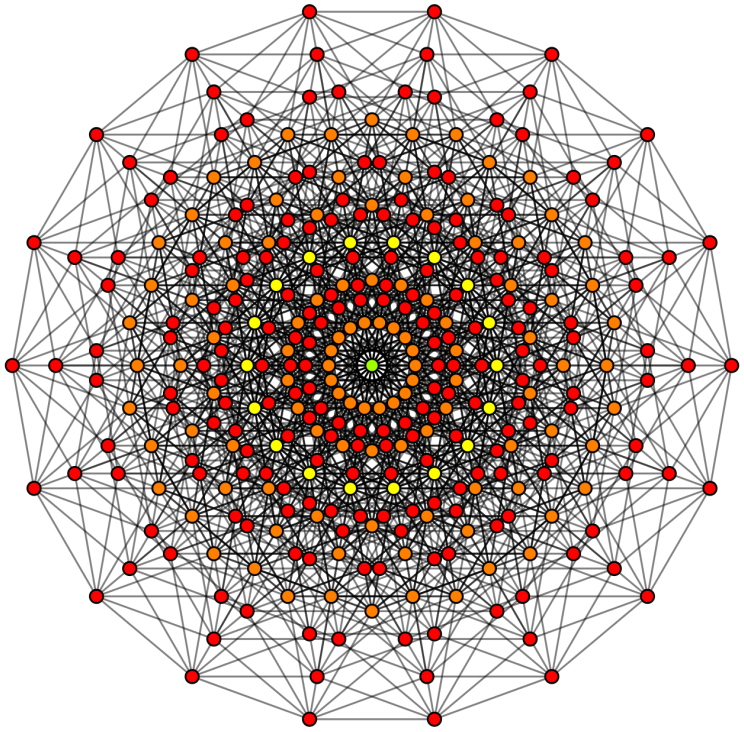
Computational Learning Theory
&
Fourier Analysis

Proceedings of the (online) Summer/Fall School

October 10–14 2022



Organizers

Alexandros Eskenazis, CNRS & Sorbonne Université

Paata Ivanisvili, University of California, Irvine

Contents

Cover picture of $\{-1,1\}^9$ courtesy of Wikipedia.

Chapter 1

Towards a proof of the Fourier-entropy conjecture?

after E. Kelman, G. Kindler, N. Lifshitz, D. Minzer, M. Safra [1]
A summary written by Alan Chang

Abstract. We sketch some ideas in recent progress towards the Fourier entropy conjecture.

1.1 Introduction

We first introduce/recall enough to state the Fourier entropy conjecture.

Definition 1.1. *On the space of functions $\{-1,1\}^n \to \mathbb{R}$, we define $\langle f, g\rangle = \mathbb{E}_x[f(x)g(x)]$ and $\|f\|_p = (\mathbb{E}_x[f(x)^p])^{1/p}$.*

Theorem 1.2. *For $S \subset [n]$, we let $\chi_S(x) = \prod_{i\in S} x_i$. The functions $(\chi_S)_S$ form an orthonormal basis. Thus, we have the Fourier expansion $f(x) = \sum_{S\subset[n]} \widehat{f}(S)\chi_S(x)$, where $\widehat{f}(S) = \langle f, \chi_S\rangle$. We also have Parseval/Plancherel: $\langle f, g\rangle = \sum_{S\subset[n]} \widehat{f}(S)\widehat{g}(S)$.*

Definition 1.3 (Influence)**.** *For $f, g : \{-1,1\}^n \to \mathbb{R}$ and $i \in [n]$, we define*

$$I_i[f] = \mathbb{E}_x\left[\left(\frac{f(x) - f(x\oplus e_i)}{2}\right)^2\right], I[f,g] = \sum_{i\in[n]} \sqrt{I_i[f]I_i[g]}, I[f] = I[f,f].$$

The key question that motivates this section is the following:

Question. *If $f : \{-1,1\}^n \to \{-1,1\}$ satisfies $I[f] \le K$, then what can we say about f?*

Here is one thing we can say.

Theorem 1.5 (KKL theorem)**.** *For all $f : \{-1,1\}^n \to \{-1,1\}$, there is an $i \in [n]$ such that $I_i[f] \ge e^{-CI[f]}$.*

We will prove Theorem 1.5 as a corollary of Lemma 1.19, below. Note that Theorem 1.5 gives a nontrival answer to Question 1.4 only if $K \lesssim \log n$. (If $I[f] \ge C' \log n$ for some sufficiently large C', then Theorem 1.5 only tells us that $I_i[f] \ge \frac{1}{n}$, but that already follows from the pigeonhole principle.) The difficulty of obtaining results when $K \gtrsim \log n$ is known as the *logarithmic barrier*. The Fourier entropy conjecture is a statement that goes beyond the logarithmic barrier.

Conjecture 1.6 (Fourier entropy conjecture)**.** *For all $f : \{-1,1\}^n \to \{-1,1\}$,*

$$H_{shannon}[\widehat{f}^2] := \sum_S \widehat{f}(S)^2 \log(1/\widehat{f}(S)^2) \lesssim I[f].$$

Conjecture 1.7 (Min-entropy conjecture)**.** *For all $f : \{-1,1\}^n \to \{-1,1\}$,*

$$H_\infty[\widehat{f}^2] := \min_S \log(1/\widehat{f}(S)^2) \lesssim I[f].$$

In [1], the authors almost prove the min-entropy conjecture; they are off by only a logarithmic factor.

Theorem 1.8 ([1, Theorem 1.1])**.** *such that for all $f : \{-1,1\}^n \to \{-1,1\}$,*

$$H_\infty[\widehat{f}^2] \lesssim I[f] \log(1 + I[f]).$$

In these notes, we present some ideas in the proof of Theorem 1.8.

1.2 Preliminaries

Definition 1.9 (Derivative)**.** *For $f : \{-1,1\}^n \to \mathbb{R}$ and $i \in [n]$, we define $\partial_i f(x) = \frac{1}{2}(f(x) - f(x \oplus e_i))$.*

Lemma 1.10. *We have* $\widehat{\partial_i f}(S) = \begin{cases} \widehat{f}(S), & \text{if } i \in S \\ 0, & \text{if } i \notin S \end{cases}$

As a consequence of Lemma 1.10, we have

$$\partial_i(f^{\le d}) = (\partial_i f)^{\le d}, \quad I_i[f] = \|\partial_i f\|_2^2 = \sum_{S \ni i} \widehat{f}(S)^2, \quad I[f] = \sum_S |S| \widehat{f}(S)^2.$$

1.2.1 Restrictions

Given $A \subset [n]$, we let $\bar{A}$ denote the complement of A. There is a natural bijection $\{-1,1\}^A \times \{-1,1\}^{\bar{A}} \to \{-1,1\}^n$, where we associate $(y,z) \in \{-1,1\}^A \times \{-1,1\}^{\bar{A}}$ to $x \in \{-1,1\}^n$ by $x_i = y_i$ if $i \in A$ and $x_i = z_i$ if $i \in \bar{A}$.

Definition 1.11. *For $f : \{-1,1\}^n \to \mathbb{R}$, $A \subset [n], z \in \{-1,1\}^{\bar{A}}$, we define $f_{\bar{A}\to z} : \{-1,1\}^A \to \mathbb{R}$ by $f_{\bar{A}\to z}(y) = f(y,z)$*

Lemma 1.12. $\langle f, g\rangle = \mathbb{E}_z\langle f_{\bar{A}\to z}, g_{\bar{A}\to z}\rangle$

Proof. $\langle f, g\rangle = \mathbb{E}_x[f(x)g(x)] = \mathbb{E}_z\mathbb{E}_y[f(y,z)g(y,z)] = \mathbb{E}_z\mathbb{E}_y[f_{\bar{A}\to z}(y)g_{\bar{A}\to z}(y)] = \mathbb{E}_z\langle f_{\bar{A}\to z}, g_{\bar{A}\to z}\rangle$. □

Lemma 1.13. *For $S \subset A$,* $\widehat{g_{\bar{A}\to z}}(S) = \sum_{T \subset \bar{A}} \widehat{g}(S \cup T)\chi_T(z)$

By the lemma above and Plancherel,

$$\mathbb{E}_z[\widehat{g_{\bar{A}\to z}}(S)^2] = \sum_{T : T \cap A = S} \widehat{g}(T)^2 \tag{1.1}$$

1.2.2 Partitions

Definition 1.14. *A partition of $[n]$ into m parts is $\mathcal{I} = (I_1, \dots, I_m)$, where $[n] = \bigsqcup_{j=1}^m I_j$*

We consider the uniform probability distribution over partitions of $[n]$ into m parts.

Lemma 1.15. *Let $S \subset [n]$, and let $\mu = |S|/m$. Then*

$$\Pr_{\mathcal{I}}\left[\forall j \in [m], (1-\epsilon)\frac{|S|}{m} \le |S \cap I_j| \le (1+\epsilon)\frac{|S|}{m}\right] \ge 1 - 2m\exp(-\frac{\epsilon^2|S|}{3m})$$

Proof. Fix $j \in [m]$. Let $\mu = \mathbb{E}|S \cap I_j| = \frac{|S|}{m}$. By the Chernoff bound,

$$\Pr_{\mathcal{I}}\left[\left||S \cap I_j| - \mu\right| \ge \epsilon\mu\right] \le 2\exp(-\epsilon^2\mu/3)$$

Apply the union bound to conclude. □

1.2.3 Hypercontractivity

Lemma 1.16 (Bonami lemma)**.** *If $f : \{-1,1\}^n \to \mathbb{R}$ has degree at most d, then $\|f\|_4 \le 3^{d/2}\|f\|_2$.*

Proof. Induct on n. □

The following lemma allows us to exchange maximum and expectation. The cost of this exchange is a factor of $e^{O(d)}$.

Lemma 1.17. *If $h_1, \dots, h_k : \{0,1\}^n \to \mathbb{R}$ are all of degree at most d, then*

$$\mathbb{E}_x\left[\max_{i\in[k]} h_i(x)^2\right] \le 3^d \max_{i\in[k]} \|h_i\|_2 \left(\mathbb{E}_x\left[\sum_{i=1}^k h_i(x)^2\right]\right)^{1/2}$$

Proof. Use Jensen, Holder, and Lemma 1.16. □

Corollary 1.18. *If $g : \{-1,1\}^n \to \mathbb{R}$ of degree d. Fix $A \subset [n]$.*

$$\mathbb{E}_z\left[\max_{S\subset A} \widehat{g_{\bar{A}\to z}}(S)^2\right] \le 3^d \|g\|_2 \max_{S\subset A}\left(\sum_{T:T\cap A=S} \widehat{g}(T)^2\right)^{1/2}$$

Proof. We apply Lemma 1.17 as follows: Let $(S_i)_i$ be an enumeration of the subsets of A, and let $h_i(z) = \widehat{g_{\bar{A}\to z}}(S_i)$. Note that $\deg h_i \le d$, and that

$$\mathbb{E}_z \sum_{i=1}^k h_i(z)^2 = \mathbb{E}_z \sum_{S\subset A} \widehat{g_{\bar{A}\to z}}(S)^2 = \sum_{S\subset A}\sum_{T:T\cap A=S} \widehat{g}(T)^2 = \sum_{S\subset[n]} \widehat{g}(S)^2 = \|g\|_2^2.$$

The result follows from Lemma 1.17 and (1.1). □

1.3 Proof ideas

Lemma 1.19. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be balanced. Let $d \in \mathbb{N}$. Let $g : \{-1,1\}^n \to \mathbb{R}$ be such that $\deg g \le d$. Then*

$$\langle f, g\rangle \le 3^{d/4} I[f,g]\left(\max_{i\in[n]} I_i[f^{\le d}]\right)^{1/8}.$$

Proof. We have

$$\langle f, g\rangle = \langle f^{\le d}, g\rangle \le \sum_{i\in[n]} \langle \partial_i f^{\le d}, \partial_i g\rangle \le \sum_{i\in[n]} \|\partial_i f^{\le d}\|_2 \|\partial_i g\|_2$$

and

$$\begin{aligned}\|\partial_i f^{\le d}\|_2^2 = \langle \partial_i f^{\le d}, \partial_i f\rangle &\le \|\partial_i f^{\le d}\|_4 \|\partial_i f\|_{4/3} \le 3^{d/2}\|\partial_i f^{\le d}\|_2 \|\partial_i f\|_{4/3} \\ &\le 3^{d/2}\|\partial_i f^{\le d}\|_2^{1/2}\|\partial_i f\|_2^{1/2}\|\partial_i f\|_{4/3} = 3^{d/2} I_i[f^{\le d}]^{1/4} I_i[f].\end{aligned}$$

(Since f is Boolean-valued, $\partial_i f$ is $\{-1,0,1\}$-valued, so $\|\partial_i f\|_{4/3}^{4/3} = \|\partial_i f\|_2^2 = I_i[f]$.) Combining the above,

$$\langle f, g\rangle \le \sum_{i\in[n]} \|\partial_i f^{\le d}\|_2 \|\partial_i g\|_2 \le 3^{d/4} \sum_{i\in[n]} I_i[f^{\le d}]^{1/8} I_i[f]^{1/2} I_i[g]^{1/2}$$ □

As a corollary, we can deduce the KKL inequality (Theorem 1.5).

Proof of Theorem 1.5. Let $d = 2I[f]$. First,

$$\sum_{|S|>d} \widehat{f}(S)^2 \le \frac{1}{d} \sum_{|S|>d} |S|\widehat{f}(S)^2 \le \frac{1}{2}, \quad \text{so} \quad \langle f, f^{\le d}\rangle = \sum_{|S|\le d} \widehat{f}(S)^2 \ge \frac{1}{2}. \tag{1.2}$$

Combining this with Lemma 1.19 gives

$$\frac{1}{2} \le 3^{d/4} I[f, f^{\le d}] \left(\max_{i\in[n]} I_i[f^{\le d}]\right)^{1/8} \le 3^{I[f]/2} I[f] \left(\max_{i\in[n]} I_i[f^{\le d}]\right)^{1/8}$$

which implies $\max_{i\in[n]} I_i[f^{\le d}] \ge e^{-CI[f]}$. □

The following is useful when you don't assume anything about influences.

Lemma 1.20. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be balanced. Let $g : \{-1,1\}^n \to \mathbb{R}$ be of degree $\le d$. Then*

$$\langle f, g\rangle \le \delta^{-d} d^d \max_S |\widehat{f}(S)||\widehat{g}(S)| + \delta^{1/8} 3^{d/4} I[f,g] \qquad \text{for all } \delta > 0.$$

Proof. We may assume without loss of generality that $\widehat{f}(S)\widehat{g}(S) \ge 0$ for all $S \subset [n]$. Let $\mathsf{HI} = \{i \in [n] : I_i[f^{\le d}] \ge \delta\}$. ("HI" stands for "high influence.") Since

$$\sum_{i\in[n]} I_i[f^{\le d}] = \sum_S |S|\widehat{f^{\le d}}(S)^2 = \sum_{|S|\le d} |S|\widehat{f}(S)^2 \le d \sum_{|S|\le d} \widehat{f}(S)^2 = d,$$

Markov's inequality implies that $|\mathsf{HI}| \le \frac{d}{\delta}$. We split

$$\langle f, g\rangle = \sum_{S\subset \mathsf{HI}} \widehat{f}(S)\widehat{g}(S) + \sum_{S\not\subset \mathsf{HI}} \widehat{f}(S)\widehat{g}(S) \le \sum_{S\subset \mathsf{HI}} \widehat{f}(S)\widehat{g}(S) + \sum_{i\notin \mathsf{HI}} \sum_{S\ni i} \widehat{f}(S)\widehat{g}(S).$$

For the first term, we use the following trivial bound. (Recall that $\deg g \le d$.)

$$\sum_{S\subset \mathsf{HI}} \widehat{f}(S)\widehat{g}(S) \le |\{S : S \subset \mathsf{HI}, |S| \le d\}| \max_{S\subset \mathsf{HI}} \widehat{f}(S)\widehat{g}(S) \le \left(\tfrac{d}{\delta}\right)^d \max_{S\subset[n]} \widehat{f}(S)\widehat{g}(S)$$

For the second term, we argue exactly as in the proof of Lemma 1.19 to get

$$\|\partial_i f^{\le d}\|_2 \|\partial_i g\|_2 \le 3^{d/4} I_i[f^{\le d}]^{1/8} I_i[f]^{1/2} I_i[g]^{1/2} \le 3^{d/4} \delta^{1/8} I_i[f]^{1/2} I_i[g]^{1/2}$$ □

In the following lemma, we think of v as the degree we are restricting to.

