sign-rank of polynomial-size DNF formulas as well as the first exponential lower bound on the size of threshold-of-majority circuits for AC^0.

8.1 Introduction

In the previous chapter, we studied the unbounded-error communication complexity of symmetric functions. We continue the study of the unbounded-error model in the context of AC^0, the class of functions $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ computable by a polynomial-size constant-depth circuit of AND, OR, and NOT gates. Since AC^0 can compute the inner product function

$$f(x,y) = \bigoplus_{i=1}^{k} (x_i \wedge y_i)$$

on $k = \log^c n$ variables for any constant $c > 1$, work by Forster [70] gives a polylogarithmic lower bound on the unbounded-error communication complexity of AC^0. No stronger bound was known. The main result of this chapter is an exponentially stronger lower bound.

THEOREM 8.1 (Razborov and Sherstov [178]). *Define*

$$f_m(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij}).$$

Then

$$U(f_m) = \Omega(m).$$

We further show that the lower bound in Theorem 8.1 is almost tight, with a matching upper bound of $O(m \log m)$. Moreover, Theorem 8.1 is optimal with respect to circuit depth: it is straightforward to verify that AC^0 circuits of depth 1 and 2 have unbounded-error complexity $O(\log m)$. As a corollary, we obtain the separations

$$\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}, \qquad \Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}},$$

solving an open problem due to Babai et al. [23]. These separations are best possible in that $\mathsf{UPP}^{\mathsf{cc}}$ trivially contains the first two levels of the polynomial hierarchy: Σ_0^{cc}, Σ_1^{cc}, Π_0^{cc}, Π_1^{cc}.

As an application of our result to circuit complexity, we prove a lower bound of $\exp\{\Omega(m)\}$ on the size of any threshold-of-majority circuit that computes the function $f_m(x, y)$ above. This is the first exponential lower bound for threshold-of-majority circuits computing a function in AC^0. It substantially generalizes and strengthens earlier work by Krause and Pudlák [132].

As a final corollary, we obtain a lower bound of $\exp\{\Omega(n^{1/3})\}$ on the sign-rank of polynomial-size DNF formulas in n variables. This lower bound nearly matches the upper bound of $\exp\{\tilde{O}(n^{1/3})\}$ due to Klivans and Servedio [122]. Our result gives the first exponential, unconditional lower bound for learning polynomial-size DNF formulas in any reasonable model.

The remainder of this chapter is organized as follows. We start with an intuitive overview of our proof in Section 8.2, comparing and contrasting our approach with the work in the previous chapter. After a technical development that spans several sections, we arrive at our main result on unbounded-error communication complexity and circuit complexity in Section 8.7. The applications to complexity classes and learning theory are discussed in the concluding two sections of this chapter.

8.2 Overview of the proof

At a high level, we adopt the approach introduced in Chapter 7 in the context of determining the unbounded-error communication complexity of symmetric functions. This approach features three main steps.

(1) First, one proves that the Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$ of interest has a *smooth orthogonalizing distribution.* This term refers to a *smooth* distribution on $\{0,1\}^n$ (i.e., one that places nontrivial weight on all but a tiny fraction of the inputs) with respect to which f is orthogonal to all low-degree polynomials.

(2) Next, one applies to f the pattern matrix method of Chapter 4, which yields a sign matrix F that has low spectral norm with respect to a smooth distribution on the matrix entries. By design, the columns of F are applications of f to various subsets of n variables from among $x_1, x_2, \ldots, x_N$, where $N \geqslant 4n$.

(3) Finally, one invokes Forster's generalized lower bound [70, 71] on the sign-rank of matrices with low spectral norm and concludes that F has high unbounded-error communication complexity.

In Chapter 7, we carried out this three-step program for each symmetric function f. Our current focus, on the other hand, is the circuit class AC^0. Accordingly, we take the function f in the above program to be a suitable DNF formula. Our proof is devoted almost entirely to the first step of the program, i.e., showing that f has a smooth orthogonalizing distribution. Once this crucial property is settled, steps 2 and 3 are straightforward. We note that the implementation of step 1 is quite nontrivial and is unrelated to the development in Chapter 7.

Having described our proof at a high level, we will now examine it in more detail, from the bottom up. Figure 8.1 illustrates the main components of our proof. A starting point in our study is an elegant result due to Minsky and Papert [153], who constructed a linear-size DNF formula that cannot be sign-represented by polynomials of low degree.

Second, we revisit a fundamental technique from approximation theory, the *interpolation bound,* which bounds a degree-d univariate polynomial p on an interval based on the values of p at $d+1$ distinct points. By combining the interpolation

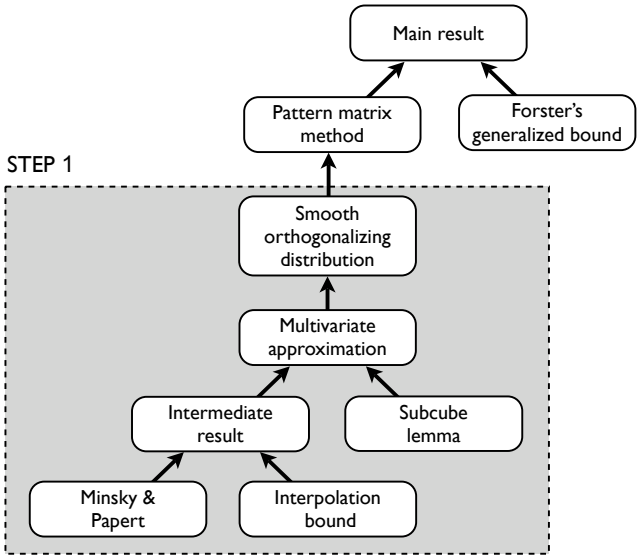

Figure 8.1: Proof outline.

bound with an adapted version of Minsky and Papert's argument, we establish a key intermediate result (Lemma 8.8). This result concerns multivariate polynomials that have nonnegligible agreement with the Minsky-Papert function and constrains their behavior on a large fraction of the inputs.

We proceed by deriving a Fourier-theoretic property common to all low-degree multivariate polynomials on $\{0,1\}^n$: we show that their values on $\{0,1\}^n$ can be conveniently bounded in terms of their behavior on certain small subcubes (Lemma 8.6). In light of this Fourier-theoretic observation, our intermediate result on multivariate polynomials takes on a much stronger form. Namely, we prove that multivariate polynomials with any nontrivial agreement with the Minsky-Papert function are highly constrained *throughout* the hypercube (Theorem 8.10). With some additional work in Section 8.5, we are able to deduce the existence of a smooth distribution on $\{0,1\}^n$ with respect to which the Minsky-Papert function is orthogonal to all low-degree polynomials. This completes step 1 of the above program, as desired.

8.3 Technical preliminaries

The following is a straightforward generalization of Minsky and Papert's symmetrization argument, Proposition 2.2.

PROPOSITION 8.2 (Razborov and Sherstov [178]). *Let $n_1, \dots, n_k$ be positive integers, $n = n_1 + \cdots + n_k$. Let $\phi\colon \{0,1\}^n \to \mathbb{R}$ be representable by a real n-variate polynomial of degree r. Write $x \in \{0,1\}^n$ as $x = (x^{(1)}, \dots, x^{(k)})$, where $x^{(i)} = (x_{n_1+\cdots+n_{i-1}+1}, \dots, x_{n_1+\cdots+n_i})$. Then there is a polynomial p on $\mathbb{R}^k$ of degree at most r such that*

$$\mathop{\mathbf{E}}_{\sigma_1 \in S_{n_1}, \dots, \sigma_k \in S_{n_k}} \Big[\phi(\sigma_1 x^{(1)}, \dots, \sigma_k x^{(k)})\Big] = p\Big(|x^{(1)}|, \dots, |x^{(k)}|\Big), \qquad x \in \{0,1\}^n.$$

Recall that vectors $v_1, \dots, v_n$ in $\mathbb{R}^r$ are said to be *in general position* if no r of them are linearly dependent. A powerful result due to Forster [70] states that any set of vectors in general position can be balanced in a useful way:

THEOREM 8.3 (Forster [70, Thm. 4.1]). *Let $U \subset \mathbb{R}^r$ be a finite set of vectors in general position, $|U| \geqslant r$. Then there is a nonsingular transformation $A \in \mathbb{R}^{r\times r}$ such that*

$$\sum_{u \in U} \frac{1}{\|Au\|^2} (Au)(Au)^{\mathsf{T}} = \frac{|U|}{r} I_r.$$

Implicit in Forster's work [70] is the following theorem, which is a key starting point in this chapter.

THEOREM 8.4 (Forster [70], implicit). *Let X, Y be finite sets, $M = [M_{xy}]_{x\in X, y\in Y}$ a real matrix ($M \neq 0$). Put $r = \operatorname{rk}_{\pm} M$. Then there is a matrix $R = [R_{xy}]_{x\in X, y\in Y}$*

such that:

$$\operatorname{rk} R = r, \tag{8.1}$$

$$M \circ R \geqslant 0, \tag{8.2}$$

$$\|R\|_\infty \leqslant 1, \tag{8.3}$$

$$\|R\|_{\mathrm{F}} = \sqrt{|X|\,|Y|/r}. \tag{8.4}$$

The notation $M \circ R \geqslant 0$ above means that all the entries in the matrix $M \circ R$ are nonnegative. The shorthand $M \neq 0$ means as usual that M is not the zero matrix. For completeness, we include a derivation of this result using Theorem 8.3.

PROOF OF THEOREM 8.4. Our treatment is closely analogous to Forster's derivation [70], p. 617. Since $M \neq 0$, it follows that $r \geqslant 1$. Fix a matrix $Q = [Q_{xy}]$ of rank r such that

$$Q_{xy} M_{xy} > 0 \qquad \text{whenever} \qquad M_{xy} \neq 0. \tag{8.5}$$

Write

$$Q = \Big[\langle u_x, v_y \rangle\Big]_{x \in X,\, y \in Y}$$

for suitable collections of vectors $\{u_x\} \subset \mathbb{R}^r$ and $\{v_y\} \subset \mathbb{R}^r$. If the vectors u_x, $x \in X$, are not already in general position, we can replace them with their slightly perturbed versions $\tilde{u}_x$ that *are* in general position. Provided that the perturbations are small enough, property (8.5) will still hold, i.e., we will have $\langle \tilde{u}_x, v_y \rangle M_{xy} > 0$ whenever $M_{xy} \neq 0$. As a result, we can assume w.l.o.g. that $\{u_x\}$ are in general position. Furthermore, a moment's reflection reveals that the vectors $\{v_y\}$ can be assumed to be all nonzero.

Since $\mathrm{rk}_{\pm} M \leqslant \mathrm{rk}\, M$, we infer that $|X| \geqslant r$. Theorem 8.3 is therefore applicable to the set $\{u_x\}$ and yields a nonsingular matrix A with

$$\sum_{x \in X} \frac{1}{\|Au_x\|^2} (Au_x)(Au_x)^{\mathsf{T}} = \frac{|X|}{r} I_r. \tag{8.6}$$

Define

$$R = \left[\frac{\langle u_x, v_y \rangle}{\|Au_x\| \, \|(A^{-1})^{\mathsf{T}} v_y\|} \right]_{x \in X,\, y \in Y}.$$

It remains to verify properties (8.1)–(8.4). Property (8.1) follows from the representation $R = D_1 Q D_2$, where D_1 and D_2 are diagonal matrices with strictly positive diagonal entries. By (8.5), we know that $R_{xy} M_{xy} > 0$ whenever $M_{xy} \neq 0$, which immediately gives us (8.2). Property (8.3) holds because

$$\frac{|\langle u_x, v_y \rangle|}{\|Au_x\| \, \|(A^{-1})^{\mathsf{T}} v_y\|} = \frac{|\langle Au_x, (A^{-1})^{\mathsf{T}} v_y \rangle|}{\|Au_x\| \, \|(A^{-1})^{\mathsf{T}} v_y\|} \leqslant 1.$$

Finally, property (8.4) will follow once we show that $\sum_x R_{xy}^2 = |X|/r$ for every $y \in Y$. So, fix $y \in Y$ and consider the unit vector $v = (A^{-1})^{\mathsf{T}} v_y / \|(A^{-1})^{\mathsf{T}} v_y\|$. We have:

$$\begin{aligned}
\sum_{x \in X} R_{xy}^2 &= \sum_{x \in X} \frac{\langle u_x, v_y \rangle^2}{\|Au_x\|^2 \, \|(A^{-1})^{\mathsf{T}} v_y\|^2} \\
&= \sum_{x \in X} \frac{(v_y^{\mathsf{T}} A^{-1})(Au_x)(Au_x)^{\mathsf{T}} (A^{-1})^{\mathsf{T}} v_y}{\|Au_x\|^2 \, \|(A^{-1})^{\mathsf{T}} v_y\|^2} \\
&= v^{\mathsf{T}} \left(\sum_{x \in X} \frac{1}{\|Au_x\|^2} (Au_x)(Au_x)^{\mathsf{T}} \right) v \\
&= \frac{|X|}{r},
\end{aligned}$$

where the last step follows from (8.6). □

8.4 A result on multivariate approximation

The purpose of this section is to establish a certain property of low-degree polynomials on $\mathbb{R}^m$ (Theorem 8.10). This property is the backbone of our main proof. A starting point in our discussion is an *interpolation bound,* i.e., a bound on the values of a polynomial on an interval given its values on a finite set of points. Results of this general form arise routinely in approximation theory. To prove the specific statement of interest to us, we follow the classical technique of interpolating the polynomial at strategically chosen points. For other uses of this technique, see Cheney [60, §7, Lem. 1] and Rivlin [185, Thm. 3.9].

LEMMA 8.5 (Razborov and Sherstov [178]). *Let $I \subset \mathbb{R}$ be an interval of length L. Let p be a polynomial of degree $d \leqslant L$ such that*

$$|p(x_i)| \leqslant 1 \qquad (i = 0, 1, \ldots, d),$$

where $x_0, x_1, \ldots, x_d \in I$ are some points with pairwise distances at least 1. Then

$$\max_{x \in I} |p(x)| \leqslant 2^d \binom{L}{d}.$$

PROOF. Without loss of generality, assume that $x_0 < x_1 < \cdots < x_d$. Fix $x \in I$. For any $k \in \{0, 1, \ldots, d\}$, we have:

$$\prod_{\substack{i=0 \\ i \neq k}}^{d} |x - x_i| \leqslant L(L-1)\cdots(L-d+1)$$

and, since $|x_i - x_k| \geqslant |i-k|$,

$$\prod_{\substack{i=0\\ i\neq k}}^{d} |x_k - x_i| \geqslant k!(d-k)!.$$

Therefore,

$$\prod_{\substack{i=0\\ i\neq k}}^{d} \frac{|x-x_i|}{|x_k-x_i|} \leqslant \frac{L(L-1)\cdots(L-d+1)}{k!(d-k)!} = \binom{L}{d}\binom{d}{k}.$$

It remains to substitute this estimate in the Lagrange interpolation formula:

$$|p(x)| = \left|\sum_{k=0}^{d} p(x_k) \prod_{\substack{i=0\\ i\neq k}}^{d} \frac{x-x_i}{x_k-x_i}\right| \leqslant \binom{L}{d} \sum_{k=0}^{d} \binom{d}{k} = 2^d \binom{L}{d}. \qquad \square$$

We now establish another auxiliary fact. It provides a convenient means to bound a function whose Fourier transform is supported on low-order characters, in terms of its behavior on low-weight inputs.

LEMMA 8.6 (Razborov and Sherstov [178]). *Let k be an integer, $0 \leqslant k \leqslant n-1$. Let $f : \{0,1\}^n \to \mathbb{R}$ be given with $\hat{f}(S) = 0$ for $|S| > k$. Then*

$$|f(1^n)| \;\leqslant\; 2^k \binom{n}{k} \max_{|x|\leqslant k} |f(x)|.$$

PROOF. Define the symmetric function $g\colon \{0,1\}^n \to \mathbb{R}$ by $g(x) = \chi_{[n]}(x)\,p(|x|)$, where

$$p(t) = \prod_{k<i<n} \frac{t-i}{n-i}.$$

The following properties of g are immediate:

$$g(1^n) = (-1)^n, \tag{8.7}$$
$$g(x) = 0 \qquad (k < |x| < n). \tag{8.8}$$

The degree of every monomial in g is between $k+1$ and n, so that

$$\hat{g}(S) = 0 \qquad (|S| \leqslant k). \tag{8.9}$$

Furthermore,

$$\sum_{|x|\leqslant k} |g(x)| = \sum_{t=0}^{k} \binom{n}{t} |p(t)| = \sum_{t=0}^{k} \binom{n}{t}\binom{n-t-1}{n-k-1} \leqslant 2^k \binom{n}{k}. \tag{8.10}$$

We are now prepared to analyze f. By (8.9),

$$\sum_{x\in\{0,1\}^n} f(x)g(x) = 0. \tag{8.11}$$

On the other hand, (8.7) and (8.8) show that

$$\sum_{x\in\{0,1\}^n} f(x)g(x) = (-1)^n f(1^n) + \sum_{|x|\leqslant k} f(x)g(x). \tag{8.12}$$

The lemma follows at once from (8.10)–(8.12). □

REMARK 8.7 (Razborov and Sherstov [178]). One can use Lemma 8.6 to bound f on inputs other than 1^n. For example, it follows immediately that $|f(y)| \leqslant 2^k \binom{|y|}{k} \max_{|x| \leqslant k} |f(x)|$, where $y \in \{0,1\}^n$ is arbitrary with $|y| > k$. We will not need this observation, however.

We are now in a position to study the approximation problem of interest to us. Define the sets

$$Z = \{0, 1, 2, \ldots, 4m^2\}^m, \qquad Z^+ = \{1, 2, \ldots, 4m^2\}^m.$$

Define $F \colon Z \to \{-1, +1\}$ by

$$F(z) = \begin{cases} -1 & \text{if } x \in Z^+, \\ 1 & \text{otherwise.} \end{cases}$$

For $u, z \in Z$, let $\Delta(u, z) = |\{i : u_i \neq z_i\}|$ be the ordinary Hamming distance. We shall prove the following intermediate result, inspired by Minsky and Papert's analysis [153] of the threshold degree of CNF formulas.

LEMMA 8.8 (Razborov and Sherstov [178]). *Let Q be a degree-d real polynomial in m variables, where $d \leqslant m/3$. Assume that*

$$F(z)Q(z) \geqslant -1 \qquad (z \in Z). \tag{8.13}$$

Then $|Q(z)| \leqslant 4^{m+d}$ at every point $z \in Z^+$ with $\Delta(u, z) < m/3$, where $u = (1^2, 3^2, 5^2, \ldots, (2m-1)^2) \in Z^+$.

PROOF. Fix $z \in Z^+$ with $\Delta(u, z) < m/3$. Define $p \in P_{2d}$ by

$$p(t) = Q(p_1(t), p_2(t), \ldots, p_m(t)),$$

where

$$p_i(t) = \begin{cases} (t - 2i + 1)^2 & \text{if } z_i = u_i \text{ (equivalently, } z_i = (2i-1)^2\text{),} \\ z_i & \text{otherwise.} \end{cases}$$

Letting $S = \{i : u_i = z_i\}$, inequality (8.13) implies that

$$p(2i-1) \geqslant -1 \qquad (i \in S), \tag{8.14}$$

$$p(2i) \leqslant 1 \qquad (i = 0, 1, \ldots, m). \tag{8.15}$$

CLAIM 8.9 (Razborov and Sherstov [178]). *Let $i \in S$. Then $|p(\xi)| \leqslant 1$ for some $\xi \in [2i-2, 2i-1]$.*

PROOF. The claim is trivial if p vanishes at some point in $[2i-2, 2i-1]$. In the contrary case, p maintains the same sign throughout this interval. As a result, (8.14) and (8.15) show that $\min\{|p(2i-2)|, |p(2i-1)|\} \leqslant 1$. □

Claim 8.9 provides $|S| > 2m/3 \geqslant 2d \geqslant \deg(p)$ points in $[0, 2m]$, with pairwise distances at least 1, at which p is bounded in absolute value by 1. By Lemma 8.5,

$$\max_{0 \leqslant t \leqslant 2m} |p(t)| \leqslant 2^{\deg(p)} \binom{2m}{\deg(p)} \leqslant 4^{m+d}.$$

This completes the proof since $Q(z) = p(0)$. □

Finally, we remove the restriction on $\Delta(u, z)$, thereby establishing the main result of this section.

THEOREM 8.10 (Razborov and Sherstov [178]). *Let Q be a degree-d real polynomial in m variables, where $d < m/3$. Assume that*

$$F(z)Q(z) \geqslant -1 \qquad (z \in Z).$$

Then

$$|Q(z)| \leqslant 16^m \qquad\qquad (z \in Z^+).$$

PROOF. As before, put $u = (1^2, 3^2, 5^2, \ldots, (2m-1)^2)$. Fix $z \in Z^+$ and define the "interpolating" function $f : \{0,1\}^m \to \mathbb{R}$ by

$$f(x) = Q(x_1 z_1 + (1-x_1)u_1, \ldots, x_m z_m + (1-x_m)u_m).$$

In this notation, we know from Lemma 8.8 that $|f(x)| \leqslant 4^{m+d}$ for every $x \in \{0,1\}^m$ with $|x| < m/3$, and our goal is to show that $|f(1^m)| \leqslant 16^m$. Since Q has degree d, the Fourier transform of f is supported on characters of order up to d. As a result,

$$\begin{aligned} |f(1^m)| &\leqslant 2^d \binom{m}{d} \max_{|x| \leqslant d} |f(x)| && \text{by Lemma 8.6} \\ &\leqslant 2^{2m+3d} \binom{m}{d} && \text{by Lemma 8.8} \\ &\leqslant 16^m. \end{aligned}$$

□

8.5 A smooth orthogonalizing distribution

The notion of a smooth orthogonalizing distribution, introduced in Chapter 7, plays a key role in this chapter as well. Let $f : \{0,1\}^n \to \{-1,+1\}$ be given. We say that a distribution μ on $\{0,1\}^n$ is *d-orthogonalizing* for f if

$$\mathop{\mathbf{E}}_{x \sim \mu}\Big[f(x)\chi_S(x)\Big] = 0 \qquad\qquad (|S| < d).$$

In words, a distribution μ is d-orthogonalizing for f if with respect to μ, the function f is orthogonal to every character of order less than d.

This section focuses on the function

$$\mathrm{MP}_m(x) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{4m^2} x_{i,j}.$$

Originally defined and studied by Minsky and Papert [153], this function played an important role in our results in Chapter 4 as well as other results in the literature [132, 163]. An explicit m-orthogonalizing distribution for MP_m is known [202]. However, our main result requires a $\Theta(m)$-orthogonalizing distribution for MP_m that is additionally *smooth,* i.e., places substantial weight on all but a tiny fraction of the points, and the distribution given in [202] severely violates the latter property. Proving the existence of a distribution that is simultaneously $\Theta(m)$-orthogonalizing and smooth is the goal of this section (Theorem 8.11).

We will view an input $x \in \{0,1\}^n = \{0,1\}^{4m^3}$ to MP_m as composed of blocks: $x = (x^{(1)}, \ldots, x^{(m)})$, where the ith block is $x^{(i)} = (x_{i,1}, x_{i,2}, \ldots, x_{i,4m^2})$. The proof that is about to start refers to the sets Z, Z^+ and the function F as defined in Section 8.4.

THEOREM 8.11 (Razborov and Sherstov [178]). *There is a $\frac{1}{3}m$-orthogonalizing distribution μ for MP_m such that $\mu(x) \geqslant \frac{1}{2}16^{-m}\,2^{-n}$ for all inputs $x \in \{0,1\}^n$ with $\mathrm{MP}_m(x) = -1$.*

PROOF. Let X be the set of all inputs with $\mathrm{MP}_m(x) = -1$, i.e.,

$$X = \{x \in \{0,1\}^n : x^{(1)} \neq 0, \ \ldots, \ x^{(m)} \neq 0\}.$$

It suffices to show that the following linear program has optimum at least $\frac{1}{2}16^{-m}$:

$$\begin{array}{lll}
\text{variables:} & \varepsilon \geqslant 0; \quad \mu(x) \geqslant 0 \text{ for } x \in \{0,1\}^n & \\
\text{maximize:} & \varepsilon & \\
\text{subject to:} & \displaystyle\sum_{x\in\{0,1\}^n} \mu(x)\mathrm{MP}_m(x)\chi_S(x) = 0 & \text{for } |S| < m/3, \\
& \displaystyle\sum_{x\in\{0,1\}^n} \mu(x) \leqslant 1, & \\
& \mu(x) \geqslant \varepsilon 2^{-n} & \text{for } x \in X.
\end{array} \tag{LP1}$$

The optimum being nonzero, it will follow by a scaling argument that any optimal solution has $\sum \mu(x) = 1$. As a result, μ will be the sought probability distribution.

For $x \in \{0,1\}^n$, we let $z(x) = (|x^{(1)}|, \ldots, |x^{(m)}|)$; note that $\mathrm{MP}_m(x) = F(z(x))$. Since the function MP_m is invariant under the action of the group $S_{4m^2} \times \cdots \times S_{4m^2}$, in view of Proposition 8.2, the dual of (LP1) can be simplified as follows:

$$\begin{array}{lll}
\text{variables:} & \text{a polynomial } Q \text{ on } \mathbb{R}^m \text{ of degree} < m/3; & \\
& \eta \geqslant 0; \quad \delta_z \geqslant 0 \text{ for } z \in Z^+ & \\
\text{minimize:} & \eta & \\
\text{subject to:} & \displaystyle\sum_{x\in X} \delta_{z(x)} \geqslant 2^n, & \\
& F(z)Q(z) \geqslant -\eta & \text{for } z \in Z, \\
& F(z)Q(z) \geqslant -\eta + \delta_z & \text{for } z \in Z^+.
\end{array} \tag{LP2}$$

The programs are both feasible and therefore have the same finite optimum. Fix an optimal solution η, Q, δ_z to (LP2). For the sake of contradiction, assume that $\eta \leqslant \frac{1}{2}16^{-m}$. Then $|Q(z)| \leqslant \frac{1}{2}$ for each $z \in Z^+$, by Theorem 8.10. From the constraints of the third type in (LP2) we conclude that $\delta_z \leqslant \frac{1}{2} + \eta < 1$ ($z \in Z^+$). This contradicts the first constraint. Thus, the optimum of (LP1) and (LP2) is at least $\frac{1}{2}16^{-m}$. □

8.6 Generalization of Forster's bound

Using Theorem 8.4, Forster [70] gave a simple proof of the following fundamental result: for any matrix $A = [A_{xy}]_{x\in X,\, y\in Y}$ with ± 1 entries,

$$\operatorname{rk}_{\pm} A \geqslant \frac{\sqrt{|X|\,|Y|}}{\|A\|}.$$

Forster et al. [71, Thm. 3] generalized this bound to arbitrary real matrices $A \neq 0$:

$$\operatorname{rk}_{\pm} A \geqslant \frac{\sqrt{|X|\,|Y|}}{\|A\|} \cdot \min_{x,y} |A_{xy}|. \tag{8.16}$$

Forster and Simon [73, §5] considered a different generalization, inspired by the notion of matrix rigidity (see, e.g., [175]). Let A be a given ± 1 matrix, and let $\tilde{A}$ be obtained from A by changing some h entries in an arbitrary fashion ($h < |X|\,|Y|$). Forster and Simon showed that

$$\operatorname{rk}_{\pm} \tilde{A} \geqslant \frac{\sqrt{|X|\,|Y|}}{\|A\| + 2\sqrt{h}}. \tag{8.17}$$

The above generalizations are not sufficient for our purposes. Before we can proceed, we need to prove the following "hybrid" bound, which combines the ideas of [70, 71, 73].

THEOREM 8.12 (Razborov and Sherstov [178]). *Let $A = [A_{xy}]_{x\in X,\, y\in Y}$ be a real matrix with $s = |X|\,|Y|$ entries ($A \neq 0$). Assume that all but h of the entries of A satisfy $|A_{xy}| \geqslant \gamma$, where h and $\gamma > 0$ are arbitrary parameters. Then*

$$\operatorname{rk}_{\pm} A \geqslant \frac{\gamma s}{\|A\|\sqrt{s} + \gamma h}.$$

PROOF. Let r denote the sign-rank of A. Theorem 8.4 supplies a matrix $R = [R_{xy}]$ with

$$\operatorname{rk} R = r, \tag{8.18}$$
$$A \circ R \geqslant 0, \tag{8.19}$$
$$\|R\|_\infty \leqslant 1, \tag{8.20}$$
$$\|R\|_{\mathrm{F}} = \sqrt{s/r}. \tag{8.21}$$

The crux of the proof is to estimate $\langle A, R\rangle$ from below and above. On the one hand,

$$\begin{aligned}
\langle A, R\rangle &\geqslant \sum_{x,y:\,|A_{xy}|\geqslant\gamma} A_{xy}R_{xy} && \text{by (8.19)}\\
&\geqslant \gamma\left(\sum_{x,y}|R_{xy}| - h\right) && \text{by (8.19), (8.20)}\\
&\geqslant \gamma\|R\|_{\mathrm{F}}^2 - \gamma h && \text{by (8.20)}\\
&= \frac{\gamma s}{r} - \gamma h && \text{by (8.21).}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\langle A, R\rangle &\leqslant \|A\| \cdot \|R\|_\Sigma && \text{by (2.5)}\\
&\leqslant \|A\| \cdot \|R\|_{\mathrm{F}}\,\sqrt{r} && \text{by (2.4), (8.18)}\\
&= \|A\|\sqrt{s} && \text{by (8.21).}
\end{aligned}$$

Comparing these lower and upper bounds on $\langle A, R\rangle$ yields the claimed estimate of $r = \operatorname{rk}_\pm A$. □

8.7 Main result and circuit consequences

At last, we are in a position to prove our main result. It will be convenient to first obtain a matrix-analytic formulation and then infer the sought statement on unbounded-error communication complexity.

THEOREM 8.13 (Razborov and Sherstov [178]). *Define*

$$f_m(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij}).$$

Then the matrix $[f_m(x,y)]_{x,y}$ *has sign-rank* $2^{\Omega(m)}$.

PROOF. Let M be the (N, n, MP_m)-pattern matrix, where $n = 4m^3$ and $N = 17^6 n$. Let P be the (N, n, μ)-pattern matrix, where μ is the distribution from Theorem 8.11. We are going to estimate the sign-rank of $M \circ P$.

By Theorem 8.11, all but a $2^{-\Omega(m^2)}$ fraction of the inputs $x \in \{0,1\}^n$ satisfy $\mu(x) \geqslant \frac{1}{2} 16^{-m} 2^{-n}$. As a result, all but a $2^{-\Omega(m^2)}$ fraction of the entries of $M \circ P$ are at least $\frac{1}{2} 16^{-m} 2^{-n}$ in absolute value. Theorem 8.12 at once implies that

$$\mathrm{rk}_{\pm}\, M \geqslant \mathrm{rk}_{\pm}\, M \circ P \geqslant \min \left\{ \frac{16^{-m} 2^{-n} \sqrt{s}}{4 \|M \circ P\|},\ 2^{\Omega(m^2)} \right\}, \tag{8.22}$$

where $s = 2^{N+n} \left(\frac{N}{n}\right)^n$ denotes the number of entries in $M \circ P$.

We now bound the spectral norm of $M \circ P$ precisely as in [204, §6]. Note first that $M \circ P$ is the (N, n, ϕ)-pattern matrix, where $\phi\colon \{0,1\}^n \to \mathbb{R}$ is given by $\phi(x) = \mathrm{MP}_m(x)\mu(x)$. Since μ is a $\frac{1}{3}m$-orthogonalizing distribution for MP_m, we have

$$\hat{\phi}(S) = 0, \qquad |S| < m/3. \tag{8.23}$$

Since $\sum_{x\in\{0,1\}^n} |\phi(x)| = 1$, Proposition 2.1 shows that

$$|\hat{\phi}(S)| \leqslant 2^{-n}, \qquad S \subseteq \{1, 2, \ldots, n\}. \tag{8.24}$$

Theorem 4.3 implies, in view of (8.23) and (8.24), that

$$\|M \circ P\| \leqslant \sqrt{s} \cdot 2^{-n} \left(\frac{N}{n}\right)^{-m/6} = 17^{-m}\, 2^{-n} \sqrt{s}.$$

Along with (8.22), this estimate shows that M has sign-rank at least $2^{\Omega(m)}$. It remains to note that M is a submatrix of $[f_{cm}(x, y)]_{x,y}$, where $c = \lceil \sqrt{8N/n} \rceil = \Theta(1)$. □

In the language of communication complexity, we have the following interpretation of our matrix-analytic result.

THEOREM 8.1 (Razborov and Sherstov [178], restated). *Define*

$$f_m(x, y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij}).$$

Then $U(f_m) = \Omega(m)$.

PROOF. Immediate from Theorems 7.2 and 8.13. □

REMARK 8.14 (Razborov and Sherstov [178]). The lower bounds in Theorems 8.1 and 8.13 are essentially optimal. To see this, note that the matrix $[f_m(x, y)]_{x,y}$ has the same sign pattern as

$$R = \left[\frac{1}{2} - \prod_{i=1}^{m} \left(\sum_{j=1}^{m^2} x_{ij}\, y_{ij}\right)\right]_{x,y}.$$

Therefore, the sign-rank of $[f_m(x,y)]_{x,y}$ does not exceed $m^{2m}+1 = 2^{O(m\log m)}$. In view of Theorem 7.2, we obtain $U(f_m) = O(m\log m)$.

These results have the following implication in circuit complexity.

THEOREM 8.15 (Razborov and Sherstov [178]). *Define*

$$f_m(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij}).$$

Let C be a depth-2 threshold circuit, with arbitrary weights at the top gate and integer weights of absolute value at most W at the bottom gates. If C computes f_m, then it has $2^{\Omega(m)}/W$ gates.

PROOF. Immediate from Theorems 7.3 and 8.13. □

Theorem 8.15 gives the first exponential lower bound for threshold-of-majority circuits computing a function in AC^0. It substantially generalizes and strengthens an earlier result of Krause and Pudlák [132, Thm. 2], who proved an exponential lower bound for threshold-of-MOD_r circuits (for any constant $r \geqslant 2$) computing a function in AC^0. Theorem 8.15 also complements our exponential lower bound for majority-of-threshold circuits computing functions in AC^0, stated as Theorem 4.19 above.

8.8 Separation of the polynomial hierarchy from $\mathsf{UPP}^{\mathsf{cc}}$

We now explore some consequences of this chapter for complexity classes, introduced in Section 3.4. A ready consequence of our work is the following separation of the polynomial hierarchy from $\mathsf{UPP}^{\mathsf{cc}}$.

THEOREM 8.16 (Razborov and Sherstov [178]).

$$\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}, \quad \Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}.$$

PROOF. The family $\{f_m\}$ of Theorem 8.13 clearly satisfies $\{f_m\} \notin \mathsf{UPP}^{\mathsf{cc}}$ and hence $\{\neg f_m\} \notin \mathsf{UPP}^{\mathsf{cc}}$. On the other hand, the memberships $\{f_m\} \in \Pi_2^{\mathsf{cc}}$ and $\{\neg f_m\} \in \Sigma_2^{\mathsf{cc}}$ follow directly from the definition of Σ_2^{cc} and Π_2^{cc}. □

Several years prior to our work, Forster [70] proved that the inner product function $\mathrm{IP}_n(x, y) = \bigoplus_{i=1}^n (x_i \wedge y_i)$ has unbounded-error communication complexity $\Theta(n)$. Since $\{\mathrm{IP}_n\} \in \mathsf{PSPACE}^{\mathsf{cc}}$, Forster's result yields the separation $\mathsf{PSPACE}^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}$. Theorem 8.16 of this chapter substantially strengthens it, showing that even the second level Σ_2^{cc}, Π_2^{cc} of the polynomial hierarchy is not contained in $\mathsf{UPP}^{\mathsf{cc}}$. This settles the open problem due to Babai et al. [23], p. 345, who asked whether $\Sigma_2^{\mathsf{cc}} \subseteq \mathsf{UPP}^{\mathsf{cc}}$. Observe that Theorem 8.16 is best possible in that $\mathsf{UPP}^{\mathsf{cc}}$ trivially contains Σ_0^{cc}, Σ_1^{cc}, Π_0^{cc}, and Π_1^{cc}.

Recall that $\mathsf{UPP}^{\mathsf{cc}}$, the class of communication problems with low sign-rank, is related to the class $\mathsf{PP}^{\mathsf{cc}}$ of communication problems with nonnegligible discrepancy. In Chapter 10, we will see that $\mathsf{PP}^{\mathsf{cc}}$ is a small subset of $\mathsf{UPP}^{\mathsf{cc}}$. Therefore, Theorem 8.16 strengthens our earlier separations

$$\Sigma_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}}, \qquad \Pi_2^{\mathsf{cc}} \not\subseteq \mathsf{PP}^{\mathsf{cc}},$$

obtained in Corollary 4.17.

8.9 Sign-rank of DNF formulas

We close this chapter with applications to learning theory. Recall from Section 7.2 that concept classes (equivalently, sign matrices) with low sign-rank admit efficient learning algorithms in the PAC model. In particular, the current fastest algorithm for PAC-learning polynomial-size DNF formulas, due to Klivans and Servedio [122], was obtained precisely by placing an upper bound of $\exp\{\tilde{O}(n^{1/3})\}$ on the sign-rank of that concept class. Our work gives a matching lower bound.

THEOREM 8.17 (Razborov and Sherstov [178]). *Let $\mathscr{C}$ be the set of all read-once (hence, linear-size) DNF formulas $f : \{0,1\}^n \to \{-1,+1\}$. Then*

$$\mathrm{rk}_{\pm}\, \mathscr{C} = \exp\{\Omega(n^{1/3})\}.$$

PROOF. Let $f_m(x,y)$ be the function from Theorem 8.13, where $m = \lfloor n^{1/3} \rfloor$. Then the matrix $[-f_m(x,y)]_{x,y}$ has sign-rank $\exp\{\Omega(n^{1/3})\}$, and each of its rows is the truth table of a read-once DNF formula $\Phi_x(y) = \neg f_m(x,y)$. □

Learning polynomial-size DNF formulas was the original challenge posed in Valiant's paper [213]. More than twenty years later, this challenge remains a central open problem in computational learning theory despite active research [214, 106, 17, 18, 217, 19, 7, 89, 136, 6, 8, 39, 150, 46, 90, 138, 5, 189, 48, 47, 211, 122]. To account for this lack of progress, several hardness results have been obtained based on complexity-theoretic assumptions [112, 10]. Theorem 8.17 complements that line of work by exhibiting an unconditional, *structural* barrier to the efficient learning of DNF formulas. In particular, it rules out a $2^{o(n^{1/3})}$-time learning algorithm based on Euclidean embeddings.

While restricted, the Euclidean embedding paradigm is quite rich and captures many PAC learning algorithms designed to date, with the notable exception [94, 38] of learning low-degree polynomials over GF(p). Furthermore, it is known [108, p. 124] that an unconditional superpolynomial lower bound for learning polynomial-size DNF formulas in the *standard* PAC model would imply that $\mathsf{P} \neq \mathsf{NP}$; thus, such a result is well beyond the reach of the current techniques.

Chapter 9

Multiparty Communication

The pattern matrix of Chapter 4 has been adapted to *multiparty* communication and has enabled substantial progress in the area. We give a detailed and integrated treatment of these developments, which we hope will serve as a helpful and self-contained reference and spur further progress in multiparty communication. Covered here are the improved lower bounds for the disjointness function due to Lee and Shraibman [140] and Chattopadhyay and Ada [59], a separation of $\mathsf{NP}^{\mathsf{cc}}$ from $\mathsf{BPP}^{\mathsf{cc}}$ due to David, Pitassi, and Viola [65], and a separation of $\mathsf{NP}^{\mathsf{cc}}$ from $\mathsf{coNP}^{\mathsf{cc}}$ and $\mathsf{coMA}^{\mathsf{cc}}$ due to Gavinsky and Sherstov [81].

9.1 Introduction

In our development so far, we have focused on two-party communication. The subject of this chapter is communication with three and more parties, a model introduced by Chandra, Furst, and Lipton [57]. Analogous to the two-party setting, the multiparty model of communication features k communicating players whose goal is to compute a given function. More precisely, one considers a Boolean function $f\colon X_1 \times \cdots \times X_k \to \{-1,+1\}$ whose arguments $x_1 \in X_1, \ldots, x_k \in X_k$ are placed on the foreheads of players 1 through k, respectively. Thus, player i sees all the arguments except for x_i. The players communicate by writing bits on a shared blackboard, visible to all. Their goal is to compute $f(x_1, \ldots, x_k)$ with minimum communication. Analogous to the two-party case, the multiparty model naturally admits definitions of deterministic $D(f)$, nondeterministic $N(f)$, co-nondeterministic $N(-f)$, and randomized $R_\varepsilon(f)$ communication complexity, along with the corresponding communication classes $\mathsf{P}^{\mathsf{cc}}_k$, $\mathsf{NP}^{\mathsf{cc}}_k$, $\mathsf{coNP}^{\mathsf{cc}}_k$, $\mathsf{BPP}^{\mathsf{cc}}_k$. We defer the formal details to Section 9.2. The multiparty model has found a variety of applications, including circuit complexity, pseudorandomness, and proof complexity [225, 93, 25, 179, 30]. This model draws its richness from the generous overlap in the players' inputs, which makes it challenging to prove lower bounds. Many fundamental questions in the multiparty model remain open despite much research.

In this chapter, we discuss progress in multiparty communication complexity enabled by the pattern matrix method. Generalized discrepancy and the pattern matrix method, presented in Chapter 4 in the context of two-party communication, readily adapt to three and more players. This adaptation was formalized by Lee

and Shraibman [140] and independently by Chattopadhyay and Ada [59], resulting in much improved lower bounds for the disjointness function. We present this development in detail in Sections 9.3 and 9.4.

The next result that we discuss is an explicit separation of the multiparty communication class NP^{cc}_k from BPP^{cc}_k for up to $k = (1-\varepsilon)\log n$ players, obtained by David, Pitassi, and Viola [65]. Here $\varepsilon > 0$ is an arbitrary constant. Since the current barrier for explicit lower bounds on multiparty communication complexity is precisely $k = \log n$, this work matches the state of the art. At the technical level, the authors of [65] combine the multiparty adaptation of the pattern matrix method with an ingenious use of the probabilistic method. We present this result in Section 9.5.

The final result discussed in this chapter, due to Gavinsky and Sherstov [81], gives an explicit separation of the multiparty communication classes NP^{cc}_k and coNP^{cc}_k for up to $k = (1-\varepsilon)\log n$ players. Prior to our work, it was unknown whether these communication classes were equal for any $k \geqslant 3$. We are further able to give an explicit separation of NP^{cc}_k from coMA^{cc}_k, the class of *Merlin-Arthur* computations [22, 24] that combines the power of nondeterminism and randomization. In particular, the latter separation subsumes the separation of NP^{cc}_k from BPP^{cc}_k by David et al. [65].

9.2 Multiparty models and complexity classes

The basic two-party models of communication, reviewed in Section 3.1, have analogues for three and more players [57]. In the case of k players, one considers a function $f\colon X_1 \times \cdots \times X_k \to \{-1,+1\}$ for some finite sets $X_1, \ldots, X_k$. A given input $(x_1, \ldots, x_k) \in X_1 \times \cdots \times X_k$ is distributed among the players by placing x_i on the forehead of player i (for $i = 1, \ldots, k$). In other words, player i knows $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$ but not x_i. The players can communicate by writing bits 0 and 1 on a shared blackboard, visible to all, according to a protocol established in advance. Analogous to the two-party case, a protocol is a fixed agreement among the k players that specifies:

(1) for each sequence of bits written on the blackboard, an output value -1 or $+1$ if the communication is over, and an indication of who is to speak next if the communication is to continue;

(2) for the player to speak next, a value 0 or 1 that is to be written on the blackboard, based on the current state of the blackboard and the $k-1$ parts of the input visible to the player (namely, $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$ for player i).

The *cost* of a protocol is the maximum number of bits written on the blackboard for any input $(x_1, \ldots, x_k)$. A protocol is said to compute f *deterministically* if the output of the protocol on input $(x_1, \ldots, x_k)$ is always $f(x_1, \ldots, x_k)$. A protocol is said to compute f *nondeterministically* if the protocol always outputs $+1$ on inputs $(x_1, \ldots, x_k) \in f^{-1}(+1)$ and outputs -1 at least on some executions for every input $(x_1, \ldots, x_k) \in f^{-1}(-1)$. The *deterministic* (respectively, *nondeterministic*) communication complexity of f is the least cost of a protocol that computes f deterministically (respectively, nondeterministically). The multiparty deterministic and nondeterministic communication complexities of f are denoted $D(f)$ and $N(f)$, respectively. The *co-nondeterministic* communication complexity of f is the quantity $N(-f)$.

In the *randomized* model, the k parties additionally have access to an unlimited supply of shared random bits. The cost of a randomized protocol is still the maximum number of bits written on the blackboard on any input. A randomized protocol is said to *compute f with error ε* if on every input $(x_1, \ldots, x_k)$, the protocol produces the correct output $f(x_1, \ldots, x_k)$ with probability at least $1-\varepsilon$. The *ε-error randomized* communication complexity of f, denoted $R_\varepsilon(f)$, is the least cost of a randomized protocol that computes f with error ε. The canonical setting is $\varepsilon = 1/3$, corresponding to *bounded-error* randomized communication complexity, but any other parameter $\varepsilon \in (0, 1/2)$ can be considered. As in the two-party case, it is useful to keep in mind that the error probability of a randomized protocol can be reduced from $1/3$ to any desired constant $\varepsilon > 0$ by executing the protocol $\Theta\left(\log \frac{1}{\varepsilon}\right)$ times and outputting the majority answer. In other words, one has

$$R_\varepsilon(f) = O\left(R_{1/3}(f) \log \frac{1}{\varepsilon}\right)$$

by basic probability, and thus the setting $\varepsilon = 1/3$ entails no loss of generality in the study of bounded-error communication complexity.

The *Merlin-Arthur* model combines the power of the randomized and nondeterministic models. A Merlin-Arthur protocol starts with a nondeterministic guess of c bits appearing on the shared blackboard, where the value of c is fixed in advance. From that point on, the players communicate using a randomized protocol. The cost of a Merlin-Arthur protocol is the sum of c and the maximum number of bits written on the blackboard after the nondeterministic guess. A Merlin-Arthur protocol is said to *compute* f *with error* ε if (1) on every input $(x_1, \ldots, x_k) \in f^{-1}(+1)$, the protocol produces the correct output with probability at least $1 - \varepsilon$ regardless of the nondeterministic guess; and (2) on every input $(x_1, \ldots, x_k) \in f^{-1}(-1)$, there is at least one nondeterministic guess for which the protocol produces the correct output with probability at least $1 - \varepsilon$. The *ε-error Merlin-Arthur* communication complexity of f, denoted $MA_\varepsilon(f)$, is the least cost of a Merlin-Arthur protocol that computes f with error ε. Arthur-Merlin computations were originally considered in [22, 24].

One defines k-party communication classes analogous to the two-party communication classes in Section 3.4. A family of functions $f_n : (\{0,1\}^n)^k \to \{-1,+1\}$, for $n = 1, 2, 3, \ldots$, belongs to $\mathsf{P}^{\mathsf{cc}}_k$ if and only if $D(f) \leqslant \log^c n$ for some constant $c > 1$ and all $n > c$. The multiparty classes $\mathsf{NP}^{\mathsf{cc}}_k$, $\mathsf{BPP}^{\mathsf{cc}}_k$, $\mathsf{MA}^{\mathsf{cc}}_k$ are defined analogously with regard to nondeterministic, $\frac{1}{3}$-error randomized, and $\frac{1}{3}$-error Merlin-Arthur communication complexity. A family of functions $\{f_n\}$ belongs to $\mathsf{coNP}^{\mathsf{cc}}_k$ if and only if $\{\neg f_n\} \in \mathsf{NP}^{\mathsf{cc}}_k$. Similarly, a family of functions $\{f_n\}$ belongs to $\mathsf{coMA}^{\mathsf{cc}}_k$ if and only if $\{\neg f_n\} \in \mathsf{MA}^{\mathsf{cc}}_k$.

The multiparty communication models reviewed above are known as *number-on-forehead* models in reference to how the input is distributed among the k players. Other generalizations of two-party communication to multiple parties have been considered in the literature, such as *number-in-hand* models [137].

9.3 Discrepancy and generalized discrepancy

The two-party discrepancy method and its generalization, discussed in Sections 3.2 and 3.3, adapt in a natural way to the multiparty setting. Fix a function $f : X_1 \times$

$\cdots \times X_k \to \{-1,+1\}$ and a distribution μ on $X_1 \times \cdots \times X_k$. The *discrepancy of f with respect to μ* is defined as

$$\operatorname{disc}_\mu(f) = \max_{\phi_1,\ldots,\phi_k} \left| \sum_{\substack{(x_1,\ldots,x_k) \\ \in X_1\times\cdots\times X_k}} \psi(x_1,\ldots,x_k) \prod_{i=1}^{k} \phi_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k) \right|,$$

where $\psi(x_1,\ldots,x_k) = f(x_1,\ldots,x_k)\mu(x_1,\ldots,x_k)$ and the maximum ranges over all functions $\phi_i \colon X_1 \times \cdots X_{i-1} \times X_{i+1} \times \cdots X_k \to \{0,1\}$, for $i = 1,2,\ldots,k$. Note that for $k = 2$, this definition is identical to the one given previously for the two-party model. We put

$$\operatorname{disc}(f) = \min_\mu \{\operatorname{disc}_\mu(f)\}.$$

We identify a function $f \colon X_1 \times \cdots \times X_k \to \{-1,+1\}$ with its *communication tensor* $F(x_1,\ldots,x_k) = [f(x_1,\ldots,x_k)]_{x_1,\ldots,x_k}$ and speak of the discrepancy of F and f interchangeably, and likewise for other complexity measures such as $R_\varepsilon(f)$. This identification is convenient because it allows us to use the tensor notation of Section 2.4, such as the Hadamard product $A \circ B$ or inner product $\langle A, B\rangle$.

Discrepancy is difficult to analyze as defined. Typically, one uses the following well-known estimate, derived by repeated applications of the Cauchy-Schwarz inequality.

THEOREM 9.1 (Discrepancy estimate [25, 62, 173]). *Fix $f \colon X_1 \times \cdots \times X_k \to \{-1,+1\}$ and a probability distribution μ on $X_1 \times \cdots \times X_k$. Put $\psi(x_1,\ldots,x_k) = f(x_1,\ldots,x_k)\mu(x_1,\ldots,x_k)$. Then*

$$\left(\frac{\operatorname{disc}_\mu(f)}{|X_1|\cdots|X_k|}\right)^{2^{k-1}} \leqslant \mathop{\mathbf{E}}_{\substack{x_1^0 \in X_1 \\ x_1^1 \in X_1}} \cdots \mathop{\mathbf{E}}_{\substack{x_{k-1}^0 \in X_{k-1} \\ x_{k-1}^1 \in X_{k-1}}} \left| \mathop{\mathbf{E}}_{x_k \in X_k} \prod_{z \in \{0,1\}^{k-1}} \psi(x_1^{z_1},\ldots,x_{k-1}^{z_{k-1}},x_k) \right|.$$

Observe that for $k = 2$ players, Theorem 9.1 reduces to its two-party analogue, Lemma 3.3.

For a function $f: X_1 \times \cdots \times X_k \to \{-1, +1\}$ and a distribution μ over $X_1 \times \cdots \times X_k$, the *$\varepsilon$-error μ-distributional* communication complexity $D^{\mu}_{\varepsilon}(f)$ is the least cost of a deterministic protocol for f whose probability of error with respect to μ is at most ε. Since a randomized protocol can be viewed as a probability distribution over deterministic protocols, we immediately have that

$$R_{\varepsilon}(f) \geqslant \max_{\mu} D^{\mu}_{\varepsilon}(f). \tag{9.1}$$

We are now prepared to state the discrepancy method for multiparty communication, which is identical to its two-party counterpart.

THEOREM 9.2 (Discrepancy method; see [137]). *For every function $f: X_1 \times \cdots \times X_k \to \{-1, +1\}$, every distribution μ on $X_1 \times \cdots \times X_k$, and every $\gamma \in (0, 1)$,*

$$R_{1/2-\gamma/2} \geqslant D^{\mu}_{1/2-\gamma/2}(f) \geqslant \log\left(\frac{\gamma}{\operatorname{disc}_{\mu}(f)}\right).$$

The two-party generalized discrepancy method, given by Theorem 3.7, extends word-for-word to the multiparty model. This extension was formalized by Lee and Shraibman [140] and independently by Chattopadhyay and Ada [59].

THEOREM 9.3 (Multiparty generalized discrepancy [140, 59]). *Let $F: X_1 \times \cdots \times X_k \to \{-1, +1\}$ be a given sign tensor, for some finite sets $X_1, \ldots, X_k$. Then for all sign tensors $H: X_1 \times \cdots \times X_k \to \{-1, +1\}$ and all probability tensors P,*

$$2^{R_{\varepsilon}(F)} \geqslant \frac{\langle F, H \circ P \rangle - 2\varepsilon}{\operatorname{disc}_P(H)}. \tag{9.2}$$

The presentation below differs slightly from the proofs in [140, 59] and is meant to emphasize the direct analogy between the two-party and multiparty cases.

PROOF OF THEOREM 9.3. Put $c = R_\varepsilon(F)$. Equation (9.1) shows that there exists a deterministic protocol $\Pi\colon X_1 \times \cdots \times X_k \to \{-1, +1\}$ with communication cost at most c and $\mathbf{P}_P[F(x_1, \ldots, x_k) \neq \Pi(x_1, \ldots, x_k)] \leqslant \varepsilon$. Then

$$\langle \Pi, H \circ P \rangle \geqslant \langle F, H \circ P \rangle - 2\varepsilon.$$

On the other hand, the ordinary discrepancy method (Theorem 9.2) states that

$$\langle \Pi, H \circ P \rangle \leqslant 2^c \operatorname{disc}_P(H).$$

Comparing the last two inequalities completes the proof. □

9.4 Lower bounds for the disjointness function

In this section, we present an extension of the pattern matrix method to the multiparty model. This extension, formalized in [58, 140, 59], closely follows the combinatorial analysis of pattern matrices in Section 4.8.

