Halfway to Halfspace Testing

by

Nathaniel Harms

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2017

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis I study the problem of *testing halfspaces* under arbitrary probability distributions, using only random samples. A *halfspace*, or *linear threshold function*, is a boolean function $f : \mathbb{R}^n \to \{\pm 1\}$ defined as the sign of a linear function; that is,

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i x_i - \theta\right)$$

where we refer to $w \in \mathbb{R}^n$ as a *weight vector* and $\theta \in \mathbb{R}$ as a *threshold*. These functions have been studied intensively since the middle of the 20^{th} century; they appear in many places, including social choice theory (the theory of voting rules), circuit complexity theory, machine learning theory, hardness of approximation, and the analysis of boolean functions.

The problem of testing halfspaces, in the sense of *property testing*, is to design an algorithm that, with high probability, decides whether an unknown function f is a halfspace function or *far* from a halfspace, using as few examples of labelled points $(x, f(x))$ as possible. In this work I focus on the problem of testing halfspaces using only random examples drawn from an arbitrary distribution, and the algorithm cannot choose the points it receives. This is in contrast with previous work on the problem, where the algorithm can query points of its choice, and the distribution was assumed to be uniform over the boolean hypercube.

Towards a solution to this problem I present an algorithm that works for rotationally invariant probability distributions (under reasonable conditions), using roughly $O(\sqrt{n})$ random examples, which is close to the known lower bound of $\Omega\left(\sqrt{n/\log n}\right)$. I further develop the algorithm to work for mixtures of two such rotationally invariant distributions and provide a partial analysis. I also survey related machine learning results, and conclude with a survey of the theory of halfspaces over the boolean hypercube, which has recently received much attention.

Acknowledgements

Thanks to my supervisor Eric Blais for suggesting this research topic, helping me along, and waiting patiently as I progressed very slowly. Thanks also to my readers Lap Chi Lau and Yaoliang Yu.

Dedication

For my parents and brother!

Table of Contents

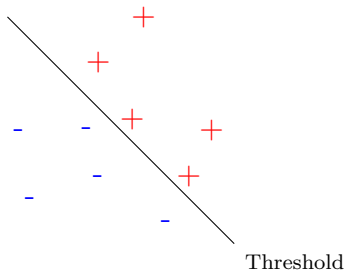
Chapter 1

Introduction

Halfspaces are objects of fundamental importance in many areas of computer science. A halfspace, or linear threshold function, is what you get when you take a linear function on points $x \in \mathbb{R}^n$, say

$$w_0 + w_1 x_1 + \cdots + w_n x_n$$

for some arbitrary weights $w_0, \ldots, w_n \in \mathbb{R}$, and label a set of points by whether the function takes a positive or negative value; in other words, a halfspace is a function that "draws a line" between two sets of points, like so:



As you can see, halfspaces are arguably the simplest way of separating two sets of points: all we must do is "draw a line" between the two groups. These functions have seen a variety of uses in the past several decades: in fact they are such basic, simple, elementary objects that this is nearly a vacuous claim, like the claim that drawing lines is important for making pictures. Solving a system of linear equations, for example, is merely the act of finding a point contained within the intersection of halfspaces. Less obviously, halfspaces

appear in such diverse fields as social choice theory, where one is interested in the properties of *voting schemes* - ways to produce consensus from a population of voters. Recently the study of boolean functions has had implications for both social choice theory and computer science, for example the "Majority is Stablest" theorem [MOO05] which was used by Khot et al. [KKMO07] to show the optimality of the famous Goemans-Williamson Max-Cut algorithm [GW95] (assuming the Unique Games Conjecture).

In computer science, there are numerous uses of linear threshold functions. A lot of work was done by electrical engineers towards understanding circuits with linear threshold gates: one can imagine a circuit that produces 1 if the total input across several input wires exceeds some threshold, and 0 otherwise (see for example [Cho61, Win71]). These circuits are still commonly studied in complexity theory [Nis93, KW15, Wil14], and many important problems remain open; in fact, understanding threshold circuits is considered a step towards answering the famous P =?NP question [Aar16]. Linear threshold circuits are also of interest in machine learning, since they are neural networks: linear threshold functions are simple models of neurons and were introduced into the study of learning, both human and machine, in the past century, beginning with Rosenblatt [Ros58]. Halfspaces have maintained their fundamental position in the field of machine learning, and the problem of efficiently learning halfspaces in the many different models of machine is still an active area of research (e.g. [ABL17, BL13, Dan15, FGKP06, GR09, KKMS08, KLS09, Lon03]).

The idea of machine learning is that we have an unknown classification of a population, and we want to "learn" the rule for separating examples into their respective classes. We see a series of classified ("labelled") examples, drawn randomly from the population, and we attempt to produce a rule that will (probably) work on future examples. However, it is necessary to assume that our classification rule belongs to some "concept class", a set of possible classification rules; for our purposes we say that the concept class is the set of all halfspaces. To learn the halfspace that classifies the population, we can just look at a large number of examples and draw a line that separates them. But what if we don't know that our target classification is in fact a halfspace? This is a situation that occurs in practice: halfspaces are nice, simple classification rules that are easy to use, but unfortunately real data may not be linearly separable.

About the problem of finding a suitable halfspace for a set of examples, Blum et al. observe that

> ...one could simply apply an LP solver to solve [it]. In practice, however, this approach is rarely used in machine learning applications. One of the main reasons is that the data often is not consistent with *any* vector w and one's goal is to simply do as well as one can [BFKV96].

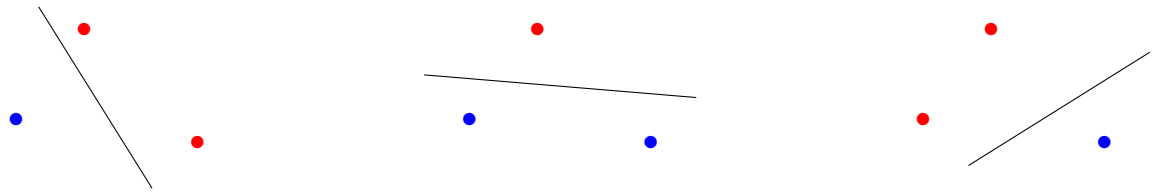
Doing as well as one can with data that may or may not be linearly separable is not a particularly noble goal, and we might hope for a way to determine whether finding a separator is even worth attempting. In other words, we want to answer the question,

> ***Question:*** *Given a small set of random examples from a population, can we test whether there is a halfspace that (nearly) correctly classifies the population as a whole?*

Or, more formally,

> ***Question:*** *For a function $f : \mathbb{R}^n \to \{\pm 1\}$, given a set of $O(\sqrt{n})$ random, labelled examples $(x, f(x))$ from an arbitrary probability distribution, can we* test *if f is a halfspace?*

I will explain the $O(\sqrt{n})$ goal, and what I mean by *test*, in Chapter 3; first, let's see why the solution is not obvious. One might be tempted to follow the naïve strategy: take your sample and try to draw a line that separates the points – if you can, say "yes", otherwise say "no". But this strategy will not work! When we have fewer than $n+2$ example points in an n-dimensional space, we can *always* separate the points with a line; for example, in 2 dimensions we can get every partition of 3 points like this:



So we have to try something more complicated, which justifies the rest of this thesis.

This question lies in the framework of *property testing*, essentially a model of making approximate decisions with incomplete information. Property testing and the analysis of boolean functions is currently a lively field of research, and halfspaces play a key role in a number of such works (e.g. [BBBY12, DJS+14, DS13, GS07, MORS09, MORS10, RS15]). Surprisingly, only a few works have tried to answer a similar question ([BBBY12, MORS09, MORS10, RS15]), and are concerned mostly with the boolean hypercube, while we are interested in solving the more general problem. So, beyond the motivation from machine learning, we hope that the study of testing halfspaces will expand our understanding of halfspaces, boolean functions, and property testing.

In this thesis I present algorithms for testing halfspaces with $O(\sqrt{n})$ random examples in some restricted settings, namely rotationally invariant probability spaces (Chapter 4),

and mixtures of rotationally invariant spaces (Chapter 4), that satisfy a condition on their "width" (defined in Chapter 4):

Theorem 4.7.1 (Informal): For any rotationally invariant distribution μ over $\mathbb{R}^n$ (satisfying some condition), there exists an algorithm that tests halfspaces using $O(\sqrt{n})$ random examples.

Theorem 5.3.1 (Informal): For any mixture of two rotationally invariant distributions over $\mathbb{R}^n$ (satisfying some condition), there exists an algorithm that tests halfspaces using $O(\sqrt{n})$ random examples.

The main tool used in these theorems is the Gap Theorem (Theorem 4.4.1), which is a simple structural theorem about the "centers of mass" of functions that has several applications and simplifies a number of proofs in the existing literature:

Theorem 4.4.1 (Informal): Let $f : \mathbb{R}^n \to \{\pm 1\}$ and let h be a halfspace with the same expected value $\mathbb{E}[h(x)] = \mathbb{E}[f(x)]$ under some probability distribution μ. Then the distance between the "centers of mass" of f and h is at least $\mathbb{P}[h(x) \neq f(x)]$ multiplied by the "width" of μ (defined in Chapter 4).

As another application of the Gap Theorem, I will generalize an existing algorithm of De *et al.* [DDFS14] for the Chow Parameters Problem in Chapter 6.

Before presenting this work, I survey in Chapter 3 some related work from machine learning theory to motivate the testing problem (Section 3.2). In Section 3.3 I show a few of the few results on testing halfspaces. After presenting my own results, I survey several recent works on the theory of halfspaces over the boolean hypercube in Chapter 7; a testing algorithm for arbitrary probability distributions must apply to the hypercube as a special case, so the rich theory of the hypercube will be important for the future work that I will outline in Chapter 8. This thesis contains some partial results: the algorithms in Chapters 4 and 5 work for nice enough distributions, but there are some examples I give where more analysis is needed. Thus the reader should consider this thesis a work in progress.

Chapter 2

Preliminaries and Centers of Mass

There are a number of standard mathematical tools used in the study of halfspaces, which I will review. I will introduce specialized tools as necessary, but the following definitions and facts are widely used in the literature and in this work.

2.1 Notation

For positive integers n I will write

$$[n] := \{1, 2, \ldots, n\}.$$

$x \sim S$ will mean that x is drawn uniformly at random from the (finite) set S, unless otherwise noted.

$x = a \pm b$ means $a - b \leq x \leq a + b$.

$x \sim \mu$ will mean that x is drawn from the distribution μ. For a distribution (i.e. probability measure) μ, I will abuse notation slightly by writing $\mu(A)$ for the measure of set A and $\mu(x)$ for the density at point x.

$\mathbb{1}[X]$ denotes the indicator function

$$\mathbb{1}[X] := \begin{cases} 1 & \text{if } X \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

To denote the size of a set S, I will use either the notation $|S|$ or $\#S$, whichever is most readable.

I will use $\langle \cdot, \cdot \rangle$ to denote the standard inner product, unless otherwise noted: for vectors $x, y \in \mathbb{R}^n$,

$$\langle x, y \rangle := \sum_{i \in [n]} x_i y_i \,.$$

The norms $\|x\|_p$ will refer to the standard L_p norms

$$\|x\|_p = \left(\sum_{i \in [n]} |x_i|^p \right)^{\frac{1}{p}}$$

and $\|x\|$ without a subscript will refer to the 2-norm.

By "boolean value" I will mean a value in $\{\pm 1\}$ rather than the common $\{0, 1\}$.

For vectors $x \in \mathbb{R}^n$, coordinates $i \in [n]$, and values $b \in \mathbb{R}$, I will write

$$x^{i \leftarrow b} := (x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n)$$

as the vector where coordinate i is set to b.

log will refer to the logarithm with base 2.

For derivatives I will use operator-style notation when it avoids confusion: for a function f of n variables I will write

$$D_i f(x_1, \ldots, x_n) := \frac{\partial}{\partial x_i} f(x_1, \ldots, x_n) \,.$$

In addition to standard $O\left(\left(\right) \cdot\right)$ notation, I will use the tilde notation to hide polylogarithmic factors:

$$\tilde{O}(f(x)) := O\left(f(x) \log^C(f(x))\right)$$

for some constant C, and

$$\tilde{\Omega}(f(x)) := \Omega\left(f(x) \log^{-C}(f(x))\right) \,.$$

2.2 Halfspaces

The object of study in this work is the *halfspace*, or *linear threshold function*, so named because these functions take positive values when some linear function is above some threshold, and negative values otherwise.

Definition 2.2.1 (Linear Threshold Function)**.** A linear threshold function (LTF), or halfspace, is any function of the form $f : \mathbb{R}^n \to \{\pm 1\}$ where for some $w \in \mathbb{R}^n$, called the *normal* of the halfspace, and for some *threshold* $\alpha \in \mathbb{R}$, for all $x \in \mathbb{R}^n$,

$$f(x) = \text{sign}(\langle w, x \rangle - \alpha)$$

where we say $\text{sign}(0) = 1$ for convenience. Here the inner product $\langle u, v \rangle$ is the standard inner product $\sum_{i \in [n]} u_i v_i$ for vectors $u, v \in \mathbb{R}^n$.

An important special case is the balanced halfspace:

Definition 2.2.2 (Balanced Function)**.** A *balanced* function $f : \mathbb{R}^n \to \{\pm 1\}$, and in particular a balanced halfspace, is a function for which $\mathbb{E}[f] = 0$.

2.3 Probability Theory and Concentration

The problems we are considering are of a fundamentally probabilistic nature, so we will require some basic probability theory.

2.3.1 Basic Definitions

I will omit some elementary definitions (σ-algebras, measures, measurable functions) that can be found, for example, in [Fel68].

Definition 2.3.1 (Probability Space)**.** A probability space is a triple (Ω, Σ, μ) where Ω is a set (called the set of 'outcomes'), Σ is a σ-algebra on Ω, and μ is a measure on Σ such that $\mu(\Omega) = 1$.

Let $(A, \mathcal{A})$ be another measurable space. Then an A-valued *random variable* X is a $(\Sigma, \mathcal{A})$-measurable function $X : \Omega \to A$.

Definition 2.3.2 (Expectation, Covariance, Variance)**.** Let $X : \Omega \to A$ be any random variable such that addition makes sense on A (e.g. $A = \mathbb{R}^n$ or another vector space). Then we define the *expected value* of X as

$$\mathbb{E}[X] := \int_\Omega X(\omega)\mu(d\omega) .$$

In the case of a discrete probability space, this is simply

$$\mathbb{E}[X] = \sum_\Omega X(\omega)\mathbb{P}[\omega] .$$

For real-valued random variables, we can further define the covariance and variance; let $Y : \Omega \to \mathbb{R}$ be another random variable. Covariance is a measure of how well correlated two random variables are:

$$\mathrm{Cov}(X, Y) := \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y] .$$

Variance defines, in an intuitive sense, how far a variable is likely to differ from its mean:

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 .$$

Note that $\mathbb{V}[X] = \mathrm{Cov}(X, X)$.

This thesis deals mostly with *centered* probability distributions:

Definition 2.3.3 (Centered Distribution)**.** A distribution μ is *centered* if $X \sim \mu$ satisfies $\mathbb{E}[X] = 0$.

2.3.2 Concentration Inequalities

These standard inequalities can be found, for example, in [BLM13].

Theorem 2.3.4 (Jensen's Inequality)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be any convex function and X a real-valued random variable. Then*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)] .$$

Theorem 2.3.5 (Markov's Inequality)**.** *Let $X : \Omega \to [0, \infty)$ be any non-negative random variable. Then*

$$\mathbb{P}[X > t] \leq \frac{\mathbb{E}[X]}{t} .$$

Theorem 2.3.6 (Chebyshev's Inequality)**.** *Let X be any random variable. Then*

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| > t\right] \leq \frac{\mathbb{V}\left[X\right]}{t^2}.$$

Theorem 2.3.7 (Multiplicative Chernoff Bound)**.** *For each $i \in [n]$, let X_i be a $\{0,1\}$-valued random variable. Let $X = \sum_{i\in[n]} X_i$. Then for $0 < \delta < 1$,*

- $\mathbb{P}\left[X \geq (1+\delta)\mathbb{E}\left[X\right]\right] \leq e^{-\frac{\delta^2 \mathbb{E}[X]}{3}}$
- $\mathbb{P}\left[X \leq (1-\delta)\mathbb{E}\left[X\right]\right] \leq e^{-\frac{\delta^2 \mathbb{E}[X]}{2}}$.

Theorem 2.3.8 (Hoeffding Bound)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that, for values $a_1, b_1, \ldots, a_n, b_n$, each X_i satisfies $X_i \in [a_i, b_i]$. Then, for $S = \sum_{i=1}^n X_i$,*

$$\mathbb{P}\left[|S - \mathbb{E}\left[S\right]| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

2.3.3 Gaussian Distribution

Perhaps the most important probability distribution for this work is the Gaussian distribution, defined as follows:

Definition 2.3.9. Let $m, \sigma^2 \in \mathbb{R}$; then the 1-dimensional Gaussian distribution with mean m and variance σ^2 is the distribution with density

$$\phi(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

The n-dimensional Gaussian distribution can be defined as the product distribution of n 1-dimensional Gaussians:

Definition 2.3.10. The n-dimensional standard Gaussian distribution over $\mathbb{R}^n$ is the distribution with density

$$\phi(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\|x\|^2}{2}}.$$

Lemma 2.3.11. *Let $X \sim \mathcal{N}(0,1)$ be a standard Gaussian. Then*

$$\mathbb{E}\left[|X|\right] = \frac{\sqrt{2}}{\sqrt{\pi}}.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[|X|\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} te^{-\frac{t^2}{2}}dt = \frac{\sqrt{2}}{\sqrt{\pi}}\int_{0}^{\infty} te^{-\frac{t^2}{2}}dt \\ &= \frac{\sqrt{2}}{\sqrt{\pi}}\int_{0}^{\infty} e^{-z}dz = \frac{\sqrt{2}}{\sqrt{\pi}} \qquad (\text{where } z = t^2/2, dz = tdt).\end{aligned}$$

□

Lemma 2.3.12 ([BLM13])**.** *Let z be drawn from the Gaussian distribution with mean 0 and variance σ^2. Then*

$$\mathop{\mathbb{P}}_{z\sim\mathcal{N}(0,\sigma^2)}\left[z \geq t\right] \leq e^{-\frac{t^2}{2\sigma^2}}.$$

2.3.4 Projections and Indiscrete Distributions

A step that will be frequently taken in this work is to project a probability distribution μ over $\mathbb{R}^n$ onto a 1-dimensional vector, producing a univariate distribution. We will use the following notation:

Definition 2.3.13 (1-Dimensional Projection)**.** Let μ be any probability distribution over $\mathbb{R}^n$, and let $w \in \mathbb{R}^n$ be any vector with $\|w\| = 1$. Then we define the measure μ_w to be the 1-dimensional projection (i.e. the pushforward measure) of μ onto w:

$$\mu_w(A) := \mu\left\{x \in \mathbb{R}^n : \langle w, x\rangle \in A\right\}.$$

An important class of distributions over $\mathbb{R}^n$ is the set of distributions μ that are "continuous" in the sense that any 1-dimensional projection μ_w is continuous. Since these themselves may not be continuous (meaning absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, see for example [Fel68]), we will instead call them *indiscrete*:

Definition 2.3.14 (Indiscrete Distributions)**.** Let μ be a probability distribution over $\mathbb{R}^n$. We say μ is an indiscrete distribution if for all $w \in \mathbb{R}^n$ with $\|w\| = 1$, μ_w is a continuous univariate distribution.

For example, the Gaussian distribution is both continuous and indiscrete, while the uniform distribution μ over the sphere is indiscrete but not continuous, since the surface of a sphere has Lebesgue measure 0 but measure 1 under μ.

2.4 Spheres

When working with spheres and Gaussians, the Gamma function is unavoidable. This function is is a generalization of the factorial:

Definition 2.4.1 (Gamma function)**.** For all $t \in \mathbb{R}$,

$$\Gamma(t) := \int_0^\infty x^{t-1} e^{-x} dx \,.$$

A generalization of the factorial function to the real numbers, the Gamma function satisfies some very well established properties:

Fact 2.4.2. *The Gamma function satisfies the following:*

1. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$*;*
2. *For all* $n \in \mathbb{N}^+ : \Gamma(n) = (n-1)!$*;*
3. *For all* $t \in \mathbb{R}, t > 1 : \Gamma(t) = (t-1)\Gamma(t-1)$*.*
4. Γ *is convex on positive reals.*

The next property is a useful approximation to the Gamma function that greatly simplifies quantities involving ratios of two Gamma functions, which often appear in the analysis of Gaussians or spheres.

Theorem 2.4.3 ([Wen48], see [Qi10], equation 2.8)**.** *Let* $x > 1$ *be a positive real and* $0 \leq \epsilon \leq 1$*. Then*

$$(x-1)^\epsilon \Gamma(x-\epsilon) \leq \Gamma(x) \leq (x-\epsilon)^\epsilon \Gamma(x-\epsilon) \,.$$

Now that we have defined the Gamma function, we can find the volume and surface area of the high-dimensional sphere (for which the calculations are standard). For these calculations the following inequality, ubiquitous in computer science, will be useful:

Fact 2.4.4. *For all* $x \in \mathbb{R}, 1 + x \leq e^x$*.*

Proof. At $x = 0$, we have $1 + x = 1 = e^x$. $\frac{d}{dx}(1+x) = 1$ while $\frac{d}{dx}e^x = e^x$ which is greater than 1 for $x > 0$ and less than 1 for $x < 0$; thus e^x grows faster than $1 + x$ for positive x and shrinks slower than $1 + x$ for negative x. □

Definition 2.4.5 (Volume and Surface Area of the n-Sphere)**.** Let λ denote the Lebesgue measure on $\mathbb{R}^n$. Then the volume and surface are of the n-sphere with radius r are, respectively,

$$V_n(r) := \lambda\left\{x \in \mathbb{R}^n : \|x\|_2 \leq r\right\} \quad \text{and} \quad S_n(r) := \frac{d}{dr}V_n(r) = \int_0^{2\pi} r d\theta_1 \cdots \int_0^{2\pi} r d\theta_{n-1}$$

where in the definition of S_n we have converted the integral

$$\int_{\{x:\|x\|=r\}} \lambda(dx) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathbb{1}\left[\|x\|_2 = r\right] \lambda(dx_1) \cdots \lambda(dx_2)$$

to polar coordinates. For notational simplicity, I will write S_{n-1} to mean $S_{n-1}(1)$.

Proposition 2.4.6. *The volume and surface area of a sphere in n dimensions with radius r are, respectively,*

$$V_n(r) = r^n \frac{\sqrt{\pi^n}}{\Gamma\left(\frac{n}{2}+1\right)} \quad \textit{and} \quad S_{n-1}(r) = r^{n-1}\frac{2\sqrt{\pi^n}}{\Gamma\left(\frac{n}{2}\right)}\,.$$

Proof. First we integrate the exponential function over the whole space:

$$\begin{aligned} I_n &:= \int_{\mathbb{R}^n} e^{-\|x\|^2}\lambda(dx) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-(\sum_{i\in[n]} x_i^2)}\lambda(dx_1)\lambda(dx_2)\cdots\lambda(dx_d) \\ &= \int_{-\infty}^{\infty} e^{-x_1^2}\int_{-\infty}^{\infty} e^{-x_2^2}\cdots\int_{-\infty}^{\infty} e^{-x_n^2} dx_1 dx_2 \cdots dx_d \\ &= \left(\int_{-\infty}^{\infty} e^{-t^2}dt\right)^n = \left(2\frac{\sqrt{\pi}}{2}\int_0^{\infty}\frac{2}{\sqrt{\pi}}e^{-t^2}dt\right)^n = \sqrt{\pi}^n\,, \end{aligned}$$

by the identity for the Gaussian error function $\int_0^{\infty}\frac{2}{\sqrt{\pi}}e^{-t^2}dt = 1$. Now we express the

integral in terms of the surface area of the sphere:

$$\begin{aligned}
I_n &= \int_0^\infty \int_{\substack{x\in\mathbb{R}^n \\ \|x\|=r}} e^{-\|x\|^2} \lambda(dx)\lambda(dr) = \int_0^\infty e^{-r^2} \int_{\substack{x\in\mathbb{R}^n \\ \|x\|=r}} \lambda(dx)\lambda(dr) \\
&= \int_0^\infty e^{-r^2} \left(\int_0^{2\pi} r d\theta_1 \int_0^{2\pi} r d\theta_2 \cdots \int_0^{2\pi} r d\theta_{n-1} \right) dr \\
&= S_{n-1}(1) \int_0^\infty r^{n-1} e^{-r^2} dr \\
&= S_{n-1}(1) \int_0^\infty t^{\frac{n-1}{2}} e^{-t} \frac{1}{2\sqrt{t}} dt \qquad (t = r^2, dt = 2rdr) \\
&= S_{n-1}(1) \frac{1}{2} \int_0^\infty t^{\frac{n}{2}-1} e^{-t} dt \\
&= S_{n-1}(1) \frac{1}{2} \Gamma\left(\frac{n}{2}\right) .
\end{aligned}$$

Combining these, we have $S_{n-1}(1) = \frac{2\sqrt{\pi^n}}{\Gamma\left(\frac{n}{2}\right)}$ and, as noted earlier,

$$S_{n-1}(r) = \int_0^{2\pi} \cdots \int_0^{2\pi} r^{n-1} d\theta_1 \cdots d\theta_{n-1} = r^{n-1} S_{n-1}(1) .$$

To compute the volume, we integrate the surface area over the radius:

$$V_n(r) = \int_0^r S_{n-1}(t) dt = S_{n-1}(1) \int_0^r t^{n-1} dt = S_{n-1}(1) \frac{r^n}{n} = r^n \frac{\sqrt{\pi^n}}{\frac{n}{2}\Gamma\left(\frac{n}{2}\right)} = r^n \frac{\sqrt{\pi^n}}{\Gamma\left(\frac{n}{2}+1\right)} . \square$$

Using this proposition and the approximation of the Gamma function, we can get a good approximation on the ratio between areas of spheres of different dimensions:

Proposition 2.4.7. *For $n \geq 3$,*

$$\frac{\sqrt{n-2}}{\sqrt{2\pi}} \leq \frac{S_{n-2}}{S_{n-1}} \leq \frac{\sqrt{n-1}}{\sqrt{2\pi}} \qquad \textit{and} \qquad \frac{1}{\sqrt{2\pi}} \leq \frac{S_{n-2}}{\sqrt{n-2} S_{n-1}} \leq \frac{\sqrt{2}}{\sqrt{2\pi}} .$$

Proof. We have

$$\frac{S_{n-2}}{S_{n-1}} = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \in \left[\frac{\sqrt{n-2}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2\pi}\Gamma\left(\frac{n-1}{2}\right)}, \frac{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2\pi}\Gamma\left(\frac{n-1}{2}\right)}, \right]$$

by Theorem 2.4.3. The second part follows from $\frac{n-1}{n-2} \leq 2$ for $n \geq 3$. $\square$

We now know the total surface area of the sphere, and it will also be important to know approximately how much area is within a "spherical cap", that is, what fraction of the sphere is above some threshold in a given direction. This quantity is important since we can use it in concentration inequalities after projecting the sphere into 1 dimension.

Proposition 2.4.8. *Let v be any vector on the unit sphere in $\mathbb{R}^n$ (where $n \geq 3$) and let $0 < a < 1$ be any threshold. Suppose w is selected uniformly at random from the unit sphere. Then*

$$\mathop{\mathbb{P}}_{x}\left[\langle x, v\rangle \geq a\right] \leq \min\left\{\frac{1}{a\sqrt{\pi(n-2)}}e^{-\frac{a^2(n-2)}{2}}, \sqrt{2}e^{-\frac{a^2(n-2)}{2}}\right\},$$

and if $a \geq \frac{\pi}{3}$ we also have

$$\mathop{\mathbb{P}}_{x}\left[\langle x, v\rangle \geq a\right] \leq \frac{\sqrt{3}}{\sqrt{2\pi n}}(2\cos^{-1}(a))^{n-1}.$$

Proof. By rotation invariance, we may assume without loss of generality that $v = e_1$, the first standard basis vector. Then $\mathbb{P}\left[\langle x, v\rangle \geq a\right] = \mathbb{P}\left[x_1 \geq a\right]$ and

$$\mathbb{P}\left[x_1 \geq a\right] = \frac{1}{S_{n-1}}\int_a^1 S_{n-2}(\sqrt{1-r^2})dr = \frac{S_{n-2}}{S_{n-1}}\int_a^1 (1-r^2)^{\frac{n-2}{2}}dr \leq \frac{S_{n-2}}{S_{n-1}}\int_a^1 e^{-\frac{r^2(n-2)}{2}}dr$$

where we have used our favorite inequality in the rightmost step. To get the first bound:

$$\begin{aligned}
\int_a^1 e^{-\frac{r^2(n-2)}{2}}dr &= \frac{\sqrt{2}}{\sqrt{n-2}}\int_{a\sqrt{(n-2)/2}}^{\sqrt{(n-2)/2}} e^{-t^2}dt \qquad (t = r\sqrt{(n-2)/2}, dt = \sqrt{(n-2)/2}dr) \\
&\leq \frac{\sqrt{2}}{\sqrt{n-2}}\int_{a\sqrt{(n-2)/2}}^{\sqrt{(n-2)/2}} \frac{t}{a\sqrt{(n-2)/2}}e^{-t^2}dt \qquad (\text{since } t \geq a\sqrt{(n-2)/2}) \\
&\leq \frac{1}{a(n-2)}\int_{a\sqrt{(n-2)/2}}^{\infty} 2te^{-t^2}dt \\
&= \frac{1}{a(n-2)}e^{-\frac{a^2(n-2)}{2}}.
\end{aligned}$$

Now for the second bound, we make use of Lemma 2.3.12:

$$\int_a^1 e^{-r^2\frac{n-2}{2}}dr = \sqrt{2\pi\sigma^2}\int_a^1 \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}dr = \sqrt{2\pi\sigma^2}\mathop{\mathbb{P}}_{z\sim\mathcal{N}(0,\sigma^2)}\left[z > a\right] \leq \frac{\sqrt{2\pi}}{\sqrt{n-2}}e^{-\frac{a^2(n-2)}{2}}$$

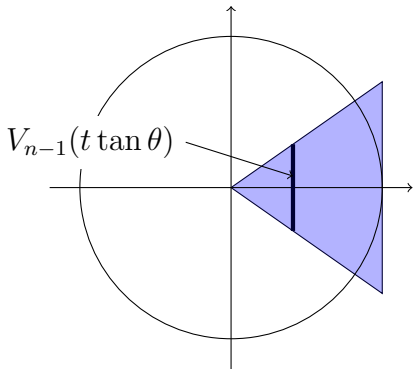


Figure 2.1: Overapproximation of a spherical cap.

where $\sigma^2 = 1/(n-2)$. Finally, we note that Theorem 2.4.3 implies

$$\frac{S_{n-2}}{S_{n-1}} = \frac{2\sqrt{\pi^{n-1}}}{\Gamma\left(\frac{n-1}{2}\right)}\frac{\Gamma\left(\frac{n}{2}\right)}{2\sqrt{\pi^n}} \leq \frac{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2\pi}\Gamma\left(\frac{n-1}{2}\right)} = \frac{\sqrt{n-1}}{\sqrt{2\pi}}\,.$$

Note that since $n \geq 3$ we have $\frac{\sqrt{n-1}}{\sqrt{n-2}} \leq \sqrt{2}$, which completes the proof of the first two bounds.

Finally, we get the last bound from [Lon94]. Observe that

$$\mathop{\mathbb{P}}_{x:\|x\|=1}\left[\langle x, v\rangle \geq a\right] = \mathop{\mathbb{P}}_{x:\|x\|\leq 1}\left[\|x\|\left\langle \frac{x}{\|x\|}, v\right\rangle \geq a\right]$$

where in the second expression the random variable x is drawn from the unit ball rather than the unit sphere, so we may bind the volume of this cone rather than the area of the cap. The angle of the cone is $\theta = \cos^{-1}(a)$ so for each $t \in [0,1]$ the radius of the 'disk' is $t\tan\theta$. We get an upper bound by stretching the cone out to radius 1, where it overapproximates the area we are interested in (see Figure 2.4). The volume of this

overapproximation is

$$\begin{aligned}\frac{1}{V_n}\int_0^1 V_{n-1}(t\tan\theta)dt &= \frac{V_{n-1}}{V_n}\int_0^1 t^{n-1}(\tan\theta)^{n-1}dt = \frac{\Gamma\left(\frac{n+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n+1}{2}\right)}(\tan\theta)^{n-1}\int_0^1 t^{n-1}dt \\ &= \frac{\Gamma\left(\frac{n+2}{2}\right)}{n\sqrt{\pi}\Gamma\left(\frac{n+1}{2}\right)}(\tan\theta)^{n-1} \leq \frac{\sqrt{n+1}}{n\sqrt{\pi}}(\tan(\cos^{-1}a))^{n-1} \\ &\leq \frac{\sqrt{3}}{\sqrt{2\pi n}}(2\cos^{-1}a)^{n-1}\,,\end{aligned}$$

where we have used Theorem 2.4.3 in the second-last inequality, and the following two facts in the final inequality:

$$\frac{n+1}{n^2} = \frac{1}{n}\left(1+\frac{1}{n}\right) \leq \frac{3}{2n}$$

so $\frac{\sqrt{n-1}}{n} \leq \frac{\sqrt{3}}{\sqrt{2n}}$, and for $\theta \leq \frac{\pi}{3}$,

$$\tan\theta = \frac{\sin\theta}{\cos\theta} \leq \frac{\theta}{1/2} = 2\theta\,. \qquad \square$$

2.5 Discrete Fourier Analysis

Recently, discrete Fourier analysis of boolean functions has led to a large number of results in the study of these functions; see, for example, the recent book by Ryan O'Donnell [O'D14]. Many of the results that I will review in this work make use of these techniques, so I will give a brief introduction to the most important concepts, all of which can be found in [O'D14].

The main idea of Fourier analysis of boolean functions is that we can decompose all boolean functions (i.e. functions $\{\pm1\}^n \to \mathbb{R}$) into a linear combination of orthonormal "basis" functions: the parity functions. The parity functions are those that multiply a subset of coordinates:

Definition 2.5.1 (Parity Functions)**.** For any $x \in \{\pm1\}^n$ and $S \subseteq [n]$,

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

Parity functions are orthonormal with respect to the following definition of an inner product:

Definition 2.5.2 (Inner Product of Functions)**.** Let $f, g \colon \{\pm 1\}^n \to \mathbb{R}$. Then

$$\langle f, g \rangle := \mathop{\mathbb{E}}_{x \sim \{\pm 1\}^n} [f(x)g(x)]$$

where x is drawn uniformly at random from $\{\pm 1\}^n$.

We can easily verify the orthonormality of these functions: for $A, B \subseteq [n], A \neq B$ we have some k that satisfies, without loss of generality, $k \in A \setminus B$. Thus

$$\langle \chi_A, \chi_B \rangle = \mathbb{E}\left[\chi_A(x)\chi_B(x)\right] = \mathbb{E}\left[\prod_{i \in A} x_i \cdot \prod_{j \in B} x_j\right] = \mathbb{E}\left[x_k\right] \mathbb{E}\left[\prod_{i \in A \setminus \{k\}} x_i \cdot \prod_{j \in B} x_j\right] = 0$$

since $\mathbb{E}\left[x_k\right] = 0$; here we have used the independence of x_i, x_j for all $i \neq j$. Therefore, since the space of all functions $\{\pm 1\}^n \to \mathbb{R}$ is of dimension 2^n (each $\{\pm 1\}^n$ vector is a "coordinate"), this set of 2^n orthonormal functions forms a basis. To find the Fourier coefficients, we simply take the inner product of the function with each basis vector:

Definition 2.5.3 (Fourier coefficients)**.** Let $S \subseteq [n]$. Then we write

$$\hat{f}(S) := \langle \chi_S, f \rangle$$

and for the coefficients of sets $S = \{i\}$ of size 1, we will write $\hat{f}(i) := \hat{f}(\{i\})$. Thus we have, for any $f : \{\pm 1\}^n \to \mathbb{R}$,

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

Some basic identities are Plancherel's and Parseval's identities:

Fact 2.5.4 (Plancherel's Identity)**.** *Let $f, g : \{\pm 1\}^n \to \mathbb{R}$. Then*

$$\langle f, g \rangle = \mathbb{E}\left[f(x)g(x)\right] = \sum_{S \subseteq [n]} \hat{f}(S)\hat{g}(S) \,.$$

This follows immediately from the orthonormality of the parity functions, and the next fact follows immediately from this:

Fact 2.5.5 (Parseval's Identity)**.** *Let* $f : \{\pm 1\}^n \to \mathbb{R}$*. Then*

$$\langle f, f \rangle = \mathbb{E}\left[f(x)^2\right] = \sum_{S \subseteq [n]} \hat{f}(S)^2$$

and if $f : \{\pm 1\}^n \to \{\pm 1\}$*,* $1 = \mathbb{E}\left[f(x)^2\right] = \sum_{S \subseteq [n]} \hat{f}(S)^2$*.*

The Fourier expansion lets us examine some important properties of boolean functions in elegant ways. We will provide some of the basic ideas for analyzing boolean functions, and will leave any more advanced concepts until they are necessary.

One essential question we might ask about a function is, How much does the output of the function depend on each coordinate? We might have, for example, a function that depends on only one coordinate, in which case we would say informally that the coordinate has a very high "influence" on the function. On the other hand, a coordinate might not matter very much to the output; maybe it is completely ignored, or maybe it is only influential for a small number of inputs. We formalize this intuitive concept of influence as follows:

Definition 2.5.6 (Influence)**.** For $f : \{\pm 1\}^n \to \{\pm 1\}$, the i^{th} *influence* is

$$\mathrm{Inf}_i(f) := \mathop{\mathbb{P}}_{x \sim \{\pm 1\}^n}\left[f(x^{i \leftarrow +1}) \neq f(x^{i \leftarrow -1})\right] .$$

The total influence of a function is

$$\mathrm{Inf}(f) := \sum_{i \in [n]} \mathrm{Inf}_i(f) .$$

For functions $f : \{\pm 1\}^n \to \mathbb{R}$, this definition no longer works very well. So we generalize the above by defining a "derivative" for each coordinate:

$$D_i f(x) = \frac{f(x^{i \leftarrow 1}) - f(x^{i \leftarrow -1})}{2} .$$

Note that $D_i f(x) = \pm 1$ for boolean-valued functions. Then we can define

$$\mathrm{Inf}_i(f) := \mathop{\mathbb{E}}_{x}\left[(D_i f(x))^2\right]$$

which is the same as the above definition when f is boolean-valued. Intuitively, the influence of a coordinate is the average effect it has on the output.

The stability of a function represents how robust the function is to noise: when we flip coordinates independently with some probability, a stable function will be unlikely to change values.

Definition 2.5.7 (Stability)**.** For $f : \{\pm 1\}^n \to \{\pm 1\}$, the *stability* of f under noise $\rho \in [-1, 1]$ is

$$\mathrm{Stab}_\rho(f) := \mathop{\mathbb{E}}_{x \sim_\rho y}[f(x)f(y)]$$

where $x \sim_\rho y$ means that x is drawn uniformly at random from $\{\pm 1\}^n$ and y is ρ-correlated with x; i.e. y is the vector obtained from x by flipping each coordinate with probability $\frac{1}{2}(1 - \rho)$.

A common characterization of the noise stability is the following:

Proposition 2.5.8. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ and $\rho \in [-1, 1]$. Then*

$$\mathrm{Stab}_\rho(f) = \sum_{S \subseteq [n]} \rho^{|S|} \hat{f}(S)^2 .$$

Proof. Consider the function $T_\rho f(x) = \mathop{\mathbb{E}}_{y \sim_\rho x}[f(y)]$. The Fourier expansion is

$$\mathop{\mathbb{E}}_{y \sim_\rho x}\left[\sum_{S \subseteq [n]} \hat{f}(S)\chi_S(y)\right] = \sum_{S \subseteq [n]} \hat{f}(S) \mathop{\mathbb{E}}_{x \sim \rho y}[\chi_S(y)] = \sum_{S \subseteq [n]} \hat{f}(S) \cdot T_\rho \chi_S(x) .$$

Considering only $T_\rho \chi_S$ and using independence, we have

$$\mathop{\mathbb{E}}_{y \sim_\rho x}\left[\prod_{i \in S} y_i\right] = \prod_{i \in S} \mathop{\mathbb{E}}_{x \sim_\rho y}[y_i] = \prod_{i \in S}\left(x_i \frac{1}{2}(1 + \rho) - x_i \frac{1}{2}(1 - \rho)\right) = \prod_{i \in S} \rho x_i = \rho^{|S|}\chi_S(x) .$$

Finally, we have $\mathrm{Stab}_\rho(f) = \mathop{\mathbb{E}}_{x, y \sim_\rho x}[f(x)f(y)] = \mathop{\mathbb{E}}_{x}\left[f(x) \mathop{\mathbb{E}}_{y \sim_\rho x}[f(y)]\right] = \mathbb{E}[f(x)T_\rho f(x)] = \langle f, T_\rho f\rangle$. □

Complementary to noise stability is noise sensitivity, or, the probability that applying noise to a function will change its value. That is, what is the probability, over all inputs, that when we flip each coordinate with probability ρ the function value will change? We can express this notion in terms of the sensitivity:

Definition 2.5.9 (Noise Sensitivity)**.** For $f : \{\pm 1\}^n \to \{\pm 1\}$, the *noise sensitivity* of f under noise $\rho \in [0, 1]$ is

$$\mathrm{NS}_\rho(f) := \frac{1}{2}\left(1 - \mathrm{Stab}_{1-2\rho}(f)\right) .$$

A well-known fact relevant to halfspaces is that, for unate functions, the influence is exactly the corresponding first-degree Fourier coefficient for that coordinate:

Definition 2.5.10. Unate Function Let $f : \{\pm 1\}^n \to \{\pm 1\}$. f is *unate* if for all coordinates $i \in [n]$, either $\forall x : f(x^{i\leftarrow -1}) \leq f(x^{i\leftarrow 1})$ or $\forall x : f(x^{i\leftarrow 1}) \leq f(x^{i\leftarrow -1})$.

Fact 2.5.11. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a unate function. Then*

$$\mathrm{Inf}_i(f) = \left|\hat{f}(i)\right| .$$

Proof. Let $i \in [n]$ and suppose that $f(x^{i\leftarrow 1}) \geq f(x^{i\leftarrow -1})$ for all x. Then

$$\begin{aligned}
\mathrm{Inf}_i(f) &= \mathbb{P}\left[f(x) \neq f(x^{\oplus i})\right] = \mathbb{P}\left[f(x^{i\leftarrow 1}) > f(x^{i\leftarrow -1})\right] \qquad \text{(monotonicity)} \\
&= \mathbb{E}\left[\mathbb{1}\left[f(x^{i\leftarrow 1}) = 1, f(x^{i\leftarrow -1}) = -1\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}\left[f(x^{i\leftarrow 1}) = 1\right] - \mathbb{1}\left[f(x^{i\leftarrow -1}) = 1\right]\right] .
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\hat{f}(i) &= \mathbb{E}\left[x_i f(x)\right] = \frac{1}{2}\Big(\mathbb{E}\left[f(x) \;\mid\; x_i = 1\right] - \mathbb{E}\left[f(x) \;\mid\; x_i = -1\right]\Big) \\
&= \frac{1}{2}\left(\mathbb{P}\left[f(x^{i\leftarrow 1}) = 1\right] - \mathbb{P}\left[f(x^{i\leftarrow 1}) = -1\right] - \mathbb{P}\left[f(x^{i\leftarrow -1}) = 1\right] + \mathbb{P}\left[f(x^{i\leftarrow -1}) = -1\right]\right) \\
&= \frac{1}{2}\left(2\mathbb{P}\left[f(x^{i\leftarrow 1}) = 1\right] - 1 + 1 - 2\mathbb{P}\left[f(x^{i\leftarrow -1}) = 1\right]\right)
\end{aligned}$$

which is equal to the influence. For the other cases, a similar proof holds. □

More specialized Fourier analysis will be described as necessary.

Chapter 3

Problem Definition and Prior Work

3.1 Problem Definition

I will now define precisely what it means to test halfspaces in the property testing framework. The idea of a property testing algorithm is to approximately solve a decision problem using a small number of queries to the input object; any property testing problem must have the following:

- A class of objects, say C; for example, the set of undirected graphs, or in this case the set of functions $\mathbb{R}^n \to \{\pm 1\}$.
- A "property" of the objects, that is, a subset $P \subset C$; in this case the set of linear threshold functions.
- A distance metric $\mathsf{dist} : C \times C \to \mathbb{R}$ that tells us "how far away" objects are from each other; in this case the probability that two functions differ on a random point. We also write
$$\mathsf{dist}(f, P) := \inf_{g \in P} \mathsf{dist}(f, g)$$
as the distance of a function to the set P. We say f is ϵ–close to P if $\mathsf{dist}(f, P) < \epsilon$ and ϵ–far otherwise.

As mentioned in the introduction, property testing is a subfield of sublinear algorithms. Sublinear algorithms are a class of hyper–efficient algorithms which are required to use sublinear resources: either sublinear time or space. Sublinear-space algorithms, such as streaming algorithms, are not allowed to store their input; sublinear-time algorithms are not allowed to even look at their entire input. We give algorithms "oracle access" to their

input, an element from the set C. By this we mean that the algorithm is given access to an "oracle" that answers queries of a certain kind about the input object. For example, an algorithm for testing graph properties might be provided an oracle that answers queries of the form "Are vertices i, j connected by an edge?" For inputs that are functions, say $f : A \to B$, the algorithm is given the oracle $x \mapsto f(x)$, for points $x \in A$, that answers the question "What is the value of f at the point x?"

With a class of objects, a property, an oracle model, and a distance metric in hand, we can define what we mean by a "testing algorithm"; since I am concerned with $\{\pm 1\}$-valued functions in this work, I provide the specialized definition for this domain:

Definition 3.1.1 (Testing Algorithm)**.** Let $\mathcal{P}$ be a property (set) of functions $\mathbb{R}^n \to \{\pm 1\}$, let μ be a probability measure on $\mathbb{R}^n$, and let O be some oracle. For two function $f, g : \mathbb{R}^n \to \{\pm 1\}$, we define the distance between f, g as $\mathsf{dist}(f, g) = \mathop{\mathbb{P}}_{x \sim \mu}[f(x) \neq g(x)]$, and the distance of f to $\mathcal{P}$ as $\mathsf{dist}(f, \mathcal{P}) = \inf_{g \in \mathcal{P}} \mathsf{dist}(f, g)$. A randomized algorithm A is an ϵ-tester for $\mathcal{P}$ under μ (with oracle O) if and only if, given oracle access to f via O, and ϵ as a parameter, it satisfies:

1. Completeness: for all $f \in \mathcal{P}$, $\mathbb{P}[A(f, \epsilon) = 1] > 2/3$
2. Soundness: for all $f \notin \mathcal{P}$ such that $\mathsf{dist}(f, \mathcal{P}) > \epsilon$, $\mathbb{P}[A(f, \epsilon) = 1] < 1/3$

where the probabilities are over the randomness of the algorithm A.