Lemma 1.21. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be balanced. Let $g : \{-1,1\}^n \to \mathbb{R}$ be homogeneous of degree d. Let $v \ll d, \delta > 0$. Then*

$$\langle f, g\rangle \le \delta^{-v} e^{O(d)} \max_{S\subset A} \left(\sum_{T\cap A = S} \widehat{g}(T)^2\right)^{1/8} + \delta^{1/8} e^{O(v)} I[f,g]$$

Proof. The idea is to apply Lemma 1.15 and Corollary 1.18. □

In the following, think of v as the degree we are restricting to, and δ_2 as a bound on the influence of coalitions. (The influence of the coalition $A \subset [n]$ is $I_A[f] = \sum_{S\supset A} \widehat{f}(S)^2$.)

Lemma 1.22 ($k = 2$ case of [1, Theorem 4.1])**.** *Let $f : \{-1,1\}^n \to \{-1,1\}$ be balanced. Let $g : \{-1,1\}^n \to \mathbb{R}$ be homogeneous of degree d. Let $v \le d, \delta > 0, \delta_2 \le \delta$. Then*

$$\langle f, g\rangle \lesssim (2/\delta_2)^{d/v} d^d \max_{|S|=d} |\widehat{f}(S)||\widehat{g}(S)| + (2/\delta)^v v^v e^{O(d)} \delta_2^{1/4} + \delta^{1/8} e^{O(v)} I[f,g].$$

We now show that Lemma 1.22 implies a weaker version of Theorem 1.8.

Corollary 1.23. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be balanced. Then*

$$H_\infty[\widehat{f^2}] \le CI[f]^{3/2} \tag{1.3}$$

Proof. First, note that the inequality (1.3) is equivalent to $M := \max_S |\widehat{f}(S)|^2 \ge \exp(-CI[f]^{3/2})$. We begin with the same inequality as (1.2):

$$\sum_{1\le d\le 2I[f]} \langle f, f^{=d}\rangle = \langle f, f^{\le 2I[f]}\rangle = \sum_{|S|\le 2I[f]} \widehat{f}(S)^2 \ge \frac{1}{2}$$

Since $\sum_{d=1}^{\infty} \frac{1}{100d^2} < \frac{1}{2}$, there exists some $1 \le d \le 2I[f]$ such that $\langle f, f^{=d}\rangle \ge \frac{1}{100d^2}$. By Lemma 1.22 applied to this d, we have

$$\frac{1}{d^2} \lesssim (2/\delta_2)^{d/v} d^d M + (2/\delta)^v v^v e^{O(d)} \delta_2^{1/4} + \delta^{1/8} e^{O(v)} I[f] \tag{1.4}$$

where v, δ, δ_2 are to be chosen later (depending on d). The idea for choosing these three parameters is as follows. We choose δ small to make the third term small. Then we choose δ_2 small to make the second term small. This leaves us with only the first term, so we get a lower bound on M. However, if we choose δ_2 too small, then the lower bound on M is not very good. So we choose v so that δ_2 does not need to be too small.

First we choose δ small enough so that

$$\log \frac{1}{\delta} \approx v + \log I[f] + \log d.$$

This choice of δ implies $\delta^{1/8} e^{O(v)} I[f] \ll \frac{1}{d^2}$. Next we choose δ_2 small enough so that

$$\log \frac{1}{\delta_2} \approx v \log \frac{1}{\delta} + v \log v + d \approx v^2 + v \log I[f] + v \log d + d \tag{1.5}$$

This choice of δ_2 implies $(2/\delta)^v v^v e^{O(d)} \delta_2^{1/4} \ll \frac{1}{d^2}$. Thus, (1.4) implies

$$\frac{1}{d^2} \lesssim (2/\delta_2)^{d/v} d^d M. \tag{1.6}$$

By choosing $v \approx \sqrt{d}$ and noting that $d \le 2I[f]$, (1.5) gives us $\log \frac{1}{\delta_2} \lesssim I[f]$ so (1.6) becomes

$$\frac{1}{I[f]^2} \lesssim e^{CI[f]^{3/2}} (2I[f])^{2I[f]} M$$

which implies the claim. □

Bibliography

[1] Kelman E., Kindler G., Lifshitz N., Minzer D., Safra M., *Towards a proof of the Fourier-entropy conjecture?*. Geometric and Functional Analysis. 2020 Aug;30(4):1097-138.

Alan Chang, Princeton University
email: **alanchang@math.princeton.edu**

Chapter 2

Learning DNF from Random Walks

after N. Bshouty, E. Mossel, R. O'Donnell and R. A. Servedio [1]
A summary written by Fan Chang

Abstract. Bshouty et al. [1] obtained the first passive learning algorithm for DNFs under the Random Walk model and the Noise Sensitivity learning model. We summary their ideas and mention further relevant researches.

2.1 Introduction

A *concept class* $\mathcal{C}$ is a set of Boolean functions. Every function $f \in \mathcal{C}$ is a concept. In the PAC model developed by Valiant [8] a learner tries to approximate with high probability an unknown concept f from a training set of N random labelled examples $\{(x_i, f(x_i))\}_{i=1}^{N}$. The examples are given by an *example oracle* $\mathrm{EX}(f, \mathcal{D})$ that returns an example $(x, f(x))$, where x is randomly sampled from a probability distribution $\mathcal{D}$ over $\{\pm 1\}^n$. A learning algorithm A for $\mathcal{C}$ takes as input an *accuracy parameter* $\varepsilon \in (0,1)$, a *confidence parameter* $\delta \in (0,1)$ and the training set and outputs a *hypothesis* h that is a good approximation of f with probability $1-\delta$. We say that a concept class $\mathcal{C}$ is *PAC-learnable* if, for every $\mathcal{D}, f, h, \delta$, when running a learning algorithm A on $N \geq N_{\mathcal{C}}$ examples generated by $\mathcal{D}$, we have that, with probability at least $1-\delta$, $\Pr_{x\sim\mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$. PAC theory introduces two parameters to classify the efficiency of a learner. The first one, $N_{\mathcal{C}}$, is information-theoretic and determines the minimum number of examples required to PAC-learnable the class $\mathcal{C}$. We refer to $N_{\mathcal{C}}$ as the *sample complexity* of the concept class $\mathcal{C}$. The second parameter, the *time complexity*, is computational and corresponds to the runtime of the best learner for the class $\mathcal{C}$. We say that a concept class is *efficiently* PAC-learnable if the running time of A is polynomial in n, $\frac{1}{\varepsilon}$ and $\frac{1}{\delta}$.

Whether the class of Boolean functions that can be expressed as polynomial size formulae in DNF is efficiently PAC-learnable **from random examples on points sampled from an unknown distribution**, or not, is one of the central unresolved questions. The best classical algorithm for this problem has running time $2^{\tilde{O}(n^{1/3})}$ [6] (The notations $\tilde{O}(g(n))$ and $\tilde{\Omega}(g(n))$ hide logarithmic factors).

- In uniform (or product) distribution model, a simple quasi-polynomial $n^{O(\log n)}$ algorithm for learning DNF expressions was found by Verbeurgt [9].
- In Membership Query model, Jackson [3] gave a polynomial time learning algorithm for DNFs over product distributions.
- Bshouty and Feldma [2] gave a polynomial time learning algorithm polynomial-sized DNF under a model of intermediate power between uniform-distribution learning and uniform-distribution learning with membership queries. (non-passive)
- Bshouty et al. [1] obtained the first passive learning algorithm for DNFs under the Random Walk model and the Noise Sensitivity learning model.
- Kalai, Samorodnitsky and Teng [5] show that DNF expressions are efficiently PAC-learnable over smoothed product distributions.

2.2 The Random Walk learning model

In (uniform) Random Walk (RW) model, our concept classes $\mathcal{C}$ and hypothesis are real-valued functions $f, h : \{\pm 1\}^n \to \mathbb{R}$. The first labeled example is $x_0 \sim \{\pm 1\}^n$. Following this, the examples are generated by a standard random walk on the hypercube. That is, if the jth example given to the learner is x, then the $(j+1)$th example will be chosen by selecting a coordinate $i \in [n]$ uniformly at random and then flipping x_i.

An equivalent model is easier to work, called the Random Walk model with *updating oracle* (URK). In URK model, again the first example is given uniformly at random. Further, the example given at time $j+1$ depends on the previous example x. The updating oracle picks a coordinate $i \in [n]$ uniformly at random, then *updates* the x_i, producing y (updating the bit x_i means replacing x_i with a uniformly random bit). Finally, the updating oracle tells the learner $(i, y, f(y))$. Note that the Random Walk model is a passive model of learning; the learner sees only randomly generated examples and has no control over the data used for learning.

Remark 2.1.

- *Membership queries (MQ) model is at least as powerful as RW model. In other words, one can show that uniform-distribution learning with MQ is strictly easier than learning in the RW model, under a standard cryptographic assumption.*
- *RW model is at least as easy as PAC learning under the uniform distribution. (This is because the updating random walk on the hypercube mixes rapidly; if a learner discards $O(n \log n)$ successive examples from the updating oracle, then the next example will be uniformly random and independent of all previous examples).*

2.3 A Noise Sensitivity learning model

Given $x \in \{\pm 1\}^n$ and $0 \leq \gamma \leq 1$, we define $N_\gamma(x)$ to be the random variable taking values in $\{\pm 1\}^n$ given by flipping each coordinate of x independently with probability γ. For $\gamma \leq \frac{1}{2}$, N_γ is equivalently defined by *updating* each coordinate in x independently with probability 2γ.

As special cases, note that $N_0(x)$ is the constant random variable x, $N_{\frac{1}{2}}(x)$ is a uniform random vector independent of x, and $N_1(x)$ is constantly the vector with Hamming distance n from x.

Definition 2.2. *Given a Boolean function $f : \{\pm 1\}^n \to \{\pm 1\}$ and $0 \leq \gamma \leq 1$, the* noise sensitivity *of f at γ is*

$$\mathrm{NS}_\gamma(f) = \Pr_{x, y = N_\gamma(x)}[f(x) \neq f(y)].$$

For $\gamma \in [0, \frac{1}{2}]$, we define the *$\gamma$-Noise Sensitivity* learning (NS) model as follows: Given a target function $f : \{\pm 1\}^n \to \mathbb{R}$, the learner has access to the "Noise Sensitivity oracle", NS-EX$_\gamma(f)$. Every time the learner asks for an example, NS-EX$_\gamma(f)$ independently chooses a random input $x \in \{\pm 1\}^n$, forms $y = N_\gamma(x)$, and tells the learner $(x, f(x), y, f(y))$. Note that this oracle is equivalent to an "updating" Noise Sensitivity oracle, in which each coordinate of x is updated with probability 2γ.

Remark 2.3. *The cases $\gamma = 0$ and $\gamma = \frac{1}{2}$ are trivially equivalent to the usual PAC model under the uniform distribution. For values $\gamma \in (0, \frac{1}{2})$, learning with NS-EX$_\gamma(f)$ is clearly at least as easy as learning under the uniform distribution.*

Let us see the relationship between NS model and RW model:

Proposition 2.4. *For any $\gamma \in [0, \frac{1}{2}]$, any γ-Noise Sensitivity learning algorithm can be simulated in the Random Walk model with only a multiplicative $O(n \log n)$ slowdown in running time.*

Our main theorem is the following:

Theorem 2.5 (Bshouty, Mossel, O'Donnell and A.Servedio [1])**.** *The class of s-term DNF formulas on n variables can be learned in the Random Walk model to accuracy ε and confidence $1 - \delta$ in time poly$(n, s, \frac{1}{\varepsilon}, \log(\frac{1}{\delta}))$.*

2.4 Performing the Bounded Sieve in the Noise Sensitivity model

We prove Theorem 2.5 by showing that polynomial-sized DNF can be learned under any γ-Noise Sensitivity learning model.

Theorem 2.6. *Let $\gamma \in (0, \frac{1}{2})$, and let $c_o = -\log(\gamma(\frac{1}{2} - \gamma))$ be a constant when γ is a constant. Then the class of polynomial-sized DNF formulas on n variables can be learned in γ-Noise Sensitivity model in time poly$(n^{c_0}, \varepsilon^{-c_0}, \log(\frac{1}{\delta}))$.*

To prove Theorem 2.6, we give an algorithm that, given access to the oracle NS-EX$_\gamma(f)$, finds all the "large" Fourier coefficient $\hat{f}(S)$ which satisfy $|S| \leq O(\log n)$.

Theorem 2.7. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a target function, and let $\gamma \in (0, \frac{1}{2})$. Fix parameters $\ell \in [n]$ and $\theta > 0$. Then there is an algorithm with running time poly$(n, [\gamma(\frac{1}{2} - \gamma)]^{-\ell}, \theta^{-1}, \|f\|_\infty, \log(\frac{1}{\delta}))$ which, given access to the oracle NS-EX$_\gamma(f)$, with probability $1 - \delta$ return a list of subsets of $[n]$ such that*

- *for each $S \subseteq [n]$, if $|S| \leq b$ and $\hat{f}(S)^2 \geq \theta$, then S is in the list; and,*
- *for each set S in the list, $|S| \leq b$ and $\hat{f}(S)^2 \geq \theta/2$.*

Bshouty and Feldman call the task performed by this algorithm the *Bounded Sieve*. It is a weak version of the Kushilevitz–Mansour algorithm which find *all* large Fourier coefficients. The proof of Theorem 2.7 need the following noise sensitivity-like quantity:

Definition 2.8. *Given $f : \{\pm 1\}^n \to \mathbb{R}, \gamma \in (0, \frac{1}{2})$, and $I \subseteq [n]$, define*

$$\mathcal{T}_\gamma^{(I)}(f) = \sum_{S \supseteq I} (1 - 2\gamma)^{|S|} \hat{f}(S)^2.$$

When f and γ are clear from context, we write simply $\mathcal{T}(I)$.

Note that $\mathcal{T}_\gamma^{(\emptyset)}(f) = 1 - 2\mathrm{NS}_\gamma(f)$. We can use the NS oracle to estimate the quantities $\mathcal{T}_\gamma^{(I)}(f)$:

Lemma 2.9. *For fixed constant $\gamma \in (0, \frac{1}{2})$ and target function $f : \{\pm 1\}^n \to \mathbb{R}$, an algorithm with access to NS-EX$_\gamma(f)$ can, with probability $1-\delta$, estimate $\mathcal{T}(I)$ to within $\pm\eta$ in running time poly$(n, \frac{1}{\gamma^{|I|}}, \|f\|_\infty, \frac{1}{\eta}, \log(\frac{1}{\delta}))$.*

Proof. Given γ and I, consider the joint probability distribution $\mathcal{D}_\gamma^{(I)}$ defined over pairs $(x, y) \in (\{\pm 1\}^n)^2$ as follows: First x is chosen uniformly at random; then y is formed by updating each coordinate of x in I with probability 1 and updating each coordinate of x not in I with probability 2γ.

Claim1. Access to these pairs and their value under f can be simulated by access to NS-EX$_\gamma(f)$, with slowdown poly$(\gamma^{-|I|})$.

Define

$$\mathcal{T}'(I) = \mathop{\mathbb{E}}_{(x,y) \in \mathcal{D}_\gamma^{(I)}} [f(x) f(y)].$$

Since we can simulate access to pairs from $\mathcal{D}_\gamma^{(I)}$ and their values under f, we can estimate $\mathcal{T}'(I)$ simply by taking many samples and averaging. By standard arguments we can compute a $\pm\eta$ approximation with probability $1 - \delta$ in running time poly$(n, \|f\|_\infty, \frac{1}{\eta}, \log(\frac{1}{\delta}))$.

Claim2. $\mathcal{T}'(I) = \sum_{S \cap I = \emptyset} (1 - 2\gamma)^{|S|} \hat{f}(S)^2$.