We start with some notation. Fix a function $\phi\colon \{0,1\}^t \to \mathbb{R}$ and an integer n with $t \mid n$. Define the (k, n, t, ϕ)-*pattern tensor* as the k-argument function $A\colon \{0,1\}^{t(n/t)^{k-1}} \times [n/t]^t \times \cdots \times [n/t]^t \to \mathbb{R}$ given by $A(x, V_1, \ldots, V_{k-1}) = \phi(x|_{V_1,\ldots,V_{k-1}})$, where we define

$$x|_{V_1,\ldots,V_{k-1}} = \big(x_{1,V_1[1],\ldots,V_{k-1}[1]},\ \ldots,\ x_{t,V_1[t],\ldots,V_{k-1}[t]}\big) \in \{0,1\}^t$$

and $V_j[i]$ denotes the ith element of the t-dimensional vector V_j. Note that we index the string x by viewing it as a k-dimensional array of $t \times (n/t) \times \cdots \times (n/t) = t(n/t)^{k-1}$ bits. This definition generalizes the pattern matrices of Section 4.8 to higher dimensions.

We are ready for the first result of this section, namely, an extension of the two-party Theorem 4.23 to the multiparty model. This extension was originally

obtained by Chattopadhyay [58] for slightly different tensors and has since been revisited in one form or another [140, 59].

THEOREM 9.4 (Discrepancy of pattern tensors [58, 140, 59]). *Fix a function $h\colon \{0,1\}^t \to \{-1,+1\}$ and a probability distribution μ on $\{0,1\}^t$. Let n be a given integer, $t \mid n$. Let H be the (k,n,t,f)-pattern tensor and let P be the $(k,n,t,2^{-t(n/t)^{k-1}+t}(n/t)^{-t(k-1)}\mu)$-pattern tensor. Let $d \geqslant 0$ be an integer such that $\widehat{\mu h}(S) = 0$ whenever $|S| < d$. If $n \geqslant 4\mathrm{e}t^2(k-1)2^{2^{k-1}}/d$, then*

$$\operatorname{disc}_P(H) \leqslant 2^{-d/2^{k-1}}.$$

PROOF (adapted from [58, 140, 59]). Consider the function $\psi\colon \{0,1\}^t \to \mathbb{R}$ given by $\psi(z) \equiv f(z)\mu(z)$. By Theorem 9.1,

$$\operatorname{disc}_P(H)^{2^{k-1}} \leqslant 2^{t2^{k-1}} \mathop{\mathbf{E}}_{\mathbf{V}} |\Gamma(\mathbf{V})|, \tag{9.3}$$

where we put $\mathbf{V} = (V_1^0, V_1^1, \ldots, V_{k-1}^0, V_{k-1}^1)$ and

$$\Gamma(\mathbf{V}) = \mathop{\mathbf{E}}_{x}\Bigg[\underbrace{\psi\left(x|_{V_1^0,V_2^0,\ldots,V_{k-1}^0}\right)}_{(\dagger)} \underbrace{\prod_{z\in\{0,1\}^{k-1}\setminus\{0^{k-1}\}} \psi\left(x|_{V_1^{z_1},V_2^{z_2},\ldots,V_{k-1}^{z_{k-1}}}\right)}_{(\ddagger)} \Bigg].$$

For a fixed choice of $\mathbf{V}$, define sets

$$A = \left\{(i, V_1^0[i], \ldots, V_{k-1}^0[i]) :\ i = 1,2,\ldots,t\right\},$$
$$B = \left\{(i, V_1^{z_1}[i], \ldots, V_{k-1}^{z_{k-1}}[i]) :\ i = 1,2,\ldots,t;\ z \in \{0,1\}^{k-1}\setminus\{0^{k-1}\}\right\}.$$

Clearly, A and B are the sets of variables featured in the expressions (†) and (‡) above, respectively. To analyze $\Gamma(\mathbf{V})$, we prove two key claims analogous to those in the two-party Theorem 4.23.

CLAIM 9.5. *Assume that* $|A \cap B| \leqslant d-1$. *Then* $\Gamma(\mathbf{V}) = 0$.

PROOF. Immediate from the fact that the Fourier transform of ψ is supported on characters of order d and higher. □

CLAIM 9.6. *Assume that* $|A \cap B| = i$. *Then* $|\Gamma(\mathbf{V})| \leqslant 2^{i2^{k-1} - t2^{k-1}}$.

PROOF. Proposition 4.22 shows that $|\Gamma(\mathbf{V})| \leqslant 2^{-t2^{k-1}} 2^{t2^{k-1} - |A \cup B|}$. Furthermore, it is straightforward to verify that $|A \cup B| \geqslant t2^{k-1} - |A \cap B|\, 2^{k-1}$. □

In view of Claims 9.5 and 9.6, inequality (9.3) simplifies to

$$\operatorname{disc}_P(H)^{2^{k-1}} \leqslant \sum_{i=d}^{t} 2^{i2^{k-1}}\, \mathbf{P}[|A \cap B| = i].$$

It remains to bound the probability $\mathbf{P}[|A \cap B| = i]$. For a fixed element a, we have $\mathbf{P}[a \in B \mid a \in A] \leqslant (k-1)t/n$ by the union bound. Moreover, given two distinct elements $a, a' \in A$, the corresponding events $a \in B$ and $a' \in B$ are independent. Therefore,

$$\mathbf{P}[|A \cap B| = i] \leqslant \binom{t}{i} \left(\frac{(k-1)t}{n}\right)^{i},$$

which yields the desired bound on $\operatorname{disc}_P(H)$. □

We now present an adaptation of the two-party Theorem 4.26 to the multiparty model, obtained by Lee and Shraibman [140] and independently by Chattopadhyay and Ada [59].

THEOREM 9.7 (Communication complexity of pattern tensors [140, 59]). *Let* $f:\{0,1\}^t \to \{-1,+1\}$ *be given with* $\deg_{1/3}(f) = d \geqslant 1$. *Let* n *be a given integer,* $t \mid n$. *Let* F *be the* (k,n,t,f)*-pattern tensor. If* $n \geqslant 4et^2(k-1)2^{2^{k-1}}/d$, *then*

$$R_{1/3}(F) \geqslant \Omega\left(\frac{d}{2^k}\right).$$

PROOF (adapted from [140, 59]). The proof is identical to that of the two-party Theorem 4.26, with pattern matrices replaced by pattern tensors. By Theorem 4.5, there is a function $h:\{0,1\}^t \to \{-1,+1\}$ and a probability distribution μ on $\{0,1\}^t$ such that

$$\widehat{\mu h}(S) = 0, \qquad |S| < d, \tag{9.4}$$

$$\sum_{z\in\{0,1\}^t} f(z)\mu(z)h(z) > \frac{1}{3}. \tag{9.5}$$

Letting P be the $(k,n,t,2^{-t(n/t)^{k-1}+t}(n/t)^{-t(k-1)}\mu)$-pattern tensor and H the (k,n,t,h)-pattern tensor, we obtain from (9.4) and Theorem 9.4 that

$$\operatorname{disc}_P(H) \leqslant 2^{-d/2^{k-1}}. \tag{9.6}$$

At the same time, one sees from (9.5) that

$$\langle F, H \circ P\rangle > \frac{1}{3}. \tag{9.7}$$

The theorem now follows from (9.6), (9.7), and the generalized discrepancy method (Theorem 9.3). □

These generalizations give an improved lower bound on the k-party communication complexity of the disjointness function $f(x) = \bigvee_{j=1}^{n} \bigwedge_{i=1}^{k} x_{i,j}$, where player i sees all the inputs except for $x_{i,1}, \ldots, x_{i,n}$.

COROLLARY 9.8 (Multiparty complexity of disjointness [140, 59]). *Define* $f(x) = \bigvee_{j=1}^{n} \bigwedge_{i=1}^{k} x_{i,j}$. *Then*

$$R_{1/3}(f) \geqslant n^{\Theta(1/k)} 2^{-\Theta(2^k)}.$$

PROOF (adapted from [140, 59]). Let t be an integer to be fixed later. Recall from Theorem 2.5 that $\deg_{1/3}(\mathrm{OR}_t) \geqslant \alpha\sqrt{t}$ for some constant $\alpha > 0$. Letting m be the smallest multiple of t such that $m \geqslant 4et^2(k-1)2^{2^{k-1}}/(\alpha\sqrt{t})$, it follows by Theorem 9.7 that the (k, m, t, OR_t)-pattern tensor has bounded-error randomized communication complexity $\Omega(\sqrt{t}/2^k)$. It remains to set t to be the largest integer such that $t(m/t)^{k-1} \leqslant n$, thereby assuring that the (k, m, t, OR_t)-pattern tensor is a subtensor of f. □

In a recent result, Beame and Huynh-Ngoc [29] build on this line of research to give stronger lower bounds for the disjointness function for certain ranges of k. Also obtained in [29] is the first polynomial lower bound on the multiparty communication complexity of an AC^0 function, for up to $k = \varepsilon \log n$ players.

9.5 Separation of $\mathsf{NP}^{\mathsf{cc}}$ from $\mathsf{BPP}^{\mathsf{cc}}$

Since the disjointness function trivially belongs to $\mathsf{NP}^{\mathsf{cc}}_k$, Corollary 9.8 separates the communication classes $\mathsf{NP}^{\mathsf{cc}}_k$ and $\mathsf{BPP}^{\mathsf{cc}}_k$ for up to $k = \Theta(\log\log n)$ parties. In this section, we present a separation [64, 65] of these classes for exponentially more players, up to $k = (1-\varepsilon)\log n$. The crucial insight in this new work is to redefine the projection operator $x|_{V_1,\ldots,V_{k-1}}$ from Section 9.4 using the probabilistic method. This removes the key bottleneck in the previous analyses [140, 59]. We will present the result in two stages, starting with an existential argument due David and Pitassi [64] followed by its derandomization due to David, Pitassi, and Viola [65].

We start with some notation. Fix natural numbers n, m with $n > m$. We let the symbol $\binom{[n]}{m}$ stand for the family of all m-element subsets of $[n]$. Let $\psi\colon \{0,1\}^m \to \mathbb{R}$ be a given function with $\sum_{z\in\{0,1\}^m} |\psi(z)| = 1$. Let d denote the least order of a nonzero Fourier coefficient of ψ. Fix a Boolean function $h\colon \{0,1\}^m \to \{-1,+1\}$ and a distribution μ on $\{0,1\}^m$ such that $\psi(z) \equiv h(z)\mu(z)$. For a mapping $\alpha\colon (\{0,1\}^n)^k \to \binom{[n]}{m}$, define a $(k+1)$-party communication problem $H_\alpha\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ by $H(x, y_1, \ldots, y_k) = h(x|_{\alpha(y_1,\ldots,y_k)})$. Analogously, define a distribution P_α on $(\{0,1\}^n)^{k+1}$ by $P_\alpha(x, y_1, \ldots, y_k) = 2^{-(k+1)n+m}\mu(x|_{\alpha(y_1,\ldots,y_k)})$.

THEOREM 9.9 (David and Pitassi [64]). *Assume that* $n \geqslant 16em^2 2^k$. *Then for a uniformly random choice of* $\alpha\colon (\{0,1\}^n)^k \to \binom{[n]}{m}$,

$$\mathop{\mathbf{E}}_{\alpha}\left[\mathrm{disc}_{P_\alpha}(H_\alpha)^{2^k}\right] \leqslant 2^{-n/2} + 2^{-d2^k+1}.$$

PROOF (adapted from [64]). By Theorem 9.1,

$$\mathrm{disc}_{P_\alpha}(H_\alpha)^{2^k} \leqslant 2^{m2^k} \mathop{\mathbf{E}}_{Y} |\Gamma(Y)|, \tag{9.8}$$

where we put $Y = (y_1^0, y_1^1, \ldots, y_k^0, y_k^1)$ and

$$\Gamma(Y) = \mathop{\mathbf{E}}_{x}\left[\prod_{z\in\{0,1\}^k} \psi\left(x|_{\alpha(y_1^{z_1}, y_2^{z_2}, \ldots, y_k^{z_k})}\right)\right].$$

For a fixed choice of Y, we will use the shorthand $S_z = \alpha(y_1^{z_1}, \ldots, y_k^{z_k})$. To analyze $\Gamma(Y)$, we prove two key claims analogous to those in Theorem 4.23.

CLAIM 9.10. *Assume that* $|\bigcup S_z| > m2^k - d2^{k-1}$. *Then* $\Gamma(Y) = 0$.

PROOF. If $|\bigcup S_z| > m2^k - d2^{k-1}$, then some S_z must feature more than $m - d$ elements that do not occur in $\bigcup_{u\neq z} S_u$. But this forces $\Gamma(Y) = 0$ since the Fourier transform of ψ is supported on characters of order d and higher. □

CLAIM 9.11. *For every* Y, $\quad |\Gamma(Y)| \leqslant 2^{-|\bigcup S_z|}$.

PROOF. Immediate from Proposition 4.22. □

In view of (9.8) and Claims 9.10 and 9.11, we have

$$\mathop{\mathbf{E}}_{\alpha}\left[\operatorname{disc}_{P_\alpha}(H_\alpha)^{2^k}\right] \leqslant \sum_{i=d2^{k-1}}^{m2^k-m} 2^i \mathop{\mathbf{P}}_{Y,\alpha}\left[\left|\bigcup S_z\right| = m2^k - i\right].$$

It remains to bound the probabilities in the last expression. With probability at least $1 - k2^{-n}$ over the choice of Y, we have $y_i^0 \neq y_i^1$ for each $i = 1, 2, \ldots, k$. Conditioning on this event, the fact that α is chosen uniformly at random means that the 2^k sets S_z are distributed independently and uniformly over $\binom{[n]}{m}$. A calculation now reveals that

$$\mathop{\mathbf{P}}_{Y,\alpha}\left[\left|\bigcup S_z\right| = m2^k - i\right] \leqslant k2^{-n} + \binom{m2^k}{i}\left(\frac{m2^k}{n}\right)^i \leqslant k2^{-n} + 8^{-i},$$

where the last step uses the fact that $i \geqslant d2^{k-1}$. □

We are ready to present a nonexplicit separation of nondeterministic and randomized multiparty communication complexity.

THEOREM 9.12 (David and Pitassi [64]). *Let* $k \leqslant (1-\varepsilon)\log_2 n$, *where* $\varepsilon > 0$ *is a given constant. Then there exists a function* $F_\alpha\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ *with* $N(F_\alpha) = O(\log n)$ *but* $R_{1/3}(F_\alpha) = n^{\Omega(1)}$. *In particular,*

$$\mathsf{NP}^{\mathsf{cc}}_k \not\subseteq \mathsf{BPP}^{\mathsf{cc}}_k.$$

PROOF (adapted from [64]). Let $m = \lfloor n^\zeta \rfloor$ for a sufficiently small constant $\zeta = \zeta(\varepsilon) > 0$. Recall from Theorem 2.5 that $\deg_{1/3}(\mathrm{OR}_m) = \Theta(\sqrt{m})$. As a result,

Theorem 4.5 guarantees the existence of a function $\psi\colon \{0,1\}^m \to \mathbb{R}$ such that:

$$\begin{aligned} &\hat{\psi}(S) = 0 && \text{for } |S| < \Theta(\sqrt{m}), \\ &\sum_{z\in\{0,1\}^m} |\psi(z)| = 1, \\ &\sum_{z\in\{0,1\}^m} \psi(z)\mathrm{OR}_m(z) > \frac{1}{3}. \end{aligned}$$

Fix a function $h\colon \{0,1\}^m \to \{-1,+1\}$ and a distribution μ on $\{0,1\}^m$ such that $\psi(z) \equiv h(z)\mu(z)$. For a mapping $\alpha\colon (\{0,1\}^n)^k \to \binom{[n]}{m}$, let H_α and P_α be as defined at the beginning of this section. Then Theorem 9.9 shows the existence of α such that

$$\mathrm{disc}_{P_\alpha}(H_\alpha) \leqslant 2^{-\Omega(\sqrt{m})}.$$

Using the properties of ψ, one readily verifies that $\langle H \circ P_\alpha, F_\alpha\rangle \geqslant 1/3$, where $F_\alpha\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ is given by $F_\alpha(x, y_1, \ldots, y_k) = \mathrm{OR}_m(x|_{\alpha(y_1,\ldots,y_k)})$. By the generalized discrepancy method (Theorem 9.3),

$$R_{1/3}(F_\alpha) \geqslant \Omega(\sqrt{m}) = n^{\Omega(1)}.$$

On the other hand, F_α has nondeterministic complexity $O(\log n)$. Namely, Player 1 (who knows $y_1, \ldots, y_k$) nondeterministically selects an element $i \in \alpha(y_1, \ldots, y_k)$ and announces i. Player 2 (who knows x) then announces x_i as the output of the protocol. □

We will now sketch an explicit construction of the function whose existence is given by Theorem 9.12.

THEOREM 9.13 (David, Pitassi, and Viola [65]). *Let $k \leqslant (1-\varepsilon)\log_2 n$, where $\varepsilon > 0$ is a given constant. Then there exists an (explicitly given) function $F\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ with $N(F) = O(\log n)$ but $R_{1/3}(F) = n^{\Omega(1)}$. In*

particular,

$$\mathsf{NP}_k^{\mathsf{cc}} \not\subseteq \mathsf{BPP}_k^{\mathsf{cc}}.$$

PROOF SKETCH (adapted from [65]). The proof proceeds by a derandomization of the choice of α in Theorem 9.12. Instead of working with a family $\{H_\alpha\}$ of functions, each given by $H_\alpha(x, y_1, \ldots, y_k) = h(x|_{\alpha(y_1,\ldots,y_k)})$, one posits a single function $H(\alpha, x, y_1, \ldots, y_k) = h(x|_{\alpha(y_1,\ldots,y_k)})$, where the new argument α is known to all players and ranges over a small, explicitly given subset A of all mappings $(\{0,1\}^n)^k \to \binom{[n]}{m}$. By choosing A to be pseudorandom, one arrives at the same qualitative conclusion as in Theorem 9.9. □

9.6 Analyzing nondeterministic and Merlin-Arthur complexity

In this section and the next, we present a separation of nondeterministic communication complexity from co-nondeterministic and further from Merlin-Arthur complexity. A starting point in our discussion is a well-known combinatorial fact about multiparty protocols. Consider a function $f: X_1 \times \cdots \times X_k \to \{-1, +1\}$. A *cylinder intersection* is any function $\chi: X_1 \times \cdots \times X_k \to \{0, 1\}$ of the form

$$\chi(x_1, \ldots, x_k) = \prod_{i=1}^{k} \phi_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k),$$

for some functions $\phi_i: X_1 \times \cdots \times X_{i-1} \times X_{i+1} \times \cdots \times X_k \to \{0, 1\}$, $i = 1, 2, \ldots, k$. Cylinder intersections play a basic role in the study of communication complexity. First of all, a low-cost deterministic protocol partitions the sets $f^{-1}(-1)$ and $f^{-1}(+1)$ into a small number of cylinder intersections:

FACT 9.14 (see Kushilevitz and Nisan [137]). *Let $f: X_1 \times \cdots \times X_k \to \{-1, +1\}$ be a given function, $c = D(f)$. Then there exist cylinder intersections $\chi_1, \ldots, \chi_{2^c}$*

such that

$$\sum \chi_i(x) \equiv \frac{1 - f(x)}{2}.$$

An analogous statement holds for nondeterministic complexity:

FACT 9.15 (see Kushilevitz and Nisan [137]). *Let $f: X_1 \times \cdots \times X_k \to \{-1, +1\}$ be a given function, $c = N(f)$. Then there exist cylinder intersections $\chi_1, \ldots, \chi_{2^c}$ such that*

$$f(x) = -1 \qquad \Leftrightarrow \qquad \sum \chi_i(x) > 1.$$

We are now ready for the first step of our proof, a technique for lower bounds on nondeterministic communication complexity inspired by the generalized discrepancy method.

THEOREM 9.16 (Gavinsky and Sherstov [81]). *Let $F: X \to \{-1, +1\}$ be given, where $X = X_1 \times \cdots \times X_k$. Fix a function $H: X \to \{-1, +1\}$ and a probability distribution P on X. Put*

$$\begin{aligned} \alpha &= P(F^{-1}(-1) \cap H^{-1}(-1)), \\ \beta &= P(F^{-1}(-1) \cap H^{-1}(+1)), \\ Q &= \log \frac{\alpha}{\beta + \operatorname{disc}_P(H)}. \end{aligned}$$

Then

$$N(F) \geqslant Q \tag{9.9}$$

and

$$MA_{1/3}(F) \geqslant \min\left\{\Omega(\sqrt{Q}),\ \Omega\left(\frac{Q}{\log\{2/\alpha\}}\right)\right\}. \tag{9.10}$$

PROOF. Put $c = N(F)$. Fix cylinder intersections $\chi_1, \chi_2, \ldots, \chi_{2^c}: X \to \{0,1\}$ guaranteed by Fact 9.15. By the definition of discrepancy,

$$\left\langle \sum \chi_i, -H \circ P \right\rangle \leqslant \sum |\langle \chi_i, -H \circ P \rangle| \leqslant 2^c \operatorname{disc}_P(H).$$

On the other hand, $\sum \chi_i$ ranges between 1 and 2^c on $F^{-1}(-1)$ and vanishes on $F^{-1}(+1)$. Therefore,

$$\left\langle \sum \chi_i, -H \circ P \right\rangle \geqslant \alpha - 2^c \beta.$$

These two inequalities force (9.9).

We now turn to the Merlin-Arthur model. Let $c = MA_{1/3}(F)$ and $\delta = \alpha 2^{-c-1}$. The first step is to improve the error probability of the Merlin-Arthur protocol by repetition from $1/3$ to δ. Specifically, following Klauck [115] we observe that there exist randomized protocols $F_1, \ldots, F_{2^c}: X \to \{0,1\}$, each a random variable of the coin tosses and each having communication cost $c' = O(c \log\{1/\delta\})$, such that the sum

$$\sum \mathbf{E}[F_i]$$

ranges in $[1-\delta, 2^c]$ on $F^{-1}(-1)$ and in $[0, \delta 2^c]$ on $F^{-1}(+1)$. As a result,

$$\left\langle \sum \mathbf{E}[F_i], -H \circ P \right\rangle \geqslant \alpha(1-\delta) - \beta 2^c - (1-\alpha-\beta)\delta 2^c. \tag{9.11}$$

Since a randomized protocol $X \to \{0,1\}$ of cost c' is a probability distribution over deterministic protocols $X \to \{0,1\}$ of cost c', and since by Fact 9.14 every such deterministic protocol is the sum of at most 2^c cylinder intersections, it follows from the definition of discrepancy that

$$\left\langle \sum \mathbf{E}[F_i], -H \circ P \right\rangle \leqslant \sum_{i=1}^{2^c} 2^{c'} \operatorname{disc}_P(H) = 2^{c+c'} \operatorname{disc}_P(H). \tag{9.12}$$

The bounds in (9.11) and (9.12) force (9.10). □

Since sign tensors H and $-H$ have the same discrepancy under any given distribution, we have the following alternate form of Theorem 9.16.

COROLLARY 9.17 (Gavinsky and Sherstov [81]). *Let $F: X \to \{-1,+1\}$ be given, where $X = X_1 \times \cdots \times X_k$. Fix a function $H: X \to \{-1,+1\}$ and a probability distribution P on X. Put*

$$\begin{aligned} \alpha &= P(F^{-1}(+1) \cap H^{-1}(+1)), \\ \beta &= P(F^{-1}(+1) \cap H^{-1}(-1)), \\ Q &= \log \frac{\alpha}{\beta + \operatorname{disc}_P(H)}. \end{aligned}$$

Then

$$N(-F) \geqslant Q$$

and

$$MA_{1/3}(-F) \geqslant \min\left\{ \Omega(\sqrt{Q}),\ \Omega\left(\frac{Q}{\log\{2/\alpha\}} \right) \right\}.$$

9.7 Separation of $\mathsf{NP}^{\mathsf{cc}}$ from $\mathsf{coNP}^{\mathsf{cc}}$ and $\mathsf{coMA}^{\mathsf{cc}}$

Using the results of the previous section, we will now obtain the desired separations in communication complexity. Following the organization of Section 9.5, we will first give an existential argument and only then sketch an explicit separation. We start by deriving a suitable analytic property of the OR function.

THEOREM 9.18 (Gavinsky and Sherstov [81]). *There is a function* $\psi\colon \{0,1\}^m \to \mathbb{R}$ *such that:*

$$\sum_{z\in\{0,1\}^m} |\psi(z)| = 1, \tag{9.13}$$

$$\hat{\psi}(S) = 0, \qquad |S| < \Theta(\sqrt{m}), \tag{9.14}$$

$$\psi(0) > \frac{1}{6}. \tag{9.15}$$

PROOF. Recall from Theorem 2.5 that $\deg_{1/3}(\mathrm{OR}_m) = \Omega(\sqrt{m})$. By Theorem 4.5, there is a function $\psi\colon \{0,1\}^m \to \mathbb{R}$ that obeys (9.13), (9.14), and additionally satisfies

$$\sum_{z\in\{0,1\}^m} \psi(z)\mathrm{OR}_m(z) > \frac{1}{3}.$$

As a result,

$$2\psi(0) = \sum_{z\in\{0,1\}^m} \psi(z)\{\mathrm{OR}_m(z)+1\} = \sum_{z\in\{0,1\}^m} \psi(z)\mathrm{OR}_m(z) > \frac{1}{3},$$

where the second equality follows from (9.14). □

In what follows, it will be convenient to reinstate the notation of Section 9.5, due to David and Pitassi [64]. Fix integers n, m with $n > m$. Let

$\psi\colon \{0,1\}^m \to \mathbb{R}$ be a given function with $\sum_{z\in\{0,1\}^m} |\psi(z)| = 1$. Let d denote the least order of a nonzero Fourier coefficient of ψ. Fix a Boolean function $h\colon \{0,1\}^m \to \{-1,+1\}$ and a distribution μ on $\{0,1\}^m$ such that $\psi(z) \equiv h(z)\mu(z)$. For a mapping $\alpha\colon (\{0,1\}^n)^k \to \binom{[n]}{m}$, define a $(k+1)$-party communication problem $H_\alpha\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ by $H_\alpha(x, y_1, \ldots, y_k) = h(x|_{\alpha(y_1,\ldots,y_k)})$. Analogously, define a distribution P_α on $(\{0,1\}^n)^{k+1}$ by $P_\alpha(x, y_1, \ldots, y_k) = 2^{-(k+1)n+m}\mu(x|_{\alpha(y_1,\ldots,y_k)})$.

We are now in position to give the promised existential separation. It may be helpful to compare the proof to follow with David and Pitassi's proof of Theorem 9.12.

THEOREM 9.19 (Gavinsky and Sherstov [81]). *Let $k \leqslant (1-\varepsilon)\log n$, where $\varepsilon > 0$ is any given constant. Then there exists a function $F_\alpha\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ such that $N(F_\alpha) = O(\log n)$ and $MA_{1/3}(-F_\alpha) = n^{\Omega(1)}$. In particular,*

$$\mathsf{NP}^{\mathsf{cc}}_k \not\subseteq \mathsf{coMA}^{\mathsf{cc}}_k, \qquad \mathsf{NP}^{\mathsf{cc}}_k \neq \mathsf{coNP}^{\mathsf{cc}}_k.$$

PROOF. Let $m = \lfloor n^\delta \rfloor$ for a sufficiently small constant $\delta = \delta(\varepsilon) > 0$. Let $\psi\colon \{0,1\}^m \to \mathbb{R}$ be as guaranteed by Theorem 9.18. For a mapping $\alpha\colon (\{0,1\}^n)^k \to \binom{[n]}{m}$, let H_α and P_α be defined in terms of ψ as described earlier in this section. Then Theorem 9.9 shows the existence of α such that

$$\operatorname{disc}_{P_\alpha}(H_\alpha) \leqslant 2^{-\Omega(\sqrt{m})}. \tag{9.16}$$

Define $F_\alpha\colon (\{0,1\}^n)^{k+1} \to \{-1,+1\}$ by $F_\alpha(x, y_1, \ldots, y_k) = \mathrm{OR}_m(x|_{\alpha(y_1,\ldots,y_k)})$. It is clear from the properties of ψ that

$$P_\alpha(F_\alpha^{-1}(+1) \cap H_\alpha^{-1}(+1)) > \frac{1}{6}, \tag{9.17}$$

$$P_\alpha(F_\alpha^{-1}(+1) \cap H_\alpha^{-1}(-1)) = 0. \tag{9.18}$$

The sought lower bound on the Merlin-Arthur complexity of $-F_\alpha$ now follows from (9.16)–(9.18) and Corollary 9.17.

On the other hand, as in the proof of David and Pitassi's Theorem 9.12, the function F_α has an efficient nondeterministic protocol. Namely, player 1 (who knows $y_1, \ldots, y_k$) nondeterministically selects an element $i \in \alpha(y_1, \ldots, y_k)$ and writes i on the shared blackboard. Player 2 (who knows x) then announces x_i as the output of the protocol. This yields the desired upper bound on the nondeterministic complexity of F_α. □

The probabilistic choice of α in Theorem 9.19 admits the same derandomization as Theorem 9.12, yielding our main result.

THEOREM 9.20 (Gavinsky and Sherstov [81]). *Let $k \leqslant (1-\varepsilon)\log n$, where $\varepsilon > 0$ is any given constant. Then there is an (explicitly given) function $F\colon (\{0,1\}^n)^k \to \{-1,+1\}$ with $N(F) = O(\log n)$ and $MA_{1/3}(-F) = n^{\Omega(1)}$. In particular,*

$$\mathsf{NP}^{\mathsf{cc}}_k \not\subseteq \mathsf{coMA}^{\mathsf{cc}}_k, \qquad \mathsf{NP}^{\mathsf{cc}}_k \neq \mathsf{coNP}^{\mathsf{cc}}_k.$$

PROOF. Identical to the derandomization given by Theorem 9.13. □

Chapter 10

Relations and Separations

In this chapter, we fully characterize the relations and gaps among three complexity measures of a communication problem: product discrepancy, nonproduct discrepancy, and sign-rank. As a corollary, we prove that the containment $\mathsf{PP}^{\mathsf{cc}} \subseteq \mathsf{UPP}^{\mathsf{cc}}$ is proper. We further prove that product discrepancy is equivalent to the statistical query dimension. Finally, we solve an open problem due to Kushilevitz and Nisan [137], exhibiting a gap of $\Theta(1)$ versus $\Theta(n)$ between the product and nonproduct distributional complexities of a function on n-bit strings.

10.1 Introduction

In previous chapters, we demonstrated the power of matrix analysis, duality, approximation theory, the Fourier transform, and other analytic methods in the study of communication complexity. In this concluding chapter on communication, we apply the analytic approach in yet another way. The work in this chapter revolves around five key complexity measures of a sign matrix A, all reviewed in the introductory Chapters 2 and 3. The first is the discrepancy $\operatorname{disc}(A)$, which is among the most important quantities in communication. The second quantity that we consider is the product discrepancy $\operatorname{disc}^{\times}(A)$, defined as the minimum discrepancy of A under product distributions. The third complexity measure is the sign-rank $\operatorname{rk}_{\pm}(A)$, a key notion in unbounded-error communication complexity. The final two complexity measures of a sign matrix discussed here are the statistical query dimension $\operatorname{sq}(A)$ and the margin complexity $\operatorname{mc}(A)$, both originating in learning theory.

Prior to our work, Ben-David et al. [35] proved that the sign-rank of a sign matrix A is essentially bounded by $O(1/\operatorname{mc}(A)^2)$. Furthermore, Linial and Shraibman [145] showed that the margin complexity and discrepancy are essentially equivalent notions, in that $\operatorname{mc}(A)$ and $1/\operatorname{disc}(A)$ are always within a factor of 8 of each other. Here, we complete these two results to the picture in Figure 10.1.

We will now traverse this schematic left to right, giving more precise statements. Our first result, proved in Section 10.3, states that the statistical query dimension and product discrepancy are essentially equivalent notions in that $\operatorname{sq}(A)$ and $1/\operatorname{disc}^{\times}(A)$ are polynomially related for all A. Next, we prove in Section 10.4 that the statistical query dimension $\operatorname{sq}(A)$ of a sign matrix is bounded by $2(\operatorname{rk}_{\pm} A)^2$. We further show that the gap between the two quantities can be essentially arbitrary. In particular, we prove that sign matrices exist of order N with statistical query

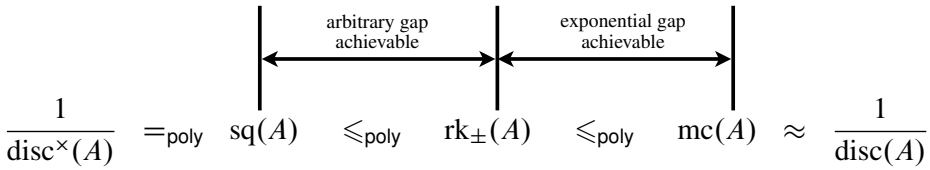

Figure 10.1: Relationships among key complexity measures.

dimension $O(1)$ and sign-rank $\Omega(N^{1-\varepsilon})$. As an application of our findings, we strengthen the current upper bound on the statistical query dimension of halfspaces in n dimensions, due to Blum et al. [36], from $n^{O(1)}$ to a near-tight $2(n+1)^2$.

Continuing in Section 10.5, we exhibit a sign matrix A with an exponential gap between $\mathrm{rk}_{\pm}(A)$ and $1/\mathrm{disc}(A)$. Independently of the author, Buhrman et al. [53] exhibited another sign matrix with such a gap, which we also present in this chapter. This exponential gap between the sign-rank and the reciprocal of the discrepancy establishes the strict containment $\mathsf{PP}^{\mathsf{cc}} \subsetneq \mathsf{UPP}^{\mathsf{cc}}$, which was open to prove since the introduction of these classes by Babai et al. [23].

We conclude this chapter in Section 10.6 by resolving a related open problem due to Kushilevitz and Nisan [137]. These authors asked whether the equality

$$R_{1/3}(f) = \max_{\mu}\{D^{\mu}_{1/3}(f)\},$$

given by Yao's minimax principle (Theorem 3.1), is approximately preserved if the maximum is taken over product distributions only, rather than all distributions μ. We give a strong negative answer to this question, proving the existence of a function $f\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ whose maximum product and nonproduct distributional complexities are $O(1)$ and $\Theta(n)$, respectively. The function f in question is particularly hard in that it has essentially the smallest possible discrepancy and the maximum unbounded-error complexity, despite the fact that its distributional complexity is constant under all product distributions.

The placement of this chapter at the end of Part I on communication complexity and just before Part II on learning theory is no accident. Indeed, an important theme in our results and methods here is the close interplay between commu-

nication complexity and learning theory. This relationship between communication and learning is natural since a sign matrix can be viewed both as a communication problem and a learning problem. We will see more such instances in Part II of this thesis.

10.2 Technical preliminaries

We start with some combinatorial and learning-theoretic background. The reader may find it useful to review the definitions and results of Section 2.5 on fundamentals of learning theory before reading on.

The matrix family $\mathscr{Z}(N,c) \subset \{-1,+1\}^{N\times N}$, due to Zarankiewicz [227], is the family of sign matrices that contain no submatrix of size $c \times c$ with all entries equal to $+1$. A classical result due to Bollobás [43] states that $\mathscr{Z}(N,c)$ contains a considerable fraction of the matrices in $\{-1,+1\}^{N\times N}$. On the other hand, Alon et al. [13] proved that all but a tiny fraction of the matrices in $\{-1,+1\}^{N\times N}$ have high sign-rank. These two results were combined in the work of Ben-David et al. [35] to the following effect.

THEOREM 10.1 (Ben-David et al. [35, Thm. 12]). *Let $c \geqslant 2$ be a fixed integer. Then all but a vanishing fraction of the matrices in $\mathscr{Z}(N,c)$ have sign-rank $\Omega(N^{1-\frac{2}{c}})$.*

A glance at the proof of Ben-David et al. [35] reveals the following somewhat more delicate result, which is what we will need in this chapter.

THEOREM 10.2 (Ben-David et al. [35], implicit). *Let $\alpha > 0$ be a suitably small absolute constant. Let c be a given integer with*

$$2 \leqslant c \leqslant \alpha\left(\frac{\log N}{\log\log N}\right)^{1/2}.$$

Then all but a vanishing fraction of the matrices in $\mathscr{Z}(N,c)$ have sign-rank at least $\alpha N^{1-\frac{2}{c}}$.

Apart from Zarankiewicz matrices, we will need a bound on the statistical query dimension of a concept class in terms of its Vapnik-Chervonenkis dimension.

LEMMA 10.3 (Sherstov [197]). *Let $\mathscr{C}$ be a concept class. Then*

$$\mathrm{sq}(\mathscr{C}) \leqslant 2^{O(\mathrm{vc}(\mathscr{C}))}.$$

PROOF. Let $\mathrm{sq}(\mathscr{C}) = d \geqslant 2$. Our goal is to show that $\mathrm{vc}(\mathscr{C}) = \Omega(\log d)$. By definition of the statistical query dimension, there is a distribution μ and functions $f_1, \ldots, f_d \in \mathscr{C}$ such that

$$\Delta_\mu(f_i, f_j) \geqslant \frac{1}{2} - \frac{1}{2d}$$

for all $i \neq j$. In particular, $\Delta_\mu(f_i, f_j) > 1/5$. Thus, Proposition 2.15 shows that learning $\mathscr{C}$ to accuracy $1/10$ and confidence $1/2$ requires $m \geqslant \Omega(\log d)$ examples. Yet by Theorem 2.14, the number of examples needed is at most $m \leqslant O(\mathrm{vc}(\mathscr{C}))$. Comparing these two estimates completes the proof. □

10.3 SQ dimension vs. product discrepancy

The purpose of this section is to prove the equivalence of two well-studied notions, which were independently defined one in learning theory and the other in communication complexity. The notions in question are the statistical query dimension and the minimum discrepancy under product distributions. We start with an observation that a concept class $\mathscr{C}$ with low statistical query dimension contains a small set of functions which, collectively, approximate all of $\mathscr{C}$ in a nontrivial way.

PROPOSITION 10.4 (Sherstov [198]). *Let $\mathscr{C}$ be a given concept class of functions $X \to \{-1, +1\}$ with $\mathrm{sq}_\mu(\mathscr{C}) = N$. Then there is a set $\mathscr{H} \subseteq \mathscr{C}$ with $|\mathscr{H}| = N$*

such that each $f \in \mathscr{C}$ *has*

$$\left|\mathop{\mathbf{E}}_{\mu}[f(x)h(x)]\right| > \frac{1}{N+1}$$

for some $h \in \mathscr{H}$.

PROOF. For a set $\mathscr{F} \subseteq \mathscr{C}$, define

$$\gamma(\mathscr{F}) = \max_{f_1, f_2} \left|\mathop{\mathbf{E}}_{\mu}[f_1(x)\, f_2(x)]\right|.$$

where the maximum ranges over distinct functions $f_1, f_2 \in \mathscr{C}$. In words, $\gamma(\mathscr{F})$ is the largest correlation between any two distinct functions in $\mathscr{F}$. Let γ^* be the minimum $\gamma(\mathscr{F})$ over all N-element subsets $\mathscr{F} \subseteq \mathscr{C}$. Let $\mathscr{H}$ be a set of N functions in $\mathscr{C}$ such that $\gamma(\mathscr{H}) = \gamma^*$ and the number of function pairs in $\mathscr{H}$ with correlation γ^* is the smallest possible (over all N-element subsets $\mathscr{F}$ with $\gamma(\mathscr{F}) = \gamma^*$).

We claim that each $f \in \mathscr{C}$ has $|\mathbf{E}_\mu[f(x)h(x)]| > 1/(N+1)$ for some $h \in \mathscr{H}$. If $f \in \mathscr{H}$, the claim is trivially true. Thus, assume that $f \notin \mathscr{H}$. There are two cases to consider. If $\gamma(\mathscr{H}) \leqslant 1/(N+1)$, then f must have correlation more than $1/(N+1)$ with some member of $\mathscr{H}$ because otherwise we would have $\gamma(\mathscr{H} \cup \{f\}) \leqslant 1/(N+1)$ and $\mathrm{sq}_\mu(\mathscr{C}) \geqslant N+1$. If $\gamma(\mathscr{H}) > 1/(N+1)$, then again f must have correlation more than $1/(N+1)$ with some member of $\mathscr{H}$ because otherwise we could improve on the number of function pairs in $\mathscr{H}$ with correlation γ^* by replacing a suitably chosen element of $\mathscr{H}$ with f. □

Our next ingredient is an expression for the discrepancy of a sign matrix under a product distribution. Recall from Section 3.2 that a *product distribution* μ on the finite Cartesian product $X \times Y$ is a distribution that can be expressed in the form $\mu(x, y) = \mu_X(x)\mu_Y(y)$ for some distributions μ_X and μ_Y over X and Y, respectively.

LEMMA 10.5 (Ford and Gál [69]). *Let* $f : X \times Y \to \{-1, +1\}$ *be a given function, where* X *and* Y *are finite sets. Let* $\mu(x, y) = \mu_X(x)\mu_Y(y)$ *be a product*

distribution over $X \times Y$. Then

$$\mathrm{disc}_\mu(f) \leqslant \sqrt{\mathop{\mathbf{E}}_{y,y' \sim \mu_Y} \left| \mathop{\mathbf{E}}_{x \sim \mu_X} [f(x,y) f(x,y')] \right|}.$$

PROOF (adapted from [69]). As shown in the proof of Lemma 3.3, there exist values $\alpha_x, \beta_y \in \{-1, +1\}$ for all x and y such that

$$\mathrm{disc}_\mu(f) \leqslant \left| \sum_{x,y} \alpha_x \beta_y \mu(x,y) f(x,y) \right|.$$

Recalling that $\mu(x,y) = \mu_X(x)\mu_Y(y)$, we obtain

$$\mathrm{disc}_\mu(f) \leqslant \left| \mathop{\mathbf{E}}_{x \sim \mu_X} \mathop{\mathbf{E}}_{y \sim \mu_Y} [\alpha_x \beta_y f(x,y)] \right|.$$

The remainder of the proof is closely analogous to Lemma 3.3. Squaring and applying the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathrm{disc}_\mu(f)^2 &\leqslant \mathop{\mathbf{E}}_{x \sim \mu_X} \left[\left(\mathop{\mathbf{E}}_{y \sim \mu_Y} [\alpha_x \beta_y f(x,y)] \right)^2 \right] \\
&= \mathop{\mathbf{E}}_{x \sim \mu_X} \mathop{\mathbf{E}}_{y,y' \sim \mu_Y} [\alpha_x^2 \beta_y \beta_{y'} f(x,y) f(x,y')] \\
&\leqslant \mathop{\mathbf{E}}_{y,y' \sim \mu_Y} \left| \mathop{\mathbf{E}}_{x \sim \mu_X} [f(x,y) f(x,y')] \right|,
\end{aligned}$$

where the last step uses the fact that $\alpha_x^2 = |\beta_y \beta_{y'}| = 1$. □

We are now in a position to prove the claimed equivalence of the statistical query dimension of a sign matrix and its minimum discrepancy under product distributions. We split the proof in two parts, corresponding to the upper and lower bound on the product discrepancy in terms of the statistical query dimension.

LEMMA 10.6 (Sherstov [198]). *Let* $A \in \{-1,+1\}^{M\times N}$. *Then*

$$\operatorname{disc}^{\times}(A) < \sqrt{\frac{2}{\operatorname{sq}(A)}}.$$

PROOF. Assume $\operatorname{sq}(A) = d$. By definition, there are d rows $f_1,\dots,f_d \in \{-1,+1\}^N$ of A and a distribution μ on $\{1,\dots,N\}$ such that

$$\left|\mathop{\mathbf{E}}_{x\sim\mu}[f_i(x)\,f_j(x)]\right| \leqslant \frac{1}{d}$$

for all $i \neq j$. Let $\mathscr{U}$ be the uniform distribution over the d rows $f_1,\dots,f_d$ of A. Then by Lemma 10.5,

$$\operatorname{disc}_{\mu\times U}(A)^2 \leqslant \mathop{\mathbf{E}}_{i,j\sim\mathscr{U}}\left|\mathop{\mathbf{E}}_{x\sim\mu}[f_i(x)\,f_j(x)]\right| \leqslant \frac{1}{d}\cdot 1 + \frac{d-1}{d}\cdot\frac{1}{d} < \frac{2}{d}. \qquad \square$$

We now derive a matching lower bound on the product discrepancy in terms of the statistical query dimension.

LEMMA 10.7 (Sherstov [198]). *Let* $A \in \{-1,+1\}^{M\times N}$. *Then*

$$\operatorname{disc}^{\times}(A) > \frac{1}{8\operatorname{sq}(A)^2}.$$

PROOF. Let $\mu\times\lambda$ be an arbitrary product distribution over $\{1,\dots,N\}\times\{1,\dots,M\}$. We will obtain a lower bound on $\operatorname{disc}_{\mu\times\lambda}(A)$ by constructing an efficient protocol for A with a suitable advantage.

Let $\operatorname{sq}(A) = d$. Then Proposition 10.4 guarantees the existence of d rows $f_1,\dots,f_d \in \{-1,+1\}^N$ in A such that each of the remaining rows f satisfies $|\mathbf{E}_\mu[f(x)\,f_i(x)]| > 1/(d+1)$ for some $i = 1,\dots,d$. This yields the following protocol for evaluating A_{yx}. Bob, who knows the row index y, sends Alice the

index i of the function f_i whose correlation with the yth row of A is the greatest in absolute value. Bob additionally sends Alice the sign $\sigma \in \{-1, +1\}$ of the correlation of f_i and the yth row of A. This communication costs $\lceil \log d \rceil + 1$ bits. Alice, who knows the column index x, announces $\sigma \cdot f_i(x)$ as the output of the protocol.

For every fixed y, the described protocol achieves advantage greater than $1/(d+1)$ over the choice x. As a result, the protocol achieves overall advantage greater than $1/(d+1)$ with respect to any distribution λ on the rows of A. Since only $\lceil \log d \rceil + 2$ bits are exchanged, we obtain the sought bound on the discrepancy by Proposition 3.2:

$$\operatorname{disc}_{\mu\times\lambda}(A) > \frac{1}{(d+1)\cdot 2^{2+\lceil \log d \rceil}} \geqslant \frac{1}{8d^2},$$

where the second inequality holds because d is an integer. □

At last, we arrive at the main result of this section.

THEOREM 10.8 (Sherstov [198]). *Let $A \in \{-1, +1\}^{M\times N}$. Then*

$$\sqrt{\frac{1}{2}\operatorname{sq}(A)} < \frac{1}{\operatorname{disc}^{\times}(A)} < 8\operatorname{sq}(A)^2.$$

PROOF. Immediate from Lemmas 10.6 and 10.7. □

Observe that the product discrepancy is the same for a sign matrix as for its transpose: $\operatorname{disc}^{\times}(A) = \operatorname{disc}^{\times}(A^{\mathsf{T}})$. In particular, Theorem 10.8 has the following interesting corollary that the rows and columns of a sign matrix have the same statistical query dimension, up to a polynomial factor.

COROLLARY 10.9 (Sherstov [198]). *Let $A \in \{-1, +1\}^{M\times N}$. Then*

$$\left(\frac{\operatorname{sq}(A)}{128}\right)^{1/4} < \operatorname{sq}(A^{\mathsf{T}}) < 128\operatorname{sq}(A)^4.$$

10.4 Product discrepancy vs. sign-rank

In this section, we analyze the relationship between the product discrepancy of a sign matrix and its sign-rank. Our main result is that the product discrepancy implies a lower bound on the sign-rank and that the gap between the two can be arbitrary. In view of the equivalence of the product discrepancy and statistical query dimension, proved in the previous section, it will be convenient to work with the statistical query dimension instead. Along the way, we will substantially sharpen the upper bound on the statistical query dimension of halfspaces derived by Blum et al. [36].

THEOREM 10.10 (Sherstov [198]). *Fix a finite set X and arbitrary functions $\phi_1, \ldots, \phi_k : X \to \mathbb{R}$. Let $\mathscr{C}$ be the set of all Boolean functions $f : X \to \{-1, +1\}$ representable as $f(x) \equiv \operatorname{sgn}(\sum_{i=1}^{k} a_i \phi_i(x))$ for some reals $a_1, \ldots, a_k$. Then*

$$\operatorname{sq}(\mathscr{C}) < 2k^2.$$

PROOF. Fix a distribution μ on X. We shall use the same technical tool, Forster's work [70] on sign-rank, as Simon [206], who proved this claim for μ uniform. Assume for simplicity that μ is rational, extension to the general case being straightforward. Then the weight $\mu(x)$ of each point x is an integral multiple of $1/M$, where M is a suitably large integer.

Let $N = \operatorname{sq}_\mu(\mathscr{C})$. By definition, there exists a set $\mathscr{F} \subseteq \mathscr{C}$ of $|\mathscr{F}| = N$ functions with $|\mathbf{E}_\mu[f(x)g(x)]| \leqslant 1/N$ for all distinct $f, g \in \mathscr{F}$. Consider the matrix $A \in \{-1, +1\}^{N \times M}$ whose rows are indexed by the functions in $\mathscr{F}$, whose columns are indexed by inputs $x \in X$ (an input x indexes exactly $\mu(x)M$ columns), and whose entries are given by $A = [f(x)]_{f,x}$. By Theorem 7.5,

$$N \leqslant \frac{(\operatorname{rk}_\pm A)^2 \, \|A\|^2}{M}. \tag{10.1}$$

We complete the proof by obtaining upper bounds on $\operatorname{rk}_\pm A$ and $\|A\|$.

We analyze the sign-rank of A first. Recall that each $f \in \mathscr{F}$ has a representation $f(x) = \operatorname{sgn}(\sum_{i=1}^{k} a_{f,i} \phi_i(x))$, where $a_{f,1}, \ldots, a_{f,k}$ are reals specific to f.

Therefore, A has the same sign pattern as the matrix

$$\left[\sum_{i=1}^{k} a_{f,i}\phi_i(x)\right]_{f,x} = \sum_{i=1}^{k} [a_{f,i}\phi_i(x)]_{f,x}.$$

The last equation shows that A is sign-representable by the sum of k matrices of rank 1, whence

$$\operatorname{rk}_{\pm} A \leqslant k. \tag{10.2}$$

We now turn to $\|A\|$. The $N \times N$ matrix AA^{T} is given by

$$AA^{\mathsf{T}} = \Big[M \cdot \mathop{\mathbf{E}}_{x\sim\mu}[f(x)g(x)]\Big]_{f,g}.$$

This means that the diagonal entries of AA^{T} are equal to M, whereas the off-diagonal entries do not exceed M/N in absolute value. As a consequence,

$$\begin{aligned}\|A\|^2 &= \|AA^{\mathsf{T}}\| \\ &\leqslant \|M \cdot I\| + \|AA^{\mathsf{T}} - M \cdot I\| \\ &\leqslant \|M \cdot I\| + \|AA^{\mathsf{T}} - M \cdot I\|_{\mathrm{F}} \\ &\leqslant M + M\sqrt{\frac{N(N-1)}{N^2}}.\end{aligned} \tag{10.3}$$

Substituting the estimates (10.2) and (10.3) in (10.1) yields the inequality

$$N \leqslant k^2\left(1 + \sqrt{1 - \frac{1}{N}}\right),$$

whence

$$N \leqslant 2k^2 - \frac{1}{4}.$$

This completes the proof for rational μ. To extend the analysis to irrational distributions μ, one considers a rational distribution $\tilde{\mu}$ that approximates μ closely enough and follows the same reasoning. We omit these standard manipulations. □

REMARK 10.11 (Sherstov [198]). Observe that the proof of Theorem 10.10 actually yields the following stronger result. For a distribution μ, let N be the size of the largest set $\{f_1, \ldots, f_N\} \subseteq \mathscr{C}$ with average (not maximum) pairwise correlation at most $1/N$:

$$\frac{1}{N(N-1)} \sum_{i \neq j} \left(\mathop{\mathbf{E}}_{\mu} [f_i(x) f_j(x)] \right)^2 \leqslant \frac{1}{N^2}.$$

Clearly, N is at least the statistical query dimension of $\mathscr{C}$. The above proof of Theorem 10.10 establishes an upper bound on this larger quantity: $N < 2k^2$.

Recall that a halfspace is a Boolean function $f : \{0,1\}^n \to \{-1,+1\}$ representable as $f(x) \equiv \operatorname{sgn}(\sum a_i x_i - \theta)$ for some fixed reals $a_1, \ldots, a_n, \theta$. Halfspaces are arguably the most studied concept class [139, 216, 123, 121, 127, 128] in computational learning theory, with applications in areas as diverse as data mining, artificial intelligence, and computer vision. In a fundamental paper, Blum et al. [36] gave a polynomial-time algorithm for learning halfspaces in the *statistical query* model under arbitrary distributions. It follows from their work that the statistical query dimension of halfspaces is $O(n^c)$, where $c > 0$ is a sufficiently large constant. As an immediate corollary to Theorem 10.10, we substantially sharpen that upper bound.

COROLLARY 10.12 (Sherstov [198]). *Let $\mathscr{C}$ be the concept class of halfspaces $\{0,1\}^n \to \{-1,+1\}$. Then*

$$\operatorname{sq}(\mathscr{C}) < 2(n+1)^2.$$

The statistical query dimension of halfspaces is at least $n+1$ under the uniform distribution, as one can easily verify by considering the halfspaces $(-1)^{x_1}$, $(-1)^{x_2}, \ldots, (-1)^{x_n}, -1$. Thus, the quadratic upper bound of Corollary 10.12 is not far from optimal. In addition to strengthening the estimate of Blum et al. [36], Corollary 10.12 has a much simpler and shorter proof that builds only on Forster's self-contained theorem [70]. The proof of Blum et al. relies on nontrivial notions from computational geometry and requires a lengthy analysis of robustness under noise. That said, the proof of Blum et al. has the substantial advantage of giving an explicit learning algorithm, unlike the existential proof in this section.

Theorem 10.10 admits the following restatement in matrix terminology.

THEOREM 10.13 (Sherstov [198]). *Let $A \in \{-1,+1\}^{M \times N}$ be an arbitrary matrix. Then*

$$\operatorname{sq}(A) < 2(\operatorname{rk}_{\pm} A)^2.$$

We now prove that the gap between the two quantities in Theorem 10.13 can be arbitrary.

THEOREM 10.14 (Sherstov [197]). *Let $\varepsilon > 0$ be an arbitrary constant. Then there exists a matrix $A \in \{-1,+1\}^{N \times N}$ with*

$$\operatorname{sq}(A) = O(1),$$

$$\operatorname{rk}_{\pm} A = \Omega(N^{1-\varepsilon}).$$

PROOF. Let $c = 2\lceil 1/\varepsilon \rceil$. By Theorem 10.1, there exists a matrix $A \in \mathscr{Z}(N,c)$ with $\operatorname{rk}_{\pm} A = \Omega(N^{1-\varepsilon})$. On the other hand, it is clear that every matrix in $\mathscr{Z}(N,c)$

has Vapnik-Chervonenkis dimension at most $2c$, whence $\mathrm{sq}(A) \leqslant 2^{O(c)} = O(1)$ by Lemma 10.3. □

Since the sign-rank of an $N \times N$ sign matrix is at most N, Theorem 10.14 gives essentially the largest possible gap between the statistical query dimension and sign-rank that can exist by definition. We close this section with a lower bound on the product discrepancy in terms of sign-rank, which will be useful later in this chapter.