For this thesis I also want to restrict the algorithm to using only random samples. That is, the oracle we give the the algorithm always produces a random point $x \sim \mu$ and its label $f(x)$.

Definition 3.1.2 (Sampling Algorithm)**.** A randomized algorithm A is a *sampling algorithm* if it uses an oracle that provides only answers $(x, f(x))$ where x is selected independently at random from some distribution μ.

However, we will allow the testing algorithm to know something about the distribution of points. I will assume that the algorithm is capable of computing the density function $\mu(x)$ for any point $x \in \mathbb{R}^n$, as well as the density of any 1-dimensional projection $\mu_w(x)$, along with the cumulative distribution function of μ_w and its inverse.

The goal for this testing algorithm should be to use as few random samples as possible. In particular, the algorithm should use much fewer than n samples when the input function has domain $\mathbb{R}^n$. In the next section I will summarize some of what is known about learning

halfspaces, with the main lesson being that roughly n random samples is sufficient to learn a halfspace. Obviously we want the testing algorithm to be much faster than the learning algorithm. (In fact one may use a learning algorithm as a tester with some minor modifications, see [Ron08].)

An idea that immediately pops into mind is to take a set of random samples and attempt to construct a linear separator (say, by linear programming). If it can be done, return "yes" and otherwise return "no". In the introduction I hinted that this approach cannot work. In this chapter I will define the VC dimension, which in particular describes the number (labelled) points in an n-dimensional space that can always be separated by a hyperplane. As it turns out, if we have fewer than n points, we can always find a linear separator! So this strategy cannot beat the sample complexity of learning.

3.2 Learning Halfspaces

The models and definitions available in the machine learning literature are so numerous that an innocent reader's mind erupts in panic, sending tremors of fear into the depths of his soul. Each learning paper introduces roughly one new model and one new headache; nevertheless, let us conquer our fear along with a few of the most important definitions.

Foremost and most easily parsable of these models is that of Probably Approximately Correct (PAC) learning, originating with the founding work of Valiant [Val84]:

Definition 3.2.1 (PAC Learning)**.** Let X be some domain set, $\mathcal{C}$ be a "concept" class of functions, and $\mathcal{H}$ be a "hypothesis" class of functions $X \to \{\pm 1\}$. Let $m : \mathbb{R}^2 \to \mathbb{N}$ and A be any algorithm that takes as input any set Q of labelled samples $Q = \{(x_i, \ell_i)\}$ for $x_i \in X, \ell_i \in \{\pm 1\}$ for each i, and produces as output a hypothesis function $h \in \mathcal{H}$. Then A is an $m(\epsilon, \delta)$-sample PAC-learning algorithm for $\mathcal{C}$ if for all $\epsilon \in (0,1), \delta \in (0, 1/2)$, all distributions D over X, and all functions $f \in \mathcal{C}$,

$$\mathop{\mathbb{P}}_{Q \sim D^m} [\mathsf{dist}_D(A(Q, f(Q)), f) > \epsilon] < \delta$$

where $(Q, f(Q)) = \{x_i, f(x_i)\}_{i \in m(\epsilon,\delta)}$.

From this basic definition, a number of generalizations and modifications are available. To determine which models are the most appropriate matches for our property testing model, we can answer the following questions (keep in mind the idea of using the testing algorithm as a preprocessing step):

Is the learner required to produce a function from the concept class? If not, so $\mathcal{C} \neq \mathcal{H}$, the learner is called "improper" and is allowed to return any function in $\mathcal{C}$ within distance ϵ of the target; otherwise the algorithm is proper (and $\mathcal{C} = \mathcal{H}$). The property testing model implies nothing for this question.

Are the sample points and labels error-free, or is there some kind of noise in the data? In this work, I am assuming that the points and labels are error-free.

What does the algorithm know about the distribution of samples? Is it completely known, does it belong to some known class, or is it completely unknown? If the learner doesn't know the distribution, it is called "distribution-free". The original definition of PAC learning assumes the learner is distribution-free, although it is common to study more restricted learning algorithms. In this work, I assume the algorithm knows the distribution. For learning halfspaces, standard distributions are the n-dimensional Gaussian distribution and the uniform distribution over the either the n-sphere or the boolean hypercube. Distribution-free testing has been studied (e.g. [GS07]) but I am not using this model in this work (I will say more about future work on distribution-free testing in Section 8).

Does the learner get a label for each sample, or must it ask for labels for specific sample points? If the learner automatically receives a label for each random sample (and cannot make queries), it is a "passive" learner; otherwise it must request labels for some of the samples and is called an "active" learner. For active learners, there is another complexity measure: the number of labels it must request. These models have their counterparts in property testing ("active testing" was defined and explored in [BBBY12]), but I am concerned with only the passive model.

Can the learner make arbitrary queries or must it wait for random samples? Learning algorithms are usually (but not always) prohibited from making queries and must wait for random samples; this is because a typical application of a learning algorithm is to learn about some real-life phenomenon, and cannot construct its own examples (a canonical example is the diagnosis of diseases; we cannot construct a new patient according to parameters we choose and then check if they have the disease). However, the prior work done on testing halfspaces *has* allowed arbitrary queries, and indeed most property testing algorithms are granted this power; this is one motivation of the current work, since these models do not match the standard learning models and there is a gap in our understanding.

Does the target function belong to the hypothesis class of the learner? In other words, is the target function *realizable*? Here is where the idea of using the testing algorithm as a preprocessing step to learning is important: if our testing algorithm accepts a function, then we are only guaranteed that the target function is *close* to a halfspace

Model		Distribution				Results	
proper	agnostic	dist. free	n-sphere	cube	exact	time	samples
✓		✓			✓	poly	$\Theta\left(\frac{n}{\epsilon}\right)$
✓			✓		✓	poly [BL13]	$\Theta\left(\frac{n}{\epsilon}\right)$ [Lon94, BL13]
✓	✓	✓			✓	NP-hard [HSV95]	$\Theta\left(\frac{n}{\epsilon^2}\right)$
✓	✓	✓			$< 418/415$	NP-hard [BDEL03]	$\Theta\left(\frac{n}{\epsilon^2}\right)$
	✓	✓			$O(n)$ [KL93]	poly	
	✓		✓		✓	poly	$n^{O\left(1/\epsilon^4\right)}$ [KKMS08]
	✓			✓	✓	poly	$n^{O\left(1/\epsilon^2\right)}$ [KKMS08]
	✓		✓		$1+\mu$	poly $\left(n^{\frac{\log^3(1/\mu)}{\mu^2}}, \frac{1}{\epsilon}\right)$	poly [Dan15]

Table 3.1: A brief summary of what we know about learning halfspaces. Here we are treating δ as constant in the final two columns. A checkmark in the 'exact' column means the algorithm is trying to achieve opt, otherwise an approximation ratio is given.

(with high probability). The learning algorithm cannot assume that the function is actually a halfspace; this type of learning algorithm is known as *agnostic*, (see [KSS94] for the origin of this definition).

The history of halfspaces and machine learning are closely intertwined, and the literature on learning halfspaces is vast; to get a sense of this landscape, we will present a selection of work from this literature, with special attention paid to those works that have the most in common with our problem of testing halfspaces. Table 3.1 collects a few of the most relevant results on learning theory.

Work on property testing is usually concerned with sample or query complexity, rather than time complexity, so I will focus on what is known about the sample complexity of learning. And, recalling the idea that a property testing algorithm may be used as a preprocessing step for a learning algorithm, we will look at the models of learning that most closely match up with the result of property testing algorithm, that is, one where the input is guaranteed only to be close to the concept class (rather than within it). But first, we will some of the basics of the field, starting with the linear programming and Perceptron algorithms.

3.2.1 The Linear Programming Solution

Ignoring for now the problem of producing a hypothesis that is probably approximately correct, we will focus on the problem of finding a linear separator for a set of sample points, under the guarantee that such a separator exists. We will see later that solving this simpler problem suffices for the PAC learning problem as a whole.

Suppose $Q_\ell = \{(x_1, \ell_1), \ldots, (x_m, \ell_m)\}$ is a set of points such that there exists a function $f(x) = \text{sign}(\langle w, x \rangle)$ satisfying $f(x_i) = \ell_i$ for all $i \in [m]$. We can write the sample points as an $m \times n$ matrix

$$A = \begin{bmatrix} —\ell_1 x_1— \\ \vdots \\ —\ell_m x_m— \end{bmatrix}$$

Then we simply want to find a vector w such that $Aw \geq \epsilon \vec{1}$ (we choose some $\epsilon > 0$ to guarantee that no point x_i will have $\langle w, x_i \rangle = 0$). Thus, using a dummy objective function, we can use an LP solver to produce such a vector w when it is guaranteed to exist, giving us a polynomial-time algorithm.

3.2.2 The Perceptron Algorithm

A *perceptron* is a formalization intended to be a simple model of a neuron: its introduction in psychology by Rosenblatt was part of an attempt to understand better how humans learn [Ros58]. In this simplified model, the neuron receives a number of weighted signals (through its dendrites), and if the total weighted sum of the signals is greater than some threshold, it produces a signal (through the axon). In short, a neuron is modelled as a linear threshold function! So one can see why they have become so important in machine learning theory.

The Perceptron algorithm is a way to produce, given a number of examples, a normal vector for a (balanced) halfspace that correctly classifies each example (assuming that such a classification is possible). It works by going through the given examples, finding one that is incorrectly classified by the intermediate hypothesis, and adjusting the hypothesis accordingly, by adding or subtracting the misclassified point to the normal vector.

The following convergence property of the perceptron algorithm is well-known (see e.g. [SSBD14]):

Theorem 3.2.2. *Let $Q_\ell = ((x_1, \ell_1), \ldots, (x_m, \ell_m))$ be a set of labelled points consistent with a (balanced) linear threshold function $f(x) = \text{sign}(\langle w, x \rangle)$. Then the perceptron algorithm*

Algorithm 1 Perceptron Algorithm

Input: $Q_\ell = ((x_1, \ell_1), \ldots, (x_m, \ell_m))$ such that the labels $\{\ell_i\}$ are consistent with a linear separator

function PERCEPTRON(Q_ℓ)
  $w_0 \leftarrow \vec{0}, t \leftarrow 0$
  while $\exists i \in [m] : \ell_i \langle w_t, x_i \rangle \leq 0$ **do**
    $w_{t+1} \leftarrow w_t + \ell_i x_i$
    $t \leftarrow t + 1$
  return w_t

will produce a vector w such that $\text{sign}(\langle w, x_i \rangle) = \ell_i$ for all i, after at most $(r/s)^2$ iterations, where $r = \max_i \|x_i\|$ is the "radius" of the set and $s = \min_i |\langle w, x_i \rangle|$ is the smallest separation between an example point and the separating hyperplane.

Proof. Let w^* be a vector with Euclidean norm $\|w^*\| = 1$ such that $\text{sign}(\langle w^*, x_i \rangle) = \ell_i$ for all i. We will show that the angle between w_t and w^* shrinks as the algorithm performs more iterations.

Let k be the index of the vector x_k on which the update is performed at iteration t. First we show that the inner product between w^*, w_t grows at each iteration. The first execution of the loop initializes $w_1 \leftarrow \ell_i x_i$ for some i, so as the base case we take $t = 1$. Here $\langle w^*, w_1 \rangle = \ell_k \langle w^*, x_k \rangle = |\langle w^*, x_k \rangle|$ since, by definition, $\ell_k = \text{sign}(\langle w^*, x_k \rangle)$. Then by induction,

$$\begin{aligned} \langle w^*, w_t \rangle &= \langle w^*, w_{t-1} + \ell_k x_k \rangle = \langle w^*, w_{t-1} \rangle + \ell_k \langle w^*, x_k \rangle \\ &= \langle w^*, w_{t-1} \rangle + |\langle w^*, x_k \rangle| \\ &\geq t \min_i |\langle w^*, x_i \rangle| = ts \end{aligned} \tag{3.1}$$

Now that we know the inner product is growing, we need to know something about the norm. In the base case we have $\|w_1\|^2 = \|x_k\|^2$, and afterwards

$$\begin{aligned} \|w_t\|^2 &= \langle w_{t-1} + \ell_k x_k, w_{t-1} + \ell_k x_k \rangle \\ &= \|w_{t-1}\|^2 + \ell_k^2 \|x_k\|^2 + 2\ell_k \langle w_{t-1}, x_k \rangle \\ &\leq t \max_i \|x_i\|^2 = tr^2 \end{aligned} \tag{3.2}$$

where we have used the fact that, by definition, $\ell_k \langle w_{t-1}, x_k \rangle \leq 0$. Thefore, combining equations (3.1) and (3.2), we get

$$1 \geq \frac{\langle w^*, w_t \rangle}{\|w_t\|} \geq \frac{ts}{r\sqrt{t}}$$

which implies $t \leq \frac{r^2}{s^2}$. □

Note that, in this version of the proof, the convergence rate of the algorithm is given in terms of the parameter $s = \min_{x \in Q}(|\langle w, x \rangle|)$. This parameter is often called the "margin" of the sample set [SSBD14]; sample sets with large margins are intuitively easier to separate, and with some margin guarantee the perceptron algorithm can be much faster than the LP algorithm (such a guarantee holds for PAC learning on the sphere [Bau90]). Unfortunately the assumption of large margins is sometimes too much of a restriction on the model.

The Perceptron algorithm predates modern definitions of learning; its original purpose was merely to find a separating hyperplane for a given set of points, with no concern for which distribution they came from or how likely it is that the hypothesis correctly classifies further points. Nevertheless, some fundamental results in learning theory prove that such algorithms can work as PAC learning algorithms. Before we see these fundamental results, we need some fundamental definitions.

3.2.3 Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis Dimension, or VC Dimension, is a property of classes of functions that has many deep connections to learning theory and other areas. It tells us exactly when the above algorithms, which are examples of *empirical risk minimizers*, suffice for PAC learning and how many samples they will need.

Essential to the definition of VC dimension is the idea of *shattering*: a set of points is shattered by a concept class (a set of functions) if every possible labelling is induced by some function in the class.

Definition 3.2.3 (Shattering)**.** Let $\mathcal{X}$ be any domain set and $\mathcal{C}$ be any set of functions $\mathcal{X} \to \{\pm 1\}$. Let $X \subseteq \mathcal{X}$ be any finite set $\{x_1, \ldots, x_m\}$; then X is *shattered* by $\mathcal{C}$ if

$$\{(h(x_1), h(x_2), \ldots, h(x_m)) \colon h \in \mathcal{C}\} = \{\pm 1\}^m$$

Or in other words, for every labelling of X there is a function $h \in \mathcal{C}$ that produces that labelling.

Definition 3.2.4 (VC Dimension)**.** Let $\mathcal{X}$ be as above. Then the VC Dimension of $\mathcal{C}$, denoted $\mathsf{VC}(\mathcal{C})$, is

$$\mathsf{VC}(\mathcal{C}) := \max\{|C| : C \subseteq \mathcal{X}, \mathcal{C} \text{ shatters } C\}$$

If arbitrarily large sets $C \subset \mathcal{X}$ can be shattered, we say $\mathsf{VC}(\mathcal{C}) = \infty$.

The importance of this definition is exemplified in the following theorem:

Theorem 3.2.5 (Fundamental Theorem of Statistical Learning [SSBD14])**.** *Let $\mathcal{X}, \mathcal{C}$ be as above. Then the following are equivalent:*

- *$\mathcal{C}$ is (standard and agnostic) PAC learnable,*
- *$\mathcal{C}$ has finite VC Dimension,*
- *Any empirical risk minimizer for $\mathcal{C}$ is a (standard and agnostic) PAC learning algorithm.*

Further, if $\mathsf{VC}(\mathcal{C}) = d < \infty$, for all accuracy and confidence parameters ϵ, δ:

- *$\mathcal{C}$ is agnostic PAC learnable with sample complexity $\Theta\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$, and*
- *$\mathcal{C}$ is PAC learnable with sample complexity $\Theta\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$.*

The third statement has to do with *empirical risk minimizers*, which is what allows us to use the Perceptron algorithm as a PAC learner. An empirical risk minimizer for $\mathcal{C}$ is an algorithm which, given a set of examples X with labels generated by f, produces a function $h \in \mathcal{C}$ that minimizes the empirical error, defined as

$$\mathsf{dist}_X(h, f) := \frac{1}{|X|} \sum_{x \in X} \mathbb{1}\left[h(x) \neq f(x)\right]$$

Assuming $\mathcal{C}$ is the set of halfspaces and $f \in \mathcal{C}$, we can see that the Perceptron algorithm is indeed an empirical risk minimizer, since it produces a halfspace whose error on the examples is 0; thus the theorem proves that the Perceptron algorithm is a PAC learning algorithm for halfspaces, with sample complexity

$$\Theta\left(\frac{\mathsf{VC}(\mathcal{C}) + \log(1/\delta)}{\epsilon}\right)$$

All that remains is to find the VC dimension of halfspaces, which is well-known to be n in the balanced case, and $n+1$ in the general case (see e.g. [SSBD14]).

Theorem 3.2.6. *The VC dimension of the set of balanced halfspaces on the domain $\mathbb{R}^n$ is n, and the VC dimension of (not necessarily balanced) halfspaces on $\mathbb{R}^n$ is $n+1$.*

To prove this, we have to show that halfspaces (respectively, balanced halfspaces) cannot shatter any set of size $n+2$ (resp. $n+1$), and there exists a set of size $n+1$ (resp. n) that

is shattered by $\mathcal{C}$. We will in fact prove something a little bit stronger, namely, that *all* sets of size $n+1$ (resp. n) are shattered by $\mathcal{C}$ (provided the set is linearly independent).

Theorem 3.2.7. *Balanced halfspaces over $\mathbb{R}^n$ shatter a set X if and only if the points in X are linearly independent.*

General halfspaces over $\mathbb{R}^n$ shatter a set X if and only if the points in X are linearly independent after embedding them in the n-dimensional affine subspace of $\mathbb{R}^{n+1}$ defined by the translation vector e_{n+1}.

Proof. Let $X = \{x_1, x_2, \ldots, x_m\} \subset \mathbb{R}^n$. Define the $m \times n$ matrix A by using each point in X as a row, and let $\ell_1, \ldots, \ell_m$ be a set of labels. There exists a balanced halfspace that produces these labels if and only if there exists a normal vector w such that

$$b = Aw$$

satisfies $\operatorname{sign}(b_i) = \operatorname{sign}(\langle x_i, w\rangle) = \ell_i$ for all $i \in [m]$. If each point (row) is linearly independent then clearly we can achieve this labelling. On the other hand, if the points are not linearly independent then for some point, say x_1, we have $x_1 = \sum_{i=2}^{n} a_i x_i$ for some coefficients $\{a_i\}$. Assuming $\ell_1 = 1$, we have

$$0 \leq \langle w, x_1\rangle = \sum_{i=2}^{n} a_i \langle w, x_i\rangle$$

so clearly we cannot achieve the labelling $\ell_1 = 1, \ell_i = -\operatorname{sign}(a_i)$, which would make the sum negative; this concludes the case where the halfspaces are balanced.

If the halfspaces are not necessarily balanced, they are of the form $h(x) = \operatorname{sign}(\langle w, x\rangle - \theta)$ for some $w \in \mathbb{R}^n, \theta \in \mathbb{R}$. For any $x \in \mathbb{R}^n$ we define the embedding $\hat{x} \in \mathbb{R}^{n+1}$ such that $\hat{x}_i = x_i$ for $i \in [n]$ and $\hat{x}_{n+1} = 1$. Then $\operatorname{sign}(\langle w, x\rangle - \theta) = \operatorname{sign}(\langle \hat{w} - \theta e_{n+1}, \hat{x}\rangle)$, so we are now looking for a *balanced* halfspace in $\mathbb{R}^{n+1}$. Therefore a solution w, θ exists if and only if the points $\widehat{X} = \{\hat{x}_1, \cdots, \hat{x}_m\}$ are linearly independent in $\mathbb{R}^{n+1}$, by the previous claim. □

3.2.4 The Sample Complexity of Learning

In 1994, Philip Long proved a lower bound of $\Omega\left(\frac{1}{\epsilon}n + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$ independently random samples required to learn a balanced halfspace on the uniform sphere [Lon94]. This was significant because it matched an earlier lower bound given by Eurenfeucht *et al.* [EHKV89] for the *distribution-free* model, where the algorithm does not know the distribution in

advance; lower bounds for the distribution-free model must be larger than for any specific distribution, since a lower bound for a specific distribution imply the distribution-free bound as well. Thus, Long's lower bound is significant since it shows that this number of samples is required even when the distribution is very simple: a distribution-free bound is allowed to depend on arbitrarily complex distributions.

Long later proved a matching upper bound on the sample complexity (although this is an "information theoretic" bound, in the sense that no attention is paid to the time complexity of the algorithm) [Lon03]. In this subsection I will summarize both these papers.

Lower Bound

The main theorem proven in [Lon94] is

Theorem 3.2.8 ([Lon94])**.** *Let μ be the uniform distribution over the unit sphere $\{x \in \mathbb{R}^n : \|x\| = 1\}$, and let A be any algorithm such that, given a set of m labelled examples $\{(x_i, f(x_i))\}_{i\in[m]}$ of a balanced halfspace function $f : \mathbb{R}^n \to \{\pm 1\}$, produces a hypothesis halfspace h. Then for any $\epsilon \in (0, 1/2), \delta \in (0, 1/2)$, and independently randomly distributed $\{x_i\}_{i\in[m]}$, if*

$$\mathop{\mathbb{P}}_{\{x_i\}}\left[\mathop{\mathbb{P}}_{x}\left[h(x) \neq f(x)\right] < \epsilon\right] \geq 1 - \delta$$

where $h = A(\{(x_i, f(x_i))\}_{i\in[m]})$, then

$$m = \Omega\left(\frac{n + \log(1/\delta)}{\epsilon}\right)$$

I will follow the original proof, which has two parts: first, show that $\Omega(n/\epsilon)$ queries are required, and second, show that $\Omega(\log(1/\delta)/\epsilon)$ are required. The first part is achieved via the probabilistic method, and the second is achieved by a general argument using "continuous hard pairs".

The first part of the proof will use the following special case of a well-known lemma that places a bound on the number of different ways a set of points can be labelled by balanced halfspaces:

Lemma 3.2.9 (Sauer-Shelah-Perles lemma, see [SSBD14])**.** *For any $\{x_i\}_{i\in[m]} \subset \mathbb{R}^n$ and concept class $\mathcal{H}$ with VC dimension $\mathsf{VC}(\mathcal{H}) = d$,*

$$|\{(f(x_1), \cdots, f(x_m)) : f \in \mathcal{H}\}| \leq \left(\frac{em}{d}\right)^d$$

In particular, for halfspaces we have $d = n + 1$ and for balanced halfspaces, $d = n$.

The first lemma that we will prove shows that we can bound the sample complexity from below using the maximum number of mutually 2ϵ-far halfspaces:

Lemma 3.2.10. *Let μ be any distribution on $\mathbb{R}^n$ (where $n \geq 2$). Let F be any set of balanced halfspaces such that $\forall f, g \in F, \mathsf{dist}_\mu(f, g) \geq 2\epsilon$ for any $\epsilon > 0$. Then the number of samples required to PAC learn balanced halfspaces under μ with accuracy ϵ and confidence $1/2$ is at least*

$$\frac{n}{e}\left(\frac{|F|}{2}\right)^{1/n}$$

Proof. Let A be any deterministic learning algorithm such that for all balanced halfspaces f and for a random set of m labelled samples $Q = \{(x_i, f(x_i))\}_{i\in[m]}$,

$$\mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}(A(Q), f) < \epsilon\right] > \frac{1}{2}$$

Now let $F \subset \mathcal{H}$ be a set of balanced halfspaces such that for each distinct $g_1, g_2 \in F, \mathsf{dist}(g_1, g_2) \geq 2\epsilon$. For any sequence of samples $Q = (x_1, \ldots, x_m)$ and function f, denote by Q_f the labelled sequence $((x_1, f(x_1)), \ldots, (x_m, f(x_m)))$. Note that for each halfspace f,

$$\mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}(A(Q_f), f) < \epsilon\right] = \mathop{\mathbb{E}}_{Q}\left[\mathbb{1}\left[\mathsf{dist}(A(Q_f), f) < \epsilon\right]\right]$$

Then, taking the sum of this indicator over F, we have

$$\sum_{f\in F} \mathop{\mathbb{E}}_{Q}\left[\mathbb{1}\left[\mathsf{dist}(A(Q_f), f) < \epsilon\right]\right] = \sum_{f\in F} \mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}(A(Q_f, f), f) < \epsilon\right] > \frac{|F|}{2}$$

On the other hand, we have

$$\sum_{f\in F} \mathop{\mathbb{E}}_{Q}\left[\mathbb{1}\left[\mathsf{dist}(A(Q_f), f) < \epsilon\right]\right] = \mathop{\mathbb{E}}_{Q}\left[\sum_{f\in F} \mathbb{1}\left[\mathsf{dist}(A(Q_f), f) < \epsilon\right]\right]$$

Suppose that functions $f, g \in F$ produce the same labelling on Q, i.e $f(x_i) = g(x_i)$ for each $i \in [m]$. Then $Q_f = Q_g$ so $A(Q_f) = A(Q_g)$. Suppose that $\mathsf{dist}(A(Q_f), f) < \epsilon$; then $\mathsf{dist}(A(Q_f), g) < \epsilon$ since $A(Q_f) = A(Q_g)$, and therefore we must have $f = g$ since otherwise $2\epsilon \leq \mathsf{dist}(f, g) \leq \mathsf{dist}(A(Q_f), f) + \mathsf{dist}(A(Q_g), g) < 2\epsilon$. Thus, each labelling arising from

some function in f contributes at most 1 to the sum for each Q. That means we can apply the Sauer-Shelah-Perles Lemma 3.2.9 to get

$$\frac{|F|}{2} < \mathop{\mathbb{E}}_{Q}\left[\sum_{f\in F} \mathbb{1}\left[\mathsf{dist}(A(Q_f), f) < \epsilon\right]\right] \leq \left(\frac{em}{n}\right)^n$$

Rearranging to isolate m completes the proof. □

To complete the first part of the lower-bound, we must now show that a large enough set F of 2ϵ-far balanced halfspaces exists when μ is the uniform distribution over the sphere. We achieve this with the probabilistic method.

Lemma 3.2.11 ([Lon94], lemma 6)**.** *Let μ be the uniform distribution over the unit sphere in $\mathbb{R}^n$. Let $0 < \epsilon < 1/2$. Then there exists a set F of balanced halfspaces such that for all $f, g \in F, \mathsf{dist}_\mu(f, g) \geq 2\epsilon$ and*

$$|F| \geq \frac{\sqrt{n}}{2}\left(\frac{1}{2\pi\epsilon}\right)^n - 1$$

Proof. Let $w_1, \ldots, w_m$ be a sequence of vectors drawn independently and uniformly at random from the unit sphere, and let $h_1, \ldots, h_m$ be the balanced halfspaces $h_i(x) = \mathrm{sign}(\langle w_i, x\rangle)$ defined by these vectors. Using the notation $\angle(w_i, w_j) = \arccos(\langle w_i, w_j\rangle)$ to denote the angle between unit vectors, the distance between any two distinct functions h_i, h_j is

$$\mathsf{dist}(h_i, h_j) = \frac{\angle(w_i, w_j)}{\pi}$$

and the probability that two random unit vectors w_i, w_j have angle at most a is the area of the spherical cap above the threshold $\cos a$; in this case, we have

$$\begin{aligned}\mathop{\mathbb{P}}_{w_i,w_j}\left[\mathsf{dist}(h_i, h_j) < \epsilon\right] &= \mathop{\mathbb{P}}_{w_i,w_j}\left[\angle(w_i, w_j) < \pi\epsilon\right] = \mathop{\mathbb{P}}_{w_i,w_j}\left[\langle w_i, w_j\rangle \geq \cos(\pi\epsilon)\right] \\ &\leq \frac{\sqrt{3}}{\sqrt{2\pi n}}(2\pi\epsilon)^{n-1} \qquad \text{(Proposition 2.4.8)}\end{aligned} \tag{3.3}$$

The expected number of pairs of halfspaces which are too close for our set is thus

$$\mathbb{E}\left[\sum_{i\neq j} \mathbb{1}\left[d(h_i, h_j) < \epsilon\right]\right] \leq \frac{\sqrt{3}m^2}{\sqrt{2\pi n}}(2\pi\epsilon)^{n-1} \leq \frac{m^2}{\sqrt{n}}(2\pi\epsilon)^{n-1}$$

so by the probabilistic method, there exists a set F' with at most this many conflicting pairs. Removing one vector from each of these conflicting pairs gives us our set F with size

$$m - \frac{m^2}{\sqrt{n}}(2\pi\epsilon)^{n-1}$$

Optimizing the value of m gives us $\frac{\sqrt{n}}{2(2\pi\epsilon)^{n-1}}$; taking the floor of this value gives us the result we want. □

Combining these two lemmas, we see that the number of samples we need is at least

$$\frac{n-1}{e}\left(\frac{|F|}{4}\right)^{1/(n-1)} \geq \frac{n-1}{e}\left(\frac{\sqrt{n}}{8}\right)^{1/(n-1)}\frac{1}{2\pi\epsilon} \geq \frac{n-1}{2\pi e\epsilon}$$

We now move on to the second part of the bound, using the idea of "continuous hard pairs" to show that we must use at least $\Omega\left(\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$ samples.

Definition 3.2.12 (Continuous Hard Pairs)**.** Let μ be a probability distribution over $\mathbb{R}^n$ and let P be a class of functions $\mathbb{R}^n \to \{\pm 1\}$. Then P has continuous hard pairs (with respect to μ) if for all $0 < \epsilon < 1$, there exist $f, g \in P$ such that $\mathsf{dist}_\mu(f, g) = \epsilon$.

Example 3.2.13. This definition clearly applies to halfspaces on the uniform sphere: for any (balanced) f with normal w, we can pick a v of the appropriate angle to w so that the halfspace defined by v has exactly ϵ distance to f.

We will prove our lower bound for all spaces and function classes that have continuous hard pairs:

Theorem 3.2.14 ([Lon94], Theorem 8)**.** *Let μ be any distribution over $\mathbb{R}^n$ and let P be any class of measurable functions $\mathbb{R}^n \to \{\pm 1\}$ that has continuous hard pairs with respect to μ. Let A be any deterministic (ϵ, δ)-PAC-learning algorithm for P. Then A requires at least*

$$\Omega\left(\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$$

labelled random samples to learn P.

Proof. Let $f, g \in P$ such that $\mathsf{dist}(f, g) = 2\epsilon$ and let $Q = \{x_i\}_{i\in[m]}$ be a set of random samples from μ; define $Q_f := ((x_1, f(x_1)), \ldots, (x_m, f(x_m)))$ and similarly $Q_g := ((x_1, g(x_1)), \ldots, (x_m, f(x_m)))$. Let MISS be the event that for all $i \in [m]$, $f(x_i) = g(x_i)$, i.e.

the event that $Q_f = Q_g$. Assuming this event occurs, the algorithm A cannot produce a hypothesis that is ϵ-close to both f and g (by the triangle inequality), so for each set Q the algorithm fails on either f or g with probability 1. Therefore

$$\mathop{\mathbb{P}}_{Q}[\mathsf{dist}(A(Q_f), f) \geq \epsilon \quad | \quad \mathsf{MISS}] + \mathop{\mathbb{P}}_{Q}[\mathsf{dist}(A(Q_g), g) \geq \epsilon \quad | \quad \mathsf{MISS}] \geq 1$$

meaning that one of these probabilities is at least $1/2$; assume without loss of generality it is the first. The probability that MISS occurs is exactly $(1-2\epsilon)^m$ since $\mathsf{dist}(f,g) = 2\epsilon$, so

$$\delta > \mathop{\mathbb{P}}_{Q}[\mathsf{dist}(A(Q_f), f) \geq \epsilon] \geq \mathop{\mathbb{P}}_{Q}[\mathsf{MISS}] \mathop{\mathbb{P}}_{Q}[\mathsf{dist}(A(Q_f), f) \geq \epsilon \quad | \quad \mathsf{MISS}] \geq \frac{1}{2}(1-2\epsilon)^m$$

Taking the log of both sides, we have

$$m \geq \frac{\ln(2\delta)}{\ln(1-2\epsilon)} \geq -\frac{\ln(2\delta)}{2\epsilon} = \frac{1}{2\epsilon}\ln\left(\frac{1}{2\delta}\right)$$

where we have used our favorite inequality $\ln(1-x) \leq -x$. □

Therefore, we need at least

$$\max\left(\frac{n-1}{2\pi e\epsilon}, \frac{1}{2\epsilon}\ln\left(\frac{1}{2\delta}\right)\right) \geq \frac{1}{2\epsilon}\left(\frac{n-1}{2\pi e} + \frac{1}{2}\ln\left(\frac{1}{2\delta}\right)\right) = \Omega\left(\frac{n}{\epsilon} + \frac{1}{\epsilon}\ln\left(\frac{1}{\delta}\right)\right)$$

samples.

Upper Bound

Now we will prove the matching upper bound by giving an algorithm that uses only $O\left(\frac{n+\log(1/\delta)}{\epsilon}\right)$ samples. In the construction of this algorithm, we will not be concerned with time complexity, since the goal is only to minimize the number of random samples. In fact, the first step of the algorithm will be to construct a (possibly exponentially large) set G containing approximations for every possible halfspace: the existence of this set is the main lemma of the paper [Lon03]:

Lemma 3.2.15 ([Lon03], Lemma 5)**.** *Let $\mathcal{H}$ be the set of all balanced halfspaces in $\mathbb{R}^n$ and let μ be the uniform distribution over the unit sphere. There exists a universal constant C such that, for all $\epsilon \in (0,1]$, there exists a set $G \subset \mathcal{H}$ satisfying, for all $h \in \mathcal{H}$:*

- $\exists g \in G, \mathsf{dist}_\mu(h,g) \leq \frac{\epsilon}{4}$

- $\forall \alpha \geq \epsilon, \#\{g \in G : \mathsf{dist}(g,h) \leq \alpha\} \leq (C\alpha/\epsilon)^{n-1}$

Proof. Let G be a set obtained by repeatedly selecting vectors arbitrarily from the unit sphere such that each new vector defines a (balanced) halfspace of distance at least $\epsilon/4$ from all previously selected vectors, until there are no more such vectors to be added; we will show that this set satisfies the above properties.

The first property must be satisfied since otherwise there would be some $h \in \mathcal{H}$ such that $\mathsf{dist}(h,g) > \epsilon/4$ for all $g \in G$; by definition, if there was such a vector, it would have been added to G.

Let $\alpha \geq \epsilon$. By the Triangle Inequality, we know that the balls of radius $\epsilon/8$ around each function $g \in G$ must be disjoint, since each distinct functions $g_1, g_2 \in G$ satisfy $\mathsf{dist}(g_1,g_2) > \epsilon/4$. Then we have the inequality

$$\#\{g \in G : \mathsf{dist}(g,h) \leq \alpha\} \leq \frac{\mathbb{P}_w[\mathsf{dist}(h,h_w) \leq \alpha + \epsilon/8]}{\mathbb{P}_w[\mathsf{dist}(g,h_w) \leq \epsilon/8]}, \tag{3.4}$$

which is achieved by the Triangle Inequality and inspection of the rightmost part of Figure 3.1: the numerator is the "volume" of the outer circle and the denominator is the volume of the circle (of radius $\epsilon/8$) centered at g. Now, bounding the first and final terms of this inequality will give us the property.

For both bounds, we need bounds on the volume of functions within distance r of a given target; more formally, using the notation $h_w(x) = \mathrm{sign}(\langle w, x\rangle)$ for any unit vector w, given an arbitrary halfspace and "radius" r, we want bounds on

$$\mathbb{P}_w[\mathsf{dist}(g,h_w) \leq r],$$

which we will achieve by bounding the volumes over- and under-approximating cones (see Figure 3.1). As before (equation (3.3)), we have the upper-bound

$$\mathbb{P}_w[\mathsf{dist}(h_w,h_u) \leq r] = \mathbb{P}_w[\langle w,u\rangle \geq \cos(r)] \leq \frac{\sqrt{3}}{\sqrt{2\pi n}}(2r)^{n-1}$$

For the lower-bound we use

$$\begin{aligned}\frac{1}{V_n}\int_0^{\cos r} V_{n-1}(t\tan r)dt &= \frac{V_{n-1}}{V_n}(\tan r)^{n-1}\int_0^{\cos r} t^{n-1}dt \\ &\geq r^{n-1}\frac{V_{n-1}}{V_n}\left[\frac{t^n}{n}\right]_0^{\cos r} = r^{n-1}\frac{V_{n-1}}{V_n}\frac{\cos^n(r)}{n}\end{aligned}$$

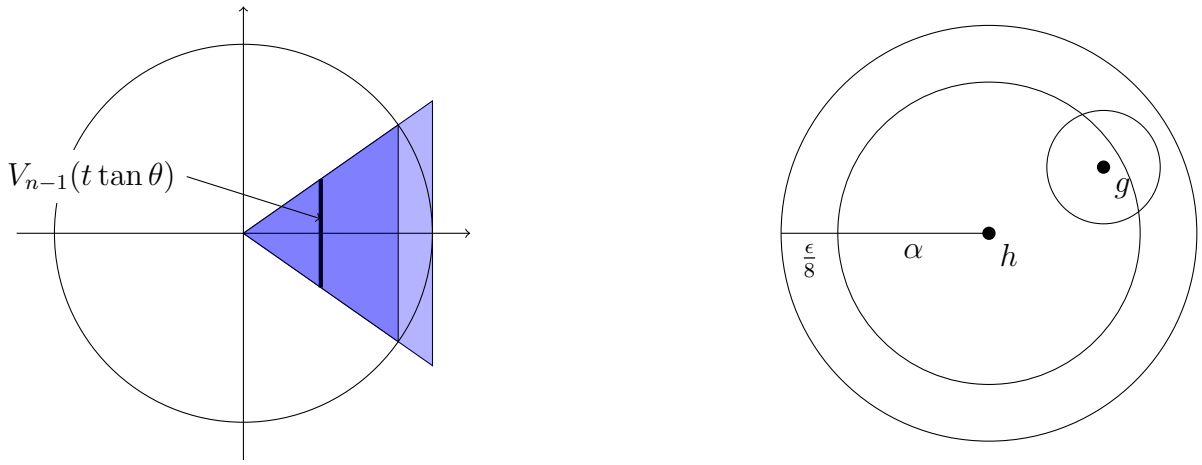


Figure 3.1: Left: estimates for the spherical cap. Right: bound on the number of functions.

For the inequality, we have used $\tan x \geq x$ for $x \in [0, 1)$. From Proposition 2.4.6 and Theorem 2.4.3 we have

$$\frac{V_{n-1}}{V_n} = \frac{\Gamma\left(\frac{n+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n+1}{2}\right)} \geq \frac{\sqrt{n}}{\sqrt{2\pi}}$$

and for $r \leq \frac{\pi}{6}, \cos(r) \geq \frac{\sqrt{3}}{2}$. Therefore the lower-bound becomes

$$\frac{V_{n-1}}{V_n} r^{n-1} \frac{\cos^n(r)}{n} \geq \frac{\sqrt{n}\sqrt{3}}{2n\sqrt{2\pi}} r^{n-1} \left(\frac{\sqrt{3}}{2}\right)^{n-1} = \frac{\sqrt{3}}{\sqrt{8\pi n}} \left(\frac{\sqrt{3}r}{2}\right)^{n-1} \tag{3.5}$$

Putting these bounds into inequality 3.4, we get

$$\#\left\{g \in G \colon \mathsf{dist}(g, h) \leq \alpha\right\} \leq 2\left(\frac{32(\alpha + \epsilon/8)}{\sqrt{3}\epsilon}\right)^{n-1} \leq 2\left(\frac{36\alpha}{\sqrt{3}\epsilon}\right)^{n-1}$$

since $\epsilon \leq \alpha$. □

Using this lemma, we can construct a learning algorithm that uses the optimal number of samples.

Algorithm 2 Sample-Efficient Learning Algorithm

Input: $Q_f = ((x_1, f(x_1)), \ldots, (x_m, f(x_m))), \epsilon \in (0,1)$

1: **function** $A(Q_f, \epsilon)$

2: Construct G as in Lemma 3.2.15.

3: Return $\operatorname{argmin}_{g \in G} \#\{(x_i, f(x_i)) \in Q_f : f(x_i) \neq g(x_i)\}$

Theorem 3.2.16. *Let A be the above algorithm, let f be any balanced halfspace, and let $\epsilon \in (0,1)$. Then for independently selected points Q drawn from the uniform distribution on the unit sphere in $\mathbb{R}^n$, where $m = |Q| = \Omega\left(\frac{n}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$,*

$$\mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}(A(Q_f), f) > \epsilon\right] \leq \delta$$

Proof. Let f be the input to the algorithm and denote by dist_Q the distance measure restricted to the samples Q:

$$\mathsf{dist}_Q(f,g) := \frac{1}{|Q|}\sum_{x \in Q} \mathbb{1}\left[f(x) \neq g(x)\right]$$

We want to bound the probability that the algorithm fails on f. Let $h \in G$ be the function, dependent on Q, that is produced by the algorithm, i.e. the function that minimizes $\mathsf{dist}_Q(f,h)$. We can overestimate the probability by considering the events (1) that a "bad" function $g \in G$ has small error on the sample, and (2) that *all* functions in G have a large error on the sample.

$$\mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}(f,h) > \epsilon\right] \leq \mathop{\mathbb{P}}_{Q}\left[\left(\exists g \in G : \mathsf{dist}(g,f) > \epsilon, \mathsf{dist}_Q(g,f) < \frac{\epsilon}{2}\right) \vee \left(\mathsf{dist}_Q(h,f) \geq \frac{\epsilon}{2}\right)\right]$$

The probability of event (2) is easy to bound; from the first property of G in Lemma 3.2.15, there is certainly some function $g \in G$ with $\mathsf{dist}(f,g) \leq \epsilon/4$, so this function must make twice as many errors on the sample as expected. The probability that this occurs is

$$\begin{aligned}\mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}_Q(h,f) \geq \frac{\epsilon}{2}\right] \leq \mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}_Q(g,f) \geq \frac{\epsilon}{2}\right] &\leq \mathop{\mathbb{P}}_{Q}\left[\frac{1}{m}\sum_{x \in Q}\mathbb{1}\left[g(x) \neq f(x)\right] \geq 2\mathsf{dist}(f,g)\right] \\ &\leq \exp\left(-\frac{m\mathsf{dist}(f,g)}{3}\right) \leq \exp\left(-\frac{m\epsilon}{12}\right),\end{aligned}$$

where the inequality on the second line is due to the multiplicative Chernoff bound (Lemma 2.3.7). Binding the probability of event (1) is more complicated; we will use the second

property of G to split the "far" functions into levels, count the number of functions in each level, and add up the probability that any one of them has small dist_Q distance.

For $1 \leq i \leq \lceil 1/\epsilon \rceil$, we will define the "$i^{\text{th}}$ level" to be $G_i := \{g \in G : i\epsilon < \mathsf{dist}(g,f) \leq (i+1)\epsilon\}$. For any $g \in G_i$, we therefore have $\epsilon/2 < \mathsf{dist}(g,f)/2i$, so by the multiplicative Chernoff bound (Theorem 2.3.7),

$$\begin{aligned}\mathop{\mathbb{P}}_{Q}\left[\mathsf{dist}_Q(g,f) < \frac{\epsilon}{2}\right] &\leq \mathop{\mathbb{P}}_{Q}\left[\frac{1}{m}\sum_{x\in Q}\mathbb{1}\left[g(x) \neq f(x)\right] < \frac{\mathsf{dist}(g,f)}{2i}\right] \leq \exp\left(-\frac{\left(1-\frac{1}{2i}\right)^2 m\mathsf{dist}(g,f)}{2}\right) \\ &\leq \exp\left(-\frac{\left(1-\frac{1}{2i}\right)^2 im\epsilon}{2}\right).\end{aligned}$$

Using the Union Bound and combining this expression with what we know about the number of functions in each level from Lemma 3.2.15, we have, for each $i > 0$,

$$\begin{aligned}\mathop{\mathbb{P}}_{Q}\left[\exists g \in G_i : \mathsf{dist}_Q(g,f) < \frac{\epsilon}{2}\right] &\leq \#\left\{g \in G : \mathsf{dist}(g,f) \leq (i+1)\epsilon\right\} \cdot \exp\left(-\frac{\left(1-\frac{1}{2i}\right)^2 im\epsilon}{2}\right) \\ &\leq (C(i+1))^{n-1}\exp\left(-\frac{\left(1-\frac{1}{2i}\right)^2 im\epsilon}{2}\right) \leq C^{n-1}\exp\left(in - i\frac{m\epsilon}{8}\right)\end{aligned}$$

where for the last inequality, we have used $(1+i)^{n-1} \leq e^{i(n-1)} \leq e^{in}$ and $1 - \frac{1}{2i} \geq 1/2$. Now we can add these all up to get

$$\begin{aligned}\mathop{\mathbb{P}}_{Q}\left[\exists i, g \in G_i : \mathsf{dist}_Q(g,f) \leq \frac{\epsilon}{2}\right] &\leq C^{n-1}\sum_{i=1}^{\lceil 1/\epsilon\rceil}\exp\left(i\left(n - \frac{m\epsilon}{8}\right)\right) \\ &= C^{n-1}\left(\frac{1-\exp\left(\left(\lceil 1/\epsilon\rceil+1\right)\left(n-\frac{m\epsilon}{8}\right)\right)}{1-\exp\left(n-\frac{m\epsilon}{8}\right)} - 1\right) \\ &= C^{n-1}\frac{1-\exp\left(\left(\lceil 1/\epsilon\rceil+1\right)\left(n-\frac{m\epsilon}{8}\right)\right) - 1 + 1 - \exp\left(n-\frac{m\epsilon}{8}\right)}{1-\exp\left(n-\frac{m\epsilon}{8}\right)} \\ &\leq C^n\frac{\exp\left(n-\frac{m\epsilon}{8}\right)}{1-\exp\left(n-\frac{m\epsilon}{8}\right)} = \frac{\exp\left((\ln(C)+1)n - \frac{m\epsilon}{8}\right)}{1-\exp\left(n-\frac{m\epsilon}{8}\right)}\end{aligned}$$

where in the final inequality we have used the fact that $C > 1$. Finally, substitute $m = \frac{8(\ln(C)+1)n}{\epsilon} + \frac{8\ln(2/\delta)}{\epsilon}$ to get

$$\frac{\exp\left(\ln(\delta/2)\right)}{1-\exp\left(\ln(\delta/2) - n\ln(C)\right)} = \frac{\delta}{2(1-\delta C^{-n})} \leq \delta$$

since $1 - \frac{\delta}{2}C^{-n} \geq 1/2$. □

Both the upper and lower bound presented here have since been extended to hold for all log-concave distributions as well. Balcan and Long [BL13] prove the following two theorems:

Theorem 3.2.17 ([BL13] Theorem 6)**.** *Let μ be a centered log-concave distribution over $\mathbb{R}^n$. There exists a constant C such that for all $d \geq 4, \delta > 0, 0 < \epsilon < 1/4$, any algorithm which outputs a halfspace that is correct on a sample of size $\frac{C}{\epsilon}(n + \log(1/\delta))$ will output a halfspace of error at most ϵ with probability at least $1 - \delta$.*

Theorem 3.2.18 ([BL13] Theorem 13)**.** *Let μ be a log-concave distribution whose covariance matrix has full rank. Then any algorithm that learns centered halfspaces (in the passive model) under μ requires a sample of size $\Omega\left(\frac{1}{\epsilon}(n + \log(1/\delta))\right)$.*

Since the uniform distributions over any convex set, the uniform n-sphere, and the Gaussian distribution are all log-concave, these results show that the general VC bounds on sample complexity hold not only for the distribution-free case, that is, the worst-case distributions, but also for the distributions we usually regard as "simple".