Let us now define $\mathcal{T}''(I) = \mathcal{T}'(\emptyset) - \mathcal{T}'(I)$, again a quantity we can estimate in running time poly$(n, \frac{1}{\gamma^{|I|}}, \|f\|_\infty, \frac{1}{\eta}, \log(\frac{1}{\delta}))$. We have

$$\mathcal{T}''(I) = \sum_{S \cap I \neq \emptyset} (1 - 2\gamma)^{|S|} \hat{f}(S)^2.$$

Thus if we compute $\mathcal{T}''(J)$ for all $J \subseteq I$, it is straightforward to calculate $\mathcal{T}(I) = \sum_{S \supseteq I} (1 - 2\gamma)^{|S|} \hat{f}(S)^2$ using inclusion-exclusion. Since there are only $2^{|I|} \leq \gamma^{-|I|}$ such subsets J, the claimed running time follws. $\square$

Next, we note that the sum of the $\mathcal{T}_\gamma^{(I)}$ values across all $|I| = j$ is not too large:

Lemma 2.10. *For any $f : \{\pm 1\}^n \to \mathbb{R}$ and $\gamma \in (0, \frac{1}{2})$, we have*

$$\sum_{|I|=j} \mathcal{T}_\gamma^{(I)} \leq \|f\|_\infty^2 (2\gamma)^{-j}.$$

Using these Lemmas we can now complete the proof of Theorem 2.7.

Proof of Theorem 2.7. **Construction.** Consider the directed graph $D = (V, E)$ where $V(D) = 2^{[n]}$ and $E(D) = \{I \sim J \text{ iff } I \subset J \text{ and } |J \setminus I| = 1\}$. The vertices I are divided into n layers according to the value of $|I|$.

Our Bounded Sieve algorithm for f performs a *breadth-first search* on this graph, starting at the vertex $I = \emptyset$. For each active vertex in the search, the algorithm estimates $\mathcal{T}(I)$ and $\hat{f}(I)^2$.

- If the estimate of $\hat{f}(I)^2$ is at least $\theta/2$ then the algorithm adds I to the list of f's large Fourier coefficient.
- The breadth-first search proceeds to the neighbors of I only if $|I| < \ell$ and the estimate of $\mathcal{T}(I)$ is at least $(1 - 2\gamma)^\ell \theta/2$.

Claim 1. The algorithm finds with high probability
(i) all Fourier coefficients $\hat{f}(S)$ with $\hat{f}(S)^2 \geq \theta$ and $|S| \leq \ell$;
(ii) The algorithm end its search within running time $\text{poly}(n, [\gamma(\frac{1}{2} - \gamma)]^{-\ell}, \theta^{-1}, \|f\|_\infty, \log(\frac{1}{\delta}))$.

To see the first claim, simply note that if $|S| \leq \ell$ and $\hat{f}(S)^2 \geq \theta$, then this Fourier coefficient contributes at least $(1 - 2\gamma)^\ell \theta$ to the value of $\mathcal{T}(I)$ for all $I \subseteq S$. Thus by the monotonicity of $\mathcal{T}$, the search will proceed all the way to S. So long as all estimations are taken to be sufficiently precise (compared to the quantity $(1 - 2\gamma)^\ell \theta/2$).

The second claim comes from Lemma 2.10 and some counting argument like the pigeonhole principle. □

In fact, we can directly estimate $\mathcal{T}_\gamma^{(I)}(f)$ under the Random Walk model.

2.5 Further research

A natural generalization of DNFs to $\mathbb{F}_m^n$ was given in : for each $1 \leq i \leq n$, choose two values $0 \leq a_i \leq b_i \leq m - 1$, and consider the *rectangle*:$[a, b] = \{x \in \mathbb{F}_m^n : a_i \leq x_i \leq b_i, \forall i\}$. An instance of UBOX is a union of rectangles. So in the Boolean case, a DNF can be seen as a union of subcubes of $\mathbb{F}_2^n$.

- Roch [7] proved that Harmonic Sieve can be performed efficiently in the Cyclic Random Walk (CRW) model (In (CRW) model, the first example is uniform over $\mathbb{F}_m^n$ and then follow a random walk where at each step, instead of picking a uniformly random coordinate to update (flip), according to a fixed cycle $(i_1, \ldots, i_n)$ running through all of $[n]$. Thus DNFs, TOPs and UBOXs are (δ, ε)-learnable in the CRW model. (polynomial-weight threshold-of-parity circuit (TOP))
- Roch [7] proved that UBOXs are (δ, ε)-learnable in the NS model, and RW model.
- Jackson and Wimmer [4] give a quasi-polynomial algorithm for learning TOP in the RW model.
- Jackson and Wimmer [4] proved that DNF formulas can be efficiently learned in the p-biased version of RW model and NS model.

Bibliography

[1] Nader H Bshouty et al., *Learning DNF from random walks.* Journal of Computer and System Sciences 71.3 (2005), pp. 250-265.

[2] Nader H. Bshouty and Vitaly Feldman, *On using extended statistical queries to avoid membership queries.* Journal of Machine Learning Research 2.3 (2002), pp. 359-395.

[3] Jeffrey C Jackson,*An efficient membership-query algorithm for learning DNF with respect to the uniform distribution.* Journal of Computer and System Sciences 55.3 (1997), pp. 414-440

[4] Jeffrey C. Jackson and Karl Wimmer, *New results for random walk learning.* Journal of Machine Learning Research 15 (2014), pp. 3635-3666.

[5] Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng, *Learning and smoothed analysis.* 2009 50th Annual IEEE Symposium on Foundations of Computer Science. IEEE. 2009, pp. 395-404.

[6] Adam R. Klivans and Rocco A. Servedio, *Learning DNF in time* $2^{\tilde{O}(n^{1/3})}$. Journal of Computer and System Sciences 68.2 (2004), pp. 303 C318.

[7] Sébastien Roch, *On learning thresholds of parities and unions of rectangles in random walk models.* Random Structures Algorithms 31.4 (2007), pp. 406-417.

[8] Leslie G Valiant, *A theory of the learnable.* Communications of the ACM 27.11 (1984), pp. 1134-1142.

[9] Karsten Verbeurgt. *Learning DNF under the uniform distribution in quasi-polynomial time.* Proceedings of the third annual workshop on Computational learning theory. 1990, pp. 314-326.

Fan Chang, shandong university
email: **cf25264@163.com**

Chapter 3

Invariance via Polynomial Decompositions

after D. Kane [1]
A summary written by Jacob Denson

Let $X \in \mathbb{R}^N$ be a random vector with independent coordinates. The *invariance principle* says that if $f : \mathbb{R}^N \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ are 'regular', then the quantities $\mathbb{E}[\psi(f(X))]$ depend only on very coarse properties of the distribution of X up to a small error. A basic instance is the central limit theorem, which says that a sum of independent random variables is approximately normally distributed, and thus independent of all properties of those variables but their mean and variance, stated below.

Theorem 3.1 (Lindeberg)**.** *Let X and A be random vectors in $\mathbb{R}^N$, each having independent coordinates, and sharing the same means and variances. Let*

$$\gamma = \sum_i \mathbb{E}|X_i|^3 + \mathbb{E}|A_i|^3.$$

Let $f(z) = z_1 + \cdots + z_N$. Then for any $\psi : \mathbb{R} \to \mathbb{R}$,

$$|\mathbb{E}[\psi(f(X))] - \mathbb{E}[\psi(f(A))]| \leq \|D^3\psi\|_{L^\infty(\mathbb{R})} \cdot \frac{\gamma_3}{6}.$$

One can still exploit this theorem to get transference principles which apply to less regular f, for instance, for quantities with a simple jump discontinuity like if $\psi(t) = \mathbf{I}(t \leq s)$, for which $\mathbb{E}[\psi(f(X))] = \mathbb{P}(f(X) \leq s)$ gives the CDF of the random variable $f(X)$, or $\psi(t) = \operatorname{sgn}(t)$, in which case $\psi(f(X))$ is called a *threshold function*. To get these theorems, we apply an additional *anticoncentration inequality*. Let's see why in Lindeberg's scenario: pick a non-negative $\eta \in C^\infty$ supported on $|t| \leq 1$ and with $\int \eta(x)\, dx = 1$, and we define $\psi_\varepsilon = \psi * \operatorname{Dil}_\varepsilon \eta$, then $\|D^3\psi\|_{L^\infty} \lesssim \varepsilon^{-3}$, and since $\psi(t) \leq \psi_\varepsilon(t-\varepsilon) \leq \psi(t-2\varepsilon)$,

$$\begin{aligned}
\mathbb{P}(f(A) \leq s) &\leq \mathbb{E}[\psi_\varepsilon(f(A) - \varepsilon)] \\
&\leq \mathbb{E}[\psi_\varepsilon(f(X) - \varepsilon)] + O\left(\varepsilon^{-3}\gamma_3\right) \\
&\leq \mathbf{P}(f(X) \leq s + 2\varepsilon) + O(\varepsilon^{-3}\gamma_3).
\end{aligned}$$

Similarily, one shows $\mathbb{P}(f(A) \leq s) \leq \mathbb{P}(f(X) \leq s - 2\varepsilon) + O(\varepsilon^{-3}\gamma_3)$. Thus to finish this argument and show that $\mathbb{P}(f(A) \leq s) \approx \mathbb{P}(f(X) \leq s)$, we must show that $\mathbb{P}(s - 2\varepsilon \leq f(X) \leq s + 2\varepsilon)$ is small, i.e. that $f(X)$ *does not concentrate*. If for simplicity we assume $f(X)$ and $f(A)$ both have variance one, then we find that $\mathbb{P}(s - 2\varepsilon \leq f(X) \leq s + 2\varepsilon) \lesssim \varepsilon$, and plugging this in gives that $|\mathbb{P}(f(A) \leq s) - \mathbb{P}(f(X) \leq s)| \lesssim \varepsilon + \varepsilon^{-3}\gamma_3$. Picking $\varepsilon = \gamma_3^{1/4}$ gives an error $O(\gamma_3^{1/4})$. We have thus proved the *Berry-Esseen theorem* by means of an *anticoncentration inequality* for the Gaussian.

The paper [1] we discuss here studies anticoncentration inequalities for random quantities $f(X)$, where f is no longer a linear sum, but a *polynomial* p with an independent vector as inputs. For instance, one might want to study $\psi(f(X))$ with $\psi(x) = \operatorname{sgn}(x)$, quantities called *polynomial threshold*

functions. There are results already existing in the literature that give general anticoncentration bounds for general polynomials of a fixed degree (a result of Carbery-Wright), and this result is tight for general polynomials. The paper [1] gives tools indicating a way to identify polynomials for which one can *improve* this anticoncentration result, via a decomposition of this polynomial. One consequence is more sophisticated invariance principles for polynomials, with better error terms if a polynomial has the right decompositions.

In the following summary, we write X, Y and Z for standard normal vectors, and let A and B be Bernoulli random vectors. We write γ for the normal Gaussian distribution on $\mathbb{R}^N$, and β for the Bernoulli distribution on $\{-1,+1\}^N$. We thus have norms $L^p_\gamma(\mathbb{R}^N)$ and $L^p_\beta(\mathbb{R}^N)$ for functions $f:\mathbb{R}^N\to\mathbb{R}$ given by

$$\|f\|_{L^p_\gamma}=\mathbb{E}[|f(X)|^p]^{1/p} \quad \text{and} \quad \|f\|_{L^p_\beta}=\mathbb{E}[|f(A)|^p]^{1/p}.$$

Similarily, we have variances $\mathrm{Var}_\gamma(f)$ and $\mathrm{Var}_\beta(f)$. The ith influence $\mathrm{Inf}_i(f)$ is defined as $\|\partial f/\partial x_i\|^2_{L^2_\gamma}$. This agrees with the standard definition of influence that occurs in the analysis of Boolean functions, in the case that f is a multilinear polynomial. A k-tensor on $\mathbb{R}^N$ is a quantity of the form

$$\sum A_S dx^{\otimes S}$$

where $\{A_S\}$ are real numbers, and S ranges over $[k]^S$. A k-tensor valued function is

$$A(x)=\sum A_S(x)dx^{\otimes S}.$$

The magnitude $|A|$ of a k-tensor is equal to $(\sum |A_S|^2)^{1/2}$, and using this we can define the L^p_γ and L^p_β norms of a k-tensor valued function in the way you would expect, i.e. as $\mathbb{E}[|A(X)|^p]^{1/p}$ and $\mathbb{E}[|A(B)|^p]^{1/p}$.

3.1 The Main Result

For general polynomials p of a fixed degree d, Carbery-Wright showed that

$$\mathbb{P}(|p(X)|\le\varepsilon\|p\|_{L^2_\gamma})\lesssim d\varepsilon^{1/d}.$$

This result is tight, for instance, if $p(x)=(x_1+\cdots+x_N)^d$, or $p(x)=q_1(x)^7+q_2(x)^7+q_1(x)^2q_2(x)^2+\delta q_3$, where q_1 and q_2 are polynomials of degree d and δ is small. But the $\varepsilon^{1/d}$ error term leads to invariance principles which have poor dependence on d, i.e. the following result.

Theorem 3.2 (Mossel, O'Donnell, Oleszkiewicz)**.** *If p is a τ-*regular *multilinear polynomial of degree d, i.e. $Inf_i(p)\le\tau\,Var_\beta(p)$ for all indices i, then*

$$|\mathbb{P}(p(X)\le t)-\mathbb{P}(p(A)\le t|\lesssim d\tau^{1/8d}.$$

Given the poor dependence on d here (tight for general inputs), to obtain better invariance principles it is useful to identify those particular scenarios in which we can improve upon the general result of Carbery-Wright, or equivalently, to identify all obstacles which make the Carbery-Wright inequality tight. Notice that the tight examples to Carbery-Wright are of the form $h(q_1,\dots,q_m)$, where h is a poorly behaved polynomial, and (q_1,q_2) has good anticoncentration results. This is indeed true of all badly behaved counterexamples up to a small error term, which is the main result to be discussed.

We begin with some definitions. We say a vector $q=(q_1,\dots,q_m)$ of polynomial functions $q_i:\mathbb{R}^N\to\mathbb{R}$ is (ε,α) *diffuse* if for any $a\in\mathbb{R}^m$,

$$\mathbb{P}(|q(X)-a|\le\varepsilon)\le\varepsilon^m\alpha.$$

Intuitively, this means the probability density of the random vector $q(X)$ has average value at most α on any box of sidelength α. We say a polynomial $p:\mathbb{R}^N\to\mathbb{R}$ has a *decomposition* into $h(q_1,\dots,q_m)$ for $h:\mathbb{R}^m\to\mathbb{R}$ and $q_i:\mathbb{R}^N\to\mathbb{R}$ if $p=h(q_1,\dots,q_m)$, and if, for any monomial $\prod_{i\in\beta}x_i^\beta$ occuring in h, and monomials $x^{\beta_1},\dots,x^{\beta_m}$ occuring in $q_1,\dots,q_m$ respectively, $\deg(\prod_{i\in\beta}x^{\beta_i})\le\deg(p)$. This is to prevent some decomposition where high degree terms in the decomposition cancel each other out, which complicates the analysis of the random variables involved. We can now state the structure result for polynomials, which gives the main result of [1].

Theorem 3.3. *For any degree d polynomial p, and any $\varepsilon, N, c > 0$, there exists a degree d polynomial p_0, a polynomial vector $q = (q_1, \ldots, q_m)$, and a polynomial $h : \mathbb{R}^m \to \mathbb{R}$ such that p_0 has a decomposition into $h(q_1, \ldots, q_m)$, and:*

- *$p \approx p_0$ in the quantitative sense that $\|p - p_0\|_{L^2_\gamma} \lesssim_{c,d,N} \varepsilon^N \|p\|_{L^2_\gamma}$.*
- *$(q_1, \ldots, q_m)$ is $(\varepsilon, \varepsilon^{-c})$ diffuse, and $m \lesssim_{c,d,N} 1$.*

The currently known dependence on m on c, d, and N is currently very poor, i.e. that $m \leq A(d + O(1), N/c)$, where A is the Ackerman function. But it is conjectured that one can find a bound which is polynomial in (dN/c).