THEOREM 10.15 (Sherstov). *Let* $A \in \{-1, +1\}^{M \times N}$. *Then*

$$\mathrm{disc}^{\times}(A) > \frac{1}{32(\mathrm{rk}_{\pm} A)^4}.$$

PROOF. Immediate from Lemma 10.7 and Theorem 10.13. □

10.5 Sign-rank vs. nonproduct discrepancy, or $\mathsf{PP}^{\mathsf{cc}} \subsetneq \mathsf{UPP}^{\mathsf{cc}}$

In this section, we examine the relationship between the sign-rank and discrepancy. We start by citing a well-known bound on the former in terms of the latter. We then show by an explicit construction that the gap between the two quantities can be exponential. In particular, we will prove the separation $\mathsf{PP}^{\mathsf{cc}} \subsetneq \mathsf{UPP}^{\mathsf{cc}}$, where the communication classes $\mathsf{PP}^{\mathsf{cc}}$ and $\mathsf{UPP}^{\mathsf{cc}}$ correspond to sign matrices with non-negligible discrepancy and low sign-rank, respectively.

There are various ways to bound the sign-rank of a matrix in terms of its discrepancy. The derivation that we present here is based on two well-known results in the literature. The first of these, due to Ben-David et al. [35], is an application of the random projection technique of Arriaga and Vempala [20].

THEOREM 10.16 (Ben-David et al. [35]). *Let* $A \in \{-1, +1\}^{M \times N}$. *Then*

$$\mathrm{rk}_{\pm} A \leqslant O(\mathrm{mc}(A)^2 \log(M + N)).$$

The second result, due to Linial and Shraibman [145], proves that margin complexity and discrepancy are essentially equivalent notions.

THEOREM 10.17 (Linial and Shraibman [145]). *Let A be a sign matrix. Then*

$$\frac{1}{8}\,\mathrm{mc}(A) \leqslant \frac{1}{\mathrm{disc}(A)} \leqslant 8\,\mathrm{mc}(A).$$

An immediate consequence of Theorems 10.16 and 10.17 is the sought relationship between the sign-rank and discrepancy.

THEOREM 10.18 (Sign-rank vs. discrepancy). *Let $A \in \{-1,+1\}^{M\times N}$. Then*

$$\mathrm{rk}_{\pm}(A) \leqslant O\left(\frac{\log(M+N)}{\mathrm{disc}(A)^2}\right).$$

One can alternately derive Theorem 10.18 by a communication-based argument, without appealing to margin complexity. The above derivation, however, is more economical for our purposes because we will have occasion to refer to Theorems 10.16 and 10.17 later on.

Now that we have seen that the sign-rank of a matrix can be bounded in terms of its discrepancy, we will show an arbitrary gap between the two quantities. Consider the Boolean function $\mathrm{GHR}_n\colon \{-1,+1\}^{4n^2} \times \{-1,+1\}^{2n} \to \{-1,+1\}$ given by

$$\mathrm{GHR}_n(x,y) = \mathrm{sgn}\left(1 + \sum_{j=0}^{2n-1} y_j \sum_{i=0}^{n-1} 2^i (x_{i,2j} + x_{i,2j+1})\right).$$

This function was defined and studied by Goldmann et al. [82] in the context of separating classes of threshold circuits. Their analysis exhibits a nonproduct distribution with respect to which $\mathrm{GHR}_n(x,y)$ has high distributional complexity:

THEOREM 10.19 (Goldmann et al. [82]). *There is an* (*explicitly given*) *nonproduct distribution* λ *such that every deterministic one-way protocol for* GHR_n *with advantage* γ *with respect to* λ *has cost at least* $\log(\gamma\, 2^{n/2}/\sqrt{n}) - O(1)$.

A key consequence of Theorem 10.19 for our purposes is as follows.

LEMMA 10.20 (Sherstov [198]). *Let* GHR_n *be as defined above. Then*

$$\mathrm{disc}(\mathrm{GHR}_n) = O(\sqrt{n}2^{-n/2}).$$

PROOF. Consider the distribution λ from Theorem 10.19. Let R be a rectangle for which the discrepancy $\mathrm{disc}_\lambda(\mathrm{GHR}_n)$ is achieved:

$$\mathrm{disc}_\lambda(\mathrm{GHR}_n) = \left| \sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}_n(x,y) \right|.$$

We claim that there is a deterministic one-way protocol for $\mathrm{GHR}_n(x,y)$ with constant cost and with advantage at least $\mathrm{disc}_\lambda(\mathrm{GHR}_n)$ with respect to λ. Namely, define

$$a = \mathrm{sgn}\left(\sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}_n(x,y) \right),$$
$$b = \mathrm{sgn}\left(\sum_{(x,y)\notin R} \lambda(x,y)\mathrm{GHR}_n(x,y) \right).$$

Consider the protocol P that outputs a if the input is in R, and outputs b otherwise. By definition, the advantage of P with respect to λ is

$$\begin{aligned}
&\sum_{x,y} \lambda(x,y)P(x,y)\mathrm{GHR}_n(x,y) \\
&\qquad = a \sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}_n(x,y) + b \sum_{(x,y)\notin R} \lambda(x,y)\mathrm{GHR}_n(x,y) \\
&\qquad \geqslant \left| \sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}_n(x,y) \right| \\
&\qquad = \mathrm{disc}_\lambda(\mathrm{GHR}_n).
\end{aligned}$$

But by Theorem 10.19, every one-way constant-cost protocol achieves advantage at most $O(\sqrt{n}2^{-n/2})$. Thus, $\mathrm{disc}_\lambda(\mathrm{GHR}_n) = O(\sqrt{n}2^{-n/2})$. □

We arrive at the sought separation of sign-rank and discrepancy.

THEOREM 10.21 (Sherstov [198]). *There is an* (*explicitly given*) *sign matrix* $A \in \{-1,+1\}^{N\times N^{\log N}}$ *with*

$$\mathrm{rk}_\pm A \leqslant \log N,$$

$$\mathrm{disc}(A)^{-1} \geqslant \Omega\left(\frac{N^{1/4}}{\sqrt{\log N}}\right).$$

PROOF. Consider the matrix $A = [\mathrm{GHR}_n(x,y)]_{x,y}$. By definition,

$$\mathrm{rk}_\pm A \leqslant \mathrm{rk}\left[1 + \sum_{j=0}^{2n-1} y_j \sum_{i=0}^{n-1} 2^i (x_{i,2j} + x_{i,2j+1})\right]_{x,y} \leqslant 2n+1.$$

On the other hand, $\mathrm{disc}(A) \leqslant O(\sqrt{n}2^{-n/2})$ by Lemma 10.20. □

Independently of the author, Buhrman et al. [53] exhibited a different function with an exponential gap between the sign-rank and the reciprocal of the discrepancy.

THEOREM 10.22 (Buhrman et al. [53]). *There is an* (*explicitly given*) *matrix* $A \in \{-1,+1\}^{N\times N}$ *with*

$$\mathrm{rk}_{\pm} A \leqslant \log N,$$

$$\mathrm{disc}(A)^{-1} \geqslant \exp\{\Omega(\log^{1/3} N)\}.$$

PROOF (adapted from [53]). Define

$$A = \left[\operatorname{sgn}\left(1 + \sum_{i=1}^{n} (-2)^i x_i y_i\right)\right]_{x,y\in\{0,1\}^n}.$$

Then it is clear that the sign-rank of A does not exceed $n+1$, whereas Theorem 4.16 states that $\mathrm{disc}(A) = \exp\{-\Omega(n^{1/3})\}$. □

Recall from Section 3.4 that $\mathsf{PP}^{\mathsf{cc}}$ and $\mathsf{UPP}^{\mathsf{cc}}$ are the communication classes that correspond to sign matrices with low sign-rank and nonnegligible discrepancy, respectively. Theorem 10.18 makes it clear that

$$\mathsf{PP}^{\mathsf{cc}} \subseteq \mathsf{UPP}^{\mathsf{cc}},$$

settling the promised containment (3.7). Moreover, Theorems 10.21 and 10.22 each immediately imply that the containment is proper:

COROLLARY 10.23 (Buhrman et al. [53], Sherstov [198]).

$$\mathsf{PP}^{\mathsf{cc}} \not\subseteq \mathsf{UPP}^{\mathsf{cc}}.$$

Another useful perspective on the results of this section the following theorem, which shows that an exponential gap is achievable between the product and nonproduct discrepancy of a function.

THEOREM 10.24 (Sherstov). *There exists an* (*explicitly given*) *Boolean function* $f\colon \{-1,+1\}^{n^2} \times \{-1,+1\}^n \to \{-1,+1\}$ *with*

$$\operatorname{disc}^{\times}(f) \geqslant \Omega(n^{-4}),$$

$$\operatorname{disc}(f) \leqslant O(\sqrt{n}2^{-n/4}).$$

Furthermore, there exists an (*explicitly given*) *Boolean function* $g\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ *such that*

$$\operatorname{disc}^{\times}(g) \geqslant \Omega(n^{-4}),$$

$$\operatorname{disc}(g) \leqslant \exp\{-\Omega(n^{1/3})\}.$$

PROOF. Let $m = \lfloor n/2 \rfloor$ and define $f(x,y) = \mathrm{GHR}_m(x,y)$. Since the sign-rank of $[f(x,y)]_{x,y}$ is at most $2m+1$, the sought properties of f follow from Theorem 10.15 and Lemma 10.20. Next, define $g(x,y) = \operatorname{sgn}(1 + \sum_{i=1}^n (-2)^i x_i y_i)$. Since the sign-rank of $[g(x,y)]_{x,y}$ is at most $n+1$, the sought properties of g follow immediately from Theorems 10.15 and 4.16. □

We conclude this section with an observation about discrepancy that will be useful later in this chapter.

PROPOSITION 10.25 (Lower bound on discrepancy). *Let* $A \in \{-1,+1\}^{M \times N}$. *Then*

$$\operatorname{disc}(A) \geqslant \frac{1}{8 \min\left\{\sqrt{M}, \sqrt{N}\right\}}.$$

PROOF. Immediate from Proposition 2.16 and Theorem 10.17. □

10.6 Product vs. nonproduct distributional complexity

Let $f: X \times Y \to \{-1, +1\}$ be a given communication problem. Yao's well-known minimax principle [224], stated as Theorem 3.1 above, gives the following relationship between the randomized and distributional communication complexities of f:

$$R_\varepsilon(f) = \max_\mu \{D^\mu_\varepsilon(f)\}.$$

This equation has been the basis for essentially all lower bounds on randomized communication complexity: one defines a probability distribution μ on $X \times Y$ and argues that the cost $D^\mu_\varepsilon(f)$ of the best deterministic protocol with error at most ε over μ must be high. The main question, then, is what distribution μ to consider. Product distributions $\mu(x, y) = \mu_X(x)\mu_Y(y)$ are particularly attractive because they are easier to analyze. Unfortunately, they do not always lead to optimal lower bounds. A standard example of this phenomenon is the disjointness function $\mathrm{DISJ}_n(x, y) = \bigvee_{i=1}^n (x_i \wedge y_i)$, whose randomized complexity is $\Theta(n)$ bits [102, 176] and whose distributional complexity is $O(\sqrt{n}\log n)$ under every product distribution [23, §8].

Let $D^\times_\varepsilon(f)$ stand for the maximum distributional complexity of f under a product distribution:

$$D^\times_\varepsilon(f) = \max_{\mu \text{ product}} \{D^\mu_\varepsilon(f)\}.$$

Motivated by the above considerations, Kushilevitz and Nisan [137, p. 37] posed the problem of estimating the gap between $D^\times_{1/3}(f)$ and $R_{1/3}(f)$. In particular, they asked whether the gap between the two quantities is at most polynomial for all f. Kremer et al. [135] studied this question in the context of *one-way* randomized protocols and obtained a separation of $O(1)$ versus $\Omega(n)$ for the function GREATER-THAN on n-bit strings. Unfortunately, a function can have vastly different communication complexity in the one-way model and the usual, two-way model. Such is the case of GREATER-THAN, whose two-way randomized complexity is a mere $O(\log n)$. On a different front, we proved in Theorem 10.24 that the product dis-

crepancy $\mathrm{disc}^{\times}(f)$ of a function can be exponentially larger than its nonproduct discrepancy $\mathrm{disc}(f)$. In particular, our work in previous sections shows that the use of nonproduct distributions is indeed essential to the discrepancy method.

In this section, we will solve the Kushilevitz-Nisan problem completely and in its original form. Specifically, we will prove the existence of a function f with $D^{\times}_{1/3}(f) = O(1)$ and $R_{1/3}(f) = \Omega(n)$.

THEOREM 10.26 (Sherstov [197]). *Let $\varepsilon > 0$ be an arbitrary constant. Then there exists a function $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ with all of the following properties:*

$$\begin{aligned} D^{\times}_{\varepsilon}(f) &= O(1), \\ R_{1/3}(f) &= \Omega(n), \\ \mathrm{disc}^{\times}(f) &= \Omega(1), \\ \mathrm{disc}(f) &= O\left(2^{-n(\frac{1}{2}-\varepsilon)}\right), \\ U(f) &= \Omega(n). \end{aligned}$$

A key aspect of Theorem 10.26 is that the function f in question has exponentially small discrepancy. Indeed, its discrepancy essentially meets the $\Omega(2^{-n/2})$ lower bound given by Proposition 10.25 for any function on n-bit strings. As a result, f has communication complexity $\Omega(n)$ not only in the randomized model, but also in the nondeterministic and various quantum models. Furthermore, the communication complexity of f remains $\Omega(n)$ even if one simply seeks a randomized or quantum protocol with advantage $2^{-n/4}$ over random guessing. Finally, f has complexity $\Omega(n)$ in the unbounded-error model, which is even more powerful. In summary, f has the highest communication complexity in all standard models, and yet the distributional method restricted to product distributions can certify at best an $\Omega(1)$ lower bound. Theorem 10.26 also improves on our previously obtained exponential separation between product and nonproduct discrepancy (Theorem 10.24), although the new function is no longer explicitly given.

As a starting point in our technical development, we recall an elegant simulation that relates the communication complexity of a sign matrix to its Vapnik-Chervonenkis dimension.

THEOREM 10.27 (Kremer et al. [135, Thm. 3.2]). *Let A be a sign matrix. Let ε be given with $0 < \varepsilon \leqslant 1/3$. Then*

$$D^{\times}_{\varepsilon}(A) = O\left(\frac{1}{\varepsilon}\operatorname{vc}(A)\log\frac{1}{\varepsilon}\right).$$

Moreover, this communication cost can be achieved with a one-way protocol.

PROOF (adapted from [135]). Let X and Y be the finite sets that index the columns and rows of A, respectively. Let $\mu(x, y) = \mu_X(x)\mu_Y(y)$ be a given product distribution. Consider the following randomized one-way protocol for A. On input $(x, y) \in X \times Y$, Alice and Bob use their shared random bits to pick points

$$x^{(1)},\ x^{(2)},\ \ldots,\ x^{(m)} \in X$$

independently at random, according to μ_X. Here m is a parameter to be fixed later. Next, Bob sends Alice the values

$$A(y, x^{(1)}),\ A(y, x^{(2)}),\ \ldots,\ A(y, x^{(m)}).$$

At this point, Alice identifies any $y' \in Y$ with

$$\begin{aligned} A(y', x^{(1)}) &= A(y, x^{(1)}), \\ A(y', x^{(2)}) &= A(y, x^{(2)}), \\ &\vdots \\ A(y', x^{(m)}) &= A(y, x^{(m)}), \end{aligned}$$

and announces $A(y', x)$ as the output of the protocol.

In the terminology of Section 2.5, the protocol amounts to Alice learning the unknown row A_y of the matrix A from random labeled examples distributed according to μ_X. By the Vapnik-Chervonenkis Theorem (Theorem 2.14), any row A'_y consistent with $m = O(\frac{1}{\varepsilon} \operatorname{vc}(A) \log \frac{1}{\varepsilon})$ labeled examples will, with probability at least $1 - \varepsilon/2$, have $\Delta_{\mu_X}(A'_y, A_y) \leqslant \varepsilon/2$. In particular, Alice's answer will be correct with probability at least $1 - \varepsilon$ (with respect to μ_X and regardless of Bob's input y).

Thus, we have obtained a randomized one-way protocol for A with cost m and error at most ε over μ. By an averaging argument, there must be a one-way deterministic protocol with the same cost and error at most ε with respect to μ. □

We are now in a position to prove the main result of this section, which is an elaboration of Theorem 10.26 above.

THEOREM 10.28. *Let ε be given, $0 < \varepsilon \leqslant 1/3$. Then there exists a function $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ with all of the following properties:*

$$D^{\times}_{\varepsilon}(f) \leqslant \Theta\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right),$$

$$R_{1/3}(f) \geqslant \Theta(n),$$

$$\operatorname{disc}^{\times}(f) \geqslant \varepsilon^{\Theta(1/\varepsilon^2)},$$

$$\operatorname{disc}(f) \leqslant O\left(2^{-n(\frac{1}{2}-\varepsilon)}\right),$$

$$U(f) \geqslant \Theta(n).$$

PROOF. Let $\alpha > 0$ be the absolute constant from Theorem 10.2. The inner product function $\mathrm{IP}_n(x,y) = \bigoplus_{i=1}^n (x_i \wedge y_i)$ satisfies $R_{1/3}(\mathrm{IP}_n) \geqslant D^{\times}_{1/3}(\mathrm{IP}_n) \geqslant \Omega(n)$ and $\operatorname{disc}^{\times}(\mathrm{IP}_n) \leqslant 2^{-n/2}$ by Proposition 3.5. One further has $\operatorname{disc}(\mathrm{IP}_n) = \Omega(2^{-n/2})$ by Proposition 10.25 as well as $U(\mathrm{IP}_n) = \Omega(n)$ by Theorem 7.5. In summary, the theorem holds for the inner product function when $\varepsilon < \frac{4}{\alpha}\sqrt{\log n / n}$.

In the contrary case, Theorem 10.2 is applicable with $c = 2\lceil 1/\varepsilon \rceil$ and ensures the existence of $A \in \mathscr{Z}(2^n, c)$ with $\mathrm{rk}_{\pm} A \geqslant \alpha 2^{n(1-\varepsilon)}$. Then

$$\begin{aligned} \mathrm{disc}(A) &\leqslant \frac{8}{\mathrm{mc}(A)} && \text{by Theorem 10.17} \\ &\leqslant O\left(\sqrt{\frac{n}{\mathrm{rk}_{\pm} A}}\right) && \text{by Theorem 10.16} \\ &\leqslant \Theta\left(2^{-n(\frac{1}{2}-\varepsilon)}\right). \end{aligned}$$

By Proposition 3.2, we immediately conclude that

$$R_{1/3}(A) \geqslant \Theta(n).$$

Since every matrix in $\mathscr{Z}(2^n, c)$ has Vapnik-Chervonenkis dimension at most $2c$, it follows from Theorem 10.27 that

$$D^{\times}_{\varepsilon}(A) \leqslant \Theta\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right). \tag{10.4}$$

In light of (10.4), Proposition 3.2 shows that

$$\mathrm{disc}^{\times}(A) \geqslant \varepsilon^{\Theta(1/\varepsilon^2)}.$$

At last, Theorem 7.2 shows that

$$U(A) = \log \mathrm{rk}_{\pm} A \pm O(1) = \Theta(n). \qquad \square$$

REMARK 10.29. The upper bound $D^{\times}_{\varepsilon}(f) \leqslant \left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$ stated in Theorem 10.28 can be achieved even by one-way protocols. This is because we bounded $D^{\times}_{\varepsilon}(f)$ using Theorem 10.27, which gives a one-way communication protocol for the task.

Chapter 11

Conclusions and Open Problems

11.1 Our contributions in communication complexity

In Chapters 3–10, we answered several fundamental questions in communication complexity. We started by developing a novel technique, the pattern matrix method, for proving communication lower bounds. The pattern matrix method converts well-studied analytic properties of Boolean functions, such their approximate degree or threshold degree, into lower bounds for the associated communication problems. Depending on the analytic property used, our method yields lower bounds in a variety of settings, including classical and quantum communication channels as well as bounded-error, unbounded-error, and small-bias protocols. We used our technique and additional analytic machinery to solve several open problems in communication complexity and circuit complexity, as follows.

First, we settled the relationship between two well-studied circuit classes, majority circuits and constant-depth AND/OR/NOT circuits, thereby solving an open problem posed in 1997 by Krause and Pudlák [132] and showing the optimality of Allender's classical simulation of constant-depth circuits [11].

Second, we obtained lower bounds for a new family of functions in the bounded-error quantum model, regardless of prior entanglement. Our results broadly subsume the celebrated work by Razborov [177], who studied symmetric functions in this context and used quite different techniques.

Third, we made progress on proving the conjectured polynomial equivalence of quantum and classical bounded-error communication. In particular, we proved this conjecture for the communication problems of computing $f(x \wedge y)$ and $f(x \vee y)$ on input $x, y \in \{0,1\}^n$, broadly subsuming previous work.

Fourth, we studied unbounded-error communication complexity. Specifically, we settled the unbounded-error communication complexity of symmetric functions and solved a longstanding open problem due to Babai et al. [23] on the comparative power of alternation and unbounded-error communication.

Fifth, we fully determined the relations and gaps among three key complexity measures of a communication problem: discrepancy, product discrepancy, and sign-rank. As as application, we solved the open problem on the distributional complexity under product and nonproduct distributions, posed in 1997 by Kushilevitz and Nisan [137]. As another application, we gave an exponential separation of the communication classes $\mathsf{PP}^{\mathsf{cc}}$ and $\mathsf{UPP}^{\mathsf{cc}}$, defined in 1986 by Babai et al. [23].

Finally, we discussed the broader impact of our work in the research community. We presented generalizations of the pattern matrix method by several authors to the multiparty model and the resulting progress on fundamental questions in communication complexity. The new consequences include improved lower bounds for the disjointness function [140, 59], an exponential separation of nondeterministic and randomized complexity [64, 65] , and near-tight communication lower bounds for constant-depth circuits [29]. We also contributed to this growing body of work a separation of nondeterministic and co-nondeterministic complexity, as well as separations involving Merlin-Arthur communication.

11.2 Open problems

There is no shortage in communication complexity of difficult, mathematically rich, and natural open problems. We conclude with several challenges that are directly related to the results and techniques of this thesis. One such concerns representations of Boolean functions by real polynomials.

OPEN PROBLEM 1. Develop strong new techniques for analyzing the approximate degree and threshold degree of Boolean functions.

The motivation behind this problem is clear in view of the applications of the approximate degree and threshold degree in the preceding chapters of this thesis as well as earlier literature. Essentially this question was posed by Aaronson in a recent tutorial [2].

To set the context for our next problem, recall from the previous chapters that the approximate degree of a function $f\colon \{0,1\}^n \to \{-1,+1\}$ characterizes the deterministic, randomized, and quantum bounded-error communication complexity of the associated pattern matrix M_f up to a polynomial. Analogously, the threshold weight of f characterizes the discrepancy of M_f. An analytic property conspicuously absent from this picture is the threshold degree, which likely characterizes the unbounded-error communication complexity of M_f. In summary, we have the following open problem [178].

OPEN PROBLEM 2. Prove that the threshold degree of a Boolean function implies a lower bound on the unbounded-error communication complexity of the corre-

sponding pattern matrix. More precisely, prove that the (n^c, n, f)-pattern matrix has unbounded-error communication complexity $\deg_{\pm}(f)^{\Omega(1)}$ for some constant $c > 1$ and every Boolean function $f : \{0,1\}^n \to \{-1,+1\}$.

Continuing, observe that the pattern matrix method relates analytic notions to combinatorial ones, viz., the analytic properties of a Boolean function to its communication complexity in different models. In the same vein, the next two open problems are well-known conjectures [147, 55] about the polynomial equivalence of analytic complexity measures and their natural combinatorial analogues. For example, one can consider the deterministic communication complexity of a sign matrix F, an essentially combinatorial notion, and its analytic counterpart $\log \operatorname{rk} F$. Similarly, one can consider the sensitivity of a Boolean function f and its counterpart block sensitivity, which is equivalent to several analytic properties of f. We have:

OPEN PROBLEM 3. Prove the log-rank conjecture of Lovász and Saks [147]. In other words, prove that every sign matrix F obeys $D(F) \leqslant (\log \operatorname{rk} F)^{O(1)} + O(1)$.

OPEN PROBLEM 4. Prove a polynomial equivalence of sensitivity and block sensitivity for every Boolean function. Formally, prove that for some constant $\alpha > 0$, every function $f : \{0,1\}^n \to \{-1,+1\}$ obeys $\operatorname{bs}(f)^{\alpha} \leqslant \operatorname{s}(f) \leqslant \operatorname{bs}(f)$.

Another natural question raised by several authors [54, 116, 205] and directly related to our work is the polynomial equivalence of the classical and quantum communication complexities for functions of the form $f(x \wedge y)$.

OPEN PROBLEM 5. Prove a polynomial equivalence of the quantum and classical bounded-error complexities for communication problems of the form $F(x,y) = f(x \wedge y)$, where $f : \{0,1\}^n \to \{-1,+1\}$ is any given function.

The next two open problems focus on computational models that, on the one hand, have long resisted attempts at proving lower bounds, but on the other hand seem very related to the communication and circuit questions successfully resolved in this thesis and earlier literature. Both problems are well-known [23, 132].

OPEN PROBLEM 6. Exhibit a family of sign matrices outside the communication classes Σ_2^{cc} and Π_2^{cc}. More ambitiously, separate the classes $\mathsf{PH}^{\mathsf{cc}}$ and $\mathsf{PSPACE}^{\mathsf{cc}}$.

OPEN PROBLEM 7. Exhibit a function $f: \{0,1\}^n \to \{-1,+1\}$ that requires a depth-2 threshold circuit of exponential size, regardless of the weights. More ambitiously, prove that the inner product function $\mathrm{IP}_n(x,y) = (-1)^{\sum x_i y_i}$ has this property.

We conclude with two natural avenues for work in multiparty communication complexity. The first involves proving near-tight lower bounds on the multiparty communication complexity of the disjointness function, matching the known upper bounds [87]. The second involves breaking the logarithmic barrier for the number of players in multiparty communication complexity.

OPEN PROBLEM 8. Derive a lower bound of $n^{\Omega(1)}/2^{O(k)}$ on the multiparty communication complexity of the disjointness function in the k-party bounded-error model.

OPEN PROBLEM 9. Exhibit a communication problem $f: (\{0,1\}^n)^k \to \{-1,+1\}$ with complexity $n^{\Omega(1)}$ in the bounded-error model with $k \gg \log n$ players.

Part II

Learning Theory

Chapter 12

Lower Bounds for PAC Learning

In this first chapter on learning theory, we focus on the problem of PAC learning intersections of halfspaces on the hypercube $\{0,1\}^n$. This problem has long resisted attack and remains a central challenge in the area. Our main result shows that in fact, under a widely believed cryptographic assumption, no efficient algorithm exists for the task. We obtain analogous hardness results for learning other concept classes, such as majority circuits and arithmetic circuits. Analytic representations of Boolean functions play a central role in our proofs.

12.1 Introduction

Recall that a halfspace on the hypercube $\{0,1\}^n$ is a Boolean function of the form $f(x) = \operatorname{sgn}(\sum a_i x_i - \theta)$, where $a_1, \ldots, a_n, \theta$ are some fixed weights. Halfspace-based learning algorithms have applications in many areas of computer science, including data mining, artificial intelligence, and computer vision. This chapter focuses on *intersections* of halfspaces, a natural extension of the concept class of halfspaces. While many efficient algorithms exist for PAC learning a single halfspace, the problem of learning the intersection of even two halfspaces remains a central challenge in computational learning theory. A variety of efficient algorithms have been developed for natural restrictions of the problem [139, 216, 121, 125]. At the same time, attempts to prove that the problem is hard have been met with limited success: all previous hardness results [40, 10] for PAC learning intersections of halfspaces applied only to the case of *proper learning*, whereby the learner's output hypothesis must itself be an intersection of halfspaces. We give a thorough review of the PAC model and other learning-theoretic background in Section 12.2.

The main contribution of this chapter is the first representation-independent hardness result for PAC learning intersections of halfspaces. Put differently, we place no restrictions on the learner's output hypothesis other than polynomial-time computability. We prove that, under a widely believed cryptographic assumption, there is no efficient algorithm for PAC learning the intersection of n^ε halfspaces on $\{0,1\}^n$. We obtain analogous results for majority circuits and arithmetic circuits. The cryptographic assumption on which our results are based is that of computational intractability for well-studied lattice-based problems, such as the unique shortest vector problem or the shortest independent vector problem. We defer a

more technical statement of our results and their extensions to the concluding two sections of the chapter.

Our proof is based on a well-known method for obtaining hardness results for a concept class $\mathscr{C}$, which consists in proving that $\mathscr{C}$ contains the *decryption functions* of a public-key cryptosystem. For this purpose, we use recent public-key cryptosystems due to Regev [183, 184], presented in detail in Section 12.3. The difficulty in this program, however, is that intersections of a polynomial number of halfspaces cannot directly compute the decryption functions of the cryptosystems that we use. In fact, the decryption functions in question contain PARITY as a subfunction, which cannot be computed by intersections of a polynomial number of any unate functions [127]. Furthermore, the decryption functions in Regev's cryptosystems perform a division or an iterated addition, requiring threshold circuits of depth 3 and 2, respectively [219, 209]. We overcome these difficulties by a purely analytic argument, namely, by exploiting nonuniform distributions on $\{0,1\}^n$ to help us with the computation. This technique allows us to draw on the computational power of intersections of *quadratic* polynomial threshold functions to compute the decryption functions, while still obtaining a hardness result for intersections of *halfspaces*. We present a detailed review of the cryptography-to-learning reduction in Section 12.4, followed by the actual implementations of the decryption functions in Sections 12.5 and 12.6.

12.2 Weak and strong PAC learning

A central model in learning theory is Valiant's PAC model [213], already reviewed briefly in Chapter 2. As is common in computational complexity, we will discuss PAC learning in the context of an infinite family $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_n, \ldots$, where $\mathscr{C}_n$ is a set of Boolean functions on $\{0,1\}^n$. It is a widely followed convention in the literature to refer to the sequence $\{\mathscr{C}_n\}$ as a *concept class*, although it is formally an infinite *sequence* of concept classes, each on a different domain. We will be using the formulation of the PAC model for arbitrary distributions. In this setting, one fixes a function $f \in \mathscr{C}_n$ and a distribution μ on $\{0,1\}^n$, both unknown to the learner, and provides the learner with $m = m(n)$ training examples $(x^{(1)}, f(x^{(1)})), \ldots, (x^{(m)}, f(x^{(m)}))$ labeled according to the unknown function f, where $x^{(1)}, \ldots, x^{(m)}$ are independent and identically distributed according to μ.

The learner is then called upon to produce, with probability at least $1 - \delta$, a hypothesis $h: \{0,1\}^n \rightarrow \{-1,+1\}$ such that $\mathbf{E}_\mu[f(x) \neq h(x)] < \varepsilon$. Here ε and δ are small positive constants, and the probability is taken over the random choice of examples and any randomized computation by the learner. The family $\{\mathscr{C}_n\}$ is said to be *efficiently PAC learnable*, or equivalently *PAC learnable in polynomial time*, if a successful learner exists that takes $m(n) < n^c$ labeled examples and runs in time at most n^c, for some constant $c > 1$ and all n. In this chapter, we will mostly be concerned with the PAC learnability of intersections of halfspaces.

The above version of the PAC model is also known as *strong* PAC learning, to distinguish it from the *weak* PAC learning model of Kearns and Valiant [105]. In the latter setting, the learner only needs to produce a hypothesis h with nonnegligible agreement with the target function f:

$$\mathop{\mathbf{P}}_{x \sim \mu}[f(x) \neq h(x)] < \frac{1}{2} - \frac{1}{n^c}$$

for a constant $c > 1$. A remarkable result, discovered by Schapire [191, 192], is that strong and weak learning are equivalent, i.e., a sequence $\{\mathscr{C}_n\}$ is strongly PAC learnable in polynomial time if and only if it is weakly PAC learnable in polynomial time. In this light, we will use the two variants of the PAC model interchangeably, as best fits the presentation. To further simplify the theorem statements, we will use the term "PAC learnable" as a shorthand for "efficiently PAC learnable." For further background on computational learning theory, we refer the reader to the text by Kearns and Vazirani [108].

Recall that throughout this thesis, we follow the convention whereby Boolean functions take on values -1 and $+1$. In this chapter, however, it will be useful to relax this rule in order to accommodate standard cryptographic notation. Specifically, we will sometimes represent the range of a Boolean function by $\{0,1\}$, where 0 and 1 represent "false" and "true," respectively.

12.3 Cryptographic preliminaries

A *lattice* in n dimensions is the set $\{a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n : a_1, \ldots, a_n \in \mathbb{Z}\}$ of all integral linear combinations of a given basis $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$. There are several well-studied computational problems on lattices. In the *unique shortest vector problem*, denoted $f(n)$-uSVP, the goal is to find a shortest nonzero vector in the lattice, provided that it is shorter by a factor of at least $f(n)$ than any other nonparallel vector. In the *shortest vector problem*, denoted $f(n)$-SVP, the goal is to approximate the length of a shortest nonzero vector within a factor of $f(n)$. Thus, uSVP is a special case of SVP, distinguished by the uniqueness condition. In the *shortest independent vector problem*, denoted $f(n)$-SIVP, the goal is to output a set of n linearly independent lattice vectors of length at most $f(n)\cdot\mathsf{opt}$, where opt is the minimum length over all sets of n linearly independent vectors from the lattice and the length of a set is defined as the length of its longest vector. Note that all three problems become harder as the approximation factor $f(n) \geqslant 1$ decreases. There is an extensive literature on the computational hardness depending on the approximation factor $f(n)$. For example, certain variants of the shortest vector problem are known to be NP-hard if $f(n)$ is a small constant. On the other hand, it is known that for larger values such as $f(n) = \sqrt{n}$, some lattice problems are unlikely to be NP-hard in that their NP-hardness would imply the collapse of the polynomial-time hierarchy. We refer the reader to the excellent survey by Regev [184] for an extensive treatment of the subject. In this chapter, we will be working with the setting $f(n) = \tilde{O}(n^{1.5})$, an approximation factor for which the above three problems are believed to be computationally difficult and in particular are not known to admit a subexponential-time solution.

We will use public-key cryptosystems due to Regev [180, 181], whose security is based on the conjectured computational hardness of the above lattice problems uSVP, SVP, and SIVP. As Regev points out [181], an advantage of these problems from a cryptographic standpoint is the equivalence of their worst-case and average-case complexity. In other words, an efficient algorithm for solving these problems on a nonnegligible (inverse-polynomial) fraction of instances yields an efficient algorithm for solving every instance. This contrasts with common number-theoretic problems such as factoring or deciding quadratic residuosity. Furthermore, lattice-based cryptosystems feature decryption functions that are completely different from modular exponentiation $d(Y) = Y^D \bmod N$, the decryption function

that is at the heart of virtually every number-theoretic cryptosystem. As a result, lattice-based cryptosystems have the potential to lead to hardness results that earlier, number-theoretic cryptosystems have not yielded.

The lattice-based cryptosystems below encrypt one-bit messages, 0 and 1. The encryption is randomized, whereas the decryption is deterministic. Let $e_{K,r}\colon \{0,1\} \to \{0,1\}^{\mathsf{poly}(n)}$ denote the encryption function corresponding to a choice of private and public keys $K = (K_{\text{priv}}, K_{\text{pub}})$ and a random string r. In discussing security, we will need the following notion.

DEFINITION 12.1 (Distinguisher). An algorithm $\mathscr{A}$ is said to distinguish between the encryptions of 0 and 1 if for some universal constant $c > 1$,

$$\left| \mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(1)) = 1] - \mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(0)) = 1] \right| \geqslant \frac{1}{n^c}.$$

We will focus on those aspects of the cryptosystems that are relevant to the hardness results of this chapter. For example, we will state the numeric ranges of public and private keys without describing the key generation procedure. We will follow the convention of denoting polynomially-bounded quantities (in n) by lowercase letters, and superpolynomial ones by capital letters.

The uSVP-based cryptosystem. We start with a lattice cryptosystem, due to Regev [180], whose security is based on the worst-case hardness of uSVP. Let n be the security parameter. Denote $N = 2^{8n^2}$ and $m = cn^2$, where c is a universal constant. Let $\gamma(n)$ be any function with $\gamma(n) = \omega(n\sqrt{\log n})$, where faster-growing functions γ correspond to worse security guarantees but also a lower probability of decryption error.

Private key: A real number H with $\sqrt{N} \leqslant H < 2\sqrt{N}$.

Public key: A vector $(A_1, \ldots, A_m, i_0)$, where $i_0 \in \{1, \ldots, m\}$ and each $A_i \in \{0, \ldots, N-1\}$.

Encryption: To encrypt 0, pick a random set $S \subseteq [m]$ and output $\sum_{i \in S} A_i \bmod N$.
To encrypt 1, pick a random $S \subseteq [m]$ and output $\lfloor A_{i_0}/2 \rfloor + \sum_{i \in S} A_i \bmod N$.

Decryption: On receipt of $W \in \{0, \ldots, N-1\}$, decrypt 0 if $\mathrm{frac}(WH/N) < 1/4$, and 1 otherwise. Here $\mathrm{frac}(a) = \min\{\lceil a \rceil - a, a - \lfloor a \rfloor\}$ denotes the distance from $a \in \mathbb{R}$ to the closest integer. By a standard argument, the security and correctness of the cryptosystem are unaffected if we change the decryption function to $\mathrm{frac}(AW) < 1/4$, where A is a representation of H/N to within $\mathsf{poly}(n)$ fractional bits.

Correctness: The probability of decryption error, over the choice of private and public keys and the randomness in the encryption, is $\exp\{-\Omega(\gamma(n)^2/m)\}$.

Regev [180] showed that breaking the above cryptosystem would yield a polynomial-time algorithm for the unique shortest vector problem. A more detailed statement follows, cf. Theorem 4.5 and Lemma 5.4 of [180].

THEOREM 12.2 (Regev [180]). *Assume that there is a polynomial-time distinguisher between the encryptions of* 0 *and* 1. *Then there is a polynomial-time solution to every instance of* $(\sqrt{n} \cdot \gamma(n))$*-uSVP.*

We will set $\gamma(n) = n \log n$ to make the probability of decryption error negligible (inverse-superpolynomial) while guaranteeing $\tilde{O}(n^{1.5})$-uSVP security. In particular, Regev's cryptosystem improves on the public-key cryptosystem of Ajtai and Dwork [9] whose security is based on the worst-case hardness of $O(n^8)$-uSVP, an easier problem than $\tilde{O}(n^{1.5})$-uSVP.

The SVP- and SIVP-based cryptosystem. The second cryptosystem [181] is based on the worst-case quantum hardness of SVP and SIVP. Let n be the security parameter. Denote by p a prime with $n^2 < p < 2n^2$, and let $m = 5(n+1)(1+2\log n)$. Let $\gamma(n)$ be any function with $\gamma(n) = \omega(\sqrt{n}\log n)$, where faster-growing functions γ correspond to worse security guarantees but also a lower probability of decryption error.

Private key: A vector $\mathbf{s} \in \mathbb{Z}_p^n$.

Public key: A sequence of pairs $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_m, b_m)$, where each $\mathbf{a}_i \in \mathbb{Z}_p^n$ and $b_i \in \mathbb{Z}_p$.

Encryption: To encrypt 0, pick $S \subseteq [m]$ randomly and output $(\sum_{i \in S} \mathbf{a}_i, \sum_{i \in S} b_i)$. To encrypt 1, pick $S \subseteq [m]$ randomly and output $(\sum_{i \in S} \mathbf{a}_i, \lfloor p/2 \rfloor + \sum_{i \in S} b_i)$. All arithmetic is modulo p.

Decryption: On receipt of $(\mathbf{a}, b) \in \mathbb{Z}_p^n \times \mathbb{Z}_p$, decrypt 0 if $b - \langle \mathbf{a}, \mathbf{s} \rangle$ is closer to 0 than to $\lfloor p/2 \rfloor$ modulo p. Decrypt 1 otherwise. All arithmetic is modulo p.

Correctness: The probability of decryption error, over the choice of private and public keys and the randomness in the encryption, is $\exp\{-\Omega(\gamma(n)^2/m)\}$.

Regev [181] showed that breaking the above cryptosystem would imply a polynomial-time quantum algorithm for solving SVP and SIVP. A more precise statement is as follows, cf. Theorem 3.1 and Lemmas 4.4, 5.4 of [181].

THEOREM 12.3 (Regev [181]). *Assume that there is a polynomial-time* (*possibly quantum*) *algorithm for distinguishing between the encryptions of* 0 *and* 1. *Then there is a polynomial-time quantum solution to* $\tilde{O}(n \cdot \gamma(n))$*-SVP and* $\tilde{O}(n \cdot \gamma(n))$*-SIVP.*

We adopt the setting $\gamma(n) = \sqrt{n} \log^2 n$ to make the probability of decryption error negligible while guaranteeing $\tilde{O}(n^{1.5})$-SVP and $\tilde{O}(n^{1.5})$-SIVP security. Observe that this second cryptosystem is preferable to the first in that it is based on the worst-case hardness of a more general lattice problem (SVP rather than uSVP). The disadvantage of the second cryptosystem is that breaking it would only yield a *quantum* algorithm for SVP, as opposed to the first cryptosystem which would yield a classical algorithm for uSVP.

12.4 From cryptography to learning theory

Cryptographic assumptions have long been used to obtain hardness results for PAC learning. In a seminal work, Goldreich et al. [84] proved the first representation-independent hardness result for PAC learning, ruling out an efficient algorithm for learning polynomial-size Boolean circuits under the uniform distribution with or without membership queries. Kearns and Valiant [105] used number-theoretic problems to obtain hardness results for NC^1 circuits, constant-depth threshold circuits

TC^0, and deterministic finite automata. Kharitonov [111] obtained hardness results for AC^1 and NC^1 circuits based on the conjectured hardness of the subset sum problem. Kharitonov [112] later used the Blum-Blum-Shub pseudorandom generator [41] to obtain a hardness result for learning the circuit classes AC^0 and TC^0 that holds even under the uniform distribution and if membership queries are allowed.

Hardness results such as the above exploit a fundamental relationship between the security of a public-key cryptosystem and the hardness of learning the associated concept class. We rederive it below for completeness and extend it to allow for errors in the decryption process. This relationship is a natural consequence of the ease of encrypting messages with the public key. A large pool of such encryptions can be viewed as a set of training examples for learning the decryption function. But learning the decryption function to a nonnegligible advantage would mean breaking the cryptosystem. Assuming that the cryptosystem is secure, we can thus conclude that it is not feasible to learn the decryption function. We formalize this intuition as follows.

LEMMA 12.4 (cf. Kearns and Valiant [105]). *Consider a public-key cryptosystem for encrypting individual bits by n-bit strings. Let $\mathscr{C}$ be a concept class that contains all the decryption functions $d_K: \{0,1\}^n \to \{0,1\}$ of the cryptosystem, one for each choice of key $K = (K_{\text{priv}}, K_{\text{pub}})$. Let*

$$\varepsilon(n) = \mathop{\mathbf{P}}_{K,r}[d_K(e_{K,r}(0)) \neq 0 \text{ or } d_K(e_{K,r}(1)) \neq 1]$$

be the probability of decryption error, over the choice of keys and randomization in the encryption. If $\mathscr{C}$ is weakly PAC-learnable in time $t(n)$ with $t(n)\varepsilon(n) = n^{-\omega(1)}$, then there is a distinguisher between the encryptions of 0 and 1 that runs in time $O(t(n))$.

PROOF. For a pair of keys $K = (K_{\text{priv}}, K_{\text{pub}})$, let $e_{K,r}: \{0,1\} \to \{0,1\}^n$ be the randomized encryption function, indexed by the choice of random string r. Let $d_K: \{0,1\}^n \to \{0,1\}$ denote the matching decryption function. We will use the assumed learnability of $\mathscr{C}$ to exhibit an algorithm $\mathscr{A}$ that runs in time $O(t(n))$ and

has

$$\mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(1)) = 1] - \mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(0)) = 1] \geqslant \frac{1}{n^c}$$

for some universal constant c, as long as $t(n)\varepsilon(n) = n^{-\omega(1)}$. The probability is taken over the choice of keys, randomness in the encryption, and any internal randomization in $\mathscr{A}$. It will thus follow that $\mathscr{A}$ is the desired distinguisher.

Algorithm $\mathscr{A}$ takes as input a pair (K_{pub}, w), where $w \in \{0,1\}^n$ is the encryption of an unknown bit. First, $\mathscr{A}$ draws $t(n)$ independent training examples, choosing each as follows.

1. Pick $b = 0$ or $b = 1$, with equal probability.
2. Pick r, an unbiased random string.
3. Create a training example $\langle e_{K,r}(b), b\rangle$.

Next, $\mathscr{A}$ passes the training examples to the assumed algorithm for learning $\mathscr{C}$. Assume no decryption error has occurred, i.e., the decryption function d_K is consistent with all the generated examples. Then the learning algorithm outputs a hypothesis h that approximates d_K with a nonnegligible advantage:

$$\mathop{\mathbf{P}}_{b,r}[h(e_{K,r}(b)) = d_K(e_{K,r}(b))] \geqslant \frac{1}{2} + \frac{1}{n^c}, \tag{12.1}$$

for some constant c. With this hypothesis in hand, algorithm $\mathscr{A}$ outputs $h(w)$ and exits. It remains to show that $\mathscr{A}$ is indeed a distinguisher. We will first handle the

case in which no decryption error occurs; call this event $\overline{E}$. Then

$$\begin{aligned}
&\mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(1)) = 1 \mid \overline{E}] - \mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(0)) = 1 \mid \overline{E}] \\
&\qquad = \mathop{\mathbf{P}}_{K,r}[h(e_{K,r}(1)) = 1] - \mathop{\mathbf{P}}_{K,r}[h(e_{K,r}(0)) = 1] \\
&\qquad = 2 \mathop{\mathbf{P}}_{K,b,r}[h(e_{K,r}(b)) = b] - 1 \\
&\qquad \geqslant 2 \left(\mathop{\mathbf{P}}_{K,b,r}[h(e_{K,r}(b)) = d_K(e_{K,r}(b))] - \mathop{\mathbf{P}}_{K,b,r}[d_K(e_{K,r}(b)) \neq b] \right) - 1 \\
&\qquad \geqslant 1 + \frac{2}{n^c} - 2\varepsilon(n) - 1 \\
&\qquad = \frac{2}{n^c} - 2\varepsilon(n).
\end{aligned}$$

We now extend the analysis to account for possible decryption errors. Observe that the likelihood of a decryption error on a run of $\mathscr{A}$ is small:

$$\begin{aligned}
\mathbf{P}[E] &= \mathop{\mathbf{E}}_{K}[\mathbf{P}[E \mid K]] \\
&\leqslant \mathop{\mathbf{E}}_{K}\left[t(n) \cdot \mathop{\mathbf{P}}_{b,r}[d_K(e_{K,r}(b)) \neq b \mid K] \right] \\
&= t(n) \cdot \mathop{\mathbf{P}}_{K,b,r}[d_K(e_{K,r}(b)) \neq b] \\
&\leqslant t(n)\varepsilon(n).
\end{aligned}$$

This upper bound on $\mathbf{P}[E]$, along with the above analysis of the error-free case, allows us to complete the proof of the desired claim, for all n large enough:

$$\begin{aligned}
&\mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(1)) = 1] - \mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(0)) = 1] \\
&\quad \geqslant \left(\mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(1)) = 1 \mid \overline{E}] - \mathop{\mathbf{P}}_{K,r}[\mathscr{A}(K_{\text{pub}}, e_{K,r}(0)) = 1 \mid \overline{E}] \right) \\
&\qquad\qquad - 2\,\mathbf{P}[E] \\
&\quad \geqslant \frac{2}{n^c} - 2\varepsilon(n) - 2t(n)\varepsilon(n) \\
&\quad \geqslant \frac{1}{n^c}.
\end{aligned}$$

□

REMARK 12.5 (Klivans and Sherstov [128]). The above proof assumes that, given consistent training examples, the learner is guaranteed to succeed in finding a hypothesis h that satisfies (12.1). This makes for a shorter and simpler analysis. In reality, we need only assume that the learner succeeds with probability $1/\mathsf{poly}(n)$, and outputs "FAIL" otherwise. To accommodate this more general setting, it suffices to have $\mathscr{A}$ output a random Boolean value whenever the learner fails.

12.5 Implementing uSVP-based decryption

The previous section demonstrated that if a public-key cryptosystem is secure, then no concept class that can implement its decryption function is efficiently PAC-learnable. In this section and the next, we will obtain implementations of the decryption functions of Section 12.3 by intersections of *quadratic threshold functions.* Naturally, this will lead to a hardness result for learning intersections of quadratic threshold functions. We will obtain the main result of this chapter by noting that intersections of quadratic threshold functions are no harder to learn than intersections of halfspaces, a claim we formalize next.

LEMMA 12.6 (Klivans and Sherstov [128]). *Assume that intersections of n^ε arbitrary* (*respectively, light*) *halfspaces on $\{0,1\}^n$ are weakly PAC-learnable. Then*

for any constant $c > 0$, intersections of n^c arbitrary (*respectively, light*) *quadratic threshold functions on $\{0,1\}^n$ are weakly PAC-learnable.*

PROOF. We will prove the "light" case only, the "arbitrary" case being closely analogous. Let $\mathscr{C}$ be the concept class of intersections of n^ε light halfspaces on $\{0,1\}^n$. Let $\mathscr{C}'$ be the concept class of intersections of n^ε light quadratic threshold functions on $\{0,1\}^n$. Finally, let $\mathscr{C}''$ be the concept class of intersections of n^c light quadratic threshold functions.

We first observe that a polynomial-time PAC-learning algorithm for $\mathscr{C}$ implies one for $\mathscr{C}'$. This is because every quadratic threshold function in the n variables $x_1, \ldots, x_n$ is a halfspace in the $n + \binom{n}{2}$ variables $x_1, \ldots, x_n$, $x_1x_2, x_1x_3, \ldots, x_{n-1}x_n$, which yields a polynomial-time map from the training and testing examples for a quadratic threshold function to those for a halfspace. This map is naturally viewed as a change of distribution in that a given distribution of $(x_1, \ldots, x_n)$ will induce another, non-uniform distribution in the $n + \binom{n}{2}$ new variables.

Finally, by a basic padding argument, the problem of PAC-learning the intersection of n^c halfspaces reduces to n^ε halfspaces for any constant $c > 0$. As a result, a polynomial-time learning algorithm for $\mathscr{C}'$ implies one for $\mathscr{C}''$. □

Recall that $\mathrm{frac}(a) = \min\{\lceil a \rceil - a, a - \lfloor a \rfloor\}$ denotes the distance from $a \in \mathbb{R}$ to the closest integer. Throughout this section, $\{a\}$ stands for the fractional part of $a \in \mathbb{R}$. Define the predicate NEAR-INT: $\mathbb{R} \to \{0,1\}$ by

$$\text{NEAR-INT}(a) = 1 \qquad \Leftrightarrow \qquad \mathrm{frac}(a) < \frac{1}{4}.$$

This predicate ignores the integral part of a, meaning that NEAR-INT(a) = NEAR-INT$(\{a\})$. Recall that the decryption function in the uSVP-based cryptosystem is $d_A(W) = \text{NEAR-INT}(AW)$, where A is a fixed real number and W is an integer input, both with a polynomial number of bits. We will demonstrate how to implement NEAR-INT(AW) with intersections of quadratic threshold functions. A critical ingredient of our implementation is the "interval trick" of Siu and Roychowdhury [209], an insightful idea that those authors used to obtain a depth-2 majority circuit for computing iterated addition.

LEMMA 12.7 (Klivans and Sherstov [128]). *Let $A > 0$ be a real number with k fractional bits. Let $f : \{0,1\}^n \to \{0,1\}$ be given by*

$$f(x) = \text{NEAR-INT}\left(A \sum_{j=0}^{n-1} x_{j+1} 2^j \right).$$

Then f can be computed by the intersection of k quadratic threshold functions with weight $O(k^4 4^k)$.

PROOF. Let $\{A\} = .b_1 b_2 \ldots b_k$ be the fractional part of A in binary, with $b_i \in \{0,1\}$ for all i. The integral part of A is irrelevant. Then

$$\left\{ A \sum_{j=0}^{n-1} x_{j+1} 2^j \right\} = \left\{ \sum_{i=1}^{k} \sum_{j=0}^{n-1} b_i x_{j+1} 2^{j-i} \right\} = \left\{ \sum_{i=1}^{k} \sum_{j=0}^{\min\{n-1, i-1\}} b_i x_{j+1} 2^{j-i} \right\},$$

where the last equation follows by dropping those terms $b_i x_{j+1} 2^{j-i}$ that are whole numbers. Define

$$S(x) = \sum_{i=1}^{k} \sum_{j=0}^{\min\{n-1, i-1\}} b_i x_{j+1} 2^{j-i}$$

so that $\{A \sum_{j=0}^{n-1} x_{j+1} 2^j\} = \{S(x)\}$. Observe that $S(x)$ is a multiple of 2^{-k} and ranges between 0 and k. We will use degree-2 PTFs to identify intervals in $[0, k]$ on which NEAR-INT$(S(x)) = 1$. A listing of the first few such intervals is as follows.

Value of $S(x)$ in binary	NEAR-INT($S(x)$)
. 0 0 0 0 … 0 0	
⋮	1
. 0 0 1 1 … 1 1	
. 0 1 0 0 … 0 0	
⋮	0
. 1 1 0 0 … 0 0	
. 1 1 0 0 … 0 1	
⋮	1
1 . 0 0 1 1 … 1 1	
1 . 0 1 0 0 … 0 0	
⋮	0
1 . 1 1 0 0 … 0 0	
1 . 1 1 0 0 … 0 1	
⋮	1
1 0 . 0 0 1 1 … 1 1	

The characteristic function of an interval $[a, b]$ is given by

$$\left(S(x) - \frac{a+b}{2}\right)^2 \leqslant \left(\frac{b-a}{2}\right)^2,$$

which has a representation as a quadratic threshold function with weight $O(k^4 4^k)$. To compute the negation of an interval, one replaces the inequality sign by a greater-than sign. Finally, there are at most $2k + 1$ intervals because every two consecutive intervals, starting at the second, cover a distance of 1 on the interval $[0, k]$. By forming the conjunction of the negations of the k intervals on which

NEAR-INT$(S(x)) = 0$, we obtain the desired representation of f as the intersection of k quadratic threshold functions with weight $O(k^4 4^k)$. □

12.6 Implementing SVP- and SIVP-based decryption

We will now show how to implement the decryption function of the other cryptosystem using quadratic threshold functions. Define the predicate NEAR-MID$_p$: $\mathbb{Z} \to \{0, 1\}$ by

$$\text{NEAR-MID}_p(a) = 1 \qquad \Leftrightarrow \qquad \left| b - \left\lfloor \frac{p}{2} \right\rfloor \right| \leqslant \min\{b, p - b\},$$

where $b \in \{0, 1, \ldots, p-1\}$ is the integer with $a \equiv b \pmod{p}$. In this notation, recall that the decryption function in the SVP- and SIVP-based cryptosystem is given by $d_{s_1,\ldots,s_n}(b, a_1, \ldots, a_n) = \text{NEAR-MID}_p(b - \sum a_i s_i)$, where all s_i, a_i, and b are integers in $\{0, \ldots, p-1\} = \mathbb{Z}_p$. We will show how to compute $d_{s_1,\ldots,s_n}$ with intersections of degree-2 PTFs.