3.2.5 L_1 Polynomial Regression

The Perceptron and linear programming methods for learning halfspaces, as well as the sample complexity bounds from the previous subsection, work only in the most pristine situations, where the points are labelled by a halfspace and there is no corruption of the points. Modern methods of learning halfspaces try to deal with more realistic situations: as I referred to in the introduction, it is often the case that we merely try to "do the best we can" with the data we have [BFKV96]. This is the motivation for agnostic learning [KSS94], where the labels are not guaranteed to be consistent with *any* function. Instead, they are included in the input distribution, so rather than having a distribution D over $\mathbb{R}^n$, say, we have a distribution D over $\mathbb{R}^n \times \{\pm 1\}$. In the distribution-free case, it is NP-hard to solve this problem even approximately, so recent works focus on specific distributions. We will review one such paper by Kalai et al. [KKMS08], which used an L_1 polynomial regression algorithm (based on earlier low-degree Fourier algorithms on the hypercube, see [LMN93]).

First we will need the definition of agnostic learning (for the rest of this section I will use x^i to refer to the i^{th} point, while x_i is the i^{th} coordinate of point x):

Definition 3.2.19 (Agnostic Learning)**.** For a domain set X, let D be a distribution over $X \times \{\pm 1\}$, and let $\mathcal{H}$ be a class of functions $X \to \{\pm 1\}$. Since we are comparing a hypothesis function to a distribution, rather than a function, we will replace the usual dist metric with

$$\mathsf{err}(f) := \mathop{\mathbb{P}}_{x,\ell \sim D} [f(x) \neq \ell]$$

For a set of labelled samples $Q_\ell = \{(x^1, \ell_1), \ldots, (x^m, \ell_m)\}$ we define the empirical error

$$\mathsf{err}_{Q_\ell}(f) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1} \left[f(x^i) \neq \ell_i \right]$$

There may not be a function in $\mathcal{H}$ that always accurately predicts the labels, so we define

$$\mathsf{opt} = \min_{f \in \mathcal{H}} \ \mathsf{err}(f)$$

Then we say an algorithm A agnostically learns $\mathcal{H}$ with respect to D, with sample complexity m, accuracy $\epsilon > 0$, and confidence $1 - \delta$ if

$$\mathop{\mathbb{P}}_{Q_\ell \sim D^m} [\mathsf{err}(A(Q_\ell)) > \mathsf{opt} + \epsilon] < \delta$$

The L_1 polynomial regression algorithm of Kalai *et al.* [KKMS08] is a simple algorithm that uses linear programming to find a low-degree polynomial that minimizes the L_1 error on the random sample, and then optimizing the threshold. The algorithm will produce a hypothesis with additive error at most ϵ as long as there is some low-degree polynomial with expected L_2 error at most ϵ^2. Therefore there are two parts to the proof: first, we must show that the algorithm is correct under the assumption of such a polynomial, and second, we must show that there is a polynomial that approximates halfspaces. The L_1 polynomial regression algorithm is shown in algorithm 3. The first step can be expressed as a linear program: we can map a point x to a vector of monomials of degree at most d: $\hat{x} = (1, x_1, \ldots, x_n, x_1^2, x_1 x_2, \ldots)$. Therefore, for a vector of coefficients $\hat{p}$, we can define the polynomial $p(x) = \langle \hat{p}, \hat{x} \rangle$. So we must only solve for these coefficients $\hat{p}$:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} u_i \\ \text{subject to:} \ \forall i \in [m], u_i & \geq \ell_i - \left\langle \hat{p}, \hat{x}^i \right\rangle \\ u_i & \geq -\left(\ell_i - \left\langle \hat{p}, \hat{x}^i \right\rangle \right) . \end{aligned}$$

Algorithm 3 L_1 Polynomial Regression

Input: $Q_\ell = ((x^1, \ell_1), \ldots, (x^m, \ell_m) \sim D^m, d > 0$

1: **function** $A(Q_f, \epsilon)$

2: Find a polynomial p with $\deg(p) < d$ that minimizes

$$\frac{1}{m} \sum_{i=1}^{m} \left| p(x^i) - \ell_i \right|$$

3: Find $t \in [-1, 1]$ minimizing

$$\frac{1}{m} \sum_{i=1}^{m} \left| \operatorname{sign}(p(x^i) - t) - \ell_i \right|$$

return $h(x) = \operatorname{sign}(p(x) - t)$

Theorem 3.2.20 ([KKMS08], Theorem 5)**.** *Let $\mathcal{H}$ be a class of functions $\mathbb{R}^n \to \{\pm 1\}$ and let D be a distribution over $\mathbb{R}^n \times \{\pm 1\}$ with marginal distribution D_X over $\mathbb{R}^n$. Suppose that for all $\epsilon > 0$ there exists d such that, for all functions $f \in \mathcal{H}$, there is some polynomial p with $\deg(p) < d$ that satisfies*

$$\mathop{\mathbb{E}}_{x,\ell \sim D} \left[(p(x) - f(x))^2 \right] \leq \epsilon^2 .$$

Then, using $O\left(n^d/\epsilon^3\right)$ labelled samples, the L_1 polynomial regression algorithm produces a hypothesis h (not necessarily in $\mathcal{H}$) such that

$$\mathop{\mathbb{E}}_{Q_\ell \sim D^m} \left[\mathsf{err}(A(Q_\ell, d)) \right] \leq \mathsf{opt} + \epsilon$$

Repeating the algorithm $r = O\left(\frac{\log(1/\delta)}{\epsilon}\right)$ times to produce $h_1, \ldots, h_r$ and then choosing h_i to minimize $\mathsf{err}_T(h_i)$ on an independent set of samples T of size $\tilde{O}\left(\frac{1}{\epsilon}\log(1/\delta)\right)$ produces a hypothesis h such that

$$\mathbb{P}\left[\mathsf{err}(h) > \mathsf{opt} + \epsilon \right] < \delta$$

Proof. Let $f \in \mathcal{H}$ be a function achieving opt and let p^* be the polynomial, assumed to exist, such that

$$\mathop{\mathbb{E}}_{x,\ell \sim D} \left[(p^*(x) - f(x))^2) \leq \epsilon^2 \right]$$

Then

$$\mathop{\mathbb{E}}_{x,\ell}\left[|p^*(x) - f(x)|\right] \leq \sqrt{\mathop{\mathbb{E}}_{x,\ell}\left[(p^*(x) - f(x))^2\right]} \leq \epsilon$$

Let p be the polynomial produced by step 1 of the algorithm. Then

$$\frac{1}{m}\sum_{i\in[m]}\left|p(x^i) - \ell_i\right| \leq \frac{1}{m}\sum_{i\in[m]}\left|p^*(x^i) - \ell_i\right| \leq \frac{1}{m}\sum_{i\in[m]}\left|f(x^i) - \ell_i\right| + \left|p^*(x^i) - f(x^i)\right|$$

where the first inequality is by definition of p and the last inequality is the triangle inequality. We want to show that the expectation is at most $\mathsf{opt} + \epsilon$ so we take the expectation of this inequality:

$$\begin{aligned}\mathop{\mathbb{E}}_{Q_\ell}\left[\frac{1}{m}\sum_{i\in[m]}\left|p(x^i) - \ell_i\right|\right] &\leq \mathop{\mathbb{E}}_{Q_\ell}\left[\frac{1}{m}\sum_{i\in[m]}\left|f(x^i) - \ell_i\right| + \left|p^*(x^i) - f(x^i)\right|\right] \\ &= \mathop{\mathbb{E}}_{x,\ell}\left[|f(x) - \ell|\right] + \mathop{\mathbb{E}}_{x,\ell}\left[|p^*(x) - f(x)|\right] \\ &\leq 2\mathsf{opt} + \epsilon\end{aligned}$$

Finally, we examine step 2 of the algorithm. Assuming that we have chosen $t \sim [-1, 1]$ uniformly at random, we have

$$\mathop{\mathbb{E}}_{t,x,\ell}\left[\mathbb{1}\left[\text{sign}(p(x) - t) \neq \ell\right]\right] = \mathop{\mathbb{P}}_{t,x,\ell}\left[\text{sign}(p(x) - t) \neq \ell\right]$$

This event can only occur if t is between $p(x)$ and ℓ, i.e. it occurs with probability $|p(x) - \ell|/2$. So

$$\mathop{\mathbb{E}}_{t,x,\ell}\left[\mathbb{1}\left[\text{sign}(p(x) - t) \neq \ell\right]\right] = \frac{1}{2}\mathop{\mathbb{E}}_{x,\ell}\left[|p(x) - \ell|\right]$$

giving us, after optimizing t,

$$\mathop{\mathbb{E}}_{Q_\ell}\left[\mathsf{err}_{Q_\ell}(A(Q_\ell, d))\right] \leq \mathsf{opt} + \epsilon/2$$

To conclude the first part of the theorem, we must show that the difference between the empirical error and true error is at most $\epsilon/2$ (we will actually aim for $\epsilon/4$, which will be important for the second part). We do this by an appeal to VC theory:

Theorem 3.2.21 ([Vap92])**.** *Let $\mathcal{H}$ be a class of functions $X \to \{\pm 1\}$ with VC dimension d, and let D be a distribution over $X \times \{\pm 1\}$. Let $Q_\ell \sim D^m$ be a set of independent samples from X. Then for all $\epsilon > 0$,*

$$\mathop{\mathbb{P}}_{Q_\ell}\left[\sup_{h\in\mathcal{H}}|\mathsf{err}_{Q_\ell}(h) - \mathsf{err}(h)| > \epsilon\right] \leq \left(\frac{2me}{d}\right)^d e^{-m\epsilon^2}$$

Denote by Δ the difference $|\mathsf{err}_{Q_\ell}(h) - \mathsf{err}(h)|$. Then we want

$$\begin{aligned}\mathop{\mathbb{E}}_{Q_\ell}[\Delta] &= \mathbb{P}[\Delta \le \epsilon/8]\,\mathbb{E}[\Delta \;\mid\; \Delta \le \epsilon/8] + \mathbb{P}[\Delta > \epsilon/8]\,\mathbb{E}[\Delta \;\mid\; \Delta > \epsilon/8] \\ &\le \frac{\epsilon}{8} + \mathbb{P}[\Delta > \epsilon/8] < \frac{\epsilon}{4}\end{aligned}$$

so we want to show that the latter probability is at most $\epsilon/8$. We can accomplish this by choosing an appropriate constant C and setting

$$m = \frac{Ch}{\epsilon^3} \ge \frac{16h}{\epsilon^2} 3\ln\left(\frac{(2Ce)^{1/3}}{\epsilon}\right)\ln\left(\frac{8}{\epsilon}\right) = \frac{16h}{\epsilon^2}\ln\left(\frac{2Ce}{\epsilon^3}\right)\ln\left(\frac{8}{\epsilon}\right) \ge \frac{16h}{\epsilon^2}\ln\left(\frac{2me}{h}\right)\ln\left(\frac{8}{\epsilon}\right)$$

where the first inequality holds when $C \ge 3 \cdot 16 \cdot (2Ce)^{1/6}\sqrt{8} \iff C^6 \ge (3\cdot 16)^6 2e8^3$, because $\sqrt{x} \ge \ln(x)$. Applying the theorem with m and $\epsilon/8$ gives us

$$\mathbb{P}[\Delta > \epsilon/4] < e^{-\ln(8/\epsilon)} = \frac{\epsilon}{8}$$

as desired. To show that $m = \mathrm{poly}(n^d/\epsilon)$ we must show that the VC dimension is at most n^d; since we are producing a degree-d polynomial over n dimensions, this is the case.

What remains to be shown is the "boosting" step from a good *expected* error to a good error with high probability. To accomplish this, we will run the above algorithm $r = O(\ln(1/\delta)/\epsilon)$ times to get hypotheses $h_1, \ldots, h_r$. Then we will use a fresh, independent set R of $O(\ln(1/\delta)/\epsilon^2)$ samples and take the final hypothesis h to be $\mathrm{argmin}_{i\in[r]}\,\mathsf{err}_R(h_i)$. By Markov's inequality, the probability that run i satisfies $\mathsf{err}(h_i) > \mathsf{opt} + \frac{7}{8}\epsilon$ is

$$\mathbb{P}\left[\forall i \in [r], \mathsf{err}(h_i) > \mathsf{opt} + \frac{7}{8}\epsilon\right] \le \left(\frac{\mathsf{opt} + (3/4)\epsilon}{\mathsf{opt} + (7/8)\epsilon}\right)^r \le \left(1 - \frac{\epsilon}{16}\right)^r \le \frac{\delta}{2}$$

for $r = \frac{16}{\epsilon}\ln(2/\delta)$. Assume that this good event occurs, so there is some hypothesis, say h_1, such that $\mathsf{err}(h_1) \le \mathsf{opt} + (7/8)\epsilon$. Then the probability that we pick a bad hypothesis is

$$\begin{aligned}\mathop{\mathbb{P}}_R[\mathsf{err}(h) > \mathsf{opt} + \epsilon] &\le \mathop{\mathbb{P}}_R[\mathsf{err}_R(h_1) > \mathsf{opt} + (15/16)\epsilon \wedge \exists i > 1 : \mathsf{err}_R(h_i) < \mathsf{opt} + (15/16)\epsilon] \\ &\le \mathop{\mathbb{P}}_R[\exists i \in [r] : |\mathsf{err}(h_i) - \mathsf{err}_R(h_i)| > \epsilon/16] \le re^{-2|R|^2\epsilon^2/(16)^2}\end{aligned}$$

where the last inequality is the union bound and Hoeffding's inequality. Thus we need $|R| = \frac{8}{\epsilon}\ln(2r/\delta) = \tilde{O}\left(\frac{1}{\epsilon}\ln(1/\delta)\right)$. □

Now all that remains is to show that for the concept classes and distributions we are interested in, a polynomial that approximates halfspaces exist. For the uniform distribution over $\{\pm 1\}^n$, this is known. We can simply take the Fourier polynomial:

Fact 3.2.22 ([LMN93])**.** *Let $\epsilon > 0$ and f be a halfspace. Then for $d = \frac{441}{\epsilon^2}$, the polynomial*

$$p = \sum_{S \subset [n] : |S| < d} \hat{f}(S) \chi_S$$

satisfies $\mathbb{E}\left[(p(x) - f(x))^2\right] \leq \epsilon^2$.

For the uniform n-sphere, we need to do some more work. We will assume the following theorem, which relies on a construction of Hermite polynomials (for now it is not necessary to define Hermite polynomials; I will define them in Subsection 3.3.1):

Theorem 3.2.23 ([KKMS08], Theorem 6)**.** *Let $d > 0$ and $\theta \in \mathbb{R}$. Then there exists a polynomial p with $\deg(p) < d$ such that*

$$\int_{-\infty}^{\infty} (p(x) - \text{sign}(x - \theta))^2 \frac{e^{-x^2}}{\sqrt{\pi}} dx \leq O\left(\frac{1}{\sqrt{d}}\right)$$

Theorem 3.2.24. *Let $\mathcal{H}$ be the set of halfspaces over the uniform n-sphere and let $d = O\left(1/\epsilon^4\right)$. For all functions $f \in \mathcal{H}$ there exists a polynomial p of degree at most d such that*

$$\mathbb{E}\left[(p(x) - f(x))^2\right] \leq \epsilon^2$$

Proof. Let w, θ be such that $f(x) = \text{sign}(\langle w, x \rangle - \theta)$, where $\|w\| = 1$. Suppose we set $p(x) = \hat{p}(\langle w, x \rangle)$ for some polynomial $\hat{p}$ to be defined later. Projecting the probability distribution onto the line w, we have

$$\begin{aligned}
\mathbb{E}\left[(p(x) - f(x))^2\right] &= \int_{-1}^{1} (\hat{p}(z) - \text{sign}(z - \theta))^2 \frac{S_{n-2}(\sqrt{1 - z^2})}{S_{n-1}} dz \\
&= \frac{S_{n-2}}{S_{n-1}} \int_{-1}^{1} (1 - z^2)^{(n-2)/2} (\hat{p}(z) - \text{sign}(z - \theta))^2 dz \\
&\leq \frac{S_{n-2}}{S_{n-1}} \int_{-\infty}^{\infty} e^{-z^2 (n-2)/2} (\hat{p}(z) - \text{sign}(z - \theta))^2 dz \,.
\end{aligned}$$

Substituting $t = z\frac{\sqrt{n-2}}{2}, dt = \frac{\sqrt{n-2}}{2} dz$:

$$\begin{aligned}
&= \frac{S_{n-2}}{S_{n-1}} \frac{2\sqrt{\pi}}{\sqrt{n-2}} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{\sqrt{\pi}} (\hat{p}(\frac{2t}{\sqrt{n-2}}) - \text{sign}(\frac{2t}{\sqrt{n-2}} - \theta))^2 dt \\
&\leq 2 \int_{-\infty}^{\infty} \frac{e^{-t^2}}{\sqrt{\pi}} (\hat{p}(\frac{2t}{\sqrt{n-2}}) - \text{sign}(t - \frac{\sqrt{n-2}}{2}\theta))^2 dt
\end{aligned}$$

where in the last inequality, we have multiplied the argument to sign by $\sqrt{n-2}/2$ which doesn't change the value, and used Proposition 2.4.7. Therefore, if we let q be the polynomial from the previous theorem, where we use $\theta \cdot \sqrt{n-2}/2$ as the threshold, and define $\hat{p}(x) = q\left(\frac{\sqrt{n-2}x}{2}\right)$, we can bound this by $O\left(1/\sqrt{d}\right) = \epsilon^2$ for the appropriate choice of constant in d. □

Combining this theorem with the L_1 polynomial regression algorithm, we see that the number of random samples required to agnostically learn halfspaces over the uniform sphere is at most

$$\frac{n^{O\left(1/\epsilon^4\right)}}{\epsilon^4}\log(1/\delta)$$

3.3 Testing Halfspaces

Now that we have seen some of the learning theory to give some context for the testing problem, I will review a few of the recent results on testing halfspaces. There are very few results on this problem; see Table 3.2 for a summary. The most important work done on this problem is the work of Matulef *et al.* [MORS10], which presents a testing algorithm for the Gaussian and uniform hypercube distributions. This algorithm uses queries, rather than samples, but some of the ideas are similar to what I will present in Chapter 4. In this section I will review the algorithm for the Gaussian space (the "MORS algorithm"), and briefly touch on more recent work of Balcan *et al.* [BBBY12] that translates the MORS Algorithm into the passive model. The MORS algorithm for the hypercube is a very long, complex algorithm that requires advanced Fourier techniques; I will review some of these techniques in Chapter 7.

3.3.1 The MORS Algorithm

The MORS algorithm for testing halfspaces over the Gaussian distribution is an elegant solution powered by a pair of elementary observations. First, there is a function U that, given the degree-0 Fourier coefficient $\mathbb{E}[f]$, tells us exactly what the 2-norm of the first-degree Fourier coefficients $\sum_{i\in[n]}\mathbb{E}[x_i f(x)]^2$ would be if f were a halfspace. Second, if f is *far* from being a halfspace, then $\sum_{i\in[n]}\mathbb{E}[x_i f(x)]^2$ is much smaller than the output of $U(\mathbb{E}[f])$.

From these observations, a simple algorithm can be extracted: first, estimate the volume $\mathbb{E}[f]$ and feed it to the function U. Then, estimate $\sum_{i\in[n]}\mathbb{E}[x_i f(x)]^2$ and compare it to

Model	Distribution	Lower	Upper
passive	Gaussian	$\Omega_\epsilon\left(\left(\frac{n}{\log n}\right)^{1/2}\right)$ [BBBY12]	$O_\epsilon\left(\sqrt{n \log n}\right)$ [BBBY12]
active	Gaussian	$\Omega_\epsilon\left(\left(\frac{n}{\log n}\right)^{1/3}\right)$ [BBBY12]	$O_\epsilon\left(\sqrt{n \log n}\right)$ [BBBY12]
queries	Gaussian	—	poly $\left(\frac{1}{\epsilon}\right)$ [MORS10]
queries	Hypercube	—	poly $\left(\frac{1}{\epsilon}\right)$ [MORS10]
queries	Dist. free	$\Omega_\epsilon\left(\left(\frac{n}{\log n}\right)^{1/5}\right)$ [GS07]	—

Table 3.2: A summary of what we know about testing halfspaces.

the output of U. If it is too large, reject, otherwise accept. I will first prove these two observations, and then show the query-based method that the authors use to estimate the 2-norm of first-degree Fourier coefficients. Combining these results will give us the correctness of algorithm 4.

Over the Gaussian space, the Fourier transform has a different meaning than for the boolean hypercube (which is explained in the preliminaries). The concept is similar, but we must select a different basis (the parity functions don't work). We will use the Hermite polynomials for this purpose. In fact, we will only explicitly require the degree-0 and degree-1 Hermite polynomials.

Definition 3.3.1 (Hermite Polynomials (See [O'D14] §11.2.))**.** Let ϕ be the standard Gaussian distribution over $\mathbb{R}^n$. For $k \in \mathbb{N}$ and $z \in \mathbb{R}$, write

$$H_k(z) := \frac{(-1)^k}{\sqrt{k!}\phi(z)} \frac{d^k}{dz^k}\phi(z)\,.$$

For a multi-index $S \in \mathbb{N}^n$ write, for each $x \in \mathbb{R}^n$,

$$H_S(x) := \prod_{i \in [n]} H_{S_i}(x_i)$$

For any distinct multi-indices $S, T \in \mathbb{N}^n$, H_S and H_T satisfy

$$\langle H_S, H_T \rangle := \mathop{\mathbb{E}}_{x \sim \gamma}\left[H_S(x) H_T(x)\right] = 0\,,$$

so the polynomials are orthogonal (in fact, orthonormal). The degree-0 polynomial is

$$H_{\vec{0}}(x) = \prod_{i \in [n]} \frac{\phi(x_i)}{\phi(x_i)} = 1$$

and for any $i \in [n]$,

$$H_{e_i} = -\frac{1}{\phi(x_i)} \cdot \frac{d}{dx_i}\phi(x_i) = -e^{-x_i^2/2} \cdot e^{-x_i^2/2}(-x_i) = x_i\,.$$

For any function $f : \mathbb{R}^n \to \mathbb{R}$, we define the Fourier coefficients as

$$\hat{f}(S) := \langle f, H_S \rangle = \mathop{\mathbb{E}}_{x\sim\gamma}[f(x)H_S(x)]$$

for each $S \in \mathbb{N}^n$, so we can write the function f as

$$f = \sum_{S\in\mathbb{N}^n} \hat{f}(S)H_S$$

Similar to the Fourier coefficients over the hypercube, we will write $\hat{f}(e_i) = \hat{f}(i)$.

The central idea of this paper is to relate the first-degree Fourier coefficients to the Gaussian isoperimetric function:

Definition 3.3.2 (Gaussian Isoperimetric Function)**.** Let $v \in (0, 1)$ and write $\phi(x)$ as the density of the Gaussian function. Let $\Phi(t) = \int_{-\infty}^{t} \phi(x)dx$ be the cumulative distribution function of the Gaussian with inverse Φ^{-1}. The Gaussian isoperimetric function is defined as

$$I(v) := \phi(\Phi^{-1}(v))$$

Matulef *et al.* define the following function obtained from the isoperimetric function [MORS10]:

$$U(v) := (2I(v))^2 = (2\phi(\Phi^{-1}(v)))^2$$

A remarkable and perhaps surprising fact about this function is that it tells us exactly how large the first-degree coefficients ought to be if f is a halfspace. We can prove this using the rotation invariance of the Gaussian distribution; I will talk more about rotation invariance in the next chapter.

Theorem 3.3.3 ([MORS10], Proposition 25)**.** *Let $f : \mathbb{R}^n \to \{\pm 1\}$ be the halfspace $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ for some $w \in \mathbb{R}^n, \theta \in \mathbb{R}$ where $\|w\| = 1$. Let ϕ be the standard n-dimensional Gaussian distribution. Let*

$$c = \phi(f^+)\mathbb{E}[x \mid f(x) = 1] + \phi(f^-)\mathbb{E}[x \mid f(x) = -1]$$

be the center of mass, and let $v = \phi(f^+) - \phi(f^-) = \mathbb{E}[f]$. *Then*

$$c = \sqrt{U(v)}w\,.$$

Proof. Let ϕ be the density of the one-dimensional standard Gaussian distribution and Φ be its cumulative distribution function. We know that c, w are parallel (Proposition 4.2.2) so $c = \|c\| w$. By rotation invariance it suffices to show $\|c\| = \sqrt{U(v)}$ for $w = e_1$. In this case we have

$$c_1 = \mathbb{E}[x_1 \operatorname{sign}(\langle w, x\rangle - \theta)] = \mathbb{E}[x_1 \operatorname{sign} x_1 - \theta]$$

and $c_i = 0$ for all $i > 1$. Since the n–dimensional Gaussian distribution is a product distribution, we may replace x_1 with a 1-dimensional Gaussian $z \sim \mathcal{N}(0,1)$:

$$\begin{aligned}\|c\| = |c_1| = \mathbb{E}[z \operatorname{sign}(z-\theta)] &= 2\int_\theta^\infty z\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz \\ &= \frac{2}{\sqrt{2\pi}}\int_{\theta^2/2} e^{-t}dt \qquad (t = z^2/2, dt = zdz) \\ &= \frac{2}{\sqrt{2\pi}}e^{-\theta^2/2} = 2\phi(\theta) = 2\phi(\Phi^{-1}(v))\end{aligned}$$

In the first equality we have used the fact that $0 = \int_{-\infty}^{\theta} y\phi(dy) + \int_\theta^\infty y\phi(dy)$. □

Even better, we know that if f is not a halfspace, then this relationsip cannot hold, and how far away the center of mass is from satisfying the equality will tell us how far away the function is from being a halfspace.

Theorem 3.3.4 ([MORS10], Theorem 26)**.** *Let* $f : \mathbb{R}^n \to \{\pm 1\}$ *satisfy* $|\mathbb{E}[f]| \le 1 - \epsilon$. *Then, writing* $\mathsf{dist}(f) := \min_h \mathsf{dist}(f,h)$ *where* h *ranges over all halfspaces,*

$$\frac{\mathsf{dist}(f)^2}{2} \le \sqrt{U(v)} - \|c\| \le \frac{\sqrt{\pi}}{\sqrt{2}}\frac{\left|U(v) - \|c\|^2\right|}{\epsilon}$$

Proof. Let

$$c = \phi(f^+)\mathbb{E}\left[x \;\middle|\; x \in f^+\right] - \phi(f^-)\mathbb{E}\left[x \;\middle|\; x \in f^-\right]\,.$$

be the center of mass. Let $h(x) := \frac{1}{\|c\|}\sum_{i\in[n]}\hat{f}(i)x_i - t$ where t is the threshold such that $\mathbb{E}[h] = \mathbb{E}[f]$. That is, h is the linear function with normal vector $c/\|c\|$ and threshold t. We will first show that

$$\mathbb{E}[h(\operatorname{sign}(h) - f)] = \sqrt{U(v)} - \|c\| \tag{3.6}$$

We have

$$\mathbb{E}\left[h(\operatorname{sign}(h)-f)\right]=\mathbb{E}\left[|h|\right]-\mathbb{E}\left[hf\right]$$

and we will look at each of these terms individually. The first term is

$$\begin{aligned}\mathbb{E}\left[|h|\right]&=\mathop{\mathbb{E}}_{z\sim\gamma^{(1)}}\left[|z-t|\right]\qquad\text{(rotational invariance)}\\&=\mathbb{E}\left[(z-t)\operatorname{sign}(z-t)\right]=\mathbb{E}\left[z\operatorname{sign}(z-t)\right]-t\mathbb{E}\left[\operatorname{sign}(z-t)\right]\\&=2\phi(t)-tv\end{aligned}$$

To compute the second term, note that for $S\in\mathbb{N}^n$ with $|S|>1$ we have

$$\hat{h}(S)=\langle h,H_S\rangle=\frac{1}{\|c\|}\sum_{i\in[n]}\hat{f}(i)\left\langle x_i,H_S(x)\right\rangle-t\left\langle 1,H_S\right\rangle=0$$

since $x_i=H_i$ and H_i,H_S are orthonormal. And for $|S|=1$ we have

$$\hat{h}(i)=\langle h,x_i\rangle=\frac{1}{\|c\|}\sum_{j\in[n]}\hat{f}(j)\mathbb{E}\left[x_jx_i\right]-t\left\langle 1,H_S\right\rangle=\frac{1}{\|c\|}\hat{f}(i)$$

Finally,

$$\hat{h}(0)=\langle h,1\rangle=\frac{1}{\|c\|}\sum_{i\in[n]}\hat{f}(i)\left\langle x_i,1\right\rangle-t\left\langle 1,1\right\rangle=-t$$

Using these facts,

$$\begin{aligned}\mathbb{E}\left[h(x)f(x)\right]&=\sum_{S\in\mathbb{N}^n}\hat{h}(S)\hat{f}(S)\qquad\text{(Plancherel's Identity)}\\&=\frac{1}{\|c\|}\sum_{i\in[n]}\hat{f}(i)^2+\hat{h}(0)\hat{f}(0)=\|c\|-tv\end{aligned}$$

which proves equation (3.6). Now, note that

$$h(x)(\operatorname{sign}(h(x))-f(x))=\begin{cases}0 & \text{if }\operatorname{sign}(h(x))=f(x)\\2\,|h(x)| & \text{if }\operatorname{sign}(h(x))\neq f(x)\end{cases}$$

so

$$\mathbb{E}\left[h(\operatorname{sign}(h)-f)\right]=2\mathbb{E}\left[|h(x)|\,\mathbb{1}\left[f(x)\neq\operatorname{sign}(h(x))\right]\right]$$

which is clearly minimized when f disagrees with $\operatorname{sign}h$ on those points x nearest t (where h is smallest). In this case, look at the interval of length $p=\mathbb{P}\left[f(x)\neq\operatorname{sign}(h(x))\right]$ centered

at t; since the density function satisfies $\phi(z) < 1/\sqrt{2\pi}$, the probability mass within this interval is at most $p/\sqrt{2\pi} < p/2$. Therefore there is at least $p/2$ mass outside this interval where $|h(x)| > p/2$ and $f(x) \neq \text{sign}(h(x))$. This shows that

$$\mathbb{E}\left[|h(x)|\,\mathbb{1}\left[f(x) \neq \text{sign}(h(x))\right]\right] > (p/2)^2$$

so combining this with equation (3.6) we have

$$\sqrt{U(v)} - \|c\| = \mathbb{E}\left[h(\text{sign}(h) - f)\right] > \frac{p^2}{2} \tag{3.7}$$

For the upper bound, we will use the inequality

$$(a-b)(a+b) = a^2 - b^2 \implies |a-b| = \frac{|a^2-b^2|}{a+b} \leq \frac{|a^2-b^2|}{b}$$

for positive a, b to get

$$\sqrt{U(v)} - \|c\| \leq \frac{\left|U(v)^2 - \|c\|^2\right|}{\sqrt{U(v)}}$$

Letting $\alpha = \Phi^{-1}(v)$ and recalling that $v \leq 1 - \epsilon$, we have

$$\epsilon \leq \mathbb{P}\left[|z| > \alpha\right] \leq e^{-\alpha^2/2} \qquad \text{(Lemma 2.3.12)}$$

from which, taking the log of both sides, we conclude that

$$\alpha^2 < 2\ln(1/\epsilon)$$

This gives us the bound since

$$\sqrt{U(v)} = 2\phi(\alpha) \geq \frac{2}{\sqrt{2\pi}} e^{\alpha^2/2} = \frac{\sqrt{2}}{\sqrt{\pi}}\epsilon\,. \qquad \square$$

The last ingredient that the algorithm needs is a way for estimating the volume $\mathbb{E}[f]$ and center of mass. The authors present a general way of estimating products of Fourier coefficients that has since seen further applications (e.g. [RS15]). Although only special cases of this theorem is required for now, I will present the general theorems since more general uses will be seen in Chapter 7.

To estimate the product, we will first need the following generalization of the well-known characterization of noise sensitivity (Proposition 2.5.8). This proof essentially copies the proof of that fact:

Lemma 3.3.5 ([MORS10], Lemma 14)**.** *Let $T \in \mathbb{N}^n$ and $\rho \in [-1, 1]$. Let $f_1, \ldots, f_p : \mathbb{R}^n \to \{\pm 1\}$ and let $x_1, \ldots, x_{p-1}$ be a set of independent random variables, and define y as a random variable chosen such that for $i \notin \operatorname{supp}(T), \mathbb{P}[y_i = 1] = 1/2$ and otherwise $\mathbb{P}[y_i = 1] = 1/2 + \rho/2$. Then*

$$\mathbb{E}[f_1(x_1) \cdots f_{p-1}(x_{p-1}) f_p(x_1 \odot \cdots x_{p-1} \odot y)] = \sum_{S \subseteq T} \rho^{|S|} \hat{f}_1(S) \ldots \hat{f}_p(S)$$

where $S \subseteq T$ if $S_i \leq T_i$ for all $i \in [n]$ and $\odot$ denotes coordinate-wise multiplication.

Proof. We first write each function in the Fourier basis (of Hermite polynomials):

$$\begin{aligned}
&\sum_{S_1, \ldots, S_p \in \mathbb{N}^n} \hat{f}_1(S_1) \cdots \hat{f}_p(S_p) \mathbb{E}\left[H_{S_1}(x_1) \cdots H_{S_{p-1}}(x_{p-1}) \cdot H_{S_p}(x_1 \odot \cdots x_{p-1} \odot y)\right] \\
&\quad = \sum_{S \subseteq T} \hat{f}_1(S) \cdots \hat{f}_p(S) \mathbb{E}\left[H_S(x_1) \cdots H_S(x_{p-1}) \cdot H_S(x_1 \odot \cdots x_{p-1} \odot y)\right] \\
&\quad = \sum_{S \subseteq T} \rho^{|S|} \hat{f}_1(S) \cdots \hat{f}_p(S)
\end{aligned}$$

where the first inequality is due to the orthogonality of the Hermite polynomials as well as the independence of each variable for coordinates $i \notin T$, and the second inequality is a property of the Hermite polynomials ([O'D14], §11.2). □

Theorem 3.3.6 ([MORS10], Lemma 15)**.** *Let $f_1, \ldots, f_p : \mathbb{R}^n \to \{\pm 1\}$ and let $\eta > 0, \delta > 0$. Then for any $T \subseteq [n]$ we can estimate*

$$\sum_{i \in T} \hat{f}_1(i) \cdots \hat{f}_p(i)$$

to within $\pm\eta$ with confidence $1 - \delta$ using $O\left(p \log(1/\delta)/\eta^4\right)$ queries.

Proof. First we empirically estimate

$$\mathbb{E}[f_1(x_1) \cdots f_p(x_p)] \qquad \text{and} \qquad \mathbb{E}[f_1(x_1) \cdots f_{p-1}(x_{p-1}) \cdot f_p(x_1 \odot \cdots \odot x_{p-1} \odot y)]$$

to within an additive η^2, where y is defined as in the previous lemma. Each quantity can be estimated using $O\left(\log(1/\delta)/\eta^4\right)$ examples by the standard Hoeffding bound, and each example requires p queries, for a total of $O\left(p \log(1/\delta)/\eta^4\right)$ queries. By the previous lemma, we can see that

$$\mathbb{E}[f_1(x_1) \cdots f_p(x_p)] = \hat{f}_1(0) \cdots f_p(0)$$

(by setting $T = [n], \rho = 0$) and

$$\mathbb{E}\left[f_1(x_1) \cdots f_{p-1}(x_{p-1}) \cdot f_p(x_1 \odot \cdots \odot x_{p-1} \odot y)\right] = \sum_{S \subseteq T} \eta^{|S|} \hat{f}_1(S) \cdots \hat{f}_p(S)$$

(setting $\rho = \eta$). Now we subtract the first from the second to get

$$\begin{aligned}
\sum_{S \subseteq T, |S|>0} \eta^{|S|} \hat{f}_1(S) \cdots \hat{f}_p(S) &= \sum_{i \in \mathrm{supp}(T)} \eta \hat{f}_1(i) \cdots \hat{f}_p(i) + \sum_{S \subseteq T, |S|>1} \eta^{|S|} \hat{f}_1(S) \cdots \hat{f}_p(S) \\
&\leq \sum_{i \in \mathrm{supp}(T)} \eta \hat{f}_1(i) \cdots \hat{f}_p(i) + \eta^2 \sum_{S \subseteq T, |S|>1} \hat{f}_1(S) \cdots \hat{f}_p(S) \\
&\leq \sum_{i \in \mathrm{supp}(T)} \eta \hat{f}_1(i) \cdots \hat{f}_p(i) + \eta^2
\end{aligned}$$

Since we have estimated this quantity to within $\pm 2\eta^2$, we can see that the total error is at most $3\eta^2$. Dividing by η gives us an estimate to within $\pm 3\eta$. Replacing η with $\eta/3$ would give us the theorem. □

With these theorems, we can now prove the correctness of the following algorithm:

Algorithm 4 MORS Algorithm for the Gaussian Distribution

Input: $\epsilon > 0$, oracle access to $f : \mathbb{R}^n \to \{\pm 1\}$

1: **function** $\mathrm{A}(\epsilon, f)$
2: Let $\tilde{v}$ be an empirical estimate of $\mathbb{E}[f]$ to within $\pm\epsilon^3$
3: Let $\tilde{\rho}$ be an estimate of $\|c\|^2 = \sum_{i \in [n]} \hat{f}(i)^2$ to within $\pm\epsilon^3$.
4: **if** $\rho - U(\tilde{v}) \leq 2\epsilon^3$ **then accept**

Theorem 3.3.7. *The above algorithm satisfies the following for all f and $\epsilon > 0$:*

1. *If f is a halfspace, A accepts with probability $\geq 2/3$,*
2. *If f is $C\epsilon$-far from all halfspaces (for some constant C), then A rejects with probability $\geq 2/3$,*
3. *A uses at most $O\left(\frac{1}{\epsilon^6}\right)$ queries.*

Proof. Let $\epsilon > 0$ and consider the case where f is a halfspace, so in particular $U(v) - \|c\|^2 = 0$. Assume that the estimation steps succeeded (which occurs with high probability). Then since $|\tilde{v} - v| \leq \epsilon^3$ and $\frac{d}{dv} U(v) < 1$ we have $|U(\tilde{v}) - U(v)| \leq \epsilon^3$, and we have also

$\left|\tilde{c}^2 - \|c\|^2\right| \leq \epsilon^3$ so the total error of $|\tilde{c} - U(\tilde{v})|$ is at most $2\epsilon^3$. Since this value ought to be 0, its estimate is at most $2\epsilon^3$ so the algorithm accepts.

Now suppose f is accepted by the algorithm. If $|v| > 1 - \epsilon$ then f is ϵ-close to a constant function, so assume this is not the case. Then by Theorem 3.3.4 we have

$$\mathsf{dist}(f)^2 \leq \sqrt{2\pi}\frac{|U(v) - \|c\||}{\epsilon} \leq \sqrt{2\pi}\frac{|U(\tilde{v}) - \|\tilde{c}\|| + 2\epsilon^3}{\epsilon} \leq 4\sqrt{2\pi}\epsilon^2$$

so $\mathsf{dist}(f) \leq 2(2\pi)^{1/4}\epsilon$.

Theorem 3.3.6 gives us the required guarantees on the query complexity. □

3.3.2 Active and Passive Testing [BBBY12]

To match the active learning model, Balcan *et al.* provide a model of active testing, in which the tester may make queries, but only from a randomly selected set of possible points [BBBY12]. While we are not concerned with this model for the current work, the paper also gives a testing algorithm for halfspaces over the Gaussian distribution that uses only passive queries and shows the first lower bound for this model.

In the previous subsection I presented the MORS algorithm of Matulef *et al.* [MORS10], which tested halfspaces over the Gaussian distribution using queries. An examination of that algorithm reveals that queries were used only to estimate the 2-norm of Fourier coefficients $\sum i \in [n]\hat{f}(i)^2$. If this estimation was replaced by a sampling estimation, we would get a sampling algorithm. This is the approach taken in [BBBY12]:

Proposition 3.3.8. *Let $f : \mathbb{R}^n \to \{\pm 1\}$. Then*

$$\mathop{\mathbb{E}}_{x,y}\left[f(x)f(y)\left\langle x, y\right\rangle\right] = \sum_{i\in[n]} \hat{f}(i)^2$$

where x, y are selected independently at random from the n-dimensional Gaussian distribution.

Proof. By linearity of expectation followed by independence of x and y,

$$\mathbb{E}\left[f(x)f(y)\left\langle x, y\right\rangle\right] = \sum_{i\in[n]} \mathbb{E}\left[f(x)f(y)x_i y_i\right] = \sum_{i\in[n]} \mathbb{E}\left[f(x)x_i\right] = \sum_{i\in[n]} \hat{f}(i)^2$$ □

Thus we can use $f(x)f(y)\langle x, y\rangle$ as an unbiased estimator. Note also that this estimator is symmetric in x, y; this is important because the analysis given by Balcan et al uses facts about U-statistics, a class of statistics which are symmetric in their arguments and therefore benefit from computing the statistic on every combination of samples. In this case, we could take m independent samples and get $m/2$ independent pairs, on which we compute the statistic and take the average. However, taking m samples we get $\binom{m}{2}$ pairs which are not independent but which are still sufficient. I will omit the U-statistics analysis since it is improved by the analysis of the algorithm in the next chapter. The other relevant result from this work is the lower bound on passive testing:

Theorem 3.3.9 (Balcan et al, Theorem 6.8 [BBBY12])**.** *Let A be a $\Theta(1)$-tester for halfspaces under the Gaussian distribution. Then*

1. *If A is a passive tester, then A requires a sample of size $\Omega\left(\left(\frac{n}{\log n}\right)^{1/2}\right)$.*
2. *If A is an active tester, then A requires $\Omega\left(\left(\frac{n}{\log n}\right)^{1/3}\right)$ labels.*

3.4 Summary

Efficiently learning halfspaces in a variety of models is still an active area of research, although classic VC dimension bounds show that a linear number of samples is required for PAC learning; I surveyed some works showing that this tight bound holds even for the easy case of the uniform sphere. From this we can conclude that a testing algorithm should aim to have a sublinear sample complexity.

I have briefly described a number of learning models and identified agnostic learning as the model which most closely matches our model of testing: it allows the target function to differ from a halfspace by a small amount, as would be guaranteed by a testing algorithm. For completeness, I surveyed a recent work that used L_1 polynomial regression to learn in this model.

Finally, I reviewed the two recent works on testing halfspaces: the first of these, [MORS10], relies on queries and uses only a constant number of labelled examples. It depends on a nice property of halfspace in the Gaussian space: the norm of their first-degree Fourier coefficients can be deduced exactly from their degree-0 coefficient. The algorithm of [BBBY12] uses this principle but replaces the queries with samples. In the next chapter, I will continue to develop this principle to design an algorithm that works for all rotationally invariant spaces, with the Gaussian space following as a special case.

Chapter 4

Testing Halfspaces in Rotation-Invariant Spaces

In this chapter I will present the first of my own contributions, a sampling algorithm for testing halfspaces over any *rotationally invariant* (RI) distribution. Rotation-invariant distributions are distributions that are the same in every direction. These are important spaces for two reasons: first, some common distributions such as the multivariate Gaussian distribution and the uniform distribution over the sphere are rotationally invariant; second, because these spaces have the special property that the norm of the center of mass of any halfspace depends only on its volume and not on its orientation. This means they are the simplest possible case for the algorithm that I will present, which relies heavily on this property.

The MORS algorithm from Chapter 3 ([MORS10]) relied on the elegant observation that the sum of squares of first-degree Fourier coefficients, $\sum_{i\in[n]} \hat{f}(i)^2$, satisfy a unique relationship to the volume $\mathbb{E}[f]$ when f is a halfspace, and that the farther a function f is from satisfying this relationship, the farther it is from a halfspace. For our goal of finding a tester for general distributions, we have to avoid Fourier techniques, since they may not generalize to all spaces. Thus, in this chapter, I will present a geometric interpretation of the first-degree Fourier coefficients as the *centers of mass*, and develop a Gap Theorem (Theorem 4.4.1) that will expand upon the observations from the previous chapter.