This structure result is closely related to the characterization of polynomials for which concentration does not occur. A polynomial for which Theorem 2 is tight (in the sense of the dependence of the result on $\tau^{-1/d}$), for an even dimension d, is the multilinear projection q of the polynomial

$$p(x_0, \ldots, x_N) = \tau x_0 + \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i\right)^d = p_0(x) + p_1(x)^d.$$

As $N \to \infty$, $\lim_{N\to\infty} \|p - q\|_{L^2_\gamma} = 0$, so results like hypercontractivity imply that the distributions of $p(X) - q(X)$ are very close, i.e. for any $\delta > 0$, there exists $N_0 > 0$ such that if $N \geq N_0$, then

$$\mathbb{P}(|p(X) - q(X)| \geq \delta A) \lesssim 2^{-c_d A^{2/d}}.$$

Now $\mathrm{Inf}_0(q) = \tau^2$, $\mathrm{Inf}_i(q) \lesssim_d 1/N$, and $\mathrm{Var}_\beta(q) \gtrsim 1 + \tau^2$. Thus q is τ-regular for $\tau \lesssim 1$, and so Theorem 2 applies to p. Note that since A is $\{-1, 1\}^{N+1}$ valued, we always have $p(A) \geq -\tau$. On the other hand, if X is Gaussian, $p_1(X_1, \ldots, X_N)$ then we can guarantee that $|p_1(X_1, \ldots, X_N)| \lesssim \tau^{1/d}$ with probability $\gtrsim \tau^{1/d}$, so $|Lq_1(X_1, \ldots, X_N)| \lesssim \tau^{1/d}$ with probability $\gtrsim \tau^{1/d}$. On the other hand, we guarantee that $\tau X_0 \leq -2\tau$ with probability $\gtrsim 1$. Thus by independence, *both properties* hold with probability $\gtrsim \tau^{1/d}$, and in this case $p(X) \leq -\tau$. Thus

$$|\mathbb{P}(p(X) \leq -\tau) - \mathbb{P}(p(A) \leq -\tau)| = \mathbb{P}(p(X) \leq -\tau) \gtrsim \tau^{1/d}.$$

What happened here? Even though the first coordinate has small influence, the Carbery-Wright result was tight for the polynomial $p_1(X)^d$, i.e. the polynomial concentrated within a $O(\tau)$ neighborhood of zero with probability $\Omega(\tau^{-1/d})$. The Boolean polynomial $p_0(A)^d$ also concentrates within a $O(\tau)$ neighborhood of zero with probability. In this situation, $p(X) \approx \tau X_0$ and $p(A) \approx \tau A_0$, and this causes a problem since X_0 and A_0 are *very* different probability distributions.

Theorem 3.4. *We say a degree d multilinear polynomial has a $(\tau, \alpha, m, \varepsilon)$ regular decomposition if there exists a polynomial p_0 of degree d such that*

$$\|p - p_0\|_{L^2_\beta} \leq \varepsilon \cdot Var_\gamma(p_0(X))^{1/2}$$

and $p_0 = h(q_1, \ldots, q_m)$, where $(q_1, \ldots, q_m)$ is a vector of multilinear polynomials which is $(\tau^{1/5}, \alpha)$ diffuse and $Inf_j(q_i) \leq \tau$ for all i and j. Under these conditions, for $0 < \tau, \varepsilon < 1/2$, we have

$$|\mathbb{P}(p(A) \leq t) - \mathbb{P}(p(X) \leq t)| \lesssim_{d,m} \tau^{1/5} \alpha \log(1/\tau)^{dm/2+1} + \varepsilon^{1/d} \log(1/\varepsilon)^{1/2}.$$

For $p_0(x) = \tau x_0$ and $p_1(x) = (x_1 + \cdots + x_N)/\sqrt{N}$, the theorem above can only apply with $\alpha = \tau^{-1/5}$, which yields a relatively useful error term of $O(\log(1/\tau))$. Thus the assumptions of this theorem avoid this kind of concentration phenomenon. To recover Theorem 2 from this result, the regularity assumption there implies that the polynomials $(x_0, x_1, \ldots, x_n, p)$ are $(\tau^{1/5}, O(d\tau^{(1/d-1)/5}))$ diffuse, and so the theorem above gives that

$$|\mathbb{P}(p(X) \leq t) - \mathbb{P}(p(A) \leq t)| \lesssim_d \tau^{1/5d} \log(1/t)^{d/2+1},$$

which is analogous to Theorem 2 in the sense that we still get a power of $\tau^{1/d}$.

Even if a polynomial does not satisfy the regularity conditions, this might only be true a 'few coordinates are bad', and we can obtain a polynomial by fixing a few values of the polynomial. Results showing this are true are called 'regularity lemmas'. Here is a result applying to the assumptions of Theorem 2.

Theorem 3.5 (Diakonix, Servedio, Tan, Wan). *If $f = sgn(p(x))$ is a polynomial threshold function of degree d, then there exists a decision tree of depth $\tau^{-1}(d\log(1/\tau))^{O(d)}$ such that a random root of this tree is τ-close to a τ-regular polynomial threshold function of degree d.*

Thus one can make a function τ regular by 'fixing' $\tau^{-1}(d\log(1/\tau))^{O(d)}$ different inputs, for most input values. [1] gets an analogous result which applies to the assumptions of the theorem above, i.e. that there is a decision tree of depth $\tau^{-1}(d\log(1/\tau))^{O(d)}$ such that with probability $1-\tau$, a random root either has a $(\tau, \tau^{-c}, O(1), O(\tau^M))$ regular decomposition, or has variance less than τ^M times the square of it's L^2_γ norm. To the former case, one can apply the theorem above by fixing variables. In the latter case, the function is roughly constant, i.e. it is incredibly highly concentrated, and thus can also be easily understood.

3.2 The Idea of the Proof

Due to space constraints, we only discuss the idea of the proof of Theorem 3. We emphasize the main principles upon which the proof lies, and the reason for such a poor dependence of m on d, c, and N, without introducing too much numerology, and also concentrating on the case where p is quadratic, since it is characteristic of the more complicated case.

The first principle is a heuristic that the author developed in a previous paper called *strong anti-concentration*. The result says that for a polynomial p, with high probability we have $p(X) \gtrsim \nabla p(X)$. Intuitively this is true because if $p(X)$ is significantly less than $\nabla p(X)$ at some value of X, a shift in the value of X will drastically effect $p(X)$, so that not many points will satisfy $p(X) \lesssim \nabla p(X)$ around this bad point. A significant part of this paper is extending this intuition to *tensors with polynomial coefficients*. For a k tensor $A = \sum A_S dx^{\otimes S}$ with low degree polynomial coefficients, the author shows that with a good probability,

$$|A_1 \otimes \cdots \otimes A_l| \gtrsim |\nabla A_1 \wedge \cdots \wedge \nabla A_l|$$

where we view $\nabla A_i = \sum D_j A_i$.

The main idea is the following. At any stage r of the algorithm, we have a decomposition $p \approx h(q_1, \ldots, q_{m_r})$, though not necessarily a diffuse decomposition. Thus, unless our argument is complete, we can find $x \in \mathbb{R}^{m_r}$ such that the diffuse property does not hold for $\mathbb{P}(|q - x| \leq \varepsilon)$. By strong anticoncentration, with large probability we have

$$\prod_{i=1}^{m_r} |q_i(X) - x_i| \gtrsim |\nabla q_1 \wedge \cdots \wedge \nabla q_{m_r}|.$$

Thus with significant probability the quantity $|\nabla q_1 \wedge \cdots \wedge \nabla q_{m_r}|$ is small.

Now let's recall some multi-linear algebra. If a family of vectors $v_1, \ldots, v_{m_r}$ is given, and $v_1 \wedge \cdots \wedge v_{m_r}$ is small, then this means these vectors are close to being linearly dependent. And since the $\{q_i\}$ are polynomials, this means, say, we can write q_1 as a function of the other q_j's, plus the products a_i and b_i introduced above, up to some small error. We then remove q_1 from the equation, and introduce the variables $\{a_i\}$ and $\{b_i\}$ into the family of q_i in the algorithm above at the next stage.

How do we ensure that keeping repeating this process will eventually give us the required decomposition? We associate with each stage of the algorithm a tuple of $d+1$ non-negative integers $(a_0, \ldots, a_d)$. These integers change on each stage of the algorithm, but it is important that the associated polynomials are *decreasing* at each step, if we give the set of all such polynomials a linear ordering by defining $(a_0, \ldots, a_d) \geq (b_0, \ldots, b_d)$ if $a_0 \geq b_0$, or $a_0 = b_0$ and $a_1 \geq b_1$, or $a_0 = b_0$, $a_1 = b_1$, and $a_2 \geq b_2$, and so on, i.e. the dictionary ordering. Because the set of all tuples $(a_0, \ldots, a_m)$ *has no infinite decreasing subsequence*, like for the non-negative integers, our algorithm must eventually terminate. But for $d > 0$, the number of steps before termination happens is unbounded, i.e. because at each stage of the algorithm we must decrease some a_i term, but we can increase the $a_{i+1}, \ldots, a_d$ terms by an arbitrary amount. But we can be slightly careful about quantifying how much this happens, which gives the bounds on the number of iterations involved, and thus the implicit constants in the algorithm, but they still grow quite large in the parameters involved.

Bibliography

[1] Daniel Kane, *A Structure Theorem For Poorly Anticoncentrated Gaussian Chaos and Applications To The Study of Polynomial Threshold Functions*, FOCS. (2012), 91-100.

Jacob Denson, UW Madison
email: jcdenson@wisc.edu

Chapter 4

On Rank Vs. Communication Complexity

after N. Nisan and A. Widgerson [2]
A summary written by Jaume de Dios Pont

Abstract. The paper studies the relationship between the communication complexity of a boolean function and the rank of the associated matrix. It gives an example exhibiting the largest gap known, and shows two related theorems.

4.1 Introduction and notation

The central object of study in this work is *Yao's two party communication complexity*. In this communication complexity model Alice and Bob want to evaluate $f(x, y)$, for a given, known, function $f : X \times Y \to \{0, 1\}$. Alice knows the value $x \in X$, and Bob $y \in Y$. The deterministic communication complexity $c(f)$ of f is the minimum number of bits they must share with each other to compute $f(x, y)$, for the hardest input (x, y):

$$c(F) := \min_{\substack{\text{Communication} \\ \text{algorhtims}}} \max_{(x,y)\in X\times Y} (\text{Bits shared})$$

Matrices, functions To each function $f : X \times Y \to \{0, 1\}$ we will assoicate a matrix of size $|X| \times |Y|$ with entries in $0, 1$, such that, under the identification $X \equiv \{1, \ldots, |X|\}$ (resp. Y), we have $M_{i,j} = f(i, j)$. We will use the matrix and function representation indistinguishably through the summary.

Other than the notation convenience of writing f in a matrix form, it is expected that there is a relationship between the complexity of f and the linear algebra of M_f. It is, for example, known that

$$\log \operatorname{rank} M \leq c(M) \leq \operatorname{rank} f \tag{4.1}$$

It is expected (and posed as a question by Lovász and Saks [1]) that

Conjecture 4.1 (Lovász-Saks,Nisan-Widgerson)**.** *For every binary matrix M, $c(M) = (\log \operatorname{rank} M)^{O(1)}$.*

The $O(1)$ term in Conjecture 4.1 was known to be necessary (or at least in the form $1 + o(1)$ since the work of Razborov [3], who gave a lower bound of the form $C(M) \geq \log \operatorname{rank} M \log \log \log \operatorname{rank} M$). We will see a stronger lower bound:

Theorem 4.2. *There exist (explicit) $(0, 1)$ matrices M_n of size $2^n \times 2^n$ such that $c(M) = \Omega(n)$, but $\log \operatorname{rank} M = O(n^\alpha)$, for $\alpha = \log_3 2 = 0.63 \ldots$.*

The rest of the work focuses on a weaker version of Conjecture 4.1. In order to describe it we need further notation:

Definition 4.3. *A subset of the entries of M will be called* monochromatic *if all of its entries are either* 0 *or* 1*. A submatrix of $A \subseteq M$ will be called monochromatic if all of its entries are either* 0 *or* 1 *(note that a choosing submatrix $A \subseteq M$ is equivalent to choosing a product subset $X' \times Y' \subseteq X \times Y$). We will denote by* $\mathrm{mono}(M)$ *the largest monochromatic subset of x.*

Definition 4.4. *As a relaxation to* mono *we define $\delta(A)$ the* advantadge *of A, as the difference $|\#\{(i,j) : A_{ij} = 1\} - \#\{(i,j) : A_{ij} = 0\}|$, and the* discrepancy *of M as the supremmum*

$$\mathrm{disc}(M) := \sup_{A \subseteq M} \frac{\delta(A)}{|M|}$$

where $|M|$ is the number of entries of M.

We have the following relationship between mono, disc and $\mathrm{c}(M)$:

Lemma 4.5. *Any matrix M is the union of at most $2^{c(M)}$ monochromatic matrices. In particular,*

$$\mathrm{disc}(M) \geq \mathrm{mono}(M) \geq 2^{-c(M)}$$

or, equivalently

$$-\log \mathrm{disc}(M) \leq -\log \mathrm{mono}(M) \leq c(M)$$

The $2^{c(M)}$ matrices arise by splitting over the the possible strings sent in the communication protocol: Every time Alice sends a new bit to Bob (similary Bob to Alice), we split the sub-matrices at the previous iteration depending on the value of the bit sent. The value of the bit sent can only depend on the information Alice (resp. Bob) has:

- All the communication exchanged (which already is encoded in the particular submatrix from the previous steps) and
- The value of $x \in X$ that Alice can see.

This guarantees that each matrix is split into sub-matrices, and not merely sub-sets
.

We have now the tools necessary to make two weaker conjectures:

Conjecture 4.6. *For every M, $-\log \mathrm{mono}(M) \leq (\log \mathrm{rank}\, M)^{O(1)}$*

Conjecture 4.7. *For every M, $-\log \mathrm{disc}(M) \leq (\log \mathrm{rank}\, M)^{O(1)}$*

Lemma 4.5 shows that we have the implications Conjecture 4.1 $\implies$ Conjecture 4.6 $\implies$ Conjecture 4.7. The paper has two parts: Proving Conjecture 4.7 and showing that in fact Conjectures 4.1 and 4.6 are equivalent. Conjecture 4.7, moreover, shows that a weaker model of communication, namely *distributional communication complexity*, where the inputs are chosen at random, and one wants to guess the value of $f(x,y)$ with nontrivial advantage over random guessing:

Corollary 4.8. *If $\mathrm{rank}(M) = r$ then there is a two-bit protocol P for which, for uniformly random inputs (x,y),*

$$\mathbb{P}[f(x,y) = P(x,y)] \geq \frac{1}{2} + \Omega(r^{-3/2})$$

4.2 Low-rank high-complexity matrices

The construction is based on functions $f : \{0,1\}^n \to \{0,1\}$ that are *fully sensitive at zero* (in the sense that $f(e_k) \neq f(0)$ for any unit vector $e_k = (0, \ldots, 0, 1, \ldots 0)$) and low *degree* (in the sense that on $\{0,1\}^n$ it is equal to a low-degree polynomial)

Lemma 4.9. *Let $n = 3^k$ There exists a boolean function $f_k : \{0,1\}^n$ that:*

- *is fully sensitive at zero*
- *Has degree* $2^k = n^{\log_3 2}$
- *Its polynomial representation has at most* $2^{O(n^{\log_3 2})}$ *monomials.*

The function f_k is constructed explicitly for $k = 1$ first, and then recursively defined as $f_{k+1}(\cdots) = f_1(f_k(\cdot), f_k(\cdot), f_k(\cdot))$.

Fix $f = f_k$ for some k. To the function $f : \{0,1\}^n \to \{0,1\}$ one associate a matrix M_f trough the communication game coming from

$$M_f(x_1 \dots x_n, y_1 \dots y_n) = f(x_1 y_1, \dots x_n y_n)$$

The rank of M_f is at most $2^{O(n^{\log_3 2})}$. This is because M_f is the sum of matrices M_i, each correspoding to a monomial of f, which have rank 1. On the other hand, the communication complexity of f is at least $\Omega(n)$:

Theorem 4.10. *[4]Let the UDISJ problem be the following: Two players are each given a subset of* $\{1 \dots n\}$*. If the sets are disjoint they must return* 1*, if the sets intersect at one point exactly, they must return* 0*. Otherwise they may return either* 0 *or* 1*.*

The communication cost of the UDISJ problem is $\Omega(n)$*.*

Since f is fully sensitive, a communication protocol for M_f solves the UDISJ problem by encoding the subsets by their characteristic functions. Therefore M_f must have $\Omega(n)$ complexity cost.