LEMMA 12.8 (Klivans and Sherstov [128]). *Let* $d_{s_1,\ldots,s_n}: \left(\{0,1\}^{\log p}\right)^{n+1} \to \{0, 1\}$ *be the function defined by*

$$d_{s_1,\ldots,s_n}(x) = \text{NEAR-MID}_p\left(\sum_{i=0}^{\log p - 1} 2^i x_{0,i} - \sum_{j=1}^{n} s_j \sum_{i=0}^{\log p - 1} 2^i x_{j,i} \right), \tag{12.2}$$

where all s_i *are integers in* $\{0, \ldots, p-1\}$. *Then* $d_{s_1,\ldots,s_n}$ *can be computed by the intersection of* $n \log p$ *quadratic threshold functions with weight* $O(p^2 n^2 \log^2 p)$.

PROOF. Define

$$S(x) = \sum_{i=0}^{\log p - 1} 2^i x_{0,i} - \sum_{j=1}^{n} \sum_{i=0}^{\log p - 1} (2^i s_j \bmod p) x_{j,i}.$$

Thus, $S(x)$ is the original weighted sum in (12.2) with the coefficients reduced modulo p. It is clear that $d_{s_1,\ldots,s_n}(x) = \text{NEAR-MID}_p(S(x))$. The integer $S(x)$ ranges between $-(p-1)n \log p$ and $p-1$, an interval of length less than $pn \log p$. As in the proof of Lemma 12.7, this range can be divided into consecutive intervals on which $d_{s_1,\ldots,s_n}(x)$ is constant (i.e., does not change value within an interval). Every two consecutive intervals cover a length of p units. Thus, there are a total of $\leqslant 2(pn \log p)/p = 2n \log p$ consecutive intervals. By selecting the $n \log p$ intervals on which $d_{s_1,\ldots,s_n}(x) = 0$ and taking the conjunction of their negations, we can compute $d_{s_1,\ldots,s_n}$ exactly. It remains to note that the negation of an interval $[a, b]$ has characteristic function

$$\left(S(x) - \frac{a+b}{2}\right)^2 > \left(\frac{b-a}{2}\right)^2,$$

which can be represented by a quadratic threshold function with weight $O(p^2 n^2 \log^2 p)$. □

We additionally observe that the decryption function in the SVP- and SIVP-based cryptosystem can be computed by a depth-3 arithmetic circuit.

LEMMA 12.9 (Klivans and Sherstov [128]). *Let* $d_{s_1,\ldots,s_n} : \left(\{0,1\}^{\log p}\right)^{n+1} \to \{0,1\}$ *be the function defined by* (12.2)*, where all* s_i *are integers in* $\{0,\ldots,p-1\}$*. Then* $d_{s_1,\ldots,s_n}$ *can be computed by a depth-*3 *arithmetic circuit of size polynomial in* p *and* n*.*

PROOF. Set $S(x)$ as in the proof of Lemma 12.8. Then $S(x)$ is an integer in the range $R = [-(p-1)n \log p,\ p-1] \cap \mathbb{Z}$ and completely determines the value of $d_{s_1,\ldots,s_n}(x)$. Now, let g be a polynomial such that $g(S(x)) = d_{s_1,\ldots,s_n}(x)$ for all Boolean inputs x. It can be constructed by interpolating $d_{s_1,\ldots,s_n}$ on the range of $S(x)$ via the Lagrange formula:

$$g(y) = \sum_{r \in R} \text{NEAR-MID}_p(r) \prod_{\substack{r' \in R, \\ r' \neq r}} \frac{y - r'}{r - r'}.$$

Since the range R contains $\mathsf{poly}(p,n)$ integers, $g(S(x))$ can be computed by a depth-3 arithmetic circuit of size $\mathsf{poly}(p,n)$ with input $S(x)$ and summation gates at the bottom. But $S(x)$ is a sum of $\mathsf{poly}(p,n)$ terms, each a singleton variable x_i or a constant. Thus, $d_{s_1,\ldots,s_n}$ can be computed directly by a depth-3 arithmetic circuit of size $\mathsf{poly}(p,n)$ with inputs x. □

12.7 Hardness of learning intersections of halfspaces

Based on the assumed computational hardness of the cryptosystems in Section 12.3 and the reductions of Sections 12.4–12.6, we are in a position to prove the desired hardness results for learning intersections of halfspaces. The reader may find it helpful to review the definition of *light* polynomial threshold functions from Section 2.2.

THEOREM 12.10 (Klivans and Sherstov [128]). *Assume that intersections of n^ε halfspaces on $\{0,1\}^n$ are PAC-learnable for some constant $\varepsilon > 0$. Then there is a polynomial-time solution to $\tilde{O}(n^{1.5})$-uSVP.*

PROOF. Let $\mathscr{C}$ denote the concept class of intersections of n^ε halfspaces. Let $\mathscr{C}'$ denote the concept class of intersections of n^c quadratic threshold functions, for a sufficiently large constant $c > 0$. By Lemma 12.6, the assumed PAC-learnability of $\mathscr{C}$ implies the PAC-learnability of $\mathscr{C}'$. By Lemma 12.7, all the decryption functions in the uSVP-based cryptosystem are in $\mathscr{C}'$. A PAC-learning algorithm for $\mathscr{C}'$ would thus yield a distinguisher between the encryptions of 0 and 1 by Lemma 12.4 and hence an efficient solution to $O(\sqrt{n} \cdot \gamma(n))$-uSVP for $\gamma(n) = n \log n$ by Theorem 12.2. □

REMARK 12.11 (Klivans and Sherstov [128]). Oded Regev observed [182] that Theorem 12.10 is also valid for light halfspaces, rather than arbitrary ones as stated. To see this, note that in Regev's first cryptosystem [180, Lem. 5.2], except with probability exponentially small in n, the quantity $\mathrm{frac}(AW)$ is bounded away from $1/4$ by a small constant. Therefore, with extremely high probability, we can ignore many of the least significant bits of AW, as these bits can only change the value of AW by $o(1)$. This allows one to restrict the sum $S(x)$ in Lemma 12.7 to contain only the terms $b_i x_j 2^{j-i}$ with $j - i > -C \log n$, for a sufficiently large constant

$C > 0$. The integral representation of the resulting threshold function would have polynomial weight, leading to hardness for intersections of light halfspaces.

THEOREM 12.12 (Klivans and Sherstov [128]). *Assume that intersections of n^ε light halfspaces on $\{0,1\}^n$ are PAC-learnable for some constant $\varepsilon > 0$. Then there is a polynomial-time quantum solution to $\tilde{O}(n^{1.5})$-SVP and $\tilde{O}(n^{1.5})$-SIVP.*

PROOF. Let $\mathscr{C}$ denote the concept class of intersections of n^ε light halfspaces, and let $\mathscr{C}'$ denote the concept class of intersections of n^c light quadratic threshold functions. By Lemma 12.6, the assumed PAC-learnability of $\mathscr{C}$ implies the PAC-learnability of $\mathscr{C}'$. By Lemma 12.8, the decryption function in the uSVP-based cryptosystem is in $\mathscr{C}'$. A PAC-learning algorithm for $\mathscr{C}'$ would thus yield a distinguisher between the encryptions of 0 and 1 by Lemma 12.4 and, as a result, an efficient quantum solution to $\tilde{O}(n \cdot \gamma(n))$-SVP and $\tilde{O}(n \cdot \gamma(n))$-SIVP for $\gamma(n) = \sqrt{n} \log^2 n$ by Theorem 12.3. □

Theorems 12.10 and 12.12 both imply a hardness result for PAC learning polynomial-size depth-2 circuits of majority gates, a concept class commonly denoted by $\widehat{\mathsf{LT}}_2$. To prove this, we will need a result regarding light threshold circuits, due to Goldmann et al. [82] and Goldmann and Karpinski [83]. Let $\widehat{\mathsf{LT}}_d$ denote the class of depth-d polynomial-size circuits of threshold gates with polynomially-bounded weights. Let $\widetilde{\mathsf{LT}}_d$ denote the class of depth-d polynomial-size threshold circuits in which only the output gate is required to have polynomially-bounded weights.

THEOREM 12.13 (Bounded vs. unbounded weights [82, 83]). *For any constant d,*

$$\widehat{\mathsf{LT}}_d = \widetilde{\mathsf{LT}}_d.$$

We are now in a position to prove the desired hardness result for depth-2 majority circuits.

THEOREM 12.14 (Klivans and Sherstov [128]). *Assume that depth-2 polynomial-size circuits of majority gates are PAC learnable. Then there is a polynomial-time*

solution to $\tilde{O}(n^{1.5})$-uSVP and polynomial-time quantum solutions to $\tilde{O}(n^{1.5})$-SVP and $\tilde{O}(n^{1.5})$-SIVP.

PROOF. Let $\wedge\widehat{\mathsf{LT}}_1$ (respectively, $\wedge\mathsf{LT}_1$) denote the concept classes of intersections of polynomially many light (respectively, arbitrary) halfspaces. By Theorems 12.10 and 12.12, it suffices to show that $\wedge\widehat{\mathsf{LT}}_1 \subseteq \widehat{\mathsf{LT}}_2$ and $\wedge\mathsf{LT}_1 \subseteq \widehat{\mathsf{LT}}_2$. The first statement is clear: each light halfspace is already a majority gate, with the inputs suitably negated or replicated, and the top gate $\mathrm{AND}(f_1, f_2, \ldots, f_s)$ can be replaced by a majority gate $\mathrm{MAJ}(-s+1, f_1, f_2, \ldots, f_s)$. To prove that $\wedge\mathsf{LT}_1 \subseteq \widehat{\mathsf{LT}}_2$, observe that $\wedge\mathsf{LT}_1 \subseteq \widetilde{\mathsf{LT}}_2$ by an analogous argument and $\widetilde{\mathsf{LT}}_2 = \widehat{\mathsf{LT}}_2$ by Theorem 12.13. □

Independently of our work, Feldman et al. [68] obtained a result similar to Theorem 12.14. Specifically, those authors proved that a polynomial-time algorithm for learning depth-2 polynomial-size majority circuits would break the Ajtai-Dwork cryptosystem. In contrast, our work makes use of more recent cryptosystems due to Regev, whose security is based on weaker assumptions.

12.8 Hardness of learning arithmetic circuits and beyond

Several efficient, sparse polynomial interpolation algorithms are known in the case when the learner has query access to the unknown polynomial [151, 193, 129]. If, in addition to membership queries, the learner can make equivalence queries, Klivans and Shpilka [119] showed how to exactly learn restricted types of depth-3 arithmetic circuits via multiplicity automata techniques [34]. Here, we show that if the learner receives random examples only, then learning depth-3 polynomial-size arithmetic circuits is as hard as solving $\tilde{O}(n^{1.5})$-SVP in quantum polynomial-time.

THEOREM 12.15 (Klivans and Sherstov [128]). *Assume that depth-3 polynomial-size arithmetic circuits are PAC-learnable in polynomial time. Then there is a polynomial-time quantum solution to $\tilde{O}(n^{1.5})$-SVP and $\tilde{O}(n^{1.5})$-SIVP.*

PROOF. Invoke Lemma 12.9 and argue precisely as was done in the proofs of Theorems 12.10 and 12.12. □

A natural question to ask is whether our approach could yield hardness results for other concept classes. Particularly interesting candidates are AC^0 and, more ambitiously, polynomial-size DNF formulas. Here we prove that the decryption functions of Regev's cryptosystems contain PARITY as a subfunction and thus are not computable in AC^0. We start with the SVP- and SIVP-based cryptosystem, which admits an easier proof.

PROPOSITION 12.16 (Klivans and Sherstov [128]). *The decryption function of the SVP- and SIVP-based cryptosystem, defined by* $f_{s_1,\ldots,s_n}(a_1,\ldots,a_n,b) = \text{NEAR-MID}_p(b - \sum a_i s_i)$*, is not in* AC^0.

PROOF. Let $x_1, x_2, \ldots, x_n \in \{0,1\}$. Note that

$$\begin{aligned}\text{NEAR-MID}_p\left(\frac{p-1}{2}\sum_{i=1}^{n} x_i\right) &= \text{NEAR-MID}_p\left(\frac{p}{2}\sum_{i=1}^{n} x_i\right)\\ &= \text{PARITY}(x_1,\ldots,x_n).\end{aligned}$$

The first equality holds because $\frac{1}{2}\sum x_i \leqslant \frac{1}{2}n \ll p$. Thus, $\text{PARITY}(x_1,\ldots,x_n)$ is a subfunction of $\text{NEAR-MID}_p(b - \sum a_i s_i)$. Since AC^0 circuits cannot compute the PARITY function [75, 91], the claim follows. □

Recall now that the decryption function in the uSVP-based cryptosystem is $d_A(X) = \text{NEAR-INT}(AX)$, where A is a fixed real number and X is an integer input. For convenience, we assume that X has $n+1$ bits rather than n.

PROPOSITION 12.17 (Klivans and Sherstov [128]). *The decryption function of the uSVP-based cryptosystem,* $d_A(X) = \text{NEAR-INT}(AX)$*, is not in* AC^0.

PROOF. We will show that $d_A(X)$ computes the PARITY function on a subset of $\Theta(n/\log n)$ bits from among $x_1,\ldots,x_n$, with the other bits set to 0. The claim will follow. Let $\Delta = 3 + \log n$ and $A = \sum_{i=0}^{n/\Delta} 2^{-i\Delta-1}$. For convenience of notation, we assume that $\Delta \mid n$. In what follows, we show that $d_A(X) =$

PARITY$(x_0, x_\Delta, x_{2\Delta}, \ldots, x_n)$ when $x_i = 0$ for all $i \notin \{0, \Delta, 2\Delta, \ldots, n\}$. We have

$$\begin{aligned} d_A(X) &= \text{NEAR-INT}\,(AX) \\ &= \text{NEAR-INT}\left(\left(\sum_{i=0}^{n/\Delta} \frac{1}{2^{i\Delta+1}}\right)\left(\sum_{j=0}^{n/\Delta} x_{j\Delta} 2^{j\Delta}\right)\right) \\ &= \text{NEAR-INT}\left(\sum_i \sum_{j>i} \frac{x_{j\Delta} 2^{j\Delta}}{2^{i\Delta+1}} + \sum_i \frac{x_{i\Delta} 2^{i\Delta}}{2^{i\Delta+1}} + \sum_i \sum_{j<i} \frac{x_{j\Delta} 2^{j\Delta}}{2^{i\Delta+1}}\right). \end{aligned}$$

The first summation features only whole numbers and can thus be dropped. The second summation is precisely $\frac{1}{2}(x_0 + x_\Delta + \cdots + x_n)$, a multiple of $\frac{1}{2}$. The third summation does not exceed $\frac{1}{8}$, by the choice of Δ and the geometric series, and thus does not affect the result. Thus,

$$\begin{aligned} d_A(X) &= \text{NEAR-INT}\Big(\frac{x_0 + x_\Delta + \cdots + x_n}{2}\Big) \\ &= \text{PARITY}(x_0, x_\Delta, \ldots, x_n). \end{aligned}$$

□

Chapter 13

Lower Bounds for Statistical Query Learning

In the previous chapter, we proved representation-independent, cryptographic hardness results for learning intersections of halfspaces on $\{0,1\}^n$. Here, we complement those results with *unconditional* lower bounds for learning intersections of halfspaces in Kearns' well-studied statistical query model [104]. In particular, we prove that any statistical-query algorithm for learning the intersection of $\sqrt{n}$ majority functions on $\{0,1\}^n$ runs in time $\exp\{\Omega(\sqrt{n})\}$. This lower bound is essentially tight and is an exponential improvement on previous work. In addition, we derive near-tight, exponential lower bounds on the threshold density of this concept class, placing it beyond the scope of Jackson's influential Harmonic sieve algorithm [98].

13.1 Introduction

Recall from the previous chapter that learning intersections of halfspaces is a fundamental and well-studied problem in computational learning theory. In addition to generalizing well-known concept classes such as DNF formulas, intersections of halfspaces are capable of representing arbitrary convex sets. While efficient algorithms exist for PAC learning a single halfspace, the problem of learning the intersection of even two halfspaces remains an unresolved challenge in the area. In the previous chapter, we proved the first representation-independent, cryptographic hardness results for PAC learning intersections of halfspaces. Furthermore, the problem is known to be NP-hard for *proper* learning, where additional restrictions are placed on the learner's output hypothesis [10].

The hardness results surveyed above are *conditional* in that they depend on widely believed but unproven assumptions from cryptography or complexity theory. Here, we complement that line of work by proving lower bounds that are *unconditional* but valid only for a restriction of the PAC model. Specifically, we study the problem of learning intersections of halfspaces in Kearns' *statistical query* model [104], described in detail in Section 13.2. The statistical query model differs from the PAC model in that instead of individual labeled examples, the learner only receives statistical information about the unknown function. Efficient algorithms in this model are particularly useful because of their inherent robustness to noisy training data. Perhaps surprisingly, virtually all known PAC learning algorithms can be adapted to work efficiently in the statistical query model.

A tight measure of the learning complexity of a concept class $\mathscr{C}$ under a given distribution μ in the statistical query model is the statistical query dimension $\mathrm{sq}_\mu(\mathscr{C})$, introduced in Section 2.5 and already used in other contexts in this thesis. We focus on $\mathsf{MAJ}_{n,k}$, the concept class of conjunctions of up to k majority functions in n variables. Our main result, derived in Section 13.6, is as follows.

THEOREM 13.1 (Klivans and Sherstov [127]). *There are (explicitly given) distributions μ on $\{0,1\}^n$ such that*

$$\mathrm{sq}_\mu(\mathsf{MAJ}_{n,k}) = \begin{cases} \max\left\{n^{\Omega(k/\log\log n)}, n^{\Omega(\log k)}\right\} & \text{if } \ k \leqslant \log n, \\ n^{\Omega(k/\log k)} & \text{if } \ \log n \leqslant k \leqslant \sqrt{n}. \end{cases}$$

We show that Theorem 13.1 is essentially tight and is an exponential improvement on previous lower bounds [37]. An illustrative instantiation of our result is as follows: for any constant $0 < \varepsilon \leqslant 1/2$, the intersection of n^ε majority functions has statistical query dimension $\exp\{\Omega(n^\varepsilon)\}$, a known upper bound being $\exp\{O(n^\varepsilon \log^3 n)\}$.

In addition, we study the threshold density of intersections of k majority functions. As we discuss in Section 13.3, threshold weight and threshold density are useful parameters of a Boolean function from the standpoint of uniform-distribution learning. In particular, the celebrated algorithm due to Jackson [98] efficiently learns any concept class with small threshold weight. In Sections 13.4 and 13.5, we show that $\mathsf{MAJ}_{n,k}$ has high threshold weight and density and is thus not amenable to Jackson's techniques. More precisely, we prove the following statement.

THEOREM 13.2 (Klivans and Sherstov [127]). *Let k be an integer, $2 \leqslant k \leqslant \sqrt{n}$. Then there is a function $f \in \mathsf{MAJ}_{n,k}$ with*

$$\mathrm{dns}(f) \geqslant n^{\Omega(k/\log k)} n^{\Omega(\log n/\log\log n)}.$$

The lower bound in Theorem 13.2 is essentially optimal, in light of earlier work by Klivans et al. [121].

13.2 Learning via statistical queries

The *statistical query* model, introduced by Kearns [104], is an elegant restriction of the PAC model. In both cases, one considers a family $\mathscr{C}$ of Boolean functions $X \to \{-1, +1\}$ defined on some finite set X. In both cases, one fixes a distribution μ on X as well as a function $f \in \mathscr{C}$. The learner, who does not know μ or f, is faced with the problem of constructing a hypothesis $h: X \to \{-1, +1\}$ that closely approximates f with respect to μ, in the sense that $\mathbf{E}_\mu[f(x) \neq h(x)] < \varepsilon$ for a small constant ε. The difference between the PAC model and the SQ model resides in the kind of information that the learner receives about the unknown function f. In the PAC model, the learner has access to individual labeled examples $(x^{(1)}, f(x^{(1)})), \ldots, (x^{(m)}, f(x^{(m)}))$, where $x^{(1)}, \ldots, x^{(m)} \in X$ are independent and distributed identically according to μ. In the SQ model, the learner no longer sees individual labeled examples. Instead, the learner may posit any polynomial-time computable function $\chi: \{0,1\}^n \times \{-1, +1\} \to \{-1, +1\}$, called a *statistic*, and receive the quantity

$$\mathop{\mathbf{E}}_{x \sim \mu}[\chi(x, f(x))], \tag{13.1}$$

up to a small additive error called the *tolerance*, τ.

In light of the Chernoff bound, the PAC learner can simulate any statistical query by drawing $\Theta(\log \frac{1}{\tau})$ labeled examples and estimating the expectation (13.1). In other words, any SQ algorithm can be simulated by a PAC algorithm, and hence the SQ model can be viewed as a restriction of the PAC model. A compelling advantage of SQ algorithms is that they admit simulations even in the PAC model with *random classification noise*, when the individual labeled examples each have a small independent probability of getting the wrong label. From that standpoint, having an efficient SQ learner for a concept class is preferable to having an efficient PAC learner. It is a remarkable fact, however, that a vast majority of efficient PAC learning algorithms designed to date have efficient counterparts in the SQ model.

In this chapter, we will mostly be concerned with lower bounds for SQ learning. Such lower bounds can be proved in terms of the statistical query dimension of the concept class, a versatile complexity measure introduced in Section 2.5 and already used in other contexts in this thesis. We have:

THEOREM 13.3 (Yang [222, Cor. 1]). *Let $\mathscr{C}$ be a concept class and μ a probability distribution such that $\mathrm{sq}_\mu(\mathscr{C}) = N$. Then if all queries are made with tolerance at least $N^{-1/3}$, at least*

$$\frac{N^{1/3}}{2} - 1$$

queries are required to learn $\mathscr{C}$ to error $1/2 - 1/(2N^{1/3})$ with respect to μ in the statistical query model.

In other words, concept classes with high statistical query dimension require many statistical queries even when one only seeks to achieve classification accuracy slightly above random guessing. Theorem 13.3 follows up and improves on an earlier result due to Blum et al. [37], who additionally proved matching upper bounds on the number of statistical queries in terms of the statistical query dimension. In this sense, the statistical query dimension $\mathrm{sq}_\mu(\mathscr{C})$ is a tight measure of the complexity of learning $\mathscr{C}$ with respect to μ in the statistical query model. We refer the reader to the textbook by Kearns and Vazirani [108] for further results on the SQ model.

13.3 Threshold weight and density in learning

The threshold weight and density of Boolean functions, reviewed in Section 2.2, are useful quantities from the standpoint of learning with respect to the uniform distribution. Key to this application is the elegant algorithm of Kushilevitz and Mansour [136], which efficiently identifies the dominant Fourier coefficients of a function.

THEOREM 13.4 (Kushilevitz and Mansour [136]). *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a given Boolean function. Let $\varepsilon > 0$ and $\delta > 0$ be given parameters. There is an explicit algorithm that runs in time polynomial in $(n/\varepsilon)\log(1/\delta)$ and outputs, with probability at least $1-\delta$, every set $S \subseteq \{1,2,\ldots,n\}$ for which $|\hat{f}(S)| \geqslant \varepsilon$. The algorithm only uses query access to f.*

The *Harmonic sieve*, a celebrated algorithm due to Jackson [98], uses Theorem 13.4 to learn the concept class $\mathscr{C}$ of functions $f:\{0,1\}^n \to \{-1,+1\}$ with $W(f) \leqslant W$ to any accuracy $\varepsilon > 0$ with respect to the uniform distribution, in time polynomial in nW/ε. In particular, Jackson's algorithm learns the concept class of polynomial-size DNF formulas under the uniform distribution, using membership queries, in polynomial time.

In light of the above, threshold weight can be viewed as a criterion for *strong* learnability with respect to the uniform distribution. Threshold density, on the other hand, can be viewed as a criterion for *weak* learnability with respect to the uniform distribution. This observation is based on the following result of Bruck [44].

THEOREM 13.5 (Bruck [44]). *Let $f:\{0,1\}^n \to \{-1,+1\}$ be a given Boolean function. Then for all d,*

$$\mathrm{dns}(f,d) \geqslant \frac{1}{\max_{|S|\leqslant d} |\hat{f}(S)|}.$$

In particular,

$$\mathrm{dns}(f) \geqslant \frac{1}{\|f\|_\infty}.$$

As an application to weak learning, we have:

PROPOSITION 13.6 (Klivans and Sherstov [127]). *Let $\mathscr{C}$ be a family of Boolean functions $f:\{0,1\}^n \to \{-1,+1\}$. Put $L = \max_{f\in\mathscr{C}} \mathrm{dns}(f)$. Then $\mathscr{C}$ is learnable to accuracy $\frac{1}{2} + \frac{1}{2L}$ under the uniform distribution in time polynomial in n and L, using membership queries.*

PROOF. Let $f \in \mathscr{C}$ be the unknown target function. In time polynomial in nL, the algorithm of Theorem 13.4 identifies all characters $\chi_S:\{0,1\}^n \to \{-1,+1\}$ with correlation $1/L$ or more with f. It thus suffices to show that $\|\hat{f}\|_\infty \geqslant 1/L$. That, in turn, is immediate from Theorem 13.5 and the fact that $\mathrm{dns}(f) \leqslant L$. □

Key to our analysis of threshold density in this chapter is the family of *bent* Boolean functions $f: \{0,1\}^n \to \{-1,+1\}$, defined by the property that $|\hat{f}(S)| = 2^{-n/2}$ for all S. In light of Parseval's identity (2.1), bent functions are precisely those Boolean functions f that minimize $\|\hat{f}\|_\infty$. Two important bent functions are the inner product function $\mathrm{IP}_n: \{0,1\}^{2n} \to \{-1,+1\}$, defined by

$$\mathrm{IP}_n(x) = \bigoplus_{i=1}^{n} (x_{2i-1} \wedge x_{2i}),$$

and the complete quadratic function $\mathrm{CQ}_n: \{0,1\}^n \to \{-1,+1\}$, defined for even n by

$$\mathrm{CQ}_n(x) = \bigoplus_{1 \leqslant i < j \leqslant n} (x_i \wedge x_j).$$

It is known [44] and easy to see that

$$\mathrm{CQ}_n(x) = \begin{cases} 1, & \text{if } |x_i| \equiv 0 \pmod 4, \\ 1, & \text{if } |x_i| \equiv 1 \pmod 4, \\ -1, & \text{otherwise.} \end{cases}$$

We have:

THEOREM 13.7 (Bruck [44]). *The function* IP_n *is bent for all* n. *The function* CQ_n *is bent for all even* n.

Our analysis of threshold density requires the following observation.

LEMMA 13.8 (Klivans and Sherstov [127]). *Let* $f: \{0,1\}^n \to \{-1,+1\}$ *be a given function. Define* $F: \{0,1\}^m \to \{-1,+1\}$ *by* $F(x) = f(\chi_1, \ldots, \chi_n)$, *where each* χ_i *is a parity function on* $\{0,1\}^m$ *or the negation of a parity function on* $\{0,1\}^m$. *Then*

$$\mathrm{dns}(F) \leqslant \mathrm{dns}(f).$$

PROOF. Immediate from the definition of threshold density and the fact that the product of characters is another character. □

Given a function $f : \{0,1\}^n \to \{-1,+1\}$, recall that a *reflection* of f is any function of the form $g(x) = f(x \oplus y)$ for some fixed $y \in \{0,1\}^n$. The lower bounds in this chapter target be the concept class $\mathsf{MAJ}_{n,k}$, defined as the set of all functions $f : \{0,1\}^n \to \{-1,+1\}$ expressible as

$$f(x) = \bigwedge_{i=1}^{k} \operatorname{sgn}\left(\sum_{j=1}^{n} \sigma_{ij}(-1)^{x_i}\right)$$

for some coefficients $\sigma_{ij} \in \{0,1\}$. In other words, $\mathsf{MAJ}_{n,k}$ is the concept class of intersections of k majority functions.

13.4 Threshold density of the intersection of two majorities

We start our threshold density analysis with the intersection of two majority functions. An essential ingredient in our proof is the following result of O'Donnell and Servedio [163, Thm. 17].

THEOREM 13.9 (O'Donnell and Servedio [163]). *Consider the Boolean function* $f(x,y) = \mathrm{MAJ}(x_1,\ldots,x_n) \wedge \mathrm{MAJ}(y_1,\ldots,y_n)$. *Then*

$$\deg_{\pm}(f) = \Omega\left(\frac{\log n}{\log\log n}\right).$$

Krause and Pudlák [132, Prop. 2.1] discovered an elegant procedure for converting Boolean functions with high threshold degree into Boolean functions with high threshold density. Their construction maps a given function $f : \{0,1\}^n \to \{-1,+1\}$ to the function $f^{\mathrm{KP}} : (\{0,1\}^n)^3 \to \{-1,+1\}$ given by

$$f^{\mathrm{KP}}(x,y,z) = f(\ldots,(\overline{z_i} \wedge x_i) \vee (z_i \wedge y_i),\ldots).$$

We have:

THEOREM 13.10 (Krause and Pudlák [132]). *For every* $f : \{0,1\}^n \to \{-1,+1\}$,

$$\operatorname{dns}(f^{\mathrm{KP}}) \geqslant 2^{\deg_\pm(f)}.$$

A final ingredient in our proof is a procedure for amplifying the threshold degree of a given function by composition with an XOR function.

LEMMA 13.11 (Klivans and Sherstov [127]). *Let* $f : \{0,1\}^n \to \{-1,+1\}$ *be a given function. Define* $F : \{0,1\}^{nk} \to \{-1,+1\}$ *by*

$$F(x) = f\left(\ldots, \bigoplus_{j=1}^{k} x_{ij}, \ldots\right).$$

Then

$$\deg_\pm(F) = k \deg_\pm(f).$$

PROOF. Our proof draws inspiration from an earlier result due to O'Donnell and Servedio [163, Thm. 13]. The upper bound on the threshold degree of F is immediate. To prove the lower bound, let $d = \deg_\pm(f)$ and note that Theorem 4.6 provides a distribution μ on $\{0,1\}^n$ with the property that $\widehat{\mu f}(S) = 0$ whenever $|S| < d$. Letting λ be the distribution on $\{0,1\}^{kn}$ given by

$$\lambda(x) = c\,\mu\left(\ldots, \bigoplus_{j=1}^{n} x_{ij}, \ldots\right),$$

where $c > 0$ is a normalizing constant, it is straightforward to verify that $\widehat{\lambda F}(S) = 0$ whenever $|S| < kd$. By Theorem 4.6, the properties of λ force the lower bound $\deg_\pm(F) \geqslant kd$. □

We are now in a position to prove the main result of this section, a superpolynomial lower bound on the threshold density of the intersection of two majorities.

THEOREM 13.12 (Klivans and Sherstov [127]). *The function* $\mathrm{MAJ}(x_1, \ldots, x_n) \wedge \mathrm{MAJ}(y_1, \ldots, y_n)$ *has threshold density* $n^{\Omega(\log n / \log\log n)}$.

PROOF. Let t and k be integers to be fixed later. Define $f : \{0,1\}^{2t} \to \{-1,+1\}$ by $f(x) = \mathrm{MAJ}(x_1, \ldots, x_t) \wedge \mathrm{MAJ}(x_{t+1}, \ldots, x_{2t})$. Consider the function $f^{\oplus} : (\{0,1\}^k)^{2t} \to \{-1,+1\}$ given by

$$f^{\oplus}(x) = f\left(\ldots, \bigoplus_{j=1}^{k} x_{i,j}, \ldots\right).$$

Lemma 13.11 implies that

$$\deg_{\pm}(f^{\oplus}) = k \deg_{\pm}(f).$$

Consider now the function $f^{\oplus \mathrm{KP}}$. For bits $a, b, c \in \{0,1\}$, we have

$$(\overline{c} \wedge a) \vee (c \wedge b) = \frac{1 + (-1)^c}{2} \cdot (-1)^a + \frac{1 - (-1)^c}{2} \cdot (-1)^b.$$

As a result,

$$f^{\oplus \mathrm{KP}}(x, y, z) \equiv \left(\prod_{i=1}^{k} q_{1,i} + \cdots + \prod_{i=1}^{k} q_{t,i} \geqslant 0\right) \bigwedge \left(\prod_{i=1}^{k} q_{t+1,i} + \cdots + \prod_{i=1}^{k} q_{2t,i} \geqslant 0\right),$$

where $q_{i,j} = (1 + (-1)^{z_{i,j}})(-1)^{x_{i,j}} + (1 - (-1)^{z_{i,j}})(-1)^{y_{i,j}}$.

The above construction shows that $f^{\oplus \mathrm{KP}}$ is computed by the intersection of two functions with threshold weight at most $2t4^k + 1$ each. Lemma 13.8 implies

that if the intersection of two majorities, each on a distinct set of $2t4^k + 1$ variables, has threshold density at most L, then $\mathrm{dns}(f^{\oplus \mathrm{KP}}) \leqslant L$. Theorem 13.10, on the other hand, implies that $f^{\oplus \mathrm{KP}}$ has threshold density at least $2^{\deg_\pm(f^\oplus)} = 2^{k \deg_\pm(f)}$. In view of Theorem 13.9 we conclude that the intersection of two majorities, each on $2t4^k + 1$ variables, has threshold density $\exp\{\Omega(k \log t / \log\log t)\}$. The theorem follows by setting $t = \lfloor \sqrt{n}/3 \rfloor$ and $k = \lfloor \frac{1}{4} \log n \rfloor$. □

Using rational approximation techniques, Beigel et al. [33] showed among other things that the function $f(x, y) = \mathrm{MAJ}(x_1, \ldots, x_n) \wedge \mathrm{MAJ}(y_1, \ldots, y_n)$ has threshold density $n^{O(\log n)}$. Thus, the lower of Theorem 13.12 is nearly tight. Several chapters later, in Theorem 16.33, we will obtain a tight lower bound of $n^{\Omega(\log n)}$ using additional techniques. We also note that a weaker lower bound of $n^{\omega(1)}$ on the threshold density of f can be obtained by using, in place of Theorem 13.9, the simpler result of Minsky and Papert [153] that $\deg_\pm(f) = \omega(1)$. That would suffice to show that the intersection of even two majorities has superpolynomial threshold density.

13.5 Threshold density of the intersection of $\omega(1)$ majorities

We continue our study of threshold density with intersections of $k = \omega(1)$ majorities. For small k, the density lower bounds in this section are somewhat weaker than those obtained in the previous section for $k = 2$. However, the proofs below rely solely on the fundamental Theorem 13.5 and are thus considerably simpler. Furthermore, it is the constructions in this section that will allow us to prove our sought SQ lower bounds.

Our first construction is based on a reduction to the inner product function.

LEMMA 13.13 (Klivans and Sherstov [127]). *Let $k \leqslant 2^{n^{o(1)}}$. Let $\alpha > 0$ be a sufficiently small absolute constant. Fix an integer $m \leqslant \alpha \log n \cdot \log k$. Then there are explicitly given functions $\chi_1, \chi_2, \ldots, \chi_n$, each a parity or the negation of a parity on $\{0,1\}^{2m}$, such that for each fixed $y \in \{0,1\}^{2m}$ one has*

$$\mathrm{IP}_m(x \oplus y) \equiv f_y(\chi_1, \chi_2, \ldots, \chi_n)$$

for some $f_y \in \mathsf{MAJ}_{n,k}$.

PROOF. Let $g_1, g_2, \ldots, g_{\log k}$ be copies of the inner product function, each on a disjoint set of variables V_i with $|V_i| = 2v$ for some $v = v(n,k)$ to be chosen later. Thus, $g = \bigoplus g_i$ is an inner product function on $2v \log k$ variables. At the same time, g is computable by the conjunction of $2^{\log k - 1} < k$ functions, each of the form $h_1 \vee h_2 \vee \cdots \vee h_{\log k}$, where $h_i \in \{g_i, \neg g_i\}$. Each conjunction $h_1 \vee h_2 \vee \cdots \vee h_{\log k}$ is logically equivalent to

$$h_1 + h_2 + \cdots + h_{\log k} \geqslant 1 - \log k,$$

which is in turn equivalent to

$$2^v h_1 + 2^v h_2 + \cdots + 2^v h_{\log k} \geqslant 2^v (1 - \log k). \tag{13.2}$$

Every h_i is a bent function on the $2v$ variables V_i, and thus $2^v h_i$ is simply the sum of the 4^v parities on V_i, each with a plus or a minus sign.

Create a new set of variables $U = \{\chi_1, \chi_2, \ldots\}$ as follows. U will contain a distinct variable for each parity on V_i (for each $i = 1, 2, \ldots, \log k$) and one for its negation. In addition, U will contain $2^v(\log k - 1) < 2^v \log k$ variables, each of which corresponds to the constant 1. As a result, each of the k polynomial threshold functions of the form (13.2) is a majority function in terms of U. Therefore, the inner product function on $2v \log k$ variables is computable by $f(\chi_1, \chi_2, \ldots)$ for some $f \in \mathsf{MAJ}_{|U|,k}$. Furthermore, for every fixed $y \in \{-1, +1\}^{2v \log k}$, the reflection $\mathrm{IP}_{v \log k}(x \oplus y)$ is computable by $f_y(\chi_1, \chi_2, \ldots)$ for some $f_y \in \mathsf{MAJ}_{|U|,k}$ This is because for each parity, $U = \{\chi_1, \chi_2, \ldots\}$ additionally contains its negation.

It remains to show that $|U| \leqslant n$. Setting $v = \frac{1}{2}(\log n - \log\log k - 2)$ yields $|U| = 2 \cdot 4^v \log k + 2^v \log k \leqslant n$. Thus, for $k \leqslant 2^{n^{o(1)}}$ the above construction computes inner product on the claimed number of variables:

$$2v \log k = (\log n - \log\log k - 2)\log k = \Omega(\log n \cdot \log k). \qquad \square$$

A consequence of Lemma 13.13 for threshold density is as follows.

THEOREM 13.14 (Klivans and Sherstov [127]). *Let* $k \leqslant 2^{n^{o(1)}}$. *Then there is a function* $f \in \mathsf{MAJ}_{n,k}$ *with*

$$\mathrm{dns}(f) \geqslant n^{\Omega(\log k)}.$$

PROOF. Let $k \leqslant 2^{n^{o(1)}}$. By Lemma 13.13, there is a function $f \in \mathsf{MAJ}_{n,k}$ and a choice of parities or negated parities $\chi_1, \ldots, \chi_n$ such that $f(\chi_1, \ldots, \chi_n)$ computes the inner product function on $2m = \Omega(\log n \cdot \log k)$ variables. Since all the Fourier coefficients of $f(\chi_1, \ldots, \chi_n)$ are 2^{-m} in absolute value, it follows from Theorem 13.5 that $f(\chi_1, \ldots, \chi_n)$ has threshold density at least $2^m = n^{\Omega(\log k)}$. By Lemma 13.8, the same lower bound holds for the function $f(x_1, \ldots, x_n)$. □

We now prove a different lower bound on the threshold density of intersection of k majorities. For $k \gg \log\log n$, the new bound will be significantly stronger than that of Theorem 13.14. The new construction is based on the complete quadratic function.

LEMMA 13.15 (Klivans and Sherstov [127]). *Let* $k \leqslant \sqrt{n}$. *Let* $\alpha > 0$ *be a sufficiently small absolute constant. Fix an integer*

$$m \leqslant \alpha \min\left\{\frac{k\log n}{\log\log n}, \frac{k\log n}{\log k}\right\}.$$

Then there are explicitly given functions $\chi_1, \chi_2, \ldots, \chi_n$, *each a parity or the negation of a parity on* $\{0,1\}^m$, *such that for every* $y \in \{0,1\}^m$ *one has*

$$\mathrm{CQ}(x \oplus y) = f_y(\chi_1, \chi_2, \ldots, \chi_n)$$

for some $f_y \in \mathsf{MAJ}_{n,k}$.

PROOF. Consider the function CQ_m. Since CQ_m depends only on the sum of the input bits, we have

$$\mathrm{CQ}(x) = 1 \qquad \Longleftrightarrow \qquad \bigwedge_{s \in S} \left(\sum (-1)^{x_i} \neq s \right),$$

where $S \subseteq \{-m, \ldots, 0, \ldots, m\}$ and $|S| \leqslant m$. The conjunction of $t = |S|/k$ of these predicates can be expressed as

$$\left(\sum (-1)^{x_i} - s_1 \right)^2 \left(\sum (-1)^{x_i} - s_2 \right)^2 \cdots \left(\sum (-1)^{x_i} - s_t \right)^2 > 0, \tag{13.3}$$

where $s_1, \ldots, s_t \in S$.

Consider the inequality $(\sum (-1)^{x_i} + m)^{2t} > 0$. Multiplying out the left member yields the sum of exactly $(2m)^{2t}$ parities, not all distinct. Construct a set of variables $U = \{\chi_1, \chi_2, \ldots\}$ to contain a variable for each of these $(2m)^{2t}$ parities and their negations. Over U, the function $(\sum (-1)^{x_i} + m)^{2t} > 0$ is a majority function. In fact, any Boolean function of the form (13.3) is a majority over U. Hence, $\mathrm{CQ}_m(x)$ is computable by $f(\chi_1, \chi_2, \ldots)$ for some $f \in \mathsf{MAJ}_{|U|,k}$. Furthermore, for every fixed $y \in \{0,1\}^m$, the reflection $\mathrm{CQ}_m(x \oplus y)$ is computable by $f_y(\chi_1, \chi_2, \ldots)$ for some $f_y \in \mathsf{MAJ}_{|U|,k}$. This is because for each parity, $U = \{\chi_1, \chi_2, \ldots\}$ additionally contains its negation. Finally, it is straightforward to check that $|U| \leqslant n$. □

As an application to threshold density, we obtain:

THEOREM 13.16 (Klivans and Sherstov [127]). *Let $k \leqslant \sqrt{n}$. Then there exists $f \in \mathsf{MAJ}_{n,k}$ with*

$$\mathrm{dns}(f) \geqslant \min \left\{ n^{\Omega(k/\log\log n)}, n^{\Omega(k/\log k)} \right\}.$$

PROOF. Let $k \leqslant \sqrt{n}$. By Lemma 13.15, there is a function $f \in \mathsf{MAJ}_{n,k}$ and a choice of parities or negations of parities $\chi_1, \ldots, \chi_n$ such that $f(\chi_1, \ldots, \chi_n)$

computes CQ_m on $m = \min\{\Omega(k \log n / \log\log n), \Omega(k \log n / \log k)\}$ variables. Since the Fourier coefficients of $f(\chi_1, \ldots, \chi_n)$ are all $2^{-m/2}$ in absolute value, Theorem 13.5 implies that $f(\chi_1, \ldots, \chi_n)$ has threshold density at least $2^{m/2}$. By Lemma 13.8, the same lower bound is valid for $f(x_1, \ldots, x_n)$. □

Combining the results for far, we obtain our main lower bound on the threshold density of intersections of majorities.

THEOREM 13.2 (Klivans and Sherstov [127], restated). *Let k be an integer, $2 \leqslant k \leqslant \sqrt{n}$. Then there is a function $f \in \mathsf{MAJ}_{n,k}$ with*

$$\mathrm{dns}(f) \geqslant n^{\Omega(k/\log k)} n^{\Omega(\log n/\log\log n)}.$$

PROOF. Immediate from Theorems 13.12 and 13.16. □

The lower bound in Theorem 13.2 is essentially optimal. In particular, Klivans et al. [121, Thm. 29] showed in earlier work that every function in $\mathsf{MAJ}_{n,k}$ has threshold density $n^{O(k \cdot \log k \cdot \log n)}$.

13.6 Statistical query dimension of intersections of majorities

Recall that the statistical query dimension tightly characterizes the weak learnability of a concept class in the statistical query model. In this section, we will explicitly construct distributions under which intersections of n^ε majorities, for any constant $0 < \varepsilon \leqslant 1/2$, have statistical query dimension $\exp\{\Omega(n^\varepsilon)\}$. This lower bound is essentially tight and is an exponential improvement on the previous construction [37], which was based on computing parity functions by intersections of halfspaces.

Throughout this section, $\mathscr{U}$ stands for the uniform distribution on $\{0,1\}^n$. An important ingredient in our result is the observation that any two distinct reflections of a bent function are orthogonal with respect to the uniform distribution. This fact is well-known in coding theory. For completeness, we include a proof.

LEMMA 13.17 (cf. Macwilliams and Sloane [149, p. 427]). *Let* $f : \{0,1\}^n \to \{-1,+1\}$ *be a bent function. Then for any distinct* $y, y' \in \{0,1\}^n$,

$$\mathop{\mathbf{E}}_{x \sim \mathscr{U}}[f(x \oplus y) f(x \oplus y')] = 0.$$

PROOF. By hypothesis, $y \oplus y' \neq 0$. Thus,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x \sim \mathscr{U}}[f(x \oplus y) f(x \oplus y')] \\
&\qquad = \mathbf{E}\left[\left(\sum_S \hat{f}(S)\chi_S(x)\chi_S(y)\right)\left(\sum_T \hat{f}(T)\chi_T(x)\chi_T(y')\right)\right] \\
&\qquad = \sum_S \sum_T \hat{f}(S)\hat{f}(T)\chi_S(y)\chi_T(y')\,\mathbf{E}[\chi_S(x)\chi_T(x)] \\
&\qquad = \sum_S \hat{f}(S)^2 \chi_S(y)\chi_S(y') \\
&\qquad = 2^{-n} \sum_S \chi_S(y \oplus y') \\
&\qquad = 0.
\end{aligned}$$

The last equality holds because on every $z \in \{0,1\}^n \setminus \{0\}$, exactly half of the characters evaluate to -1 and the other half, to 1. □

The following is a simple consequence of Lemma 13.17.

THEOREM 13.18 (Klivans and Sherstov [127]). *Let* $\mathscr{C}$ *denote the concept class of bent functions on* n *variables. Then*

$$\mathrm{sq}_{\mathscr{U}}(\mathscr{C}) = 2^n.$$

PROOF. Fix a bent function f and consider its 2^n reflections, themselves bent functions. By Lemma 13.17, any two of them are orthogonal. □

Consider a function $h\colon \{0,1\}^n \to \{0,1\}^n$. The *$h$-induced distribution on $\{0,1\}^n$*, denoted by $h \circ \mathscr{U}$, is the distribution given by

$$(h \circ \mathscr{U})(z) = \mathop{\mathbf{P}}_{x \sim \mathscr{U}}[h(x) = z]$$

for any $z \in \{0,1\}^n$. Put differently, $h \circ \mathscr{U}$ is the uniform distribution over the multiset $h(\{0,1\}^n)$.

PROPOSITION 13.19 (Klivans and Sherstov [127]). *Let $f, g\colon \{0,1\}^n \to \{-1,+1\}$ and $h\colon \{0,1\}^n \to \{0,1\}^n$ be arbitrary functions. Then*

$$\mathop{\mathbf{E}}_{x \sim h \circ \mathscr{U}}[f(x)g(x)] = \mathop{\mathbf{E}}_{x \sim \mathscr{U}}[f(h(x))g(h(x))].$$

PROOF. By definition of $h \circ \mathscr{U}$, picking a random input according to $h \circ \mathscr{U}$ is equivalent to picking $x \in \{0,1\}^n$ uniformly at random and outputting $h(x)$. □

We are ready to prove our claimed lower bound on the statistical query dimension of $\mathsf{MAJ}_{n,k}$.

THEOREM 13.1 (Klivans and Sherstov [127], restated). *There are (explicitly given) distributions μ on $\{0,1\}^n$ such that*

$$\mathrm{sq}_\mu(\mathsf{MAJ}_{n,k}) = \begin{cases} \max\left\{n^{\Omega(k/\log\log n)}, n^{\Omega(\log k)}\right\} & \text{if } k \leqslant \log n, \\ n^{\Omega(k/\log k)} & \text{if } \log n \leqslant k \leqslant \sqrt{n}. \end{cases}$$

PROOF. Let $k \leqslant \log n$. Fix an integer $m = \Omega(\log n \cdot \log k)$ and parity functions $\chi_1, \chi_2, \ldots, \chi_n$ as in Lemma 13.13. Then there is a family $\mathscr{F} = \{f_1, f_2, \ldots, f_{4^m}\} \subset \mathsf{MAJ}_{n,k}$ such that every reflection $\mathrm{IP}_m(x \oplus y)$ can be represented as $f_i(\chi_1(x), \chi_2(x), \ldots, \chi_n(x))$ for some i.

Define $h\colon \{0,1\}^n \to \{0,1\}^n$ by $h(x) = (\chi_1(x), \chi_2(x), \ldots, \chi_n(x))$. Then for every two distinct $f_i, f_j \in \mathscr{F}$, we have

$$\mathop{\mathbf{E}}_{x\sim h\circ\mathscr{U}}[f_i(x)\,f_j(x)] = \mathop{\mathbf{E}}_{x\sim\mathscr{U}}[f_i(\chi_1(x),\ldots,\chi_n(x))\,f_j(\chi_1(x),\ldots,\chi_n(x))]$$

by Proposition 13.19 and therefore further

$$\mathop{\mathbf{E}}_{x\sim h\circ\mathscr{U}}[f_i(x)\,f_j(x)] = 0$$

by Lemma 13.17. In words, every pair of functions in $\mathscr{F}$ are orthogonal under the distribution $h \circ \mathscr{U}$. Therefore, $\mathrm{sq}_{h\circ\mathscr{U}}(\mathsf{MAJ}_{n,k}) \geqslant |\mathscr{F}| = 4^m = n^{\Omega(\log k)}$ for $k \leqslant \log n$. Moreover, the distribution $h \circ \mathscr{U}$ has an explicit description: pick a random $x \in \{0,1\}^n$ and return the n-bit string $(\chi_1(x), \ldots, \chi_n(x))$, where $\chi_1, \ldots, \chi_n$ are the explicitly given parity functions from Lemma 13.13. Applying an analogous argument to Lemma 13.15 yields the alternate lower bound $\mathrm{sq}(\mathsf{MAJ}_{n,k}) = \min\{n^{\Omega(k/\log k)}, n^{\Omega(k/\log\log n)}\}$ for $k \leqslant \sqrt{n}$. □

Our lower bounds on the statistical query dimension of intersections of majority functions are essentially tight, as we now observe.

THEOREM 13.20 (Klivans and Sherstov [127]). *For every distribution μ on $\{0,1\}^n$,*

$$\mathrm{sq}_\mu(\mathsf{MAJ}_{n,k}) \leqslant n^{O(k\cdot\log k\cdot\log n)}.$$

PROOF. Klivans et al. [121, Thm. 29] show that every $f \in \mathsf{MAJ}_{n,k}$ has threshold degree $d = O(k \cdot \log k \cdot \log n)$. Thus, every $f \in \mathsf{MAJ}_{n,k}$ is a halfspace in terms of the parity functions of degree at most d. It follows that the statistical query dimension of $\mathsf{MAJ}_{n,k}$ is at most the statistical query dimension of halfspaces in $\sum_{i=0}^{d}\binom{n}{i} \leqslant n^{O(k\cdot\log k\cdot\log n)}$ dimensions. It remains to recall Corollary 10.12, which states that the statistical query dimension of halfspaces on $\{0,1\}^D$ is $2(D+1)^2$ or less under every distribution. □

We conclude this section with an application of our results to sign-rank.

THEOREM 13.21 (Sherstov). *The concept class* $\mathsf{MAJ}_{n,k}$ *satisfies*

$$\mathrm{rk}_{\pm}(\mathsf{MAJ}_{n,k}) = \begin{cases} \max\left\{n^{\Omega(k/\log\log n)}, n^{\Omega(\log k)}\right\} & \text{if } k \leqslant \log n, \\ n^{\Omega(k/\log k)} & \text{if } \log n \leqslant k \leqslant \sqrt{n}. \end{cases}$$

PROOF. Immediate from Theorems 10.13 and 13.1. □

Chapter 14

Lower Bounds for Agnostic Learning

This chapter studies the *agnostic model*, a powerful abstraction of learning from noisy training data. Both algorithmic results and lower bounds for this model have seen limited progress. We contribute several new lower bounds, ruling out the use of the current techniques for learning concept classes as simple as decision lists and disjunctions. Along the way we relate agnostic learning, via the pattern matrix method, to our work on communication complexity as well as to an algorithmic problem known as *approximate inclusion-exclusion.*

14.1 Introduction

Valiant's PAC model has been a central model in computational learning theory and the focus of much algorithmic and complexity-theoretic research. From a practical standpoint, however, the PAC model is somewhat problematic in that the learning algorithm requires noise-free, uncorrupted evaluations of the unknown function on a rather large sample of points. Such training data can be difficult to obtain, for various reasons such as human error and measurement error. One model of noisy training data, referred to as random classification noise, was already mentioned in the previous chapter on statistical query learning. Here, we consider the much more realistic and challenging *agnostic model*, reviewed in Section 14.2. The agnostic learner no longer assumes that some function in the concept class $\mathscr{C}$ perfectly matches the training data, hence the term *agnostic*. The objective, then, is to identify a function $f \in \mathscr{C}$ that matches the training data almost as well as the best member of $\mathscr{C}$.

Designing efficient algorithms in the agnostic model is notoriously difficult. Nevertheless, progress on proving lower bounds has also been scarce. The purpose of this chapter is to prove several such lower bounds. Kalai et al. [101] discovered what appears to be the only efficient, general-purpose algorithm for agnostic learning to date. This algorithm efficiently learns concept classes $\mathscr{C}$ with low approximate rank. We study approximate rank in Sections 14.3 and 14.4 and develop versatile techniques for bounding it from below. Our proofs use the pattern matrix method as well as Fourier analysis and the basics of perturbation theory for real matrices. In particular, we derive tight, exponential lower bounds on the approximate rank of disjunctions, decision lists, and majority functions, all well-studied concept classes in learning theory. It follows that an approach fundamentally different from

that of Kalai et al. [101] will be needed to solve the agnostic learning problem even for these concept classes. In Section 14.5, we additionally relate the approximate rank to the statistical query dimension, linking the two learning models.