I present an algorithm which tests the halfspace property under arbitrary (known) rotation-invariant distribution that uses roughly $\sqrt{n}$ random queries in nice enough distributions. This should be compared with the known $\Omega\left(\sqrt{n/\log n}\right)$ lower bound of Balcan *et al.* for the Gaussian distribution (which is "nice enough", as we will see)[BBBY12]. The algorithm

estimates the Euclidean distance from the origin of the center of mass of the given function f and compares it to what its "proper" value should be, if the function were a halfspace. This approach works in the rotation-invariant setting because the "proper" distance is determined by the total number of +1-labelled points: regardless of the direction in which the halfspace points, this distance will be the same.

The basic idea behind the algorithm is simple and relies on finding a quantity (the center-norm) that satisfies a few requirements:

1. it is maximized when f is a halfspace;
2. it is robust, in the sense that functions close to halfspaces are close to having the maximum value (section 4.4);
3. for any function f, it can be computed from an easily-estimatable quantity (the volume $\mathbb{E}[f]$, Sections 4.5 and 4.6).

From here, the algorithm (Section 4.7) is easy: estimate the volume $\mathbb{E}[f]$ and use it to compute the center-norm; compare the center-norm to its maximum value and accept or reject accordingly.

At the end of the chapter, I will discuss some shortcomings of the algorithm which arise in spaces that are somehow unreasonable, namely, those whose "width" shrinks as the dimension increases. Width, essentially the maximal concentration of a distribution, is a definition I introduce in Section 4.3.

4.1 Centers of Mass

The main tool that I will use for dealing with halfspaces is the *center of mass*:

Definition 4.1.1 (Center of Mass)**.** We will be working a lot with the "centers of mass" of functions and set. For a probability measure μ and a measurable set $S \subset \mathbb{R}$ of points, we will define the "centers of mass" to be

$$\mathsf{Com}(S) := \mathop{\mathbb{E}}_{x\sim\mu}[x \mid x \in S] = \frac{1}{\mu(S)}\int_S x\mu(dx)\,.$$

Or, in other words, it is the average point in S. For a function $f : \mathbb{R}^n \to \{\pm 1\}$ we will use the following notation:

$$f^+ := \{x : f(x) = 1\} \qquad\qquad f^- := \{x : f(x) = -1\}$$

so that $\mathsf{Com}(f^+) = \mathsf{Com}(\{x : f(x) = 1\}), \mathsf{Com}(f^-) = \mathsf{Com}(\{x : f(x) = -1\})$. We will say

$$\mathsf{Com}(f) := \mathbb{E}[xf(x)] = \mu(f^+)\mathsf{Com}\left(f^+\right) - \mu(f^-)\mathsf{Com}\left(f^-\right) .$$

I will refer to the norm $\|\mathsf{Com}(f)\|_2$ as the *center-norm* of f.

Here is an oft-used property of centers of mass:

Proposition 4.1.2. *Let μ be any centered probability distribution over $\mathbb{R}^n$, i.e. $\mathop{\mathbb{E}}_{x\sim\mu}[x] = 0$. Then for any measurable subset $S \subseteq \mathbb{R}^n$ we have*

$$\mu(S)\mathbb{E}[x \mid x \in S] = -\mu(\bar{S})\mathbb{E}[x \mid x \notin S] .$$

Proof.

$$0 = \mathbb{E}[x] = \mu(S)\mathbb{E}[x \mid x \in S] + \mu(\bar{S})\mathbb{E}[x \mid x \notin S] . \qquad \square$$

The centers of mass have a few interesting connections with the Fourier spectrum: observe that for any boolean function, the center of mass is the vector of degree-1 Fourier coefficients:

Fact 4.1.3. *Let $f : \{\pm1\}^n \to \{\pm1\}$ be any boolean function and let $c = \mathsf{Com}(f)$ be its center of mass. Then for all $i \in [n]$,*

$$c_i = \langle e_i, \mathbb{E}[xf(x)]\rangle = \mathbb{E}[x_i f(x)] = \hat{f}(i) .$$

Thus $\|\mathsf{Com}(f)\|^2 = \sum_{i\in[n]} \hat{f}(i)^2$. And by Fact 2.5.11, we have

Fact 4.1.4. *Let $f : \{\pm1\}^n \to \{\pm1\}$ be any unate function (see Definition 2.5.10) with center $c = \mathsf{Com}(f)$. Then for all $i \in [n]$*

$$|c_i| = \left|\hat{f}(i)\right| = \mathrm{Inf}_i(f)$$

and as a consequence,

$$\mathrm{Inf}(f) = \|\mathsf{Com}(f)\|_1 \qquad \text{and} \qquad \max_i \mathrm{Inf}_i(f) = \|\mathsf{Com}(f)\|_\infty .$$

More advanced properties of the centers of mass will be explored in later chapters.

4.2 Rotation Invariance

Rotationally invariant (RI) spaces are those whose density function depends only on the Euclidean distance of a point from the center:

Definition 4.2.1 (Rotation-Invariant)**.** A probability measure μ on $\mathbb{R}^n$ is *rotation–invariant* (RI) if a random variable $x \sim \mu$ can be written as $x = rv$ where v is a unit vector drawn uniformly at random from the unit sphere and $r \sim \mu_R$ is a real number drawn independently from some distribution μ_R over $\mathbb{R}_{>0}$. That is, a distribution is RI if the orientation and 2–norm of a random vector are independent.

For example, the standard n-dimensional Gaussian distribution is rotation invariant: a point x has density $\frac{1}{\sqrt{(2\pi)^n}} e^{-\|x\|^2/2}$, which depends only on its 2–norm. We will construct an algorithm that, for any (known) RI space, will be able to test whether an unkown function is a halfspace or far from all halfspaces.

The reason RI spaces are a good place to start is that there is a crucial equivalence between centers of mass and weight vectors for halfspaces for these distributions:

Proposition 4.2.2. *Let μ be any RI distribution over $\mathbb{R}^n$ and let $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ be any halfspace. Then* $\mathsf{Com}(f)$ *is parallel to w.*

Proof. This is easy to see if μ is uniform over unit sphere: for any "ring" of points x such that $\langle w, x\rangle = t$ for some $-1 \le t \le 1$, we have $\mathop{\mathbb{E}}_x [x \mid \langle w, x\rangle = t] = tw$ so

$$\mathbb{E}[x \,\text{sign}(\langle w, x\rangle - \theta)] = \int_{-1}^{1} \text{sign}(t-\theta) \mathop{\mathbb{E}}_x [x \mid \langle w, x\rangle = t] \, \mu_w(t) = w \cdot \int_{-1}^{1} t \,\text{sign}(t-\theta) \mu_w(t)$$

The integral depends only on θ (see Proposition A.0.10 for the calculation of $\mu_w(t)$) so we can denote it by $I(\theta)$ and the expectation is $I(\theta) \cdot w$. Now for any RI distribution we have by definition

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mu}[x \,\text{sign}(\langle w, x\rangle - \theta)] &= \mathop{\mathbb{E}}_{r\sim\mu_R, v}[rv \,\text{sign}(\langle w, rv\rangle - \theta)] = \mathop{\mathbb{E}}_{r\sim\mu_R}\left[r \cdot \mathop{\mathbb{E}}_v [v \,\text{sign}(\langle w, v\rangle - \theta/r)]\right] \\ &= \mathop{\mathbb{E}}_{r\sim\mu_R}[r \cdot I(\theta/r) \cdot w] = \mathop{\mathbb{E}}_{r\sim\mu_R}[r \cdot I(\theta/r)] \cdot w \end{aligned}$$

□

An important observation is that the relationship used by the prior testing algorithms [BBBY12, MORS10] does not immediately hold in RI spaces. As a simple example, we can show that even a mixture of Gaussians with the same mean is no longer a Gaussian:

Example 4.2.3 (Mixtures of Gaussians aren't Gaussian)**.** Let

$$\phi_1(t) := \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{t^2}{2\sigma_1^2}} \qquad \phi_2(t) := \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{t^2}{2\sigma_2^2}}$$

be the densities of two Gaussians with mean 0 and variances σ_1, σ_2 respectively. We want to find σ such that for all $x \in \mathbb{R}$,

$$\phi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} = \lambda \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{t^2}{2\sigma_1^2}} + (1-\lambda) \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{t^2}{2\sigma_2^2}} .$$

Taking the natural log of both sides yields

$$-\frac{t^2}{2\sigma^2} - \ln(\sqrt{2\pi\sigma^2}) = -\frac{t^2}{2\sigma_1^2} - \frac{t^2}{2\sigma_2^2} - \ln(\sqrt{2\pi\sigma_1^2}) - \ln(\sqrt{2\pi\sigma_2^2}) + \ln(\lambda) + \ln(1-\lambda) .$$

Negating both sides and rearranging gives

$$\frac{t^2}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) = \frac{1}{2}\ln\left(\frac{4\pi^2\sigma_1^2\sigma_2^2}{2\pi\sigma^2}\right) - \ln(\lambda(1-\lambda)) .$$

Only the left side depends on t so this equality can only hold for all t if both sides are 0, i.e. $\sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2}$. Then, setting the right side to 0, we have

$$\frac{1}{2}\ln\left(2\pi\sigma_1^2\sigma_2^2(\sigma_1^{-2} + \sigma_2^{-2}\right) = \ln(\lambda(1-\lambda)) .$$

Thus we merely need to pick σ_1, σ_2 and λ appropriately to force a contradiction.

4.3 Width, Anticoncentration, and Margins

The main theorem that allows the tester to work relates the distance between two functions to the *width* of a 1-dimensional projection. To develop the definition of the width, we will introduce two other definitions that have appeared in recent work on halfspaces: the Lévy anticoncentration function and the margin. These two definitions will not have immediate application for RI probability distributions, but they will be important when we move beyond these simple spaces to the hypercube (see Chapter 7).

The Lévy anticoncentration function tells us the greatest probability mass of a ball of given radius r; in 1 dimension, this function is defined as follows:

Definition 4.3.1 (Lévy Anticoncentration Function [DS13])**.** Let $w \in \mathbb{R}^n$ be an arbitrary weight vector and let $r \in \mathbb{R}_+$ be some radius. Let μ be some probability distribution over $\mathbb{R}^n$. (In this chapter, we will assume μ is the uniform distribution over $\{\pm 1\}^n$.) The Lévy anticoncentration function of w at r is

$$p_r(w) := \sup_{\theta \in \mathbb{R}} \mathop{\mathbb{P}}_{x \sim \mu} [|\langle w, x \rangle - \theta| \leq r] \,.$$

This function is related to anticoncentration inequalities and the Littlewood–Offord problem, the significance of which will be described in more detail in Chapter 7. A related definition for halfspaces that appears, for example, in [OS11], is that of the *margin* of a halfspace:

Definition 4.3.2 (Margin)**.** Let $w \in \mathbb{R}^n$ be any vector satisfying $\|w\| = 1$ and let $\theta \in \mathbb{R}$. Let $r > 0$. The *r-margin* of the hyperplane defined by w, θ is the set of all points within distance r of the hyperplane:

$$\text{margin}_r(w, \theta) := \{x \in \mathbb{R}^n : |\langle w, x \rangle - \theta| \leq r\} \,.$$

These are easily seen to be related by the identity

$$p_r(w) = \sup_{\theta \in \mathbb{R}} \mu(\text{margin}_r(w, \theta)) \,.$$

Using these concepts, I will define the *ϵ-width* of a 1-dimensional projection of a distribution μ onto the vector w:

Definition 4.3.3 (Width)**.** Let μ be any probability distribution over $\mathbb{R}^n$, $w \in \mathbb{R}^n$ such that $\|w\| = 1$, and let $\epsilon \in (0, 1]$. We define the *ϵ-width* of the distribution to be

$$W_\mu(w, \epsilon) := \sup\{r \geq 0 : p_r(w) \leq \epsilon\} \,.$$

When μ is clear from context, we will drop the subscript μ.

As an example of width, consider the n-dimensional Gaussian distribution. Because of rotation invariance, the ϵ-width for any unit vector w will be the same:

Proposition 4.3.4. *Let $w \in \mathbb{R}^n$ such that $\|w\| = 1$ and let $0 < \epsilon < 1$. Suppose μ is that standard n-dimensional Gaussian distribution. Then*

$$W(w, \epsilon) \geq \sqrt{2\pi} \cdot \epsilon \,.$$

Proof. For any $\theta \in \mathbb{R}$ and radius r, we have

$$\mu(\text{margin}_r(w,\theta)) = \mu\{x : |\langle w, x\rangle - \theta| \leq r\} \leq \frac{1}{\sqrt{2\pi}} \cdot r$$

since μ_w is a 1-dimensional Gaussian. Thus for $r = \sqrt{2\pi}\epsilon$,

$$p_r(w) = \sup_\theta \mu(\text{margin}_r(w,\theta)) \leq \epsilon$$

so $W(w,\epsilon) \geq r = \sqrt{2\pi} \cdot \epsilon$. □

I will also define the variance of the 1-dimensional projection and the norm, which will be important quantities upon which the sample complexity will depend:

Definition 4.3.5. Let μ be any distribution on $\mathbb{R}^n$ and let w be any unit vector. Then we write

$$V(w) := \mathop{\mathbb{V}}_{x\sim\mu}[\langle w, x\rangle] = \mathop{\mathbb{E}}_{x\sim\mu}\left[\langle w, x\rangle^2\right],$$

and

$$V := \max_{w:\|w\|=1} V(w)$$

is the maximum variance over all 1-dimensional projections. For rotationally invariant distributions, we have $V = V(w)$ for all unit vectors w.

Any rotationally invariant distribution can be thought of as a mixture of uniform distributions over spheres of different radii, along with a distribution over radii. I will define R to be the variance of this distribution over radii, which corresponds to the expected norm squared for general distributions:

Definition 4.3.6. Let μ be any distribution on $\mathbb{R}^n$. We will write

$$R := \mathbb{E}\left[\|x\|^2\right].$$

The relationship between the width, variance, and expected norm of RI distributions is important. One interesting identity for RI distributions is the following:

Proposition 4.3.7. *Let μ be any rotationally invariant distribution on $\mathbb{R}^n$, and let V^* be the variance of the 1-dimensional projection of the unit sphere. Then*

$$V = RV^*.$$

Proof. For the sphere of radius r, the 1-dimensional variance is $V_r = r^2 V^*$ (Proposition A.0.8). Then the 1-dimensional variance of μ is

$$V = \mathop{\mathbb{E}}_{r}[V_r] = \mathop{\mathbb{E}}_{r}\left[r^2 V^*\right] = V^* \mathbb{E}[r] = V^* \mathbb{E}\left[\|x\|^2\right]$$

where r is drawn from the distribution over radii defined by μ. □

Since the sphere is the atomic building-block of RI distributions, I will calculate the width and variance and compare them to width and variance of the Gaussian distribution. The calculations can be found in Appendix A.

Example 4.3.8. The uniform sphere with radius r (respectively $\sqrt{n}$) satisfies the following:

1. $\mathbb{E}[\|x\|] = r$ (resp. $\sqrt{n}$);
2. $R = r^2$ (resp. n);
3. $V = r^2 \cdot \frac{2\pm1}{n+1}$ (resp. $\frac{n(2\pm1)}{n+1}$);
4. $W(\epsilon) = \Omega\left(\frac{\sqrt{n}\cdot\epsilon}{r}\right)$ (resp. $\Omega(\epsilon)$).

Example 4.3.9. The standard Gaussian distribution satisfies the following:

1. $\mathbb{E}[\|x\|] \approx \sqrt{n}$;
2. $R = n$;
3. $V = 1$;
4. $W(\epsilon) = \Omega(\epsilon)$.

4.4 The Gap Theorem

The following theorem is the most important ingredient of the algorithm; it relates the width, center-norm, and distance between two functions. It is phrased in such a way as to be usable in non-rotation-invariant spaces as well, which we will see in Chapter 5. The theorem itself was inspired by a result of Eldan ([Eld15], Corollary 4) although its proof differs significantly and it does not depend on the Gaussian distribution. Note also the similarity to Theorem 3.3.4; that theorem relied upon Fourier analysis and Hermite polynomials for its proof, which we are trying to avoid.

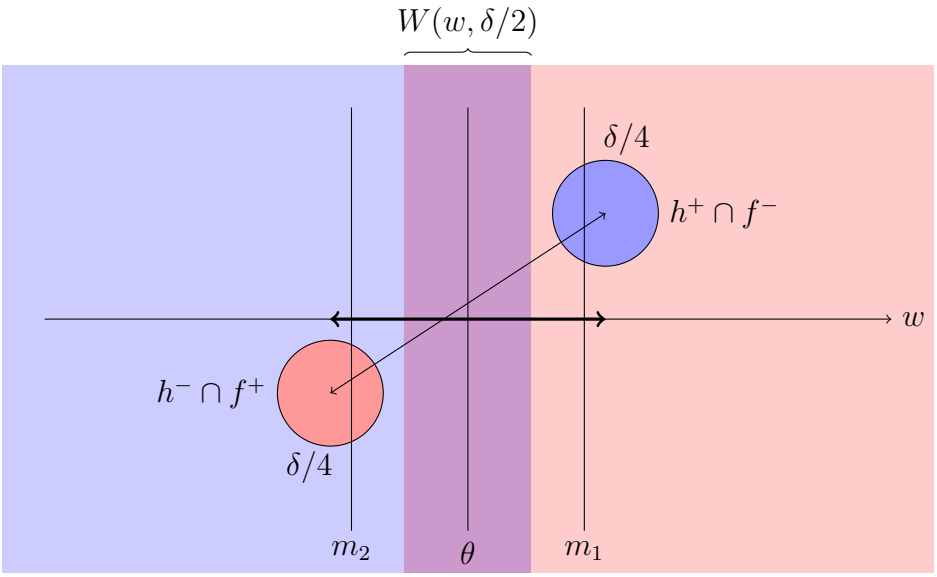


Figure 4.1: Proof of the Gap Theorem

Theorem 4.4.1. *Let μ be any probability measure on $\mathbb{R}^n$. Suppose $f : \mathbb{R}^n \to \{\pm 1\}$ is any measurable function and $h(x) = \text{sign}(\langle w, x\rangle - \theta)$ is a halfspace such that $\|w\| = 1$ and $\mathbb{E}[h] = \mathbb{E}[f]$. Then*

$$\|\mathsf{Com}(h) - \mathsf{Com}(f)\| \geq \frac{\mathsf{dist}(f,h)}{2\cos(\alpha)} W_\mu\left(w, \frac{\mathsf{dist}(f,h)}{2}\right)$$

where α is the angle between $\mathsf{Com}(h) - \mathsf{Com}(f)$ and the normal vector w.

Proof. For simplicity, let $\delta = \mathsf{dist}(f,h)$. Taking the inner product with w, the difference is

$$\begin{aligned}
&\|\mathbb{E}[x(h(x) - f(x))]\| \\
&\quad = \left\|\mu(h^+ \cap f^-)\mathbb{E}\left[2x \;\middle|\; x \in h^+ \cap f^-\right] - \mu(h^- \cap f^+)\mathbb{E}\left[2x \;\middle|\; x \in h^- \cap f^+\right]\right\| \\
&\quad = \frac{2}{\cos\alpha} \cdot \left(\mu(h^+ \cap f^-)\mathbb{E}\left[\langle w, x\rangle \;\middle|\; x \in h^+ \cap f^-\right] - \mu(h^- \cap f^+)\mathbb{E}\left[\langle w, x\rangle \;\middle|\; x \in h^- \cap f^+\right]\right)
\end{aligned}$$

Let m_1 be the median of $\langle w, x\rangle$, under the condition that $x \in h^+ \cap f^-$. Then

$$\begin{aligned}
&\mathbb{E}\left[\langle w,x\rangle \;\middle|\; x \in h^+ \cap f^-\right] \\
&\quad = \frac{1}{2}\mathbb{E}\left[\langle w,x\rangle \;\middle|\; x \in h^+ \cap f^-, \langle w,x\rangle < m_1\right] + \frac{1}{2}\mathbb{E}\left[\langle w,x\rangle \;\middle|\; x \in h^+ \cap f^-, \langle w,x\rangle \geq m_1\right] \\
&\quad \geq \frac{1}{2}\theta + \frac{1}{2}m_1
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\langle w,x\rangle \;\middle|\; x \in h^- \cap f^+\right] \leq \frac{1}{2}\theta + \frac{1}{2}m_2$$

where m_2 is the median of $\langle w, x\rangle$ under the condition $x \in h^- \cap f^+$. Note also that $\mathbb{E}[f] = \mathbb{E}[h]$ and $\mathbb{E}[f] = 2\mu(f^+) - 1, \mathbb{E}[h] = 2\mu(h^+) - 1$, which implies

$$0 \;=\; \mu(h^+) - \mu(f^+) \;=\; \mu(h^+ \cap f^-) - \mu(h^- \cap f^+)$$

so $\mu(h^+ \cap f^-) = \mu(h^- \cap f^+)$. Now $\delta = \mathsf{dist}(f,h) = \mu(h^+ \cap f^-) + \mu(h^- \cap f^+)$ so $\mu(h^+ \cap f^-) = \mu(h^- \cap f^+) = \delta/2$. Rewriting the difference gives us

$$2\frac{1}{\cos(\alpha)} \cdot \frac{\delta}{4}\left(\theta - \theta + m_1 - m_2\right) = \frac{\delta}{2\cos(\alpha)}(m_1 - m_2)\,.$$

Finally, note that since $\mu(h^+ \cap f^-) = \delta/2$ and m_1 is its median, the set $\{x : \langle w, x\rangle \in [\theta, m_1]\}$ must have measure at least $\delta/4$. Similarly, the set $\{x : \langle w, x\rangle \in [m_2, \theta)\}$ must have measure at least $\delta/4$, so $m_1 - m_2 \geq W_\mu(w, \delta/2)$, which completes the proof. □

As we will see later (Chapter 7), there are a few theorems for the hypercube that have a similar flavor, but do not have the requirement that $\mathbb{E}[f] = \mathbb{E}[h]$ and instead prove a bound that depends on the difference $\mathbb{E}[f] - \mathbb{E}[h]$.

Question 4.4.2. *Is there a version of the Gap Theorem that depends on $\mathbb{E}[f] - \mathbb{E}[h]$ instead of requiring that this difference is 0?*

4.5 Finding the Center from the Volume

To define the algorithm, we need to define a function ξ that tells us what the center of mass "ought to be", given the volume; i.e. we want a function that looks like

$$\xi(v) = \|\mathsf{Com}(h)\|_2^2$$

where h is a halfspace with volume $\mathbb{E}[h_v] = v$. The reason rotation-invariant distributions are nice is that no matter which direction the normal vector is pointing, the value of this function will be the same. This function ought to exist since for a halfspace $f(x) = \text{sign}(\langle w, x\rangle - \theta)$, $\mathbb{E}[f]$ is monotonically decreasing as the threshold θ increases. Thus there is an inverse function $\mathbb{E}[f] \mapsto \theta$, and from θ one may compute the center-norm. I will formalize this argument in this section.

Proposition 4.5.1. *Let μ be a rotationally invariant probability distribution with 1-dimensional projection μ_w onto an arbitrary unit vector w, and let $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ be a halfspace with volume $v = \mathbb{E}[f]$. Then*

$$\frac{dv}{d\theta} = \frac{d}{d\theta}\left(2\int_\theta^\infty \mu_w(z)dz - 1\right) = -2\mu_w(\theta)$$

and therefore the mapping $v \mapsto \theta$ is invertible on the support of μ_w; we will call this inverse function Φ^{-1}, so $\theta = \Phi^{-1}(v)$.

Proof. The equality holds because

$$v = \mathbb{E}[f] = \mathbb{P}[f(x) = 1] - \mathbb{P}[f(x) = -1] = 2\mathbb{P}[f(x) = 1] - 1\,. \qquad \square$$

Using this fact we can get our function:

Lemma 4.5.2. *Let μ be any rotationally invariant probability distribution over $\mathbb{R}^n$ with 1-dimensional projection μ_w onto an arbitrary unit vector w. There exists a function ξ such that for any halfspace $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ with volume $v = \mathbb{E}[f]$,*

$$\sqrt{\xi(\mathbb{E}[f])} = 2\int_\theta^\infty z\mu_w(z)dz = \|\mathsf{Com}(f)\|\,.$$

Furthermore, on the values $v \in (-1, 1)$, $\sqrt{\xi}$ has derivative

$$\frac{d}{dv}\sqrt{\xi(v)} = \theta\,.$$

Therefore $\|\mathsf{Com}(f)\| = \sqrt{\xi(\mathbb{E}[f])}$ is convex with respect to $\mathbb{E}[f]$ with its maximum achieved when the threshold satisfies $\theta = 0$, i.e. when $\mathbb{E}[f] = 0$.

Proof. The identity $\|\mathsf{Com}(f)\| = 2\int_\theta^\infty z\mu_w(z)dz$ holds due to rotation invariance:

$$\|\mathsf{Com}(f)\| = \langle w, \mathsf{Com}(f)\rangle = 2\left\langle w, \mu(f^+)\mathsf{Com}\left(f^+\right)\right\rangle = 2\int_{\{x:\langle w,x\rangle \geq \theta\}} \langle w, x\rangle\,\mu(dx)$$

where the equalities are, respectively, the fact that w and $\mathsf{Com}(f)$ are parallel (Proposition 4.2.2), Proposition 4.1.2, and linearity of expectation. Now rearranging the integral produces the identity.

That $2\int_\theta^\infty z\mu_w(z)dz$ is a function of $\mathbb{E}[f]$ is a consequence of the fact that $\theta = \Phi^{-1}(\mathbb{E}[f])$ (Proposition 4.5.1).

For the derivative, since the mapping $v \mapsto \theta$ is invertible, we have by Proposition 4.5.1

$$\frac{d}{dv}\sqrt{\xi(v)} = 2\left(\frac{dv}{d\theta}\right)^{-1}\frac{d}{d\theta}\int_\theta^\infty z\mu_w(z)dz = 2\frac{1}{-2\mu_w(\theta)}(-\theta\mu_w(\theta)) = \theta\,. \qquad \square$$

To determine how accurately we must estimate $\mathbb{E}[f]$ to get a good estimation of $\|\mathsf{Com}(f)\|$, we should know the derivative of the ξ function. I will relate this derivative to the variance of the 1-dimensional projection.

Lemma 4.5.3. *Let μ be a rotationally invariant probability distribution over $\mathbb{R}^n$, and let ξ be the function from the previous lemma. Then*

$$\left|\frac{d}{dv}\xi(v)\right| \le 2V\,.$$

Proof. First, by definition

$$\frac{d}{dv}\xi(v) = \frac{d}{dv}(\sqrt{\xi(v)})^2 = 2\sqrt{\xi(v)}\frac{d}{dv}\sqrt{\xi(v)} = 2\sqrt{\xi(v)}\cdot\Phi^{-1}(v)\,.$$

from Lemma 4.5.2. Expanding $\sqrt{\xi(v)}$,

$$\frac{d}{dv}\xi(v) = 4\theta\int_\theta^\infty z\mu_w(z)dz \le 4\int_\theta^\infty z^2\mu_w(z)dz \le 4\int_0^\infty z^2\mu_w(z)dz = 2\int_{-\infty}^\infty z^2\mu_w(z)dz\,.$$

This latter term is the variance of a 1-dimensional projection, which is at most $2V$ By symmetry, if $\theta < 0$ (i.e. $v > 0$) then $\frac{d}{dv}\xi(v) = -\frac{d}{dv}\xi(-v) \ge -2V$ by the above argument. $\square$

We can get some basic lower and upper bounds on ξ with respect to the width and variance:

Proposition 4.5.4. *Let μ be a rotationally invariant probability distribution over $\mathbb{R}^n$ and let $0 < \epsilon < 1$. Let $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$ be a halfspace with volume $|\mathbb{E}[f]| < 1 - \epsilon$. Then*

$$\|\mathsf{Com}(f)\| = \sqrt{\xi(\mathbb{E}[f])} \ge \epsilon\cdot W_\mu(1-\epsilon)\,.$$

Proof. Assume without loss of generality that $\mathbb{E}[f] \geq 0$. Let $v' = 1 - \epsilon$ and $v^+ = 1 - \epsilon/2, v^- = \epsilon/2$ so that $v' = v^+ - v^-$. Since $\sqrt{\xi(v)}$ is convex with its maximum achieved at 0, we have $\sqrt{\xi(v)} > \sqrt{\xi(v')}$ because $v < v'$. Let θ be the threshold such that $\mathbb{E}[\text{sign}(\langle w, x\rangle - \theta)] = v'$ (for any w since the space is rotationally invariant), i.e. $\theta = \Phi^{-1}(1-\epsilon)$ and $v^+ = \int_\theta^\infty \mu_w(z)dz$. Note that $v \geq 0$ so $\theta < 0$; therefore

$$\int_\theta^{-\theta} \mu_w(z)dz = 1 - \epsilon$$

so $2\theta \geq W(1-\epsilon)$. Then

$$\sqrt{\xi(v)} \geq 2\int_\theta^\infty z\mu_w(z)dz \geq 2\theta\int_\theta^\infty \mu_w(z)dz = 2\theta\epsilon \geq \epsilon W(1-\epsilon)\,. \qquad \square$$

Proposition 4.5.5. *Let μ be any rotationally invariant distribution over $\mathbb{R}^n$ and let $v \in (-1, 1)$. Then for any arbitrary unit vector w,*

$$\sqrt{\xi(v)} \leq \mathbb{E}[|\langle w, x\rangle|] \leq \sqrt{V}\,.$$

Proof. We know that $\sqrt{\xi(v)}$ is maximized at $v = 0$, when the threshold $\theta = 0$, so we need only consider this case. Taking the 1-dimensional projection μ_w along any axis w gives us

$$\sqrt{\xi(0)} = 2\int_0^\infty z\mu_w(z)dz = \int_{-\infty}^\infty |z|\mu_w(z)dz = \mathop{\mathbb{E}}_{z\sim\mu_w}[|z|] = \mathop{\mathbb{E}}_{x}[|\langle w, x\rangle|]$$

The final inequality holds because for all centered random variables x,

$$\mathbb{E}[|x|] = \sqrt{\mathbb{E}[|x|]^2} \leq \sqrt{\mathbb{E}[x^2]} = \sqrt{\mathbb{V}[x]}$$

by Jensen's inequality (Theorem 2.3.4). $\square$

4.6 Estimating the Norm of the Centroid

Now we have the first two ingredients of the algorithm: the ξ function and the Gap Theorem. The final ingredient is a way of estimating $\|\mathsf{Com}(f)\|$ using only samples. There are two observations that I use to develop the estimator. First is that for independent $x, y \sim \mu$, $\mathbb{E}[\langle xf(x), yf(y)\rangle] = \langle \mathsf{Com}(f), \mathsf{Com}(f)\rangle = \|\mathsf{Com}(f)\|^2$ by linearity of expectation, and second is that with a set of m random samples, we actually have $\binom{m}{2}$ examples

of $\langle xf(x), yf(y)\rangle$ instead of just $m/2$ (but these pairs are not all independent). I combine this with Chebyshev's inequality to show that we need only $\sqrt{n}$ samples.

To use Chebyshev's inequality, we will need to compute the variance of the estimator. We will calculate a general formula for the variance since we will reuse it in the next chapter:

Lemma 4.6.1. *Let μ be a probability distribution over $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \{\pm 1\}$ be a measurable function with center $c = \mathbb{E}[xf(x)]$. Let $\{x_1, x_2, \ldots, x_m\}$ be independently chosen random variables from μ. Then*

$$\mathbb{V}\left[\binom{m}{2}^{-1} \sum_{i<j} f(x_i)f(x_j)\langle x_i, x_j\rangle\right] = 4 \cdot \frac{m-2}{m(m-1)}\left(\mathbb{E}\left[\langle x, c\rangle^2\right] - \|c\|^4\right) + \frac{2}{m(m-1)}\left(\mathbb{E}\left[\langle x, y\rangle^2\right] - \|c\|^4\right)$$

Proof. Let $M = \binom{m}{2}$ for convenience. By definition:

$$\mathbb{V}\left[M^{-1}\sum_{i<j} f(x_i)f(x_j)\langle x_i, x_j\rangle\right] = M^{-2}\mathbb{E}\left[\left(\sum_{i<j} f(x_i)f(x_j)\langle x_i, x_j\rangle\right)^2\right] - \|c\|^4 .$$

Expanding the expectation, we get

$$\sum_{i<j,k<\ell} f(x_i)f(x_j)f(x_k)f(x_\ell)\langle x_i, x_j\rangle\langle x_k, x_\ell\rangle , \tag{4.1}$$

which can be broken down into 3 cases: either i, j, k, ℓ are all distinct, or the 4 variables take 3 distinct values, or they take only 2 distinct values.

The first case, in which i, j, k, ℓ are all distinct, will occur $\binom{m}{4} \cdot 6$ times in the sum, since there are $\binom{m}{4}$ choices of values, and 6 ways of ordering the variables, since $i < j$ and $k < \ell$. In this case, the expectation of the summand is

$$\mathbb{E}[f(x_i)f(x_j)f(x_k)f(x_\ell)\langle x_i, x_j\rangle\langle x_k, x_\ell\rangle] = \langle c, c\rangle^2 = \|c\|^4$$

by linearity of expectation.

The third case, in which $i = k, j = \ell$, will occur $\binom{m}{2}$ times in the sum, since we must choose 2 distinct values for i, j. In this case,

$$\begin{aligned}\mathbb{E}[f(x_i)f(x_j)f(x_k)f(x_\ell)\langle x_i, x_j\rangle\langle x_k, x_\ell\rangle] &= \mathbb{E}\left[f(x_i)^2 f(x_j)^2\langle x_i, x_j\rangle^2\right] \\ &= \mathbb{E}\left[\langle x_i, x_j\rangle^2\right] .\end{aligned}$$

Finally, the second case, in which there are 3 distinct values taken by the indices, occurs $\binom{m}{2}^2 - \binom{m}{2} - 6 \cdot \binom{m}{4}$ times, which is

$$\begin{aligned}\binom{m}{2}^2 - \binom{m}{2} - 6 \cdot \binom{m}{4} &= \binom{m}{2}\left(\frac{m(m-1)}{2} - \frac{2}{2} - \frac{(m-2)(m-3)}{2}\right) \\ &= \frac{1}{2}\binom{m}{2}\left(m^2 - m - 2 - m^2 + 5m - 6\right) \\ &= \binom{m}{2}(2m-4) \\ &= 2\frac{m(m-1)(m-2)}{2} = 6 \cdot \binom{m}{3}.\end{aligned}$$

In this case, we may assume without loss of generality that $i = k, j \neq \ell$. Then

$$\begin{aligned}\mathbb{E}\left[f(x_i)^2 f(x_j) f(x_\ell) \langle x_i, x_j\rangle \langle x_i, x_\ell\rangle\right] &= \mathbb{E}\left[\langle x_i, x_j f(x_j)\rangle \langle x_i, x_\ell f(x_\ell)\rangle\right] \\ &= \mathbb{E}\left[\langle x_i, c\rangle^2\right]\end{aligned}$$

Putting these cases together, we get that the variance is

$$\begin{aligned}&M^{-2}\left(6 \cdot \binom{m}{4}\|c\|^4 + 6 \cdot \binom{m}{3}\mathbb{E}\left[\langle x, c\rangle^2\right] + \binom{m}{2}\mathbb{E}\left[\langle x, y\rangle^2\right]\right) - \|c\|^4 \\ &= M^{-2}\left(6 \cdot \binom{m}{3}\left(\mathbb{E}\left[\langle x, c\rangle^2\right] - \|c\|^4\right) + \binom{m}{2}\left(\mathbb{E}\left[\langle x, y\rangle^2\right] - \|c\|^4\right)\right)\end{aligned}$$

where we have used the fact that $6\binom{m}{4} + 6\binom{m}{3} + \binom{m}{2} = \binom{m}{2}^2$. Now writing out $M, \binom{m}{3}, \binom{m}{2}$ in terms of m proves the lemma. □

Lemma 4.6.2. *Let μ be any RI probability distribution over $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \{\pm 1\}$. For all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, independent of n, we can estimate $\|\mathsf{Com}(f)\|^2$ to within $\pm\epsilon$ using at most*

$$O\left(\frac{\sqrt{RV}}{\epsilon\sqrt{\delta}}\right)$$

random samples, where $R = \mathbb{E}\left[\|x\|^2\right]$.

Proof. Let $\{x_i\}_{i\in[m]}$ the set of independent random sample points drawn from μ, where m is some number to be determined later. We will use $\binom{m}{2}^{-1}\sum_{i<j}\langle f(x^i)x^i, f(x^j)x^j\rangle$ as our estimator: note that

$$\mathbb{E}\left[\langle f(x)x, f(y)y\rangle\right] = \langle \mathbb{E}\left[f(x)x\right], \mathbb{E}\left[f(y)y\right]\rangle = \langle c, c\rangle = \|c\|^2$$

for $c = \mathsf{Com}(f) = \mathbb{E}[f(x)x]$. Letting $M = \binom{m}{2}$ for simplicity, we will use Chebyshev's inequality to show that

$$\mathbb{P}\left[\left|M^{-1}\sum_{i<j}\left\langle f(x^i)x^i, f(x^j)x^j\right\rangle - \|c\|^2\right| > \eta\right] < \delta$$

for an appropriate choice of m. From Lemma 4.6.1, we know the variance is

$$4\cdot\frac{m-2}{m(m-1)}\left(\mathbb{E}\left[\langle x,c\rangle^2\right] - \|c\|^4\right) + \frac{2}{m(m-1)}\left(\mathbb{E}\left[\langle x,y\rangle^2\right] - \|c\|^4\right),$$

so we need only calculate $\mathbb{E}\left[\langle x,c\rangle^2\right], \mathbb{E}\left[\langle x,y\rangle^2\right]$. Using rotational invariance, we may let u be drawn uniformly randomly from the unit sphere, and write

$$\mathbb{E}\left[\langle x_i,x_j\rangle^2\right] = \mathbb{E}\left[\|x_j\|^2\langle x_i,u\rangle^2\right] = RV. \tag{4.2}$$

Again using rotational invariance, we can let u be drawn from the unit sphere and write

$$\mathbb{E}\left[\langle x_i,c\rangle^2\right] = \|c\|^2\,\mathbb{E}\left[\langle x_i, c/\|c\|\rangle^2\right] = \|c\|^2 V. \tag{4.3}$$

The variance is then

$$\begin{aligned}
&4\cdot\frac{m-2}{m(m-1)}\|c\|^2\left(V-\|c\|^2\right) + \frac{2}{m(m-1)}\left(RV-\|c\|^4\right)\\
&\quad\le \frac{m-2}{m(m-1)}V^2 + \frac{2}{m(m-1)}RV
\end{aligned}$$

since the first term is maximized when $\|c\|^2 = V/2$. Using the identity $V = RV^*$ (Proposition 4.3.7), where V^* is the 1-dimensional variance of the unit sphere, and the inequality $V^* \le \frac{3}{n+1}$ (Proposition A.0.8), we get the upper bound

$$\frac{RV}{m(m-1)}\left((m-2)V^*+2\right) \le \frac{RV}{m(m-1)}\left(\frac{3(m-2)}{n+1}+2\right)$$

Note that the second term is bounded by a constant as long as $m \le O(n)$, so setting $m = \Theta\left(\frac{\sqrt{RV}}{\epsilon\sqrt{\delta}}\right)$ will suffice, as long as $\sqrt{RV} \le O(n)$, since ϵ, δ are independent of n. Again using the identities $R = RV^*$ and $V^* = \Theta(1/n)$, we can rewrite this condition as $\frac{R}{n} \le O(n)$ or $R \le O(n^2)$.

Setting the constant in $m = \Theta\left(\frac{\sqrt{RV}}{\epsilon\sqrt{\delta}}\right)$ appropriately, using Chebyshev's inequality gives us

$$\mathbb{P}\left[\left|M^{-1}\sum_{i<j} f(x_i)f(x_j)\left\langle x_i, x_j\right\rangle - \|c\|^2\right| > \epsilon\right] \leq \frac{RV}{m(m-1)\epsilon^2}\left(2 + 3\frac{m-2}{n+1}\right) \leq \delta\,,$$

which completes the proof. □

Chebyshev's inequality is quite basic and it is possible that more advanced concentration inequalities could produce better results here. However, recall that there is a $\Omega\left(\sqrt{n/\log n}\right)$ lower bound for passive testing in the Gaussian space ([BBBY12], see Subsection 3.3.2) while the above estimator uses roughly $\sqrt{n}$ samples, so more advanced methods can only significantly improve the dependence on ϵ.

4.7 Algorithm

Mixing all the ingredients together, we finally get the algorithm for testing halfspaces in RI distributions. Below I will write $\mu_{max} := \max_z \mu_w(z)$ as the maximum density of the 1-dimensional projection of μ.

Algorithm 5 RI Tester

Input: $f : \mathbb{R}^n \to \{\pm 1\}, \epsilon \in (0,1)$

1: **function** μ-HALFSPACE TESTER(f, ϵ)
2: Let $\epsilon_1 \leftarrow \frac{\epsilon}{4\mu_{max}}$;
3: Let $\tilde{v}$ be an empirical estimate of $\mathbb{E}[f]$ to within $\pm\frac{1}{2V}\epsilon_1^3$;
4: Let $\tilde{c}^2$ be an estimate of $\|c\|^2 = \|\mathbb{E}[xf(x)]\|^2$ to within $\pm\epsilon_1^3$;
5: **if** $\xi(\tilde{v}) - \tilde{c}^2 \leq 2\epsilon_1^3$ **then accept**;

Theorem 4.7.1. *Algorithm 5 satisfies the following properties: for any f and $0 < \epsilon < 1/2$,*

1. *If f is a halfspace then A accepts with probability at least $2/3$,*
2. *If f is ϵ-far from all halfspaces then A rejects with probability at least $2/3$, and*
3. *A requires at most $O\left(\frac{\sqrt{RV}\mu_{max}^3}{\epsilon^3} + \frac{V^2\mu_{max}^6}{\epsilon^6}\right)$ labelled samples.*

Proof. By Lemma 4.6.2, with $O\left(\frac{\sqrt{RV}}{\epsilon_1^3}\right)$ samples we can estimate $\|c\|^2$ to within $\pm\epsilon_1^3$ (with constant probability, say 5/6).

To estimate the volume $\mathbb{E}[f]$ we can use standard Hoeffding bounds. From Lemma 4.5.3 we know that $\frac{d}{dv}\xi(v) \leq 2V$ so to estimate $\xi(v)$ within $\pm\epsilon_1^3$ we need to estimate v to within $\pm\epsilon_1^3/2V$. Then taking random samples $x_1, \ldots, x_m$ we have

$$\mathop{\mathbb{P}}_{X}\left[\frac{1}{m}\left|\sum_i f(x_i)\right| > \eta\right] \leq 2\exp\left(-\frac{m\eta^2}{2}\right),$$

which is at most 1/6 when

$$m = \frac{2}{\eta^2}\ln(12) = O\left(\frac{1}{\eta^2}\right) = O\left(\frac{V^2}{\epsilon_1^6}\right).$$

Assume that both estimations are successful, which occurs with probability at least $1 - (1/6 + 1/6) = 2/3$. Assume f is a halfspace. Then the total error in $\xi(\tilde{v}) - \tilde{c}^2$ is at most $2\epsilon_1^3$, so the algorithm will accept.

Now suppose f is accepted by the algorithm, so $\xi(\tilde{v}) - \tilde{c}^2 \leq 2\epsilon_1^3$. If $|v| \geq 1 - 2\epsilon$ then either $v^+ \leq \epsilon$ or $v^- \leq \epsilon$ so f is ϵ-close to a constant function and would be correctly accepted, so we may assume $|v| < 1 - 2\epsilon$. We may also assume, for contradiction, that $\mathsf{dist}(f, h) > \epsilon$, since otherwise the algorithm was correct to accept.

Suppose h is the halfspace whose center is parallel to the center of f with $\mathbb{E}[h] = \mathbb{E}[f]$. Let $\mu_{max} = \max_z \mu_w(z)$ denote the maximum density of the projection. From the Gap Theorem and the inequality $(a-b) \leq (a-b)(a+b)/b = (a^2-b^2)/b$ for $a > b > 0$, we have

$$W(\mathsf{dist}(f,h)/2)\mathsf{dist}(f,h) \leq \frac{\xi(v) - \|c\|^2}{\sqrt{\xi(v)}} \leq \frac{\xi(\tilde{v}) - \tilde{c}^2 + 2\epsilon_1^3}{\sqrt{\xi(v)}} \leq \frac{4\epsilon_1^2}{W(1-2\epsilon)},$$

where the final inequality is due to Proposition 4.5.4 and the assumption that $|v| \leq 1 - 2\epsilon$. Since $\mathsf{dist}(f, h) > \epsilon$ by assumption, we have $W(\mathsf{dist}(f,h)/2) \geq W(\epsilon/2)$ and we can write

$$\mathsf{dist}(f,h) \leq \frac{4\epsilon_1^2}{W(\epsilon/2)\cdot W(1-2\epsilon)},$$

We want to show a contradiction, which we get by setting

$$\epsilon_1 = \frac{1}{2}\sqrt{\epsilon \cdot W(\epsilon/2)\cdot W(1-2\epsilon)} \geq \frac{\epsilon}{4\mu_{max}} \tag{4.4}$$

where we have used the inequality $W(a) \geq a/\mu_{max}$ and the fact that $1 - 2\epsilon \geq 1/2$ since $\epsilon < 1/4$. □

The dependence on μ_{max} in the sample complexity is not ideal. For many rotationally invariant spaces, such as the Gaussian space or the uniform sphere of radius $\Theta(\sqrt{n})$, μ_{max} will be a constant along with the variance V. Unfortunately, this may not hold for all RI spaces, as I will show in the next example.

Example 4.7.2. Let μ_1, μ_2 be the uniform distributions over spheres of respective radii r_1, r_2, and for $\lambda \in (0,1)$ let μ be the mixture $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$. We want to restrict this distribution so that it is comparable in scale to the Gaussian distribution, and we will look at two ways to do this: first, dictating that $V = 1$, and second, that $\mathbb{E}[\|x\|] = \sqrt{n}$.

Suppose that $1 = V = \lambda V_1 + (1-\lambda)V_2$. From Proposition A.0.9 we have

$$1 = V^* \left(\lambda r_1^2 + (1-\lambda) r_2^2\right)$$

where $V^* = (2 \pm 1)/(n+1)$ is the variance for the unit sphere (Proposition A.0.8), so $\lambda r_1^2 + (1-\lambda) r_2^2 \approx n$. Now for any choice of r_1 (say, very small) and constant λ, we can choose r_2 to satisfy $(1-\lambda) r_2^2 \approx n - \lambda r_1^2$, say $r_2 \approx \sqrt{n}$ (note that we cannot set $r_2 = \omega(\sqrt{n})$ unless λ depends on n). Thus have constructed a distribution where at least a constant λ mass occurs within radius r_1, which is arbitrarily small; this implies that $W(\epsilon) < r_1$ is arbitrarily small for any constant $\epsilon < \lambda$. This plays havoc with the Gap Theorem so the accuracy we need in the algorithm explodes. In this example we also have $\mathbb{E}[\|x\|] = \lambda r_1 + (1-\lambda) r_2 = \Theta(r_2) = \Theta(\sqrt{n})$ so as we would expect, the other restriction has an identical weakness.