4.3 Conjectures 4.1 and 4.6 are equivalent

Conjecture 4.6 states that every matrix M of rank r and m entries has a *large* monochromatic sub matrix A with at least $\delta_r |M|$ entries, for $\delta := (\exp(\log^C r))^{-1}$. Using this large submatrix one can partition M (up to permutation)

$$M = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$$

Since A has rank 1, we can bound $\operatorname{rank} B + \operatorname{rank} C \leq \operatorname{rank} M$. Assuming (without loss of generality) that $\operatorname{rank} B \leq \operatorname{rank} C$ then $\operatorname{rank}(A|B) \leq 2 + \frac{1}{2}\operatorname{rank} M$. We use this to define a recursive algorithm, by spending one bit to determine whether the input belogs to $(A|B)$ or $(C|D)$, and inducting on the resulting submatrix. Let $L(m, r)$ be the number of leaves (terminal states) of this algorithm starting with a matrix with m entries and rank r. The protocol presented gives a recurrence

$$L(m, r) \leq L(m, \frac{1}{2}r + 2) + L((1 - \delta)m, r) + 1$$

from which, by induction (using the cases $m = 1$ and $r = 1$) one sees that $L(m, r) \leq \exp(\log^{k+1} r)$. One can bound the amount of necessary communication by the logarithm of the number of leaves, giving the theorem.

4.4 Conjecture 4.7 holds

In fact, we will show something slightly stronger, namely

Theorem 4.11. *For every matrix* M*,* $1/disc\ (M) = O(rank\ M^{3/2})$*. Therefore*

$$-\log disc(M) \leq \frac{3}{2}\log(rank\ M) + O(1)$$

Trough this section, we change the notation to ± 1*-valued matrices instead of* $\{0,1\}$*-valued matrices. This has the advantadge that now*

$$|M| disc\ M = \max_{u,v\ (0,1)-vectors} u^t M v$$

The proof of the theorem goes trough the following steps:

- By choosing a square sub-matrix of the same rank if necessary, assume that M is square of size $n \times n$.
- If M has low rank r (number of $\neq 0$ eigenvalues of M^tM), its $l^2 \to l^2$ operator norm (largest eigenvalue of M^tM) must be large compared to its Hilbert-Schmidt norm (equal to the sum of eigenvalues of M^tM). Since the Hilbert-Schmidt norm is the sum of the entries squared, there must be unit (in l^2) vectors x, y such that x^tMy is large ($\gtrsim nr^{-1/2}$).
- Pruning the entries of x, y that are $\geq \sqrt{8r/n}$ (making them zero) does not change the value of x^tMy significantly (at most by a factor of 2, by Cauchy-Schwartz). Rescaling these pruned vectors one can build u, v with l^∞ norm equal to 1 such that u^tMv is large, at least $\frac{1}{2} \cdot \frac{n}{r^{1/2}} \cdot \frac{n}{8r}$.
- The entries of u, v can be assumed to be ± 1, as otherwise one can enlarge them to enlarge the value of u^tMv.
- Let $u = u_+ - u_-$ be the decomposition into two vectors of the positive and negative parts, and similarly for v. Then u^tM^v is the sum of 4 terms of the form $u_\pm^t M v_\pm$. At least one of those terms must be of size $\frac{n^2}{16r^{3/2}}$

Bibliography

[1] Lovász, L. and Saks, M., *Lattices, mobius functions and communications complexity.* 29th Annual Symposium on Foundations of Computer Science (pp. 81-90). IEEE Computer Society. (1988)

[2] Nisan, N and Wigderson, A, *On rank vs. communication complexity* Combinatorica, 15, 557–565 (1995).

[3] Razborov, A. *The gap between the chromatic number of a graph and the rank of its adjacency matrix is superlinear.* Discrete mathematics 108 : 393-396. (1992)

[4] Kalyanasundaram, B, and Georg S. *The probabilistic communication complexity of set intersection.* SIAM Journal on Discrete Mathematics 5.4, 545-557. (1992)

Jaume de Dios Pont, UCLA
email: **jdedios@math.ucla.edu**

Chapter 5

A structure theorem for Boolean functions with small total influences

after Hamed Hatami [1]
A summary written by Jacek Jakimiuk

Abstract. We show that on every product probability space, Boolean functions with small total influences are essentially the ones that are almost measurable with respect to certain natural sub σ-algebras.

5.1 Introduction

We call a function Boolean if its range is $\{0,1\}$, or equivalently, if it is an indicator of some set. The influence of a variable on a Boolean function measures the sensitivity of the function with respect to the changes in that variable. There are many areas where influences of Boolean functions are studied, such as statistical physics, probability theory, computer science, combinatorics and economics. In particular, Boolean functions with small total influence arise frequently in many situations. Our purpose is to essentially characterize these functions. We define specific class of Boolean functions with small total influences, those are measurable with respect to certain σ-algebras, and show that every Boolean function with small total influence can be effectively approximated by the functions of this class.

Let us give some notations and definitions. Denote $[n] = \{1, \ldots, n\}$. In this summary $X = (\Omega, \mathcal{F}, \mu)$ is always a probability space and X^n means product space with the product measure μ^n. For every $x = (x_1, \ldots, x_n) \in X^n$ and $S \subset [n]$ denote as $x_S = (x_i : i \in S) \in X^S$ the restriction of x to the coordinates of S. If $S, T \subset X$, $S \cap T = \emptyset$, then (x, y) denotes the unique element $z \in X^{S \cup T}$ with $z_S = x$ and $z_T = y$. We treat functions $g : X^S \to \mathbb{R}$ also as functions defined on X^n with $g(x) = g(x_S)$.

For measurable function $f : X^n \to \{0,1\}$ we define influence of the j-th variable on f as

$$I_f(j) = \mathbb{P}(f(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n) \neq f(x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n)),$$

where x_i, y_i are iid random variables taking values in X according to its probability measure. We also define total influence as

$$I_f = \sum_{j=1}^{n} I_f(j).$$

By measurability of f it is clear that $I_f(j)$ is well defined.

We are particularly interested in the case $\Omega = \{0,1\}$ with Bernoulli distribution μ_p defined by $\mu_p(\{1\}) = p$ and $\mu_p(\{0\}) = 1 - p$ for $0 < p < 1$. The p-biased distribution means product measure μ_p^n. We call function $f : \{0,1\}^n \to \{0,1\}$ increasing if $\forall_i x_i \leq y_i$ implies $f(x) \leq f(y)$. For $A \subset [n]$ we say that $f : X^n \to \mathbb{R}$ depends only on coordinates in A if $x_A = y_A$ implies $f(x) = f(y)$.

Remark 5.1. *In many situations we can assume that Ω is a finite set. Indeed, for any measurable function $f : X^n \to \{0,1\}$ and arbitrary small $\varepsilon > 0$ we can find a finite σ-algebra $\mathcal{G} \subset \mathcal{F}$ and $g : \Omega^n \to \{0,1\}$, measurable with respect to the product σ-algebra generated by $\mathcal{G}$, such that $\mathbb{P}(f(x) \neq g(x)) < \varepsilon$. Then obviously $|I_f - I_g| < 2n\varepsilon$.*

5.2 Main results

Consider the case of the p-biased distribution. The first, intuitive attempt to characterize functions with small total influences is to relate them with functions depending only on a few number of coordinates (see [2], [3], [4] and [5]). Following example shows, that this attempt is not sufficient if p is small.

Example 5.2. *Let $p = n^{-1}$ and $f = \mathbf{1}_{\Omega^n \setminus \{0\}}$. Then $I_f(1) = I_f(2) = \ldots = I_f(n) \leq 2p$, so $I_f \leq 2$. See that f has no variable with large influence and f does not depend only on a small set of coordinates. Indeed, take constant size set $A \subset [n]$, we have $\mathbb{E}(f(x)|x_A = 0) = 1 - (1-p)^{n-|A|} = 1 - e^{-1} \pm o(1)$ and $\mathbb{P}(x_A = 0) \geq 1 - p|A| = 1 - o(1)$, hence $\|f - g\|_1 \geq e^{-1} - o(1)$ for every g depending only on coordinates in A (x is treaten as random variable and all asymptotics are meant for $n \to \infty$).*

Let us return to general case. We begin our attempt with defining a certain class of Boolean functions. Let $\mathcal{J} = \{J_S\}_{S \subset [n]}$ be any set of measurable functions $J_S : X^S \to \{0,1\}$. Define the map $J_{\mathcal{J}} : X^n \to \mathcal{P}([n])$ as $J_{\mathcal{J}}(x) = \bigcup_{S \subset [n], J_S(x)=1} S$. Let $\mathcal{F}_{\mathcal{J}}$ be the sub σ-algebra induced by the map $x \mapsto (J_{\mathcal{J}}(x), x_{J_{\mathcal{J}}(x)})$.

Definition 5.3. *Let $k > 0$. A k-pseudo-junta is a function $f : X^n \to \{0,1\}$ that is measurable with respect to $\mathcal{F}_{\mathcal{J}}$ for some $\mathcal{J}$ satisfying*

$$\int |J_{\mathcal{J}}(x)| dx \leq k.$$

Example 5.4. *Let $A \subset [n]$, $|A| \leq k$. Then every measurable Boolean function that depends only on coordinates in A is k-pseudo-junta. Indeed, take $J_A \equiv 1$ and $J_S \equiv 0$ for $S \neq A$. Then $J_{\mathcal{J}} \equiv A$, hence f is measurable with respect to $\mathcal{F}_{\mathcal{J}}$. Obviously $\int |J_{\mathcal{J}}(x)| dx = |A| \leq k$.*

Above example shows that our attempt is a generalization of previous. Now we give a simple proof that pseudo-juntas has small total influences.

Proposition 5.5. *Let f be a k-pseudo-junta. Then $I_f \leq 2k$.*

Proof. Using notation from the definition of total influence, let $x = (x_1, \ldots, x_n)$ and $x^{(j)} = (x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n)$. It follows from the definition of k-pseudo-junta that if $f(x) \neq f(x^{(j)})$, then $j \in J_{\mathcal{J}}(x) \cup J_{\mathcal{J}}(x^{(j)})$. Hence

$$I_f = \sum_{j=1}^{n} \mathbb{P}\left(f(x) \neq f(x^{(j)}\right) \leq \sum_{j=1}^{n} \mathbb{P}\left(j \in J_{\mathcal{J}}(x) \cup J_{\mathcal{J}}(x^{(j)})\right) \leq 2 \sum_{j=1}^{n} \mathbb{P}\left(j \in J_{\mathcal{J}}(x)\right) \leq$$

$$\leq 2 \int |J_{\mathcal{J}(x)}| dx \leq 2k.$$

□

Our main result is essentialy the inverse theorem of above proposition.

Theorem 5.6. *Consider a measurable function $f : X^n \to \{0,1\}$. For every $\varepsilon > 0$ there exists a $e^{10^{15}\varepsilon^{-3}\lceil I_f \rceil^3}$-pseudo-junta $h : X^n \to \{0,1\}$ such that $\|f - h\| \leq \varepsilon$.*

Example 5.7. *Consider f as in example 2.1. Define $J_S = \mathbf{1}_{x_S = (1,\ldots,1)}$. Then $J_{\mathcal{J}}(x) = \{i : x_i = 1\}$, hence $\int |J_{\mathcal{J}}(x)| dx = pn = 1$. Furthermore, $\mathcal{F}_{\mathcal{J}}$ is an original discrete σ-algebra, so f is measurable with respect to it. Hence in Theorem 5.6 we can take $h = f$.*

This result can be improved in the case of the p-biased distribution.

Theorem 5.8. *Suppose that in theorem 5.6 we have $X = (\{0,1\}, \mu_p)$. Then in the statement of Theorem 5.6 function h is a $e^{10^{10}\varepsilon^{-2}\lceil I_f \rceil^2}$-pseudo-junta.*

In fact proof in the p-biased case is much simpler than general case, but is based on the similar idea. Therefore we firstly prove Theorem 5.8 in section 5.4, and then, in section 5.5, we prove Theorem 5.6.

5.3 Generalized Walsh expansion

Before proving main theorems, we briefly review some basic facts about generalized Walsh expansion. Let $L_2(X^n) = \{f : X^n \to \mathbb{C} : \int |f(x)|^2 dx < \infty\}$.

Definition 5.9. *The generalized Walsh expansion of a function $f \in L_2(X^n)$ is the unique expansion $f = \sum_{S\subset[n]} F_S$ that satisfies the following two properties:*

1. *For every $S \subset [n]$ the function f depends only on coordinates in S;*
2. *$\int F_S dx_i \equiv 0$ for every $i \in S \subset [n]$.*

It follows from the above definition that for every $T \subset [n]$ we have $\int f dx_{[n]\setminus T} = \sum_{S\subset T} F_S$, hence $F_S(y) = \sum_{T\subset S}(-1)^{|S\subset T|} \int f(y_T, x_{[n]\setminus T}) dx_{[n]\setminus T}$. It shows uniqueness of generalized Walsh expansion.

It is easy to see, that functions F_S are pairwise orthogonal, hence we have some nice properties such as Parseval's identity. We can use them to prove some connections between generalized Walsh expansion and influence. The following identity holds:

$$I_f(i) = 2\sum_{S\ni i} \|F_S\|_2^2.$$

It easily gives us

$$I_f = 2\sum_{S\subset[n]} |S|\|F_S\|_2^2.$$

5.4 Proof of Theorem 5.8

In this summary we give only sketch of the proof. Let $f = \sum_{S\subset[n]} F_S$ be a generalized Walsh expansion of f. To make the proof more clear we divide it into steps.

1. Firstly we want to simplify the expansion by removing some insignificant terms from it in such way that remaining terms have nice properties.
2. Let $\mathcal{S}$ be a family of remaining terms from previous point and let $g = \sum_{S\in\mathcal{S}} F_S$. Now we want to find set of functions $\mathcal{J}$ (such as in definition of pseudo-junta) such that $\|g - \mathbb{E}(g|\mathcal{F}_\mathcal{J})\|_2$ is small. This is the most difficult and technical step. Here we use the fact, that in the p-biased case we have $F_S(x) = \hat{f}(S)\prod_{i=1}^n r(x_i)$, where $r(0) = -\sqrt{\frac{p}{1-p}}$, $r(1) = \sqrt{\frac{1-p}{p}}$ and $\hat{f}(S)$ are real constants bounded dependently only on p and S.
3. Now we can easily finish the proof. We have

   $$\|f - \mathbb{E}(f|\mathcal{F}_\mathcal{J})\|_2^2 \le \|f - \mathbb{E}(g|\mathcal{F}_\mathcal{J})\|_2^2 \le 2\|f-g\|_2^2 + 2\|g - \mathbb{E}(g|\mathcal{F}_\mathcal{J})\|_2^2,$$

   where the right hand of inequality is small by previous steps. It is easy to check that function h defined by

   $$h(x) = 1 \text{ if } \mathbb{E}(f|\mathcal{F}_\mathcal{J})(x) > \frac{1}{2}, \quad h(x) = 0 \text{ if } \mathbb{E}(f|\mathcal{F}_\mathcal{J})(x) \le \frac{1}{2}$$

   is desired pseudo-junta.

5.5 Proof of Theorem 5.6

Again, we give only a sketch. Steps 1 and 3 from the previous proof proceed exactly in the same way, but since in the general case functions F_S are not as well behaved as in the p-biased case, we need to modify step 2. So assume that we have functions F_S and family $\mathcal{S}$ as in the proof of Theorem 5.8. We define functions G_S in such way that $\|G_S - F_S\|_2$ is small and G_S satisfy some bounds similar to those satisfied by F_S in the p-biased case. Now we can proceed as in the previous case by defining suitable σ-algebra and proving necessary error bounds. Again, this is the most technically difficult part of the proof.

Bibliography

[1] H. Hatami, A structure theorem for Boolean functions with small total influences, Annals of Mathematics 176 (2012), 509-533

[2] M. Talagrand, On Russo's approximate zero-one law, Ann. Probab. 22 (1994), 1576–1587

[3] E. Friedgut and G. Kalai, Every monotone graph property has a sharp threshold, Proc. Amer. Math. Soc. 124 (1996), 2993–3002

[4] J. Bourgain and G. Kalai, Influences of variables and threshold intervals under group symmetries, Geom. Funct. Anal. 7 (1997), 438–461

[5] E. Friedgut, Boolean functions with low average sensitivity depend on few coordinates, Combinatorica 18 (1998), 27–35

Jacek Jakimiuk, University of Warsaw
email: **j.jakimiuk4@student.uw.edu.pl**

Chapter 6

Learning Low-Degree Functions From a Logarithmic Number of Random Queries

after A. Eskenazis and P. Ivanisvili [4]
A summary written by Dylan Langharst

Abstract. We show that a bounded function f on the Hamming Cube with magnitude at most 1 and degree at most d can be determined with $g(n,\varepsilon,\delta)C^{d^{3/2}\sqrt{\log d}}$ random queries, where $C>1$ is a finite, universal constant, n is the dimension, ε is the L_2-accuracy, $1-\delta$ is the confidence and g is an explicit function.