Another natural approach to learning in the agnostic model and other models is to use hypotheses $h\colon \{0,1\}^n \to \{-1,+1\}$ that can be expressed as low-degree polynomials. In Section 14.6, we prove that this approach fails for various common concept classes. Building on our analysis, we relate agnostic learning in Section 14.7 to the *approximate inclusion-exclusion* problem. This problem, due to Linial and Nisan [143], requires one to estimate the probability $\mathbf{P}[f(A_1, \dots, A_n)]$ for given events $A_1, \dots, A_n$ and a given Boolean function f, using only the probabilities of intersections $\bigcap_{i \in S} A_i$ for small sets S. In the concluding section of this chapter, we determine the ε-approximate degree for every symmetric Boolean function and every $\varepsilon \leqslant 1/3$, which we use to give a full solution of the approximate inclusion-exclusion problem for symmetric functions.

14.2 Agnostic learning model

In our discussion of statistical query learning in the previous chapter, we mentioned a generalization of the PAC model whereby the learner receives training examples of the form $(x^{(1)}, \xi_1 f(x^{(1)})), \dots, (x^{(m)}, \xi_m f(x^{(m)}))$, where $\xi_1, \dots, \xi_m$ are independent random variables taking on -1 with a small probability ε and taking on $+1$ with the complementary probability $1 - \varepsilon$. For $\varepsilon = 0$, this model is identical to Valiant's original PAC model. For $0 < \varepsilon < 1/2$, the new model is commonly described as PAC learning with *random classification noise*. This noise regime is arguably the most benign from a learning standpoint.

Agnostic learning is a more realistic, adversarial model of noisy learning proposed by Kearns et al. [107]. Let $\mathscr{C}$ be a concept class of functions $X \to \{-1,+1\}$ for some finite set X. Let λ be a distribution on $X \times \{-1,+1\}$, unknown to the learner. The learner receives training examples $(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})$ distributed independently according to λ. Let

$$\mathsf{opt} = \max_{f \in \mathscr{C}} \left\{ \mathop{\mathbf{P}}_{(x,y)\sim\lambda} [f(x) = y] \right\}$$

be the performance of the function in $\mathscr{C}$ that best agrees with the training data. The learner in this model is called upon to produce, with probability at least $1 - \delta$, a hypothesis $h: X \to \{-1, +1\}$ with near-optimal agreement with the training data:

$$\mathop{\mathbf{P}}_{(x,y)\sim\lambda}[h(x) = y] \geqslant \mathsf{opt} - \varepsilon,$$

where ε is an error parameter fixed in advance. If such a learning algorithm can be found, one says that $\mathscr{C}$ is *learnable to accuracy* ε. We stress that the hypothesis h need not be a member of $\mathscr{C}$. The term *agnostic learning* points to the fact that the learner can no longer assume the existence of a function in the concept class that matches the data perfectly, as was the case in PAC learning. The objective is simply to be competitive with the most accurate classifier from $\mathscr{C}$. As usual, the goal is to have an algorithm that is efficient with respect to both training data and running time. Ideally, one would want an algorithm with running time and training data requirements polynomial in $\log|\mathscr{C}|$, $\log|X|$, $1/\varepsilon$, and $1/\delta$.

Designing efficient algorithms in the agnostic model is difficult. Kalai et al. [101] discovered what appears to be the only efficient, general-purpose algorithm for agnostic learning to date. The algorithm works efficiently whenever the concept class satisfies a natural analytic property, as follows.

THEOREM 14.1 (Kalai et al. [101], implicit). *Fix a constant $\varepsilon > 0$ and a concept class $\mathscr{C}$ of functions $\{0,1\}^n \to \{-1,+1\}$. Assume that there are functions $\phi_1, \ldots, \phi_r: \{0,1\}^n \to \mathbb{R}$ such that for every $f \in \mathscr{C}$,*

$$\min_{\phi\in\operatorname{span}\{\phi_1,\ldots,\phi_r\}} \|f - \phi\|_\infty \leqslant \frac{1}{3}.$$

Assume further that each $\phi_i(x)$ is computable in polynomial time. Then $\mathscr{C}$ is agnostically learnable to accuracy ε in time polynomial in r and n.

Let $\mathscr{C}$ be a concept class of functions $X \to \{-1, +1\}$, for some finite set X. We define the *ε-approximate rank* of $\mathscr{C}$, denoted $\mathrm{rk}_\varepsilon\,\mathscr{C}$, to be the ε-approximate rank of the characteristic matrix $M_{\mathscr{C}} = [f(x)]_{f\in\mathscr{C}, x\in X}$. One contribution of this

chapter is to derive exponential lower bounds on the approximate rank of concept classes as simple as disjunctions and majority functions, thereby showing that an approach fundamentally different from Theorem 14.1 is needed to efficiently learn those concept classes in the agnostic model. Note that we will not use the assumption of polynomial-time computability in Theorem 14.1, which makes our results stronger.

14.3 Analyzing the approximate rank

In this section, we take a closer look at the ε-approximate rank as a function of ε and develop some techniques for analyzing it. Let M be a sign matrix. Suppose that we have an estimate of $\mathrm{rk}_E M$ for some E with $0 < E < 1$. Can we use this information to obtain a nontrivial upper bound on $\mathrm{rk}_\varepsilon M$, where $0 < \varepsilon < E$? It turns out that we can. We first recall that the sign function can be approximated well by a low-degree polynomial.

FACT 14.2. *Let E be given, $0 < E < 1$. Then for each integer $d \geqslant 1$, there exists $p \in P_d$ such that*

$$|p(t) - \operatorname{sgn} t| \leqslant 8\sqrt{d}\left(1 - \frac{(1-E)^2}{16}\right)^d \qquad (1 - E \leqslant |t| \leqslant 1 + E).$$

Fact 14.2 is implicit in Rudin's proof [188, Thm. 7.26] of the Weierstrass approximation theorem. Subtler, improved versions of Fact 14.2 can be readily found in the approximation literature. As an application to approximate rank, we have the following result.

THEOREM 14.3 (Klivans and Sherstov [126]). *Let M be a sign matrix. Let E, ε be given with $0 < \varepsilon < E < 1$. Then*

$$\mathrm{rk}_\varepsilon M \leqslant (\mathrm{rk}_E M)^d,$$

where d is any positive integer with

$$8\sqrt{d}\left(1 - \frac{(1-E)^2}{16}\right)^d \leqslant \varepsilon.$$

PROOF. The key idea of the proof is to improve the quality of the approximating matrix by applying a suitable polynomial to its entries. Prior to our work, this technique was used by Alon [12] in the simpler setting of *one-sided* uniform approximation.

Specifically, let d be as stated. By Fact 14.2, there exists $p \in P_d$ with

$$|p(t) - \operatorname{sgn} t| \leqslant \varepsilon \qquad (1 - E \leqslant |t| \leqslant 1 + E).$$

Let A be a real matrix with $\|A - M\|_\infty \leqslant E$ and $\operatorname{rk} A = \operatorname{rk}_E M$. Then the matrix $B = [p(A_{ij})]_{i,j}$ obeys $\|B - M\|_\infty \leqslant \varepsilon$. Since p is a polynomial of degree d, elementary linear algebra shows that $\operatorname{rk} B \leqslant (\operatorname{rk} A)^d$. □

COROLLARY 14.4 (Klivans and Sherstov [126]). *Let M be a sign matrix. Let ε, E be constants with $0 < \varepsilon < E < 1$. Then*

$$\operatorname{rk}_\varepsilon M \leqslant (\operatorname{rk}_E M)^c,$$

where $c = c(\varepsilon, E)$ is a constant.

The above corollary shows that the choice of the constant $0 < \varepsilon < 1$ affects $\operatorname{rk}_\varepsilon M$ by at most a polynomial factor. When such factors are unimportant, we will adopt $\varepsilon = 1/3$ as a canonical setting. We will now develop some techniques for analyzing the approximate rank. An important tool in this chapter is the well-known Hoffman-Wielandt inequality [85, Thm. 8.6.4], which states that small perturbations to the entries of a matrix result in small perturbations to its singular values. This inequality has seen numerous previous uses in the complexity theory literature [146, 103, 73].

THEOREM 14.5 (Hoffman-Wielandt inequality). *Let $A, B \in \mathbb{R}^{m\times n}$. Then*

$$\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(B))^2 \leqslant \|A - B\|_{\mathrm{F}}^2.$$

In particular, if $\operatorname{rk} B = k$ *then*

$$\sum_{i\geqslant k+1} \sigma_i(A)^2 \leqslant \|A - B\|_{\mathrm{F}}^2.$$

As an application of the Hoffman-Wielandt inequality, we obtain a lower bound on the approximate trace norm, which in view of Proposition 2.12 implies a lower bound on the approximate rank.

LEMMA 14.6 (Klivans and Sherstov [126]). *Let $M = [f(x \oplus y)]_{x,y}$, where $f\colon \{0,1\}^n \to \{-1,+1\}$ is a given function and the indices x, y range over $\{0,1\}^n$. Then for all $\varepsilon \geqslant 0$,*

$$\|M\|_{\Sigma,\varepsilon} \geqslant 2^n(\|\hat{f}\|_1 - \varepsilon 2^{n/2}). \tag{14.1}$$

In particular,

$$\operatorname{rk}_\varepsilon M \geqslant \left(\frac{\|\hat{f}\|_1 - \varepsilon 2^{n/2}}{1+\varepsilon}\right)^2 \tag{14.2}$$

provided that $\|\hat{f}\|_1 > \varepsilon 2^{n/2}$.

PROOF. Let $N = 2^n$ be the order of M. Fix a matrix A with $\|A - M\|_\infty \leqslant \varepsilon$. By the Hoffman-Wielandt inequality,

$$N^2\varepsilon^2 \geqslant \|A - M\|_F^2 \geqslant \sum_{i=1}^{N}(\sigma_i(A) - \sigma_i(M))^2 \geqslant \frac{1}{N}(\|A\|_\Sigma - \|M\|_\Sigma)^2,$$

so that $\|A\|_\Sigma \geqslant \|M\|_\Sigma - N^{3/2}\varepsilon$. Since the choice of A was arbitrary, we conclude that

$$\|M\|_{\Sigma,\varepsilon} \geqslant \|M\|_\Sigma - N^{3/2}\varepsilon. \tag{14.3}$$

It is well-known [142, p. 458] that the singular values of M/N are precisely the absolute values of the Fourier coefficients of f. Indeed,

$$M = Q \begin{bmatrix} N\hat{f}(\varnothing) & & \\ & \ddots & \\ & & N\hat{f}([n]) \end{bmatrix} Q^{\mathsf{T}},$$

where $Q = N^{-1/2}[\chi_S(x)]_{x,S}$ is an orthogonal matrix. In particular, $\|M\|_\Sigma = N\|\hat{f}\|_1$. Together with (14.3), this completes the proof of (14.1). Finally, (14.2) follows from (14.1) and Proposition 2.12. □

We will now use the pattern matrix method to obtain a different lower bound on the approximate trace norm and approximate rank.

THEOREM 14.7 (Sherstov [203]). *Let F be the (n, t, f)-pattern matrix, where $f\colon \{0,1\}^t \to \{-1,+1\}$ is given. Let $s = 2^{n+t}(n/t)^t$ be the number of entries in F. Then for every $\varepsilon \in [0,1)$ and every $\delta \in [0,\varepsilon]$,*

$$\|F\|_{\Sigma,\delta} \geqslant (\varepsilon - \delta)\left(\frac{n}{t}\right)^{\deg_\varepsilon(f)/2}\sqrt{s} \tag{14.4}$$

and

$$\mathrm{rk}_{\delta}\, F \geqslant \left(\frac{\varepsilon-\delta}{1+\delta}\right)^2 \left(\frac{n}{t}\right)^{\deg_{\varepsilon}(f)}. \tag{14.5}$$

PROOF. We may assume that $\deg_{\varepsilon}(f) \geqslant 1$, since otherwise f is a constant function and the claims hold trivially by taking $\Psi = F$ in Proposition 2.11. Construct Ψ as in the proof of Theorem 4.8. Then the lower bound on $\|F\|_{\Sigma,\delta}$ follows from (4.16), (4.18), and Proposition 2.11. Finally, (14.5) follows from (14.4) and Proposition 2.12. □

We prove an additional lower bound in the case of small-bias approximation.

THEOREM 14.8 (Sherstov [203]). *Let F be the (n,t,f)-pattern matrix, where $f\colon\{0,1\}^t \to \{-1,+1\}$ is given. Let $s = 2^{n+t}(n/t)^t$ be the number of entries in F. Then for every $\gamma \in (0,1)$ and every integer $d \geqslant 1$,*

$$\|F\|_{\Sigma,1-\gamma} \geqslant \gamma \min\left\{\left(\frac{n}{t}\right)^{d/2}, \left(\frac{W(f,d-1)}{2t}\right)^{1/2}\right\}\sqrt{s} \tag{14.6}$$

and

$$\mathrm{rk}_{1-\gamma}\, F \geqslant \left(\frac{\gamma}{2-\gamma}\right)^2 \min\left\{\left(\frac{n}{t}\right)^{d}, \frac{W(f,d-1)}{2t}\right\}. \tag{14.7}$$

In particular,

$$\|F\|_{\Sigma,1-\gamma} \geqslant \gamma \left(\frac{n}{t}\right)^{\deg_{\pm}(f)/2}\sqrt{s} \tag{14.8}$$

and

$$\mathrm{rk}_{1-\gamma} F \geqslant \left(\frac{\gamma}{2-\gamma}\right)^2 \left(\frac{n}{t}\right)^{\deg_\pm(f)}. \tag{14.9}$$

PROOF. Construct Ψ as in the proof of Theorem 4.11. Then the claimed lower bound on $\|F\|_{\Sigma,\delta}$ follows from (4.26), (4.28), and Proposition 2.11. Now (14.7) follows from (14.6) and Proposition 2.12. Finally, (14.8) and (14.9) follow by taking $d = \deg_\pm(f)$ in (14.6) and (14.7), respectively, since $W(f, d-1) = \infty$ in that case. □

Recall that Theorem 4.3 gives an easy way to calculate the trace norm and rank of a pattern matrix. In particular, it is straightforward to verify that the lower bounds in (14.5) and (14.7) are close to optimal for various choices of $\varepsilon, \delta, \gamma$. For example, one has $\|F - A\|_\infty \leqslant 1/3$ by taking F and A to be the (n, t, f)- and (n, t, ϕ)-pattern matrices, where $\phi\colon \{0,1\}^t \to \mathbb{R}$ is any polynomial of degree $\deg_{1/3}(f)$ with $\|f - \phi\|_\infty \leqslant 1/3$.

14.4 Approximate rank of specific concept classes

We will now apply the techniques of the previous section to specific concept classes. The proofs below differ from those in the original publication [126], which chronologically preceded the discovery of the pattern matrix method [203]. Using the pattern matrix method, we are able to considerably simplify the original proofs. We start with the concept class of disjunctions.

THEOREM 14.9 (Klivans and Sherstov [126]). *Let $\mathscr{C}$ be the concept class of functions $f\colon \{0,1\}^n \to \{-1,+1\}$ of the form $f(x) = \bigvee_{i \in S} x_i$ for some subset $S \subseteq \{1, 2, \ldots, n\}$. Then*

$$\mathrm{rk}_{1/3}(\mathscr{C}) = \exp\{\Omega(\sqrt{n})\}.$$

PROOF. The characteristic matrix of $\mathscr{C}$ is $M = [\mathrm{OR}_n(x \wedge y)]_{x,y}$, where the indices x, y range over $\{0,1\}^n$. Note that M contains the $(2m, m, \mathrm{OR}_m)$-pattern matrix as a submatrix, where $m = \lfloor n/4 \rfloor$. The proof is now complete in view of Theorems 2.5 and 14.7 and Corollary 14.4. □

The above theorem shows in particular that the algorithm of Kalai et al. [101], which agnostically learns the concept class of disjunctions in time $\exp\{\tilde{\Theta}(\sqrt{n})\}$, is essentially optimal. We now prove an incomparable result for the concept class of decision lists.

THEOREM 14.10 (Klivans and Sherstov [126]). *Let $\mathscr{C}$ denote the concept class of functions $f : \{0,1\}^n \to \{-1,+1\}$ representable as*

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n} (-2)^i x_i y_i\right)$$

for some $y \in \{0,1\}^n$. Then for a sufficiently small absolute constant $\alpha > 0$,

$$\mathrm{rk}_{1-\exp(-\alpha n^{1/3})}(\mathscr{C}) \geqslant \exp(\alpha n^{1/3}).$$

PROOF. The characteristic matrix of $\mathscr{C}$ is $M = [\mathrm{OMB}_n(x \wedge y)]_{x,y}$, where the function $\mathrm{OMB}_n : \{0,1\}^n \to \{-1,+1\}$ is given by (4.34). Put $m = \lfloor n/4 \rfloor$. A well-known result due to Beigel [32] shows that $W(\mathrm{OMB}_m, \alpha m^{1/3}) \geqslant \exp(\alpha m^{1/3})$ for some absolute constant $\alpha > 0$. Since the $(2m, m, \mathrm{OMB}_m)$-pattern matrix is a submatrix of M, the proof is complete in view of Theorem 14.8. □

Note that the concept class in Theorem 14.10 is a subclass of decision lists, a heavily studied family of functions in learning theory. Comparing the results of Theorems 14.9 and 14.10 for small constant ε, we see that Theorem 14.9 is stronger in that it gives a better lower bound for a simpler concept class. On the other hand, Theorem 14.10 is stronger in that it remains valid in the broad range $0 \leqslant \varepsilon \leqslant 1 - \exp\{-\Theta(n^{1/3})\}$, whereas the ε-approximate rank in Theorem 14.9 is easily seen to be at most n for all $\varepsilon \geqslant 1 - \frac{1}{2n}$.

As a final application, we consider the concept class of majority functions. Here we prove a lower bound of $\Omega(2^n/n)$ on the approximate rank, which is the best of our three constructions and nearly matches the trivial upper bound of 2^n.

THEOREM 14.11 (Klivans and Sherstov [126]). *Let $\mathscr{C}$ denote the concept class of functions $f : \{0,1\}^n \to \{-1,+1\}$ representable as $f(x) = \mathrm{MAJ}_n(x \oplus y)$ for some $y \in \{0,1\}^n$. Then*

$$\mathrm{rk}_{\alpha/\sqrt{n}}(\mathscr{C}) \geqslant \Omega\left(\frac{2^n}{n}\right) \tag{14.10}$$

for a sufficiently small absolute constant $\alpha > 0$. Also,

$$\mathrm{rk}_{1/3}(\mathscr{C}) = \exp\{\Omega(n)\}. \tag{14.11}$$

PROOF. The characteristic matrix of $\mathscr{C}$ is $M = [\mathrm{MAJ}_n(x \oplus y)]_{x,y}$. The Fourier spectrum of the majority function has been extensively studied by various authors. In particular, it is well-known [142, §7] that

$$\|\widehat{\mathrm{MAJ}_n}\|_1 = \Omega\left(\frac{2^{n/2}}{\sqrt{n}}\right).$$

Now (14.10) follows by Lemma 14.6.

Finally, it is straightforward to verify that M contains the $(2m, m, \mathrm{MAJ}_m)$-pattern matrix as a submatrix, where $m = \lfloor n/4 \rfloor$. Hence, (14.11) follows at once from Theorems 2.5 and 14.7 and Corollary 14.4. □

14.5 Approximate rank vs. statistical query dimension

Let M be a sign matrix. Recall that $\mathrm{rk}_\varepsilon M \geqslant \mathrm{rk}_\pm M$ for every ε with $0 \leqslant \varepsilon < 1$. In other words, the sign rank of a matrix is a lower bound on its approximate rank. Furthermore, we proved in Theorem 10.13 that $\mathrm{rk}_\pm M > \sqrt{\mathrm{sq}(M)/2}$. Combining

these two facts, we obtain

$$\mathrm{rk}_\varepsilon M > \sqrt{\frac{\mathrm{sq}(M)}{2}}, \qquad 0 \leqslant \varepsilon < 1.$$

In this section, we take a closer look at the relationship between the statistical query dimension and approximate rank. First, we will strengthen the above lower bound to $\mathrm{rk}_\varepsilon M \geqslant \Omega(\mathrm{sq}(M))$ for constant ε. Second, we will construct an explicit matrix with an exponential gap between its approximate rank and statistical query dimension. The novelty in the second case is not the exponential separation per se, which follows from Theorem 10.14, but rather the fact that such a separation is achieved for an explicit matrix.

A starting point in our analysis is the relationship between the statistical query dimension of a concept class $\mathscr{C}$ and the approximation of $\mathscr{C}$ in the ℓ_2 norm, which is also of some independent interest.

THEOREM 14.12 (Klivans and Sherstov [126]). *Let $\mathscr{C}$ be a concept class of functions $X \to \{-1,+1\}$, for some finite set X. Let μ be a distribution over X. Suppose there exist functions $\phi_1, \ldots, \phi_r : X \to \mathbb{R}$ such that each $f \in \mathscr{C}$ has $\mathbf{E}_{x\sim\mu}[(f(x) - \sum_{i=1}^r \alpha_i \phi_i(x))^2] \leqslant \varepsilon$ for some reals $\alpha_1, \ldots, \alpha_r$. Then*

$$r \geqslant (1-\varepsilon)d - \sqrt{d},$$

where $d = \mathrm{sq}_\mu(\mathscr{C})$.

PROOF. By the definition of the statistical query dimension, there exist functions $f_1, \ldots, f_d \in \mathscr{C}$ with

$$\left| \mathop{\mathbf{E}}_{\mu}[f_i(x) f_j(x)] \right| \leqslant \frac{1}{d}$$

for all $i \neq j$. For simplicity, assume that μ is a distribution with rational weights (extension to the general case is straightforward). Then there is an integer $k \geqslant 1$

such that each $\mu(x)$ is an integral multiple of $1/k$. Construct the $d \times k$ sign matrix

$$M = [f_i(x)]_{i,x},$$

whose rows are indexed by the functions $f_1, \ldots, f_d$ and whose columns are indexed by inputs $x \in X$ (a given input x indexes exactly $k\mu(x)$ columns). It is easy to verify that $MM^{\mathsf{T}} = [k \mathbf{E}_\mu[f_i(x) f_j(x)]]_{i,j}$, and thus

$$\|MM^{\mathsf{T}} - k \cdot I\|_{\mathrm{F}} < k. \tag{14.12}$$

The existence of $\phi_1, \ldots, \phi_r$ implies the existence of a rank-r real matrix A with $\|M - A\|_{\mathrm{F}}^2 \leqslant \varepsilon k d$. On the other hand, Theorem 14.5 (the Hoffman-Wielandt inequality) guarantees that $\|M - A\|_F^2 \geqslant \sum_{i=r+1}^{d} \sigma_i(M)^2$. Combining these two inequalities yields:

$$\begin{aligned}
\varepsilon k d &\geqslant \sum_{i=r+1}^{d} \sigma_i(M)^2 = \sum_{i=r+1}^{d} \sigma_i(MM^{\mathsf{T}}) \\
&\geqslant k(d-r) - \sum_{i=r+1}^{d} |\sigma_i(MM^{\mathsf{T}}) - k| \\
&\geqslant k(d-r) - \left(\sum_{i=r+1}^{d} (\sigma_i(MM^{\mathsf{T}}) - k)^2 \right)^{1/2} \sqrt{d-r} \\
&\geqslant k(d-r) - \|MM^{\mathsf{T}} - k \cdot I\|_F \sqrt{d-r} && \text{by Theorem 14.5} \\
&\geqslant k(d-r) - k\sqrt{d} && \text{by (14.12).}
\end{aligned}$$

We have shown that $\varepsilon d \geqslant (d-r) - \sqrt{d}$, which is precisely what the theorem claims. To extend the proof to irrational distributions μ, one considers a rational distribution $\tilde{\mu}$ suitably close to μ and repeats the above analysis. □

We are now in a position to relate the statistical query dimension to the approximate rank and exhibit an exponential gap between the two quantities.

THEOREM 14.13 (Klivans and Sherstov [126]). *Let $\mathscr{C}$ be a concept class of functions $X \to \{-1,+1\}$ for some finite set X. Then for $0 \leqslant \varepsilon < 1$,*

$$\mathrm{rk}_\varepsilon(\mathscr{C}) \geqslant (1-\varepsilon^2)\,\mathrm{sq}(\mathscr{C}) - \sqrt{\mathrm{sq}(\mathscr{C})}. \tag{14.13}$$

Moreover, there exists an (explicitly given) concept class $\mathscr{A}$ with

$$\begin{aligned} \mathrm{sq}(\mathscr{A}) &\leqslant O(n^2), \\ \mathrm{rk}_{1/3}(\mathscr{A}) &\geqslant \exp\{\Omega(n)\}. \end{aligned}$$

PROOF. Let $r = \mathrm{rk}_\varepsilon(\mathscr{C})$. Then there are functions $\phi_1, \ldots, \phi_r$ such that every function $f \in \mathscr{C}$ has $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant \varepsilon$ for some reals $\alpha_1, \ldots, \alpha_r$. As a result,

$$\mathop{\mathbf{E}}_{x\sim\mu}\left[\left(f(x) - \sum_{i=1}^r \alpha_i \phi_i(x)\right)^2\right] \leqslant \varepsilon^2$$

for every distribution μ. By Theorem 14.12,

$$r \geqslant (1-\varepsilon^2)\,\mathrm{sq}_\mu(\mathscr{C}) - \sqrt{\mathrm{sq}_\mu(\mathscr{C})}.$$

Maximizing over μ establishes (14.13).

To prove the second part, let $\mathrm{MAJ}_n\colon \{0,1\}^n \to \{-1,+1\}$ be the majority function and let $\mathscr{A}$ be the family of all its reflections $\mathrm{MAJ}_n(x \oplus y)$, $y \in \{0,1\}^n$. Theorem 14.11 shows that $\mathscr{A}$ has the stated approximate rank. To bound its statistical query dimension, note that each function in $\mathscr{A}$ can be pointwise approximated within error $1 - 1/n$ by a linear combination of the functions $(-1)^{x_1}, \ldots, (-1)^{x_n}$. Therefore, (14.13) implies that $\mathrm{sq}(\mathscr{A}) \leqslant O(n^2)$. □

14.6 Learning via low-degree polynomials

A natural approach to learning in the agnostic model and other models is to consider only those hypotheses that depend on few variables. One tests each such hypothesis against the training data and outputs the one with the least error. This technique is attractive in that the hypothesis space is small and well-structured, making it possible to efficiently identify the best approximation to the observed examples. The question then becomes, what advantage over random guessing can such hypotheses guarantee? In this section we prove that, even when learning the simplest concept classes, one is forced to use hypotheses that depend on many variables, all others having zero advantage over random guessing. The following definition formalizes the subject of our study.

DEFINITION 14.14 (Sherstov [199]). Let $f : \{0,1\}^n \to \{-1,+1\}$ be a given function. Define

$$\Gamma(f,k) = \max_{\lambda} \left\{ \mathop{\mathbf{P}}_{(x,y)\sim\lambda}[f(x) = y] \right\},$$

where the maximum is taken over all distributions λ on $\{0,1\}^n \times \{-1,+1\}$ such that

$$\mathop{\mathbf{P}}_{(x,y)\sim\lambda}[g(x) = y] = \frac{1}{2}$$

for every function $g : \{0,1\}^n \to \{-1,+1\}$ that depends on k or fewer variables.

Observe that the maximization in Definition 14.14 is over a nonempty compact set that contains the uniform distribution. We will see, among other things, that $\Gamma(\mathrm{OR}_n, \Theta(\sqrt{n})) \geqslant 0.99$. In other words, even though the training examples have 99% agreement with the target function OR_n, no hypothesis that depends on few variables can match the data better than random. An analogous statement holds for any Boolean function with high approximate degree.

LEMMA 14.15 (Sherstov [199]). *Let λ be a distribution on $\{0,1\}^n \times \{-1,+1\}$. Then for every function $f : \{0,1\}^n \to \{-1,+1\}$,*

$$\mathop{\mathbf{P}}_{(x,y)\sim\lambda}[f(x) = y] = \frac{1}{2} + \frac{1}{2} \sum_{x\in\{0,1\}^n} \{\lambda(x,1) - \lambda(x,-1)\} f(x).$$

PROOF. We have

$$\begin{aligned}
\mathop{\mathbf{P}}_{(x,y)\sim\lambda}[f(x) = y] &= \mathop{\mathbf{P}}_{(x,y)\sim\lambda}[f(x) = y = -1] + \mathop{\mathbf{P}}_{(x,y)\sim\lambda}[f(x) = y = +1] \\
&= \sum_{x\in\{0,1\}^n} \frac{1 - f(x)}{2} \cdot \lambda(x,-1) + \sum_{x\in\{0,1\}^n} \frac{1 + f(x)}{2} \cdot \lambda(x,1) \\
&= \frac{1}{2} + \frac{1}{2} \sum_{x\in\{0,1\}^n} \{\lambda(x,1) - \lambda(x,-1)\} f(x). \qquad \square
\end{aligned}$$

Recall from Section 2.2 that the symbol $E(f,k)$ stands for the least error in a uniform approximation of the function f by a polynomial of degree at most k. We will now show that $\Gamma(f,k)$ and $E(f,k)$ are closely related quantities.

THEOREM 14.16 (Sherstov [199]). *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a given function. Then*

$$\Gamma(f,k) = \frac{E(f,k) + 1}{2}.$$

PROOF. By Lemma 14.15,

$$\Gamma(f,k) = \frac{1}{2} + \frac{1}{2} \max_{\lambda} \left\{ \sum_{x\in\{0,1\}^n} (\lambda(x,1) - \lambda(x,0)) f(x) \right\},$$

where the maximum is over all distributions λ on $\{0,1\}^n \times \{-1,+1\}$ such that

$$\sum_{x \in \{0,1\}^n} (\lambda(x,1) - \lambda(x,-1)) g(x) = 0$$

for every function $g: \{0,1\}^n \to \{-1,+1\}$ that depends on k or fewer variables. As λ ranges over all distributions, the function $\lambda(x,1) - \lambda(x,-1)$ ranges over all functions $\psi: \{0,1\}^n \to \mathbb{R}$ with $\|\psi\|_1 \leqslant 1$. Equivalently,

$$\Gamma(f,k) = \frac{1}{2} + \frac{1}{2} \max_{\psi} \left\{ \sum_{x \in \{0,1\}^n} \psi(x) f(x) \right\},$$

where the maximum is over all $\psi: \{0,1\}^n \to \mathbb{R}$ such that $\|\psi\|_1 \leqslant 1$ and $\hat{\psi}(S) = 0$ for $|S| \leqslant k$. By Theorem 4.4, the proof is complete. □

Rephrasing the previous theorem yields the main result of this section.

THEOREM 14.17 (Sherstov [199]). *Let $f: \{0,1\}^n \to \{-1,+1\}$ be a given function. Then for each $k \geqslant 0$, there is a distribution λ on $\{0,1\}^n \times \{-1,+1\}$ such that*

$$\mathop{\mathbf{P}}_{(x,y) \sim \lambda} [f(x) = y] \geqslant \frac{1 + E(f,k)}{2}$$

and

$$\mathop{\mathbf{P}}_{(x,y) \sim \lambda} [g(x) = y] = \frac{1}{2}$$

for every $g: \{0,1\}^n \to \{-1,+1\}$ that depends on at most k variables.

PROOF. Immediate from Definition 14.14 and Theorem 14.16. □

In particular, Theorem 14.17 settles the promised statement about Boolean functions with high approximate degree. To place this result in the framework of agnostic learning, consider any concept class $\mathscr{C}$ that contains many functions with high approximate degree. For example, one could fix a nonconstant symmetric function $f:\{0,1\}^n \to \{-1,+1\}$ and consider the concept class $\mathscr{C}$ of $\binom{2n}{n}$ functions, each being a copy of f applied to a separate set of n variables from among $x_1, x_2, \ldots, x_{2n}$:

$$\mathscr{C} = \Big\{ f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) : \quad 1 \leqslant i_1 < i_2 < \cdots < i_n \leqslant 2n \Big\}.$$

Theorem 14.17 now supplies scenarios when *some* member of $\mathscr{C}$ matches the training data almost perfectly, and yet every hypothesis that depends on few variables is completely useless. This result generalizes earlier work by Tarui and Tsukiji [211], who obtained a special case of Theorem 14.17 for the OR function.

14.7 Relationship to approximate inclusion-exclusion

Let $A_1, A_2, \ldots, A_n$ be events in a probability space. The well-known inclusion-exclusion principle allows one to compute the probability of $A_1 \cup \cdots \cup A_n$ using the probabilities of various intersections of $A_1, A_2, \ldots, A_n$:

$$\mathbf{P}[A_1 \cup \cdots \cup A_n] = \sum_i \mathbf{P}[A_i] - \sum_{i<j} \mathbf{P}[A_i \cap A_j] + \sum_{i<j<k} \mathbf{P}[A_i \cap A_j \cap A_k] - \cdots + (-1)^{n+1} \mathbf{P}[A_1 \cap \cdots \cap A_n].$$

A moment's reflection shows that knowledge of every term in this summation is necessary in general for an exact answer [143]. It is therefore natural to wonder if one can closely approximate $\mathbf{P}[\bigcup A_i]$ using the probabilities of intersections of up to k events, where $k \ll n$. This problem, due to Linial and Nisan [143], is known as *approximate inclusion-exclusion.* Linial and Nisan studied this question and gave near-tight bounds on the least approximation error as a function of k. A

follow-up article by Kahn, Linial, and Samorodnitsky [100] improved those bounds to optimal.

While $A_1 \cup \cdots \cup A_n$ is an important event, it is not the only one of interest. For example, we might be interested in the probability that *most* of the events $A_1, \ldots, A_n$ occur, or the probability that an *odd number* of the events from among $A_1, \ldots, A_n$ occur. More generally, let $f : \{0,1\}^n \to \{-1,+1\}$ be a given Boolean function. The problem of interest to us in this section is that of estimating

$$\mathbf{P}[f(A_1, \ldots, A_n) = -1]$$

given the probabilities $\mathbf{P}[\bigcap_{i \in S} A_i]$ for $|S| \leqslant k$. In what follows, we will adopt the shorthand $\mathbf{P}[f(A_1, \ldots, A_n)] = \mathbf{P}[f(A_1, \ldots, A_n) = -1]$.

Our approach departs from the previous methods [143, 100], which are specialized to the case $f = \mathrm{OR}_n$. First, we will show that the inclusion-exclusion problem for a given function f is exactly equivalent to a classical approximation problem. Specifically, define

$$\Delta(f,k) = \frac{1}{2} \sup \left\{ \mathop{\mathbf{P}}_{\mathscr{P}_1}[f(A_1, \ldots, A_n)] \; - \; \mathop{\mathbf{P}}_{\mathscr{P}_2}[f(B_1, \ldots, B_n)] \right\},$$

where the supremum is over all probability spaces $\mathscr{P}_1$ and $\mathscr{P}_2$, over all events $A_1, \ldots, A_n$ in $\mathscr{P}_1$, and over all events $B_1, \ldots, B_n$ in $\mathscr{P}_2$, such that

$$\mathop{\mathbf{P}}_{\mathscr{P}_1}\left[\bigcap_{i \in S} A_i\right] = \mathop{\mathbf{P}}_{\mathscr{P}_2}\left[\bigcap_{i \in S} B_i\right], \qquad |S| \leqslant k. \tag{14.14}$$

In words, the quantity $\Delta(f,k)$ is the optimal error achievable in approximating $\mathbf{P}[f(A_1, \ldots, A_n)]$. We will show:

THEOREM 14.18 (Sherstov [199]). *Let $f\colon \{0,1\}^n \to \{-1,+1\}$ be a given function. Then for all k,*

$$\Delta(f,k) = \frac{E(f,k)}{2}.$$

Theorem 14.18 states that the approximate inclusion-exclusion problem for a given function f is equivalent to the problem of approximating f by a multivariate polynomial of degree up to k. Note the similarity of this result to our earlier Theorem 14.16 on agnostic learning. This similarity is no accident; the two results are proved via the same duality transformation, revealing a relationship between agnostic learning and a seemingly disjoint algorithmic question. In the next section, we will solve the approximation problem for all symmetric functions and all k, thereby solving the corresponding inclusion-exclusion problem.

In the remainder of this section, we focus on proving Theorem 14.18. First, we will first show that the arbitrary probability spaces in the definition of $\Delta(f,k)$ can be restricted to probability distributions on $\{0,1\}^n$.

DEFINITION 14.19 (Sherstov [199]). Let $E_1, \ldots, E_n$ be events in a probability space $\mathscr{P}$. The *distribution on $\{0,1\}^n$ induced by $\mathscr{P}, E_1, \ldots, E_n$* is defined as

$$\mu(x) = \mathbf{P}\left[\bigcap_{i:x_i=0} \overline{E_i} \quad \bigcap_{i:x_i=1} E_i\right].$$

PROPOSITION 14.20 (Sherstov [199]). *Let $E_1, \ldots, E_n$ be events in a probability space $\mathscr{P}$. Let μ be the distribution on $\{0,1\}^n$ induced by $\mathscr{P}, E_1, \ldots, E_n$. Then for every $g\colon \{0,1\}^n \to \{-1,+1\}$,*

$$\mathbf{P}[g(E_1, \ldots, E_n)] = \frac{1}{2} - \frac{1}{2} \mathop{\mathbf{E}}_{x \sim \mu}[g(x)].$$

PROOF. We have:

$$\begin{aligned}\mathbf{P}[g(E_1,\dots,E_n)] &= \sum_{x\in\{0,1\}^n} \frac{1-g(x)}{2}\,\mathbf{P}\left[\bigcap_{i:x_i=0}\overline{E_i} \quad \bigcap_{i:x_i=1} E_i\right] \\ &= \sum_{x\in\{0,1\}^n} \frac{1-g(x)}{2}\mu(x) \\ &= \frac{1}{2} - \frac{1}{2}\mathop{\mathbf{E}}_{x\sim\mu}[g(x)].\end{aligned}$$

□

For a set $S \subseteq [n]$, define $\mathrm{AND}_S\colon\{0,1\}^n \to \{-1,+1\}$ by $\mathrm{AND}_S(x) = \bigwedge_{i\in S} x_i$. In particular, $\mathrm{AND}_\varnothing \equiv -1$.

LEMMA 14.21 (Sherstov [199]). *Let $f\colon\{0,1\}^n \to \{-1,+1\}$ be a given function. Then*

$$\Delta(f,k) = \frac{1}{4}\max_{\alpha,\beta}\left\{\mathop{\mathbf{E}}_{x\sim\alpha}[f(x)] - \mathop{\mathbf{E}}_{x\sim\beta}[f(x)]\right\}, \tag{14.15}$$

where the maximum is taken over all probability distributions α, β on $\{0,1\}^n$ such that $\mathbf{E}_{x\sim\alpha}[\mathrm{AND}_S(x)] = \mathbf{E}_{x\sim\beta}[\mathrm{AND}_S(x)]$ for $|S| \leqslant k$.

PROOF. Fix probability spaces $\mathscr{P}_1, \mathscr{P}_2$, events $A_1,\dots,A_n$ in $\mathscr{P}_1$, and events $B_1,\dots,B_n$ in $\mathscr{P}_2$, such that (14.14) holds. Let α and β be the distributions on $\{0,1\}^n$ induced by $\mathscr{P}_1, A_1,\dots,A_n$ and $\mathscr{P}_2, B_1,\dots,B_n$, respectively. Then by Proposition 14.20,

$$\frac{1}{2}\mathop{\mathbf{E}}_{x\sim\alpha}[f(x)] - \frac{1}{2}\mathop{\mathbf{E}}_{x\sim\beta}[f(x)] = \mathop{\mathbf{P}}_{\mathscr{P}_1}[f(A_1,\dots,A_n)] - \mathop{\mathbf{P}}_{\mathscr{P}_2}[f(B_1,\dots,B_n)]$$

and

$$\mathop{\mathbf{E}}_{x\sim\alpha}[\mathrm{AND}_S(x)] = \mathop{\mathbf{E}}_{x\sim\beta}[\mathrm{AND}_S(x)], \qquad |S| \leqslant k.$$

Letting δ stand for the right-hand side of (14.15), we conclude that $\Delta(f,k) \leqslant \delta$.

It remains to show that $\Delta(f,k) \geqslant \delta$. Given a probability distribution μ on $\{0,1\}^n$, there is an obvious discrete probability space $\mathscr{P}$ and events $E_1,\dots,E_n$ in it that induce μ: simply let $\mathscr{P} = \{0,1\}^n$ with E_i defined to be the event that $x_i = 1$, where $x \in \{0,1\}^n$ is distributed according to μ. This allows us to reverse the argument of the previous paragraph (again using Proposition 14.20) and show that $\Delta(f,k) \geqslant \delta$. □

With $\Delta(f,k)$ thus simplified, we are in a position to prove the main result of this section. It is instructive to compare the proof to follow with the proof of Theorem 14.16.

PROOF OF THEOREM 14.18. As α and β in the statement of Lemma 14.21 range over all distributions on $\{0,1\}^n$, the function $\{\alpha(x)-\beta(x)\}/2$ ranges over all functions $\psi\colon\{0,1\}^n \to \mathbb{R}$ with $\|\psi\|_1 \leqslant 1$ and $\hat{\psi}(\varnothing) = 0$. As a result, we can restate Lemma 14.21 as follows:

$$\Delta(f,k) = \frac{1}{2}\max_{\psi}\left\{\sum_{x\in\{0,1\}^n} f(x)\psi(x)\right\},$$

where the maximum is taken over all functions $\psi\colon\{0,1\}^n \to \mathbb{R}$ with $\|\psi\|_1 \leqslant 1$ and $\hat{\psi}(S) = 0$ for $|S| \leqslant k$. By Theorem 4.4, the proof is complete. □

14.8 High-accuracy approximation of symmetric functions

The purpose of this section is to determine the ε-approximate degree for every symmetric function and every $\varepsilon \leqslant 1/3$. This result is of independent interest, given the

role of uniform approximation in this thesis. Our primary application of this result will be to complete the solution of the approximate inclusion-exclusion problem, started in the previous section. Our main theorem on approximation is as follows.

THEOREM 14.22 (Sherstov [199]). *Let $f\colon \{0,1\}^n \to \{-1,+1\}$ be a given function with $f(x) \equiv D(\sum x_i)$ for some predicate $D\colon \{0,1,\ldots,n\} \to \{-1,+1\}$. Let $\varepsilon \in [2^{-n}, 1/3]$ be given. Then*

$$\deg_\varepsilon(f) = \tilde{\Theta}\left(\sqrt{n(\ell_0(D)+\ell_1(D))} + \sqrt{n\log(1/\varepsilon)}\right),$$

where $\ell_0(D) \in \{0,1,\ldots,\lfloor n/2\rfloor\}$ and $\ell_1(D) \in \{0,1,\ldots,\lceil n/2\rceil\}$ are the smallest integers such that D is constant in the range $[\ell_0(D), n-\ell_1(D)]$. Furthermore, the approximating polynomial for each D and ε is given explicitly.

The $\tilde{\Theta}$ notation in the above statement suppresses logarithmic factors. In words, Theorem 14.22 rather fully characterizes the uniform approximation of symmetric Boolean functions. It is a broad generalization of several earlier results in the literature. The first of these is Paturi's Theorem 2.5, which states that

$$\deg_{1/3}(f) = \Theta\left(\sqrt{n(\ell_0(D)+\ell_1(D))}\right)$$

in the notation of Theorem 14.22. Unfortunately, Paturi's result and its proof give no insight into the behavior of the ε-approximate degree for vanishing ε. Another relevant result is due to Kahn et al. [100], who conducted an in-depth study of the case $f = \mathrm{OR}_n$. They showed that

$$\deg_\varepsilon(\mathrm{OR}_n) = \tilde{\Theta}(\sqrt{n\log(1/\varepsilon)}), \qquad 2^{-n} \leqslant \varepsilon \leqslant \frac{1}{3}.$$

Using different techniques, Buhrman et al. [52] gave the final answer for $f = \mathrm{OR}_n$:

$$\deg_\varepsilon(\mathrm{OR}_n) = \Theta(\sqrt{n\log(1/\varepsilon)}), \qquad 2^{-n} \leqslant \varepsilon \leqslant \frac{1}{3}.$$

Thus, our work generalizes the above results to every symmetric function and every $\varepsilon \in [2^{-n}, 1/3]$. Theorem 14.22 has another, more revealing interpretation. In view of Paturi's work, it can be restated as

$$\deg_\varepsilon(f) = \tilde{\Theta}(\deg_{1/3}(f) + \sqrt{n \log(1/\varepsilon)}), \qquad 2^{-n} \leqslant \varepsilon \leqslant \frac{1}{3}, \tag{14.16}$$

where f is any nonconstant symmetric function. In words, past a certain threshold, the dependence of the ε-approximate degree on ε is essentially the same for all nonconstant symmetric functions. This threshold varies from one function to another and equals the degree required for approximation within $1/3$.

In the remainder of this section, we prove Theorem 14.22. We will establish the upper and lower bounds in this result separately, as Lemma 14.26 and Lemma 14.29 below. To simplify notation, we will speak of the approximate degree $\deg_\varepsilon(D)$ of a predicate $D: \{0, 1, \ldots, n\} \to \{-1, +1\}$ rather than the approximate degree of its corresponding symmetric function. Here $\deg_\varepsilon(D)$ is of course defined as the least degree of a univariate polynomial p such that $|D(i) - p(i)| \leqslant \varepsilon$ for $i = 0, 1, \ldots, n$. In view of Proposition 2.2, the ε-approximate degree of a predicate D is equal to the ε-approximate degree of its corresponding symmetric Boolean function $f(x) = D(\sum x_i)$. This justifies our switch from the latter to the former.

Our proofs make heavy use of Chebyshev polynomials, which is not surprising given their fundamental role in approximation. The other key ingredient is interpolation, which here amounts to multiplying an imperfect approximant $p(t)$ by another polynomial $q(t)$ that zeroes out p's mistakes. This interpolation technique is well-known [21, 100] and is vital to exploiting the discrete character of the problem: we are interested in approximation over the discrete set of points $\{0, 1, \ldots, n\}$ rather than the stronger continuous setting, $[0, n]$. Kahn et al. [100], who obtained the special case of Theorem 14.22 for $f = \mathrm{OR}_n$, also used Chebyshev polynomials and interpolation, although in a simpler and different way.

We start by recalling a few properties of Chebyshev polynomials, whose proofs can be found in any standard textbook on approximation theory, e.g., Rivlin [185] and Cheney [60].

FACT 14.23 (Chebyshev polynomials). *The d^{th} Chebyshev polynomial, $T_d(t)$, has degree d and satisfies the following properties:*

$$T_d(1) = 1 \tag{14.17}$$

$$|T_d(t)| \leqslant 1 \qquad (-1 \leqslant t \leqslant 1) \tag{14.18}$$

$$T_d'(t) \geqslant d^2 \qquad (t \geqslant 1) \tag{14.19}$$

$$T_d(1+\delta) \geqslant \tfrac{1}{2} \cdot 2^{d\sqrt{2\delta}} \qquad (0 \leqslant \delta \leqslant 1/2) \tag{14.20}$$

$$2 \leqslant T_{\lceil a \rceil}\left(1 + \tfrac{1}{a^2}\right) \leqslant 7 \qquad (a \geqslant 1) \tag{14.21}$$

At the heart of our construction is the following technical lemma, which gives an efficient method for approximating a given predicate D everywhere except in the vicinity of points where D changes value.

LEMMA 14.24 (Sherstov [199]). *Let $\ell \geqslant 0$, $\Delta \geqslant 1$, and $d \geqslant 1$ be integers with $\ell + \Delta \leqslant n/2$. Then there is an (explicitly given) polynomial $p(t)$ of degree at most $22(d+1)\sqrt{n(\ell+\Delta)}/\Delta$ with*

$$p(n-\ell) = 1$$

and

$$|p(t)| \leqslant 2^{-d}, \qquad t \in [0,n] \setminus (n-\ell-\Delta, n-\ell+\Delta).$$

PROOF. Define

$$p_1(t) = T_{\left\lceil \sqrt{\frac{n-\ell-\Delta}{\ell+\Delta}} \right\rceil}\left(\frac{t}{n-\ell-\Delta}\right).$$

One readily verifies the following properties of p_1:

$$\left.\begin{array}{ll}
p_1([0, n-\ell-\Delta]) \subseteq [-1,1] & \text{by (14.18);} \\
p_1([n-\ell-\Delta, n]) \subseteq [1,7] & \text{by (14.17), (14.19), (14.21);} \\
p_1'(t) \geqslant \dfrac{1}{\ell+\Delta} \text{ for } t \geqslant n-\ell-\Delta & \text{by (14.19);} \\
p_1(n-\ell) - p_1(n-\ell-\Delta) \geqslant \dfrac{\Delta}{\ell+\Delta} & \text{by previous line;} \\
p_1(n-\ell+\Delta) - p_1(n-\ell) \geqslant \dfrac{\Delta}{\ell+\Delta} & \text{likewise.}
\end{array}\right\} \tag{14.22}$$

Now consider the polynomial defined by

$$p_2(t) = \left(\frac{p_1(t) - p_1(n-\ell)}{8}\right)^2.$$

In view of (14.22), this new polynomial satisfies

$$p_2(n-\ell) = 0$$

and

$$p_2(t) \in \left[\frac{\Delta^2}{64(\ell+\Delta)^2}, 1\right], \qquad t \in [0,n] \setminus (n-\ell-\Delta, n-\ell+\Delta).$$

Finally, define

$$p_3(t) = T_{\left\lceil \frac{8(d+1)(\ell+\Delta)}{\sqrt{2}\Delta} \right\rceil}\left(1 + \frac{\Delta^2}{64(\ell+\Delta)^2} - p_2(t)\right).$$

Using (14.20) and the properties of p_2, one sees that $p(t) = p_3(t)/p_3(n-\ell)$ is the desired polynomial. □

There are a large number of distinct predicates on $\{0, 1, \ldots, n\}$. To simplify the analysis, we would like to work with a small family of functions that have simple structure yet allow us to efficiently express any other predicate. A natural choice is the family $\mathrm{EXACT}_\ell \colon \{0, 1, \ldots, n\} \to \{0, 1\}$ for $\ell = 0, 1, \ldots, n$, where

$$\mathrm{EXACT}_\ell(t) = \begin{cases} 1 & \text{if } t = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

For a moment, we shall focus on an explicit construction for EXACT_ℓ.

LEMMA 14.25 (Sherstov [199]). *Let* $0 \leqslant \ell \leqslant n/2$. *Then for* $\varepsilon \leqslant 1/3$,

$$\begin{aligned} \deg_\varepsilon(\mathrm{EXACT}_\ell) &= \deg_\varepsilon(\mathrm{EXACT}_{n-\ell}) \\ &= O\left(\sqrt{n(\ell+1)}\log n + \sqrt{n\log(1/\varepsilon)\log n}\right). \end{aligned}$$

PROOF. The first equality in the statement of the lemma is obvious, and we concentrate on the second. We may assume that $\ell \leqslant n/\log^2 n$ and $\log(1/\varepsilon) \leqslant n/\log n$, since otherwise the claim is trivial. Set

$$\Delta = \left\lceil \frac{\log(1/\varepsilon)}{\log n} \right\rceil, \qquad d = 3\Delta \lceil \log n \rceil.$$

Our assumptions about ℓ and ε imply that $\ell + \Delta \ll n/2$, and thus Lemma 14.24 is applicable. Denote by $p(t)$ the polynomial constructed in Lemma 14.24. Define

$$q(t) = \prod_{\substack{i=-(\Delta-1),\ldots,(\Delta-1) \\ i \neq 0}} (t - (n - \ell + i)).$$

We claim that the polynomial given by

$$r(t) = \frac{1}{q(n-\ell)} \cdot p(t)q(t)$$

is the sought approximation to $\mathrm{EXACT}_{n-\ell}$. Indeed, it is easy to verify that $r(t)$ has the desired degree. For $t \in \{0, 1, \ldots, n\} \setminus \{n-\ell-(\Delta-1), \ldots, n-\ell+(\Delta-1)\}$,

$$|r(t) - \mathrm{EXACT}_{n-\ell}(t)| = |r(t)| \leqslant n^{2(\Delta-1)} \cdot \frac{1}{2^d} \leqslant \varepsilon.$$

Since $r(t) = \mathrm{EXACT}_{n-\ell}(t)$ for all remaining t, the proof is complete. □

We are now in a position to prove the sought upper bound on the approximate degree of any predicate.

LEMMA 14.26 (Sherstov [199]). *Let $D\colon \{0, 1, \ldots, n\} \to \{-1, +1\}$ be a given predicate. Then for $\varepsilon \leqslant 1/3$,*

$$\deg_\varepsilon(D) \leqslant O\left(\sqrt{n(\ell_0(D) + \ell_1(D))}\,\log n + \sqrt{n\log(1/\varepsilon)\log n}\right).$$

Moreover, the approximating polynomial is given explicitly.

PROOF. Without loss of generality, we can assume that $D(\lceil n/2 \rceil) = 1$ (otherwise, work with the negation of D). For $\ell = 0, 1, \ldots, n$, let $p_\ell(t)$ denote the polynomial that approximates $\mathrm{EXACT}_\ell(t)$ pointwise to within $\varepsilon/(2n)$, as constructed in Lemma 14.25. Put

$$p(t) = 1 - 2 \sum_{\ell\,:\,D(\ell)=-1} p_\ell(t).$$

Then clearly $p(t)$ approximates D pointwise to within ε. It remains to place an upper bound on the degree of p:

$$\begin{aligned}
\deg_\varepsilon(D) &\leqslant \deg p \\
&\leqslant \max_{\substack{\ell\,:\,D(\ell)=-1,\\ \ell<\lceil n/2\rceil}} \{\deg p_\ell\} \quad + \quad \max_{\substack{\ell\,:\,D(\ell)=-1,\\ \ell>\lceil n/2\rceil}} \{\deg p_{n-\ell}\} \\
&\leqslant O\left(\left(\sqrt{n\ell_0(D)} + \sqrt{n\ell_1(D)}\right)\log n + \sqrt{n\log(n/\varepsilon)\log n}\right) \\
&\leqslant O\left(\sqrt{n(\ell_0(D)+\ell_1(D))}\log n + \sqrt{n\log(1/\varepsilon)\log n}\right),
\end{aligned}$$

where the third inequality follows by Lemma 14.25. □

We now turn to the matching lower bounds on the approximate degree of predicates, which are substantially easier to obtain. Our proof uses a reduction to the function EXACT_0, for which Kahn et al. [100] have already obtained a near-tight lower bound.