This example shows that a sample complexity of $\sqrt{n}$ is achieved for distributions whose width satisfies $W(\epsilon) \geq C \cdot \epsilon$ (for any constant $C > 0$), but those with sublinear width will require further study. The relationship between the width, variance, center-norm and sample complexity is an interesting topic for future work. Since we can still learn halfspaces with roughly n samples on these extreme spaces, I expect to find a testing algorithm with $\sqrt{n}$ sample complexity as well:

Question 4.7.3. *How can we improve the algorithm to use fewer samples on spaces with sublinear width?*

4.7.1 Distance Metrics

There is a subtle observation that we can make about this tester: it is slightly stronger than is required for the definition of property testing algorithms. The distance metric

that we usually use for property testing is the L_1 metric, i.e. the distance between two boolean-valued functions f, g is $\mathsf{dist}(f, g) = \mathbb{P}[f(x) \neq g(x)]$. However, as a consequence of the Gap Theorem, a function f is correctly rejected by Algorithm 5 if it is ϵ-far from the halfspace h *with the same volume* — this is not necessarily the closest halfspace to f, so it is possible that f is less than ϵ-far from being a halfspace. Thus the algorithm correctly rejects in some cases where acceptance is acceptable. So it is possible that the tester works for some distance metrics that are more difficult than L_1.

Question 4.7.4. *Does this testing algorithm work for other, more difficult distance metrics?*

4.7.2 Easy Extensions

To conclude this chapter, I make the easy observation that the algorithm, with hardly any modification, can be used for probability spaces that are linear transformations of rotationally-invariant spaces. This is because any linear threshold function will remain a linear threshold function after such a transformation, and our model assumes that the algorithm knows the distribution; that is, we could tailor the algorithm to apply the inverse of the transformation.

Chapter 5

Beyond Rotation-Invariance

The algorithm for rotationally invariant spaces relied on the fact that the center of mass is always the same length, no matter which direction the weight vector is pointing. Abandoning rotational invariance means abandoning this nice property, so we need to adapt our algorithm to this inconvenience. For RI distributions, we found a function ξ such that $\xi(\mathbb{E}[f]) = \|\mathsf{Com}(f)\|^2$. To adapt the algorithm to other spaces, we can try to find an analogous function $\xi(p_1, p_2, \dots) = \|\mathsf{Com}(f)\|^2$, depending on some parameters p_i that are easy to estimate, that determines what the center-norm of the halfspace ought to be. The number of parameters should be small: if we have more than n parameters we might as well learn the function.

In this chapter I will look at mixtures of 2 rotationally invariant spaces μ_1, μ_2, and develop a 2-parameter function depending on the volumes $v_1 = \mathbb{E}_{\mu_1}[f]$ and $v_2 = \mathbb{E}_{\mu_2}[f]$ due to each component distribution. The algorithm can then work as before: estimate the parameters v_1, v_2 and compute $\xi(v_1, v_2)$ to find what the center-norm ought to be; then estimate and compare the center-norm, with correctness guaranteed by the Gap Theorem. I will show that a small modification of the previous algorithm will work for the mixture under the condition that the components are not too far apart (at most $n^{1/8}$) and that the derivatives of the new ξ function are not too large; fully understanding these restrictions is left for future work.

As an example of why rotation variance can be troublesome, consider the simple statement "The coordinate i with the largest weight w_i should have the largest influence on the function value". This statement is quite intuitive and indeed it holds for many distributions, including RI distributions and the uniform hypercube:

Proposition 5.0.5. *Let μ be any probability distribution over $\mathbb{R}^n$ that is coordinate-wise symmetric, i.e. whose density satisfies, for all $x \in \mathbb{R}^n$ and $i \in [n]$, $\mu(x^{i\leftarrow x_i}) = \mu(x^{i\leftarrow -x_i})$. Suppose $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ is a halfspace with center of mass $c = \mathsf{Com}(f)$. Then for all $i, j \in [n]$, $\text{sign}(w_i) = \text{sign}(c_i)$, and if $|w_i| \geq |w_j|$ then $|c_i| \geq |c_j|$.*

Proof. Without loss of generality, consider w_1 and assume $w_1 \geq 0$. Fix any $\alpha = \sum_2^n w_i x_i$. Note that for all x_1 such that $|w_1 x_1| \leq |\theta - \alpha|$, the point $-x_1$ has the same probability and also satisfies this condition; therefore $\mathbb{E}\left[w_1 x_1 f(x) \;\middle|\; |w_1 x_1| \leq |\theta - \alpha|\right] = 0$. In the opposite case, if $w_1 x_1 > |\theta - \alpha|$ then $x_1 > 0$ and $f(x) = 1$ and if $w_1 x_1 < |\theta - \alpha|$ then $x_1 < 0$ and $f(x) = -1$, so

$$\mathbb{E}\left[w_1 x_1 f(x) \;\middle|\; |w_1 x_1| > |\theta - \alpha|\right] > 0\,.$$

This proves $\text{sign}(w_i) = \text{sign}(\hat{f}(i))$.

For the second part, assume without loss of generality that $i = 1, j = 2$, and $w_1 \geq w_2 \geq 0$. Fix any $\alpha = \theta - \sum_{i=3}^n w_i x_i$. Consider

$$\hat{f}(1) - \hat{f}(2) = \mathbb{E}\left[(x_1 - x_2)\,\text{sign}(w_1 x_1 + w_2 x_2 - \alpha)\right]\,.$$

Let x_1, x_2 be such that the term in the expectation is negative. We will show that the mapping $(x_1, x_2) \mapsto (x_2, x_1)$ is a one-to-one mapping from negative examples to positive ones. Consider the expression

$$(w_1 x_1 + w_2 x_2) - (w_1 x_2 + w_2 x_1) = (w_1 - w_2)(x_1 - x_2)\,.$$

There are two cases: first, if $x_1 < x_2$ and $w_1 x_1 + w_2 x_2 \geq \alpha$. Then this expression is at most 0, so $\alpha \leq (w_1 x_1 + w_2 x_2) \leq (w_1 x_2 + w_2 x_1)$, and the new point is a positive example. In the second case, $x_1 > x_2$ and $w_1 x_1 + w_2 x_2 < \alpha$, so the expression is at least 0, $\alpha > (w_1 x_1 + w_2 x_2) \geq (w_1 x_2 + w_2 x_1)$, and this is also a positive point. □

The requirement that the distribution be not only symmetric (i.e. $\mu(x) = \mu(-x)$ for all x) but coordinate-wise symmetric ($\mu(x^{i\leftarrow +1}) = \mu(x^{i\leftarrow -1})$) seems too strong; however, we can easily construct an example distribution that is symmetric but for which the fact does not hold:

Example 5.0.6. Let $w = \frac{1}{\sqrt{2}}(1 + \epsilon, 1 - \epsilon)$ and consider the uniform distribution over the ellipse $x_1^2 + 4x_2^2 \leq 1$, illustrated in Figure 5.1:

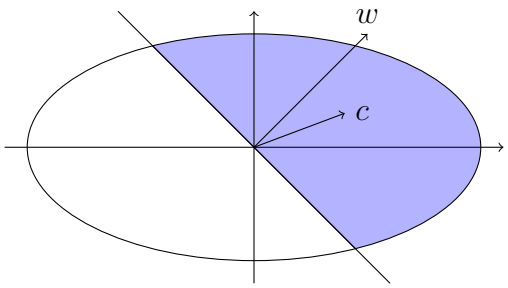


Figure 5.1: A symmetric distribution where the order of $\mathsf{Com}(f)$ coordinates differs from the order of the w coordinates.

5.1 Mixtures of Rotationally Invariant Distributions

In this section I will extend the tools used for rotationally invariant distributions to mixtures of 2 rotationally invariant distributions; a good example to keep in mind is a mixture of Gaussians.

Definition 5.1.1. In this chapter, we will be considering distributions $\mu = \frac{1}{2}(\mu_1 + \mu_2)$ over $\mathbb{R}^n$ where μ_1, μ_2 are RI (and share the same σ–algebra) and have means $m, -m$ respectively; I will usually write $M := \|m\|$. It will be useful to consider the distributions $\hat{\mu}_1, \hat{\mu}_2$ which are the centered copies of μ_1 and μ_2; formally, for all x,

$$\hat{\mu}_1(x) := \mu_1(x + m)\,, \qquad \hat{\mu}_2(x) := \mu_2(x + -m)\,.$$

I will write $R_i := \mathop{\mathbb{E}}_{x\sim\hat{\mu}_i}\left[\|x\|^2\right]$ as the "radius" of each (centered) component, and $V_i := \mathop{\mathbb{E}}_{x\sim\hat{\mu}_i}\left[\langle w, x\rangle^2\right]$ (where w is an arbitrary unit vector), as used in the previous chapter. I will also write ξ_1, ξ_2 to be the respective functions defined in Lemma 4.5.2. For simplicity, I will write

$$\Xi := \left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right)\,.$$

Keeping the same notation as in the previous chapter, I will also use, when considering a function $f : \mathbb{R}^n \to \{\pm 1\}$,

$$v_i := \mathop{\mathbb{E}}_{x\sim\mu_i}\left[f(x)\right]\,.$$

As in the previous chapter, we shall write

$$V_i := \sup_{w:\|w\|=1} \mathop{\mathbb{V}}_{x\sim\hat{\mu}_i}\left[\langle w, x\rangle\right] = \mathop{\mathbb{E}}_{x\sim\hat{\mu}_i}\left[\langle w, x\rangle^2\right]$$

where in the last expression w is an arbitrary unit vector, and equality holds due to rotational invariance.

For a unit vector w, we can calculate the 1-dimensional variance along w by splitting it into the component distributions:

Proposition 5.1.2. *Let μ be as in Definition 5.1.1, and let $w \in \mathbb{R}^n$ be an arbitrary unit vector. Then*

$$V_\mu(w) = \frac{V_1 + V_2}{2} + \langle m, w \rangle^2 .$$

Proof. By definition,

$$V(w) = \mathop{\mathbb{E}}_{x \sim \mu} \left[\langle x, w \rangle^2 \right] = \frac{1}{2} \left(\mathop{\mathbb{E}}_{x \sim \mu_1} \left[\langle x, w \rangle^2 \right] + \mathop{\mathbb{E}}_{x \sim \mu_2} \left[\langle w, x \rangle^2 \right] \right) .$$

First considering $x \sim \mu_1$ and writing $x = u + m$ where $u \sim \hat{\mu}_1$ is drawn from the centered copy of μ_1, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{x \sim \mu_1} \left[\langle x, w \rangle^2 \right] &= \mathop{\mathbb{E}}_{u \sim \hat{\mu}_1} \left[\langle u + m, w \rangle^2 \right] \\ &= \mathbb{E} \left[(\langle u, w \rangle + \langle m, w \rangle)^2 \right] \\ &= \mathbb{E} \left[\langle u, w \rangle^2 + 2 \langle u, w \rangle \langle m, w \rangle + \langle m, w \rangle^2 \right] \\ &= V_1 + \langle m, w \rangle^2 \end{aligned}$$

since $\mathbb{E}[u] = \vec{0}$ in the final equality. Doing the same for μ_2 completes the proof. □

Following the same strategy, we can compute the value for R in terms of R_1 and R_2:

Proposition 5.1.3. *In the same setting as the previous proposition,*

$$R = \frac{R_1 + R_2}{2} + \|m\|^2 .$$

Proof. Again by definition, we have

$$R = \mathop{\mathbb{E}}_{x \sim \mu} \left[\|x\|^2 \right] = \frac{1}{2} \left(\mathop{\mathbb{E}}_{x \sim \mu_1} \left[\|x\|^2 \right] + \mathop{\mathbb{E}}_{x \sim \mu_2} \left[\|x\|^2 \right] \right) .$$

First considering $x \sim \mu_1$ and writing $x = u + m$ for $u \sim \hat{\mu}_1$ drawn from the centered copy of μ_1, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mu_1}\left[\|x\|^2\right] &= \mathop{\mathbb{E}}_{u\sim\hat{\mu}_1}\left[\langle u+m, u+m\rangle\right] \\ &= \mathbb{E}\left[\langle u,u\rangle + 2\langle u,m\rangle + \langle m,m\rangle\right] \\ &= R_1 + \|m\|^2 ,\end{aligned}$$

since $\mathbb{E}[u] = \vec{0}$ in the final inequality. Doing the same for μ_2 completes the proof. □

Now I construct the ξ function that maps the separate volumes to the center-norm. Before diving into the calculations, note that it seems natural for $\mathsf{Com}(f)$ to simply be $\frac{1}{2}(\hat{c}_1 + \hat{c}_2)$, the average of the centers due to the separate components. However, this turns out to be incorrect since the the component distributions are not centered.

Lemma 5.1.4. *Let μ be as in Definition 5.1.1, and let $f(x) = \operatorname{sign}(\langle w, x\rangle - \theta)$. Then*

$$\mathsf{Com}(f) = \frac{1}{2}\left(\Xi w + (v_1 - v_2)m\right) .$$

Proof. Let $c = \mathop{\mathbb{E}}_{x\sim\mu}[xf(x)]$ and $c_i = \mathop{\mathbb{E}}_{x\sim\mu_i}[xf(x)]$, so $c = \frac{c_1+c_2}{2}$. Since μ_1, μ_2 are rotationally invariant about their means, we may simplify their analysis by centering them. Denote by $\hat{x}$ the point $x - m$, so $\mathop{\mathbb{E}}_{x\sim\mu_1}[\hat{x}] = \mathop{\mathbb{E}}_{x\sim\mu_1}[x - m] = 0$. We also define a function $\hat{f}$ such that $\hat{f}(\hat{x}) = f(x)$. This function over the centered space has center

$$\hat{c}_1 = \mathop{\mathbb{E}}_{x\sim\mu_1}\left[\hat{x}\hat{f}(\hat{x})\right] = \mathop{\mathbb{E}}_{x\sim\mu_1}[(x-m)f(x)] = c_1 - v_1 m \tag{5.1}$$

Note that the volume of $\hat{f}$ over the centered distribution is $\mathop{\mathbb{E}}_{x\sim\mu_1}\left[\hat{f}(\hat{x})\right] = \mathop{\mathbb{E}}_{x\sim\mu_1}[f(x)] = v_1$. Since the distribution of $\hat{x}$ is centered and rotation-invariant, we know that there exists a function ξ_1 such that $\xi_1(v_1) = \|\hat{c}_1\|^2$. Examining $\hat{f}$, we see that for all x,

$$\begin{aligned}\hat{f}(\hat{x}) = f(x) &= \operatorname{sign}(\langle w, x\rangle - \theta) \\ &= \operatorname{sign}\left(\langle w, x\rangle - \langle w, m\rangle + \langle w, m\rangle - \theta\right) \\ &= \operatorname{sign}\left(\langle w, x-m\rangle + \langle w, m\rangle - \theta\right) \\ &= \operatorname{sign}\left(\langle w, \hat{x}\rangle + \langle w, m\rangle - \theta\right)\end{aligned}$$

This means $\hat{f}$ is a halfspace whose center (with respect to the centered distribution) is parallel to w. Thus $\hat{c}_1 = \|\hat{c}_1\| w = \sqrt{\xi_1(v_1)}w$. Combining this with equation (5.1), we know that

$$c_1 = \hat{c}_1 + v_1 m = \sqrt{\xi_1(v_1)}w + v_1 m .$$

Performing a similar transformation on μ_2 we get

$$c_2 = \hat{c}_2 + v_2(-m) = \sqrt{\xi_2(v_2)}w - v_2 m$$

with $\|\hat{c}_2\|^2 = \xi_2(v_2)$. Thus

$$\begin{aligned} c = \frac{c_1 + c_2}{2} &= \frac{1}{2}\left(\sqrt{\xi_1(v_1)}w + v_1 m + \sqrt{\xi_2(v_2)}w - v_2 m\right) \\ &= \frac{1}{2}\left(\left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right)w + (v_1 - v_2)m\right) . \end{aligned}$$

□

Theorem 5.1.5. *In the same setting as the previous lemma, there exists a function $\xi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $\|\mathsf{Com}\,(f)\|^2 = \xi(v_1, v_2)$, given by*

$$\begin{aligned} \xi(v_1, v_2) := \frac{1}{4}\Bigg(&\left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right)^2 \\ &+ (v_1 - v_2)(\theta_1 - \theta_2)\left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right) + (v_1 - v_2)^2 M^2\Bigg) . \end{aligned}$$

Proof. Let c, c_1, c_2 be as defined in Lemma 5.1.4. Then

$$\|c\|^2 = \langle c, c\rangle = \frac{1}{4}\left(\Xi^2 + 2\Xi(v_1 - v_2)\langle w, m\rangle + (v_1 - v_2)^2\|m\|^2\right) .$$

This is clearly a function of v_1, v_2 and $\langle w, m\rangle$, so all that remains is to show that $\langle w, m\rangle$ is also a function of v_1, v_2. We do this by noting that the centered halfspaces have thresholds $\theta_1 = \langle w, m\rangle - \theta$ and $\theta_2 = \langle w, -m\rangle - \theta = -\langle w, m\rangle - \theta$. Thus, since the centered spaces are rotationally invariant, we can use the functions Φ_1^{-1} and Φ_2^{-1} to compute

$$\theta_1 = \Phi_1^{-1}(v_1) \qquad\qquad \theta_2 = \Phi_2^{-1}(v_2)\,.$$

Finally, observe that

$$\theta_1 - \theta_2 = \langle w, m\rangle - \theta + \langle w, m\rangle + \theta = 2\langle w, m\rangle$$

so dividing by 2 gives us what we want. □

We have v_1 and v_2 as the parameters for ξ, which means v_1 and v_2 should be easily estimated with samples. The next lemma shows how to do this:

Lemma 5.1.6. *Let $\mu_1, \ldots, \mu_k$ be probability distributions over $\mathbb{R}^n$ (with the same σ–algebras)[1], and let $\mu = \sum_{i\in[k]} \lambda_i \mu_i$ where $\sum_{i\in[k]} \lambda_i = 1$ and $\lambda_i > 0$ for all $i \in [k]$. Let*

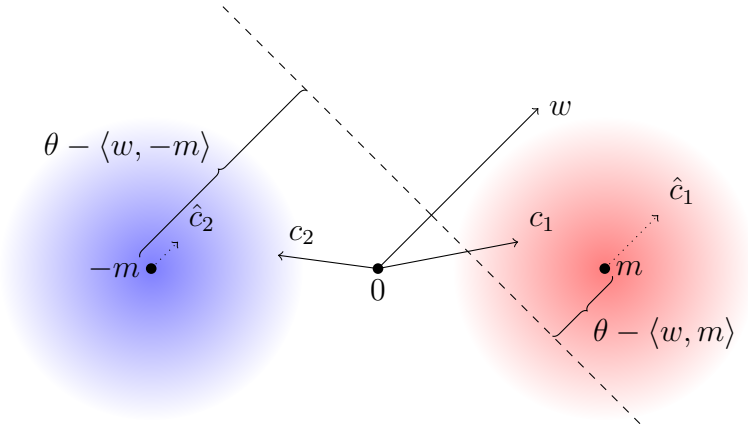


Figure 5.2: The center of mass in a mixture of RI distributions.

$f : \mathbb{R}^n \to \{\pm 1\}$ and define $v_i := \mathop{\mathbb{E}}_{x \sim \mu_i}[f(x)]$ for each i. Then, writing $\mu_i(x)$ as the density of μ_i with respect to μ,

$$v_i = \mathop{\mathbb{E}}_{x \sim \mu}\left[\frac{\mu_i(x)}{\mu(x)} f(x)\right] .$$

Proof. Let S be the support of $\mu = \sum_{i \in [k]} \lambda_i \mu_i$ (note that the support S_i of μ_i satsifies $S_i \subseteq S$). Then

$$\mathop{\mathbb{E}}_{x \sim \mu}\left[\frac{\mu_i(x)}{\mu(x)} f(x)\right] = \int_S \frac{\mu_i(x)}{\mu(x)} f(x) \mu(x) \mu(dx) = \int_S f(x) \mu_i(x) \mu(dx) = v_i \,. \qquad \square$$

Recall that for the correctness of the RI algorithm, we made use of the fact that we could pick a halfspace whose center was parallel to the center of f, and use this halfspace in the Gap Theorem. This let us use the identity

$$\|\mathsf{Com}\,(h) - \mathsf{Com}\,(f)\| = \|\mathsf{Com}\,(h)\| - \|\mathsf{Com}\,(f)\| \,.$$

To establish the same identity for mixtures, I will first establish the continuity of the center of mass as a function of the weight vector and volume:

[1] A technical condition for this lemma is that the mixture should comprise distributions of the same "type"; it wouldn't, for example, make sense to mix discrete and continuous distributions (their σ-algebras would not match up).

Lemma 5.1.7. *Let μ be as in Definition 5.1.1, and let $h(w,v) = \text{sign}(\langle w, \cdot\rangle - \theta_v)$ be the mapping from normal vectors w with $\|w\| = 1$ and volumes $v \in [-1,1]$ to the halfspace with normal w and volume v. Then:*

1. *For all v, $w \mapsto \mathsf{Com}(h(w,v))$ is continuous at w,*
2. *For all w, $v \mapsto \mathsf{Com}(h(w,v))$ is continuous at v.*

Proof. Fix $v \in [-1,1]$ and let $w \in \mathbb{R}^n$ such that $\|w\| = 1$. From Lemma 5.1.4 we have

$$\mathsf{Com}(h(w,v)) = \frac{1}{2}\left(\left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right)w + (v_1 - v_2)m\right),$$

and $v_1 = \int_{\theta_1}^{\infty} \mu_{1,w}(dx)$. μ_1 is indiscrete since it is rotationally invariant, meaning $\mu_{1,w}$ is absolutely continuous with respect to the Lebesgue measure. Thus v_1 is continuous at $\theta_1 = \theta - \langle w, m\rangle = \Phi_w^{-1}(v) - \langle w, m\rangle$, which is itself a continuous function of w. The same holds for v_2. Thus $\mathsf{Com}(h(w,v))$ is a composition of functions continuous at w, so it is continuous.

Next fix $w \in \mathbb{R}^n$ with $\|w\| = 1$. By a similar argument, we know v_1, v_2 are continuous functions of v since $\theta_1 = \Phi^{-1}(v) - \langle w, m\rangle$ and $\Phi_w^{-1}(v)$ is continuous since μ is indiscrete. Thus $\mathsf{Com}(h(w,v))$ is again a composition of functions continuous at v, so it is continuous. □

Using continuity, we can show that for any vector $c \in \mathbb{R}^n$ and volume v, there exists a halfspace with volume v whose center is parallel to c:

Lemma 5.1.8. *Let μ be as in Definition 5.1.1. Let $c \in \mathbb{R}^n$ such that $\|c\| = 1$ and let $v \in (-1,1)$. Then there exists a halfspace h such that $\mathbb{E}[h] = v$ and $\langle \mathsf{Com}(h), c\rangle = \|\mathsf{Com}(h)\|$.*

Proof. We first show that a halfspace with volume v exists whose center has the same angle to m as c, i.e. $\left\langle \frac{\mathsf{Com}(h)}{\|\mathsf{Com}(h)\|}, m \right\rangle = \langle c, m\rangle$. This holds since $w \mapsto \mathsf{Com}(h)$ is continuous by the previous lemma, so

$$w \mapsto \left\langle m, \frac{\mathsf{Com}(h)}{\|\mathsf{Com}(h)\|} \right\rangle$$

is a composition of continuous functions of w. Setting $w = m/\|m\|$ we have $\left\langle m, \frac{\mathsf{Com}(h)}{\|\mathsf{Com}(h)\|} \right\rangle = \|m\|$ while setting $w = -m/\|m\|$ gets us $\left\langle m, \frac{\mathsf{Com}(h)}{\|\mathsf{Com}(h)\|} \right\rangle = -\|m\|$. Thus for some w the inner product is $\langle m, c\rangle \in [-\|m\|, \|m\|]$ by the Intermediate Value Theorem.

Now that we have a halfspace with the correct angle to m, we note that since μ is rotationally invariant around the axis m, we merely need to rotate w about this axis to get the correct normal vector. □

I expect that this fact is true for general indiscrete spaces (but obviously not for finite discrete spaces, in which there are only a finite number of possible halfspaces):

Conjecture 5.1.9. *Let μ be any indiscrete distribution over $\mathbb{R}^n$ and let $c \in \mathbb{R}^n$ be any unit vector and $v \in (-1, 1)$. Then there exists a halfspace h with volume v whose center is parallel to c.*

To estimate the center-norm to within, say, $\pm\eta$ using approximations of v_1 and v_2, we must know how accurate our estimates of v_1 and v_2 should be.

Lemma 5.1.10. *Let μ be the mixture and ξ the function defined in Theorem 5.1.5. Suppose that $v = (v_1 + v_2)/2$ and v_1 satisfies $|v_1| \leq 1 - \epsilon$. Then the partial derivatives of ξ are bounded by*

$$|D_i\xi(v_1, v_2)| \leq O\left(M^2 + \frac{M\sqrt{V'} + V'}{\sqrt{\epsilon}} + \frac{\sqrt{V'}}{\mu_i(\sqrt{2V'/\epsilon} + 2M)}\right),$$

where $V' = \max(V_1, V_2)$.

Proof. Recall the definition of the function

$$\xi(v_1, v_2) := \frac{1}{4}\left(\Xi^2 + (v_1 - v_2)\Xi(\theta_1 - \theta_2) + (v_1 - v_2)^2 M^2\right),$$

where $\theta_1 = \langle w, m\rangle - \theta, \theta_2 = \langle w, -m\rangle - \theta = -\langle w, m\rangle - \theta$. Also recall that the distributions μ_1, μ_2 are symmetric about their respective means, so the 1-dimensional projections μ_1^1, μ_2^1 are symmetric. In particular, this implies that the partial derivatives of ξ with respect to v_1, v_2 are symmetric: $D_1\xi(v_1, v_2) = -D_1\xi(-v_1, v_2)$. From Proposition 4.5.5 we get

$$\Xi \leq \sqrt{V_1} + \sqrt{V_2}\,.$$

The last term has the easiest derivative,

$$\frac{\partial}{\partial v_1}(v_1 + v_2)^2 M^2 = 2M^2(v_1 - v_2) \leq 4M^2\,,$$

which shows that the estimation will depend on the separation parameter.

The derivative of the first term is

$$\frac{\partial}{\partial v_1}\Xi^2 = \frac{\partial}{\partial v_1}\left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right)^2 = \frac{d}{dv_1}\xi_1(v_1) + 2\sqrt{\xi_2(v_2)}\frac{d}{dv_1}\sqrt{\xi_1(v_1)}$$
$$= 2\Xi\frac{d}{dv_1}\sqrt{\xi_1(v_1)} = 2\Xi\theta_1$$

by Lemma 4.5.2.

The second term's derivative is

$$\frac{\partial}{\partial v_1}(v_1 - v_2)\left(\sqrt{\xi_1(v_1)} + \sqrt{\xi_2(v_2)}\right)(\Phi_1^{-1}(v_1) - \Phi_2^{-1}(v_2))$$
$$= \Xi(\theta_1 - \theta_2) + (v_1 - v_2)\theta_1(\theta_1 - \theta_2) - (v_1 - v_2)\Xi\frac{1}{2\mu_1(\theta_1)}$$

These two terms depend on the threshold θ_1. We can get a bound on θ_1 using the assumption that $|v_1| \leq 1 - \epsilon$, so

$$\frac{\epsilon}{2} = \mathbb{P}\left[z \geq \theta_1\right] \tag{5.2}$$

We can easily bound this with Chebyshev's inequality:

$$\frac{\epsilon}{2} \leq \frac{V_1}{\theta_1^2}$$

so $\theta_1 \leq \sqrt{2V_1/\epsilon}$. Thus we can bound the first and second terms in the derivative by

$$2\Xi\theta_1 \leq 2(\sqrt{V_1} + \sqrt{V_2})\sqrt{2V_1/\epsilon} = O\left(\frac{\max(V_1, V_2)}{\sqrt{\epsilon}}\right)$$

and

$$2(\sqrt{V_1} + \sqrt{V_2})M + 4M\sqrt{2V_1/\epsilon} + \frac{\sqrt{V_1} + \sqrt{V_2}}{\mu_1(\sqrt{2V_1/\epsilon})} = O\left(\frac{M\sqrt{\max(V_1, V_2)}}{\sqrt{\epsilon}} + \frac{\sqrt{\max(V_1, V_2)}}{\mu_1(\sqrt{2V_1/\epsilon})}\right).$$

For the partial derivative with respect to v_2, we cannot use equation (5.2), and instead must use the fact that $\theta_2 \leq \theta_1 + 2M$ to get the same bound as above but with $\mu_2(\sqrt{2V_1/\epsilon} + 2M)$ in the final denominator. □

These basic bounds are, I expect, far from optimal. For one thing, this simplistic bound does not take advantage of any cancellations that could occur (which may be intricate in unbalanced cases where the component distributions are very dissimilar). Another

immediate target for improvement is the use of Chebyshev's inequality for equation (5.2). In this equation, z is drawn from the projection of a 1-dimensional RI distribution so there should be significantly more structure to exploit. Consider the Gaussian distribution for example:

Example 5.1.11. We can use Lemma 2.3.12 to get

$$\frac{\epsilon}{2} = \mathbb{P}\left[z \geq \theta\right] \leq \exp\left(-\theta^2/2\right)$$

so $\theta \leq \sqrt{2\ln(2/\epsilon)}$. This gives

$$\mu_w(\theta) \geq \frac{1}{\sqrt{2\pi}}\exp\left(-2\ln(2/\epsilon)/2\right) = \sqrt{2/\pi}\epsilon$$

In fact, the concentration bound in equation (5.2) appears to be most evasive in the same situations that are tough for the RI algorithm; it will be interesting to learn more about this.

Question 5.1.12. *What is the best bound on these derivatives? Can we get much better bounds when $W(\epsilon) \geq C\cdot\epsilon$ (for any constant C)? What happens when we restrict $M \leq n^{1/8}$?*

The $n^{1/8}$ question comes from the next section, where we will see that this restriction naturally arises from the analysis of the center–norm estimator.

5.2 Estimating the Center–Norm

Now it is necessary to update the estimation guarantee, Lemma 4.6.2. All that is required is to apply the formula Lemma 4.6.1 and compute the two expectations in that formula. In this section I will use the notation $\bar{c} = c/\left\|c\right\|$ and $\bar{m} = m/\left\|m\right\|$. Then we get the following lemma:

Lemma 5.2.1. *Let μ be as in Definition 5.1.1, and let $f : \mathbb{R}^n \to \{\pm 1\}$ be any measurable function with center $c = \mathsf{Com}\left(f\right)$. Let*

$$B = \frac{R_1 + R_2}{2} \cdot \frac{V_1 + V_2}{2} + M^2(V_1 + V_2 + M^2)\,.$$

Then the variance of the estimator $\binom{k}{2}^{-1} \sum_{i<j} f(x_i) f(x_j) \langle x_i, x_j \rangle$ *for independent random vectors* $x_1, \ldots, x_k \sim \mu$ *is*

$$4\frac{k-2}{k(k-1)} \|c\|^2 \left(V_\mu(\bar{c}) - \|c\|^2\right) + 2\frac{1}{k(k-1)} B\,.$$

Proof. From Lemma 4.6.1, the total variance of the estimator $\binom{k}{2}^{-1} \sum_{i<j} \langle f(x_i) f(x_j) \langle x_i, x_j \rangle \rangle$ is

$$\binom{k}{2}^{-2} \left(6 \cdot \binom{k}{3} \left(\mathbb{E}\left[\langle x, c\rangle^2\right] - \|c\|^4\right) + \binom{k}{2} \left(\mathbb{E}\left[\langle x, y\rangle^2\right] - \|c\|^4\right)\right),$$

so we need only compute $\mathbb{E}\left[\langle x, y\rangle^2\right]$ and $\mathbb{E}\left[\langle x, c\rangle^2\right]$. Splitting the first expectation into its components, we get

$$\mathop{\mathbb{E}}_{x,y\sim\mu}\left[\langle x, y\rangle^2\right] = \frac{1}{4}\mathop{\mathbb{E}}_{x,y\sim\mu_1}\left[\langle x, y\rangle^2\right] + \frac{1}{4}\mathop{\mathbb{E}}_{x,y\sim\mu_2}\left[\langle x, y\rangle^2\right] + \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mu_1, y\sim\mu_2}\left[\langle x, y\rangle^2\right].$$

We can focus on the more general third term. Write $x = u + m, y = v - m$ where $u \sim \hat{\mu}_1, v \sim \hat{\mu}_2$ are drawn from the centered copies of their respective distributions. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{u\sim\hat{\mu}_1, v\sim\hat{\mu}_2}\left[\langle u + m, v - m\rangle^2\right] &= \mathbb{E}\left[\left(\langle u, v\rangle - \langle u, m\rangle + \langle v, m\rangle - \langle m, m\rangle\right)^2\right] \\
&= \mathbb{E}\left[\langle u, v\rangle^2 + \langle u, m\rangle^2 + \langle v, m\rangle^2 + \langle m, m\rangle^2\right] \\
&= \mathbb{E}\left[\|v\|^2 \langle u, \bar{v}\rangle^2 + M^2 \langle u, \bar{m}\rangle^2 + M^2 \langle v, \bar{m}\rangle^2 + M^4\right] \\
&= R_2 V_1 + M^2 V_1 + M^2 V_2 + M^4 \\
&= R_1 R_2 V^* + M^2 R_1 V^* + M^2 R_2 V^* + M^4
\end{aligned}$$

where we have used $\mathbb{E}[u] = \mathbb{E}[v] = \vec{0}$ in the second equality to eliminate all cross-terms, leaving only the squares. By setting $\mu_1 = \mu_2$ we can specialize this equation for the other two cases:

$$\mathop{\mathbb{E}}_{x,y\sim\mu_1}\left[\langle x, y\rangle^2\right] = R_1^2 V^* + 2M^2 R_1 V^* + M^4$$

and similar for μ_2. Combining these gives us

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,y\sim\mu}\left[\langle x, y\rangle^2\right] &= \frac{1}{4}\left(R_1^2 V^* + 2M^2 R_1 V^* + R_2^2 V^* + 2M^2 R_2 V^* + 2M^4\right) \\
&\quad + \frac{1}{2}\left(R_1 R_2 V^* + M^2(R_1 + R_2) V^* + M^4\right) \\
&= \frac{1}{4}\left(R_1^2 V^* + R_2^2 V^* + 2R_1 R_2 V^* + 4M^2(R_1 + R_2) V^* + 4M^4\right) \\
&= \left(\frac{R_1 + R_2}{2}\right)^2 V^* + M^2\left(V_1 + V_2 + M^2\right) = B\,.
\end{aligned}$$

For the other expectation, we can plainly see

$$\mathop{\mathbb{E}}_{x\sim\mu}\left[\langle x,c\rangle^2\right] = \|c\|^2\,\mathbb{E}\left[\langle x,\bar{c}\rangle^2\right] = \|c\|^2\,V_\mu(\bar{c})\,.$$

(Compare to the RI setting, where this was $\|c\|^2\,V$.) Thus the variance is

$$\begin{aligned}&\binom{k}{2}^{-2}\left(6\cdot\binom{k}{3}\|c\|^2\left(V(\bar{c})-\|c\|^2\right)+\binom{k}{2}B\right)\\ &\quad= 4\frac{k-2}{k(k-1)}\|c\|^2\left(V(\bar{c})-\|c\|^2\right)+\frac{2}{k(k-1)}(B-\|c\|^4)\,.\end{aligned}$$ □

In the RI setting, we can set $k = O\left(\frac{\sqrt{RV}}{\epsilon\sqrt{\delta}}\right) = O\left(\frac{\sqrt{n}}{\epsilon\sqrt{\delta}}\right)$ to get ϵ accuracy with confidence $1-\delta$ (Lemma 4.6.2), and we would like a similar result here. I will analyze the two terms of the variance and show that we can get a similar result when we constrain the separation parameter (we will require that $M = O\left(n^{1/8}\right)$.

The first observation is that we can get a simple worst–case bound on B:

Proposition 5.2.2. *Let μ be as in Definition 5.1.1 satisfying $R_1 = R_2 = n$. Then*

$$B = \Theta\left(M^4 + n\right)\,.$$

Proof. The condition $R_1 = R_2 = n$ implies $V_1 + V_2 = \Theta(1)$ (Propositions 4.3.7 and A.0.8), so

$$B = \frac{R_1+R_2}{2}\cdot\frac{V_1+V_2}{2} + M^2(V_1+V_2+M^2) = \Theta\left(n + M^4 + M^2\right)\,.$$ □

Recall that we want the variance to be roughly $\epsilon^2\delta$ and m to have dependence $\sqrt{n}$ on n. Then we should have $B - \|c\|^4 = O\left(\sqrt{n}\right)$. We can assume that f is ϵ–far from constant but even in this case we can have $\|c\| = 0$, so we cannot make any assumption on $\|c\|$. Therefore, to get the sample complexity we want, we need the restriction $M = O\left(n^{1/4}\right)$, and this is the best we can do with this estimator and analysis.

We must also check that this restriction is sufficient to bound the first term: recall that this term is

$$4\frac{k-2}{k(k-1)}\|c\|^2\left(\mathbb{E}\left[\langle x,\bar{c}\rangle^2\right] - \|c\|^2\right)\,,$$

or more simply, we can consider

$$\frac{1}{k}\|c\|^2\left(\mathbb{E}\left[\langle x,\bar{c}\rangle^2\right] - \|c\|^2\right)\,.$$

Our goal is to have $k = O(\sqrt{n})$, ignoring ϵ and δ factors. Thus we must have

$$\|c\|^2 \left(\mathbb{E}\left[\langle x, \bar{c}\rangle^2\right] - \|c\|^2\right) = O\left(\sqrt{n}\right) ;$$

recall that this holds for RI distributions, since $V = \mathbb{E}\left[\langle x, \bar{c}\rangle^2\right]$ was a constant. Unfortunately, for mixtures, $\mathbb{E}\left[\langle x, \bar{c}\rangle^2\right]$ is not constant so we need to do more work.

Proposition 5.2.3. *In the same setting as Lemma 5.2.1,*

$$\|c\|^2 \left(\mathbb{E}\left[\langle x, \bar{c}\rangle^2\right] - \|c\|^2\right) = O\left(M^4\right) .$$

Proof. The 1-dimensional variance, from Proposition 5.1.2, is

$$V(\bar{c}) = \mathbb{E}\left[\langle x, \bar{c}\rangle^2\right] = \frac{V_1 + V_2}{2} + \langle \bar{c}, m\rangle^2 = O\left(M^2\right)$$

since $V_1 + V_2$ is a constant. The center–norm satisfies

$$\|c\|^2 = \frac{1}{4}\left(\Xi^2 + 2\Xi(v_1 - v_2)\langle w, m\rangle + (v_1 - v_2)^2 M^2\right) = O\left(M^2\right)$$

since $\Xi \leq V_1 + V_2 = O(1)$ (Propositions 4.5.5, 4.3.7 and A.0.8), so

$$\|c\|^2 \left(\mathbb{E}\left[\langle x, \bar{c}\rangle^2\right] - \|c\|^2\right) = O\left(M^4\right) . \qquad \square$$

The following example shows that this is essentially tight: the term can be as large as $\Theta(M^4)$.

Example 5.2.4. We have $c = \frac{1}{2}(\Xi w + (v_1 - v_2)M)$, and

$$\|c\|^2 = \frac{1}{4}\left(\Xi^2 + 2\Xi(v_1 - v_2)\langle w, m\rangle + (v_1 - v_2)^2 M^2\right) .$$

Pick any halfspace such that $(v_1 - v_2) = \Theta(1)$, so $\|c\| = \Theta(M)$. Assume $\langle w, \bar{m}\rangle \geq 0$ and $v_1 \geq v_2$, which is a consistent assumption. Then

$$\langle \bar{c}, \bar{m}\rangle = \frac{1}{\|c\|}\langle c, \bar{m}\rangle = \frac{\Xi\langle w, \bar{m}\rangle + (v_1 - v_2)M}{\|c\|} = \Theta\left(\frac{(v_1 - v_2)M}{M}\right) = \Theta(1)$$

since $v_1 - v_2$ is constant. Then $V(\bar{c}) = \Theta(M^2)$.

We can certainly pick w such that $\langle w, \bar{m}\rangle = O\left(1/\sqrt{n}\right)$ and $(v_1 - v_2)$ is bounded by a constant (in fact $\langle w, \bar{m}\rangle = 1/\sqrt{n}$ works for a mixture of two spheres of radius $\sqrt{n}$). Therefore this example shows another important fact: even under the natural restriction that $|\langle w, m\rangle| = O(1)$ (i.e. the volumes due to both components are bounded away from constant), we can have $|\langle \bar{c}, m\rangle| = \Omega(M)$.

Now, since we have an example where

$$\frac{1}{k}\|c\|^2\left(\mathbb{E}\left[\langle x, \bar{c}\rangle^2\right] - \|c\|^2\right) = \Theta\left(\frac{M^4}{k}\right),$$

we see that, ignoring ϵ and δ, we must set $k = \Theta(M^4)$. Thus $M^4 = \sqrt{n}$ so $M = n^{1/8}$ is asymptotically the largest separation parameter that the estimator can tolerate with this analysis. I summarize the above discussion with the following lemma:

Lemma 5.2.5. *Let μ be as in Definition 5.1.1 satisfying $\|m\| = n^{1/8}$ and $R_1 = R_2 = n$. Let $f : \mathbb{R}^n \to \{\pm 1\}$ be any measurable function, and let $0 < \epsilon, \delta < 1$. Then we can estimate $\|\mathsf{Com}(f)\|^2$ to within $\pm\epsilon$ with confidence $1 - \delta$ using at most*

$$O\left(\frac{\sqrt{n}}{\epsilon^2\delta}\right)$$

random samples.

Proof. Let $x_1, x_2, \ldots, x_k \sim \mu$ be a set of independent random points. From Lemma 5.2.1 and Propositions 5.2.2 and 5.2.3, we know that the variance of the estimator

$$E := \binom{k}{2}^{-1} \sum_{i<j} f(x_i) f(x_j) \langle x_i, x_j\rangle$$

is at most

$$O\left(\frac{1}{k}M^4 + \frac{1}{k^2}(M^4 + n)\right) = O\left(\frac{\sqrt{n}}{k} + \frac{n}{k^2}\right)$$

by the assumption $M = n^{1/8}$. By Chebyshev's inequality,

$$\mathbb{P}\left[\left|E - \|c\|^2\right| \geq \epsilon\right] \leq \frac{\mathbb{V}[E]}{\epsilon^2} = O\left(\frac{\sqrt{n}}{k\epsilon^2} + \frac{n}{k^2\epsilon^2}\right).$$

Setting $k = \Theta\left(\frac{\sqrt{n}}{\epsilon^2\delta}\right)$ with an appropriate constant bounds this probability by δ, as we want. □

5.3 Algorithm

Now that we have explored the relationship between the center of mass and the volume in these mixture distributions, we can modify the algorithm to make use of these new properties. We have extended the ξ function from a function of 1 variable (the volume) to a function of 2 variables (the volumes contributed by each distribution in the mixture) and so the most important change to the algorithm is the estimation of two volume parameters instead of 1.

Algorithm 6 Tester for RI Mixtures

Input: $f : \mathbb{R}^n \to \{\pm 1\}, \epsilon \in (0, 1)$

1: **function** μ-Halfspace Tester(f, ϵ)
2: Let $\epsilon_1 = \frac{1}{2}\sqrt{\epsilon \cdot W(\epsilon/2) \cdot W(1-2\epsilon)}$;
3: Let $D \leftarrow \inf_{v_2} \frac{1}{D_1\xi(1-\epsilon, v_2)}$ be the bound from Lemma 5.1.10;
4: Let $\tilde{v}_1$ be an empirical estimate of $\mathop{\mathbb{E}}_{\mu_1}[f]$ to within $\pm D\epsilon_1^3$;
5: Let $\tilde{v}_2$ be an empirical estimate of $\mathop{\mathbb{E}}_{\mu_2}[f]$ to within $\pm D\epsilon_1^3$;
6: Let $\tilde{c}^2$ be an estimate of $\|c\|^2 = \|\mathbb{E}[xf(x)]\|^2$ to within $\pm\epsilon_1^3$;
7: **if** $\xi(\tilde{v}_1, \tilde{v}_2) - \tilde{c}^2 \leq 2\epsilon_1^3$ **then accept**

Theorem 5.3.1. *Algorithm 6 satisfies the following properties: Let μ be as in Definition 5.1.1 such that $R_1 = R_2 = n$ and $\|m\| \leq n^{1/8}$. For all $\epsilon > 0$ and f satisfying $|\mathbb{E}[f]| \leq 1 - \epsilon$,*

1. *If f is a halfspace then A accepts with probability at least $2/3$,*
2. *If f is ϵ-far from all halfspaces then A rejects with probability at least $2/3$, and*
3. *A requires at most $O\left(\frac{\sqrt{n}\mu_{max}^3}{\epsilon^6} + \frac{D^2\mu_{max}^6}{\epsilon^6}\right)$ labelled samples.*

where D is the bound from Lemma 5.1.10.

Proof. This proof is essentially the same as the proof of Theorem 4.7.1. We first use Lemma 5.2.5: with $O\left(\frac{\sqrt{n}}{\epsilon^6}\right)$ samples we can estimate $\|c\|^2$ to within $\pm\epsilon^3$ (with constant probability, say $5/6$).