6.1 Introduction

We first recall the classical fact from analysis: consider an integrable function $f:[-\pi,\pi]\to\mathbb{R}$. The *Fourier expansion* of f is then given by

$$f(x)=\sum_{j\in\mathbb{Z}}\hat{f}(j)e^{jix},\quad \hat{f}(j)=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)e^{-jix}dx.$$

The concept of Fourier analysis has been extended to real valued functions on the n-dimensional Hamming Cube. That is, let $f:\{-1,1\}^n\to\mathbb{R}$. Then, the *Fourier-Walsh expansion* of f is given by, with $w_S(x)=\prod_{i\in S}x_i$,

$$f(x)=\sum_{S\subset\{1,\dots,n\}}\hat{f}(S)w_S(x),\quad \hat{f}(S)=\frac{1}{2^n}\sum_{y\in\{-1,1\}^n}f(y)w_S(y). \tag{6.1}$$

For simplicity, $\mathcal{C}$ will denote the set of real-valued functions on the Hamming cube. For $f\in\mathcal{C}$, f has degree at most $d\in\{1,\dots,n\}$ if $\hat{f}(S)=0$ whenever $|S|>d$. A classical problem in the field is the learning problem: given a source of *examples* $(x,f(x))$, $x\in\{-1,1\}^n, f\in\mathcal{C}$, we will construct a *hypothesis* function $h\in\mathcal{C}$ which approximates f up to some error. Here, we will use the *random query model*, where we have N independent examples chosen uniformly at random from the Hamming Cube $\{-1,1\}^n$, and wish to construct our (random) function h such that $\|h-f\|_{L_2}^2<\varepsilon$ with probability at least $1-\delta$, where $\varepsilon,\delta\in(0,1)$ are accuracy and confidence parameters.

This problem has been studied for decades, with various restrictions on the range of f. Let $\mathcal{C}_b^d$ denote the set of bounded functions $f:\{-1,1\}^n\to[-1,1]$ of degree at most d. Then, the Low-Degree Algorithm [6] shows that for $f\in\mathcal{C}_b^d$, there exists an algorithm producing an ε-approximation of f with probability at least $1-\delta$ using $N=\frac{2n^d}{\varepsilon}\log\left(\frac{2n^d}{\delta}\right)$ samples; such a result is said to be $\mathcal{O}_{\varepsilon,\delta,d}(n^d\log n)$. This estimate was improved in [5] by deriving new bounds on the ℓ_1-size of the Fourier spectrum of bounded functions

to then show that $N = \mathcal{O}_{\varepsilon,\delta,d}(n^{d-1}\log n)$ examples suffice to learn $\mathcal{C}_b^d$. Our first main result shows that $N = \mathcal{O}_{\varepsilon,\delta,d}(\log n)$ suffices.

Theorem 6.1. *Fix $\varepsilon, \delta \in (0,1), n \in \mathbb{N}, d \in \{1,\dots,n\}$ and a function $f \in \mathcal{C}_b^d$. If $N \in \mathbb{N}$ satisfies*

$$N \geq \min\left\{\frac{\exp(Cd^{3/2}\sqrt{\log d})}{\varepsilon^{d+1}}, \frac{4dn^d}{\varepsilon}\right\}\log\left(\frac{n}{\delta}\right)$$

for a large numerical constant $C \in (0,\infty)$, then N uniformly random independent queries of examples $(x, f(x))$, $x \in \{-1,1\}^n$, suffices for the construction of a random function $h \in \mathcal{C}$ satisfying the condition $\|h - f\|_{L_2}^2 < \varepsilon$ with probability at least $1-\delta$.

We next have the following result: for every $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and $d \in \mathbb{N}$, there exists $B_d^{\mathbb{K}} \in (0,\infty)$ such that for every $n \in \mathbb{N}$ and every coefficients $c_\alpha \in \mathbb{K}$, $\alpha \in (\mathbb{N}\cup\{0\})^n$ such that $|\alpha| \leq d$, one has

$$\left(\sum_{|\alpha|\leq d} |c_\alpha|^{\frac{2d}{d+1}}\right)^{\frac{d+1}{2d}} \leq B_d^{\mathbb{K}} \max\left\{\left|\sum_{|\alpha|\leq d} c_\alpha x^\alpha\right| : x \in \mathbb{K}^n, \|x\|_{\ell_\infty^n(\mathbb{K})} \leq 1\right\}, \tag{6.2}$$

and $\frac{2d}{d+1}$ is the smallest exponent for which the optimal constant in (6.2) is independent of the number of variables n of the polynomial. This result called be derived from an extension of Littlewood's $\frac{4}{3}$-inequality [7] shown by Bohnenblust and Hille [2]. For the constants, $B_d^{\mathbb{K}}$, it is known that $\limsup_{d\to\infty}\left(B_d^{\mathbb{R}}\right)^{1/d} = 1+\sqrt{2}$ and $B_d^{\mathbb{C}} \leq C^{\sqrt{d\ln d}}$ for a finite constant $C > 1$. We will use this result in the following way: restricting (6.2) to real multilinear polynomials, convexity shows that the maximum is obtained at some point in $x \in \{-1,1\}^n$. Combining this with equation (6.1), we deduce there exists an optimal constant, denoted $B_d^{\{\pm 1\}}$ and first explored in [1, p. 175], such that

$$\left(\sum_{S\subseteq\{1,\dots,n\}} |\hat{f}(S)|^{\frac{2d}{d+1}}\right)^{\frac{d+1}{2d}} \leq B_d^{\{\pm1\}}\|f\|_{L_\infty}. \tag{6.3}$$

It is known [3] that there exists a universal constant $\kappa \in (0,\infty)$ such that $B_d^{\{\pm1\}} \leq \exp(\kappa\sqrt{d\log d})$. In this work, we show the following.

Theorem 6.2. *Fix $\varepsilon, \delta \in (0,1), n \in \mathbb{N}, d \in \{1,\dots,n\}$ and a function $f :\in \mathcal{C}_b^d$. If $N \in \mathbb{N}$ satisfies*

$$N \geq \frac{e^8 d^2}{\varepsilon^{d+1}}\left(B_d^{\{\pm1\}}\right)^{2d}\log\left(\frac{n}{\delta}\right),$$

then given N uniformly random independent queries of pairs $(x, f(x))$, where $x \in \{-1,1\}^n$, one can construct a random function $h \in \mathcal{C}$ satisfying $\|h-f\|_{L_2}^2 < \varepsilon$ with probability at least $1-\delta$.

6.2 Proofs

Proof of Theorem 2. Fix a parameter $b \in (0,\infty)$ and denote by

$$N_b \overset{\text{def}}{=} \left\lceil \frac{2}{b^2}\log\left(\frac{2}{\delta}\sum_{k=0}^{d}\binom{n}{k}\right)\right\rceil. \tag{6.4}$$

Let $X_1,\dots,X_{N_b}$ be independent random vectors, each uniformly distributed on $\{-1,1\}^n$. For a subset $S \subseteq \{1,\dots,n\}$ with $|S| \leq d$ consider the empirical Walsh coefficient of f, given by

$$\alpha_S = \frac{1}{N_b}\sum_{j=1}^{N_b} f(X_j)\, w_S(X_j). \tag{6.5}$$

Since α_S is a sum of bounded i.i.d. random variables and $\mathbb{E}[\alpha_S] = \hat{f}(S)$, the Chernoff bound gives

$$\forall S \subseteq \{1,\dots,n\}, \quad \mathbb{P}\left\{\left|\alpha_S - \hat{f}(S)\right| > b\right\} \leq 2\exp\left(-N_b b^2/2\right). \tag{6.6}$$

Let G_b be a symbol denoting the event

$$|\alpha_S - \hat{f}(S)| \leq b, \text{ for every } S \subseteq \{1,\dots,n\} \text{ with } |S| \leq d.$$

Then, by using the union bound and taking into account that f has degree at most d, we get

$$\mathbb{P}\{G_b\} \geq 1 - 2\sum_{k=0}^{d}\binom{n}{k}\exp\left(-N_b b^2/2\right) \geq 1-\delta, \tag{6.7}$$

where the last inequality follows from the definition of N_b, (6.4). Next, fix an additional parameter $a \in (b,\infty)$ and consider the random collection of sets given by

$$\mathcal{S}_a \stackrel{\text{def}}{=} \{S \subseteq \{1,\dots,n\} : |\alpha_S| \geq a\} \tag{6.8}$$

Observe that if the event G_b holds, then

$$|\hat{f}(S)| \begin{cases} \leq \left|\alpha_S - \hat{f}(S)\right| + |\alpha_S| < a+b & \forall S \notin \mathcal{S}_a \\ \geq |\alpha_S| - \left|\alpha_S - \hat{f}(S)\right| \geq a-b & \forall S \in \mathcal{S}_a. \end{cases} \tag{6.9}$$

Finally, consider the random function $h_{a,b} \in \mathcal{C}$ given by

$$\forall x \in \{-1,1\}^n, \quad h_{a,b}(x) \stackrel{\text{def}}{=} \sum_{S\in\mathcal{S}_a} \alpha_S w_S(x).$$

We deduce that

$$\begin{aligned} |S_a| &\leq (a-b)^{-\frac{2d}{d+1}} \sum_{S\in\mathcal{S}_a} |\hat{f}(S)|^{\frac{2d}{d+1}} \leq (a-b)^{-\frac{2d}{d+1}} \sum_{S\subseteq\{1,\dots,n\}} |\hat{f}(S)|^{\frac{2d}{d+1}} \\ &\leq (a-b)^{-\frac{2d}{d+1}} \left(B_d^{\{\pm 1\}}\right)^{\frac{2d}{d+1}}, \end{aligned} \tag{6.10}$$

where the first inequality follows from the second case of (6.9) and the third inequality follows from (6.3). Therefore, on the event G_b, we obtain

$$\begin{aligned} \|h_{a,b} - f\|_{L_2}^2 &= \sum_{S\subseteq\{1,\dots,n\}} \left|\hat{h}_{a,b}(S) - \hat{f}(S)\right|^2 = \sum_{S\in\mathcal{S}_a} \left|\alpha_S - \hat{f}(S)\right|^2 + \sum_{S\notin\mathcal{S}_a} |\hat{f}(S)|^2 \\ &< |\mathcal{S}_a|\, b^2 + (a+b)^{\frac{2}{d+1}} \sum_{S\notin\mathcal{S}_a} |\hat{f}(S)|^{\frac{2d}{d+1}} \\ &\leq \left(B_d^{\{\pm 1\}}\right)^{\frac{2d}{d+1}} \left((a-b)^{-\frac{2d}{d+1}} b^2 + (a+b)^{\frac{2}{d+1}}\right), \end{aligned} \tag{6.11}$$

where the first inequality follows from the first case of (6.9) and the second inequality from (6.2) and (6.10). Choosing $a = b(1+\sqrt{d+1})$, we deduce that

$$\left\|h_{b(1+\sqrt{d+1}),b} - f\right\|_{L_2}^2 < \left(B_d^{\{\pm 1\}}\right)^{\frac{2d}{d+1}} b^{\frac{2}{d+1}} \left((d+1)^{-\frac{d}{d+1}} + (2+\sqrt{d+1})^{\frac{2}{d+1}}\right). \tag{6.12}$$

By choosing $b^2 \leq e^{-5} d^{-1} \varepsilon^{d+1} \left(B_d^{\{\pm 1\}}\right)^{-2d}$, (6.12) yields that $\left\|h_{b(1+\sqrt{d+1}),b} - f\right\|_{L_2}^2 < \varepsilon$ due to the inequality

$$(d+1)^{-\frac{d}{d+1}} + (2+\sqrt{d+1})^{\frac{2}{d+1}} \leq \left(e^4(d+1)\right)^{\frac{1}{d+1}} \quad \text{for all} \quad d \geq 1,$$

which can be readily verified from convexity or rudimentary analysis. Inserting this restriction on b into (6.4), we obtain that, given

$$N = \left\lceil \frac{e^6 d \left(B_d^{\{\pm 1\}}\right)^{2d}}{\varepsilon^{d+1}} \log\left(\frac{2}{\delta}\sum_{k=0}^{d}\binom{n}{k}\right)\right\rceil \tag{6.13}$$

random queries, the random function $h_{b(1+\sqrt{d+1}),b}$ satisfies $\left\|h_{b(1+\sqrt{d+1}),b} - f\right\|_{L_2}^2 < \varepsilon$ with probability at least $1-\delta$. Elementary estimates such as

$$\sum_{k=0}^{d}\binom{n}{k} \le \sum_{k=0}^{d}\frac{n^k}{k!} = \sum_{k=0}^{d}\frac{d^k}{k!}\left(\frac{n}{d}\right)^k \le \left(\frac{en}{d}\right)^d$$

then yield the conclusion of the theorem.

□

Proof of Theorem 1. For $\varepsilon < \frac{\exp(C\sqrt{d\log d})}{n}$, the result follows from the Low-Degree algorithm of [6]. For $\varepsilon \ge \frac{\exp(C\sqrt{d\log d})}{n}$, the result follows from Theorem 2 combined with the bound $B_d^{\{\pm 1\}} \le \exp(\kappa\sqrt{d\log d})$ from [3]. □

6.3 Concluding Remarks

Recall that the Chebyshev polynomials of the first kind are a sequence of orthogonal polynomials $\{T_d\}$ such that $T_d(\cos\theta) = \cos(d\theta)$. Given a function $f \in \mathcal{C}$, its *Rademacher* projection on the level $\ell \in \{1,\ldots,n\}$ is defined as

$$\mathrm{Rad}_\ell f(x) = \sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=\ell}} \hat{f}(S) w_S(x) \tag{6.14}$$

The following was shown, and, furthermore, is asymptotically sharp [5]: let $f \in \mathcal{C}$ be a function of degree d. Then, for every $\ell \in \{1,\ldots,d\}$,

$$\left\|\mathrm{Rad}_\ell f\right\|_{L_\infty} \le \begin{cases} \frac{\left|T_d^{(\ell)}(0)\right|}{\ell!}\cdot\|f\|_{L_\infty}, & \text{if } (d-\ell) \text{ is even} \\ \frac{\left|T_{d-1}^{(\ell)}(0)\right|}{\ell!}\cdot\|f\|_{L_\infty}, & \text{if } (d-\ell) \text{ is odd} \end{cases}. \tag{6.15}$$

In particular, (6.15) implies that if f has degree at most d then

$$\forall \ell \in \{1,\ldots,d\}, \quad \left\|\mathrm{Rad}_\ell f\right\|_{L_\infty} \le \frac{d^\ell}{\ell!}\cdot\|f\|_{L_\infty}. \tag{6.16}$$

In the same work [5], the following was shown: if $f \in \mathcal{C}_b^d$, then for every $\ell \in \{1,\ldots,d\}$ one has

$$\sum_{S\subseteq\{1,\ldots,n\}} \left|\widehat{\mathrm{Rad}_\ell f}(S)\right| = \sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=\ell}} |\hat{f}(S)| \le n^{\frac{\ell-1}{2}} d^\ell e^{\binom{\ell+1}{2}}. \tag{6.17}$$

We can use a Bohnenblust-Hille type inequality from [3] to improve (6.17).

Corollary 6.3. *Let $n \in \mathbb{N}$ and $d \in \{1,\ldots,n\}$. Then, every $f \in \mathcal{C}_b^d$ satisfies*

$$\sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=\ell}} |\hat{f}(S)| \le \binom{n}{\ell}^{\frac{\ell-1}{2\ell}} e^{\kappa\sqrt{\ell\log\ell}}\frac{d^\ell}{\ell!} \le n^{\frac{\ell-1}{2}} d^\ell \ell^{-c\ell} \tag{6.18}$$

for some universal constant $c \in (0,1)$.