THEOREM 14.27 (Kahn et al. [100]). *For every polynomial p of degree $k < n$,*

$$\max_{i=0,1,\ldots,n} |\mathrm{EXACT}_0(i) - p(i)| \geqslant n^{-\Theta(k^2/n)}.$$

COROLLARY 14.28. *Let ε be given with $2^{-\Theta(n\log n)} \leqslant \varepsilon \leqslant 1/3$. Then*

$$\deg_\varepsilon(\mathrm{EXACT}_0) \geqslant \Omega\left(\sqrt{\frac{n\log(1/\varepsilon)}{\log n}}\right).$$

We are now in a position to prove the desired lower bound on the approximate degree of any given predicate.

LEMMA 14.29 (Sherstov [199]). *Let* $D\colon\{0,1,\ldots,n\} \to \{-1,+1\}$ *be a nonconstant predicate. Then for each* ε *with* $2^{-\Theta(n\log n)} \leqslant \varepsilon \leqslant 1/3$,

$$\deg_\varepsilon(D) \geqslant \Omega\left(\sqrt{n(\ell_0(D)+\ell_1(D))} + \sqrt{\frac{n\log(1/\varepsilon)}{\log n}}\right).$$

PROOF. In view of Theorem 2.5, it suffices to show that

$$\deg_\varepsilon(D) \geqslant \Omega\left(\sqrt{\frac{n\log(1/\varepsilon)}{\log n}}\right). \tag{14.23}$$

Abbreviate $\ell = \ell_0(D)$ and assume without loss of generality that $\ell \geqslant 1$ (otherwise work with $\ell = \ell_1(D)$). We can additionally assume that $\ell \leqslant n/5$ since otherwise the claim follows trivially from Theorem 2.5. Consider the function $\mathrm{EXACT}_0\colon\{0,1,\ldots,\lfloor n/5\rfloor\} \to \{0,1\}$. By Corollary 14.28,

$$\deg_\varepsilon(\mathrm{EXACT}_0) \geqslant \Omega\left(\sqrt{\frac{n\log(1/\varepsilon)}{\log n}}\right) \tag{14.24}$$

On the other hand,

$$\mathrm{EXACT}_0(t) = \frac{1}{2} - \frac{1}{2}D(\ell)D(t+\ell-1),$$

so that

$$\deg_\varepsilon(\mathrm{EXACT}_0) \leqslant \deg_\varepsilon(D). \tag{14.25}$$

Equations (14.24) and (14.25) imply (14.23), thereby completing the proof. □

At this point, we have proved the lower and upper bounds in Theorem 14.22. We will now combine this theorem with our work in the previous section to solve the inclusion-exclusion problem for all symmetric functions.

THEOREM 14.30 (Sherstov [199]). *Let $f\colon\{0,1\}^n \to \{-1,+1\}$ be a nonconstant function with $f(x) \equiv D(\sum x_i)$ for some predicate $D\colon\{0,1,\ldots,n\} \to \{-1,+1\}$. Let $\ell = \ell_0(D) + \ell_1(D)$, where $\ell_0(D)$ and $\ell_1(D)$ are as defined in Theorem 14.22. Then*

$$\begin{aligned}
&\Delta(f,k) = \Theta(1) && \text{if} \quad k \leqslant \Theta(\sqrt{n\ell}),\\
&\Delta(f,k) \in \left[2^{-\Theta\left(\frac{k^2\log n}{n}\right)}, 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}\right] && \text{if} \quad \Theta(\sqrt{n\ell}\log n) \leqslant k \leqslant \Theta(n).
\end{aligned}$$

Furthermore, for every $k \geqslant \Theta(\sqrt{n\ell}\log n)$, there are reals $a_0, a_1, \ldots, a_k$, computable in time polynomial in n, such that

$$\left|\, \mathbf{P}[f(A_1,\ldots,A_n)] \quad - \quad \sum_{j=0}^{k} a_j \sum_{S:|S|=j} \mathbf{P}\left[\bigcap_{i\in S} A_i\right] \,\right| \leqslant 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}$$

for any events $A_1,\ldots,A_n$ in any probability space $\mathscr{P}$.

PROOF. By Proposition 2.2, Theorem 2.5, and Lemmas 14.26 and 14.29,

$$E(f,k) \in \begin{cases} \Theta(1) & \text{if} \quad k \leqslant \Theta(\sqrt{n\ell}),\\ \left[2^{-\Theta\left(\frac{k^2\log n}{n}\right)}, 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}\right] & \text{if} \quad \Theta(\sqrt{n\ell}\log n) \leqslant k \leqslant \Theta(n). \end{cases}$$

In view of Theorem 14.18, this proves the claim regarding $\Delta(f,k)$. We now turn to the claim regarding $a_0, a_1, \ldots, a_k$. For $k \geqslant \Theta(\sqrt{n\ell}\log n)$, Lemma 14.26 gives an

explicit univariate polynomial $p \in P_k$ such that

$$|f(x) - p(x_1 + \cdots + x_n)| \leqslant 2^{-\Theta\left(\frac{k^2}{n \log n}\right)}, \qquad x \in \{0,1\}^n. \tag{14.26}$$

Fix a probability space $\mathscr{P}$ and events $A_1, \ldots, A_n$ in it. Let μ be the distribution on $\{0,1\}^n$ induced by $\mathscr{P}, A_1, \ldots, A_n$. We claim that the quantity

$$\frac{1}{2} - \frac{1}{2} \mathop{\mathbf{E}}_{x \sim \mu} [p(x_1 + \cdots + x_n)]$$

is the desired approximant of $\mathbf{P}[f(A_1, \ldots, A_n)]$. Indeed,

$$\begin{aligned}
\frac{1}{2} - \frac{1}{2} \mathop{\mathbf{E}}_{x \sim \mu} [p(x_1 + \cdots + x_n)] &= \mathop{\mathbf{E}}_{x \sim \mu} \left[\sum_{j=0}^{k} a_j \sum_{|S|=j} \prod_{i \in S} x_i \right] \\
&= \sum_{j=0}^{k} a_j \sum_{|S|=j} \mathop{\mathbf{E}}_{x \sim \mu} \left[\prod_{i \in S} x_i \right] \\
&= \sum_{j=0}^{k} a_j \sum_{|S|=j} \mathbf{P} \left[\bigcap_{i \in S} A_i \right],
\end{aligned}$$

where the reals $a_0, a_1, \ldots, a_k$ are uniquely determined by the polynomial p, itself explicitly given. It is also clear that $a_0, a_1, \ldots, a_k$ can be computed from the coefficients of p in time polynomial in n. Therefore, the quantity $\mathbf{E}_{x \sim \mu}[p(\sum x_i)]$ has the desired representation. It remains to verify that it approximates $\mathbf{P}[f(A_1, \ldots, A_n)]$ as claimed:

$$\begin{aligned}
&\left| \mathbf{P}[f(A_1, \ldots, A_n)] - \frac{1}{2} + \frac{1}{2} \mathop{\mathbf{E}}_{x \sim \mu} [p(x_1 + \cdots + x_n)] \right| \\
&\qquad = \left| \frac{1}{2} \mathop{\mathbf{E}}_{x \sim \mu} [f(x) - p(x_1 + \cdots + x_n)] \right| \leqslant 2^{-\Theta\left(\frac{k^2}{n \log n}\right)},
\end{aligned}$$

where the equality holds by Proposition 14.20 and the inequality by (14.26). □

REMARK 14.31. Equation (14.16) in this section determines the ε-approximate degree of every symmetric function to within logarithmic factors. In a recent follow-up result, de Wolf [221] improved our bounds to a tight answer:

$$\deg_\varepsilon(f) = \Theta\left(\deg_{1/3}(f) + \sqrt{n\log(1/\varepsilon)}\right)$$

for every nonconstant symmetric f and every $\varepsilon \in [2^{-n}, 1/3]$. By Theorem 14.18, this automatically leads to sharper bounds for the inclusion-exclusion problem. De Wolf's argument, short and elegant, is based on quantum query complexity.

Chapter 15

Lower Bounds for Sign-Representation

In this chapter, we will take an in-depth look at the sign-representation of Boolean functions by real polynomials. We will prove that, for any Boolean functions f and g, the intersection $f(x) \wedge g(y)$ has threshold degree $O(d)$ if and only if there exist rational functions F, G of degree $O(d)$ with $\|f - F\|_\infty + \|g - G\|_\infty < 1$. This characterization extends to conjunctions of three and more functions as well as various other compositions. This result is of interest because of the applications of the threshold degree in previous chapters of this thesis and in earlier literature. As a concrete application in the next chapter, we will solve an open problem in learning theory, due to Klivans [120], on the threshold degree of the intersection of two halfspaces.

15.1 Introduction

Recall that the threshold degree of a function $f : X \to \{-1, +1\}$, for some finite set $X \subset \mathbb{R}^n$, is the least degree of a real polynomial with $f(x) \equiv \operatorname{sgn} p(x)$. In words, the threshold degree $\deg_\pm(f)$ is the least degree of a real polynomial that represents f in sign. The formal study of this complexity measure and of sign-representations in general began in 1969 with the seminal monograph of Minsky and Papert [153], where the threshold degree was analyzed for several common functions. Since then, the notion of threshold degree has found a variety of applications. Paturi and Saks [166] and later Siu et al. [210] used Boolean functions with high threshold degree to obtain size-depth trade-offs for threshold circuits. The well-known result, due to Beigel et al. [33], that PP is closed under intersection is also naturally interpreted in terms of threshold degree. In another development, Aspnes et al. [21] used the notion of threshold degree and its relaxations to obtain oracle separations for PP and to give an insightful new proof of classical lower bounds for AC^0. Krause and Pudlák [132, 133] used random restrictions to show that the threshold degree gives lower bounds on the weight and density of perceptrons and their generalizations, which are well-studied computational models. A variety of other applications of the threshold degree are discovered in this thesis, including discrepancy bounds (Chapter 4), unbounded-error communication complexity (Chapters 7 and 8), and approximate rank (Chapter 14).

Apart from complexity theory, the threshold degree of Boolean functions is of interest in computational learning. In this context, low threshold degree trans-

lates into efficient learnability [122, 121]. Specifically, functions with low threshold degree can be efficiently PAC-learned under arbitrary distributions via linear programming. An illustrative example is the current fastest algorithm for PAC-learning polynomial-size DNF formulas, due to Klivans and Servedio [122], which is based precisely on an upper bound on the threshold degree of this concept class. Klivans et al. [121] showed that intersections of light halfspaces also have low threshold degree, thereby giving an efficient PAC algorithm for this class as well.

Despite the role of sign-representations in learning theory and complexity theory, progress in understanding the threshold degree has been slow and difficult [153, 21, 163, 164]. The main contribution of this chapter is a strong new technique for estimating the threshold degree. To set the stage for our results, consider the special but illustrative case of the conjunction of two functions. In other words, we are given functions $f: X \to \{-1, +1\}$ and $g: Y \to \{-1, +1\}$ for some finite sets $X, Y \subset \mathbb{R}^n$ and would like to determine the threshold degree of their conjunction, $(f \wedge g)(x, y) = f(x) \wedge g(y)$. A simple and elegant method for sign-representing $f \wedge g$, due to Beigel et al. [33], is to use rational approximation. Specifically, let $p_1(x)/q_1(x)$ and $p_2(y)/q_2(y)$ be rational functions of degree d that approximate f and g, respectively, in the following sense:

$$\max_{x \in X} \left| f(x) - \frac{p_1(x)}{q_1(x)} \right| + \max_{y \in Y} \left| g(y) - \frac{p_2(y)}{q_2(y)} \right| < 1. \tag{15.1}$$

Then

$$f(x) \wedge g(y) \equiv \operatorname{sgn}\{1 + f(x) + g(y)\} \equiv \operatorname{sgn}\left\{1 + \frac{p_1(x)}{q_1(x)} + \frac{p_2(y)}{q_2(y)}\right\}. \tag{15.2}$$

Multiplying the last expression in braces by the positive quantity $q_1(x)^2 q_2(y)^2$ gives

$$f(x) \wedge g(y) \equiv \operatorname{sgn}\left\{q_1(x)^2 q_2(y)^2 + p_1(x) q_1(x) q_2(y)^2 + p_2(y) q_1(x)^2 q_2(y)\right\},$$

whence $\deg_{\pm}(f \wedge g) \leqslant 4d$. In summary, if f and g can be approximated as in (15.1) by rational functions of degree at most d, then the conjunction $f \wedge g$ has threshold degree at most $4d$.

A natural question to ask is whether there exists a better construction. After all, given a sign-representing polynomial $p(x, y)$ for $f(x) \wedge g(y)$, there is no reason to expect that p arises from the sum of two independent rational functions as in (15.2). Indeed, x and y can be tightly coupled inside $p(x, y)$ and can interact in complicated ways. In Section 15.4 we prove that, surprisingly, no such interactions can beat the simple construction above. In other words, the sign-representation based on rational functions always achieves the optimal degree, up to a small constant factor:

THEOREM 15.1 (Sherstov [200]). *Let $f: X \to \{-1, +1\}$ and $g: Y \to \{-1, +1\}$ be given functions, where $X, Y \subset \mathbb{R}^n$ are arbitrary finite sets. Assume that f and g are not identically false. Let $d = \deg_{\pm}(f \wedge g)$. Then there exist degree-$4d$ rational functions*

$$\frac{p_1(x)}{q_1(x)}, \quad \frac{p_2(y)}{q_2(y)}$$

that satisfy (15.1).

Via repeated applications of this theorem, we obtain in Section 15.5 analogous results for conjunctions $f_1 \wedge f_2 \wedge \cdots \wedge f_k$ of any Boolean functions $f_1, f_2, \ldots, f_k$ and any k. We then further extend our results to compositions $F(f_1, \ldots, f_k)$ for various F other than $F = \mathrm{AND}_k$, such as halfspaces and read-once AND/OR/NOT formulas.

Previously, it was a substantial challenge to analyze the threshold degree even for compositions of the form $f \wedge g$. Indeed, we are only aware of the work in [153, 163], where the threshold degree of $f \wedge g$ was studied for the special case $f = g = \mathrm{MAJ}_n$. The main difficulty in those previous works was analyzing the unintuitive interactions between f and g. Our results remove this difficulty, even in the general setting of compositions $F(f_1, f_2, \ldots, f_k)$ for arbitrary $f_1, f_2, \ldots, f_k$ and various combining functions F. In other words, one can study the base func-

tions $f_1, f_2, \ldots, f_k$ individually, in isolation; once their rational approximability is understood, one immediately obtains lower bounds on the threshold degree of $F(f_1, f_2, \ldots, f_k)$. As an application of Theorem 15.1 in the next chapter, we will construct two halfspaces on $\{0,1\}^n$ with threshold degree $\Omega(\sqrt{n})$, improving exponentially on previous work and solving an open problem due to Klivans [120].

15.2 Background and definitions

Throughout this chapter, the symbol t refers to a real variable, whereas u, v, w, x, y, z refer to vectors in $\mathbb{R}^n$ and in particular in $\{0,1\}^n$. For a subset $X \subseteq \mathbb{R}^n$, we adopt the notation $-X = \{-x : x \in X\}$. We say that a set $X \subseteq \mathbb{R}^n$ is *closed under negation* if $X = -X$. Given a function $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}^n$, we say that f is *odd* (respectively, *even*) if X is closed under negation and $f(-x) = -f(x)$ for all $x \in X$ (respectively, $f(-x) = f(x)$ for all $x \in X$).

Given functions $f : X \to \{-1,+1\}$ and $g : Y \to \{-1,+1\}$, recall that the function $f \wedge g : X \times Y \to \{-1,+1\}$ is given by $(f \wedge g)(x,y) = f(x) \wedge g(y)$. The function $f \vee g$ is defined analogously. Observe that in this notation, $f \wedge f$ and f are completely different functions, the former having domain $X \times X$ and the latter X. These conventions extend in the obvious way to any number of functions. For example, $f_1 \wedge f_2 \wedge \cdots \wedge f_k$ is a Boolean function with domain $X_1 \times X_2 \times \cdots \times X_k$, where X_i is the domain of f_i. Generalizing further, we let the symbol $F(f_1, \ldots, f_k)$ denote the Boolean function on $X_1 \times X_2 \times \cdots \times X_k$ obtained by composing a given function $F : \{-1,+1\}^k \to \{-1,+1\}$ with the functions $f_1, f_2, \ldots, f_k$.

Finally, we establish some notation for rational approximation. Consider a function $f : X \to \{-1,+1\}$, where $X \subseteq \mathbb{R}^n$ is an arbitrary set. For $d \geqslant 0$, we define

$$R(f,d) = \inf_{p,q} \sup_{x \in X} \left| f(x) - \frac{p(x)}{q(x)} \right|,$$

where the infimum is over multivariate polynomials p and q of degree up to d such that q does not vanish on X. In words, $R(f,d)$ is the least error in an approximation of f by a multivariate rational function of degree up to d. We will also take an

interest in the quantity

$$R^+(f,d) \;=\; \inf_{p,q}\,\sup_{x\in X}\left|f(x)-\frac{p(x)}{q(x)}\right|,$$

where the infimum is over multivariate polynomials p and q of degree up to d such that q is positive on X. These two quantities are related as follows:

$$R^+(f,2d) \leqslant R(f,d) \leqslant R^+(f,d). \tag{15.3}$$

The second inequality here is trivial. The first follows from the fact that every rational approximant $p(x)/q(x)$ of degree d gives rise to a degree-$2d$ rational approximant with the same error and a positive denominator, namely, $\{p(x)q(x)\}/q(x)^2$.

The infimum in the definitions of $R(f,d)$ and $R^+(f,d)$ cannot in general be replaced by a minimum [185], even when X is finite subset of $\mathbb{R}$. This is in contrast to the more familiar setting of a finite-dimensional normed linear space, where least-error approximants are guaranteed to exist.

15.3 Auxiliary results on uniform approximation

In this section, we prove a number of auxiliary facts about uniform approximation and its limiting case, sign-representation. This preparatory work will set the stage for the proofs of our main results in later sections. We start by spelling out the exact relationship between the rational approximation and sign-representation of a Boolean function.

THEOREM 15.2 (Sherstov [200]). *Let $f\colon X \to \{-1,+1\}$ be a given function, where $X \subset \mathbb{R}^n$ is compact. Then for every integer d,*

$$\deg_{\pm}(f) \leqslant d \quad\Leftrightarrow\quad R^+(f,d) < 1.$$

PROOF. For the forward implication, let p be a polynomial of degree at most d such that $f(x)p(x) > 0$ for every $x \in X$. Letting $M = \max_{x\in X} |p(x)|$ and $m = \min_{x\in X} |p(x)|$, we have

$$R^+(f,d) \leqslant \sup_{x\in X} \left| f(x) - \frac{p(x)}{M} \right| \leqslant 1 - \frac{m}{M} < 1.$$

For the converse, fix a degree-d rational function $p(x)/q(x)$ with $q(x) > 0$ on X and $\sup_X |f(x) - \{p(x)/q(x)\}| < 1$. Then clearly $f(x)p(x) > 0$ on X. □

Our next observation amounts to reformulating the rational approximation of Boolean functions in a way that is more analytically pleasing.

THEOREM 15.3 (Sherstov [200]). *Let $f: X \to \{-1,+1\}$ be a given function, where $X \subset \mathbb{R}^n$ is compact and $\deg_\pm(f) < \infty$. Then for every integer $d \geqslant \deg_\pm(f)$, one has*

$$R^+(f,d) = \inf_{c\geqslant 1} \frac{c^2-1}{c^2+1},$$

where the infimum is over all $c \geqslant 1$ for which there exist polynomials p, q of degree up to d such that $0 < \frac{1}{c}q(x) \leqslant f(x)p(x) \leqslant cq(x)$ on X.

PROOF. In view of Theorem 15.2, the quantity $R^+(f,d)$ is the infimum over all $\varepsilon < 1$ for which there exist polynomials p and q of degree up to d such that $0 < (1-\varepsilon)q(x) \leqslant f(x)p(x) \leqslant (1+\varepsilon)q(x)$ on X. Equivalently, one may require that

$$0 < \frac{1-\varepsilon}{\sqrt{1-\varepsilon^2}}\, q(x) \leqslant f(x)p(x) \leqslant \frac{1+\varepsilon}{\sqrt{1-\varepsilon^2}}\, q(x).$$

Letting $c = c(\varepsilon) = \sqrt{(1+\varepsilon)/(1-\varepsilon)}$, the theorem follows. □

Our next result shows that if a degree-d rational approximant achieves error ε in approximating a given Boolean function, then a degree-$2d$ approximant can achieve error as small as ε^2.

THEOREM 15.4 (Sherstov [200]). *Let $f: X \to \{-1, +1\}$ be a given function, where $X \subseteq \mathbb{R}^n$. Let d be a given integer. Then*

$$R^+(f, 2d) \leqslant \left(\frac{\varepsilon}{1 + \sqrt{1 - \varepsilon^2}} \right)^2,$$

where $\varepsilon = R(f, d)$.

PROOF. The theorem is clearly true for $\varepsilon = 1$. For $0 \leqslant \varepsilon < 1$, consider the univariate rational function

$$S(t) = \frac{4\sqrt{1 - \varepsilon^2}}{1 + \sqrt{1 - \varepsilon^2}} \cdot \frac{t}{t^2 + (1 - \varepsilon^2)}.$$

Calculus shows that

$$\max_{1-\varepsilon \leqslant |t| \leqslant 1+\varepsilon} |\operatorname{sgn} t - S(t)| = \left(\frac{\varepsilon}{1 + \sqrt{1 - \varepsilon^2}} \right)^2.$$

Fix a sequence $A_1, A_2, \ldots$ of rational functions of degree at most d such that $\sup_{x \in X} |f(x) - A_m(x)| \to \varepsilon$ as $m \to \infty$. Then $S(A_1(x)), S(A_2(x)), \ldots$ is the sought sequence of approximants to f, each a rational function of degree at most $2d$ with a positive denominator. □

COROLLARY 15.5 (Sherstov [200]). *Let $f: X \to \{-1, +1\}$ be a given function, where $X \subseteq \mathbb{R}^n$. Then for all integers $d \geqslant 1$ and reals $t \geqslant 2$,*

$$R^+(f, td) \leqslant R(f, d)^{t/2}.$$

PROOF. If $t = 2^k$ for some integer $k \geqslant 1$, then repeated applications of Theorem 15.4 yield $R^+(f, 2^k d) \leqslant R(f, 2^{k-1}d)^2 \leqslant \cdots \leqslant R(f,d)^{2^k}$. The general case follows because $2^{\lfloor \log t \rfloor} \geqslant t/2$. □

15.4 Threshold degree of conjunctions of functions

In this section, we prove our main results on conjunctions of Boolean functions. Recall that a key challenge will be, given a sign-representation $\phi(x, y)$ of a composite function $f(x) \wedge g(y)$, to suitably break down ϕ and recover individual rational approximants of f and g. We now present an ingredient of our solution, namely, a certain fact about pairs of matrices based on Farkas' Lemma. For the time being, we will formulate this fact in a clean and abstract way.

THEOREM 15.6 (Sherstov [200]). *Fix matrices $A, B \in \mathbb{R}^{m \times n}$ and a real $c \geqslant 1$. Consider the following system of linear inequalities in $u, v \in \mathbb{R}^n$:*

$$\left.\begin{aligned} \frac{1}{c} Au \leqslant Bv \leqslant cAu, \\ u \geqslant 0, \\ v \geqslant 0. \end{aligned}\right\} \tag{15.4}$$

If $u = v = 0$ is the only solution to (15.4), then there exist vectors $w \geqslant 0$ and $z \geqslant 0$ such that

$$w^{\mathsf{T}} A + z^{\mathsf{T}} B > c(z^{\mathsf{T}} A + w^{\mathsf{T}} B).$$

PROOF. If $u = v = 0$ is the only solution to (15.4), then linear programming duality implies the existence of vectors $w \geqslant 0$ and $z \geqslant 0$ such that $w^{\mathsf{T}} A > cz^{\mathsf{T}} A$ and $z^{\mathsf{T}} B > cw^{\mathsf{T}} B$. Adding the last two inequalities completes the proof. □

For clarity of exposition, we first prove the main result of this section for the case of *two* Boolean functions at least one of which is *odd*. While this case seems restricted, we will see that it captures the full complexity of the problem.

THEOREM 15.7 (Sherstov [200]). *Let $f: X \to \{-1,+1\}$ and $g: Y \to \{-1,+1\}$ be given functions, where $X, Y \subset \mathbb{R}^n$ are arbitrary finite sets. Assume that $f \not\equiv 1$ and $g \not\equiv 1$. Let $d = \deg_{\pm}(f \wedge g)$. If f is odd, then*

$$R^+(f, 2d) + R^+(g, d) < 1.$$

PROOF. We first collect some basic observations. Since $f \not\equiv 1$ and $g \not\equiv 1$, we have $\deg_{\pm}(f) \leqslant d$ and $\deg_{\pm}(g) \leqslant d$. Therefore, Theorem 15.2 implies that

$$R^+(f, d) < 1, \qquad R^+(g, d) < 1. \tag{15.5}$$

In particular, the theorem holds if $R^+(g, d) = 0$. In the remainder of the proof, we assume that $R^+(g, d) = \varepsilon$, where $0 < \varepsilon < 1$.

By hypothesis, there exists a degree-d polynomial ϕ with the property that $f(x) \wedge g(y) = \operatorname{sgn} \phi(x, y)$ for all $x \in X$, $y \in Y$. Define

$$X^- = \{x \in X : f(x) = -1\}.$$

Since X is closed under negation and f is odd, we have $f(x) = 1 \Leftrightarrow -x \in X^-$. We will make several uses of this fact in what follows, without further mention.

Put

$$c = \sqrt{\frac{1 + (1-\delta)\varepsilon}{1 - (1-\delta)\varepsilon}},$$

where $\delta \in (0, 1)$ is sufficiently small. Since $R^+(g, d) > (c^2 - 1)/(c^2 + 1)$, we know by Theorem 15.3 that there cannot exist polynomials p, q of degree up to d such that

$$0 < \frac{1}{c} q(y) \leqslant g(y) p(y) \leqslant c q(y), \qquad y \in Y. \tag{15.6}$$

We claim, then, that there cannot exist reals $a_x \geqslant 0$, $x \in X$, not all zero, such that

$$\frac{1}{c} \sum_{x \in X^-} a_{-x} \phi(-x, y) \leqslant g(y) \sum_{x \in X^-} a_x \phi(x, y) \leqslant c \sum_{x \in X^-} a_{-x} \phi(-x, y), \quad y \in Y.$$

Indeed, if such reals a_x were to exist, then (15.6) would hold for the polynomials $p(y) = \sum_{x \in X^-} a_x \phi(x, y)$ and $q(y) = \sum_{x \in X^-} a_{-x} \phi(-x, y)$. In view of the nonexistence of the a_x, Theorem 15.6 applies to the matrices

$$\Big[\phi(-x, y)\Big]_{y \in Y,\, x \in X^-}, \qquad \Big[g(y)\phi(x, y)\Big]_{y \in Y,\, x \in X^-}$$

and guarantees the existence of nonnegative reals λ_y, μ_y for $y \in Y$ such that

$$\begin{aligned} &\sum_{y \in Y} \lambda_y \phi(-x, y) + \sum_{y \in Y} \mu_y g(y) \phi(x, y) \\ &\qquad > c \left(\sum_{y \in Y} \mu_y \phi(-x, y) + \sum_{y \in Y} \lambda_y g(y) \phi(x, y) \right), \qquad x \in X^-. \end{aligned} \tag{15.7}$$

Define polynomials α, β on X by

$$\begin{aligned} \alpha(x) &= \sum_{y \in g^{-1}(-1)} \{\lambda_y \phi(-x, y) - \mu_y \phi(x, y)\}, \\ \beta(x) &= \sum_{y \in g^{-1}(1)} \{\lambda_y \phi(-x, y) + \mu_y \phi(x, y)\}. \end{aligned}$$

Then (15.7) can be restated as

$$\alpha(x) + \beta(x) > c\{-\alpha(-x) + \beta(-x)\}, \qquad x \in X^-.$$

Both members of this inequality are nonnegative, and thus $\{\alpha(x) + \beta(x)\}^2 > c^2\{-\alpha(-x) + \beta(-x)\}^2$ for $x \in X^-$. Since in addition $\alpha(-x) \leqslant 0$ and $\beta(-x) \geqslant 0$ for $x \in X^-$, we have

$$\{\alpha(x) + \beta(x)\}^2 > c^2\{\alpha(-x) + \beta(-x)\}^2, \qquad x \in X^-.$$

Letting $\gamma(x) = \{\alpha(x) + \beta(x)\}^2$, we see that

$$R^+(f, 2d) \leqslant \max_{x \in X} \left| f(x) - \frac{c^2 + 1}{c^2} \cdot \frac{\gamma(-x) - \gamma(x)}{\gamma(-x) + \gamma(x)} \right| \leqslant \frac{1}{c^2} < 1 - \varepsilon,$$

where the final inequality holds for all $\delta \in (0, 1)$ small enough. □

REMARK 15.8 (Sherstov [200]). In Theorem 15.7 and elsewhere in this thesis, the degree of a multivariate polynomial $p(x_1, x_2, \ldots, x_n)$ is defined as the greatest total degree of any monomial of p. A related notion is the *partial degree* of p, which is the maximum degree of p in any one of the variables $x_1, x_2, \ldots, x_n$. One readily sees that the proof of Theorem 15.7 applies unchanged to this alternate notion. Specifically, if the conjunction $f(x) \wedge g(y)$ can be sign-represented by a polynomial of partial degree d, then there exist rational functions $F(x)$ and $G(y)$ of partial degree $2d$ such that $\|f - F\|_\infty + \|g - G\|_\infty < 1$. In the same way, the program of Section 15.5 carries over, with cosmetic changes, to the notion of partial degree. Analogously, our proofs apply to hybrid definitions of degree, such as partial degree over blocks of variables. Other, more abstract notions of degree can also be handled. In the remainder of the chapter, we will maintain our focus on total degree and will not elaborate further on its generalizations.

As promised, we will now remove the assumption, made in Theorem 15.7, about one of the functions being odd. The result that we are about to prove settles Theorem 15.1 from the introduction.

THEOREM 15.9 (Sherstov [200]). *Let $f: X \to \{-1, +1\}$ and $g: Y \to \{-1, +1\}$ be given functions, where $X, Y \subset \mathbb{R}^n$ are arbitrary finite sets. Assume that $f \not\equiv 1$*

and $g \not\equiv 1$. *Let* $d = \deg_{\pm}(f \wedge g)$. *Then*

$$R^+(f, 4d) + R^+(g, 2d) < 1 \tag{15.8}$$

and, by symmetry,

$$R^+(f, 2d) + R^+(g, 4d) < 1.$$

PROOF. It suffices to prove (15.8). Define $X' \subset \mathbb{R}^{n+1}$ by $X' = \{(x, 1), (-x, -1) : x \in X\}$. It is clear that X' is closed under negation. Let $f' \colon X' \to \{-1, +1\}$ be the odd Boolean function given by

$$f'(x, b) = \begin{cases} f(x), & b = 1, \\ -f(-x), & b = -1. \end{cases}$$

Let ϕ be a polynomial of degree no greater than d such that $f(x) \wedge g(y) \equiv \operatorname{sgn} \phi(x, y)$. Fix an input $\tilde{x} \in X$ such that $f(\tilde{x}) = -1$. Then $f'(x, b) \wedge g(y) \equiv \operatorname{sgn} \{K(1 + b)\phi(x, y) + \phi(-x, y)\phi(\tilde{x}, y)\}$ for a large enough constant $K \gg 1$, whence

$$\deg_{\pm}(f' \wedge g) \leqslant 2d.$$

Theorem 15.7 now yields $R^+(f', 4d) + R^+(g, 2d) < 1$. Since $R^+(f, 4d) \leqslant R^+(f', 4d)$ by definition, the proof is complete. □

Finally, we obtain an analogue of this result for a conjunction of three and more functions.

THEOREM 15.10 (Sherstov [200]). *Let* $f_1, f_2, \ldots, f_k$ *be given Boolean functions on finite sets* $X_1, X_2, \ldots, X_k \subset \mathbb{R}^n$, *respectively. Assume that* $f_i \not\equiv 1$ *for* $i =$

$1, 2, \ldots, k$. *Let* $d = \deg_{\pm}(f_1 \wedge f_2 \wedge \cdots \wedge f_k)$. *Then*

$$\sum_{i=1}^{k} R^{+}(f_i, D) < 1$$

for $D = 8d \log 2k$.

PROOF. Since $f_1, f_2, \ldots, f_k \not\equiv 1$, it follows that for each pair of indices $i < j$, the function $f_i \wedge f_j$ is a subfunction of $f_1 \wedge f_2 \wedge \cdots \wedge f_k$. Theorem 15.9 now shows that for each $i < j$,

$$R^{+}(f_i, 4d) + R^{+}(f_j, 4d) < 1. \tag{15.9}$$

Without loss of generality, $R^{+}(f_1, 4d) = \max_{i=1,\ldots,k} R^{+}(f_i, 4d)$. Abbreviate $\varepsilon = R^{+}(f_1, 4d)$. By (15.9),

$$R^{+}(f_i, 4d) < \min\left\{1 - \varepsilon, \frac{1}{2}\right\}, \qquad i = 2, 3, \ldots, k.$$

Now Corollary 15.5 implies that

$$\sum_{i=1}^{k} R^{+}(f_i, D) \leqslant \varepsilon + \sum_{i=2}^{k} R^{+}(f_i, 4d)^{1+\log k} < 1. \qquad \square$$

15.5 Threshold degree of other compositions

As we will now see, the development in Section 15.4 applies to many combining functions other than conjunctions. Disjunctions are an illustrative starting point. Consider two Boolean functions $f: X \to \{-1, +1\}$ and $g: Y \to \{-1, +1\}$, where $X, Y \subset \mathbb{R}^n$ are finite sets and $f, g \not\equiv -1$. Let $d = \deg_{\pm}(f \vee g)$. Then, we claim

that

$$R^+(f, 4d) + R^+(g, 4d) < 1. \tag{15.10}$$

To see this, note first that the function $f \vee g$ has the same threshold degree as its negation, $\overline{f} \wedge \overline{g}$. Applying Theorem 15.9 to the latter function shows that

$$R^+(\overline{f}, 4d) + R^+(\overline{g}, 4d) < 1.$$

This is equivalent to (15.10) since approximating a function is the same as approximating its negation: $R^+(\overline{f}, 4d) = R^+(f, 4d)$ and $R^+(\overline{g}, 4d) = R^+(g, 4d)$. As in the case of conjunctions, (15.10) can be strengthened to

$$R^+(f, 2d) + R^+(g, 2d) < 1$$

if at least one of f, g is known to be odd. These observations carry over to disjunctions of multiple functions, $f_1 \vee f_2 \vee \cdots \vee f_k$.

The above discussion is still too specialized. In what follows, we consider composite functions $h(f_1, f_2, \ldots, f_k)$, where $h: \{-1, +1\}^k \to \{-1, +1\}$ is any given Boolean function. We will shortly see that the results of the previous sections hold for various h other than $h = \mathrm{AND}_k$ and $h = \mathrm{OR}_k$.

We start with some notation and definitions. Let $f, h: \{-1, +1\}^k \to \{-1, +1\}$ be given Boolean functions. Recall that f is called a *subfunction* of h if for some fixed strings $y, z \in \{-1, +1\}^k$, one has

$$f(x) = h(\ldots, (x_i \wedge y_i) \vee z_i, \ldots)$$

for each $x \in \{-1, +1\}^k$. In words, f can be obtained from h by replacing some of the variables $x_1, x_2, \ldots, x_k$ with fixed values, -1 or $+1$.

DEFINITION 15.11 (Sherstov [200]). A function $F\colon\{-1,+1\}^k \to \{-1,+1\}$ is AND-*reducible* if for each pair of indices i, j, where $1 \leqslant i \leqslant j \leqslant k$, at least one of the eight functions

$$\begin{aligned} &x_i \wedge x_j, && x_i \vee x_j,\\ &x_i \wedge \overline{x_j}, && x_i \vee \overline{x_j},\\ &\overline{x_i} \wedge x_j, && \overline{x_i} \vee x_j,\\ &\overline{x_i} \wedge \overline{x_j}, && \overline{x_i} \vee \overline{x_j} \end{aligned}$$

is a subfunction of $F(x)$.

THEOREM 15.12 (Sherstov [200]). *Let $f_1, f_2, \ldots, f_k$ be nonconstant Boolean functions on finite sets $X_1, X_2, \ldots, X_k \subset \mathbb{R}^n$, respectively. Let $F\colon\{-1,+1\}^k \to \{-1,+1\}$ be an* AND-*reducible function. Put $d = \deg_{\pm}(F(f_1, f_2, \ldots, f_k))$. Then*

$$\sum_{i=1}^{k} R^+(f_i, D) < 1$$

for $D = 8d \log 2k$.

PROOF. Since F is AND-reducible, it follows that for each pair of indices $i < j$, one of the following eight functions is a subfunction of $F(f_1, \ldots, f_k)$:

$$\begin{aligned} &f_i \wedge f_j, && f_i \vee f_j,\\ &f_i \wedge \overline{f_j}, && f_i \vee \overline{f_j},\\ &\overline{f_i} \wedge f_j, && \overline{f_i} \vee f_j,\\ &\overline{f_i} \wedge \overline{f_j}, && \overline{f_i} \vee \overline{f_j}. \end{aligned}$$

By Theorem 15.9 and the opening remarks of this section,

$$R^+(f_i, 4d) + R^+(f_j, 4d) < 1.$$

The remainder of the proof is identical to the proof of Theorem 15.10, starting at equation (15.9). □

In summary, the development in Section 15.4 naturally extends to compositions $F(f_1, f_2, \ldots, f_k)$ for various F. For a function $F: \{-1, +1\}^k \to \{-1, +1\}$ to be AND-reducible, F must clearly depend on all of its inputs. This necessary condition is often sufficient, for example when F is a read-once AND/OR/NOT formula or a halfspace. As a result, we have the following corollary of Theorem 15.12.

COROLLARY 15.13 (Sherstov [200]). *Let $f_1, f_2, \ldots, f_k$ be nonconstant Boolean functions on finite sets $X_1, X_2, \ldots, X_k \subset \mathbb{R}^n$, respectively. Let $F: \{-1, +1\}^k \to \{-1, +1\}$ be a halfspace or a read-once AND/OR/NOT formula. Assume that F depends on all of its k inputs and that the composition $F(f_1, f_2, \ldots, f_k)$ has threshold degree d. Then there is a degree-D rational function p_i/q_i on X_i, $i = 1, 2, \ldots, k$, such that*

$$\sum_{i=1}^{k} \max_{x_i \in X_i} \left| f_i(x_i) - \frac{p_i(x_i)}{q_i(x_i)} \right| < 1,$$

where $D = 8d \log 2k$.

REMARK (Sherstov [200]). If more information is available about the combining function F, Theorem 15.12 can be generalized to let some of $f_1, \ldots, f_k$ be constant functions. For example, some or all of the functions $f_1, \ldots, f_k$ in Theorem 15.10 can be identically true. Another direction for generalization is as follows. In Definition 15.11, one considers all the $\binom{k}{2}$ distinct pairs of indices (i, j). If one happens to know that f_1 is harder to approximate than $f_2, \ldots, f_k$, then one can relax Definition 15.11 to examine only the $k - 1$ pairs $(1, 2), (1, 3), \ldots, (1, k)$. We do not formulate these extensions as theorems, the fundamental technique being already clear.

15.6 Additional observations

Our results in this chapter can be viewed as a technique for proving lower bounds on the threshold degree of composite functions $F(f_1, f_2, \ldots, f_k)$. We make this view explicit in the following statement, which is the contrapositive of Theorem 15.12.

THEOREM 15.14 (Sherstov [200]). *Let $f_1, f_2, \ldots, f_k$ be nonconstant Boolean functions on finite sets $X_1, X_2, \ldots, X_k \subset \mathbb{R}^n$, respectively. Let $F: \{-1,+1\}^k \to \{-1,+1\}$ be an* AND*-reducible function. Suppose that $\sum R^+(f_i, D) \geqslant 1$ for some integer D. Then*

$$\deg_{\pm}(F(f_1, f_2, \ldots, f_k)) > \frac{D}{8 \log 2k}. \tag{15.11}$$

REMARK 15.15 (Sherstov [200]). Theorem 15.14 is close to optimal. For example, when $F = \mathrm{AND}_k$, the lower bound in (15.11) is tight up to a factor of $\Theta(k \log k)$. This can be seen by the well-known argument [33] described in the introduction. Specifically, fix an integer D such that $\sum R^+(f_i, D) < 1$. Then there exists a rational function $p_i(x_i)/q_i(x_i)$ on X_i, for $i = 1, 2, \ldots, k$, such that q_i is positive on X_i and

$$\sum_{i=1}^{k} \max_{x_i \in X_i} \left| f_i(x_i) - \frac{p_i(x_i)}{q_i(x_i)} \right| < 1.$$

As a result,

$$\bigwedge_{i=1}^{k} f_i(x_i) \equiv \operatorname{sgn}\left(k - 1 + \sum_{i=1}^{k} f_i(x_i)\right) \equiv \operatorname{sgn}\left(k - 1 + \sum_{i=1}^{k} \frac{p_i(x_i)}{q_i(x_i)}\right).$$

Multiplying by $\prod q_i(x_i)$ yields

$$\bigwedge_{i=1}^{k} f_i(x_i) \equiv \operatorname{sgn}\left((k-1)\prod_{i=1}^{k} q_i(x_i) + \sum_{i=1}^{k} p_i(x_i) \prod_{j\in\{1,\ldots,k\}\setminus\{i\}} q_j(x_j) \right),$$

whence $\deg_{\pm}(f_1 \wedge f_2 \wedge \cdots \wedge f_k) \leqslant kD$. This settles our claim regarding $F = \mathrm{AND}_k$. For arbitrary AND-reducible functions $F\colon \{-1,+1\}^k \to \{-1,+1\}$, a similar argument shows that the lower bound in (15.11) is tight up to a polynomial in k; cf. Theorem 31 of Klivans et al. [121].

We close this section with one additional result.

THEOREM 15.16 (Sherstov [200]). *Let $f\colon X \to \{-1,+1\}$ be a given function, where $X \subset \mathbb{R}^n$ is finite. Then for every integer $k \geqslant 2$,*

$$\deg_{\pm}(\underbrace{f \wedge f \wedge \cdots \wedge f}_{k}) \leqslant (8k\log k)\cdot \deg_{\pm}(f \wedge f). \tag{15.12}$$

PROOF. Put $d = \deg_{\pm}(f \wedge f)$. Theorem 15.9 implies that $R^+(f, 4d) < 1/2$, whence $R^+(f, 8d\log k) < 1/k$ by Corollary 15.5. By the argument applied earlier in Remark 15.15, this proves the theorem. □

To illustrate, let $\mathscr{C}$ be a given family of Boolean functions on $\{0,1\}^n$. Then Theorem 15.16 shows that the task of constructing a sign-representation for the intersections of up to k members from $\mathscr{C}$ reduces to the case $k = 2$. In other words, solving the problem for $k = 2$ essentially solves it for all k. The dependence on k in (15.12) is tight up to a factor of $16\log k$ by the work of Minsky and Papert [153], even in the simple case when $f = \mathrm{OR}_n$.

Chapter 16

Lower Bounds for Intersections of Two Halfspaces

In this chapter, we study the structural complexity of intersections of two halfspaces, a central concept class in computational learning theory. Specifically, we construct two halfspaces on $\{0,1\}^n$ whose intersection has threshold degree $\Theta(\sqrt{n})$, an exponential improvement on previous lower bounds. This solves an open problem due to Klivans [120] and rules out the use of perceptron-based techniques for PAC-learning the intersection of even two halfspaces. We also prove that the intersection of two majority functions has threshold degree $\Omega(\log n)$, which is tight and settles a conjecture of O'Donnell and Servedio [163]. We obtain these results by a detailed study of the rational approximation of halfspaces, along with the relationship between rational approximation and sign-representation proved in the previous chapter.

16.1 Introduction

The technical focus of this chapter is on the uniform approximation of Boolean functions by rational functions over the reals. Rational approximation, surveyed in detail in Section 16.2, is a fascinating topic with a variety of applications to complexity theory and learning theory. We will primarily be interested in the rational approximation of halfspaces. Specifically, consider the function $f\colon\{-1,+1\}^{n^2} \to \{-1,+1\}$ given by

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{n} 2^i x_{ij}\right), \tag{16.1}$$

which is known in the literature as the *canonical* halfspace. A key technical contribution of this chapter is to determine the least degree required for approximating f uniformly within ε by a rational function, for any given ε. Analogously, we determine the least degree required for approximating the majority function by a rational function within any given error ε. This development spans Sections 16.3–16.7.

In the concluding section, we present applications of our work to computational learning theory. These applications are based on the relationship between rational approximation and sign-representation, proved in the previous chapter. Specifically, let f be an arbitrary Boolean function, and let the symbol $f \wedge f$

denote the conjunction of two independent copies of f. Among other things, we proved in Chapter 15 that the threshold degree of the conjunction $f \wedge f$ is always within a constant factor of the least degree of a rational function that approximates f within $1/3$. As a consequence, we obtain the following result.

THEOREM 16.1 (Sherstov [200]). *Let* $f : \{-1,+1\}^{n^2} \to \{-1,+1\}$ *be given by* (16.1). *Then*

$$\deg_{\pm}(f \wedge f) = \Omega(n).$$

The lower bound in Theorem 16.1 is tight and matches the upper bounds of Beigel et al. [33]. Previously, only an $\Omega(\log n/\log\log n)$ lower bound was known on the threshold degree of the intersection of two halfspaces, due to O'Donnell and Servedio [163], preceded in turn by an $\omega(1)$ lower bound of Minsky and Papert [153].

Theorem 16.1 is of interest in learning theory. By the results of Klivans et al. [122, 121], Boolean functions with low threshold degree can be efficiently PAC learned under arbitrary distributions, by expressing such an unknown function as a perceptron with unknown weights and solving the associated linear program. Klivans et al. [121] showed that intersections of *light* halfspaces have low threshold degree, thereby giving an efficient PAC algorithm for this class. That result raised hopes that intersections of arbitrary k halfspaces on the hypercube have low threshold degree, for k small. Recall from previous chapters that no efficient algorithm is known for PAC learning the intersection of even $k = 2$ halfspaces, despite much effort and known solutions to some restrictions of the problem [139, 216, 121, 125].

Motivated by these considerations, Klivans [120, §7] posed the problem of proving a lower bound of $\Omega(\log n)$ or higher on the threshold degree of the intersection of two halfspaces. Theorem 16.1 solves this problem with a lower bound of $\Omega(\sqrt{n})$, showing that perceptron-based techniques will not yield a subexponential algorithm for PAC learning the intersection of even two halfspaces. It is the first unconditional, structural lower bound for this learning problem. Prior to our work, all known hardness results [40, 10, 128, 113] were based on complexity-theoretic assumptions. We complement it with the following result.

THEOREM 16.2 (Sherstov [200]). *Let* $f:\{-1,+1\}^{n^2} \to \{-1,+1\}$ *be given by* (16.1)*. Then*

$$\deg_{\pm}\left(f \wedge \mathrm{MAJ}_{\lceil\sqrt{n}\rceil}\right) = \Theta(\sqrt{n}).$$

In words, even if one of the halfspaces in Theorem 16.1 is replaced by a majority function, the threshold degree will remain high, resulting in a challenging learning problem. Finally, we have:

THEOREM 16.3 (Sherstov [200]). *The intersection of two majority functions satisfies*

$$\deg_{\pm}(\mathrm{MAJ}_n \wedge \mathrm{MAJ}_n) = \Omega(\log n).$$

Theorem 16.3 is tight, matching the construction of Beigel et al. [33]. It settles a conjecture of O'Donnell and Servedio [163], who proved a lower bound of $\Omega(\log n/\log\log n)$ with completely different techniques and conjectured that the true answer was $\Omega(\log n)$.

16.2 Rational approximation and its applications

The study of rational approximation dates back to the remarkable 1877 article by E. I. Zolotarev [228], a student of P. L. Chebyshev. Interest in the subject was revived a century later when D. J. Newman [156] obtained surprisingly accurate rational approximants for several common functions in $\mathscr{C}[-1,1]$, such as $|x|$ and x^{α} for rational $\alpha > 0$. In particular, Newman's work contributed an efficient rational approximant for the sign function, since $\operatorname{sgn} x = |x|/x$ for $x \neq 0$. Newman's discovery inspired considerable progress in the area, as surveyed in the monograph of Petrushev and Popov [168].

Newman's work on rational approximation has also found important applications in theoretical computer science, including the proof due to Beigel et al. [33] that PP is closed under intersection, circuit lower bounds due to Paturi

and Saks [166] and Siu et al. [210], and PAC learning algorithms due to Klivans et al. [121]. In addition, rational approximation of Boolean functions exactly captures the quantum query complexity in a natural model called *postselection*, due to Aaronson [1].

Finally, lower and upper bounds for rational approximation are of interest because of their relationship to polynomial approximation. Consider a Boolean function $f: X \to \{-1, +1\}$, for some finite set $X \subset \mathbb{R}^n$. By analogy with the polynomial case, define $\widetilde{\deg}_\varepsilon(f)$ to be the least degree of a rational function A such that $\|f - A\|_\infty \leqslant \varepsilon$. For all $0 < \varepsilon < 1$, the following relationships are well-known and straightforward to verify:

$$\frac{1}{2}\deg_\pm(f) \leqslant \widetilde{\deg}_\varepsilon(f) \leqslant \deg_\varepsilon(f).$$

Furthermore, our work in the previous chapter shows that

$$\deg_\pm(f \wedge f) = \Theta(\widetilde{\deg}_{1/3}(f)). \tag{16.2}$$

To summarize, the study of rational approximation contributes both lower and upper bounds for polynomial representations, as well as a tight characterization for functions of the form $f \wedge f$.

Despite this motivation, the rational approximation of Boolean functions remains poorly understood and has seen little progress since Newman's seminal paper in 1964. To illustrate some of the counterintuitive phenomena involved in rational approximation, consider the OR function on the hypercube $\{0,1\}^n$. Recall from Theorem 2.5 that $\deg_{1/3}(\mathrm{OR}_n) = \Theta(\sqrt{n})$, meaning that a polynomial of degree $\Theta(\sqrt{n})$ is required for approximation within $1/3$. At the same time, we claim that $\widetilde{\deg}_\varepsilon(\mathrm{OR}_n) = 1$ for all $0 < \varepsilon < 1$. Indeed, let

$$A_M(x) = \frac{1 - M(x_1 + \cdots + x_n)}{1 + M(x_1 + \cdots + x_n)}.$$

Then $\|f - A_M\|_\infty \to 0$ as $M \to \infty$. This example illustrates that proving lower bounds for rational functions can be a difficult and unintuitive task.

In this chapter, we study the rational approximation of halfspaces. Our main technical contribution is the following result on the canonical halfspace.

THEOREM 16.4 (Sherstov [200]). *Let* $f : \{-1,+1\}^{n^2} \to \{-1,+1\}$ *be given by*

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{n} 2^i x_{ij}\right).$$

Then for $1/3 < \varepsilon < 1$,

$$\operatorname{rdeg}_\varepsilon(f) = \Theta\left(1 + \frac{n}{\log\{1/(1-\varepsilon)\}}\right).$$

Furthermore, for all $\varepsilon > 0$,

$$\operatorname{rdeg}_\varepsilon(f) \leqslant 64n\lceil \log_2 n \rceil + 1.$$

In particular, Theorem 16.4 shows that a rational function of degree $\Theta(n)$ is necessary and sufficient for approximating the canonical halfspace within $1/3$. The best previous degree lower bound for constant-error approximation of any halfspace was $\Omega(\log n / \log\log n)$, obtained implicitly in [163]. We complement Theorem 16.4 with a full solution for another common halfspace, the majority function.

THEOREM 16.5 (Sherstov [200]). *The majority function satisfies*

$$\operatorname{rdeg}_\varepsilon(\mathrm{MAJ}_n) = \begin{cases} \Theta\left(\log\left\{\dfrac{2n}{\log(1/\varepsilon)}\right\} \cdot \log\dfrac{1}{\varepsilon}\right), & 2^{-n} < \varepsilon < 1/3, \\ \Theta\left(1 + \dfrac{\log n}{\log\{1/(1-\varepsilon)\}}\right), & 1/3 \leqslant \varepsilon < 1. \end{cases}$$

The upper bound in Theorem 16.5 is relatively straightforward. Indeed, an upper bound of $O(\log\{1/\varepsilon\}\log n)$ for $\varepsilon < 1/3$ was known and used in the complexity literature long before our work [166, 210, 33, 121], and we only somewhat tighten that upper bound and extend it to all ε. Our primary contribution in Theorem 16.5, then, is a matching *lower* bound on the degree, which requires considerable effort. The closest previous line of research concerns *continuous* approximation of the sign function on $[-1,-\varepsilon]\cup[\varepsilon,1]$, which unfortunately gives no insight into the discrete case. For example, the lower bound derived by Newman [156] in the continuous setting is based on the integration of relevant rational functions with respect to a suitable weight function, which has no meaningful discrete analogue.

16.3 Technical background

The reader may find it helpful to review the definitions and background results in Section 15.2, which are directly relevant to our work in this chapter. We will need a number of additional conventions and results. If $\mu_1,\dots,\mu_k$ are probability distributions on finite sets $X_1,\dots,X_k$, respectively, then $\mu_1\times\cdots\times\mu_k$ stands for the probability distribution on $X_1\times\cdots\times X_k$ given by

$$(\mu_1\times\cdots\times\mu_k)(x_1,\dots,x_k)=\prod_{i=1}^{k}\mu_i(x_i).$$

The following combinatorial identity is well-known.

FACT 16.6. *For every integer $n\geqslant 1$ and every polynomial $p\in P_{n-1}$,*

$$\sum_{i=0}^{n}\binom{n}{i}(-1)^i p(i)=0.$$

This fact can be verified by repeated differentiation of the real function

$$(t-1)^n = \sum_{i=0}^{n} \binom{n}{i} (-1)^{n-i} t^i$$

at $t = 1$, as explained in [163].

We now recall Newman's classical construction of a rational approximant to the sign function [156].

THEOREM 16.7 (Newman [156]). *Fix $N > 1$. Then for every integer $k \geqslant 1$, there is a rational function $S(t)$ of degree k such that*

$$\max_{1 \leqslant |t| \leqslant N} |\operatorname{sgn} t - S(t)| \leqslant 1 - N^{-1/k} \tag{16.3}$$

and the denominator of S is positive on $[-N,-1] \cup [1,N]$.

PROOF (adapted from Newman [156]). Consider the univariate polynomial

$$p(t) = \prod_{i=1}^{k} \left(t + N^{(2i-1)/(2k)}\right).$$

By examining every interval $[N^{i/(2k)}, N^{(i+1)/(2k)}]$, where $i = 0, 1, \ldots, 2k-1$, one sees that

$$p(t) \geqslant \frac{N^{1/(2k)} + 1}{N^{1/(2k)} - 1} |p(-t)|, \qquad 1 \leqslant t \leqslant N. \tag{16.4}$$

Letting

$$S(t) = N^{-1/(2k)} \cdot \frac{p(t) - p(-t)}{p(t) + p(-t)},$$

one has (16.3). The positivity of the denominator of S on $[-N,-1]\cup[1,N]$ is a consequence of (16.4). □

A useful consequence of Newman's theorem is the following general statement on decreasing the error in rational approximation.