To estimate the volumes $v_i = \mathop{\mathbb{E}}_{\mu_i}[f]$ we can use standard Hoeffding bounds. From Lemma 5.1.10 we know that $D_1\xi, D_2\xi$ are bounded by D^{-1}. Then to estimate $\xi(v_1, v_2)$ within $\pm\epsilon_1^3$ we need to estimate v_1, v_2 to within $\pm D\epsilon_1^3$. By Lemma 5.1.6 we may use $\frac{\mu_i}{\mu} f$ as an unbiased estimator. Since $\mu = (\mu_1 + \mu_2)/2$ we must have that $\mu_1(x), \mu_2(x) \leq 2\mu(x)$ so

$\left|\frac{\mu_1(x)}{\mu(x)} f(x)\right| \leq 2$. Then taking random samples $X = \{x_1, \ldots, x_k\}$ we have, for either v_1 or v_2,

$$\mathop{\mathbb{P}}_{X}\left[\frac{1}{k}\left|\sum_i \frac{\mu_1(x_i)}{\mu(x_i)} f(x_i)\right| > \eta\right] \leq 2\exp\left(-\frac{k\eta^2}{16}\right)$$

which is at most $1/12$ when

$$k = \frac{16}{\eta^2}\ln(12) = O\left(\frac{1}{\eta^2}\right) = O\left(\frac{D^2}{\epsilon_1^6}\right)$$

Running the same procedure (with a new set of samples) for v_2 gives us $\tilde{v}_2$.

Following the proof of Theorem 4.7.1, which we may do since we may choose a halfspace h with the same volume and parallel center by Lemma 5.1.8 we have

$$\mathsf{dist}(f, h) \leq \frac{4\epsilon_1^2}{W(\epsilon/2) \cdot W(1-2\epsilon)},$$

if f is accepted by the algorithm, with the conclusion that we can set

$$\epsilon_1 = \frac{1}{2}\sqrt{\epsilon \cdot W(\epsilon/2) \cdot W(1-2\epsilon)} \geq \frac{\epsilon}{4\mu_{max}}. \qquad \square$$

As I noted in the previous chapter, this dependence on μ_{max} doesn't seem ideal, but for "reasonable" distributions, with parameters close to what we'd expect for properly scaled spaces, our desired $\sqrt{n}$ sample complexity is recovered. With stronger bounds on the derivatives of ξ, the algorithm should achieve similar performance to the one for single RI distributions, with similar caveats. These bounds might arise from the restriction $M \leq n^{1/8}$, especially under the condition that $|v_1|, |v_2| < 1 - \epsilon$; the derivatives become unwieldy when this condition is not satisfied. Finally, we would like to remove the restriction that $M \leq n^{1/8}$; ideally, a separation of $M = \sqrt{n}$ should be allowed:

Question 5.3.2. *Can we improve this tester or its analysis to work when $M = \sqrt{n}$?*

Chapter 6

The Chow Parameters Problem

This chapter takes a detour away from testing halfspace towards a related problem about halfspaces, known as the Chow Parameters Problem, that was introduced in the 1960s. The main purpose of this diversion is to show that the Gap Theorem is interesting independent of its use in property testing, by applying it to another problem. I will review recent work of De *et al.* [DDFS14] that presents an efficient algorithm for solving the Chow Parameters Problem on the hypercube. Afterwards, in Section 6.2, I will show that we can use the Gap Theorem (Theorem 4.4.1) to generalize this algorithm, and in the process I show that the Gap Theorem applies to bounded functions as well as ± 1–valued functions (Theorem 6.2.1). The generalization is a work in progress: with a more powerful version of the Gap Theorem that tolerates small differences in the volume of the functions (see Question 4.4.2), we would get the full generalization.

The Chow Parameters Problem (CPP) is this: Given the center of mass c and volume v of a halfspace, can we efficiently find a weight vector w and threshold θ that define a halfspace with center c and volume v? Recall that the center of mass and the normal vector are always parallel in rotationally invariant spaces (Proposition 4.2.2), making the problem trivial; on the other hand, we have seen that this fails when the space loses its rotational invariance (recall Example 5.0.6). Thus the problem is an instance of a broader question:

Question 6.0.3. *What is the relationship between the normal vector of a halfspace and its center of mass?*

The motivation for the CPP is grounded in a very influential theorem about halfspaces. In 1961, Chao-Kong Chow [Cho61] showed that for linear threshold functions on the hyper-

cube, the center of mass and volume (i.e. the degree 0 and 1 Fourier coefficients) uniquely determine the function within the set of all boolean functions. That is, any function that shares these values with a halfspace is itself that same halfspace. This fact is not unique to the uniform distribution over the hypercube; it continues to hold for *any* probability space, as we show below by a simple application of the Gap Lemma:

Theorem 6.0.4 (Chow's Theorem (Informal[1]))**.** *Let μ be any probability distribution over $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \{\pm 1\}$ be the halfspace $f(x) = \text{sign}(\langle w, x\rangle - \theta)$. Let $g : \mathbb{R}^n \to \{\pm 1\}$ be any function such that $\mathsf{Com}(g) = \mathsf{Com}(f)$ and $\mathbb{E}[g] = \mathbb{E}[f]$. Then $\mathbb{P}_x[f(x) \neq g(x)] = 0$.*

Proof. Since the volumes of f, g are the same and $\|\mathsf{Com}(h) - \mathsf{Com}(f)\| = 0$, Theorem 4.4.1 dictates that $\mathsf{dist}(f, g) = 0$ as long as $W(w, \mathsf{dist}(f, g)/2) > 0$. This holds for indiscrete distributions, and for finite discrete distributions one may perturb w slightly without affecting f to achieve the same. □

Chow's paper posed the question of determining, from a given center c and volume v, whether there exists a halfspace with those parameters. The importance of this question is perhaps best illustrated with a surprisingly natural application that arises in social choice theory.

Suppose we are tasked with designing, for, say, a coalition of countries, a voting system in which each country makes a vote that is weighted by its population. We can model any voting system as a boolean function: the i^{th} country makes the vote $x_i \in \{\pm 1\}$ and the result is either a yes $(+1)$ or a no (-1). We want each vote to influence the outcome in a way that is proportional to the country's population; for instance, a country with a population of 10 million will have a greater influence on the result of the vote than a country with a population of only 1 million. Recall that we already have a definition of "influence" for boolean functions (definition 2.5.6): the probability that the country's vote will be the deciding vote, assuming that all possible votes are equally likely (the "impartial culture assumption" [O'D14]). Finally, recall that these influences are exactly the first-degree Fourier coefficients of the boolean function, when the function is monotone (as in this case: more 'yes' votes cannot flip the outcome from 'yes' to 'no'). Then given the list of influences of each country, we would want to find weights that would give rise to the desired voting system: this is exactly the Chow Parameters Problem. Problems such as this are surveyed in [Kur16].

[1] This theorem is marked as informal since there may be technical analytic conditions on the distribution excluding some pathological cases.

Question 6.0.5 ((Inexact) Chow Parameters Problem)**.** *Let f be the linear threshold function with $c = \mathsf{Com}(f)$ and $v = \mu(f)$. Given approximations $\tilde{c}, \tilde{v}$ such that $|\tilde{c}_i - c_i| < \delta$ and $|\tilde{v} - v| < \delta$, can we (efficiently) compute a vector w and threshold θ such that $g(x) = \operatorname{sign}\langle w, x\rangle - \theta$ satisfies $\mathsf{dist}(f, g) < \epsilon$? Or, setting $\epsilon, \delta = 0$, can we exactly compute a function g such that $\mathsf{dist}(f, g) = 0$ when given c, v?*

Recently, a pair of papers have given efficient algorithms for the inexact version of this problem; we will review the most recent of these.

6.1 A Solution for the Hypercube

Two recent papers have presented solutions to the approximate CPP on the boolean hypercube: an earlier work by O'Donnell and Servedio [OS11] and a later work by De *et al.* [DDFS14] that offers an improvement; I will focus on the latter. The main theorem of that work is:

Theorem 6.1.1 ([DDFS14] Theorem 1)**.** *There exists a randomized algorithm A that, for any halfspace $f : \{\pm 1\}^n \to \{\pm 1\}$ with Chow parameters*

$$\hat{f} = (\mathbb{E}[f], \mathbb{E}[x_1 f(x)], \ldots, \mathbb{E}[x_n f(x)])$$

and any $\epsilon > 0$, satisfies the following:

1. *Given $\alpha \in \mathbb{R}^{n+1}$ such that $\left\|\hat{f} - \alpha\right\|_2 \leq \kappa(\epsilon)$, A produces a halfspace h (defined by a vector w and threshold θ) such that $\mathsf{dist}(f, h) \leq \epsilon$ with probability at least $1 - \delta$, and*
2. *A runs in time $\tilde{O}\left(n^2 \cdot \operatorname{poly} 1/\kappa(\epsilon)\right) \cdot \log(1/\delta)$*

where $\kappa(\epsilon) = 2^{-O\left(\log^3(1/\epsilon)\right)}$.

This algorithm makes use of a fact that relates the Chow parameters to the distance between two functions: this is essentially an analogue of the Gap Theorem for the special case of the boolean hypercube; I will say more about theorems like this in Chapter 7.

Theorem 7.5.13: Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a halfspace with Chow parameters $\hat{f}$ and let $g : \{\pm 1\}^n \to \{\pm 1\}$ be any function, with Chow parameters $\hat{g}$. For all $\epsilon > 0$, if $\left\|\hat{f} - \hat{g}\right\|_2 < \epsilon$, then

$$\mathsf{dist}(f, g) \leq 2^{-\Omega\left(\log^{1/3}(1/\epsilon)\right)}.$$

The algorithm will work by constructing a *linear bounded function* as an intermediate step:

Definition 6.1.2 (Linear Bounded Function)**.** Define the *truncation function* as:

$$\text{trunc}(x) := \begin{cases} \text{sign}(x) & \text{if } |x| \geq 1 \\ x & \text{if } |x| < 1\,. \end{cases}$$

Then a *linear bounded function (LBF)* is a function of the form

$$f(x) = \text{trunc}(\langle w, x\rangle - \theta)$$

for some $w \in \mathbb{R}^n, \theta \in \mathbb{R}$.

In the proof below, we will frequently use the notation $\hat{f} = (\mathbb{E}[f], \mathbb{E}[x_1 f(x)], \ldots, \mathbb{E}[x_n f(x)])$ to refer to the $(n+1)$-dimensional vector of Chow parameters, and we will write $\hat{f}(i)$ to be the i^{th} parameter. Since there are $n+1$ Chow parameters but our points have dimension n, we will use the slightly nonstandard notation

$$\left\langle \hat{f}, x \right\rangle = \hat{f}(0) + \sum_{i=1}^{n} \hat{f}(i) x_i$$

when $x \in \mathbb{R}^n$.

Theorem 6.1.3 ([DDFS14], Theorem 10)**.** *There exists a randomized algorithm* ChowReconstruct *that, for every function $f : \{\pm 1\}^n \to \{\pm 1\}$, every $\epsilon, \delta > 0$, and every vector $\alpha \in \mathbb{R}^{n+1}$ satisfying $\left\|\alpha - \hat{f}\right\| \leq \epsilon$, has the following properties:*

1. ChowReconstruct *produces an LBF g such that, with probability at least $1 - \delta$, $\left\|\hat{f} - \hat{g}\right\| \leq 6\epsilon$,*
2. *the weight vector $w \in \mathbb{R}^{n+1}$ defining g satisfies $w = \kappa v$ for some $\kappa \in \mathbb{R}$ and v is a vector of integers with $\|v\| = O(\sqrt{n}/\epsilon^3)$,*
3. ChowReconstruct *runs in time $\tilde{O}\left(\frac{n^3}{\epsilon^4}\right)\log(1/\delta)$ and uses*

$$O\left(\frac{n^2}{\epsilon^4}\log\left(\frac{n+1}{\delta\epsilon}\right)\right)$$

random samples.

The algorithm and its proof follow a simple intuition: we can start with the (approximate) Chow parameters $\alpha \in \mathbb{R}^{n+1}$, and use these as our initial weight vector. Since the center of

mass and the normal vector are not necessarily parallel, this initial weight vector may not give us the function we want. But we can try to iteratively improve our guess by adjusting the weights using the difference between the current Chow parameter vector $\hat{g}_t$ and the target vector α. We will show that this process will in fact converge to a suitably close function.

$$\mathbb{E}\left[x_i(\left\langle \hat{f}, x \right\rangle + v)\right] = \sum_{j\in[n]} \hat{f}(j)\mathbb{E}\left[x_i x_j\right] + \mathbb{E}\left[x_i\right] v = \hat{f}(i)\,.$$

Our goal is to generalize the result to probability distributions other than the hypercube. With this in mind, observe that the following proof of correctnes of the ChowReconstruct algorithm does not depend on the hypercube in any meaningful way.

For notational simplicity, if we have a 0-indexed weight vector $w \in \mathbb{R}^{n+1}$ and a point $x \in \mathbb{R}^n$, we will write $\langle w, x\rangle = \sum_{i\in[n]} w_i x_i$, leaving the coordinate w_0 out of the sum.

Algorithm 7 ChowReconstruct

Input: $\alpha \in \mathbb{R}^{n+1}$ satisfying $\left\|\hat{f} - \alpha\right\| \le \epsilon$, $\epsilon > 0, \delta > 0$

function ChowReconstruct(α, ϵ, δ)
    Define $g_0(x) := 0, g_0' := 0$.
    for $t \ge 0$ **do**
        $\forall 0 \le i \le n$, let $\tilde{g}_t(i)$ be an estimate of $\hat{g}_t(i)$ with accuracy $\pm\frac{\epsilon}{4\sqrt{n+1}}$ and confidence $1 - C\cdot\frac{\epsilon}{n+1}$.
        if $\|\tilde{g}_t - \alpha\| \le 4\epsilon$ **then return** g_t.
        else
            Define $h_t(x) := \langle(\alpha - \tilde{g}_t), x\rangle$.
            Define $g_{t+1}' := g_t' + \frac{1}{2}h_t$.
            Define $g_{t+1} := \text{trunc}(g_{t+1}')$.

Proof. For simplicity, we will leave out the proof of part 2.

Termination: Suppose that ChowReconstruct has terminated on some iteration t, so $\|\tilde{g}_t - \alpha\| \le 4\epsilon$. Then

$$\left\|\hat{f} - \hat{g}_t\right\| \le \left\|\hat{f} - \alpha\right\| + \|\tilde{g}_t - \alpha\| + \|\hat{g}_t - \tilde{g}_t\| \le \epsilon + 4\epsilon + \sqrt{(n+1)\frac{\epsilon^2}{4^2(n+1)}} \le 6\epsilon$$

Note that we are being somewhat inexact with the final inequality; this is because we have ignored the requirement of finding an integer vector for part 2.

Potential Function: Before defining the potential function that records our progress, we will make an observation: if we assume that the parameters of our estimate $\hat{g}_t$ are still, say, ϵ-far from the target parameters $\hat{f}$, then

$$\mathbb{E}\left[(f - g_t)\left((\hat{f}(0) - \hat{g}_t(0)) + \sum_{i=1}^{n}(\hat{f}(i) - \hat{g}_t(i))x_i\right)\right]$$
$$= (\hat{f}(0) - \hat{g}_t(0))^2 + \sum_{i=1}^{n}(\hat{f}(i) - \hat{g}_t(i))^2 = \left\|\hat{f} - \hat{g}_t\right\|_2^2$$

is at least ϵ. If we can define a potential function that depends on this, then we could show that the potential decreases by roughtly ϵ in each iteration, proving convergence. Using our approximations instead, we will show the slightly worse inequality

$$\mathbb{E}\left[(f - g_t)h_t\right] \geq \rho\left(\rho - \frac{3}{2}\epsilon\right) \tag{6.1}$$

where we define

$$\rho := \|\alpha - \hat{g}_t\| \ .$$

By linearity of expectation, we easily have

$$\begin{aligned}\mathbb{E}\left[(f - g_t)h_t\right] &= \sum_{i=0}^{n}\mathbb{E}\left[(f(x) - g_t(x))(\alpha_i - \tilde{g}_t(i))x_i\right] \\ &= \sum_{i=0}^{n}\mathbb{E}\left[(x_i f(x) - x_i g_t(x))(\alpha_i - \tilde{g}_t(i))\right] \\ &= \sum_{i=0}^{n}(\hat{f}(i) - \hat{g}_t(i))(\alpha_i - \tilde{g}_t(i))\end{aligned}$$

We want ρ to appear, so we need $(\alpha_i - \tilde{g}_t(i))^2$; we get this by adding and subtracting $(\alpha_i - \tilde{g}_t(i))$ into the left factor:

$$\begin{aligned}\mathbb{E}\left[(f - g_t)h_t\right] &= \sum_{i=0}^{n}(\hat{f}(i) - \hat{g}_t(i) + (\alpha_i - \tilde{g}_t(i)) - (\alpha_i - \tilde{g}_t(i)))(\alpha_i - \tilde{g}_t(i)) \\ &= \sum_{i=0}^{n}(\hat{f}(i) - \alpha_i + \tilde{g}_t(i) - g_t(i))(\alpha_i - \tilde{g}_t(i)) + (\alpha_i - \tilde{g}_t(i))^2 \\ &= \sum_{i=0}^{n}(\hat{f}(i) - \alpha_i)(\alpha_i - \tilde{g}_t(i)) + (\tilde{g}_t(i) - g_t(i))(\alpha_i - \tilde{g}_t(i)) + \rho^2\end{aligned}$$

We will use a trick with the Cauchy-Schwarz inequality. We start with an application of the inequality on the negative of the term (note that the first term is negated):

$$\sum_{i=0}^{n}(\alpha_i - \hat{f}(i))(\alpha_i - \tilde{g}_t(i)) \leq \left\|\hat{f} - \alpha\right\| \cdot \|\alpha - \tilde{g}_t\| \leq \epsilon\rho\,.$$

Then we can negate both sides to get

$$\sum_{i=0}^{n}(\hat{f}(i) - \alpha_i)(\alpha_i - \tilde{g}_t(i)) \geq -\epsilon\rho\,.$$

Using the same trick on the second term gives us

$$\sum_{i=0}^{n}(\tilde{g}_t(i) - \hat{g}_t(i))(\alpha_i - \tilde{g}_t(i)) \geq -\|\tilde{g}_t - \hat{g}_t\| \cdot \|\alpha - \tilde{g}_t\| \geq -\frac{\epsilon}{2}\rho$$

which gives us equation (6.1)

Recall $h_t = 2g'_{t+1} - 2g'_t$. We want to define the potential such that the difference $P(t+1) - P(t)$ to satisfies the following:

1. $P(t+1) - P(t) \leq -C$ for some $C > 0$ independent of the dimension, i.e. the potential decreases at each iteration,
2. $P(t) \geq 0$ for all t, and
3. $P(0) = 1$,

which would prove that the algorithm terminates.

Property 1: $(P(t+1) - P(t) \leq -C)$. To get this property, we can try to use equation (6.1); we would want the potential difference $P(t+1) - P(t)$ to include the (negative) term

$$-\mathbb{E}\left[(f - g_t)h_t\right] = -\mathbb{E}\left[(f - g_t)(2g'_{t+1} - 2g'_t)\right] = \mathbb{E}\left[-2fg'_{t+1} + 2fg'_t + 2g_tg'_{t+1} - 2g_tg'_t\right]$$

but this cannot be split into separate (additive) parts for t and $t+1$. We fix this by filling out missing terms and eliminating the combined term $2g_tg'_{t+1}$:

$$-2fg'_{t+1} + 2g_{t+1}g'_{t+1} + 2fg'_t - 2g_tg'_t + 2g_tg'_{t+1} - 2g_tg'_{t+1}$$

which separates into

$$-(f - g_t)(2g'_{t+1} - 2g'_t) + 2g'_{t+1}(g_{t+1} - g_t) = 2g'_{t+1}(g_{t+1} - f) - 2g'_t(g_t - f)\,.$$

Unfortunately our potential difference would now be

$$P(t+1)-P(t) = \mathbb{E}\left[-(f-g_t)h_t + 2g'_{t+1}(g_{t+1}-g_t)\right] = -\mathbb{E}\left[(f-g_t)h_t\right] + 2\mathbb{E}\left[g'_{t+1}(g_{t+1}-g_t)\right]$$

which may be larger than 0 when, say, $\left|g'_{t+1}\right|$ is very large. We correct this by further adding the term $g_t^2 - g_{t+1}^2$. This gives us the potential difference

$$\begin{aligned} &-2fg'_{t+1} + 2g_{t+1}g'_{t+1} + 2fg'_t - 2g_tg'_t + 2g_tg'_{t+1} - 2g_tg'_{t+1} + g_t^2 - g_{t+1}^2 + f^2 - f^2 \\ &\quad = -(f-g_t)h_t + 2g'_{t+1}(g_{t+1}-g_t) + (g_{t+1}+g_t)(g_{t+1}-g_t) \\ &\quad = -(f-g_t)h_t + (2g'_{t+1}+g_{t+1}+g_t)(g_{t+1}-g_t) \end{aligned} \tag{6.2}$$

which separates into

$$\begin{aligned} &(f+g_{t+1})(f-g_{t+1}) - 2g'_{t+1}(f-g_{t+1}) - (f+g_t)(f-g_t) + 2g'_t(f-g_t) \\ &\quad = (f-g_{t+1})(f+g_{t+1}-2g'_{t+1}) - (f-g_t)(f+g_t-2g'_t) \end{aligned}$$

giving us the potential function

$$P(t) := \mathbb{E}\left[(f-g_t)(f+g_t-2g'_t)\right] \tag{6.3}$$

Now we will show that this potential function decreases in each iteration. Assuming the algorithm has not terminated, we know that $\rho > 4\epsilon$.

The change in the potential at each step is, from equation (6.2),

$$P(t+1) - P(t) = \mathbb{E}\left[(f-g_{t+1})(f-2g'_{t+1}+g_{t+1})\right] - \mathbb{E}\left[(f-g_t)(f-2g'_t+g_t)\right]$$

We want to show the following inequality:

$$(g_{t+1}(x)-g_t(x))(2g'_{t+1}(x)-g_t(x)-g_{t+1}(x)) \le \frac{h_t(x)^2}{2} \tag{6.4}$$

For this, we will use the following easily-verified inequalities for real numbers a, b:

$$\begin{aligned} |\mathrm{trunc}(a) - \mathrm{trunc}(b)| &\le |a-b| \\ |a - \mathrm{trunc}(b)| &\le |a-b| \text{ when } |b| \ge 1 \ge |a| \\ |a - \mathrm{trunc}(a)| &\le |a-b| \text{ when } |a| \ge 1 \ge |b| \end{aligned} \tag{6.5}$$

If $\left|g'_{t+1}(x)\right|, |g'_t(x)| \ge 1$ and have the same sign, the left term is 0. Suppose $\left|g'_{t+1}(x)\right|, |g'_t(x)| \ge 1$ with opposite signs, so $g_{t+1}(x) - g_t(x) = \pm 2$. Then

$$\left|2g'_{t+1}(x) - g_{t+1}(x) - g_t(x)\right| = 2\left|g'_{t+1}(x)\right| \le 2\left|g'_{t+1}(x) - g'_t(x)\right| = |h_t(x)|$$

and

$$|g_{t+1}(x) - g_t(x)| = 2 \leq \left|g'_{t+1}(x) - g_t(x)\right| \leq \frac{1}{2}\left|h_t(x)\right|$$

so the inequality holds for this case. Now we need to consider only the case where (1): $\left|g'_{t+1}(x)\right| < 1$, or (2): $|g'_t(x)| < 1$. First observe that

$$|g_{t+1}(x) - g_t(x)| = \left|\text{trunc}(g'_{t+1}(x)) - \text{trunc}(g'_t(x))\right| \leq \left|g'_{t+1}(x) - g'_t(x)\right| = \left|h_t(x)\right|/2$$

by equation 6.5

If both (1) and (2) occur, then

$$\left|2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)\right| = \left|2g'_{t+1}(x) - g'_t(x) - g'_{t+1}(x)\right| = \left|g'_{t+1}(x) - g'_t(x)\right| = \frac{1}{2}\left|h_t(x)\right|$$

so we are done. If only (1) occurs, then $g_{t+1}(x) = g'_{t+1}(x)$ so

$$\left|2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)\right| = \left|g'_{t+1}(x) - g_t(x)\right| \leq \left|g'_{t+1}(x) - g'_t(x)\right| = \frac{1}{2}\left|h_t(x)\right|$$

by equation (6.5). Finally, if only (2) occurs, then $g_t(x) = g'_t(x)$ so

$$\begin{aligned}\left|2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)\right| &\leq \left|g'_{t+1}(x) - g'_t(x)\right| + \left|g'_{t+1}(x) - g_{t+1}(x)\right| \\ &= \frac{1}{2}\left|h_t(x)\right| + \left|g'_{t+1}(x) - g'_t(x)\right| \\ &\leq \left|h_t(x)\right|\end{aligned}$$

by the triangle inequality and equation (6.5). This proves the inequality (6.4). Using this inequality, we have

$$\mathbb{E}\left[(g_{t+1} - g_t)(2g'_{t+1} - g_t - g_{t+1})\right] \leq \frac{1}{2}\mathbb{E}\left[h_t^2\right]$$

and

$$\mathbb{E}\left[h_t^2\right] = \sum_{i,j=1}^{n}(\alpha(i) - \tilde{g}_t(i))(\alpha(j) - \tilde{g}_t(j))\mathbb{E}\left[x_i x_j\right] + (\alpha(0) - \tilde{g}_t(0))^2 = \sum_{i=0}^{n}(\alpha(i) - \tilde{g}_t(i))^2 = \rho^2$$

(where in first inequality we have used that $\mathbb{E}\left[x_i\right] = \mathbb{E}\left[x_j\right] = 0$ to isolate the term $(\alpha(0) - \tilde{g}_t(0))$). We combine this inequality with the previous inequality (6.1) to get

$$P(t+1) - P(t) \leq -\rho\left(\rho - \frac{3}{2}\epsilon\right) + \frac{\rho^2}{2} = -\frac{\rho^2}{2} + \frac{3}{2}\rho\epsilon \leq -\frac{4}{2}\rho\epsilon + \frac{3}{2}\rho\epsilon = -\frac{1}{2}\epsilon\rho \leq -2\epsilon^2 \quad (6.6)$$

since $\rho > 4\epsilon$ by the termination condition, so the potential decreases by at least $2\epsilon^2$ in every iteration.

Property 2: ($P(t) \geq 0$ for all $t \geq 0$). We must show that $P(t) = \mathbb{E}\left[(f - g_t)(f - 2g_t' + g_t)\right] \geq 0$. This is easy since if $|g_t'(x)| < 1$ then $(f(x)-g(x))(f(x)-2g_t'(x)+g_t(x)) = f(x)^2-g_t'(x)^2 > 0$, and otherwise either $f(x) - g_t(x) = 0$ or $\operatorname{sign}(f(x) - g_t(x)) = \operatorname{sign}(g_t')$.

Property 3: ($P(0) \leq 1$). This is easy to verify since $g_0 = 0$ so

$$P(0) = \mathbb{E}\left[(f - 0)(f + 0)\right] = \mathbb{E}\left[f^2\right] = 1\,.$$

Time and Sample Complexity: Combining property 3 with equation 6.6 and the property that $P(t) \geq 0$ for all t, we can see that the algorithm will terminate after at most $(2\epsilon^2)^{-1}$ iterations.

For each iteration, we must estimate $n + 1$ parameters to accuracy $\eta = \frac{\epsilon}{4\sqrt{n+1}}$, and the failure probability for each estimate must be at most $\delta \cdot 2\epsilon^2/(n+1)$ to ensure that the total failure probability is at most δ (by the union bound). For this we can use the Hoeffding bound. With m samples $X = \{x_i\}_{i\in[m]}$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{x\in X} x_i g(x) - \hat{g}(i)\right| > \eta\right] \leq 2\exp\left(-\frac{m\eta^2}{2}\right) \tag{6.7}$$

and this should be at most $2\delta\epsilon^2/(n+1)$, so we need

$$m = \frac{2}{\eta^2}\log\left(\frac{n+1}{\delta\epsilon^2}\right) = \frac{2(n+1)}{\epsilon^2}\log\left(\frac{n+1}{\delta\epsilon^2}\right)$$

samples for each of the $n + 1$ parameters in each of the $(2\epsilon^2)^{-1}$ iterations. Evaluating a linear threshold function at a point takes $O(n)$ and it must be done $m \cdot (n+1)/(2\epsilon^2)$ times, so the total time complexity is

$$O\left(\frac{n^3}{\epsilon^4}\log\left(\frac{n+1}{\delta\epsilon}\right)\right)\,. \qquad \square$$

It remains to show the main theorem:

Theorem 6.1.1. There exists a randomized algorithm A that, for any halfspace $f : \{\pm1\}^n \to \{\pm1\}$ with Chow parameters $\hat{f} = (\mathbb{E}[f], \mathbb{E}[x_1 f(x)], \ldots, \mathbb{E}[x_n f(x)])$, and any $\epsilon > 0$, satisfies the following:

1. Given $\alpha \in \mathbb{R}^{n+1}$ such that $\left\|\hat{f} - \alpha\right\|_2 \leq \kappa(\epsilon)$, A produces a halfspace h (defined by a vector w and threshold θ) such that $\mathsf{dist}(f, h) \leq \epsilon$ with probability at least $1 - \delta$, and
2. A runs in time $\tilde{O}\left(n^3 \cdot \operatorname{poly} 1/\kappa(\epsilon)\right) \cdot \log(1/\delta)$

where $\kappa(\epsilon) = 2^{-O\left(\log^3(1/\epsilon)\right)}$.

Proof. Assume we have a halfspace f and a vector α satisfying the hypothesis, i.e. $\left\|\hat{f} - \alpha\right\| \leq \kappa(\epsilon)$. Using the ChowReconstruct algorithm with arguments α and $\kappa(\epsilon)$, and Theorem 6.1.3, we obtain an LBF $g(x) = \operatorname{trunc}\left(\langle w, x\rangle\right)$ for some $w \in \mathbb{R}^{n+1}$ such that:

1. $\left\|\hat{f} - \hat{g}\right\| \leq 6\kappa(\epsilon)$, and
2. we have used $\tilde{O}\left(\frac{n^2}{\kappa(\epsilon)^4}\right) \log(1/\delta)$ time steps.

By Theorem 7.5.13, since $\left\|\hat{f} - \hat{g}\right\| \leq 6\kappa(\epsilon)$, we have

$$\mathsf{dist}(f, g) \leq 2^{-O\left(\log^{1/3}(1/\kappa(\epsilon))\right)} \leq \epsilon/2$$

where we have chosen the constant in κ appropriately. Finally, let $h = \operatorname{sign}(\langle w, x\rangle)$ and note that

$$|f(x) - h(x)| = \begin{cases} |f(x) - g(x)| & \text{if } |g(x)| = 1 \\ \leq 2\,|f(x) - g(x)| & \text{if } |g(x)| < 1 \end{cases}$$

so $\mathsf{dist}(f, h) \leq 2\mathsf{dist}(f, g) \leq \epsilon$. □

6.2 An Application of the Gap Theorem

The algorithm as presented by [DDFS14] works for the uniform distribution over the boolean hypercube. But observe that the proof of correctness of the ChowReconstruct algorithm does not depend on this distribution in any meaningful way, so the algorithm's correctness is instantly generalized to any probability distribution over $\mathbb{R}^n$. All that remains is to show an analogue of Theorem 6.1.1 by replacing Theorem 7.5.13 with a version of the Gap Theorem that holds for linear bounded functions rather than linear threshold functions. I do this next:

Theorem 6.2.1. *Let μ be any indiscrete probability measure on $\mathbb{R}^n$. Suppose $h : \mathbb{R}^n \to \{\pm 1\}$ is a halfspace with normal vector w and threshold θ such that $\|w\| = 1$. If $f : \mathbb{R}^n \to [-1, 1]$ is any (measurable) bounded function satisfying $\mathbb{E}[h] = \mathbb{E}[f]$, then*

$$\|\mathsf{Com}(h) - \mathsf{Com}(f)\| \geq \frac{\mathsf{dist}(f,h)}{2\cos(\alpha)} W_\mu\left(w, \frac{\mathsf{dist}(f,h)}{2}\right).$$

Proof. We will prove this theorem by defining a function $g : \mathbb{R}^n \to \{\pm 1\}$ so that the proof for f reduces to the proof for g. We first write

$$\begin{aligned}\|\mathsf{Com}(h) - \mathsf{Com}(f)\| &= \|\mathbb{E}[x(h(x) - f(x))]\| = \frac{1}{\cos\alpha} \cdot \langle w, \mathbb{E}[x(h(x) - f(x))]\rangle \\ &= \frac{1}{\cos\alpha} \cdot \left\langle w, \mu(h^+)\mathbb{E}\left[x(1 - f(x)) \;\middle|\; x \in h^+\right] - \mu(h^-)\mathbb{E}\left[x(f(x) - 1) \;\middle|\; x \in h^-\right]\right\rangle\end{aligned}$$

By continuity, we can define thresholds a_1, a_2 such that

$$\begin{aligned}\mu\{x : \langle w, x\rangle \in (\theta, a_1)\} &= \int_{h^+} \frac{1 - f(x)}{2}\mu(dx) \\ \mu\{x : \langle w, x\rangle \in (a_2, \theta)\} &= \int_{h^-} \frac{1 + f(x)}{2}\mu(dx)\end{aligned}$$

Define the function

$$g(x) = \begin{cases} 1 & \text{if } a_2 < \langle w, x\rangle < \theta \text{ or } a_1 < \langle w, x\rangle \\ -1 & \text{if } \langle w, x\rangle \leq a_2 \text{ or } \theta \leq \langle w, x\rangle \leq a_1. \end{cases}$$

For the reduction to work, we must verify the following three properties:

1. $\mathbb{E}[|h(x) - g(x)|] = \mathbb{E}[|h(x) - f(x)|]$ (i.e. $\mathsf{dist}(f,h) = \mathsf{dist}(g,h)$), and
2. $\mathbb{E}[h] = \mathbb{E}[g]$, and
3. $\langle w, \mathbb{E}[x(h(x) - f(x))]\rangle \geq \langle w, \mathbb{E}[x(h(x) - g(x))]\rangle$

Property 1 is easy:

$$\begin{aligned}\mathbb{E}[|h(x) - g(x)|] &= 2\mu(h^+ \cap g^-) + 2\mu(h^- \cap g^+) \\ &= 2\left(\int_{h^+} \frac{(1 - f(x))}{2}\mu(dx) + \int_{h^-} \frac{(1 + f(x))}{2}\mu(dx)\right) \\ &= \int_{h^+} |h(x) - f(x)|\,\mu(dx) + \int_{h^-} |h(x) - f(x)|\,\mu(dx) \\ &= \mathbb{E}[|h(x) = f(x)|]\end{aligned}$$

From the definitions

$$\mu(h^+ \cap g^-) = \frac{1}{2}\int_{h^+}(1 - f(x))\mu(dx) = \frac{1}{2}\mu(h^+) - \frac{1}{2}\int_{h^+} f(x)\mu(dx)$$
$$\mu(h^- \cap g^+) = \frac{1}{2}\int_{h^-}(1 + f(x))\mu(dx) = \frac{1}{2}\mu(h^-) + \frac{1}{2}\int_{h^-} f(x)\mu(dx)$$

we get

$$\begin{aligned}\int_{h^+} f(x)\mu(dx) &= \mu(h^+) - 2\mu(h^+ \cap g^-) = \mu(h^+ \cap g^+) - \mu(h^+ \cap g^-)\\ \int_{h^-} f(x)\mu(dx) &= 2\mu(h^- \cap g^+) - \mu(h^-) = \mu(h^- \cap g^+) - \mu(h^- \cap g^-)\end{aligned} \tag{6.8}$$

Thus we get property 2, since:

$$\begin{aligned}\mathbb{E}[g] &= \mu(g^+) - \mu(g^-) = \mu(g^+ \cap h^+) + \mu(g^+ \cap h^-) - \mu(g^- \cap h^+) - \mu(g^- \cap h^-)\\ &= \int f(x)\mu(dx) = \mathbb{E}[f]\end{aligned}$$

For property 3, we will look separately at h^+ and h^- and take the difference between the two inner products:

$$\begin{aligned}&\mu(h^+)\left\langle w, \mathbb{E}\left[x(1 - f(x)) \;\middle|\; x \in h^+\right]\right\rangle - \mu(h^+)\left\langle w, \mathbb{E}\left[x(1 - g(x)) \;\middle|\; x \in h^+\right]\right\rangle\\ &\quad= \int_{h^+}\langle w, x\rangle (1 - f(x))\mu(dx) - \int_{h^+}\langle w, x\rangle (1 - g(x))\mu(dx)\\ &\quad= \int_{h^+\cap g^-}\langle w, x\rangle (1 - f(x))\mu(dx) + \int_{h^+\cap g^+}\langle w, x\rangle (1 - f(x))\mu(dx) - \int_{h^+\cap g^-} 2\langle w, x\rangle \mu(dx)\end{aligned}$$

Since $2 = (1 - f(x)) + (1 + f(x))$ we have

$$\int_{h^+\cap g^-} 2\langle w, x\rangle \mu(dx) = \int_{h^+\cap g^-}\langle w, x\rangle (1 - f(x))\mu(dx) + \int_{h^+\cap g^-}\langle w, x\rangle (1 + f(x))\mu(dx)$$

and our expression becomes

$$\begin{aligned}&\int_{h^+\cap g^+}\langle w, x\rangle (1 - f(x))\mu(dx) - \int_{h^+\cap g^-}\langle w, x\rangle (1 + f(x))\mu(dx)\\ &\qquad\geq a_1\left(\int_{h^+\cap g^+}(1 - f(x))\mu(dx) - \int_{h^+\cap g^-}(1 + f(x))\mu(dx)\right)\\ &\qquad= a_1\left(\mu(h^+ \cap g^+) - \mu(h^+ \cap g^-) - \int_{h^+} f(x)\mu(dx)\right)\end{aligned}$$

which is 0 by equation (6.8). A similar proof shows the same for h^- using a_2. This proves the theorem. □

To get the query complexity of the ChowReconstruct algorithm, we need to know how many queries are required to estimate the values $\hat{f}(i) = \mathbb{E}[x_i f(x)]$ for each i (as well as $\mathbb{E}[f]$). Since the class of indiscrete spaces is very general, it isn't useful to prove a bound on the query complexity for all these spaces at once. Instead, I will simply show a template, into which more specific concentration inequalities can be inserted to get bounds on the complexity:

Theorem 6.2.2. *Let μ be any indiscrete distribution over $\mathbb{R}^n$. The* ChowReconstruct *algorithm satisfies the following: for every function $f : \mathbb{R}^n \to \{\pm 1\}$, every $\epsilon, \delta > 0$, and every vector $\alpha \in \mathbb{R}^{n+1}$ satisfying $\left\|\alpha - \hat{f}\right\| \leq \epsilon$, if the values $\mathbb{E}[f]$ and $\mathbb{E}[x_i f(x)]$ can be estimated to accuracy $\pm\frac{\epsilon}{4\sqrt{n+1}}$ and confidence $O(\epsilon/(n+1))$ using at most $m(\epsilon, \delta)$ queries, then:*

1. ChowRecontsruct *produces an LBF g such that, with probability at least $1 - \delta$, $\left\|\hat{f} - \hat{g}\right\| \leq 6\epsilon$,*
2. ChowReconstruct *runs in time $O\left(\frac{n^2}{\epsilon^2} m(\epsilon, \delta)\right)$ and uses*

$$O\left(\frac{n}{\epsilon^2} \cdot m(\epsilon, \delta)\right)$$

random samples.

Proof. The proof of Theorem 6.1.3 depends on the distribution only when providing guarantees for the estimates of $\hat{g}_t$. Thus we merely replace equation (6.7) with any suitable concentration inequality, giving us a requirement of $m(\epsilon, \delta)$ samples for each estimate by assumption. □

Using the Gap Theorem for bounded functions, Theorem 6.2.1, we can then make the final step similar to [DDFS14]. The Gap Theorem applies to the distance between a function and a halfspace with the same volume, so for now we have to restrict the result to balanced functions on symmetric distributions, for which the threshold (0) and volume (0) are known:

Theorem 6.2.3. *There exists a randomized algorithm A that, for any symmetric probability distribution μ over $\mathbb{R}^n$, any balanced halfspace $f : \{\pm 1\}^n \to \{\pm 1\}$ with Chow parameters*

$$\hat{f} = (\mathbb{E}[f], \mathbb{E}[x_1 f(x)], \ldots, \mathbb{E}[x_n f(x)])$$

and any $\epsilon > 0$, satisfies the following:

1. *Given $\alpha \in \mathbb{R}^{n+1}$ such that $\left\|\hat{f} - \alpha\right\|_2 \leq \kappa(\epsilon)$, A produces a halfspace h (defined by a vector w and threshold θ) such that $\mathsf{dist}(f, h) \leq \epsilon$ with probability at least $1 - \delta$, and*
2. *A runs in time $\tilde{O}\left(\frac{n^2}{\kappa(\epsilon)^4} \cdot m(\kappa(\epsilon), \delta)\right)$,*

where $\kappa(\epsilon) = \epsilon \cdot W(\epsilon/2)/12$.

Proof. Following the same reasoning as the proof of Theorem 6.1.1, we obtain from ChowReconstruct an LBF g satisfying $\left\|\hat{f} - \hat{g}\right\| \leq 6\kappa(\epsilon)$. Set $\kappa(\epsilon) = \epsilon \cdot \frac{W(\epsilon/2)}{12}$ and assume for contradiction that $\mathsf{dist}(f, g) > \epsilon/2$. By the Gap Theorem, we have

$$W(\epsilon/2) \cdot \mathsf{dist}(f, g) \leq W(\mathsf{dist}(f, g)/2) \cdot \mathsf{dist}(f, g) \leq \left\|\hat{f} - \hat{g}\right\| \leq 6\kappa(\epsilon)\,.$$

Dividing both sides by $W(\epsilon/2)$ we get an upper bound of $\epsilon/2$, which is a contradiction, so $\mathsf{dist}(f, g) < \epsilon/2$. Following Theorem 6.1.1 we get a halfspace h with $\mathsf{dist}(f, h) \leq \epsilon$. □

This theorem is essentially a preliminary theorem. With an improved Gap Theorem that tolerates differences in volumes, we will be able to get the general version of this theorem as well. This repeats Question 4.4.2: Is there a version of the Gap Theorem that depends on $\mathbb{E}[f] - \mathbb{E}[h]$ instead of requiring that this difference is 0?

Chapter 7

The Boolean Hypercube

Any algorithm for testing halfspaces on arbitrary distributions obviously must generalize an algorithm for testing halfspaces on the boolean hypercube. Finding this generalization is left, in this thesis, as future work. This future work might be helped along significantly by the existing literature on halfspaces over the hypercube, which is the topic for this chapter; I will survey several modern works on halfspaces with the purpose of identifying facts and techniques that might be useful for designing a general halfspace tester.

A secondary purpose of this chapter is to show that, despite intensive study, halfspaces over the hypercube lack a coherent theory: we know many facts, but the relationships between these facts are known only at a very informal level, for example "Noise stable halfspaces should be more regular than sensitive ones". There are many questions to answer which are valid directions for future work. Of particular interest are those related to the centers of mass and the width, which would help us apply the Gap Theorem.

I will begin in Section 7.1 by exploring the idea of *regularity*, a concept that is ubiquitous in the literature, along with the related *critical index method* which has been the backbone of many recent works. Section 7.2 covers work on noise sensitivity, such as the famous "Majority is Stablest" theorem of [MOO05]. Next, Section 7.3 discusses some relationships between halfspaces and *juntas*, a very important concept in the study of boolean functions. In Section 7.4 I briefly describe some work on approximating halfspaces using halfspaces with integer weights. To conclude the chapter, in Section 7.5 I discuss some bounds on the sum-of-squares of first-degree Fourier coefficients (interpreted as the center-norm), and a few theorems with the same flavor as the Gap Theorem that relate the center-norm to the distance between two functions. Finally, I briefly describe some initial relationships between my concept of width (Definition 4.3.3) and the existing literature in Section 7.6.

For this chapter, we will be considering functions of the form $f : \{\pm 1\}^n \to \{\pm 1\}$, and unless otherwise noted, we assume the uniform distribution over $\{\pm 1\}^n$ when we are talking about a probability space. A few obvious differences between this setting and the continuous settings we have been considering in the prior chapters are that there are now a finite number of possible halfspaces, and that two different weight vectors might produce the same function; these new properties provoke a few new questions that we will briefly touch upon.

7.1 Regularity, Anticoncentration, and Critical Indices

Many recent papers on halfspaces have used the *critical index method*, a method for separating halfspaces into different cases based on their *regularity*. The power of this method comes from the use of central limit theorems and anticoncentration to simplify the analysis.

7.1.1 Regularity and Central Limit Theorems

The main tool used in recent advancements in the theory of the hypercube is *regularity*. Regularity is a property of vectors of real numbers:

Definition 7.1.1 (Regularity)**.** Let $w \in \mathbb{R}^n$ be any vector, and let $0 \leq \tau \leq 1$. We say w is τ-regular if

$$|w_i| \leq \tau \cdot \|w\|_2$$

for all $i \in [n]$.

Essentially, the idea is that we want a convenient way to distinguish between those vectors that have a lot of variation in their entries and those that don't. To get a sense for why we would be interested in this, we can keep in mind the example of the weight vector w for a halfspace: if there are a few large weights, then these weights will dominate the value of the linear form $\langle w, x \rangle$ and nearly dictate the function value. For concreteness, compare the two extreme cases: first, where $w_1 = 1$ and $w_i = 0$ for all $i > 1$, which is not τ-regular for any $\tau < 1$ and produces the dictator function $f(x) = x_1$; second, the case where $w_1 = w_2 = \ldots w_n = 1/\sqrt{n}$ which is maximally regular (i.e. $1/\sqrt{n}$-regular) and produces the majority function; for convenience, we will write

$$\textsc{Maj}_n(x) := \text{sign}\left(\left\langle \frac{1}{\sqrt{n}}\vec{1}, x \right\rangle\right)$$

for the majority function.

There are two vectors of interest when talking about halfspaces: the weight vector and the center of mass. Thus there are two notions of regularity that we must keep in mind: regularity of the weight vector, which we call "weight-regularity", and regularity of the center of mass, which we call "Fourier-regularity" due to the equivalence between the center of mass and the first-degree Fourier coefficients. Formally:

Definition 7.1.2 (Weight– and Fourier–regularity)**.** Let $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ and $\tau \in [0,1]$. We say f is τ-weight-regular if $|w_i| \leq \tau \cdot \|w\|$ for all $i \in [n]$.

Now let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any boolean function. We say f is τ-Fourier-regular if $\left|\hat{f}(i)\right| = |\mathbb{E}[xf(x)]| \leq \tau$ for all $i \in [n]$. (Note that we do not quite use the above definition of regularity in this case, since it may not be the case that $\sum_{i\in[n]} \hat{f}(i)^2 = 1$.)