Proof. We see that

$$\begin{aligned}\sum_{\substack{S\subseteq\{1,\dots,n\}\\|S|=\ell}}|\hat{f}(S)| &\leq \binom{n}{\ell}^{\frac{\ell-1}{2\ell}}\left(\sum_{S\subseteq\{1,\dots,n\}}\left|\widehat{\mathrm{Rad}_\ell f}(S)\right|^{\frac{2\ell}{\ell+1}}\right)^{\frac{\ell+1}{2\ell}}\\ &\leq \binom{n}{\ell}^{\frac{\ell-1}{2\ell}}\exp(\kappa\sqrt{\ell\log\ell})\,\|\mathrm{Rad}_\ell\, f\|_{L_\infty}\\ &\leq \binom{n}{\ell}^{\frac{\ell-1}{2\ell}}\exp(\kappa\sqrt{\ell\log\ell})\frac{d^\ell}{\ell!},\end{aligned}$$

where the first inequality follows from Hölder's inequality, the second from [3], and the third from (6.16). Thus, we have established the first inequality of (6.18). The second inequality follows from (6.15) and the elementary bound $\binom{n}{\ell}\leq\left(\frac{ne}{\ell}\right)^\ell$. □

Theorem 1 shows that bounded functions $f:\{-1,1\}^n\to[-1,1]$ of degree at most d can be learned with accuracy at most ε and confidence at least $1-\delta$ from $N=\mathcal{O}_{\varepsilon,d}(\log(n/\delta))$ random queries. We conclude by stating without proof that this estimate is sharp for small enough values of δ.

Proposition 6.4. *Suppose that bounded linear functions $f:\{-1,1\}^n\to[-1,1]$ can be learned with accuracy at most $\frac{1}{2}$ and confidence at least $1-\frac{1}{2n}$ from N random queries. Then $N>\log_2 n$.*

Bibliography

[1] Blei R., *Analysis in integer and fractional dimensions*, vol. 71 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2001.

[2] Bohnenblust, H.F. and Hille, E., *On the absolute convergence of Dirichlet series*, Ann. of Math. (2), 32 (1931), pp.600-622.

[3] Defant, A., Mastyło, M., and Pérez, A., *On the Fourier spectrum of functions on Boolean cube*, Math. Ann. 374, (2019), pp. 653–680.

[4] Eskenazis, A. and Ivanisvili, P., *Learning Low-Degree Functions from a Logarithmic Number of Random Queries*, Arxiv, (2022).

[5] Iyer, S., Rao, A., Reis, V., Rothvoss, T. and Yehudayoff A., *Tight bounds on the Fourier growth of bounded functions on the hypercube*, to appear in ECCC 2021.

[6] Linial, N., Mansour, Y., and Nisan, N., *Constant depth circuits, Fourier transform, and learnability*, J. Assoc. Comput. Mach., 40 (1993), pp. 607-620.

[7] Littlewood, J.E., *On bounded bilinear forms in an infinite number of variables*, Q.J. Math., os-1 (1930), pp.164-174.

Dylan Langharst, KSU
email: **dlanghar@kent.edu**

Chapter 7

The Carbery-Wright inequalities for polynomial norms and distributions

after T. Carbery and J. Wright [1]
A summary written by Caleb Marshall

Abstract. We introduce the Carbery-Wright inequalities for polynomials $p : \mathbb{R}^n \to X$ taking values in a Banach space. This involves the discussion of localization results for convex bodies, an examination of classical polynomial norm inequalities, and representation theorems for plurisubharmonic functions and their boundary values.

7.1 Introduction

Let $(X, ||\cdot||)$ be a (real or complex) Banach space and let $\mathcal{P}_{d,n}$ be the space of polynomials $p : \mathbb{R}^n \to X$ taking values in X and of degree at most d. These polynomials induce functionals $p^{\#}(x) := ||p(x)||^{1/d}$ on X, and we let $\mathcal{P}^{\#}_{d,n}$ denote the space of all such functionals. For a convex body $K \subset \mathbb{R}^n$ of unit measure and each $0 < q < \infty$, we define a standard L^q "norm" on $\mathcal{P}^{\#}_{d,n}$ as $||p^{\#}||_q := \big(\int_K ||p(x)||^{\frac{q}{d}} dx\big)^{1/q}$. Also let $||p^{\#}||_0 := \exp \int_K \log p^{\#}(x) dx$ and take $||p^{\#}||_\infty$ to be the usual supremum norm of $p^{\#}$ over K.

As $\mathcal{P}_{d,n}$ is finite-dimensional, the norms $||\cdot||_q$ are equivalent. When $r \leq q$, Hölder's inequality gives the inequality $||p^{\#}||_r \leq ||p^{\#}||_q$ with optimal constant 1. This summary discusses the optimal constant for the reverse inequality. The following is due to Carbery and Jim Wright in [1].

Theorem 7.1 (Carbery-Wright Inequality)**.** *Let $p : \mathbb{R}^n \to X$ be a polynomial of degree at most d, let K be a convex body in $\mathbb{R}^n$ of volume 1 and let $0 \leq r \leq q \leq \infty$. Then there exists an absolute constant C, independent of p, d, K, n, q, r and X such that*

$$||p^{\#}||_q \leq C \frac{\big[nB(n, q+1)\big]^{\frac{1}{q}}}{\big[nB(n, r+1)\big]^{\frac{1}{r}}} ||p^{\#}||_r, \tag{7.1}$$

where $B(z_1, z_2) := \int_0^1 t^{z_1 - 1}(1-t)^{z_2 - 1} dt$ denotes the classical beta function.

Choosing $r \leq 1$ leads to a strong distributional inequality, which has many applications in convex geometry, probability and analysis.

Theorem 7.2 (Distributional Carbery-Wright)**.** *Let $p : \mathbb{R}^n \to X$ be a polynomial of degree at most d, let K be a convex body in $\mathbb{R}^n$ of volume 1 and let $0 \leq q \leq \infty$. Then there exists an absolute constant C, independent of p, d, K, n, q, r and X so that for any $\alpha > 0$,*

$$||p^{\#}||_q \; \alpha^{-1} \big|\{x \in K : p^{\#}(x) \leq \alpha\}\big| \leq Cn\big[nB(n, q+1)\big]^{\frac{1}{q}} \tag{7.2}$$

Up to numerical estimation of the constant C, the Carbery-Wright inequalities are optimal over arbitrary convex bodies. Moreover, Stirling's approximation leads to a closed-form expression of Theorems 1 and 2 upper bounds in terms of n, q and r.

The exceedingly general results of [1] are proven in three steps.

1. An extremal result over convex bodies attributed to Kannan, Lovász, and Simonovits, which reduces L^q estimates over K to weighted inequalities over line segments in $\mathbb{R}^n$.
2. Application of classical L^q norm estimates for polynomials to prove certain weighted $L^q \to L^r$ norm inequalities for polynomials $p : \mathbb{R} \to \mathbb{C}$.
3. Representation theorems and inequalities for the boundary values $\tilde{u}\big|_{\mathbb{R}^n}$ of plurisubharmonic functions $\tilde{u} : \mathbb{C}^n \to \mathbb{R}$. In particular, the scalar valued arguments for polynomials $p : \mathbb{R} \to \mathbb{C}$ are adaptable to representations of such functions.

The following sections give an overview of these three reductions, which we frame as the central innovations of [1].

7.2 An extremal result for convex bodies

The first reduction is an examination of localization result for convex bodies, attributed to R. Kannan, L. Lovás and M. Simonovits in [5]. We introduce the original formulation below.

A *needle* $N := (I, \ell)$ is a line segment $[a, b] \subset \mathbb{R}^n$ together with a nonnegative linear function $\ell : [a, b] \to \mathbb{R}_+$, not identically zero. If f is an integrable function on I, then we define

$$\int_N f := \int_0^{|b-a|} f(a + tu)[\ell(a + tu)]^{n-1} dt,$$

where $u := \frac{b-a}{|b-a|}$. Kannan,Lovász, and Simonovits obtain the following duality of needles and convex bodies in $\mathbb{R}^n$.

Proposition 7.3 (Kannan-Lovász-Simonovits Localization Lemma)**.** *Let f_1, f_2, f_3, f_4 be four nonnegative continous functions on $\mathbb{R}^n$ and let $\alpha, \beta > 0$. Then the following are equivalent:*

1. *For every convex body K in $\mathbb{R}^n$,*

$$\left(\int_K f_1\right)^\alpha \left(\int_K f_2\right)^\beta \leq \left(\int_K f_3\right)^\alpha \left(\int_K f_4\right)^\beta$$

2. *For every needle N in $\mathbb{R}^n$,*

$$\left(\int_N f_1\right)^\alpha \left(\int_N f_2\right)^\beta \leq \left(\int_N f_3\right)^\alpha \left(\int_N f_4\right)^\beta$$

The idea of converting estimates for convex bodies into local estimates for line segments is utilized in many problems in convex geometry, including more recent work on the isoperimetric problem, the problem which originally motivated [5]. As an aside, these "needle decompositions" resemble wave packet decompositions often utilized in Euclidean harmonic analysis to quasi-localize certain norm estimates (see [3] and [4] for examples).

7.3 The scalar valued inequalities

Carbery and Wright utilize Proposition 7.3 to reduce Theorems 1 and 2 to weighted $L^q \to L^r$ inequalities for polynomials $p : \mathbb{R} \to \mathbb{C}$. We present the reductions of Theorem 1 and Theorem 2 below, alongside the main theorems utilized in their proof.

Theorem 1 is *equivalent* to the following proposition.

Proposition 7.4. *Let $p : \mathbb{R} \to \mathbb{C}$ be a polynomial of degree at most d, $n \in \mathbb{N}$ and $0 \leq q \leq \infty$. Then, for every $\lambda > 1$ and $\alpha > 0$, there is an absolute constant C independent of the above parameters such that,*

$$\left(\frac{\int_0^1 |p(t)|^{\frac{q}{d}}(\lambda - t)^{n-1}dt}{\int_0^1 (\lambda - t)^{n-1}dt}\right)^{\frac{1}{q}} \leq C\frac{\big[nB(n, q+1)\big]^{\frac{1}{q}}}{\big[nB(n, r+1)\big]^{\frac{1}{r}}}\left(\frac{\int_0^1 |p(t)|^{\frac{r}{d}}(\lambda - t)^{n-1}dt}{\int_0^1 (\lambda - t)^{n-1}dt}\right)^{\frac{1}{r}}.$$

The reduction to scalar-valued inequalities allows one to apply standard L^q-norm inequalities for univariate polynomials, such as the following elementary inequality.

Lemma 7.5. *There is an absolute constant C so that if $p : \mathbb{R} \to \mathbb{C}$ is a polynomial of degree at most d, if $0 \leq r \leq q \leq \infty$, and if $t \geq u$, then*

$$\left(\frac{1}{t}\int_0^t |p|^{\frac{q}{d}}\right)^{\frac{1}{q}} \leq C\frac{t}{u}\left(\frac{1}{u}\int_0^u |p|^{\frac{r}{d}}\right)^{\frac{1}{r}}$$

We call such an inequality "elementary" as it relies solely on first-principles of finite L^q norms of polynomials, such as the inclusion of of L^q in L^r when $r \leq q$ and the fact that univariate polynomials of degree at-most d satisfy a quasi-homogeneity condition of order d.

Applying the Kannan-Lovász-Simonovits localization lemma reduces Theorem 2 to the following weighted distributional inequality.

Proposition 7.6. *Let $p : \mathbb{R} \to \mathbb{C}$ be a polynomial of degree at most d, $n \in \mathbb{N}$, $\lambda \geq 1$, and $0 \leq q \leq \infty$. Then there exists an absolute constant C independent of the involved parameters so that for any $\alpha > 0$,*

$$\left(\frac{\int_0^1 |p(t)|^{\frac{q}{d}}(\lambda - t)^{n-1}dt}{\int_0^1 (\lambda - t)^{n-1}dt}\right)^{\frac{1}{q}} \frac{\alpha^{-\frac{1}{d}}\int_0^1 \chi_\alpha(t)(\lambda - t)^{n-1}dt}{\int_0^1 (\lambda - t)^{n-1}dt} \leq Cn\big[nB(n, q+1)\big]^{\frac{1}{q}},$$

where $\chi_\alpha(t)$ is the indicator function of the set $\{x \in K : |p(x)| \geq \alpha\}$.

When $q = \infty$ and $n = 1$, Proposition 6 simplifies to the following.

Lemma 7.7. *There is an absolute constant C so that for all polynomials $p : \mathbb{R} \to \mathbb{C}$ of degree at most d and all intervals $I \subset \mathbb{R}$,*

$$||p||_{L^\infty(I)}^{\frac{1}{d}}\ \alpha^{-\frac{1}{d}}\big|\{x \in I : |p(x)| \leq \alpha\}\big| \leq C\big|I\big|.$$

This easily-stated distributional inequality had been known for at least seventy years prior to the publication of [1]. It is a consequence of the following scalar-valued Cartan inequality.

Proposition 7.8 (Cartan's polynomial inequality, [2])**.** *Let $w_1, ..., w_d$ be d points in the complex plane and let $h > 0$. Then the set of points $z \in \mathbb{C}$ such that the inequality*

$$\prod_{j=1}^{d} |z - w_j| \leq h^d \tag{7.3}$$

holds can be covered by at most d discs $Q_1, ..., Q_d$, whose radii sum to $2eh$.

For monic polynomials $p : \mathbb{R} \to \mathbb{C}$ of degree $k \geq 1$, inequality (7.3) implies that for any interval $I \subset \mathbb{R}$,

$$\big|\{x \in I : |p(x)| \leq \alpha\}\big| \leq C\alpha^{\frac{1}{k}},$$

as the intersection of I with the family Q_j provided by Cartan's lemma has length at most $2 \cdot (2e\alpha)^{\frac{1}{d}}$. Proposition 6 for general n and q is proven by combining the Remez-type inequality for the L^q norm, and using Lemma 7 to estimate the distribution function χ_α.

7.4 The vector valued inequalities

To obtain the general vector valued inequalities of Theorems 1 and 2, Carbery and Wright work over the following class of functions.

Definition 7.9. *A function $u : \mathbb{R}^n \to \mathbb{R}$ is of class $\mathcal{L}$ if there exists a plurisubharmonic function $\tilde{u} : \mathbb{C}^n \to \mathbb{R}$ with $\limsup_{|z|\to\infty} \frac{\tilde{u}(z)}{\log|z|} \leq 1$ and $u = \tilde{u}\big|_{\mathbb{R}^n}\tilde{u}$.*

Functions of class $\mathcal{L}$ admit representations with probability measures in $\mathbb{C}^n$. When $n = 1$, any function $u(x)$ of class $\mathcal{L}$ is representable as,

$$u(x) = A + \int \log\big|x - \zeta\big| d\mu(\zeta)$$

where $A \in \mathbb{C}$ and μ is a probability measure on $\mathbb{C}$. This representation applies the theory of logarithmic potentials to the functions of class $\mathcal{L}$. In particular, we have the following variant of Cartan's Lemma.

Proposition 7.10 (Cartan's lemma for logarithmic potentials, [6])**.** *Let μ be a Borel probability measure on $\mathbb{C}$ and set,*

$$u(z) := \int \log\big|z - \zeta\big| d\mu(\zeta).$$

Then, for any $0 < \alpha < 1$, the set of points $z \in \mathbb{C}$ such that $|u(z)| \leq \log\alpha$ can be covered by discs $Q_j \subset \mathbb{C}$ the sum of whose radii is bounded above by 5α.

Proposition 10 allows one to prove the analogous distributional inequality of Lemma 7 for functions $\exp(u)$ where u is of class $\mathcal{L}$; a similar Remez-type inequality of Lemma 5 is also available. Reworking Proposition 6 with a general function $\exp(u) : \mathbb{R} \to \mathbb{C}$ with u of class $\mathcal{L}$ leads to the following.

Theorem 7.11. *Let $u : \mathbb{R}^n \to \mathbb{R}$ be a function of class $\mathcal{L}$, $0 \leq q \leq \infty$ and K be a convex body in $\mathbb{R}^n$ of volume 1. Then there exists an absolute constant C independent of q, K, n and u so that*

$$\big|\big|e^u\big|\big|_{L^q(K)} \big|\big|e^{-u}\big|\big|_{L^{1,\infty}(K)} \leq Cn\big[nB(n, q+1)\big]^{\frac{1}{q}}$$

A result similar to Theorem 1, with functions $\exp(u)$ with u of class $\mathcal{L}$ replacing the polynomial $p : \mathbb{R}^n \to \mathbb{C}$, also holds.