THEOREM 16.8. *Let $f: X \to \{-1,+1\}$ be given, where $X \subseteq \mathbb{R}^n$. Let d be a given integer, $\varepsilon = R(f,d)$. Then for $k = 1,2,3,\ldots,$*

$$R(f,kd) \leqslant 1 - \left(\frac{1-\varepsilon}{1+\varepsilon}\right)^{1/k}.$$

PROOF. We may assume that $\varepsilon < 1$, the theorem being trivial otherwise. Let S be the degree-k rational approximant to the sign function for $N = (1+\varepsilon)/(1-\varepsilon)$, as constructed in Theorem 16.7. Let $A_1, A_2, \ldots, A_m, \ldots$ be a sequence of rational functions on X of degree at most d such that $\sup_X |f - A_m| \to \varepsilon$ as $m \to \infty$. The theorem follows by considering the sequence of approximants $S(A_m(x)/\{1-\varepsilon\})$ as $m \to \infty$. □

The following result generalizes Minsky and Papert's symmetrization technique, given by Propositions 2.2 and 8.2, to rational functions.

PROPOSITION 16.9 (Sherstov [200]). *Let $n_1, \ldots, n_k$ be positive integers, and α, β distinct reals. Let $G: \{\alpha,\beta\}^{n_1} \times \cdots \times \{\alpha,\beta\}^{n_k} \to \{-1,+1\}$ be a function such that $G(x_1,\ldots,x_k) \equiv G(\sigma_1 x_1, \ldots, \sigma_k x_k)$ for all $\sigma_1 \in S_{n_1}, \ldots, \sigma_k \in S_{n_k}$. Let d be a given integer. Then for each $\varepsilon > R^+(G,d)$, there exists a rational function p/q on $\mathbb{R}^k$ of degree at most d such that for all x in the domain of G, one has*

$$\left| G(x) - \frac{p(\ldots, x_{i,1} + \cdots + x_{i,n_i}, \ldots)}{q(\ldots, x_{i,1} + \cdots + x_{i,n_i}, \ldots)} \right| < \varepsilon$$

and $q(\ldots, x_{i,1} + \cdots + x_{i,n_i}, \ldots) > 0$.

PROOF. Clearly, we may assume that $\varepsilon < 1$. Using the linear bijection $(\alpha,\beta) \leftrightarrow (0,1)$ if necessary, we may further assume that $\alpha = 0$ and $\beta = 1$. Since $\varepsilon >$

$R^+(G,d)$, there are polynomials P, Q of degree up to d such that for all x in the domain of G, one has $Q(x) > 0$ and

$$(1-\varepsilon)Q(x) < G(x)P(x) < (1+\varepsilon)Q(x).$$

By Proposition 8.2, there exist polynomials p, q on $\mathbb{R}^k$ of degree at most d such that

$$\mathop{\mathbf{E}}_{\sigma_1 \in S_{n_1}, \ldots, \sigma_k \in S_{n_k}} \left[P\left(\sigma_1 x_1, \ldots, \sigma_k x_k\right) \right] = p\left(\ldots, x_{i,1} + \cdots + x_{i,n_i}, \ldots \right)$$

and

$$\mathop{\mathbf{E}}_{\sigma_1 \in S_{n_1}, \ldots, \sigma_k \in S_{n_k}} \left[Q\left(\sigma_1 x_1, \ldots, \sigma_k x_k\right) \right] = q\left(\ldots, x_{i,1} + \cdots + x_{i,n_i}, \ldots \right)$$

for all x in the domain of G. Then the required properties of p and q follow immediately from the corresponding properties of P and Q. □

16.4 Upper bounds for the approximation of halfspaces

The lower bounds in Theorem 16.4, for the rational approximation of the canonical halfspace, are considerably more involved than the upper bounds. To help build some intuition in the former case, we prove the upper bounds first.

We showed in Section 16.2 that $R^+(\mathrm{OR}_n, 1) = 0$. A similar example is the function $\mathrm{OMB}_n \colon \{0,1\}^n \to \{-1,+1\}$ given by (4.34). Indeed, letting

$$A_M(x) = \frac{1 + \sum_{i=1}^n (-M)^i x_i}{1 + \sum_{i=1}^n M^i x_i},$$

we have $\|\mathrm{OMB}_n - A_M\|_\infty \to 0$ as $M \to \infty$. Thus,

$$R^+(\mathrm{OMB}_n, 1) = 0.$$

With this construction in mind, we now turn to the canonical halfspace. We start with an auxiliary result that generalizes the above argument.

LEMMA 16.10 (Sherstov [200]). *Let* $f : \{0, \pm 1, \pm 2\}^n \to \{-1, +1\}$ *be the function given by* $f(z) = \mathrm{sgn}(1 + \sum_{i=1}^n 2^i z_i)$. *Then*

$$R^+(f, 64) = 0.$$

PROOF. Consider the deterministic finite automaton in Figure 16.1. The automaton has two terminal states (labeled "+" and "−") and three nonterminal states (the start state and two additional states). We interpret the output of the automaton to be $+1$ and -1 at the two terminal states, respectively, and 0 otherwise. A string $z = (z_n, z_{n-1}, \ldots, z_1, 0) \in \{0, \pm 1, \pm 2\}^{n+1}$, when read by the automaton left to right, forces it to output exactly $\mathrm{sgn}(\sum_{i=1}^n 2^i z_i)$. If the automaton is currently at a nonterminal state, this state is determined uniquely by the last two symbols read. Hence, the output of the automaton on input $z = (z_n, z_{n-1}, \ldots, z_1, 0) \in \{0, \pm 1, \pm 2\}^{n+1}$ is given by

$$\mathrm{sgn}\left(\sum_{i=0}^n 2^i \alpha(z_{i+2}, z_{i+1}, z_i)\right)$$

for a suitable map $\alpha : \{0, \pm 1, \pm 2\}^3 \to \{0, -1, +1\}$, where we adopt the shorthand $z_{n+1} = z_{n+2} = z_0 = 0$. Put

$$A_M(z) = \frac{1 + \sum_{i=0}^n M^{i+1} \alpha(z_{i+2}, z_{i+2}, z_i)}{1 + \sum_{i=0}^n M^{i+1} |\alpha(z_{i+2}, z_{i+2}, z_i)|}.$$

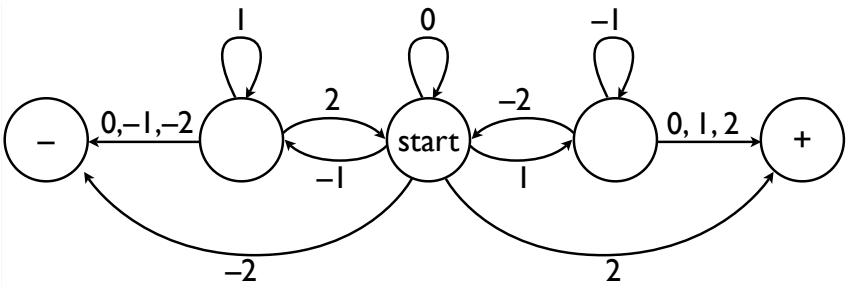

Figure 16.1: Finite automaton for the proof of Lemma 16.10.

By interpolation, the numerator and denominator of A_M can be represented by polynomials of degree no more than $4 \times 4 \times 4 = 64$. On the other hand, we have $\|f - A_M\|_\infty \to 0$ as $M \to \infty$. □

We are now prepared to prove our desired upper bounds for halfspaces.

THEOREM 16.11 (Sherstov [200]). *Let* $f : \{-1,+1\}^{nk} \to \{-1,+1\}$ *be the function given by*

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{k} 2^i x_{ij}\right). \tag{16.5}$$

Then

$$R^+(f, 64k\lceil \log k \rceil + 1) = 0. \tag{16.6}$$

In addition, for all integers $d \geqslant 1$,

$$R^+(f,d) \leqslant 1 - (k2^{n+1})^{-1/d}. \tag{16.7}$$

In particular, Theorem 16.11 settles all upper bounds on $\operatorname{rdeg}_\varepsilon(f)$ in Theorem 16.4.

PROOF OF THEOREM 16.11. Theorem 16.7 immediately implies (16.7) in view of the representation (16.5). It remains to prove (16.6). In the degenerate case $k = 1$, we have $f \equiv x_{n1}$ and thus (16.6) holds. In what follows, we assume that $k \geqslant 2$ and put $\Delta = \lceil \log k \rceil$. We adopt the convention that $x_{ij} \equiv 0$ for $i > n$. For $\ell = 0, 1, 2, \ldots$, define

$$S_\ell = \sum_{i=1}^{\Delta} \sum_{j=1}^{k} 2^{i-1} x_{\ell\Delta+i,j}.$$

Then

$$\sum_{i=1}^{n} \sum_{j=1}^{k} 2^{i-1} x_{ij} = \left(S_0 + 2^{2\Delta} S_2 + 2^{4\Delta} S_4 + 2^{6\Delta} S_6 + \cdots\right) + \left(2^{\Delta} S_1 + 2^{3\Delta} S_3 + 2^{5\Delta} S_5 + 2^{7\Delta} S_7 + \cdots\right). \tag{16.8}$$

Now, each S_ℓ is an integer in $[-2^{2\Delta} + 1, 2^{2\Delta} - 1]$ and therefore admits a representation as

$$S_\ell = z_{\ell,1} + 2z_{\ell,2} + 2^2 z_{\ell,3} + \cdots + 2^{2\Delta-1} z_{\ell,2\Delta},$$

where $z_{\ell,1}, \ldots, z_{\ell,2\Delta} \in \{-1, 0, +1\}$. Furthermore, each S_ℓ only depends on $k\Delta$ of the original variables x_{ij}, whence $z_{\ell,1}, \ldots, z_{\ell,2\Delta}$ can all be viewed as polynomials of degree at most $k\Delta$ in the original variables. Rewriting (16.8),

$$\sum_{i=1}^{n} \sum_{j=1}^{k} 2^{i-1} x_{ij} = \left(\sum_{i \geqslant 1} 2^{i-1} z_{\ell(i),j(i)} \right) + \left(\sum_{i \geqslant \Delta+1} 2^{i-1} z_{\ell'(i),j'(i)} \right)$$

for appropriate indexing functions $\ell(i), \ell'(i), j(i), j'(i)$. Thus,

$$f(x) \equiv \operatorname{sgn}\left(1 + \sum_{i=1}^{\Delta} 2^i \underbrace{z_{\ell(i),j(i)}}_{} + \sum_{i \geqslant \Delta+1} 2^i \underbrace{\left(z_{\ell(i),j(i)} + z_{\ell'(i),j'(i)}\right)}_{}\right).$$

Since the underbraced expressions range in $\{0, \pm 1, \pm 2\}$ and are polynomials of degree at most $k\Delta$ in the original variables, Lemma 16.10 implies (16.6). □

16.5 Preparatory analytic work on halfspaces

This section sets the stage for our lower bounds with some preparatory results about halfspaces. It will be convenient to establish some additional notation, for use in this section only. Here, we typeset real vectors in boldface ($\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}, \mathbf{v}$) to better distinguish them from scalars. The ith component of a vector $\mathbf{x} \in \mathbb{R}^n$ is denoted by $(\mathbf{x})_i$, while the symbol $\mathbf{x}_i$ is reserved for another *vector* from some enumeration. In keeping with this convention, we let $\mathbf{e}_i$ denote the vector with 1 in the ith component and zeroes everywhere else. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the vector $\mathbf{xy} \in \mathbb{R}^n$ is given by $(\mathbf{xy})_i \equiv (\mathbf{x})_i (\mathbf{y})_i$. More generally, for a polynomial p on $\mathbb{R}^k$ and vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, we define $p(\mathbf{x}_1, \ldots, \mathbf{x}_k) \in \mathbb{R}^n$ by $(p(\mathbf{x}_1, \ldots, \mathbf{x}_k))_i = p((\mathbf{x}_1)_i, \ldots, (\mathbf{x}_k)_i)$. The expectation of a random variable $\mathbf{x} \in \mathbb{R}^n$ is defined componentwise, i.e., the vector $\mathbf{E}[\mathbf{x}] \in \mathbb{R}^n$ is given by $(\mathbf{E}[\mathbf{x}])_i \equiv \mathbf{E}[(\mathbf{x})_i]$.

For convenience, we adopt the notational shorthand $\alpha^0 = 1$ for all $\alpha \in \mathbb{R}$. In particular, if $\mathbf{x} \in \mathbb{R}^n$ is a given vector, then $\mathbf{x}^0 = (1, 1, \ldots, 1) \in \mathbb{R}^n$. A scalar $\alpha \in \mathbb{R}$, when interpreted as a vector, stands for $(\alpha, \alpha, \ldots, \alpha)$. This shorthand allows one to speak of $\operatorname{span}\{1, \mathbf{z}, \mathbf{z}^2, \ldots, \mathbf{z}^k\}$, for example, where $\mathbf{z} \in \mathbb{R}^n$ is a given vector.

LEMMA 16.12 (Sherstov [200]). *Let positive integers N and m be given. Let $\alpha_0, \alpha_1, \ldots, \alpha_{4m}$ be suitable reals. Then for each $\mathbf{b} \in \{0,1\}^N$, there exists a probability distribution $\mu_{\mathbf{b}}$ on $\{0, \pm 1, \ldots, \pm m\}^N$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mu_{\mathbf{b}}} [(2\mathbf{v} + \mathbf{b})^d] = (\alpha_d, \alpha_d, \ldots, \alpha_d), \qquad d = 0, 1, 2, \ldots, 4m.$$

PROOF. Let λ_0 and λ_1 be the distributions on $\{0, \pm 1, \ldots, \pm m\}$ given by

$$\lambda_0(t) = 16^{-m} \binom{4m+1}{2m+2t}$$

and

$$\lambda_1(t) = 16^{-m} \binom{4m+1}{2m+2t+1}.$$

Then for $d = 0, 1, \ldots, 4m$, one has

$$\begin{aligned} \mathop{\mathbf{E}}_{t\sim\lambda_0}[(2t)^d] - \mathop{\mathbf{E}}_{t\sim\lambda_1}[(2t+1)^d] &= 16^{-m} \sum_{t=0}^{4m+1} (-1)^t \binom{4m+1}{t} (t-2m)^d \\ &= 0, \end{aligned} \tag{16.9}$$

where (16.9) holds by Fact 16.6. Now, let

$$\mu_{\mathbf{b}} = \lambda_{(\mathbf{b})_1} \times \lambda_{(\mathbf{b})_2} \times \cdots \times \lambda_{(\mathbf{b})_N}.$$

Then in view of (16.9), the lemma holds by letting $\alpha_d = \mathbf{E}_{\lambda_0}[(2t)^d]$ for $d = 0, 1, 2, \ldots, 4m$. □

Using the previous lemma, we will now establish another auxiliary result pertaining to halfspaces.

LEMMA 16.13 (Sherstov [200]). *Put*

$$\mathbf{z} = (-2^n, -2^{n-1}, \ldots, -2^0, 2^0, \ldots, 2^{n-1}, 2^n) \in \mathbb{R}^{2n+2}.$$

There are random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1} \in \{0, \pm 1, \pm 2, \ldots, \pm(3n+1)\}^{2n+2}$ *such that*

$$\sum_{i=1}^{n+1} 2^{i-1} \mathbf{x}_i \equiv \mathbf{z} \tag{16.10}$$

and

$$\mathbf{E}\left[\prod_{i=1}^{n} \mathbf{x}_i^{d_i}\right] \in \operatorname{span}\{(1, 1, \ldots, 1)\} \tag{16.11}$$

for $d_1, \ldots, d_n \in \{0, 1, \ldots, 4n\}$.

PROOF. Let

$$\mathbf{x}_i = 2\mathbf{y}_i - \mathbf{y}_{i-1} + \mathbf{e}_{n+1+i} - \mathbf{e}_{n+2-i}, \qquad i = 1, 2, \ldots, n+1,$$

where $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n+1}$ are suitable random variables with $\mathbf{y}_0 \equiv \mathbf{y}_{n+1} \equiv 0$. Then property (16.10) is immediate. We will construct $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n+1}$ such that the remaining property (16.11) holds as well.

Let $N = 2n + 2$ and $m = n$ in Lemma 16.12. Then reals $\alpha_0, \alpha_1, \ldots, \alpha_{4n}$ exist with the property that for each $\mathbf{b} \in \{0, 1\}^{2n+2}$, a probability distribution $\mu_{\mathbf{b}}$ can be found on $\{0, \pm 1, \ldots, \pm n\}^{2n+2}$ such that

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mu_{\mathbf{b}}} [(2\mathbf{v} + \mathbf{b})^d] = \alpha_d (1, 1, \ldots, 1), \qquad d = 0, 1, \ldots, 4n. \tag{16.12}$$

Now, we will specify the distribution of $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_n$ by giving an algorithm for generating $\mathbf{y}_i$ from $\mathbf{y}_{i-1}$. First, recall that $\mathbf{y}_0 \equiv \mathbf{y}_{n+1} \equiv 0$. The algorithm for generating $\mathbf{y}_i$ given $\mathbf{y}_{i-1}$ $(i = 1, 2, \ldots, n)$ is as follows.

(1) Let $\mathbf{u}$ be the unique integer vector such that $2\mathbf{u} - \mathbf{y}_{i-1} + \mathbf{e}_{n+1+i} - \mathbf{e}_{n+2-i} \in \{0,1\}^{2n+2}$.

(2) Draw a random vector $\mathbf{v} \sim \mu_{\mathbf{b}}$, where $\mathbf{b} = 2\mathbf{u} - \mathbf{y}_{i-1} + \mathbf{e}_{n+1+i} - \mathbf{e}_{n+2-i}$.

(3) Set $\mathbf{y}_i = \mathbf{v} + \mathbf{u}$.

One easily verifies that $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n+1} \in \{0, \pm 1, \ldots, \pm 3n\}^{2n+2}$.

Let R denote the resulting joint distribution of $(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n+1})$. Let $i \leqslant n$. Then conditioned on any fixed value of $(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{i-1})$ in the support of R, the random variable $\mathbf{x}_i$ is by definition independent of $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}$ and is distributed identically to $2\mathbf{v} + \mathbf{b}$, for some fixed vector $\mathbf{b} \in \{0,1\}^{2n+2}$ and a random variable $\mathbf{v} \sim \mu_{\mathbf{b}}$. In view of (16.12), we conclude that

$$\mathbf{E}\left[\prod_{i=1}^{n} \mathbf{x}_i^{d_i}\right] = (1,1,\ldots,1)\prod_{i=1}^{n} \alpha_{d_i}$$

for all $d_1, d_2, \ldots, d_n \in \{0,1,\ldots,4n\}$, which establishes (16.11). It remains to note that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \{-2n, -2n+1, \ldots, -1, 0, 1, \ldots, 2n, 2n+1\}^{2n+2}$, whereas $\mathbf{x}_{n+1} = -\mathbf{y}_n + \mathbf{e}_{2n+2} - \mathbf{e}_1 \in \{0, \pm 1, \ldots, \pm(3n+1)\}^{2n+2}$. □

At last, we arrive at the main theorem of this section, which will play a crucial role in our analysis of the rational approximation of halfspaces.

THEOREM 16.14 (Sherstov [200]). *For $i = 0, 1, 2, \ldots, n$, define*

$$A_i = \left\{ (x_1, \ldots, x_{n+1}) \in \{0, \pm 1, \ldots, \pm(3n+1)\}^{n+1} : \quad \sum_{j=1}^{n+1} 2^{j-1} x_j = 2^i \right\}.$$

Let $p(x_1, \ldots, x_{n+1})$ be a real polynomial with sign $(-1)^i$ throughout A_i $(i = 0, 1, 2, \ldots, n)$ and sign $(-1)^{i+1}$ throughout $-A_i$ $(i = 0, 1, 2, \ldots, n)$. Then

$$\deg p \geqslant 2n + 1.$$

PROOF. For the sake of contradiction, suppose that p has degree no greater than $2n$. Put $\mathbf{z} = (-2^n, -2^{n-1}, \ldots, -2^0, 2^0, \ldots, 2^{n-1}, 2^n)$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ be the random variables constructed in Lemma 16.13. By (16.11) and the identity $\mathbf{x}_{n+1} \equiv 2^{-n}\mathbf{z} - \sum_{i=1}^{n} 2^{i-n-1}\mathbf{x}_i$, we have

$$\mathbf{E}[p(\mathbf{x}_1, \ldots, \mathbf{x}_{n+1})] \in \operatorname{span}\{1, \mathbf{z}, \mathbf{z}^2, \ldots, \mathbf{z}^{2n}\},$$

whence

$$\mathbf{E}[p(\mathbf{x}_1, \ldots, \mathbf{x}_{n+1})] = q(\mathbf{z})$$

for a univariate polynomial $q \in P_{2n}$. By (16.10) and the sign behavior of p, we have

$$\operatorname{sgn} q(2^i) = (-1)^i, \qquad i = 0, 1, 2, \ldots, n,$$

and

$$\operatorname{sgn} q(-2^i) = (-1)^{i+1}, \qquad i = 0, 1, 2, \ldots, n.$$

Therefore, q has at least $2n + 1$ roots. Since $q \in P_{2n}$, we arrive at a contradiction. It follows that the assumed polynomial p does not exist. □

16.6 Lower bounds for the approximation of halfspaces

The purpose of this section is to prove that the canonical halfspace cannot be approximated well by a rational function of low degree. A starting point in our discussion is a criterion for inapproximability by low-degree rational functions, which is applicable not only to halfspaces but any odd Boolean functions on Euclidean space.

THEOREM 16.15 (Sherstov [200]). *Fix a nonempty finite subset* $S \subset \mathbb{R}^m$ *with* $S \cap -S = \varnothing$. *Define* $f : S \cup -S \to \{-1, +1\}$ *by*

$$f(x) = \begin{cases} +1, & x \in S, \\ -1, & x \in -S. \end{cases}$$

Let ψ *be a real function such that*

$$\psi(x) > \delta |\psi(-x)|, \qquad x \in S, \tag{16.13}$$

for some $\delta \in (0, 1)$ *and*

$$\sum_{S \cup -S} \psi(x) u(x) = 0 \tag{16.14}$$

for every polynomial u *of degree at most* d. *Then*

$$R^+(f, d) \geqslant \frac{2\delta}{1 + \delta}.$$

PROOF. Fix polynomials p, q of degree at most d such that q is positive on $S \cup -S$. Put

$$\varepsilon = \max_{S \cup -S} \left| f(x) - \frac{p(x)}{q(x)} \right|.$$

We assume that $\varepsilon < 1$ since otherwise there is nothing to show. For $x \in S$,

$$(1 - \varepsilon) q(x) \leqslant p(x) \leqslant (1 + \varepsilon) q(x) \tag{16.15}$$

and

$$(1-\varepsilon)q(-x) \leqslant -p(-x) \leqslant (1+\varepsilon)q(-x). \tag{16.16}$$

Consider the polynomial $u(x) = q(x)+q(-x)+p(x)-p(-x)$. Equations (16.15) and (16.16) show that for $x \in S$, one has $u(x) \geqslant (2-\varepsilon)\{q(x)+q(-x)\}$ and $|u(-x)| \leqslant \varepsilon\{q(x)+q(-x)\}$, whence

$$u(x) \geqslant \left(\frac{2}{\varepsilon}-1\right)|u(-x)|, \qquad x \in S. \tag{16.17}$$

We also note that

$$u(x) > 0, \qquad x \in S. \tag{16.18}$$

Since u has degree at most d, we have by (16.14) that

$$\sum_{x\in S}\{\psi(x)u(x)+\psi(-x)u(-x)\} = \sum_{S\cup -S}\psi(x)u(x) = 0,$$

whence

$$\psi(x)u(x) \leqslant |\psi(-x)u(-x)|$$

for some $x \in S$. At the same time, it follows from (16.13), (16.17), and (16.18) that

$$\psi(x)u(x) > \delta\left(\frac{2}{\varepsilon}-1\right)|\psi(-x)u(-x)|, \qquad x \in S.$$

We immediately obtain $\delta(\{2/\varepsilon\}-1) < 1$, as was to be shown. □

REMARK 16.16 (Sherstov [200]). The method of Theorem 16.15 amounts to reformulating (16.17) and (16.18) as a linear program and exhibiting a solution to its dual. The presentation above does not explicitly use the language of linear programs or appeal to duality, however, because our goal is solely to prove the correctness of our method and not its completeness.

Using the criterion of Theorem 16.15 and our preparatory work in Section 16.5, we now establish a key lower bound for the rational approximation of halfspaces within constant error.

THEOREM 16.17 (Sherstov [200]). *Let* $f\colon \{0, \pm 1, \ldots, \pm(3n+1)\}^{n+1} \to \{-1, +1\}$ *be given by*

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n+1} 2^i x_i\right).$$

Then

$$R^+(f, n) = \Omega(1).$$

PROOF. Let $A_0, A_1, \ldots, A_n$ be as defined in Theorem 16.14. Put $A = \bigcup A_i$ and define $g\colon A \cup -A \to \{-1, +1\}$ by

$$g(x) = \begin{cases} (-1)^i, & x \in A_i, \\ (-1)^{i+1}, & x \in -A_i. \end{cases}$$

Then $\deg_{\pm}(f) > 2n$ by Theorem 16.14. As a result, Theorem 4.6 guarantees the existence of a function $\phi\colon A \cup -A \to \mathbb{R}$, not identically zero, such that

$$\phi(x)g(x) \geqslant 0, \qquad x \in A \cup -A, \tag{16.19}$$

and

$$\sum_{A \cup -A} \phi(x) u(x) = 0 \tag{16.20}$$

for every polynomial u of degree at most $2n$. Put

$$p(x) = \prod_{j=0}^{n-1} \left(-2^j \sqrt{2} + \sum_{i=1}^{n+1} 2^{i-1} x_i \right)$$

and

$$\psi(x) = (-1)^n \{\phi(x) - \phi(-x)\} p(x).$$

Define $S = A \backslash \psi^{-1}(0)$. Then $S \neq \varnothing$ by (16.19) and the fact that ϕ is not identically zero on $A \cup -A$. For $x \in S$, we have $\psi(-x) \neq 0$ and

$$\frac{|\psi(x)|}{|\psi(-x)|} = \frac{|p(x)|}{|p(-x)|} > \left(\prod_{i=1}^{\infty} \frac{2^{i/2} - 1}{2^{i/2} + 1} \right)^2 > \exp(-9\sqrt{2}),$$

where the final step uses the bound $(a - 1)/(a + 1) > \exp(-2.5/a)$, valid for $a \geqslant \sqrt{2}$. It follows from (16.19) and the definition of p that ψ is positive on S. Hence,

$$\psi(x) > \exp(-9\sqrt{2})\, |\psi(-x)|, \qquad x \in S. \tag{16.21}$$

For any polynomial u of degree no greater than n, we infer from (16.20) that

$$\sum_{S \cup -S} \psi(x) u(x) = (-1)^n \sum_{A \cup -A} \{\phi(x) - \phi(-x)\} u(x) p(x) = 0. \tag{16.22}$$

Since f is positive on S and negative on $-S$, the proof is now complete in view of (16.21), (16.22), and Theorem 16.15. □

We have reached the main result of this section, which generalizes Theorem 16.17 to any subconstant approximation error and to halfspaces on the hypercube.

THEOREM 16.18 (Sherstov [200]). *Let* $F\colon \{-1,+1\}^{m^2} \to \{-1,+1\}$ *be given by*

$$F(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{m}\sum_{j=1}^{m} 2^i x_{ij}\right). \tag{16.23}$$

Then for $d < m/14$,

$$R(F,d) \geqslant 1 - 2^{-\Theta(m/d)}. \tag{16.24}$$

Observe that Theorem 16.18 settles the lower bounds in Theorem 16.4.

PROOF OF THEOREM 16.18. We may assume that $m \geqslant 14$, the claim being trivial otherwise. Consider the function $G\colon \{-1,+1\}^{(n+1)(6n+2)} \to \{-1,+1\}$ given by

$$G(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n+1}\sum_{j=1}^{6n+2} 2^i x_{ij}\right),$$

where $n = \lfloor (m-2)/6 \rfloor$. For every $\varepsilon > R^+(G,n)$, Proposition 16.9 provides a rational function A on $\mathbb{R}^{n+1}$ of degree at most n such that, on the domain of G,

$$\left| G(x) - A\left(\ldots, \sum_{j=1}^{6n+2} x_{ij}, \ldots\right)\right| < \varepsilon$$

and the denominator of A is positive. Letting f be the function in Theorem 16.17, it follows that $|f(x_1, \ldots, x_{n+1}) - A(2x_1, \ldots, 2x_{n+1})| < \varepsilon$ on the domain of f, whence

$$R^+(G, n) = \Omega(1). \tag{16.25}$$

We now claim that either $G(x)$ or $-G(-x)$ is a subfunction of F. For example, consider the following substitution for the variables x_{ij} for which $i > n+1$ or $j > 6n+2$:

$$\begin{aligned} x_{mj} &\leftarrow (-1)^j, && (1 \leqslant j \leqslant m), \\ x_{ij} &\leftarrow (-1)^{j+1}, && (n+1 < i < m, \quad 1 \leqslant j \leqslant m), \\ x_{ij} &\leftarrow (-1)^{j+1}, && (1 \leqslant i \leqslant n+1, \quad j > 6n+2). \end{aligned}$$

After this substitution, F is a function of the remaining variables x_{ij} and is equivalent to $G(x)$ if m is even, and to $-G(-x)$ if m is odd. In either case, (16.25) implies that

$$R^+(F, n) = \Omega(1). \tag{16.26}$$

Theorem 16.8 shows that

$$R(F, n/2) \leqslant 1 - \left(\frac{1 - R(F, d)}{2}\right)^{1/\lfloor n/(2d) \rfloor}$$

for $d = 1, 2, \ldots, \lfloor n/2 \rfloor$, which yields (16.24) in light of (15.3) and (16.26). □

16.7 Rational approximation of the majority function

The goal of this section is to determine $R^+(\mathrm{MAJ}_n, d)$ for each integer d, i.e., to determine the least error to which a degree-d multivariate rational function can approximate the majority function. As is frequently the case with symmetric Boolean functions, the multivariate problem of analyzing $R^+(\mathrm{MAJ}_n, d)$ is equivalent to a univariate question. Specifically, given an integer d and a finite set $S \subset \mathbb{R}$, we define

$$R^+(d, S) = \inf_{p,q} \max_{t \in S} \left| \operatorname{sgn} t - \frac{p(t)}{q(t)} \right|,$$

where the infimum ranges over $p, q \in P_d$ such that q is positive on S. In other words, we study how well a rational function of a given degree can approximate the sign function over a finite support. We give a detailed answer to this question in the following theorem:

THEOREM 16.19 (Sherstov [200]). *Let n, d be positive integers. Abbreviate $R = R^+(d, \{\pm 1, \pm 2, \ldots, \pm n\})$. For $1 \leqslant d \leqslant \log n$,*

$$\exp\left\{-\Theta\left(\frac{1}{n^{1/(2d)}}\right)\right\} \leqslant R < \exp\left\{-\frac{1}{n^{1/d}}\right\}.$$

For $\log n < d < n$,

$$R = \exp\left\{-\Theta\left(\frac{d}{\log(2n/d)}\right)\right\}.$$

For $d \geqslant n$,

$$R = 0.$$

Moreover, the rational approximant is constructed explicitly in each case.

Theorem 16.19 is the main result of this section. We establish it in the next two subsections, giving separate treatment to the cases $d \leqslant \log n$ and $d > \log n$ (see Theorems 16.21 and 16.26, respectively). In the concluding subsection, we give the promised proof that $R^+(d, \{\pm 1, \ldots, \pm n\})$ and $R^+(\mathrm{MAJ}_n, d)$ are essentially equivalent.

Low-degree approximation

We start by specializing the criterion of Theorem 16.15 to the problem of approximating the sign function on the set $\{\pm 1, \pm 2, \ldots, \pm n\}$.

THEOREM 16.20 (Sherstov [200]). *Let d be an integer, $0 \leqslant d \leqslant 2n-1$. Fix a nonempty subset $S \subseteq \{1, 2, \ldots, n\}$. Suppose that there exists a real $\delta \in (0,1)$ and a polynomial $r \in P_{2n-d-1}$ that vanishes on $\{-n, \ldots, n\} \setminus (S \cup -S)$ and obeys*

$$(-1)^t r(t) > \delta |r(-t)|, \qquad t \in S. \tag{16.27}$$

Then

$$R^+(d, S \cup -S) \geqslant \frac{2\delta}{1+\delta}. \tag{16.28}$$

PROOF. Define $f : S \cup -S \to \{-1, +1\}$ by $f(t) = \operatorname{sgn} t$. Define $\psi : S \cup -S \to \mathbb{R}$ by $\psi(t) = (-1)^t \binom{2n}{n+t} r(t)$. Then (16.27) takes on the form

$$\psi(t) > \delta |\psi(-t)|, \qquad t \in S. \tag{16.29}$$

For every polynomial u of degree at most d, we have

$$\sum_{S \cup -S} \psi(t) u(t) = \sum_{t=-n}^{n} (-1)^t \binom{2n}{n+t} r(t) u(t) = 0 \tag{16.30}$$

by Fact 16.6. Now (16.28) follows from (16.29), (16.30), and Theorem 16.15. □

Using Theorem 16.20, we will now determine the optimal error in the approximation of the majority function by rational functions of degree up to $\log n$. The case of higher degrees will be settled in the next subsection.

THEOREM 16.21 (Sherstov [200]). *Let d be an integer, $1 \leqslant d \leqslant \log n$. Then*

$$\exp\left\{-\Theta\left(\frac{1}{n^{1/(2d)}}\right)\right\} \leqslant R^+(d, \{\pm 1, \pm 2, \ldots, \pm n\}) < \exp\left\{-\frac{1}{n^{1/d}}\right\}.$$

PROOF. The upper bound is immediate from Newman's Theorem 16.7. For the lower bound, put $\Delta = \lfloor n^{1/d} \rfloor \geqslant 2$ and $S = \{1, \Delta, \Delta^2, \ldots, \Delta^d\}$. Define $r \in P_{2n-d-1}$ by

$$r(t) = (-1)^n \prod_{i=0}^{d-1} (t - \Delta^i \sqrt{\Delta}) \prod_{i \in \{-n,\ldots,n\} \setminus (S \cup -S)} (t - i).$$

For $j = 0, 1, 2, \ldots, d$,

$$\frac{|r(\Delta^j)|}{|r(-\Delta^j)|} = \prod_{i=0}^{j-1} \frac{\Delta^j - \Delta^i \sqrt{\Delta}}{\Delta^j + \Delta^i \sqrt{\Delta}} \prod_{i=j}^{d-1} \frac{\Delta^i \sqrt{\Delta} - \Delta^j}{\Delta^i \sqrt{\Delta} + \Delta^j} > \left(\prod_{i=1}^{\infty} \frac{\Delta^{i/2} - 1}{\Delta^{i/2} + 1} \right)^2$$
$$> \exp\left\{-5 \sum_{i=1}^{\infty} \frac{1}{\Delta^{i/2}}\right\} > \exp\left\{-\frac{18}{\sqrt{\Delta}}\right\},$$

where we used the bound $(a-1)/(a+1) > \exp(-2.5/a)$, valid for $a \geqslant \sqrt{2}$. Since $\operatorname{sgn} r(t) = (-1)^t$ for $t \in S$, we conclude that

$$(-1)^t r(t) > \exp\left\{-\frac{18}{\sqrt{\Delta}}\right\} |r(-t)|, \qquad t \in S.$$

Since in addition r vanishes on $\{-n,\dots,n\}\setminus(S\cup -S)$, we infer from Theorem 16.20 that $R^+(d,S\cup -S)\geqslant \exp\{-18/\sqrt{\Delta}\}$. □

High-degree approximation

In the previous subsection, we determined the least error in approximating the majority function by rational functions of degree up to $\log n$. Our goal here is to solve the case of higher degrees.

We start with some preparatory work. First, we need to accurately estimate products of the form $\prod_i(\Delta^i+1)/(\Delta^i-1)$ for all $\Delta>1$. A suitable *lower* bound was already given by Newman [156, Lem. 1]:

LEMMA 16.22 (Newman [156]). *For all $\Delta>1$,*

$$\prod_{i=1}^{n}\frac{\Delta^i+1}{\Delta^i-1}>\exp\left\{\frac{2(\Delta^n-1)}{\Delta^n(\Delta-1)}\right\}.$$

PROOF. Immediate from the bound $(a+1)/(a-1)>\exp(2/a)$, which is valid for $a>1$. □

We will need a corresponding upper bound:

LEMMA 16.23 (Sherstov [200]). *For all $\Delta>1$,*

$$\prod_{i=1}^{\infty}\frac{\Delta^i+1}{\Delta^i-1}<\exp\left\{\frac{4}{\Delta-1}\right\}.$$

PROOF. Let $k\geqslant 0$ be an integer. By the binomial theorem, $\Delta^i\geqslant(\Delta-1)i+1$ for integers $i\geqslant 0$. As a result,

$$\prod_{i=1}^{k}\frac{\Delta^i+1}{\Delta^i-1}\leqslant\prod_{i=1}^{k}\frac{1}{i}\left(i+\frac{2}{\Delta-1}\right)\leqslant\binom{k+\lceil\frac{2}{\Delta-1}\rceil}{k}.$$

Also,

$$\prod_{i=k+1}^{\infty} \frac{\Delta^i + 1}{\Delta^i - 1} < \prod_{i=0}^{\infty} \left(1 + \frac{2}{(\Delta^{k+1} - 1)\Delta^i}\right) < \exp\left\{\frac{2\Delta}{(\Delta^{k+1} - 1)(\Delta - 1)}\right\}.$$

Setting $k = k(\Delta) = \left\lfloor \frac{2}{\Delta - 1} \right\rfloor$, we conclude that

$$\prod_{i=1}^{\infty} \frac{\Delta^i + 1}{\Delta^i - 1} < \exp\left\{\frac{C}{\Delta - 1}\right\},$$

where

$$C = \sup_{\Delta > 1} \left\{ (\Delta - 1) \ln \binom{k(\Delta) + \left\lceil \frac{2}{\Delta - 1} \right\rceil}{k(\Delta)} + \frac{2\Delta}{\Delta^{k(\Delta)+1} - 1} \right\} < 4. \qquad \square$$

We will also need the following binomial estimate.

LEMMA 16.24 (Sherstov [200]). *Put* $p(t) = \prod_{i=1}^{n} \left(t - i - \frac{1}{2}\right)$. *Then*

$$\max_{t=1,2,\ldots,n+1} \left| \frac{p(-t)}{p(t)} \right| \leqslant \Theta(16^n).$$

PROOF. For $t = 1, 2, \ldots, n+1$, we have

$$|p(t)| = \frac{(2t-2)!(2n-2t+2)!}{4^n (t-1)!(n-t+1)!}, \quad |p(-t)| = \frac{t!(2n+2t+1)!}{4^n (2t+1)!(n+t)!}.$$

As a result,

$$\left|\frac{p(-t)}{p(t)}\right| = \frac{t}{2t+1}\cdot\frac{\binom{2n+2t+1}{2t}\binom{2n+1}{n+t}}{\binom{2t-2}{t-1}\binom{2n-2t+2}{n-t+1}} \leqslant \frac{\Theta\left(\frac{2^{4n}}{\sqrt{n}}\right)\Theta\left(\frac{2^{2n}}{\sqrt{n}}\right)}{\Theta\left(\frac{2^{2n}}{n}\right)},$$

which gives the sought bound. □

Our construction requires one additional ingredient.

LEMMA 16.25 (Sherstov [200]). *Let n, d be integers, $1 \leqslant d \leqslant n/55$. Put $p(t) = \prod_{i=1}^{d-1}(t - d\Delta^i\sqrt{\Delta})$, where $\Delta = (n/d)^{1/d}$. Then*

$$\min_{j=1,\ldots,d}\left|\frac{p(\lfloor d\Delta^j\rfloor)}{p(-\lfloor d\Delta^j\rfloor)}\right| > \exp\left\{-\frac{4\ln 3d}{\ln(n/d)} - \frac{8}{\sqrt{\Delta}-1}\right\}.$$

PROOF. Fix $j = 1, 2, \ldots, d$. Then for each $i = 1, 2, \ldots, j-1$,

$$d\Delta^j - d\Delta^i\sqrt{\Delta} \geqslant d\left(\Delta^{j-i-\frac{1}{2}} - 1\right) \geqslant \frac{1}{2}(j-i)\ln\frac{n}{d},$$

and thus

$$\begin{aligned}\prod_{i=1}^{j-1}\left(1 - \frac{1}{d\Delta^j - d\Delta^i\sqrt{\Delta}}\right) &\geqslant \exp\left\{-\frac{4}{\ln(n/d)}\sum_{i=1}^{j-1}\frac{1}{j-i}\right\} \\ &\geqslant \exp\left\{-\frac{4\ln 3d}{\ln(n/d)}\right\}. \qquad (16.31)\end{aligned}$$

For brevity, let ξ stand for the final expression in (16.31). Since $1 \leqslant d \leqslant n/55$, we have $\lfloor d\Delta^j \rfloor - d\Delta^{j-1}\sqrt{\Delta} > 1$. As a result,

$$\begin{aligned}
\left| \frac{p(\lfloor d\Delta^j \rfloor)}{p(-\lfloor d\Delta^j \rfloor)} \right| &\geqslant \prod_{i=1}^{j-1} \frac{d\Delta^j - 1 - d\Delta^i\sqrt{\Delta}}{d\Delta^j + d\Delta^i\sqrt{\Delta}} \prod_{i=j}^{d-1} \frac{d\Delta^i\sqrt{\Delta} - d\Delta^j}{d\Delta^i\sqrt{\Delta} + d\Delta^j} \\
&\geqslant \xi \prod_{i=1}^{j-1} \frac{d\Delta^j - d\Delta^i\sqrt{\Delta}}{d\Delta^j + d\Delta^i\sqrt{\Delta}} \prod_{i=j}^{d-1} \frac{d\Delta^i\sqrt{\Delta} - d\Delta^j}{d\Delta^i\sqrt{\Delta} + d\Delta^j} \qquad \text{by (16.31)} \\
&> \xi \left(\prod_{i=1}^{\infty} \frac{\Delta^{i/2} - 1}{\Delta^{i/2} + 1} \right)^2 \\
&\geqslant \xi \exp\left\{ -\frac{8}{\sqrt{\Delta} - 1} \right\},
\end{aligned}$$

where the last inequality holds by Lemma 16.23. □

We have reached the main result of this subsection.

THEOREM 16.26 (Sherstov [200]). *Let d be an integer, $\log n < d \leqslant n - 1$. Then*

$$R^+(d, \{\pm 1, \pm 2, \ldots, \pm n\}) = \exp\left\{ -\Theta\left(\frac{d}{\log(2n/d)} \right) \right\}.$$

Also,

$$R^+(n, \{\pm 1, \pm 2, \ldots, \pm n\}) = 0.$$

PROOF. The final statement in the theorem follows at once by considering the rational function $\{p(t) - p(-t)\}/\{p(t) + p(-t)\}$, where $p(t) = \prod_{i=1}^{n}(t + i)$.

Now assume that $\log n < d < n/55$. Let

$$k = \left\lceil \frac{d}{\log(n/d)} \right\rceil, \qquad \Delta = \left(\frac{n}{d}\right)^{1/d}.$$

Define sets

$$\begin{aligned} S_1 &= \{1, 2, \ldots, k\}, \\ S_2 &= \{\lfloor d\Delta^i \rfloor : i = 1, 2, \ldots, d\}, \\ S &= S_1 \cup S_2. \end{aligned}$$

Consider the polynomial

$$r(t) = (-1)^n r_1(t) r_2(t) \prod_{i \in \{-n,\ldots,n\} \setminus (S \cup -S)} (t - i),$$

where

$$r_1(t) = \prod_{i=1}^{k} \left(t - i - \frac{1}{2} \right), \qquad r_2(t) = \prod_{i=1}^{d-1} (t - d\Delta^i \sqrt{\Delta}).$$

We have:

$$\begin{aligned} \min_{t \in S} \left| \frac{r(t)}{r(-t)} \right| &\geqslant \min_{i=1,\ldots,k+1} \left| \frac{r_1(i)}{r_1(-i)} \right| \cdot \min_{i=1,\ldots,d} \left| \frac{r_2(\lfloor d\Delta^i \rfloor)}{r_2(-\lfloor d\Delta^i \rfloor)} \right| \\ &> \exp\left\{ -\frac{Cd}{\log(n/d)} \right\} \end{aligned}$$

by Lemmas 16.24 and 16.25, where $C > 0$ is an absolute constant. Since $\operatorname{sgn} p(t) = (-1)^t$ for $t \in S$, we can restate this result as follows:

$$(-1)^t r(t) > \exp\left\{ -\frac{Cd}{\log(n/d)} \right\} |r(-t)|, \qquad t \in S.$$

Since r vanishes on $\{-n,\dots,n\}\setminus(S\cup -S)$ and has degree $\leqslant 2n-1-d$, we infer from Theorem 16.20 that $R^+(d,S\cup -S)\geqslant \exp\{-Cd/\log(n/d)\}$. This proves the lower bound for the case $\log n < d < n/55$.

To handle the case $n/55\leqslant d\leqslant n-1$, a different argument is needed. Let

$$r(t)=(-1)^n\, t\prod_{i=1}^{d}\left(t-i-\frac{1}{2}\right)\prod_{i=d+2}^{n}(t^2-i^2).$$

By Lemma 16.24, there is an absolute constant $C>1$ such that

$$\left|\frac{r(t)}{r(-t)}\right|>C^{-d},\qquad t=1,2,\dots,d+1.$$

Since $\operatorname{sgn} r(t)=(-1)^t$ for $t=1,2,\dots,d+1$, we conclude that

$$(-1)^t r(t)>C^{-d}|r(-t)|,\qquad t=1,2,\dots,d+1.$$

Since the polynomial r vanishes on $\{-n,\dots,n\}\setminus\{\pm1,\pm2,\dots,\pm(d+1)\}$ and has degree $2n-1-d$, we infer from Theorem 16.20 that

$$R^+(d,\{\pm1,\pm2,\dots,\pm(d+1)\})\geqslant C^{-d}.$$

This settles the lower bound for the case $n/55\leqslant d\leqslant n-1$.

It remains to prove the upper bound for the case $\log n < d\leqslant n-1$. Here we always have $d\geqslant 2$. Letting $k=\lfloor d/2\rfloor$ and $\Delta=(n/k)^{1/k}$, define $p\in P_{2k}$ by

$$p(t)=\prod_{i=1}^{k}(t+i)\prod_{i=1}^{k}(t+k\Delta^i).$$

Fix any point $t \in \{1, 2, \ldots, n\}$ with $p(-t) \neq 0$. Letting i^* be the integer with $k\Delta^{i^*} < t < k\Delta^{i^*+1}$, we have:

$$\frac{p(t)}{|p(-t)|} > \prod_{i=0}^{i^*} \frac{k\Delta^{i^*+1} + k\Delta^i}{k\Delta^{i^*+1} - k\Delta^i} \prod_{i=i^*+1}^{k} \frac{k\Delta^i + k\Delta^{i^*}}{k\Delta^i - k\Delta^{i^*}} \geqslant \prod_{i=1}^{k} \frac{\Delta^i + 1}{\Delta^i - 1}$$
$$> \exp\left\{\frac{2(\Delta^k - 1)}{\Delta^k(\Delta - 1)}\right\},$$

where the last inequality holds by Lemma 16.22. Substituting $\Delta = (n/k)^{1/k}$ and recalling that $k \geqslant \Theta(\log n)$, we obtain $p(t) > A|p(-t)|$ for $t = 1, 2, \ldots, n$, where

$$A = \exp\left\{\Theta\left(\frac{k}{\log(n/k)}\right)\right\}.$$

As a result, $R^+(2k, \{\pm 1, \pm 2, \ldots, \pm n\}) \leqslant 2A/(A^2 + 1)$, the approximant in question being

$$\frac{A^2 - 1}{A^2 + 1} \cdot \frac{p(t) - p(-t)}{p(t) + p(-t)}.$$ □

Equivalence of the majority and sign functions

It remains to prove the promised equivalence of the majority and sign functions, from the standpoint of approximating them by rational functions on the discrete domain. We have:

THEOREM 16.27. *For every integer d,*

$$R^+(\mathrm{MAJ}_n, d) \leqslant R^+(d - 2, \{\pm 1, \pm 2, \ldots, \pm\lceil n/2\rceil\}), \tag{16.32}$$
$$R^+(\mathrm{MAJ}_n, d) \geqslant R^+(d, \{\pm 1, \pm 2, \ldots, \pm\lfloor n/2\rfloor\}). \tag{16.33}$$

PROOF. We prove (16.32) first. Fix a degree-$(d-2)$ approximant $p(t)/q(t)$ to $\operatorname{sgn} t$ on $S = \{\pm 1, \ldots, \pm\lceil n/2 \rceil\}$, where q is positive on S. For small $\delta > 0$, define

$$A_\delta(t) = -\frac{t^2 p(t) - \delta}{t^2 q(t) + \delta}.$$

Then A_δ is a rational function of degree at most d whose denominator is positive on $S \cup \{0\}$. Finally, we have $A_\delta(0) = 1$ and

$$\lim_{\delta \to 0} \max_{t \in S} |-\operatorname{sgn} t - A_\delta(t)| = \max_{t \in S} \left| \operatorname{sgn} t - \frac{p(t)}{q(t)} \right|,$$

which yields the needed approximant for $\mathrm{MAJ}_n(x)$, namely, $A_\delta(\sum x_i - \lfloor n/2 \rfloor)$.

We now turn to the proof of the lower bound, (16.33). Fix a degree-d approximant $P(x)/Q(x)$ for MAJ_n, where $Q(x) > 0$ for $x \in \{0,1\}^n$. Let ε denote the error of this approximant, $\varepsilon \leqslant 1$. Then

$$(1-\varepsilon)Q(x) \leqslant P(x) \leqslant (1+\varepsilon)Q(x)$$

whenever $\sum x_i \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$, and

$$(1-\varepsilon)Q(x) \leqslant -P(x) \leqslant (1+\varepsilon)Q(x)$$

whenever $\sum x_i \in \{\lfloor n/2 \rfloor + 1, \ldots, n\}$. Now, Proposition 2.2 guarantees the existence of univariate polynomials $p, q \in P_d$ such that for all $x \in \{0,1\}^n$, one has $p(\sum x_i) = \mathbf{E}_{\sigma \in S_n}[P(\sigma x)]$ and $q(\sum x_i) = \mathbf{E}_{\sigma \in S_n}[Q(\sigma x)]$. In view of the previous two inequalities for P and Q, we obtain:

$$\begin{aligned} (1-\varepsilon)q(t) \leqslant p(t) \leqslant (1+\varepsilon)q(t), &\qquad t = 0, 1, \ldots, \lfloor n/2 \rfloor; \\ (1-\varepsilon)q(t) \leqslant -p(t) \leqslant (1+\varepsilon)q(t), &\qquad t = \lfloor n/2 \rfloor + 1, \ldots, n. \end{aligned}$$

Thus,

$$\max_{t=\pm 1,\pm 2,\ldots,\pm\lfloor n/2\rfloor} \left| \operatorname{sgn} t - \frac{-p(t+\lfloor n/2\rfloor)}{q(t+\lfloor n/2\rfloor)} \right| \leqslant \varepsilon .$$

Since q is positive on $\{0, 1, \ldots, n\}$, this completes the proof of (16.33). □

REMARK 16.28 (Sherstov [200]). The proof that we gave for the upper bound, (16.32), illustrates a useful property of univariate rational approximants $A(t) = p(t)/q(t)$ on a finite set S. Specifically, given such an approximant and a point $t^* \notin S$, there exists an approximant $\tilde{A}$ with $\tilde{A}(t^*) = a$ for any prescribed value a and $\tilde{A} \approx A$ everywhere on S. One such construction is

$$\tilde{A}(t) = \frac{(t-t^*)p(t) + a\delta}{(t-t^*)q(t) + \delta}$$

for an arbitrarily small constant $\delta > 0$. Note that $\tilde{A}$ has degree only 1 higher than the degree of the original approximant, A. This phenomenon is in sharp contrast to approximation by polynomials, which do not possess this corrective ability.

16.8 Threshold degree of the intersections of two halfspaces

In this section, we prove our main results on the sign-representation of intersections of halfspaces and majority functions. The following elegant observation, described informally in Section 15.1, relates sign-representation and rational approximation.

THEOREM 16.29 (Beigel et al. [33]). *Let $f : X \to \{-1, +1\}$ and $g : Y \to \{-1, +1\}$ be given functions, where $X, Y \subset \mathbb{R}^n$ are finite sets. Let d be an integer with $R^+(f, d) + R^+(g, d) < 1$. Then*

$$\deg_{\pm}(f \wedge g) \leqslant 2d.$$

PROOF. Fix rational functions $p_1(x)/q_1(x)$ and $p_2(y)/q_2(y)$ of degree at most d such that q_1 and q_2 are positive on X and Y, respectively, and

$$\max_{x\in X}\left|f(x)-\frac{p_1(x)}{q_1(x)}\right|+\max_{y\in Y}\left|g(y)-\frac{p_2(y)}{q_2(y)}\right|<1.$$

Then

$$f(x)\wedge g(y)\equiv \operatorname{sgn}\{1+f(x)+g(y)\}\equiv \operatorname{sgn}\left\{1+\frac{p_1(x)}{q_1(x)}+\frac{p_2(y)}{q_2(y)}\right\}.$$

Multiplying the last expression by the positive quantity $q_1(x)q_2(y)$, we obtain $f(x)\wedge g(y)\equiv \operatorname{sgn}\{q_1(x)q_2(y)+p_1(x)q_2(y)+p_2(y)q_1(x)\}$. □

Recall that we established a converse to Theorem 16.29, namely, Theorem 15.9 in the previous chapter. We are now in a position to prove the main results of this section, stated as Theorems 16.1–16.3 above.

THEOREM 16.30 (Sherstov [200]). *Let* $f:\{-1,+1\}^{n^2}\to\{-1,+1\}$ *be given by*

$$f(x)=\operatorname{sgn}\left(1+\sum_{i=1}^{n}\sum_{j=1}^{n}2^i x_{ij}\right).$$

Then

$$\deg_{\pm}(f\wedge f)=\Omega(n), \tag{16.34}$$

$$\deg_{\pm}(\mathrm{MAJ}_n\wedge \mathrm{MAJ}_n)=\Omega(\log n). \tag{16.35}$$

PROOF. By Theorem 16.18, we have $R^+(f,\varepsilon n)\geqslant 1/2$ for some constant $\varepsilon>0$, which settles (16.34) in view of Theorem 15.9.