We have arrived at the first new instance of Question 6.0.3:

Question 7.1.3. *How closely related are the notions of weight– and Fourier–regularity?*

We will see some attempts to answer this question later (Theorems 7.1.8, 7.1.9), but first I will give some reasons for why this definition of regularity is useful.

As we will see, regularity is important for the study of the hypercube because of its connection with the Central Limit Theorem; in particular, a few variations on the Berry-Esséen theorem demonstrate the usefulness of regularity:

Theorem 7.1.4 (Berry-Esséen Theorem; [OS11] Theorem 2.7, [DJS+14] theorem 2.2)**.** *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbb{E}[X_i] = 0$ for all i, $\sqrt{\sum_{i\in[n]} \mathbb{E}[X_i^2]} = \sigma$, and $\sum_{i\in[n]} \mathbb{E}\left[|X_i|^3\right] = \rho$. Write $S = \frac{1}{\sigma}\sum_{i\in[n]} X_i$ and $F(t) = \mathbb{P}[S \leq t]$ as the cumulative distribution function of S. Let $\Phi(x)$ be the c.d.f of the standard Gaussian distribution. Then for all $t \in \mathbb{R}$,*

$$|F(t) - \Phi(t)| \leq C\frac{\rho}{\sigma^3}$$

where $C < \frac{1}{2}$ is a universal constant (see [She11]).

Corollary 7.1.5 ([OS11] Corollary 2.8, [DJS+14] Fact 2.4)**.** *Let $x = (x_1, \ldots, x_n)$ be a vector of independent random ± 1 variables and let $w \in \mathbb{R}^n$. Suppose $|w_i| \leq \tau \cdot \|w\|$ for all*

$i \in [n]$. *Then for any interval* $[a, b]$,

$$\left| \mathbb{P}\left[\langle w, x\rangle \in [a, b]\right] - \Phi\left(\left[\frac{a}{\|w\|}, \frac{b}{\|w\|}\right]\right)\right| \leq 2\tau\,.$$

We can also bound the expected values of regular linear forms:

Theorem 7.1.6 ([MORS10] Proposition 32)**.** *Let* $0 < \tau$ *and suppose* $w \in \mathbb{R}^n$ *such that* $|w_i| \leq \tau$ *for all* $i \in [n]$. *For any* $\theta \in \mathbb{R}$,

$$\mathop{\mathbb{E}}_{x\sim\{\pm 1\}^n}\left[|\langle w, x\rangle - \theta|\right] = \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,1)}\left[|\|w\| X - \theta|\right] \pm O(\tau)$$

where X *is a standard Gaussian.*

Another consequence of Berry–Esséen is our first anticoncentration inequality:

Theorem 7.1.7 ([MORS10] Theorem 30, [Ser07] Theorem 2.2)**.** *Let* $w \in \mathbb{R}^n$ *be* τ*-regular, and let* $\lambda \geq \tau$. *Then for any* $\theta \in \mathbb{R}$,

$$\mathbb{P}\left[|\langle w, x\rangle - \theta| \leq \lambda\right] \leq 6\lambda / \|w\|\,.$$

Note that the corollary and the anticoncentration inequality use the definition of regularity. This theorem (or a slight variation) has been used often (for example, [DDFS14, DS13, MORS10, OS11, Ser07, DJS+14]), and we will see some of these applications, starting with the promised theorems relating weight-regularity to Fourier-regularity; the first of these theorems shows that weight-regularity implies Fourier-regularity with only a constant factor loss.

Theorem 7.1.8 ([MORS10] Theorem 38)**.** *Let* $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$ *be a halfspace with* $\|w\| = 1$ *that is* τ*-weight-regular. Then* f *is* $O(\tau)$*-Fourier-regular.*

Proof. Assume without loss of generality that $\tau = |w_1| \geq |w_2| \geq \ldots |w_n|$. We want to find a bound on $\hat{f}(i) = \mathbb{E}\left[xf(x)\right]$ for each i. We will use the fact that for unate functions, $\left|\hat{f}(i)\right| = |\mathrm{Inf}_i(f)|$ (Fact 2.5.11). Consider $\mathrm{Inf}_1(f) = \mathbb{P}\left[f(x^{1\leftarrow -1}) \neq f(x^{1\leftarrow +1})\right]$. Since $w_1 = \tau$, $f(x^{1\leftarrow -1}) \neq f(x^{1\leftarrow +1})$ if and only if $|\sum_{i=2}^n w_i x_i| < \tau$. Let $w' = (w_2, \ldots, w_n)$ with $\|w'\| = \sqrt{1-\tau^2}$. Since $|w_2| \leq |w_1| \leq \tau$, we have $|w_2| \leq \frac{\tau}{\sqrt{1-\tau^2}}\|w'\|$, so w' is

$\tau/\sqrt{1-\tau^2}$-regular. Then by corollary 7.1.5:

$$\begin{aligned}\mathrm{Inf}_1(f) = \mathbb{P}\left[\left|\sum_{i=2}^{n} w_i x_i\right| < \tau\right] &\leq \Phi\left(\left[-\frac{\tau}{\sqrt{1-\tau^2}}, \frac{\tau}{\sqrt{1-\tau^2}}\right]\right) + 2\frac{\tau}{\sqrt{1-\tau^2}} \\ &\leq \frac{\sqrt{2}\tau}{\sqrt{\pi}\sqrt{1-\tau^2}} + 2\frac{\tau}{\sqrt{1-\tau^2}} \\ &= O(\tau)\end{aligned}$$

where in the second inequality, we have used the fact that the density of the Gaussian is at most $1/\sqrt{2\pi}$, and in the last equality we have used $\tau \leq 1/2$, since otherwise if $\tau > 1/2$ we easily have $\mathrm{Inf}_i(f) \leq 1 \leq 2\tau$. □

In the opposite direction we have the following theorem, which shows that poor weight-regularity implies poor Fourier-regularity (in other words, large weights imply large Fourier coefficients). In this direction, the loss depends also upon the bias of the function: this dependence is necessary since we may have a function with very large weights but that is nearly a constant function.

Theorem 7.1.9 ([MORS10] Theorem 39, [KKMO07] Proposition 8)**.** *Let $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$ be a halfspace with $\|w\| = 1$ and let $\tau = \max_i |w_i|$. Then*

1. *For all $0 < \epsilon < 1$, $|\mathbb{E}[f]| \leq 1 - \epsilon$. Then $\max_i \left|\hat{f}(i)\right| \geq \Omega\left(\tau\epsilon^6 \log(1/\epsilon)\right)$.*
2. *For $|\mathbb{E}[f]| = 0, \max_i \left|\hat{f}(i)\right| \geq \Omega(\tau)$.*

Proof. We will show the easy second case, since the first case is much more complicated. In this case, assume that $|w_1| \geq \cdots \geq |w_n|$ (which implies $\left|\hat{f}(1)\right| \geq \cdots \geq \left|\hat{f}(n)\right|$). Let $w_T = (w_2, \ldots, w_n), x_T = (x_2, \ldots, x_n)$. Using the identities

$$\left|\hat{f}(i)\right| = \mathrm{Inf}_i(f) = \mathbb{P}\left[f(x^{1\leftarrow+1}) \neq f(x^{1\leftarrow-1})\right] = \mathbb{P}\left[|\langle w_T, x_T\rangle| < \tau\right]$$

which hold because we may assume the threshold of f is 0, we can use Berry-Esséen on the latter probability to get

$$\mathbb{P}\left[|\langle w_T, x_T\rangle| < \tau\right] = \mathbb{P}\left[|X| < \tau\right] \pm O\left(\tau/\sqrt{1-\tau^2}\right)$$

since w_T is $\frac{\tau}{\sqrt{1-\tau^2}}$-regular ($\|w_T\| = \sqrt{1-\tau^2}$). This probability is at least $\frac{2\tau}{\sqrt{2\pi}}$, which completes the proof of the easy case. □

The authors predict that $\tau\epsilon$ may be the optimal bound for this theorem:

Question 7.1.10. *Can we replace* $\Omega(\tau\epsilon^6 \log(1/\epsilon))$ *with* $\Omega(\tau\epsilon)$ *in the above theorem?*

I will omit the full proof of this theorem since it occupies a substantial portion of the substantial paper; however, I will note that this proof is the first example that we shall see of the "critical index" method.

7.1.2 The Critical Index Method

The "critical index method" is a method of proof that was first used in [Ser07] and has since appeared in many recent works [DDFS14, DJS+14, Dia10, DS13, FGRW12, GOWZ10, MORS10, OS11]. The critical index of a vector is the first index for which the remainder of the vector is regular:

Definition 7.1.11 (Critical Index)**.** Let $w \in \mathbb{R}^n$ be a vector such that $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$, and let $0 < \tau < 1$. The τ-critical index of w is the first index k such that

$$|w_k| \leq \tau \cdot \sqrt{\sum_{i=k}^{n} w_i^2}\,.$$

Proofs using the critical index usually have the same high-level structure: for some threshold K (usually about $K = \tilde{O}(1/\tau^2)$), examine the case where the τ-critical index k is larger than K (the vector is "very irregular"), and the case where k is smaller than K (the vector has a small number of exceptional coordinates). Typically the coordinates are divided into the "head" coordinates $H = \{i : i < \min(K, k)\}$ and "tail" coordinates $T = \{i : i \geq \min(K, k)\}$; the advantage of separating the vectors into groups with either large or small critical index is that, if the critical index is large, the tail must have a very small 2-norm and thus a very small influence over the function value, and if the critical index is small, then we can approximate the tail using a Gaussian variable. The next fact shows formally that the 2-norm of the tail must shrink exponentially with the critical index:

Fact 7.1.12 ([Ser07] Lemma 4.2, [DDFS14] Fact 22)**.** *Let* $w \in \mathbb{R}^n$ *such that* $|w_1| \geq \cdots \geq |w_n|$*, let* $0 < \tau \leq 1$*, and let* k *be the* τ*-critical index of* w*. Write* $\sigma_i = \sqrt{\sum_{j=i}^{n} w_i^2}$*. Then for any* $a < b \in [k]$*,*

$$\sigma_b \leq (1 - \tau^2)^{\frac{b-a}{2}} \sigma_a$$

and in particular, $\sigma_b \leq (1-\tau^2)^{\frac{b-1}{2}} \|w\|$.

Proof. For $b = 1$, the conclusion clearly holds. For $1 < b \leq k$ we have, by definition

$$w_{b-1}^2 = \sigma_{b-1}^2 - \sigma_b^2 > \tau^2 \sigma_{b-1}^2$$

which implies

$$\sigma_b^2 < \sigma_{b-1}^2 - \tau^2 \sigma_{b-1}^2 = (1-\tau^2)\sigma_{b-1}^2 \leq (1-\tau^2)(1-\tau^2)^{b-a-1}\sigma_a^2$$

by induction. □

To give an idea why $K = \tilde{O}(1/\tau^2)$ is commonly used as a threshold, consider the example where $k = 1/\tau^2$. Then we can set $w_1 = w_2 = \cdots = w_{k-1} = 1/\sqrt{k-1} > \tau$, which gives us a vector which satisfies $|w_i| > \tau \geq \tau\sigma_i$ for each $i < k$, so it is by definition "irregular", yet each nonzero coordinate is the same. Pushing the critical index larger than $1/\tau^2$ forces the coordinates to separate from each other. For a further strengthening of this intuition, consider applying the Hoeffding bound (Theorem 2.3.8) to the tail: using $w_T = (w_k, \ldots, w_n)$ to denote the vector of tail coordinates (and similar notation for x_T),

$$\mathbb{P}\left[|\langle w_T, x_T\rangle| \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma_k^2}\right).$$

Thus with a large critical index, σ_k will be very small and $\langle w_T, x_T\rangle$ will be very tightly concentrated around 0. This means that its contribution to the function value will be dominated by the head coordinates; we will formalize this intuition in the sections on junta theorems and integer weights theorems. For now, we will introduce a few anticoncentration inequalities that can be applied to the irregular head of the vector; this complements the concentration of the tail for the following reason: suppose we truncate the function $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ to the only the head variables, producing the function $g(x) = \text{sign}\left(\sum_{i\in H} w_i x_i - \theta\right)$. How close is g to f?

Writing $f(x) = \text{sign}\left(\sum_{i\in H} w_i x_i + \sum_{i\in T} w_i x_i - \theta\right)$, we can see that if $\left|\sum_{i\in H} w_i x_i - \theta\right| > \left|\sum_{i\in T} w_i x_i\right|$, then the variables in T do not contribute to the function value and $g(x) = f(x)$. Thus, taking the contrapositive, we can get a bound on the probability that $f(x) \neq g(x)$: for all $\eta \in \mathbb{R}$,

$$\mathbb{P}\left[f(x) \neq g(x)\right] \leq \mathbb{P}\left[\left|\sum_{i\in H} w_i x_i - \theta\right| < \eta\right] + \mathbb{P}\left[\left|\sum_{i\in T} w_i x_i\right| > \eta\right] \tag{7.1}$$

We have already seen that we can apply the Hoeffding bound to the second term, since it has small 2-norm. But the head variables H are not regular, so we will need more versatile anticoncentration inequalities to deal with this term.

7.1.3 Anticoncentration Inequalities

First we recall the Lévy anticoncentration function, which I used in Chapter 4 to define the width:

Definition 4.3.1: Let $w \in \mathbb{R}^n$ be an arbitrary weight vector and let $r \in \mathbb{R}_+$ be some radius. Let μ be some probability distribution over $\mathbb{R}^n$. (In this chapter, we will assume μ is the uniform distribution over $\{\pm 1\}^n$.) The Lévy anticoncetration function of w at r is

$$p_r(w) := \sup_{\theta \in \mathbb{R}} \mathop{\mathbb{P}}_{x \sim \mu} [|\langle w, x \rangle - \theta| \leq r] \,.$$

An anticoncentration inequality is an upper bound on this function. The problem of finding bounds on this function (or a number of its generalizations) is called the "Littlewood–Offord problem"; Diakonikolas and Servedio introduced such "Littlewood–Offord theorems" to the critical index method in [DS13]; in particular, they used the following two anticoncentration theorems of Erdős and Halász; the first applies when all coordinates of the vector are large:

Theorem 7.1.13 (Erdős [Erd45])**.** *Let $w \in \mathbb{R}^n$ and $r > 0$ such that $|w_i| \geq r$ for all $i \in [n]$. Then*

$$p_r(w) \leq \frac{1}{2^n} \binom{n}{\lceil n/2 \rceil} = O\left(\frac{1}{\sqrt{n}}\right) .$$

The second, stronger theorem applies when the gaps between each coordinate are large:

Theorem 7.1.14 (Halász [Hal77])**.** *Let $w \in \mathbb{R}^n$ and $r > 0$ such that for all $i \neq j \in [n], |w_i - w_j| \geq r$. Then*

$$p_r(w) \leq O\left(\frac{1}{n^{3/2}}\right) .$$

We will explore the application of these theorems in more detail in the section on integer weight theorems (see Theorems 7.4.3 and 7.4.4). Note, for now, that these theorems cannot be applied to our case analysis of the critical index quite yet; we don't have the required guarantees on the structure of the head weights. To simplify the application of these theorems, Diakonikolas and Servedio prove the following anticoncentration extension, which will let us extract a subsequence of head variables that satisfies the requirements of the anticoncentration inequalities:

Lemma 7.1.15 ([DS13] Lemma 21)**.** *Let $w \in \mathbb{R}^n$ and $r \in \mathbb{R}_+$. Suppose $k \leq n$ and let $H = [k], T = [n] \setminus H$. Then*

$$p_r(w) \leq p_r(w_H) \,.$$

Proof. Fix any $x_{k+1}, \ldots, x_n$ and let $\alpha = \sum_{i=k+1}^{n} w_i x_i$. Since each x_i is independent, for any $\theta \in \mathbb{R}$ we have

$$\mathop{\mathbb{P}}_{x}\left[|\langle w_H, x_H\rangle + \langle w_T, x_T\rangle - \theta| \le r \quad | \quad \langle w_T, x_T\rangle = \alpha\right] = \mathop{\mathbb{P}}_{x_H}\left[|\langle w_H, x_H\rangle - (\theta - \alpha)| \le r\right] \le p_r(w_H)\,.$$

Therefore,

$$\begin{aligned}\mathbb{P}\left[|\langle w, x\rangle - \theta| \le r\right] &= \mathop{\mathbb{E}}_{x_T}\left[\mathbb{P}\left[|\langle w_H, x_H\rangle + \langle w_T, x_T\rangle - \theta| \le r \quad | \quad \langle w_T, x_T\rangle\right]\right] \\ &\le \mathop{\mathbb{E}}_{x_T}\left[p_r(w_H)\right] = p_r(w_H)\end{aligned}$$

□

Complementing Erdős' theorem, there is a lemma of O'Donnell and Servedio which shows that halfspaces can be adjusted slightly to get a lower-bound on the weights:

Lemma 7.1.16 ([OS11] Lemma 5.1)**.** *Let $f(x) = \mathrm{sign}(w_0 + w_1 x_1 + \cdots + w_n x_n)$ and assume $|w_1| \ge \cdots \ge |w_n|$. Let $0 < \epsilon < 1/2$ and $k \in [n]$. Write $\sigma_k = \sqrt{\sum_{i=k}^{n} w_i^2}$. If $\sigma_k > 0$, there exist numbers $v_0, v_1, \ldots, v_{k-1}$ such that $|v_1| \ge \cdots \ge |v_{k-1}| \ge |w_k|$,*

$$g(x) = \mathrm{sign}(v_0 + v_1 x_1 + \cdots + v_{k-1} x_{k-1} + w_k x_k + \ldots w_n x_n)$$

is ϵ-close to f and for $i = 0, \ldots, k-1$,

$$|v_i| \le k^{(k+1)/2} \cdot \sqrt{3\ln(2/\epsilon)} \cdot \sigma_k\,.$$

We get what we want by re-normalizing the weights:

Fact 7.1.17 ([DS13] claim 23)**.** *Let $f(x) = \mathrm{sign}(w_0 + \langle w, x\rangle)$ be any halfspace, $0 < \epsilon < 1/2$, and let $k \in [n]$ such that $0 < \sigma_k := \sqrt{\sum_{i=k}^{n} w_i^2}$. Then there exist numbers $v_0, v_1, \ldots, v_n$ such that $g(x) = \mathrm{sign}(v_0 + \langle v, x\rangle)$ satisfies $\mathsf{dist}(f, g) < \epsilon$,*

$$|v_k| \ge \frac{1}{\sqrt{3k^{k+2}\ln(2/\epsilon)}}$$

and $1 = |v_1| \ge \cdots \ge |v_n|$.

Proof. Since $\sigma_k > 0$ we may apply the lemma to get a set of weights $v'_0, \ldots, v'_{k-1}$ such that $|v'_i| \le |v'_1|$ for all $i \in [k-1]$ and

$$|v'_1| \le \sqrt{3k^{k+1}\ln(2/\epsilon)} \cdot \sigma_k \le \sqrt{3k^{k+1}\ln(2/\epsilon)} \cdot \sqrt{k w_k^2}$$

since $w_i^2 \leq w_k^2$ for $i > k$. Now we normalize the weights so that the largest weight is 1, which means we divide each weight by $|v_1|$, giving us the new weight

$$|v_k| = \frac{|w_k|}{|v'_k|} \geq \frac{|w_k|}{|w_k|\sqrt{3k^{k+2}\ln(2/\epsilon)}} = \frac{1}{\sqrt{3k^{k+2}\ln(2/\epsilon)}}\,. \qquad \square$$

While I won't go through the proof of the lemma in detail, I will mention that the idea is to construct a linear program from f, with variables v_i, whose solution is a set of weights that satisfy all 2^n constraints constructed from the requirement $\langle v, x\rangle + v_0 = \langle w, x\rangle + w + 0$ for each point x; then relax all of the constraints for points where $|\langle w, x\rangle + w_0| \geq \sqrt{3\ln(2/\epsilon)}\cdot\sigma_k$. Because of the relaxation, the resulting weights may produce a function g that differs slightly from f, which is why the theorem allows g to merely approximate f.

In order to use the stronger theorem of Halász, Diakonikolas and Servedio prove an analogue of the above lemma that lets us find gaps between the weights rather than just a lower-bound:

Lemma 7.1.18 ([DS13] Lemma 26)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any linear threshold function that depends on all of its variables. There exists a representation of f using weights that (when sorted) satisfy $1 = |w_1| \geq \cdots \geq |w_n|$, and, letting $\Delta_i = |w_i| - |w_{i+1}|$ for $i \in [n-2]$, the k^{th} largest Δ_i is at least $\frac{1}{(2n+2)^{2k+8}}$.*

Since this lemma produces an exact representation and its proof also relies on the analysis of a linear program (albeit a different one than the previous lemma), one might wonder if a strengthening is possible when only an ϵ-approximation function is required:

Question 7.1.19. *Can this lemma be improved under the relaxation that the new function be only ϵ-close to f?*

7.1.4 Estimating the Regularity

So far, our discussion of regularity has been limited to its utility in structural theorems. But using regularity to divide a problem into a number of cases is also a fruitful algorithmic technique that was used extensively by Matulef *et al.* in the prior work on testing halfspaces with queries [MORS10]. To see this, observe that if an algorithm could find $\|\mathsf{Com}\,(f)\|_\infty = \max_i \left|\hat{f}(i)\right|$, it would know exactly the Fourier-regularity of the function f. Finding $\|\mathsf{Com}\,(f)\|_\infty$ might be infeasible for an efficient algorithm, since estimating

each first-degree Fourier coefficient and taking the maximum would require at least $\Omega(n)$ samples (for constant accuracy); instead, we observe that we can approximately classify the regularity using lower norms:

Fact 7.1.20. *Let $w \in \mathbb{R}^n$ be a τ-regular vector with $\|w\|_2 = 1$, and let $p \geq 2$. Then*

$$\|w\|_p = \left(\sum_i |w_i|^p\right)^{1/p} \leq \left(\tau^{p-2} \sum_i w_i^2\right)^{1/p} = \tau^{1-2/p}.$$

Now suppose w is not τ-regular, so, without loss of generality, $|w_1| > \tau$. Then

$$\|w\|_p = \left(\sum_i |w_i|^p\right)^{1/p} > (\tau^p)^{1/p} = \tau.$$

Thus with $p > 1$ we get close to the ability to distinguish τ-regular vectors from those that are far from τ-regular. Combining this fact with the technique used by [MORS10] to efficiently approximate p-norms of the first-degree Fourier coefficients (Theorem 3.3.6), an algorithm can make case distinctions efficiently using queries. In [MORS10], the algorithm uses estimates the 4-norm for a number of random restrictions to f in order to check the regularity of f on several sub-cubes. The pertinent structural fact is that, since the head variables H dominate the tail variables T, most assignments of the variables in H produce a very regular function on the variables in T:

Fact 7.1.21 ([MORS10] Proposition 63)**.** *Let $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$ be a halfspace and let $H = \{i \in [n] : \mathrm{Inf}_i(f) \geq \tau\}$. Let $\alpha \in (0,1)$; then at least an α fraction of restrictions $\rho : H \to \{\pm 1\}$ produce a function f_ρ that is at least (τ/α)-Fourier-regular.*

Proof. Without loss of generality, assume $|w_1| \geq \ldots \geq |w_n|$ (so $\mathrm{Inf}_1(f) \geq \ldots \mathrm{Inf}_n(f)$ as well, Proposition 5.0.5); this means $H = [k-1]$ where k is the τ-critical index of $\hat{f}$. Note that for any restriction $\rho : H \to \{\pm 1\}$, the function $f_\rho(x) = \mathrm{sign}(\langle w_T, x_T\rangle + \langle w_H, \rho\rangle - \theta)$ has weight vector w_T with $|w_k| \geq \ldots \geq |w_n|$, so its maximum influence is $\max_i \mathrm{Inf}_i(f_\rho) = \mathrm{Inf}_k(f_\rho)$. Then

$$\tau \geq \mathrm{Inf}_k(f) = \mathop{\mathbb{E}}_{\rho \in \{\pm 1\}^H}[\mathrm{Inf}_k(f_\rho)] = \mathop{\mathbb{E}}_{\rho \in \{\pm 1\}^H}\left[\max_i \mathrm{Inf}_i(f_\rho)\right]$$

so by Markov's inequality,

$$\mathop{\mathbb{P}}_{\rho \in \{\pm 1\}^H}\left[\max_i \mathrm{Inf}_i(f_\rho) \geq \tau/\alpha\right] \leq \frac{\alpha}{\tau}\mathop{\mathbb{E}}_{\rho}\left[\max_i \mathrm{Inf}_i(f_\rho)\right] \leq \alpha. \qquad \square$$

7.2 Sensitivity and Stability

Recall the definitions of stability and noise sensitivity:

Definition 2.5.7: For $f : \{\pm 1\}^n \to \{\pm 1\}$, the *stability* of f under noise $\rho \in [-1, 1]$ is

$$\mathrm{Stab}_\rho(f) := \mathop{\mathbb{E}}_{x \sim_\delta y}[f(x)f(y)] = \mathop{\mathbb{P}}_{x \sim_\delta y}[f(x) = f(y)] - \mathop{\mathbb{P}}_{x \sim_\delta y}[f(x) \neq f(y)]$$

where $\delta = \frac{1}{2}(1 - \rho)$ and $x \sim_\delta y$ means that x is drawn uniformly at random from $\{\pm 1\}^n$ and y is δ-correlated with x; i.e. y is the vector obtained from x by flipping each coordinate with probability $\delta = \frac{1}{2}(1 - \rho)$.

Definition 2.5.9: For $f : \{\pm 1\}^n \to \{\pm 1\}$, the *noise sensitivity* of f under noise $\rho \in [0, 1]$ is

$$\mathrm{NS}_\rho(f) := \frac{1}{2}\left(1 - \mathrm{Stab}_{1-2\rho}(f)\right) = \mathop{\mathbb{P}}_{x \sim_\rho y}[f(x) \neq f(y)] \ .$$

As a first example, we can easily compute the stability of the dictator function $f(x) = x_1$:

$$\mathrm{NS}_\rho(f) = \mathop{\mathbb{P}}_{x \sim_\rho y}[f(x) \neq f(y)] = \mathop{\mathbb{P}}_{x \sim_\rho y}[x_1 \neq y_1] = \rho$$

by definition. Before computing the stability of the majority function to compare this with, we will need another extension of the Berry-Esséen theorem that works for 2-dimensional Gaussians:

Theorem 7.2.1 ([DJS+14] Theorem 2.5, [MORS10] Theorem 68)**.** *Let $0 < \tau \leq 1$ and let $w \in \mathbb{R}^n$ be a τ-regular vector. Suppose x, y be drawn from $\{\pm 1\}^n$ be picking x uniformly at random and picking y be flipping each bit of x independently with probability ρ. Then for any intervals $I_1, I_2 \subseteq \mathbb{R}$,*

$$\mathop{\mathbb{P}}_{x \sim_\rho y}[\langle w, x\rangle \in I_1, \langle w, y\rangle \in I_2] = \mathop{\mathbb{P}}_{X \sim_\rho Y}[X \in I_1, Y \in I_2] \pm O(\tau)$$

where X, Y are ρ-correlated standard Gaussians.

Fact 7.2.2 ([O'D14])**.** *Let $\rho \in (0, 1)$. Then*

$$\mathrm{Stab}_\rho(\textsc{Maj}_n) = 1 - \frac{2}{\pi}\arccos\rho + o(1)\,.$$

Proof. Before proving the fact for the boolean function, we will prove it for a 1-dimensional Gaussian instead. Let x, y be ρ-correlated standard Gaussian variables. For a fresh Gaussian z, we can write $y = \rho x + \sqrt{1-\rho^2}z$; to see this, note that

$$\mathbb{E}[xy] = \mathbb{E}\left[\rho x^2 + \sqrt{1-\rho^2}xz\right] = \rho$$

since x, z are independent, and for $v = (\rho, \sqrt{1-\rho^2})$ we see that $y = \langle v, (x,z)\rangle$ which is the 1-dimensional projection of a 2-dimensional Gaussian (x,z) onto the line v; thus y is a standard Gaussian. By the same trick, we can see that $x = \langle e_1, (x,z)\rangle$, so in fact x, y are projections of a 2-dimensional Gaussian onto unit vectors with angle $\arccos\langle e_1, v\rangle = \arccos\rho$. Then $\mathop{\mathbb{P}}_{x,z}[\mathrm{sign}(\langle e_1,(x,z)\rangle) \neq \mathrm{sign}(\langle v,(x,z)\rangle)]$ is the probability that (x,z) falls within the shaded region in figure 7.2, which we can easily see is $2\frac{\arccos\rho}{2\pi} = \frac{1}{\pi}\arccos\rho$ (this is known as Sheppard's formula [O'D14]). Returning to the boolean vectors x, y, the rest of

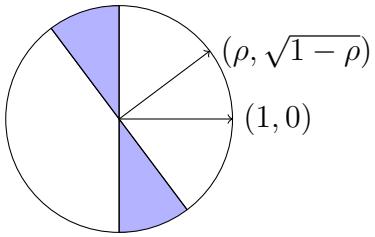


Figure 7.1: An illustration of Sheppard's formula.

the proof follows because the inner products $\langle w, x\rangle, \langle w, y\rangle$ behave like Gaussians for large n; taking $\tau = 1/\sqrt{n}$ and using Theorem 7.2.1, we have

$$\begin{aligned}\mathop{\mathbb{P}}_{x\sim_\rho y}[\mathrm{sign}(\langle w,x\rangle) \neq \mathrm{sign}(\langle w,y\rangle)] &= \mathop{\mathbb{P}}_{X\sim_\rho Y}[\mathrm{sign}(X) \neq \mathrm{sign}(Y)] \pm O\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{\pi}\arccos\rho \pm O\left(\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{7.2}$$

where X, Y are ρ-correlated Gaussians, which completes the proof. □

As we will do several times, we will use these two examples as evidence that the majority and dictator functions are the two extreme cases, and guess that as regularity decrease, stability increases.

Looking at the above calculation, one can see that the same asymptotic behaviour is exhibited by any function that is $\tau(n)$-regular for some $\tau(n) = o(1)$ that shrinks with

n. Khot *et al.*, for their well-known proof that the Goemans–Williamson algorithm for approximating MAX-CUT is optimal (under the Unique Games Conjecture) [KKMO07], conjectured that this simple observation can be improved to depend on the regularity parameter, giving a bound on the stability of a halfspace as a function of its regularity. The conjecture was soon proved by Mossel *et al.* [MOO05]:

Theorem 7.2.3 (Majority is Stablest, [MOO05], see [O'D14] section 11.7)**.** *Let $\rho \in (0,1)$ and let $f : \{\pm 1\}^n \to [-1,1]$ be any bounded function with $\mathbb{E}[f] = 0$ and $\max_i \mathrm{Inf}_i(f) \leq \tau$ for some $0 < \tau \leq 1$. Then*

$$\mathrm{Stab}_\rho(f) \leq 1 - \frac{2}{\pi} \arccos \rho + o_\tau(1)$$

(where $o_\tau(1)$ is a function that is asymptotically 0 with respect to τ).

Keeping in mind that our goal is a theory of halfspaces for arbitrary distributions, it may be worth noting that [MOO05] actually prove the above theorem for a much wider class of discrete distributions than the uniform distribution over $\{\pm 1\}^n$.

Returning to our intuition that the majority and dictator functions are the extreme cases, one might be motivated to make the following conjecture (and maybe to give it a contradictory name):

Conjecture 7.2.4 (Majority is Least Stable [BKS99], see [O'D14])**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any linear threshold function. Then for all $\rho \in [0,1]$,*

$$\mathrm{Stab}_\rho(f) \geq \mathrm{Stab}_\rho(\textsc{Maj}_n)\,.$$

But intuition is wrong in this case! Recently Jain provided a counterexample (and referred to prior unpublished counterexamples by Gopi and Kane):

Example 7.2.5 ([Jai17])**.** Let $f : \{\pm 1\}^5 \to \{\pm 1\}$ be the function

$$f(x) = \mathrm{sign}(2x_1 + 2x_2 + x_3 + x_4 + x_5)\,.$$

This function has $\|\mathsf{Com}\,(f)\|^2 = \frac{44}{64}$ while $\|\mathsf{Com}\,(\textsc{Maj})_5\|^2 = \frac{45}{64}$. This serves as a counterexample since

$$\mathrm{Stab}_\rho(f) = \sum_{S \subseteq [n]} \rho^{|S|} \hat{f}(S)^2 = \mathbb{E}[f] + \rho \|\mathsf{Com}\,(f)\|^2 + O\left(\rho^2\right)$$

so $\mathrm{Stab}_\rho(f) < \mathrm{Stab}_\rho(\textsc{Maj}_5)$ for small enough ρ.

However, this example is extreme in the sense that ρ is at the edge of its domain and the difference in stability is only $1/64$, leaving open the possibility that the conjecture is true asymptotically, or in some other slightly weakened form. We will refer to this example later, in the section on center of mass theorems (see Conjecture 7.5.2).

There are some positive results supporting the intuition that regular halfspaces should be more sensitive to noise; Diakonikolas *et al.*, using the critical index method in their proof of a "junta theorem" that relates noise sensitivity to junta properties (see Theorem 7.3.8), prove that very regular halfspaces cannot be too stable:

Theorem 7.2.6 ([DJS+14] Lemma 3.3)**.** *Let $0 < \tau \leq 1/2$ and let $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ be a halfspace such that w is τ-regular and $|\mathbb{E}[f]| < 1 - \delta$. Then*

$$\text{NS}_\tau(f) = \Omega\left(\delta^{\frac{1}{1-\tau}}\sqrt{\log(1/\delta)} \cdot \sqrt{\tau}\right) - O(\tau) \ .$$

This theorem is not too hard to see in the case that f is unbiased (i.e. $\theta = 0$). Here we just observe that

$$\begin{aligned}\text{NS}_\tau(f) &= \mathop{\mathbb{P}}_{x \sim_\tau y}[\text{sign}(\langle w, x\rangle) \neq \text{sign}(\langle w, y\rangle)] = \mathop{\mathbb{P}}_{X \sim_\tau Y}[\text{sign}(X) \neq \text{sign}(Y)] \pm O(\tau) \\ &= \frac{1}{\pi}\arccos(1 - 2\tau) \pm O(\tau)\end{aligned}$$

as in equation (7.2) and note that $\arccos(a) \geq \sqrt{(1-a)^2 + (1-a^2)} = \sqrt{2(1-a)}$ for any $a \in [-1, 1]$.

7.3 Juntas

Juntas are objects of fundamental importance to the study of boolean functions. The word "junta" means a small (military) group of people who rule a country (especially by means of force). We use the name "dictator" for a function that depends only on a single variable, and we will similarly apply "junta" to boolean functions that depend only on a small number of variables:

Definition 7.3.1 (Junta)**.** We will define juntas for the special case of boolean functions; an analogous definition would hold in more general cases. Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a boolean function and let $i \in [n]$. We will say f depends on coordinate i if there exists $x \in \{\pm 1\}^n$ such that $f(x^{i \leftarrow +1}) \neq f(x^{i \leftarrow -1})$. Then f is a *k-junta* if there are at most k coordinates upon which f depends.

Juntas are important because the complexity of an algorithm operating on a function, say, a property testing algorithm, often depends on the dimension; knowing we have a k-junta can reduce the effective dimension from n to k. In fact, it is usually enough to say that a function f is approximated by a junta: if f itself is not a junta but is very close to a junta g, we can get good enough results by doing computations on (or proving theorems about) g. Since small juntas are much easier to work with, one can see why the problem of testing juntas has been so important to the study of sublinear algorithms; see [Bla09, FKR+02].

The definition of a junta can be rephrased in terms of the influence: if a coordinate i does not affect the function value for any point x, then $\mathrm{Inf}_i(f) = 0$; thus f is a k-junta when at most k coordinates have nonzero influence. Consider some extreme examples: for a dictator function, say $f(x) = x_1$, only one coordinate has nonzero influence, and its influence is $\mathbb{E}[x_1 f(x)] = \mathbb{E}[x_1^2] = 1$, so the total influence $\sum_i \mathrm{Inf}_i(f) = 1$. In the other extreme, the least "dictator-like" or "junta-like" function (that is to say, the most democratic function) is the majority function, for which every coordinate has influence $\Theta(1/\sqrt{n})$ (see Proposition 7.5.1) adding up to a total influence of $\Theta(\sqrt{n})$ (which is in fact the greatest total influence of all unate functions [O'D14]). From these examples, one might suspect that functions with smaller total influence are closer to juntas or dictatorships. This suspicion is confirmed by the famous theorem of Friedgut:

Theorem 7.3.2 (Friedgut's Junta Theorem [Fri98], see [DS13, O'D14])**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a boolean function with total influence $\mathrm{Inf}(f) = \sum_{i \in [n]} \mathrm{Inf}_i(f)$, and let $0 < \epsilon < 1$. Then f is ϵ-close to a $2^{O(\mathrm{Inf}(f)/\epsilon)}$-junta.*

This theorem holds for all boolean functions, and it is interesting to consider special cases of functions for which stronger "junta theorems" might hold. I will survey a few theorems of this kind that pertain to linear threshold functions.

First we see that a direct improvement over Friedgut's theorem can be obtained for halfspaces:

Theorem 7.3.3 ([DS13] Theorem 1)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any linear threshold function, and let $0 < \epsilon < 1$. Then f is ϵ-close to an $\mathrm{Inf}(f)^2 \cdot \mathrm{poly}(1/\epsilon)$-junta.*

This theorem is strengthened by the simple fact if a halfspace is close to a junta, the junta is in fact a halfspace itself, as a consequence of the following well-known fact:

Proposition 7.3.4. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ and $0 < \epsilon < 1$. Suppose f is ϵ-close to a k-junta on the coordinate $H \subseteq [n]$. Denote by (x_H, y_T) the string where coordinate i takes*

value x_i if $i \in H$ and y_i otherwise. Then f is ϵ-close to the function

$$g(x) = \text{sign}\left(\mathop{\mathbb{E}}_{y_T}\left[f(x_H, y_T)\right]\right)$$

which averages over the assignments to the variables in T.

Proof. Assume without loss of generality that $\text{Inf}_1(f) \geq \text{Inf}_2(f) \geq \cdots \geq \text{Inf}_n(f)$, let $H = [k]$, and suppose f is ϵ-close to a junta g that depends on the variables $J \subseteq [n], |J| = k$. If $J \neq H$, we will show that we get a closer approximation to f by replacing a coordinate $i \in J \setminus H$ with any coordinate $j \in H \setminus J$.

Let x_H be any assignment to the coordinates H. If $\mathop{\mathbb{P}}_{x_T}\left[f(x_H, x_T) = 1\right] \geq \mathop{\mathbb{P}}_{x_T}\left[f(x_H, x_T) = -1\right]$ then the probability that $g(x_H, x_T) \neq f(x_H, x_T)$ is minimized when

$$g(x_H, x_T) = 1 = \text{sign}(\mathop{\mathbb{E}}_{x_T}\left[f(x_H, x_T)\right]);$$

the same holds in the opposite case. □

Corollary 7.3.5. *Let $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ be a halfspace and suppose f is ϵ-close to a k-junta on the coordinates H. Then f is ϵ-close to the function* $\text{sign}(\langle w_H, x_H\rangle - \theta)$.

Proof. From the above proposition, we have that f is ϵ-close to the function

$$g(x_H) = \text{sign}(\mathop{\mathbb{E}}_{x_T}\left[\text{sign}(\langle w_H, x_H\rangle + \langle w_T, x_T\rangle - \theta)\right].$$

For any $x_H \in \{\pm 1\}^H$, let $\alpha = \langle w_H, x_H\rangle - \theta$. Suppose $\mathop{\mathbb{E}}_{x_T}\left[\text{sign}(\langle w_T, x_T\rangle + \alpha)\right] \geq 0$ so $g(x_H) = 1$. Then

$$\mathop{\mathbb{P}}_{x_T}\left[\langle w_T, x_T\rangle \geq -\alpha\right] \geq \mathop{\mathbb{P}}_{x_T}\left[\langle w_T, x_T\rangle < -\alpha\right]$$

so by symmetry it must be the case that $-\alpha \leq 0$; thus $\alpha \geq 0$ and $1 = g(x_H) = \text{sign}(\langle w_H, x_H\rangle - \theta)$. By a similar argument, the same holds in the opposite case, so $g(x_H) = \text{sign}(\langle w_H, x_H\rangle - \theta)$. □

The proof of Theorem 7.3.3 relies on a prior theorem that relates juntas to regularity. Recall from the previous section on regularity that the majority function is as regular as possible and the dictator function as irregular as possible, in terms of both weight-regularity and Fourier-regularity, so like above one might suspect that more regular functions are more democratic. And Fourier-regularity is tightly connected with influence since for a

halfspace f, $\left|\hat{f}(i)\right| = \mathrm{Inf}_i(f)$ so the total influence is $\mathrm{Inf}(f) = \sum_i \mathrm{Inf}_i(f) = \|\mathsf{Com}(f)\|_1$, so this further suggests a deep relationship between juntas and regularity. In the paper that introduced the critical index method, Servedio proved something of this form:

Theorem 7.3.6 ([Ser07], case IIa)**.** *Let $0 < \epsilon < 1/2$. There is a threshold $K = \tilde{O}\left(\frac{1}{\epsilon^2}\right)$ such that if f is a halfspace with ϵ-critical index $k \geq K$ then f is ϵ-close to a K-junta.*

The next theorem of O'Donnell and Servedio nicely expands this relationship and works towards answering the question of how closely related the center of mass and the weight vector are:

Theorem 7.3.7 ([OS11] Theorem 7.1)**.** *There is a polynomial $\tau(\epsilon) = \mathrm{poly}(\epsilon)$ such for every $0 < \epsilon < 1/2$ and every halfspace $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$, if we write $H = \{i \in [n] : \left|\hat{f}(i)\right| \geq \tau(\epsilon)^2\}$ and $T = [n] \setminus H$, and assume w is normalized such that $\sum_{i\in T} w_i^2 = 1$, then either*

1. *f is ϵ-close to a junta on the coordinates H, or*
2. *f is ϵ-close to the halfspace*

$$g(x) = \mathrm{sign}\left(\langle w_H, x_H\rangle + \sum_{i\in T} \frac{\hat{f}(i)}{\sigma} x_i - \theta\right)$$

where $\sigma^2 = \sum_{i\in T} \hat{f}(i)^2$.

In the previous section, we saw that regularity and noise sensitivity are related. In this section we have seen that regularity is related to junta properties, so it makes sense to look for a relationship between noise sensitivity and junta properties. Before seeing what is known about this relationship, we can use informal reasoning to predict that since irregular halfspaces are less democratic and are less noise-stable, we should see that less noise-stable halfspaces are less democratic, i.e. closer to juntas. This was established by Diakonikolas *et al.* :

Theorem 7.3.8 ([DJS+14] Theorem 1.3)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any halfspace and let $\epsilon, \delta > 0$ be sufficiently small (i.e. less than some universal constant). If*

$$\mathrm{NS}_\epsilon(f) \leq C\delta^{\frac{2-\epsilon}{1-\epsilon}}\sqrt{\epsilon}$$

then f is ϵ-close to an $O\left(\frac{1}{\epsilon^2}\log(1/\epsilon)\log(1/\delta)\right)$-junta.

The proof of this theorem follows the critical index method, showing for each case that either a contradiction occurs (see Theorem 7.2.6) or the function is close to a junta (see Theorem 7.3.6).

Putting aside regularity for now, let's consider the rest of the Fourier spectrum. We have seen that halfspaces maximize the first-degree Fourier coefficients, which means they minimize the contribution from the higher spectrum. Any k-junta f on the coordinates H is going to have $\hat{f}(S) = 0$ for any set $S \subseteq [n]$ such that $S \cap H \neq S$, since for any coordinate $i \in S \setminus H$ we will have $\mathop{\mathbb{E}}_{x_i}[x_i b] = 0$ for any value $b = f(x)\prod_{j \in S \setminus \{i\}} x_j$. Thus $\sum_{S:|S| \geq k} \hat{f}(S) = 0$. Going in the opposite direction, there is the theorem of Nisan and Szegedy:

Theorem 7.3.9 ([NS94] Theorem 1, see [DS13] Theorem 15)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a function with maximum Fourier degree k. Then f is a $k2^k$-junta.*

This is improved for halfspaces by Diakonikolas and Servedio:

Fact 7.3.10 ([DS13] Proposition 17)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a halfspace with maximum Fourier degree k. Then f is a $(2k-1)$-junta.*

Moving on to approximations by juntas, Bourgain showed that we can soften the condition if we allow some error:

Theorem 7.3.11 ([Bou02], see [DS13] Theorem 16)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a function such that $\sum_{S:|S|>k} \hat{f}(S)^2 \leq \frac{\epsilon}{k}^{1/2-o(1)}$ for some k. Then f is ϵ-close to a $2^O(k) \cdot \mathrm{poly}(1/\epsilon)$-junta.*

Since the restriction to halfspaces of Nisan and Szegedy's theorem results in an exponential improvement, Diakonikolas and Servedio conjecture that a similar exponential improvement over Bourgain's theorem can be achieved:

Conjecture 7.3.12 ([DS13] Conjecture 2)**.** *If f is a halfspace with $\sum_{S:|S|>k} \hat{f}(S)^2 \leq \frac{\epsilon}{k}^{1/2-o(1)}$, is f ϵ-close to a $\mathrm{poly}(k/\epsilon)$-junta?*

7.4 Integer Weights

One way to simplify a halfspace is to reduce the number of variables that it depends on, as we have seen in the previous section. In this section we will see another way to simplify halfspaces, which is to restrict it to having bounded integer weights. This sort of simplification is useful for questions that arise in the discrete setting, such as counting or iteration (see [Ser07] for applications). These theorems are (probably) not as important for our purposes as those in other sections, so I present them merely as examples of the proof methods; in particular, the critical index method and anticoncentration were used by Diakonikolas and Servedio to get a dramatically simplified proof of Theorem 7.4.3 and subsequently to improve it to Theorem 7.4.4 [DS13].

The first theorems on this topic were concerned with exact representations of linear threshold functions; in this case there are examples where the weights must be exponential, and tight asymptotic bounds on the exponent is known:

Theorem 7.4.1 ([MTT61, Hås94], see [Ser07]). *There exist halfspaces $f : \{\pm 1\}^n \to \{\pm 1\}$ such that for any integer weights w, if w is a realization of f then $\max_i |w_i| = 2^{\Theta(n \log n)}$.*

As we've seen before, it is possible to get much better bounds if we look for approximations of f rather than trying to exactly represent f. Before we see the first of these theorems, I will note that the work of Diakonikolas and Servedio on this problem rely on the following lemma that connects anticoncentration to integer weights:

Lemma 7.4.2 ([DS13] Lemma 22). *Let $f(x) = \text{sign}(\langle w, x\rangle - \theta)$ be any halfspace. For any $r > 0$ and $0 < \epsilon < 1/2$, if the Lévy anticoncentration function (definition 4.3.1) satisfies $p_r(w) \leq \epsilon$ then there exists a halfspace g satisfying $\mathsf{dist}(f, g) < 2\epsilon$ with integer weights of magnitude at most $O\left(\max_i |w_i| \cdot \sqrt{n \ln(1/\epsilon)} \cdot r\right)$.*

Now we see the first theorem that bounds the integer weights of an approximator to f:

Theorem 7.4.3 ([Ser07] Theorem 1.1, see [DS13] Theorem 24). *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any halfspace. For any $0 < \epsilon < 1$ there exists a halfspace $g(x) = \text{sign}(\langle w, x\rangle - \theta)$ with integer weights w such that $\mathsf{dist}(f, g) < \epsilon$ and*

$$\|w\|_2^2 \leq n \cdot 2^{\tilde{O}(1/\epsilon^2)}.$$

Proof. The original proof uses the critical index method; however, we will follow the simplified proof of [DS13] that uses anticoncentration. The first step is to use Fact 7.1.17 to get an

ϵ-approximator g with (sorted) weights $1 = |w_1| \geq \cdots \geq |w_n|$ satisfing $|w_k| \geq \frac{1}{\sqrt{3k^{k+2}\ln(2/\epsilon)}}$, where k will be chosen later.

Next we set $r = \frac{1}{\sqrt{3k^{k+2}\ln(2/\epsilon)}}$, and by Erdős' theorem (Theorem 7.1.13) and the extension lemma (Lemma 7.1.15) we have $p_r(w) \leq O\left(1/\sqrt{k}\right)$ which is at most ϵ for $k = \Omega\left(1/\epsilon^2\right)$ (we will assume $1/\epsilon^2 \leq n$).

Finally, Lemma 7.4.2 implies that there is a representation with maximum weight

$$O\left(\frac{\sqrt{n\ln(1/\epsilon)}}{r}\right) = O\left(\sqrt{nk^{k+2}}\ln(1/\epsilon)\right) = O\left(\sqrt{n}2^{\frac{1}{2}\log(k)(k+2)}\ln(1/\epsilon)\right)$$

which is what we want since $k = 1/\epsilon^2$. □

Combining the anticoncentration method with the critical index method, Diakonikolas and Servedio improved the dependence on ϵ:

Theorem 7.4.4 ([DS13] Theorem 2)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any halfspace. For any $0 < \epsilon < 1$ there exists a halfspace $g(x) = \operatorname{sign}(\langle w, x\rangle - \theta)$ with integer weights w such that $\mathsf{dist}(f, g) < \epsilon$ and for all $i \in [n]$,*

$$|w_i| \leq n^{3/2} \cdot 2^{\tilde{O}\left(\frac{1}{\epsilon^{2/3}}\right)}.$$

7.5 Centers of Mass and Distances

In rotationally-invariant spaces, we have an easy way to find what the center-norm is supposed to be; in the hypercube, such an easy method is not readily apparent. This section surveys some of what is known about the centers of mass of halfspaces over the hypercube. Along with this, there have also been a number of theorems relating the difference between Chow vectors (first-degree Fourier coefficients) to the difference between two functions, along the same lines as the Gap Theorem.

7.5.1 Centers of Mass

First, to get some intuition for the centers of mass, we will look at balanced halfspaces whose weight vector is in $\{0, \pm 1\}^n$. I will show that for these special halfspaces, the center norm rapidly approaches $\sqrt{2/\pi}$ as the weight $\|w\|_1$ increases:

Proposition 7.5.1. *Lef $f(x) = \operatorname{sign}(\langle w, x\rangle)$ be a halfspace where $w \in \{0, \pm 1\}^n$ such that $1 < k = \|w\|_1$ is odd. Then*

$$\frac{2}{\pi} \leq \|\mathsf{Com}\,(f)\|^2 \leq \frac{2}{\pi}\frac{k}{k-1}\,.$$

Proof. Assume without loss of generality that $w = \overbrace{1\cdots 1}^{k}0\cdots 0$. We will compute $\|\mathsf{Com}\,(f^+)\|^2$ since

$$\mathsf{Com}\,(f) = \frac{1}{2}\left(\mathsf{Com}\left(f^+\right) - \mathsf{Com}\left(f^-\right)\right) = \mathsf{Com}\left(f^+\right)\,.$$

We compute the square:

$$\left\|\mathbb{E}\left[x \;\middle|\; x \in f^+\right]\right\|^2 = \sum_{i\in[n]} \mathop{\mathbb{E}}_{x}\left[x_i \;\middle|\; x \in f^+\right]^2 = \sum_{i\in[k]} \mathop{\mathbb{E}}_{x}\left[x_i \;\middle|\; x \in f^+\right]^2\,.$$

In the second inequality we have used the fact that for $i > k, \mathbb{P}\left[x_i = 1 \;\middle|\; \langle x, w\rangle \geq 0\right] = \mathbb{P}\left[x_i = -1 \;\middle|\; \langle x, w\rangle \geq 0\right]$, which we can see by the fact that $\langle x, w\rangle = \langle x', w\rangle$ where x' is x with coordinate i flipped. For $i \in [k], \mathbb{E}\left[x_i \;\middle|\; x \in f^+\right] = (p - (1-p))^2 = (2p-1)^2$ where $p = \mathbb{P}\left[x_i = 1 \;\middle|\; x \in f^+\right]$, which is the same for each i. Therefore the norm is

$$\left\|\mathbb{E}\left[x \;\middle|\; x \in f^+\right]\right\|^2 = k(2p-1)^2$$

so what remains is to compute p.

We count the number of points for which $x_1 = 1$. We are only concerned with the first k coordinates, so we ignore the last $n-k$ coordinates (which would introduce a factor of 2^{n-k} in the size of each set, which cancel out when computing the probability). So there are 2^{k-1} remaining points after fixing the first coordinate. Note that $k-1$ is even since k is odd; if more than half of these $k-1$ coordinates are $+1$, the whole sum is positive and negating these coordinates will make the whole sum negative. So there is a 1:1 correspondence between the strings that are mostly positive and mostly negative, and we only have to worry about the strings that are half positive and half negative, which is exactly $\binom{k-1}{(k-1)/2}$. Therefore we take half of all 2^{k-1} strings and add the evenly-weighted strings that were left out:

$$2^{k-1}p = \#\left\{x : x_1 = 1, \sum_{i\in[k]} x_i > 0\right\} = 2^{k-2} + \frac{1}{2}\binom{k-1}{\frac{k-1}{2}}$$

Using the identity $\binom{n}{n/2} = \frac{2^n\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n+2}{2}\right)}$ and the inequality $\Gamma\left(\frac{n+1}{2}\right) \in \left[\frac{\sqrt{n-1}}{\sqrt{2}}\Gamma\left(\frac{n}{2}\right), \frac{\sqrt{n}}{\sqrt{2}}\Gamma\left(\frac{n}{2}\right)\right]$ (Theorem 2.4.3), we have

$$\binom{k-1}{\frac{k-1}{2}} = \frac{2^{k-1}\Gamma\left(\frac{k}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{k+1}{2}\right)} \in \left[\frac{\sqrt{2}\cdot 2^{k-1}}{\sqrt{\pi k}}, \frac{\sqrt{2}\cdot 2^{k-1}}{\sqrt{\pi(k-1)}}\right].$$

This gives us

$$2^{k-1}p = 2^{k-2} + \frac{1}{2}\left[\frac{\sqrt{2}\cdot 2^{k-1}}{\sqrt{\pi k}}, \frac{\sqrt{2}\cdot 2^{k-1}}{\sqrt{\pi(k-1)}}\right] = 2^{k-2}\left(1 + \frac{\sqrt{2}}{\sqrt{\pi}}\left[\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k-1}}\right]\right).$$

Finally,

$$k(2p-1)^2 = k\left(1 + \frac{\sqrt{2}}{\sqrt{\pi}}\left[\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k-1}}\right] - 1\right)^2 = \frac{2}{\pi}k\left[\frac{1}{k}, \frac{1}{k-1}\right]. \qquad \square$$

Since all weight vectors are convex combinations of these special vectors, we might suspect that, perhaps, *all* balanced halfspaces have centers with norm at least $\sqrt{2/\pi}$. Recall example 7.2.5, which showed that for $f(x) = \text{sign}(2x_1 + 2x_2 + x_3 + x_4 + x_5)$, $\|\mathsf{Com}(f)\|_2^2 = 44/64$ while $\textsc{Maj}_5$ has $\|\mathsf{Com}(\textsc{Maj}_n)\|_2^2 = 45/64$; this showed that the corner vector $\vec{1}$ does not always produce the halfspace with least center-norm; however $44/64 > 2/\pi$ so the suspicion may still be true. This leads us to the following open problem, catalogued by O'Donnell in [O'D14], that extends the intuition to unbalanced halfspaces:

Conjecture 7.5.2. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any halfspace. Then*

$$\mathbb{E}[f]^2 + \sum_{i\in[n]} \hat{f}(i)^2 \geq \frac{2}{\pi}.$$

The vectors $\{0, \pm 1\}^n$ are the most regular weight vectors; using central limit theorems on other regular vectors gets us a similar estimate as a function of the regularity:

Theorem 7.5.3 (Improvement on [MORS10] Theorem 33)**.** *Let $f(x) = \text{sign}(\langle w, x\rangle)$ be a balanced, τ-weight-regular halfspace with $\|w\| = 1$. Then $\left|\|\mathsf{Com}(f)\| - \frac{\sqrt{2}}{\sqrt{\pi}}\right| \leq O(\tau)$ and $\left|\|\mathsf{Com}(f)\|^2 - \frac{2}{\pi}\right| \leq O(\tau)$.*

Proof. Let $c = \|\mathsf{Com}(f)\|$ for simplicity. We start with the identity

$$c = \mu(f^+)\mathbb{E}[\langle w, x\rangle \mid \langle w, x\rangle \geq 0] - \mu(f^-)\mathbb{E}[\langle w, x\rangle \mid \langle w, x\rangle < 0] = \mathbb{E}[|\langle w, x\rangle|].$$

Now we may use Theorem 7.1.6 to get

$$\mathbb{E}\left[|\langle w, x\rangle|\right] = \mathbb{E}\left[|X|\right] \pm O\left(\tau\right)$$

where $X \sim \mathcal{N}(0,1)$ is a standard Gaussian, and $\mathbb{E}\left[|X|\right] = \sqrt{2/\pi}$ (Lemma 2.3.11). Finally, since $\left|c - \sqrt{2/\pi}\right| \leq O\left(\tau\right)$ and $c \leq 1$ we have

$$\left|c^2 - 2/\pi\right| = \left|c - \sqrt{2/\pi}\right| \left(c + \sqrt{2/\pi}\right) = O\left(\tau\right). \qquad \square$$

The goal is to compare these centers of mass to the centers of arbitrary functions. There are a few results giving upper bounds on the centers of arbitrary boolean functions, going back to Talagrand:

Theorem 7.5.4 ([Tal96], see [KKMO07] Theorem 17)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any boolean function with $\mathbb{P}\left[f(x) = 1\right] = p \leq 1/2$. Then*

$$\left\|\mathsf{Com}\left(f\right)\right\|^2 \leq O\left(p^2 \log(1/p)\right).$$

Khot *et al.* provide several theorems of this type, in which the ubiquitous value $2/\pi$ becomes more prominent:

Theorem 7.5.5 ([KKMO07] Theorem 6)**.** *Let $f : \{\pm 1\}^n \to [-1, 1]$ be a bounded function with $\max_i \mathrm{Inf}_i(f) \leq \delta$. Then*

$$\left\|\mathsf{Com}\left(f\right)\right\|^2 \leq \frac{2}{\pi} + 2\delta(1 - \sqrt{2/\pi}).$$

Finally, recall the Gaussian isoperimetric function $I(v) := \phi(\Phi^{-1}(v))$.

Theorem 7.5.6 ([KKMO07] Theorem 7)**.** *Let $f : \{\pm 1\}^n \to [-1, 1]$ be a bounded function with $\max_i \mathrm{Inf}_i(f) \leq \delta$. For $v = \frac{1}{2}(1 - \mathbb{E}\left[f\right])$,*

$$\left\|\mathsf{Com}\left(f\right)\right\|^2 \leq 4(I(v) + \epsilon)^2$$

where $\epsilon = O\left(\delta\right) \cdot \max\left(1, \sqrt{|\Phi^{-1}(v)|}\right)$.

Note that, ignoring the error term ϵ that necessarily depends on the regularity, the right side of this theorem is exactly the function $U(v) = (2I(v))^2$ of Matulef *et al.* [MORS10] that gives the center-norm of a threshold function in Gaussian space.

Matulef *et al.*, in their paper on testing halfspaces, expand this relationship between the isoperimetric function and the center-norm of regular halfspaces. To motivate the next theorem, observe that by taking a halfspace $f(x) = \operatorname{sign}\langle w, x\rangle - \theta$ and applying two different restrictions $\rho_1, \rho_2 : T \to \{\pm 1\}$ for some set of coordinates, say $T = \{k+1, \ldots, n\}$, gives us two halfspaces with the same weight vector $w' = (w_1, \ldots, w_k)$ but different thresholds $\theta_1 = \theta - \sum_{i=k+1}^{n} w_i \rho_1(i)$ and $\theta_2 = \theta - \sum_{i=k+1}^{n} w_i \rho_2(i)$. Keep in mind also the fact that applying a restriction to coordinates with large weights is likely to produce a weight-regular halfspace (Fact 7.1.21).

Theorem 7.5.7 ([MORS10] Theorem 48)**.** *Let $w \in \{\pm 1\}^n$ be a weight vector, $\theta_1, \theta_2 \in \mathbb{R}$ be thresholds, and let $0 < \tau$ be a sufficiently small regularity parameter. Suppose $f_1(x) = \operatorname{sign}(\langle w, x\rangle - \theta_1), f_2(x) = \operatorname{sign}(\langle w, x\rangle - \theta_2)$ are both τ-Fourier-regular halfspaces (i.e. $\max_i \operatorname{Inf}_i(f_1), \max_i \operatorname{Inf}_i(f_1) \leq \tau$). Then*

$$|\mathsf{Com}\,(f_1) - U(\mathbb{E}\,[f_1])| \leq \tau^{1/6}$$

and

$$\left| \left(\sum_{i=1}^{n} \hat{f}_1(i) \hat{f}_2(i) \right)^2 - U(\mathbb{E}\,[f_1]) U(\mathbb{E}\,[f_2]) \right| \leq \tau^{1/6}\,.$$

7.5.2 Distance

The theorems in the last section were of the form "If f is a sufficiently regular halfspace, then f has center-norm close to X (roughtly $\sqrt{2/\pi}$)". A testing algorithm also needs to use the converse of this type of theorem, that is, a theorem of the form "If f is sufficiently regular and has center-norm close to X, then f is close to a halfspace". These are the theorems I will cover in this subsection.

Note that the template for theorems all have the condition that the functions are Fourier-regular. There are two reasons for this: first, this condition makes the calculations easier, because of the central limit theorems! And the second reason is that, for the purposes of the testing algorithm in [MORS10], it actually suffices to work with regular functions: this is because the algorithm works on randomly constructed restrictions to the highly influential coordinates, using the principle that these restrictions should be highly regular (Fact 7.1.21).

The relevant theorems from MORS are the following, which establish the converse for balanced halfspaces and then extend it to unbalanced ones:

Theorem 7.5.8 ([MORS10] Theorem 34)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any boolean function that is τ-Fourier-regular (i.e. $\left|\hat{f}(i)\right| \leq \tau$ for all $i \in [n]$). Suppose $\left|\|\mathsf{Com}(f)\|^2 - 2/\pi\right| \leq \gamma$. Write $c = \mathsf{Com}(f)$. Then the halfspace $g(x) = \mathrm{sign}(\langle c, x\rangle)$ satisfies $\mathsf{dist}(f, g) \leq O\left(\sqrt{\tau + \gamma}\right)$.*

Theorem 7.5.9 ([MORS10] Theorem 49)**.** *Let $0 < \tau < 1$ and suppose that $f, g : \{\pm 1\}^n \to \{\pm 1\}$ are boolean functions satisfying:*

1. *f is τ-Fourier-regular and $|\mathbb{E}[f]| \leq 1 - \tau^{2/9}$,*
2. *$\left|\|\mathsf{Com}(f)\|^2 - U(\mathbb{E}[f])\right| \leq \tau$,*
3. *$\left|\left(\sum_{i=1}^n \hat{f}(i)\hat{g}(i)\right)^2 - U(\mathbb{E}[f]) \cdot U(\mathbb{E}[g])\right| \leq \tau$,*
4. *$\sum_{i=1}^n \hat{f}(i)\hat{g}(i) \geq -\tau$.*

Write $c = \mathsf{Com}(f)$ and let $\hat{c} = c/\|c\|$. Then for some $\theta \in \mathbb{R}$, g is $O\left(\tau^{1/9}\right)$-close to the halfspace $\mathrm{sign}(\langle \hat{c}, x\rangle - \theta)$.

A number of works have achieved spiritually similar results in the case where the functions are not necessarily regular. A theorem of Birkendorf *et al.* is our first example of such a theorem, and is also an example of the application of integer weight theorems:

Theorem 7.5.10 ([BDJ+98], see [Ser07] Theorem 6.1)**.** *Let $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$ for some threshold $\theta \in \mathbb{R}$ and integer weights $w \in \mathbb{R}^n$ with sum $W = \sum_{i=1}^n |w_i|$. Suppose $g : \{\pm 1\}^n \to \{\pm 1\}$ is any boolean function such that, for all $i = 0, 1, \ldots, n$ and any $0 < \epsilon < 1$,*

$$\left|\hat{g}(i) - \hat{f}(i)\right| \leq \frac{\epsilon}{W}.$$

Then $\mathsf{dist}(f, g) \leq \epsilon$.

Remember that, for some extreme examples, we might have $W = 2^{\Omega(n \log n)}$ (Theorem 7.4.1, in which case the requirements for this theorem are very strong. Goldberg proved another theorem of this type, with the same motivation, that does not depend on the integer weights:

Theorem 7.5.11 ([Gol06] Theorem 4, see [Ser07] Theorem 6.2)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any halfspace and suppose $g : \{\pm 1\}^n \to \{\pm 1\}$ is any boolean function such that, for any $0 < \epsilon < 1$ and all $i = 0, 1, \ldots, n$, satisfies*

$$\left|\hat{g}(i) - \hat{f}(i)\right| \leq \left(\frac{\epsilon}{n}\right)^{O(\log(n/\epsilon)\log(1/\epsilon))}.$$

Then $\mathsf{dist}(f, g) \leq \epsilon$.

Following the critical index method, Servedio shows a similar theorem (compare the bound to the values from Theorem 7.4.3 of the same paper):

Theorem 7.5.12 ([Ser07] Theorem 1.2)**.** *Let* $f : \{\pm 1\}^n \to \{\pm 1\}$ *be any halfspace and suppose* $g : \{\pm 1\}^n \to \{\pm 1\}$ *is any boolean function that, for any* $0 < \epsilon < 1$ *and all* $i = 0, 1, \ldots, n$*, satisfies*

$$\left|\hat{g}(i) - \hat{f}(i)\right| \leq 1/\left(n \cdot 2^{\tilde{O}(1/\epsilon^2)}\right).$$

Then $\mathsf{dist}(f, g) \leq \epsilon$.

De *et al.*, for the purpose of designing an algorithm for the Chow Parameters Problem, improve these bounds with their main structural theorem (recall the notation $\hat{f} = (\hat{f}(0), \hat{f}(1), \ldots, \hat{f}(n))$):

Theorem 7.5.13 ([DDFS14] Theorem 7)**.** *Let* $f : \{\pm 1\}^n \to \{\pm 1\}$ *be any halfspace and let* $g : \{\pm 1\}^n \to [-1, 1]$ *be any bounded function such that, for any* $0 < \epsilon < 1$*,*

$$\left\|\hat{f} - \hat{g}\right\| \leq \epsilon^{O\left(\log^2(1/\epsilon)\right)} = 2^{-O\left(\log^3(1/\epsilon)\right)}.$$

Then $\mathsf{dist}(f, g) \leq \epsilon$.

The proof of this theorem combines the strategy of Goldberg (Theorem 7.5.11) with the critical index method. Writing the Gap Theorem in this manner yields the following: for any halfspace $f(x) = \mathrm{sign}(\langle w, x\rangle - \theta)$ and bounded function g satisfying $\mathbb{E}[g] = \mathbb{E}[f]$ (i.e. $\hat{f}(0) = \hat{g}(0)$), if

$$\left\|\hat{f} - \hat{g}\right\| \leq O\left(\frac{\epsilon}{W(w, \epsilon/2)}\right)$$

then $\mathsf{dist}(f, g) \leq \epsilon$. This shows that if we can find a good bound on the width of 1-dimensional projections of the hypercube, we may be able to improve these results:

Question 7.5.14. *Can we improve distance bounds for the hypercube using the Gap Theorem and the 1-dimensional width?*

For the Gaussian space, the Gap Theorem gives a bound of $\mathsf{dist}(f, g) \leq O(\sqrt{\epsilon})$ (since $W(\epsilon/2) \geq \Omega(\epsilon)$); the authors of [DDFS14] show that no similar bound of ϵ^C for any constant $C > 0$ can hold, although there is still room for improvement.

7.6 Margins and Width

Recall the definition of width:

Definition 4.3.3: Let μ be any probability distribution over $\mathbb{R}^n$, $w \in \mathbb{R}^n$ such that $\|w\| = 1$, and let $\epsilon \in (0, 1]$. The *ϵ-width* of the distribution is

$$W_\mu(w, \epsilon) := \sup\{r \geq 0 : p_r(w) \leq \epsilon\}$$

To use the Gap Theorem for the hypercube, we would need to find bounds on the width. As a first example, observe that the hypercube presents some difficulty that doesn't occur in continuous spaces: picking $w = \frac{1}{\sqrt{n}}\vec{1}$ for even n, we can see that $\binom{n}{n/2}$ points lie exactly on the plane $\{x \in \{\pm 1\}^n : \langle w, x\rangle = 0\}$. Thus if ϵ to smaller than this fraction of points, picking any radius $r > 0$ gives $p_r(w) \geq \binom{n}{n/2}2^{-n} > \epsilon$ so the width is 0. We need to avoid these degenerate cases!

Two strategies from earlier in the chapter can be used to achieve this: first, for a halfspace f, we can find alternative weight vectors that produce the same function but with better width. Second, we can construct an approximator g that has better width, which should give even stronger results.

Before delving into some theorems that might help us bound the width, consider the case that the vector w is τ-regular. Then we can use central limit theorems to move to the Gaussian distribution:

Fact 7.6.1. *Let $0 < \epsilon < 1$ and $0 < \tau < \epsilon/2$. Suppose $w \in \mathbb{R}^n$ is a τ-regular vector such that $\|w\| = 1$. Then*

$$W(w, \epsilon) \geq \sqrt{2\pi}(\epsilon - 2\tau)$$

Proof. For any $r > 0$ and $\theta \in \mathbb{R}$, corollary 7.1.5 gives us

$$\mathbb{P}\left[|\langle w, x\rangle - \theta| \leq r\right] = \mathbb{P}\left[|X - \theta| \leq r\right] \pm 2\tau$$

Therefore, by definition,

$$\begin{aligned} W(w, \epsilon) &= \sup\{r \geq 0 : \sup_\theta \mathbb{P}\left[|\langle w, x\rangle - \theta| \leq r\right] \leq \epsilon\} \\ &\geq \sup\{r \geq 0 : \sup_\theta \mathbb{P}\left[|X - \theta| \leq r\right] \leq \epsilon - 2\tau\} \end{aligned}$$

where X is a standard Gaussian. This is the $(\epsilon - 2\tau)$-width of the Gaussian, which is at least $\sqrt{2\pi}(\epsilon - 2\tau)$. □

One reason why the definition of width is convenient is that anticoncentration inequalities imply bounds: suppose we have $p_r(w) \leq \epsilon$ for some r, ϵ. Then we clearly have $W(w, \epsilon) \geq r$. Together with the extension lemma (Lemma 7.1.15, we can find some subset $H \subseteq [n]$ with the bound $p_r(w_H) \leq \epsilon$ and extend this to a width bound. Unfortunately, I note that simply applying the anticoncentration inequalities along with the structural lemmas in section 7.1 does not give an improvement over the distance Theorem 7.5.13.

A few recent works have relied on margins for their proofs, and as a consequence have included some bounds that relate to the width. O'Donnell and Servedio provide the following two theorems:

Theorem 7.6.2 ([OS11] Theorem 3.1)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be any linear threshold function and let $r > 0$ be sufficiently small. Then there is a linear threshold function $g(x) = \text{sign}(\langle w, x\rangle - \theta)$ satisfying $\mathsf{dist}(f, g) \leq 2^{-1/r}$ such that*

$$\mu(\text{margin}_r(w, \theta)) \leq \tilde{O}\left(\frac{1}{\sqrt{\log(1/r)}}\right)$$

Theorem 7.6.3 ([OS11] Theorem 4.2)**.** *Let $0 < \tau < 1/2$ and $t \geq 1$, with $K = \lceil C\frac{t}{\tau^2}\ln(t/\tau)\rceil$ for some absolute constant C. Suppose $w \in \mathbb{R}^n$ is a vector with τ-critical index $k \geq K$ and let $\sigma_k = \sqrt{\sum_{i=k}^n}$. Then for $r = \sqrt{t}\cdot\sigma_k$,*

$$p_r(w) \leq O\left(2^{-t}\right)$$

Chapter 8

Conclusions and Future Work

I have presented testing algorithms for rotationally invariant probability spaces and for mixtures of two of these spaces; clearly there is a very large gap between my current results and the ultimate goal of having efficient testing algorithms for arbitrary probability distributions. There is a lot of work remaining; in this chapter I will summarize some of the questions that I asked in earlier chapters, and suggest a few strategies that might yield progress towards the ultimate goal.

Larger Mixtures

In Chapter 5 I presented an algorithm that works for mixtures of two rotationally invariant distributions. An obvious direction to pursue is to extend the algorithm to work for arbitrary mixtures of k rotationally invariant distributions (where k is a constant independent of the dimension). This would be a good step to take, since this would include mixtures of Gaussians, which are commonly studied.

This extension would require:

1. Finding a general formula for $\xi(v_1, v_2, \ldots, v_k)$ when there are k RI distributions.
2. Computing bounds on the number of samples required to estimate this quantity.

The Gap Theorem and the algorithmic template used for 2-mixtures would finish the job.

Narrow Distributions

In Chapter 4, I gave an example of an RI distribution for which the width is sublinear, i.e. for any constant $C > 0$, there is a small enough $\epsilon > 0$ such that $W(\epsilon) < C \cdot \epsilon$. For these "narrow" distributions, the sample complexity of the RI halfspace tester can be much larger than the desired $O(\sqrt{n})$, but they seem "unreasonable" since they are not "properly scaled", relative to the Gaussian distribution. Here I repeat the question posed in that chapter:

Question 4.7.3: How can we improve the algorithm to use fewer samples on spaces with sublinear width?

Extended Gap Theorem

The Gap Theorem (Theorem 4.4.1) compares the center-norms of a function and a halfspace with the same measure. In Chapter 4, I asked:

Question 4.4.2: Is there a version of the Gap Theorem that depends on $\mathbb{E}[f] - \mathbb{E}[h]$ instead of requiring that this difference is 0?

An extension of this theorem that tolerates differences of measure would yield results for the Chow Parameters Problem (see Chapter 6). Moreover, several theorems about the hypercube, such as in Section 7.5, have a similar flavor to the Gap Theorem while tolerating these differences of measure. An improvement to the Gap Theorem along with more accurate knowledge of the width of the hypercube might lead to simplifications of this theory.

Hypercube and Regularity

In Chapter 7 we saw a number of questions and open problems about the structure of the hypercube whose solutions may help find simpler testing algorithms and generalizations, and are also of independent interest. In particular, the relationship between weight vectors and centers of mass remains mysterious, as I mentioned in the chapter on the Chow Parameters Problem, which is part of this mystery:

Question 6.0.3: *What is the relationship between the normal of a halfspace and its center? Specifically,*

1. *How close must the two vectors be? This has consequences for the Gap Theorem.*

2. *How closely related are weight– and Fourier–regularity?*
3. *How small can the center-norm be? Is it at least $\sqrt{2/\pi}$?*
4. *How can we compute the weights from the centers? This is the Chow Parameters Problem.*

How to apply the Gap Theorem in discrete spaces is a difficult problem, of which the hypercube would provide a good special case:

Question 8.0.4. *What is the tightest version of the Gap Theorem that holds for the hypercube? This is related to anticoncentration inequalities and the Littlewood-Offord problem; a thorough search of the literature on this problem might yield stronger tools.*

Regularity is a concept that has been very useful for proving theorems about the hypercube, especially combined with the critical index method. It might be interesting to explore different definitions of regularity to see if they are equally as powerful; for example, rewriting the definition of τ-regularity, we see that a vector $w \in \mathbb{R}^n$ is τ-regular if

$$\|w\|_\infty \leq \tau \cdot \|w\|_2 \ .$$

This is a useful definition because it is closely related to central limit theorems which depend on the 2-norm. However, it is the additive structure of w that determines its anticoncentration properties; we can easily see, for example, that when we change the weights w, the center of mass will stay the same until some point crosses the hyperplane: this occurs when $\langle w, x \rangle = \theta$. That means changes in center occur exactly when we can partition the weights $\{w_i\}$ into two sets P, N (positive and negative) such that $\sum_{i \in P} w_i - \sum_{i \in N} w_i = \theta$. This additive structure is more accurately described by the 1-norm of w than the 2-norm, so we might consider the L_1 regularity, defined as

$$\|w\|_\infty \leq \tau \cdot \|w\|_1 \ ,$$

or more generally

$$\|w\|_p \leq \tau \cdot \|w\|_q \ ,$$

which gives us a whole spectrum of regularity properties to work with. From these regularity definitions would follow analogous critical index methods, which might prove fruitful.

Distribution-Free Testing

The model I have used in this thesis lets the algorithm use knowledge of the distribution (e.g. it can compute inverse cumulative distribution functions). This is quite a strict

limitation on the use of the algorithms: from a practical perspective, it is easy to imagine that one would not necessarily know what distribution their data came from. A more practical model would be to hide the distribution from the algorithm: this is the model of distribution-free testing (see e.g. [GS07]). The general approach that I have used in this work, estimating the center-norm and comparing it to its proper value, might not work in this model: estimating the proper value might require too much information about the distribution to yield a sublinear sample complexity.

Question 8.0.5. *Is there a distribution-free tester for halfspaces with sublinear sample complexity?*

Lower Bounds

I have not provided any new lower bounds for this problem; in Chapter 3 I catalogued a few lower bounds related to the problem; in particular, there is a bound of $\Omega\left(\sqrt{n/\log n}\right)$ from [BBBY12]. I suspect that the $\log n$ factor is unnecessary (i.e. the dependence on n in my algorithm is optimal).

Question 8.0.6. *Fixing ϵ to be a constant, is the sample complexity of testing halfspaces $\Theta(\sqrt{n})$ (for RI and mixtures of RI distributions)?*

Further, this lower bound does not depend on ϵ, and it may be interesting to find the optimal dependence on ϵ.

Even more interesting would be to determine if *all* probability spaces have a halfspace tester; one can imagine that there might be spaces where, say, estimating the volume and center-norm gives away enough information to learn the halfspace. In this case, a tester would be useless since we could skip right to learning.

Question 8.0.7. *Are there any probability spaces where testing and learning have (roughly) the same sample complexity?*

Iterative Methods

The classic perceptron algorithm (see section 3.2.2) works by iteratively improving a hypothesis, represented by a weight vector: taking a labelled point, it checks whether the

current hypothesis correctly labels the point, and udpates the hypothesis vector by adding or subtracting the point if the classification is incorrect.

Later, I presented the ChowReconstruct algorithm of [DDFS14] for the Chow Parameters Problem (Chapter 6) which begins with the center of mass as a hypothetical weight vector and updates this hypothesis in every iteration by adding to it the difference between the current center of mass and target Chow parameters. I will call these two algorithms examples of *iterative methods*, which start with some weight vector and repeatedly update it to get closer to some target; these algorithmic methods might also yield iterative proof techniques that could aid in the quest to establish a theory of halfspaces. For example, analyzing ChowReconstruct, one might be able to discover some facts about how close the center of mass and weight vector must be to each other.

Another potential use of the iterative method would be to reduce a conjecture C, say for example, the conjecture that $\mathbb{E}[f] + \|\mathsf{Com}(f)\|_2^2 \geq 2/\pi$ for all halfspaces (Conjecture 7.5.2), to the following form: Is C true for halfspaces that satisfy the termination conditions for an algorithm A? and is C preserved at every iteration of A? A starting point would be an algorithm that computes the center of mass, c, of a halfspace, and the uses c as the weight vector for the next iteration; we can then terminate on halfspaces whose weight vector and center of mass are parallel. Experiments with this algorithm have revealed some promising but elusive patterns.

References

[Aar16] Scott Aaronson. $P \stackrel{?}{=} NP$. In *Open Problems in Mathematics*, pages 1–122. Springer, 2016.

[ABL17] Pranjal Awasthi, Maria-Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):50, 2017.

[Bau90] Eric B Baum. The perceptron algorithm is fast for non-malicious distributions. In *Advances in Neural Information Processing Systems*, pages 676–685, 1990.

[BBBY12] Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active property testing. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 21–30. IEEE, 2012.

[BDEL03] Shai Ben-David, Nadav Eiron, and Philip M Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

[BDJ+98] Andreas Birkendorf, Eli Dichterman, Jeffrey Jackson, Norbert Klasner, and Hans Ulrich Simon. On restricted-focus-of-attention learnability of boolean functions. *Machine Learning*, 30(1):89–123, 1998.

[BFKV96] Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 330–338. IEEE, 1996.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of boolean functions and applications to percolation. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, (90):5–44, 1999.

[BL13] Maria-Florina Balcan and Philip M Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, pages 288–316, 2013.

[Bla09] Eric Blais. Testing juntas nearly optimally. In *Proceedings of the forty-first annual ACM symposium on Theory of Comuting*, pages 151–158. ACM, 2009.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

[Bou02] Jean Bourgain. On the Distribution of the Fourier Spectrum of Boolean Functions. *Israel Journal of Mathematics*, 131(1):269–276, 2002.

[Cho61] Chao-Kong Chow. On the characterization of threshold functions. In *Switching Circuit Theory and Logical Design, 1961. SWCT 1961. Proceedings of the Second Annual Symposium on*, pages 34–38. IEEE, 1961.

[Dan15] Amit Daniely. A ptas for agnostically learning halfspaces. In *COLT*, pages 484–502, 2015.

[DDFS14] Anindya De, Ilias Diakonikolas, Vitaly Feldman, and Rocco A Servedio. Nearly Optimal Solutions for the Chow Parameters Problem and Low-weight Approximation of Halfspaces. *Journal of the ACM (JACM)*, 61(2):11, 2014.

[Dia10] Diakonikolas, Ilias and Gopalan, Parikshit and Jaiswal, Ragesh and Servedio, Rocco A and Viola, Emanuele. Bounded Independence Fools Halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

[DJS+14] Ilias Diakonikolas, Ragesh Jaiswal, Rocco A Servedio, Li-Yang Tan, and Andrew Wan. Noise stable halfspaces are close to very small juntas. *Chicago Journal Of Theoretical Computer Science*, 2015:4, 2014.

[DS13] Ilias Diakonikolas and Rocco A Servedio. Improved approximation of linear threshold functions. *Computational Complexity*, 22(3):623–677, 2013.

[EHKV89] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A General Lower Bound on the Number of Examples Needed for Learning. *Information and Computation*, 82(3):247–261, 1989.

[Eld15] Ronen Eldan. A two-sided estimate for the gaussian noise stability deficit. *Inventiones mathematicae*, 201(2):561–624, 2015.

[Erd45] Paul Erdős. On a lemma of Littlewood and Offord. *Bulletin of the American Mathematical Society*, 51(12):898–902, 1945.

[Fel68] William Feller. *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons New York, 1968.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 563–574. IEEE, 2006.

[FGRW12] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal of Computing*, 41(6):1558–1590, 2012.

[FKR+02] Eldar Fischer, Guy Kindler, Dana Ron, Shmuel Safra, and Alex Samorodnitsky. Testing juntas [combinatorial property testing]. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 103–112. IEEE, 2002.

[Fri98] Ehud Friedgut. Boolean Functions with Low Average Sensitivity Depend on Few Coordinates. *Combinatorica*, 18(1):27–35, 1998.

[Gol06] Paul M Goldberg. A Bound on the Precision Required to Estimate a Boolean Perceptron from its Average Satisfying Assignment. *SIAM Journal on Discrete Mathematics*, 20(2):328–343, 2006.

[GOWZ10] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *Computational Complexity (CCC), 2010 IEEE 25th Annual Conference on*, pages 223–234. IEEE, 2010.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[GS07] Dana Glasner and Rocco A Servedio. Distribution-Free Testing Lower Bounds for Basic Boolean Functions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 494–508. Springer, 2007.

[GW95] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[Hal77] G Halász. Estimates for the concentration function of combinatorial number theory and probability. *Periodica Mathematica Hungarica*, 8(3-4):197–211, 1977.

[Hås94] Johan Håstad. On the size of weights of threshold gates. *SIAM Journal on Discrete Mathematics*, 7(3):484–492, 1994.

[HSV95] Klaus-Uwe Hoffgen, Hans-Ulrich Simon, and Kevin S Vanhorn. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50(1):114–125, 1995.

[Jai17] Vishesh Jain. A Counterexample to the "Majority is Least Stable" Conjecture. *arXiv preprint arXiv:1703.07657*, 2017.

[KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal Inapproximability Results for MAX-CUT and Other 2-Variable CSPs? *SIAM Journal of Computing*, 37(1):319–357, 2007.

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KL93] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[KLS09] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

[KSS94] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[Kur16] Sascha Kurz. The inverse problem for power distributions in committees. *Social Choice and Welfare*, 47(1):65–88, 2016.

[KW15] Daniel M Kane and Ryan Williams. Super-Linear Gate and Super-Quadratic Wire Lower Bounds for Depth-Two and Depth-Three Linear Threshold Circuits. *arXiv preprint arXiv:1511.07860*, 2015.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

[Lon94] Philip M Long. On the sample complexity of PAC learning half-spaces against the uniform distribution. *IEEE transactions on neural networks/a publication of the IEEE Neural Networks Council*, 6(6):1556–1559, 1994.

[Lon03] Philip M Long. An upper bound on the sample complexity of PAC-learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

[MOO05] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Foundations of*

Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on, pages 21–30. IEEE, 2005.

[MORS09] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A Servedio. Testing±1-Weight Halfspaces. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 646–657. Springer, 2009.

[MORS10] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A Servedio. Testing halfspaces. *SIAM Journal on Computing*, 39(5):2004–2047, 2010.

[MTT61] Saburo Muroga, Iwao Toda, and Satoru Takasu. Theory of majority decision elements. *Journal of the Franklin Institute*, 271(5):376–418, 1961.

[Nis93] Noam Nisan. The Communication Complexity of Threshold Gates. *Combinatorics, Paul Erdos is Eighty*, 1:301–315, 1993.

[NS94] Noam Nisan and Mario Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[OS11] Ryan O'Donnell and Rocco A Servedio. The Chow Parameters Problem. *SIAM Journal on Computing*, 40(1):165–199, 2011.

[Qi10] Feng Qi. Bounds for the ratio of two gamma functions. *Journal of Inequalities and Applications*, 2010(1), 2010.

[Ron08] Dana Ron. Property testing: A learning theory perspective. *Foundations and Trends® in Machine Learning*, 1(3):307–402, 2008.

[Ros58] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[RS15] Dana Ron and Rocco A Servedio. Exponentially improved algorithms and lower bounds for testing signed majorities. *Algorithmica*, 72(2):400–429, 2015.

[Ser07] Rocco A Servedio. Every linear threshold function has a low-weight approximator. *Computational Complexity*, 16(2):180–209, 2007.

[She11] Irina Shevtsova. On the absolute constants in the Berry–Esseen type inequalities for identitically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[Tal96] Michel Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap92] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

[Wen48] J.G. Wendel. Note on the gamma function. *The American Mathematical Monthly*, 55(9):563–564, 1948.

[Wil14] Ryan Williams. New Algorithms and Lower Bounds for Circuits with Linear Threshold Gates. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 194–202. ACM, 2014.

[Win71] Robert O Winder. Chow parameters in threshold logic. *Journal of the ACM (JACM)*, 18(2):265–289, 1971.

APPENDICES

Appendix A

Width and Variance of Spheres and Gaussians

Proposition A.0.8. *Let μ be the uniform distribution over the n-dimensional unit sphere, where $n \geq 3$. Then*

$$V = \frac{2}{n+1} \pm \frac{1}{n+1}$$

Proof. We can write the surface area S_{n-1} as the integral of $S_{n-2}\left(\sqrt{1-z^2}\right)$ over the range $[-1, 1]$:

$$1 = \frac{S_{n-1}}{S_{n-1}} = \frac{1}{S_{n-1}} \int_{-1}^{1} S_{n-2}\left(\sqrt{1-z^2}\right) dz = \frac{S_{n-2}}{S_{n-1}} \int_{-1}^{1} (1-z^2)^{\frac{n-2}{2}} dz\,.$$

The variance is $V = \mathbb{E}\left[z^2\right]$ for z chosen from the distribution with density $\frac{S_{n-2}\left(\sqrt{1-z^2}\right)}{S_{n-1}}$, so

$$V = \frac{S_{n-2}}{S_{n-1}} \int_{-1}^{1} z^2 (1-z^2)^{\frac{n-2}{2}} dz\,.$$

Thus we can easily compute $1 - V$:

$$\begin{aligned} 1 - V = \frac{S_{n-2}}{S_{n-1}} \int_{-1}^{1} (1-z^2)(1-z^2)^{\frac{n-2}{2}} dz &= \frac{S_{n-2}}{S_{n-1}} \int_{-1}^{1} (1-z^2)^{\frac{n}{2}} dz \\ &= \frac{S_{n-2}}{S_{n-1}} \sqrt{\pi} \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+3}{2}\right)} \\ \text{Using Proposition 2.4.6:} \qquad &= \frac{2n\Gamma\left(\frac{n}{2}\right)^2}{(n^2-1)\Gamma\left(\frac{n-1}{2}\right)^2} \end{aligned}$$

which is between

$$\frac{n(n-2)}{(n+1)(n-1)} \qquad \text{and} \qquad \frac{n(n-1)}{(n+1)(n-1)}$$

by Theorem 2.4.3. Then V is between

$$\frac{1}{n+1} \qquad \text{and} \qquad \frac{(n+1)(n-1) - n(n-2)}{(n+1)(n-1)} \leq \frac{3}{n+1}\,. \qquad \square$$

Proposition A.0.9. *Let μ be the uniform distribution over the n–dimensional sphere with radius r, where $n \geq 3$. Let V^* be the projection variance for the unit sphere (from the previous proposition). Then*

$$V = r^2 V^*\,.$$

Proof.

$$\begin{aligned} V = \int_{-1}^{1} z^2 \frac{S_{n-2}\left(\sqrt{r-z^2}\right)}{S_{n-1}(r)} dz &= \frac{S_{n-2}}{S_{n-1}} \int_{-r}^{r} z^2 \frac{(r^2-z^2)^{\frac{n-2}{2}}}{r^{n-1}} dz \\ &= \frac{S_{n-2}}{S_{n-1}} \int_{-r}^{r} z^2 \frac{(r^2-z^2)^{\frac{n-2}{2}}}{r \cdot (r^2)^{\frac{n-2}{2}}} dz \\ &= \frac{S_{n-2}}{S_{n-1}} \int_{-r}^{r} z^2 \frac{(1-z^2/r^2)^{\frac{n-2}{2}}}{r} dz \\ t = z/r: \quad &= \frac{S_{n-2}}{S_{n-1}} \int_{-1}^{1} r^2 t^2 (1-t^2)^{\frac{n-2}{2}} dt = r^2 V^*\,. \end{aligned}$$

$\square$

Proposition A.0.10. *Let μ be the uniform distribution over the n–dimensional sphere with radius r (where $n \geq 3$). Then for any unit vector w, the density μ_w is*

$$\mu_w(z) = \frac{S_{n-2}\left(\sqrt{r^2-z^2}\right)}{S_{n-1}(r)} = \frac{S_{n-2}}{S_{n-1}} \frac{(1-z^2/r^2)^{\frac{n-2}{2}}}{r}$$

which is within

$$\frac{\sqrt{n-2}}{\sqrt{2\pi}}\frac{(1-z^2/r^2)^{\frac{n-2}{2}}}{r} \qquad \textit{and} \qquad \frac{\sqrt{n-1}}{\sqrt{2\pi}}\frac{(1-z^2/r^2)^{\frac{n-2}{2}}}{r},$$

and has maximum $\mu_w(0)$ *between* $\frac{\sqrt{n-2}}{\sqrt{2\pi}\cdot r}$ *and* $\frac{\sqrt{n-1}}{\sqrt{2\pi}\cdot r}$.

Proposition A.0.11. *For the* n*–dimensional standard Gaussian distribution, the width satisfies*

$$\sqrt{2\pi}\cdot\epsilon \leq W(\epsilon) \leq \sqrt{2\ln\left(\frac{2}{1-\epsilon}\right)}.$$

Proof. The lower bound is because the maximum density ϕ_{max} of the projection is $1/\sqrt{2\pi}$ and $W(\epsilon) \geq \epsilon/\phi_{max}$. For the upper bound we can use Lemma 2.3.12:

$$1-\epsilon = 2\mathbb{P}\left[z \geq W\right] \leq 2\exp\left(-\frac{W^2}{2}\right)$$

so $W \leq \sqrt{2\ln(2/(1-\epsilon))}$. □

Proposition A.0.12. *Let* μ *be the uniform distribution over the* n*–dimensional sphere with radius* r*. Then the width satisfies*

$$\sqrt{2\pi}\cdot\frac{r\epsilon}{\sqrt{n-1}} \leq W(\epsilon) \leq \sqrt{\frac{2r}{n-2}\ln\left(\frac{\sqrt{8}}{1-\epsilon}\right)}$$

Proof. Same as above, but using Proposition 2.4.8. □