Analogously, Theorems 16.19 and 16.27 show that $R^+(\mathrm{MAJ}_n,\varepsilon\log n)\geqslant 1/2$ for some constant $\varepsilon>0$, which settles (16.35) in view of Theorem 15.9. □

REMARK 16.31 (Sherstov [200]). The lower bounds (16.34) and (16.35) are tight and match the constructions due to Beigel et al. [33]. These matching upper bounds can be proved as follows. By Theorem 16.18, we have $R^+(f, Cn) < 1/2$ for some constant $C > 0$, which shows that $\deg_\pm(f \wedge f) = O(n)$ by Theorem 16.29. Analogously, Theorems 16.19 and 16.27 imply that $R^+(g, C \log n) < 1/2$ for some constant $C > 0$, which shows that $\deg_\pm(g \wedge g) = O(\log n)$ by Theorem 16.29.

Furthermore, Theorem 16.29 generalizes immediately to the conjunction of $k \geqslant 3$ functions. In particular, the lower bounds in (16.34) and (16.35) remain tight for intersections $f \wedge f \wedge \cdots \wedge f$ and $g \wedge g \wedge \cdots \wedge g$ featuring any constant number of functions.

We give one additional result, featuring the intersection of the canonical halfspace with a majority function.

THEOREM 16.32 (Sherstov [200]). *Let* $f\colon \{-1,+1\}^{n^2} \to \{-1,+1\}$ *be given by*

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{n} 2^i x_{ij}\right).$$

Let $g\colon \{0,1\}^{\lceil\sqrt{n}\rceil} \to \{-1,+1\}$ *be the majority function on* $\lceil\sqrt{n}\rceil$ *bits. Then*

$$\deg_\pm(f \wedge g) = \Theta(\sqrt{n}). \tag{16.36}$$

PROOF. We prove the lower bound first. Let $\varepsilon > 0$ be a suitably small constant. By Theorem 16.18, we have $R^+(f, \varepsilon\sqrt{n}) \geqslant 1 - 2^{-\sqrt{n}}$. By Theorems 16.19 and 16.27, we have $R^+(g, \varepsilon\sqrt{n}) \geqslant 2^{-\sqrt{n}}$. In view of Theorem 15.9, these two facts imply that $\deg_\pm(f \wedge g) = \Omega(\sqrt{n})$.

We now turn to the upper bound. It is clear that $R^+(g, \lceil\sqrt{n}\rceil) = 0$ and $R^+(f, 1) < 1$. It follows by Theorem 16.29 that $\deg_\pm(f \wedge g) = O(\sqrt{n})$. □

Theorems 16.30 and 16.32 are of course also valid with respect to disjunctions rather than conjunctions.

16.9 Threshold density revisited

Recall from Section 13.3 that threshold density is another complexity measure of sign-representations that is of interest in computational learning theory. In this section, we will prove that the intersections of halfspaces in Theorems 16.30 and 16.32 have not only high threshold degree but also high threshold density. As a starting point, we derive a tight lower bound for the intersection of two majority functions.

THEOREM 16.33 (Sherstov [200]). *The majority function satisfies*

$$\operatorname{dns}(\mathrm{MAJ}_n \wedge \mathrm{MAJ}_n) = n^{\Omega(\log n)}.$$

PROOF. The idea is to mimic the proof of Theorem 13.12, where we obtained a somewhat weaker density bound of $n^{\Omega(\log n/\log\log n)}$. We will be able to strengthen it by using our newly available lower bound of $\Omega(\log n)$ on the threshold degree of the intersection of two majority functions.

Let t and k be integers to be fixed later. Define $f\colon\{0,1\}^{2t} \to \{-1,+1\}$ by $f(x) = \mathrm{MAJ}(x_1,\ldots,x_t) \wedge \mathrm{MAJ}(x_{t+1},\ldots,x_{2t})$. Consider the function $f^{\oplus}\colon(\{0,1\}^k)^{2t} \to \{-1,+1\}$ given by

$$f^{\oplus}(x) = f\left(\ldots,\bigoplus_{j=1}^{k} x_{i,j},\ldots\right).$$

Lemma 13.11 implies that

$$\deg_{\pm}(f^{\oplus}) = k\deg_{\pm}(f).$$

Consider now the function $f^{\oplus\,\mathrm{KP}}$, where the $^{\mathrm{KP}}$ operator is as described in Section 13.4. For bits $a,b,c \in \{0,1\}$, we have

$$(\overline{c} \wedge a) \vee (c \wedge b) = \frac{1+(-1)^c}{2}\cdot(-1)^a + \frac{1-(-1)^c}{2}\cdot(-1)^b.$$

As a result,

$$f^{\oplus \mathrm{KP}}(x, y, z) \equiv \left(\prod_{i=1}^{k} q_{1,i} + \cdots + \prod_{i=1}^{k} q_{t,i} \geqslant 0\right) \bigwedge \left(\prod_{i=1}^{k} q_{t+1,i} + \cdots + \prod_{i=1}^{k} q_{2t,i} \geqslant 0\right),$$

where $q_{i,j} = (1 + (-1)^{z_{i,j}})(-1)^{x_{i,j}} + (1 - (-1)^{z_{i,j}})(-1)^{y_{i,j}}$.

The above construction shows that $f^{\oplus \mathrm{KP}}$ is computed by the intersection of two functions with threshold weight at most $2t4^k + 1$ each. Lemma 13.8 implies that if the intersection of two majorities, each on a distinct set of $2t4^k + 1$ variables, has threshold density at most L, then $\mathrm{dns}(f^{\oplus \mathrm{KP}}) \leqslant L$. Theorem 13.10, on the other hand, implies that $f^{\oplus \mathrm{KP}}$ has threshold density at least $2^{\deg_{\pm}(f^{\oplus})} = 2^{k \deg_{\pm}(f)}$. In view of (16.35) we conclude that the intersection of two majorities, each on $2t4^k + 1$ variables, has threshold density $\exp\{\Omega(k \log t)\}$. The theorem follows by setting $t = \lfloor \sqrt{n}/3 \rfloor$ and $k = \lfloor \frac{1}{4} \log n \rfloor$. □

Recall from Section 13.4 that density lower bound in Theorem 16.33 is tight and matches the construction of Beigel et al. [33]. We now turn to intersections of halfspaces with high threshold degree.

THEOREM 16.34 (Sherstov [200]). *Let* $f_n \colon \{-1, +1\}^{n^2} \to \{-1, +1\}$ *be given by*

$$f_n(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{n} \sum_{j=1}^{n} 2^i x_{ij}\right).$$

Then

$$\mathrm{dns}(f_n \wedge f_n) = \exp\{\Omega(n)\}, \tag{16.37}$$

$$\mathrm{dns}(f_n \wedge \mathrm{MAJ}_{\lceil \sqrt{n} \rceil}) = \exp\{\Omega(\sqrt{n})\}. \tag{16.38}$$

REMARK 16.35. In the proof below, it will be useful to keep in mind the following straightforward observation. Fix functions $f, g: \{-1,+1\}^k \to \{-1,+1\}$ and define functions $f', g': \{-1,+1\}^k \to \{-1,+1\}$ by $f'(x) = -f(-x)$ and $g'(y) = -g(-y)$. Then we have $f'(x) \wedge g'(y) \equiv -(f(-x) \wedge g(-y))\, f(-x) g(-y)$, whence $\mathrm{dns}(f' \wedge g') \leqslant \mathrm{dns}(f \wedge g)\, \mathrm{dns}(f)\, \mathrm{dns}(g)$ and thus

$$\mathrm{dns}(f \wedge g) \geqslant \frac{\mathrm{dns}(f' \wedge g')}{\mathrm{dns}(f)\, \mathrm{dns}(g)}. \tag{16.39}$$

Similarly, we have $f(x) \wedge g'(y) \equiv (f(x) \wedge g(-y))\, f(x)$, whence

$$\mathrm{dns}(f \wedge g) \geqslant \frac{\mathrm{dns}(f \wedge g')}{\mathrm{dns}(f)}. \tag{16.40}$$

To summarize, (16.39) and (16.40) allow one to analyze the threshold density of $f \wedge g$ by analyzing the threshold density of $f' \wedge g'$ or $f' \wedge g$ instead. Such a transition will be helpful in our case.

PROOF OF THEOREM 16.34. Put $m = \lfloor n/4 \rfloor$. It is straightforward to verify that the function ${f_m}^{\mathrm{KP}}: (\{-1,+1\}^{m^2})^3 \to \{-1,+1\}$ has the representation

$${f_m}^{\mathrm{KP}}(x,y,z) = \mathrm{sgn}\left(1 + \sum_{i=1}^{m} \sum_{j=1}^{m} 2^i (x_{ij} + y_{ij} + x_{ij} z_{ij} - y_{ij} z_{ij})\right).$$

As a result,

$$\begin{aligned}
\mathrm{dns}(f_{4m} \wedge f_{4m}) &\geqslant \mathrm{dns}({f_m}^{\mathrm{KP}} \wedge {f_m}^{\mathrm{KP}}) && \text{by Lemma 13.8} \\
&= \mathrm{dns}((f_m \wedge f_m)^{\mathrm{KP}}) \\
&\geqslant 2^{\deg_{\pm}(f_m \wedge f_m)} && \text{by Theorem 13.10} \\
&\geqslant \exp\{\Omega(m)\} && \text{by Theorem 16.30.}
\end{aligned}$$

By the same argument as in Theorem 16.18, the function f_{4m} is a subfunction of $f_n(x)$ or $-f_n(-x)$. In the former case, (16.37) is immediate from the lower bound on $\mathrm{dns}(f_{4m} \wedge f_{4m})$. In the latter case, (16.37) follows from the lower bound on $\mathrm{dns}(f_{4m} \wedge f_{4m})$ and Remark 16.35.

The proof of (16.38) is entirely analogous. □

Krause and Pudlák's method in Theorem 13.10 naturally generalizes to linear combinations of conjunctions rather than parity functions. In other words, if a function $f : \{0,1\}^n \to \{-1,+1\}$ has threshold degree d and $f^{\mathrm{KP}}(x,y,z) \equiv \mathrm{sgn}(\sum_{i=1}^{N} \lambda_i T_i(x,y,z))$ for some conjunctions $T_1, \ldots, T_N$ of the literals $x_1, y_1, z_1, \ldots, x_n, y_n, z_n, \neg x_1, \neg y_1, \neg z_1, \ldots, \neg x_n, \neg y_n, \neg z_n$, then $N \geqslant 2^{\Omega(d)}$. With this remark in mind, Theorems 16.33 and 16.34 and their proofs adapt easily to this alternate definition of density.

Chapter 17

Conclusions and Open Problems

17.1 Our contributions in learning theory

In Chapters 12–16, we answered several fundamental questions in computational learning theory, ruling out efficient learning algorithms for well-studied concept classes in the PAC model, statistical query model, and agnostic model. Our analytic approach additionally allowed us to exhibit several new relations between learning theory and communication complexity. A more detailed review follows.

First, we proved that a polynomial-time algorithm for PAC learning the intersection of n^ε halfspaces on $\{0,1\}^n$ with respect to arbitrary distributions would violate standard cryptographic assumptions. More precisely, we related the problem of learning intersections of halfspaces to shortest-vector problems on lattices, which are believed to be intractable.

Second, we obtained an unconditional, exponential lower bound on the complexity of learning the intersection of n^ε halfspaces in Kearns' statistical query model. This result complements and is incomparable with our hardness result for the PAC model: the statistical query result does not rely on any complexity-theoretic assumptions, whereas the PAC result applies to a more powerful learning model.

These two results left open the possibility of an efficient learning algorithm for the intersection of k halfspaces for k small, such as $k = 2$. We addressed this question in the context of sign-representation by polynomials. In particular, we proved a lower bound of $\Omega(\sqrt{n})$ on the threshold degree of the intersection of two halfspaces on $\{0,1\}^n$, which is an exponential improvement on previous work [153, 163] and solves an open problem posed by Klivans [120]. This result exposes the limitations of polynomial-based techniques in computational learning and points to the need for techniques that do not rely on polynomials.

Fourth, we studied the problem of learning natural concept classes in the agnostic model, which allows adversarial corruption of the training data. Our main conclusion was that the known techniques based on the approximate rank [101] and low-degree polynomials [211] require exponential time to learn concept classes as simple as decision lists and disjunctions.

Finally, our analytic approach has allowed us to exploit known relations between learning theory and communication complexity as well as to discover new ones. In Chapter 8, we studied the unbounded-error communication complexity of AC^0 and obtained as a corollary an exponential lower bound on the sign-rank of

linear-size DNF formulas, improving on the previous quasipolynomial lower bound and essentially matching the known upper bound [122]. In Chapter 10, we studied relations among key complexity measures of a communication problem and in so doing discovered the equivalence of product discrepancy and the statistical query dimension. As another illustration, our proofs in Chapter 14 on agnostic learning heavily exploit communication techniques from Part I of this thesis, most notably the pattern matrix method.

17.2 Open problems

There are several inviting avenues for future work, both on the complexity-theoretic front and on the algorithmic front. A starting point in our discussion is uniform-distribution learning.

OPEN PROBLEM 1. Design a polynomial-time algorithm for PAC learning DNF formulas of polynomial size with respect to the uniform distribution, from random examples only.

Jackson's Harmonic sieve [98] solves this learning problem in polynomial time but only if query access is available to the unknown function. A counterpart to the Harmonic sieve is Verbeurgt's algorithm [217] that uses random examples only but runs in worst-case time $n^{\Theta(\log n)}$.

Another open problem in uniform-distribution learning concerns intersections of halfspaces. Klivans et al. [121] gave a polynomial-time algorithm for learning intersections of any constant number of halfspaces with respect to the uniform distribution, from random examples alone. Extending this result to a superconstant number of halfspaces on $\{0,1\}^n$ would be of great interest. In fact, even a weak learning algorithm for the problem would be of interest.

OPEN PROBLEM 2. Design a polynomial-time algorithm for learning the intersection of $\omega(1)$ halfspaces on $\{0,1\}^n$ to accuracy $1/2 - n^{-O(1)}$ with respect to the uniform distribution.

The Fourier spectrum of the intersection of $\omega(1)$ halfspaces on $\{0,1\}^n$ is not as well understood as that of polynomial-size DNF formulas. While the inter-

section of $O(1)$ halfspaces always has a Fourier coefficient of magnitude $n^{-O(1)}$ or greater [121], the corresponding question is open for $\omega(1)$ halfspaces.

OPEN PROBLEM 3. Prove or disprove: the intersection of k halfspaces on $\{0,1\}^n$ has a Fourier coefficient of magnitude at least $n^{-O(1)}$, for some function $k = k(n)$ with $k = \omega(1)$.

It is of great interest to place a nontrivial upper bound on the statistical query dimension of polynomial-size DNF formulas and, more generally, AC^0 circuits. Apart from the significance of such results in learning theory, they would have far-reaching consequences in communication complexity [198]. A more concrete formulation follows.

OPEN PROBLEM 4. Prove or disprove: polynomial-size DNF formulas have statistical query dimension $\exp\{\log^{O(1)} n\}$ under all distributions. Prove or disprove an analogous statement for the circuit class AC^0.

Similarly, placing a nontrivial upper bound on the statistical query dimension of the intersection of a constant number of halfspaces would be a significant step toward more efficient distribution-free learning.

OPEN PROBLEM 5. Determine the statistical query dimension of the intersection of a constant number of halfspaces on $\{0,1\}^n$ under worst-case distributions.

Analytic representations of natural concept classes also merit further study. We would like to note two problems that seem particularly relevant to our work in Chapter 16 and elsewhere in this thesis.

OPEN PROBLEM 6. Improve the threshold degree lower bound for the intersection of two halfspaces from $\Omega(\sqrt{n})$ to $\Omega(n)$, or show that threshold degree $o(n)$ suffices.

OPEN PROBLEM 7. Determine the threshold degree of the circuit class AC^0.

While polynomial representations of concepts have played an important role in learning theory, the lower bounds in Chapters 12–16 point to the limitations of

polynomial-based learning. Developing an alternative to this paradigm is a natural algorithmic goal for the near future.

OPEN PROBLEM 8. Develop new techniques for learning DNF formulas, intersections of halfspaces, and other natural concept classes that do not rely on representations of the concepts by polynomials.

Appendix

Appendix A

List of Symbols

For the reader's convenience, we provide a table of all notation and technical symbols used in this thesis.

Symbol	*Meaning*	*Pages*
$[n]$	the set $\{1, 2, \ldots, n\}$	23
$\mathbf{1}_S,\ e_S$	the characteristic vector of $S \subseteq \{1, 2, \ldots, n\}$	23, 32
e_i	the characteristic vector of $\{i\}$	32
$\lvert x \rvert$	the Hamming weight $\sum x_i$	24
$x\lvert_S$	projection of $x \in \{0,1\}^n$ onto the set $S \subset \{1, 2, \ldots, n\}$	23
$\mathscr{V}(n,t)$	system of t-subsets of $\{1, 2, \ldots, n\}$ in pattern matrices	66
P_d	the family of univariate polynomials of degree up to d	24
S_n	the symmetric group on n elements	26
σx	the string $(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$	26
$\hat{f}(S)$	Fourier transform of a function $f\colon \{0,1\}^n \to \mathbb{R}$	25
fg	pointwise product of functions $f, g\colon \{0,1\}^n \to \mathbb{R}$	26
χ_S	character of the Fourier transform on $\mathbb{Z}_2^n$	25
$E(f,d)$	least error in a degree-d uniform approximation of f	27
$R(f,d)$	least error in a uniform approximation of f by a degree-d rational function	323
$R^+(f,d)$	least error in a uniform approximation of f by a degree-d rational function with positive denominator	323
$W(f,d)$	degree-d threshold weight of f	28
$\mathrm{dns}(f,d)$	degree-d threshold density of f	28
$W(f)$	threshold weight of f	28
$\mathrm{dns}(f)$	threshold density of f	29

Symbol	*Meaning*	*Pages*
$\deg_\varepsilon(f)$	ε-approximate degree of f	32
$\deg_\pm(f)$	threshold degree of f	32
$\mathrm{mon}(f)$	monomial count of f	32
$\mathrm{s}(f)$	sensitivity of f	32
$\mathrm{bs}(f)$	block sensitivity of f	32
$\mathrm{bs}_\ell(f)$	ℓ-block sensitivity of f	32
$\mathrm{zbs}(f)$	zero block sensitivity of f	32
$\mathrm{dt}(f)$	decision tree complexity of f	33
$\mathrm{rk}\, A$	rank of a real matrix A	35
$\mathrm{rk}_\varepsilon\, A$	ε-approximate rank of a real matrix A	36
$\mathrm{rk}_\pm\, A$	sign-rank of a real matrix A	37
$\mathrm{tr}\, A$	trace of a square real matrix A	36
$\sigma_i(A)$	the ith largest singular value of a real matrix A	35
$\langle A, B\rangle$	inner product of real matrices or tensors A and B	36
$A \circ B$	Hadamard product of real matrices or tensors A and B	36
$A \otimes B$	Kronecker product of real matrices A and B	36
$\|\cdot\|_\infty$	ℓ_∞ norm on real functions and matrices	24, 35
$\|\cdot\|_1$	ℓ_1 norm on real functions and matrices	24, 35
$\|\cdot\|$	Euclidean norm on vectors or spectral norm on matrices	36
$\|\cdot\|_\mathrm{F}$	Frobenius norm on matrices	35
$\|\cdot\|_\Sigma$	trace norm on matrices	35
$\|\cdot\|_{\Sigma,\varepsilon}$	ε-approximate trace norm on matrices	36
$\mathrm{vc}(A)$	Vapnik-Chervonenkis dimension of the sign matrix A	40
$\mathrm{mc}(A)$	margin complexity of the sign matrix A	42
$\mathrm{sq}(A)$	statistical query (SQ) dimension of the sign matrix A	42
$D(f)$	deterministic communication complexity of f	48
$N(f)$	nondeterministic communication complexity of f	48
$R_\varepsilon(f)$	ε-error randomized communication complexity of f	48
$D^\mu_\varepsilon(f)$	ε-error μ-distributional communication complexity of f	50
$Q_\varepsilon(f)$	ε-error quantum communication complexity of f without prior entanglement	98
$Q^*_\varepsilon(f)$	ε-error quantum communication complexity of f with prior entanglement	98

Symbol	*Meaning*	*Pages*
$U(f)$	unbounded-error communication complexity of f	136
$MA_\varepsilon(f)$	ε-error Merlin-Arthur communication complexity of f	196
$\mathrm{disc}_\mu(f)$	discrepancy of f with respect to μ	49
$\mathrm{disc}(f)$	minimum discrepancy of f under any distribution	49
$\mathrm{disc}^\times(f)$	minimum discrepancy of f under a product distribution	49
P^{cc}	sign matrices/tensors with low deterministic complexity	57, 196
$\mathsf{NP}^{\mathsf{cc}}$	sign matrices/tensors with low nondeterministic complexity	59, 196
$\mathsf{coNP}^{\mathsf{cc}}$	sign matrices/tensors with low co-nondeterministic complexity	59, 196
$\mathsf{BPP}^{\mathsf{cc}}$	sign matrices/tensors with low randomized complexity	57, 196
$\mathsf{MA}^{\mathsf{cc}}$	sign matrices/tensors with low Merlin-Arthur complexity	196
$\mathsf{coMA}^{\mathsf{cc}}$	sign matrices/tensors with low co-Merlin-Arthur complexity	196
$\mathsf{PP}^{\mathsf{cc}}$	sign matrices with nonnegligible discrepancy	57
$\mathsf{UPP}^{\mathsf{cc}}$	sign matrices with low sign-rank	58
$\mathsf{PH}^{\mathsf{cc}}$	polynomial hierarchy in communication	59
Σ_k^{cc}, Π_k^{cc}	kth level of the polynomial hierarchy	59
$\mathsf{PSPACE}^{\mathsf{cc}}$	polynomial space in communication	59

Bibliography

[1] S. Aaronson. Quantum computing, postselection, and probabilistic polynomial-time. *Proceedings of the Royal Society A*, 461(2063):3473–3482, 2005.

[2] S. Aaronson. The polynomial method in quantum and classical computing. In *Proc. of the 49th Symposium on Foundations of Computer Science (FOCS)*, page 3, 2008.

[3] S. Aaronson and A. Ambainis. Quantum search of spatial regions. *Theory of Computing*, 1(1):47–79, 2005.

[4] S. Aaronson and Y. Shi. Quantum lower bounds for the collision and the element distinctness problems. *J. ACM*, 51(4):595–605, 2004.

[5] H. Aizenstein, A. Blum, R. Khardon, E. Kushilevitz, L. Pitt, and D. Roth. On learning read-k-satisfy-j DNF. *SIAM J. Comput.*, 27(6):1515–1530, 1998.

[6] H. Aizenstein, L. Hellerstein, and L. Pitt. Read-thrice DNF is hard to learn with membership and equivalence queries. In *Proc. of the 33rd Symposium on Foundations of Computer Science (FOCS)*, pages 523–532, 1992.

[7] H. Aizenstein and L. Pitt. Exact learning of read-twice DNF formulas. In *Proc. of the 32nd Symposium on Foundations of Computer Science (FOCS)*, pages 170–179, 1991.

[8] H. Aizenstein and L. Pitt. Exact learning of read-k disjoint DNF and not-so-disjoint DNF. In *Proc. of the 5th Conf. on Computational Learning Theory (COLT)*, pages 71–76, 1992.

[9] M. Ajtai and C. Dwork. A public-key cryptosystem with worst-case/average-case equivalence. In *Proc. of the 29th Symposium on Theory of Computing (STOC)*, pages 284–293, 1997.

[10] M. Alekhnovich, M. Braverman, V. Feldman, A. R. Klivans, and T. Pitassi. Learnability and automatizability. In *Proc. of the 45th Symposium on Foundations of Computer Science (FOCS)*, pages 621–630, 2004.

[11] E. Allender. A note on the power of threshold circuits. In *Proc. of the 30th Symposium on Foundations of Computer Science (FOCS)*, pages 580–584, 1989.

[12] N. Alon. Problems and results in extremal combinatorics, Part I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[13] N. Alon, P. Frankl, and V. Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *Proc. of the 26th Symposium on Foundations of Computer Science (FOCS)*, pages 277–280, 1985.

[14] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[15] A. Ambainis. Polynomial degree and lower bounds in quantum complexity: Collision and element distinctness with small range. *Theory of Computing*, 1(1):37–46, 2005.

[16] A. Ambainis, L. J. Schulman, A. Ta-Shma, U. V. Vazirani, and A. Wigderson. The quantum communication complexity of sampling. *SIAM J. Comput.*, 32(6):1570–1585, 2003.

[17] D. Angluin. Negative results for equivalence queries. *Machine Learning*, 5:121–150, 1990.

[18] D. Angluin, M. Frazier, and L. Pitt. Learning conjunctions of Horn clauses. *Machine Learning*, 9:147–164, 1992.

[19] D. Angluin and M. Kharitonov. When won't membership queries help? *J. Comput. Syst. Sci.*, 50(2):336–355, 1995.

[20] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Mach. Learn.*, 63(2):161–182, 2006.

[21] J. Aspnes, R. Beigel, M. L. Furst, and S. Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):135–148, 1994.

[22] L. Babai. Trading group theory for randomness. In *Proc. of the 17th Symposium on Theory of Computing (STOC)*, pages 421–429, 1985.

[23] L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory. In *Proc. of the 27th Symposium on Foundations of Computer Science (FOCS)*, pages 337–347, 1986.

[24] L. Babai and S. Moran. Arthur-Merlin games: A randomized proof system, and a hierarchy of complexity classes. *J. Comput. Syst. Sci.*, 36(2):254–276, 1988.

[25] L. Babai, N. Nisan, and M. Szegedy. Multiparty protocols, pseudorandom generators for logspace, and time-space trade-offs. *J. Comput. Syst. Sci.*, 45(2):204–232, 1992.

[26] Z. Bar-Yossef, T. S. Jayram, and I. Kerenidis. Exponential separation of quantum and classical one-way communication complexity. *SIAM J. Comput.*, 38(1):366–384, 2008.

[27] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[28] R. Beals, H. Buhrman, R. Cleve, M. Mosca, and R. de Wolf. Quantum lower bounds by polynomials. *J. ACM*, 48(4):778–797, 2001.

[29] P. Beame and D.-T. Huynh-Ngoc. Multiparty communication complexity of AC^0. In *Electronic Colloquium on Computational Complexity (ECCC)*, July 2008. Report TR08-061.

[30] P. Beame, T. Pitassi, and N. Segerlind. Lower bounds for Lovász-Schrijver systems and beyond follow from multiparty communication complexity. *SIAM J. Comput.*, 37(3):845–869, 2007.

[31] R. Beigel. The polynomial method in circuit complexity. In *Proc. of the Eigth Annual Conference on Structure in Complexity Theory*, pages 82–95, 1993.

[32] R. Beigel. Perceptrons, PP, and the polynomial hierarchy. *Computational Complexity*, 4:339–349, 1994.

[33] R. Beigel, N. Reingold, and D. A. Spielman. PP is closed under intersection. *J. Comput. Syst. Sci.*, 50(2):191–202, 1995.

[34] A. Beimel, F. Bergadano, N. H. Bshouty, E. Kushilevitz, and S. Varricchio. Learning functions represented as multiplicity automata. *J. ACM*, 47(3):506–530, 2000.

[35] S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.*, 3:441–461, 2003.

[36] A. Blum, A. M. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[37] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proc. of the 26th Symposium on Theory of Computing (STOC)*, pages 253–262, 1994.

[38] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.

[39] A. Blum and S. Rudich. Fast learning of k-term DNF formulas with queries. *J. Comput. Syst. Sci.*, 51(3):367–373, 1995.

[40] A. L. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127, 1992.

[41] L. Blum, M. Blum, and M. Shub. A simple unpredictable pseudo-random number generator. *SIAM J. Comput.*, 15(2):364–383, 1986.

[42] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[43] B. Bollobás. *Extremal Graph Theory*. Academic Press, New York, 1978.

[44] J. Bruck. Harmonic analysis of polynomial threshold functions. *SIAM J. Discrete Math.*, 3(2):168–177, 1990.

[45] J. Bruck and R. Smolensky. Polynomial threshold functions, AC^0 functions, and spectral norms. *SIAM J. Comput.*, 21(1):33–42, 1992.

[46] N. H. Bshouty. Exact learning via the monotone theory. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 302–311, 1993.

[47] N. H. Bshouty. A subexponential exact learning algorithm for DNF using equivalence queries. *Inf. Process. Lett.*, 59(1):37–39, 1996.

[48] N. H. Bshouty. Simple learning algorithms using divide and conquer. *Computational Complexity*, 6(2):174–194, 1997.

[49] N. H. Bshouty, E. Mossel, R. O'Donnell, and R. A. Servedio. Learning DNF from random walks. *J. Comput. Syst. Sci.*, 71(3):250–265, 2005.

[50] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf. Quantum fingerprinting. *Phys. Rev. Lett.*, 87(16), 2001. Article no. 167902.

[51] H. Buhrman, R. Cleve, and A. Wigderson. Quantum vs. classical communication and computation. In *Proc. of the 13th Symposium on Theory of Computing (STOC)*, pages 63–68, 1998.

[52] H. Buhrman, R. Cleve, R. de Wolf, and C. Zalka. Bounds for small-error and zero-error quantum algorithms. In *Proc. of the 40th Symposium on Foundations of Computer Science (FOCS)*, pages 358–368, 1999.

[53] H. Buhrman, N. K. Vereshchagin, and R. de Wolf. On computation and communication with small bias. In *Proc. of the 22nd Conf. on Computational Complexity (CCC)*, pages 24–32, 2007.

[54] H. Buhrman and R. de Wolf. Communication complexity lower bounds by polynomials. In *Proc. of the 16th Conf. on Computational Complexity (CCC)*, pages 120–130, 2001.

[55] H. Buhrman and R. de Wolf. Complexity measures and decision tree complexity: A survey. *Theor. Comput. Sci.*, 288(1):21–43, 2002.

[56] A. Chakrabarti, Y. Shi, A. Wirth, and A. C.-C. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proc. of the 42nd Symposium on Foundations of Computer Science (FOCS)*, pages 270–278, 2001.

[57] A. K. Chandra, M. L. Furst, and R. J. Lipton. Multi-party protocols. In *Proc. of the 15th Symposium on Theory of Computing (STOC)*, pages 94–99, 1983.

[58] A. Chattopadhyay. Discrepancy and the power of bottom fan-in in depth-three circuits. In *Proc. of the 48th Symposium on Foundations of Computer Science (FOCS)*, pages 449–458, 2007.

[59] A. Chattopadhyay and A. Ada. Multiparty communication complexity of disjointness. In *Electronic Colloquium on Computational Complexity (ECCC)*, January 2008. Report TR08-002.

[60] E. W. Cheney. *Introduction to Approximation Theory*. Chelsea Publishing, New York, 2nd edition, 1982.

[61] B. Chor and O. Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.*, 17(2):230–261, 1988.

[62] F. R. K. Chung and P. Tetali. Communication complexity and quasi randomness. *SIAM J. Discrete Math.*, 6(1):110–123, 1993.

[63] S. Cook, C. Dwork, and R. Reischuk. Upper and lower time bounds for parallel random access machines without simultaneous writes. *SIAM J. Comput.*, 15(1):87–97, 1986.

[64] M. David and T. Pitassi. Separating NOF communication complexity classes RP and NP. In *Electronic Colloquium on Computational Complexity (ECCC)*, February 2008. Report TR08-014.

[65] M. David, T. Pitassi, and E. Viola. Improved separations between nondeterministic and randomized multiparty communication. In *Proc. of the 12th Intl. Workshop on Randomization and Computation (RANDOM)*, pages 371–384, 2008.

[66] R. de Wolf. A brief introduction to Fourier analysis on the Boolean cube. *Theory of Computing, Graduate Surveys*, 1:1–20, 2008.

[67] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*, volume 303. Springer-Verlag, Berlin, 1993.

[68] V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami. New results for learning noisy parities and halfspaces. In *Proceedings of the 47th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 563–574, 2006.

[69] J. Ford and A. Gál. Hadamard tensors and lower bounds on multiparty communication complexity. In *Proc. of the 32nd International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1163–1175, 2005.

[70] J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.*, 65(4):612–625, 2002.

[71] J. Forster, M. Krause, S. V. Lokam, R. Mubarakzjanov, N. Schmitt, and H.-U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *Proc. of the 21st Conf. on Foundations of Software Technology and Theoretical Computer Science (FST TCS)*, pages 171–182, 2001.

[72] J. Forster, N. Schmitt, H. U. Simon, and T. Suttorp. Estimating the optimal margins of embeddings in Euclidean half spaces. *Mach. Learn.*, 51(3):263–281, 2003.

[73] J. Forster and H. U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.*, 350(1):40–48, 2006.

[74] Y. Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995.

[75] M. L. Furst, J. B. Saxe, and M. Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17(1):13–27, 1984.

[76] D. Gavinsky. Classical interaction cannot replace a quantum message. In *Proc. of the 40th Symposium on Theory of Computing (STOC)*, pages 95–102, 2008.

[77] D. Gavinsky, J. Kempe, I. Kerenidis, R. Raz, and R. de Wolf. Exponential separations for one-way quantum communication complexity, with applications to cryptography. In *Proc. of the 39th Symposium on Theory of Computing (STOC)*, pages 516–525, 2007.

[78] D. Gavinsky, J. Kempe, O. Regev, and R. de Wolf. Bounded-error quantum state identification and exponential separations in communication complexity. In *Proc. of the 38th Symposium on Theory of Computing (STOC)*, pages 594–603, 2006.

[79] D. Gavinsky, J. Kempe, and R. de Wolf. Strengths and weaknesses of quantum fingerprinting. In *Proc. of the 21st Conf. on Computational Complexity (CCC)*, pages 288–298, 2006.

[80] D. Gavinsky and P. Pudlák. Exponential separation of quantum and classical non-interactive multi-party communication complexity. In *Proc. of the 23rd Conf. on Computational Complexity (CCC)*, pages 332–339, 2008.

[81] D. Gavinsky and A. A. Sherstov. NP and MA do not contain coNP in multiparty communication complexity. Manuscript, 2009.

[82] M. Goldmann, J. Håstad, and A. A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

[83] M. Goldmann and M. Karpinski. Simulating threshold circuits by majority circuits. *SIAM J. Comput.*, 27(1):230–246, 1998.

[84] O. Goldreich, S. Goldwasser, and S. Micali. How to construct random functions. *J. ACM*, 33(4):792–807, 1986.

[85] G. H. Golub and C. F. V. Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

[86] P. Gordan. Über die Auflösung linearer Gleichungen mit reellen Coefficienten. *Mathematische Annalen*, 6:23–28, 1873.

[87] V. Grolmusz. The BNS lower bound for multi-party protocols in nearly optimal. *Inf. Comput.*, 112(1):51–54, 1994.

[88] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, and G. Turán. Threshold circuits of bounded depth. *J. Comput. Syst. Sci.*, 46(2):129–154, 1993.

[89] T. R. Hancock. Learning 2μ DNF formulas and $k\,\mu$ decision trees. In *Proc. of the 6th Conf. on Computational Learning Theory (COLT)*, pages 199–209, 1991.

[90] T. R. Hancock. Learning $k\,\mu$ decision trees on the uniform distribution. In *Proc. of the 6th Conf. on Computational Learning Theory (COLT)*, pages 352–360, 1993.

[91] J. Håstad. *Computational limitations of small-depth circuits*. PhD thesis, Massachusetts Institute of Technology, 1987.

[92] J. Håstad. On the size of weights for threshold gates. *SIAM J. Discret. Math.*, 7(3):484–492, 1994.

[93] J. Håstad and M. Goldmann. On the power of small-depth threshold circuits. *Computational Complexity*, 1:113–129, 1991.

[94] D. P. Helmbold, R. H. Sloan, and M. K. Warmuth. Learning integer lattices. *SIAM J. Comput.*, 21(2):240–266, 1992.

[95] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, New York, 1986.

[96] P. Høyer and R. de Wolf. Improved quantum communication complexity bounds for disjointness and equality. In *Proc. of the 19th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 299–310, 2002.

[97] A. D. Ioffe and V. M. Tikhomirov. Duality of convex functions and extremum problems. *Russ. Math. Surv.*, 23(6):53–124, 1968.

[98] J. C. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *J. Comput. Syst. Sci.*, 55(3):414–440, 1997.

[99] S. Jukna. *Extremal Combinatorics with Applications in Computer Science*. Springer-Verlag, Berlin, 2001.

[100] J. Kahn, N. Linial, and A. Samorodnitsky. Inclusion-exclusion: Exact and approximate. *Combinatorica*, 16(4):465–477, 1996.

[101] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proc. of the 46th Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.

[102] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

[103] B. Kashin and A. A. Razborov. Improved lower bounds on the rigidity of Hadamard matrices. *Matematicheskie zametki*, 63(4):535–540, 1998. In Russian.

[104] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proc. of the 25th Symposium on Theory of Computing (STOC)*, pages 392–401, 1993.

[105] M. Kearns and L. Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. *J. ACM*, 41(1):67–95, 1994.

[106] M. J. Kearns, M. Li, L. Pitt, and L. G. Valiant. On the learnability of Boolean formulae. In *Proc. of the 19th Symposium on Theory of Computing (STOC)*, pages 285–295, 1987.

[107] M. J. Kearns, R. E. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

[108] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, 1994.

[109] C. Kenyon and S. Kutin. Sensitivity, block sensitivity, and ℓ-block sensitivity of Boolean functions. *Information and Computation*, 189(1):43–53, 2004.

[110] I. Kerenidis and R. Raz. The one-way communication complexity of the Boolean hidden matching problem. Available at arxiv.org/abs/quant-ph/0607173, 2006.

[111] M. Kharitonov. Cryptographic lower bounds for learnability of Boolean functions on the uniform distribution. In *Proceedings of the 5th Workshop on Computational Learning Theory (COLT)*, pages 29–36, 1992.

[112] M. Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proc. of the 25th Symposium on Theory of Computing*, pages 372–381, 1993.

[113] S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *Proc. of the 40th Symposium on Theory of Computing (STOC)*, pages 345–354, 2008.

[114] H. Klauck. Lower bounds for quantum communication complexity. In *Proc. of the 42nd Symposium on Foundations of Computer Science (FOCS)*, pages 288–297, 2001.

[115] H. Klauck. Rectangle size bounds and threshold covers in communication complexity. In *Proc. of the 18th Conf. on Computational Complexity (CCC)*, pages 118–134, 2003.

[116] H. Klauck. Lower bounds for quantum communication complexity. *SIAM J. Comput.*, 37(1):20–46, 2007.

[117] H. Klauck, A. Nayak, A. Ta-Shma, and D. Zuckerman. Interaction in quantum communication and the complexity of set disjointness. In *Proc. of the 33rd Symposium on Theory of Computing (STOC)*, pages 124–133, 2001.

[118] H. Klauck, R. Špalek, and R. de Wolf. Quantum and classical strong direct product theorems and optimal time-space tradeoffs. *SIAM J. Comput.*, 36(5):1472–1493, 2007.

[119] A. Klivans and A. Shpilka. Learning arithmetic circuits via partial derivatives. In *Proc. of the 16th Conf. on Computational Learning Theory (COLT)*, pages 463–476, 2003.

[120] A. R. Klivans. *A Complexity-Theoretic Approach to Learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

[121] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

[122] A. R. Klivans and R. A. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *J. Comput. Syst. Sci.*, 68(2):303–318, 2004.

[123] A. R. Klivans and R. A. Servedio. Learning intersections of halfspaces with a margin. In *Proc. of the 17th Conf. on Computational Learning Theory (COLT)*, pages 348–362, 2004.

[124] A. R. Klivans and R. A. Servedio. Toward attribute efficient learning of decision lists and parities. *J. Machine Learning Research*, 7:587–602, 2006.

[125] A. R. Klivans and R. A. Servedio. Learning intersections of halfspaces with a margin. *J. Comput. Syst. Sci.*, 74(1):35–48, 2008.

[126] A. R. Klivans and A. A. Sherstov. A lower bound for agnostically learning disjunctions. In *Proc. of the 20th Conf. on Learning Theory (COLT)*, pages 409–423, 2007.

[127] A. R. Klivans and A. A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2–3):97–114, 2007.

[128] A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. *J. Comput. Syst. Sci.*, 75(1):2–12, 2009.

[129] A. R. Klivans and D. A. Spielman. Randomness efficient identity testing of multivariate polynomials. In *Proceedings of the 33rd Symposium on Theory of Computing (STOC)*, pages 216–223, 2001.

[130] M. Krause. Geometric arguments yield better bounds for threshold circuits and distributed computing. *Theor. Comput. Sci.*, 156(1–2):99–117, 1996.

[131] M. Krause. On the computational power of Boolean decision lists. *Computational Complexity*, 14(4):362–375, 2006.

[132] M. Krause and P. Pudlák. On the computational power of depth-2 circuits with threshold and modulo gates. *Theor. Comput. Sci.*, 174(1–2):137–156, 1997.

[133] M. Krause and P. Pudlák. Computing Boolean functions by polynomials and threshold circuits. *Comput. Complex.*, 7(4):346–370, 1998.

[134] I. Kremer. Quantum communication. Master's thesis, Hebrew University, Computer Science Department, 1995.

[135] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

[136] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993.

[137] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, New York, 1997.

[138] E. Kushilevitz and D. Roth. On learning visual concepts and DNF formulae. *Machine Learning*, 24(1):65–85, 1996.

[139] S. Kwek and L. Pitt. PAC learning intersections of halfspaces with membership queries. *Algorithmica*, 22(1/2):53–75, 1998.

[140] T. Lee and A. Shraibman. Disjointness is hard in the multi-party number-on-the-forehead model. In *Proc. of the 23rd Conf. on Computational Complexity (CCC)*, pages 81–91, 2008.

[141] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[142] N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.

[143] N. Linial and N. Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10(4):349–365, 1990.

[144] N. Linial and A. Shraibman. Lower bounds in communication complexity based on factorization norms. In *Proc. of the 39th Symposium on Theory of Computing (STOC)*, pages 699–708, 2007.

[145] N. Linial and A. Shraibman. Learning complexity vs communication complexity. *Combinatorics, Probability & Computing*, 18(1-2):227–245, 2009.

[146] S. V. Lokam. Spectral methods for matrix rigidity with applications to size-depth trade-offs and communication complexity. *J. Comput. Syst. Sci.*, 63(3):449–473, 2001.

[147] L. Lovász and M. E. Saks. Lattices, Möbius functions and communication complexity. In *Proc. of the 29th Symposium on Foundations of Computer Science (FOCS)*, pages 81–90, 1988.

[148] L. Lovász and M. E. Saks. Communication complexity and combinatorial lattice theory. *J. Comput. Syst. Sci.*, 47(2):322–349, 1993.

[149] F. J. Macwilliams and N. J. A. Sloane. *The Theory of Error Correcting Codes*. North Holland Publishing, Amsterdam, 1977.

[150] Y. Mansour. An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *J. Comput. Syst. Sci.*, 50(3):543–550, 1995.

[151] Y. Mansour. Randomized interpolation and approximation of sparse polynomials. *SIAM J. Comput.*, 24(2):357–368, 1995.

[152] K. Mehlhorn and E. M. Schmidt. Las Vegas is better than determinism in VLSI and distributed computing. In *Proc. of the 14th Symposium on Theory of Computing (STOC)*, pages 330–337, 1982.

[153] M. L. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, Mass., 1969.

[154] E. Mossel, R. O'Donnell, and R. A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci.*, 69(3):421–434, 2004.

[155] J. Myhill and W. H. Kautz. On the size of weights required for linear-input switching functions. *IRE Trans. on Electronic Computers*, 10(2):288–290, 1961.

[156] D. J. Newman. Rational approximation to $|x|$. *Michigan Math. J.*, 11(1):11–14, 1964.

[157] I. Newman. Private vs. common random bits in communication complexity. *Inf. Process. Lett.*, 39(2):67–71, 1991.

[158] N. Nisan. CREW PRAMs and decision trees. *SIAM J. Comput.*, 20(6):999–1007, 1991.

[159] N. Nisan. The communication complexity of threshold gates. In *Combinatorics, Paul Erdős is Eighty*, pages 301–315, 1993.

[160] N. Nisan and M. Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994.

[161] A. B. J. Novikoff. On convergence proofs on perceptrons. In *Proc. of the Symposium on the Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[162] R. O'Donnell. Some topics in analysis of Boolean functions. In *Electronic Colloquium on Computational Complexity (ECCC)*, May 2008. Report TR08-055.

[163] R. O'Donnell and R. A. Servedio. New degree bounds for polynomial threshold functions. In *Proc. of the 35th Symposium on Theory of Computing (STOC)*, pages 325–334, 2003.

[164] R. O'Donnell and R. A. Servedio. Extremal properties of polynomial threshold functions. *J. Comput. Syst. Sci.*, 74(3):298–312, 2008.

[165] R. Paturi. On the degree of polynomials that approximate symmetric Boolean functions. In *Proc. of the 24th Symposium on Theory of Computing (STOC)*, pages 468–474, 1992.

[166] R. Paturi and M. E. Saks. Approximating threshold circuits by rational functions. *Inf. Comput.*, 112(2):257–272, 1994.

[167] R. Paturi and J. Simon. Probabilistic communication complexity. *J. Comput. Syst. Sci.*, 33(1):106–123, 1986.

[168] P. P. Petrushev and V. A. Popov. *Rational Approximation of Real Functions*. Cambridge University Press, Cambridge, 1987.

[169] V. V. Podolskii. Perceptrons of large weight. In *Proc. of the Second International Computer Science Symposium in Russia (CSR)*, pages 328–336, 2007.

[170] V. V. Podolskii. A uniform lower bound on weights of perceptrons. In *Proc. of the Third International Computer Science Symposium in Russia (CSR)*, pages 261–272, 2008.

[171] R. Raz. Fourier analysis for probabilistic communication complexity. *Comput. Complex.*, 5(3/4):205–221, 1995.

[172] R. Raz. Exponential separation of quantum and classical communication complexity. In *Proc. of the 31st Symposium on Theory of Computing (STOC)*, pages 358–367, 1999.

[173] R. Raz. The BNS-Chung criterion for multi-party communication complexity. *Comput. Complex.*, 9(2):113–122, 2000.

[174] A. A. Razborov. Bounded-depth formulae over the basis $\{\&, \oplus\}$ and some combinatorial problems. *Complexity Theory and Applied Mathematical Logic*, vol. "Problems of Cybernetics":146–166, 1988. In Russian.

[175] A. A. Razborov. On rigid matrices. Manuscript in Russian, available at `http://www.mi.ras.ru/~razborov/rigid.pdf`, June 1989.

[176] A. A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[177] A. A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67(1):145–159, 2003.

[178] A. A. Razborov and A. A. Sherstov. The sign-rank of AC^0. In *Proc. of the 49th Symposium on Foundations of Computer Science (FOCS)*, pages 57–66, 2008.

[179] A. A. Razborov and A. Wigderson. $n^{\Omega(\log n)}$ lower bounds on the size of depth-3 threshold circuits with AND gates at the bottom. *Inf. Process. Lett.*, 45(6):303–307, 1993.

[180] O. Regev. New lattice based cryptographic constructions. In *Proc. of the 35th Symposium on Theory of Computing (STOC)*, pages 407–416, 2003.

[181] O. Regev. On lattices, learning with errors, random linear codes, and cryptography. In *Proceedings of the 37th Symposium on Theory of Computing (STOC)*, pages 84–93, 2005.

[182] O. Regev. Personal communication, 2006.

[183] O. Regev. Lattice-based cryptography. In *Proceedings of the 26th Annual International Cryptology Conference (CRYPTO)*, volume 4117, pages 131–141, 2006.

[184] O. Regev. On the complexity of lattice problems with polynomial approximation factors. Survey prepared for the LLL+25 conference. Available at `http://www.cs.tau.ac.il/~odedr`, June 2007.

[185] T. J. Rivlin. *An Introduction to the Approximation of Functions*. Dover Publications, New York, 1981.

[186] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[187] D. Rubinstein. Sensitivity vs. block sensitivity of Boolean functions. *Combinatorica*, 15(2):297–299, 1995.

[188] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.

[189] Y. Sakai and A. Maruoka. Learning monotone log-term DNF formulas. In *Proc. of the 7th Conf. on Computational Learning Theory (COLT)*, pages 165–172, 1994.

[190] M. E. Saks. Slicing the hypercube. *Surveys in Combinatorics*, pages 211–255, 1993.

[191] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[192] R. E. Schapire. *The Design and Analysis of Efficient Learning Algorithms*. MIT Press, Cambridge, Mass., 1992.

[193] R. E. Schapire and L. Sellie. Learning sparse multivariate polynomials over a field with queries and counterexamples. *J. Comput. Syst. Sci.*, 52(2):201–213, 1996.

[194] A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, Inc., New York, 1998.

[195] A. A. Sherstov. A discrepancy-based proof of Razborov's quantum lower bounds. Technical Report TR-07-33, The University of Texas at Austin, Department of Computer Sciences, July 2007.

[196] A. A. Sherstov. Powering requires threshold depth 3. *Inf. Process. Lett.*, 102(2–3):104–107, 2007.

[197] A. A. Sherstov. Communication complexity under product and nonproduct distributions. In *Proc. of the 23rd Conf. on Computational Complexity (CCC)*, pages 64–70, 2008.

[198] A. A. Sherstov. Halfspace matrices. *Comput. Complex.*, 17(2):149–178, 2008. Preliminary version in 22nd CCC, 2007.

[199] A. A. Sherstov. Approximate inclusion-exclusion for arbitrary symmetric functions. *Computational Complexity*, 18(2):219–247, 2009. Preliminary version in 23nd CCC, 2008.

[200] A. A. Sherstov. The intersection of two halfspaces has high threshold degree. In *Proc. of the 50th Symposium on Foundations of Computer Science (FOCS)*, 2009. To appear.

[201] A. A. Sherstov. On quantum-classical equivalence for composed communication problems. Manuscript at quant-ph/0906.1399, 2009.

[202] A. A. Sherstov. Separating AC^0 from depth-2 majority circuits. *SIAM J. Comput.*, 38(6):2113–2129, 2009. Preliminary version in 39th STOC, 2007.

[203] A. A. Sherstov. The pattern matrix method for lower bounds on quantum communication. In *Proc. of the 40th Symposium on Theory of Computing (STOC)*, pages 85–94, 2008.

[204] A. A. Sherstov. The unbounded-error communication complexity of symmetric functions. In *Proc. of the 49th Symposium on Foundations of Computer Science (FOCS)*, pages 384–393, 2008.

[205] Y. Shi and Y. Zhu. Quantum communication complexity of block-composed functions. *Quantum Information & Computation*, 9(5–6):444–460, 2009.

[206] H. U. Simon. Spectral norm in learning theory: Some selected topics. In *Proc. of the 17th Conf. on Algorithmic Learning Theory (ALT)*, pages 13–27, 2006.

[207] K.-Y. Siu and J. Bruck. On the power of threshold circuits with small weights. *SIAM J. Discrete Math.*, 4(3):423–435, 1991.

[208] K.-Y. Siu, J. Bruck, T. Kailath, and T. Hofmeister. Depth efficient neural networks for division and related problems. *IEEE Transactions on Information Theory*, 39(3):946–956, 1993.

[209] K.-Y. Siu and V. P. Roychowdhury. On optimal depth threshold circuits for multiplication and related problems. *SIAM J. Discrete Math.*, 7(2):284–292, 1994.

[210] K.-Y. Siu, V. P. Roychowdhury, and T. Kailath. Rational approximation techniques for analysis of neural networks. *IEEE Transactions on Information Theory*, 40(2):455–466, 1994.

[211] J. Tarui and T. Tsukiji. Learning DNF by approximating inclusion-exclusion formulae. In *Proc. of the 14th Conf. on Computational Complexity (CCC)*, pages 215–221, 1999.

[212] C. D. Thompson. Area-time complexity for VLSI. In *Proc. of the 11th Symposium on Theory of Computing (STOC)*, pages 81–88, 1979.

[213] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[214] L. G. Valiant. Learning disjunction of conjunctions. In *Proc. of the 9th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 560–566, 1985.

[215] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[216] S. Vempala. A random sampling based algorithm for learning the intersection of halfspaces. In *Proc. of the 38th Symposium on Foundations of Computer Science (FOCS)*, pages 508–513, 1997.

[217] K. A. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *Proc. of the 3rd Workshop on Computational Learning Theory (COLT)*, pages 314–326, 1990.

[218] N. K. Vereshchagin. Lower bounds for perceptrons solving some separation problems and oracle separation of AM from PP. In *Proc. of the Third Israel Symposium on Theory of Computing and Systems (ISTCS)*, pages 46–51, 1995.

[219] I. Wegener. Optimal lower bounds on the depth of polynomial-size threshold circuits for some arithmetic functions. *Inf. Process. Lett.*, 46(2):85–87, 1993.

[220] R. de Wolf. Personal communication, October 2007.

[221] R. de Wolf. A note on quantum algorithms and the minimal degree of ε-error polynomials for symmetric functions. *Quantum Information and Computation*, 8(10):943–950, 2008.

[222] K. Yang. New lower bounds for statistical query learning. *J. Comput. Syst. Sci.*, 70(4):485–509, 2005.

[223] A. C.-C. Yao. Some complexity questions related to distributive computing. In *Proc. of the 11th Symposium on Theory of Computing (STOC)*, pages 209–213, 1979.

[224] A. C.-C. Yao. Lower bounds by probabilistic arguments. In *Proc. of the 24th Symposium on Foundations of Computer Science (FOCS)*, pages 420–428, 1983.

[225] A. C.-C. Yao. On ACC and threshold circuits. In *Proc. of the 31st Symposium on Foundations of Computer Science (FOCS)*, pages 619–627, 1990.

[226] A. C.-C. Yao. Quantum circuit complexity. In *Proc. of the 34th Symposium on Foundations of Computer Science (FOCS)*, pages 352–361, 1993.

[227] K. Zarankiewicz. Problem P 101. *Colloq. Math.*, 2:116–131, 1951.

[228] E. I. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Izvestiya Imp. Akad. Nauk*, 30(5), 1877.

Vita

Alexander Alexandrovich Sherstov was born in Karaganda, the Soviet Union. He finished Karaganda's High School No. 1 in June 1999, with a specialization in mathematics and physics, and entered the American University in Bulgaria. Starting in his sophomore year, Alexander continued his undergraduate education at Hope College, Michigan, and graduated in May 2003 with a Bachelor of Science degree in computer science. Three months later, he joined the doctoral program in the Department of Computer Sciences at the University of Texas at Austin.

Permanent address: 1 University Station C0500, Austin, TX 78712-0233.

This manuscript was typed by the author.