To prove the Carbery-Wright inequalities in full generality, we observe that whenever $p : \mathbb{R}^n \to X$ is a polynomial taking values in a Banach space, then $u(x) := \frac{1}{d}||p(x)||$ is a function of class $\mathcal{L}$. Since, in this case, $\exp(u(x)) := ||p(x)||^{\frac{1}{d}}$, Theorem 11 implies the distributional Carbery-Wright inequality of Theorem 2. An analogous transference holds for the norm inequality of Theorem 1.

Bibliography

[1] T. Carbery, J. Wright, *Distributional and L^q norm inequalities over convex bodies in $\mathbb{R}^n$*. Math. Res. Letters **8**(3) (2001), 233–248.

[2] H. Cartan, *Su les systemes de fonctions holomorphes a varietes lineaires lacunaires et leurs applications*, Ann. Sci. Ecole Norm. Sup. **45** (1928), 225–346.

[3] L. Guth, A. Iosevich, Y. Ou, H. Wang, *On Falconer's distance set problem in the plane*. Invent. Math. **219**(1), (2020).

[4] L. Guth, D. Maldague, H. Wang. *Improved decoupling for the parabola*. arXiv:2009.07953 (2020).

[5] R. Kannan, L. Lovász and M. Simonovits, *Isoperimetric problems for convex bodies and a localization lemma*. Discrete Comput. Geom. **13** (1995), 541–559.

[6] B.Y. Levin, *Lectures on entire functions*. Trans. Math. Mono. (1996).

Caleb Marshall, University of British Columbia
email: **calebdoesmath@gmail.com**

Chapter 8

Learning DNF in Time $2^{\widetilde{O}(n^{1/3})}$

after Klivans and Servedio [1]
A summary written by Shivam Nadimpalli

Abstract. We show how to represent any s-term DNF over n variables as a polynomial threshold function of degree $O(n^{1/3} \log s)$, matching (up to a logarithmic factor) a lower bound obtained by Minsky and Papert. As a consequence of this, we can obtain the fastest known algorithm for learning polynomial sized DNFs, one of the central problems in computational learning theory.

8.1 Introduction

We assume familiarity with the PAC model of learning, and refer the reader to [7] for background. We set the stage with some preliminary definitions. A Boolean function $f : \{0,1\}^n \to \{\pm 1\}$ is a *degree-d polynomial threshold function (PTF)* if there exists a degree-d polynomial $p : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = \text{sign}(p(x)) \qquad \text{for all } x \in \{0,1\}^n.$$

Recall that a *disjunctive normal form (DNF)* is a Boolean function that is an "OR of ANDs," or more formally, is a function $f : \{0,1\}^n \to \{\pm 1\}$ given by

$$f(x) = \bigvee_{i=1}^{s} T_i \qquad \text{where} \qquad T_i = \bigwedge_{j=1}^{t_i} l_{i,j}$$

where the literal $l_{i,j}$ is either x_k or $\neg x_k$ for some $k \in [n]$. We call each T_i a *term*, and say that f is a s-term DNF; if $t_i \leq t$ for all $i \in [s]$, then we say that f is a s-term t-DNF. Finally, a *read-once* DNF is a DNF which contains at most one occurance of each variable.

In [6], Minsky and Papert proved, among other things, that there exists a read-once DNF formula which cannot be computed by any PTF of degree less than $\Omega(n^{1/3})$. Klivans and Servedio [1] show that Minsky and Papert's lower bound is tight up to a logarithmic factor.

Theorem 8.1. *Any s-term DNF over $\{0,1\}^n$ can be expressed as a polynomial threshold function of degree $O(n^{1/3} \log s)$.*

Theorem 8.1 immediately implies the fastest known algorithm for learning polynomial size DNFs (i.e. DNFs with poly(n) number of terms), one of the central open problems in computational learning theory first introduced by Valiant [3]; this is immediate from the following.

Proposition 8.2. *Let $\mathcal{C}$ be a class of functions which can be expressed as a degree-d PTF over $\{0,1\}^n$. Then there is a PAC learning algorithm for $\mathcal{C}$ which runs in time $n^{O(d)}$.*

Proposition 8.2 is not too difficult to prove via a "kernel trick," viewing a degree-d PTF over $\{0,1\}^n$ as a degree-1 PTF (or a *linear threshold function*—LTF for short) over the space of all multilinear monomials of degree at most d, and then recalling that it is easy to PAC learn LTFs via linear programming in $\mathrm{poly}(n)$ time.

Because of Proposition 8.2, upper bounds on the degree of PTFs computing a DNF translate directly into bounds on the running time of a DNF learning algorithm. Indeed, this sheds a new perspective on previous algorithms for learning DNF as well:

- The main structural result of Bshouty [4] implies that any s-term DNF can be expressed as a PTF of degree $O((n \log n \log s)^{1/2})$; and
- The techniques of Tarui and Tsukiji [5] imply that any s-term DNF can be expressed as a PTF of degree $O(n^{1/2} \log s)$.

8.2 Proving Theorem 8.1

The proof of Theorem 8.1 relies on the following theorem.

Theorem 8.3. *Any s-term t-DNF can be expressed as a PTF of degree $O(t^{1/2} \log s)$.*

We present a complete proof of Theorem 8.3 in Section 8.2.1. Theorem 8.1 follows by combining Theorem 8.3 with a decomposition technique due to Bshouty [4], we sketch this in Section 8.2.2.

8.2.1 Representing s-term t-DNFs as PTFs

In this section we will prove Theorem 8.3. Our proof will make use of the Chebyshev polynomials of the first kind; these polynomials have found several applications in approximation theory and numerical analysis [8].

Proposition 8.4. *The d^{th} Chebyshev polynomial of the first kind, written $C_d(x)$ is a univariate degree-d polynomial which satisfies*

1. *$|C_d(x)| \leq 1$ for all $x \in [-1,1]$, with $C_d(1) = 1$; and*
2. *$C'_d(x) \geq d^2$ for $x > 1$, with $C'_d(1) = d^2$.*

We turn to proving Theorem 8.3.

Proof of Theorem 8.3. Let f be an s-term t-DNF, i.e.

$$f(x) = \bigvee_{i=1}^{s} T_i \qquad \text{where} \qquad T_i = \bigwedge_{j=1}^{t_i} l_{i,j}$$

where the $l_{i,j}$ are literals. We have $t_i \leq t$ for all $i \in [s]$. For $i \in [s]$, define S_i as

$$S_i(x) = \sum_{j=1}^{t_i} \varphi(l_{i,j}) \qquad \text{where} \qquad \varphi(l_{i,j}) = \begin{cases} x_k & l_{i,j} = x_k \text{ for some } k \\ 1 - x_k & l_{i,j} = \neg x_k \text{ for some } k \end{cases}.$$

Note that $S_i(x)$ is a degree-1 polynomial. We next define the polynomial

$$Q_i(x) = C_d\left(\frac{S_i(x)}{t_i}\left(1 + \frac{1}{t}\right)\right) \qquad \text{setting } d := \lceil t^{1/2} \rceil.$$

Items 1 and 2 of Fact 9 respectively imply that

1. If $S_i(x)/t_i \in [0, 1 - \frac{1}{t}]$, then $|Q_i(x)| \leq 1$; and
2. If $S_i(x)/t_i = 1$, then $Q_i(x) \geq 2$.

Consider the polynomial threshold function sign $\left(P(x) - s - \frac{1}{2}\right)$ where we define

$$P(x) := \sum_{i=1}^{s} Q_i(x)^{\log 2s}.$$

As Q_i is a polynomial of degree $d = \lceil t^{1/2} \rceil$ and each S_i is of degree 1, it follows that P has degree $O(t^{1/2} \log s)$. We next show that P does indeed compute f correctly. Fix an arbitrary element $x \in \{0,1\}^n$:

- If $f(x) = 0$, then $S_i/t_i \in [0, 1 - \frac{1}{t}]$ for all $i \in [s]$; consequently, $|Q_i(x)| \leq 1$ for all i, and so $P(x) \leq s$.
- If $f(x) = 1$, then $S_i/t_i = 1$ for some $i \in [s]$; consequently $Q_i(x) \geq 2$, and so $Q_i(x)^{\log 2s}$ contributes $2s$ to $P(x)$. As $Q_i(x) \geq -1$ for all i, it follows that $P(x) \geq s + 1$.

This completes the proof. □

8.2.2 DNFs to Decision Trees to PTFs

In this section, we describe (at a high level) how to go from Theorem 8.3 to Theorem 8.1. Recall that a *decision tree* $\mathcal{T}$ is a representation of a Boolean function as a binary tree whose each internal node is labelled with a variable x_i and each leaf node is labelled with a Boolean function. The value $\mathcal{T}(x)$ is then computed as follows: At each internal node labelled by x_i, if $x_i = 0$ we take the left branch, and otherwise take the right branch; at a leaf node labelled with ℓ, we output $\ell(x)$.

Lemma 8.5 (Lemma 10 of [1])**.** *Let $f : \{0,1\}^n \to \{\pm 1\}$ be an s-term DNF. For all $t \in [n]$, f can be expressed as a decision tree $\mathcal{T}$ where*

- *Each leaf of $\mathcal{T}$ contains an s-term t-DNF; and*
- *$\mathcal{T}$ has rank at most $(2n/t) \ln s + 1$.*

We sketch the proof of Lemma 8.5: Let $T_1, \ldots, T_p$ be the terms of f of size at least t. By the first moment method, there must be some variable x_i that appears (either negated or un-negated) in at least $\frac{pt}{n}$ of these terms; we start constructing $\mathcal{T}$ by placing x_i at the root and recursing on the functions obtained from f by restricting $x_i \leftarrow 0$ and $x_i \leftarrow 1$. The recursion terminates when a DNF with no terms larger than t is obtained.

Finally, we very briefly sketch a proof of Theorem 8.1. Set $t := n^{2/3}$. From Lemma 8.5 and Theorem 8.3, we know that f can be expressed as a decision tree $\mathcal{T}$ of rank $(2n/t) \ln s + 1$ where each leaf contains a degree-$O(t^{1/2} \log s)$ PTF. It is straightforward to check that $\mathcal{T}$ can be represented as a *decision list* each output of which is a leaf node of $\mathcal{T}$ (in this case, a degree-$O(t^{1/2} \log s)$ PTF). So as to keep this summary short, we do not formally describe the decision list model; representing a decision list as a PTF computing f is a straightforward construction that can be found in Section 3.3 of [1].

Bibliography

[1] A. Klivans and R. A. Servedio, *Learning DNF in time $2^{\widetilde{O}(n^{1/3})}$*. Proceedings of the Thirty-Third Annual Symposium on Theory of Computing 258-265 (2001).

[2] L. Pitt and L. Valiant, *Computational limitations on learning from examples.* Journal of the ACM, Volume 35 Number 4 965-984 (1988).

[3] L. Valiant, *A theory of the learnable.* Communications of the ACM, Volume 27 1134-1142 (1984).

[4] N. Bshouty, *A subexponential exact learning algorithm for DNF using equivalence queries.* Information Processing Letters 59 37-39 (1996).

[5] J. Tarui and T. Tsukiji, *Learning DNF by approximating inclusion-exclusion formulae.* Proceedings of IEEE Conference on Computational Complexit, 215-220 (1999).

[6] M. Minsky and S. Papert, *Perceptrons.* MIT Press (1968).

[7] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory.* MIT Press (1994).

[8] E. W. Cheney. *Introduction to approximation theory.* McGraw-Hill (1966).

Shivam Nadimpalli, Columbia University
email: **sn2855@columbia.edu**

Chapter 9

Agnostically Learning Halfspaces

after A. Kalai, A. Klivans, Y. Mansour, R. Servedio [5]
A summary written by Lucas Pesenti

Abstract. We consider the task of *agnostically learning halfspaces* under distributional assumptions. Given samples from a nice distribution and *arbitrary* $\{-1, 1\}$ labels, the goal is to output a hypothesis that classifies the data on the true distribution nearly as well as the best halfspace does. We present an algorithm based on *polynomial regression* that achieves near-optimal sample and time complexity guarantees.

9.1 Introduction

A *halfspace* of $\mathbb{R}^n$ is a function $h : \mathbb{R}^n \to \{\pm 1\}$ of the form

$$h(\mathbf{x}) = \text{sign}(\langle \mathbf{x}, \mathbf{u} \rangle - t),$$

where $\mathbf{u} \in \mathbb{R}^n$ and $t \in \mathbb{R}$. The function sign is defined by $\text{sign}(z) = 1$ if $z \geq 0$ and $\text{sign}(z) = -1$ if $z < 0$. We let $\mathcal{H}$ be the set of all halfspaces of $\mathbb{R}^n$.

Learning halfspaces is a good toy model for classification that has been extensively studied in the machine learning community since the 1950s. In the noiseless setting, that is, if the learner receives samples of the form $(\mathbf{x}, h(\mathbf{x}))$ for an unknown $h \in \mathcal{H}$, the problem of learning h can be reduced to finding a feasible point in a polytope described by linear inequalities, for which there exist efficient algorithms. We focus on a noisy version of the problem, called *agnostic learning*, in which the data is not necessarily labeled according to a halfspace. The goal of the learner is to output a classifier that is competitive with the best halfspace classifier of the data.

Formally, the learner is given i.i.d. samples $(\mathbf{x}, y)$ from a (partially) unknown distribution $\mathcal{D}$ over $(\mathbf{x}, y) \in \mathbb{R}^n \times \{\pm 1\}$. We assume that the learner has a priori information about the marginal distribution $\mathcal{D}_{\mathbf{x}}$ of the examples $\mathbf{x}$ under $\mathcal{D}$ — more specifically, that $\mathbf{x}$ is uniform on the n-dimensional hypercube $\{\pm 1\}^n$. On the other hand, the distribution of y given $\mathbf{x}$ is arbitrary.

We measure the loss $\text{err}(f) \in [0, 1]$ of a hypothesis $f : \mathbb{R}^n \to \{\pm 1\}$ by the probability of missclassification under the true distribution:

$$\text{err}(f) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(f(\mathbf{x}) \neq y).$$

The goal of the learner is to output a hypothesis $f : \mathbb{R}^n \to \{\pm 1\}$ that achieves error $\text{err}(f)$ as close as possible to

$$\text{opt} = \min_{h\in\mathcal{H}} \text{err}(h). \tag{9.1}$$

Our main focus in this note is the following algorithmic result for agnostically learning halfspaces. We use the $\tilde{O}$ notation to hide logarithmic factors.

Theorem 9.1 ([5]). *Suppose that $\mathcal{D}_{\mathbf{x}}$ is the uniform distribution on $\{\pm 1\}^n$. There is an algorithm that, given access to $m = n^{\tilde{O}(1/\varepsilon^2)}$ samples from $\mathcal{D}$, outputs in time $m^{O(1)}$ a hypothesis $f : \mathbb{R}^n \to \{\pm 1\}$ with expected error[1] at most opt $+ \varepsilon$.*

Let us underline at this point that the hypothesis output by the algorithm will not necessarily be a halfspace — we only require that it is an efficiently evaluable function (this is usually called *improper* learning). In fact, here the hypothesis will be a low-degree polynomial threshold function (that is, of the form $f(\mathbf{x}) = \operatorname{sign}(p(\mathbf{x}) - t)$, where $p(\mathbf{x})$ is a low-degree polynomial).

Adversarial label noise To get some intuition for the assumptions of Theorem 9.1, consider the following special case. Suppose that instead of getting samples $(\mathbf{x}, h(\mathbf{x}))$ from a halfspace $h \in \mathcal{H}$, an adversary first chooses an η-fraction of the samples and flips their label. Theorem 9.1 applies in this setting and opt $= \eta$ corresponds to the noise rate.

9.2 Agnostic learning via polynomial regression

We start by giving some intuition for the algorithm behind Theorem 9.1. Another way to write the error of a hypothesis $f : \mathbb{R}^n \to \{\pm 1\}$ is

$$\operatorname{err}(f) = \frac{1}{2}\,\mathbb{E}_{\mathcal{D}}\,|f(\mathbf{x}) - y|.$$

This might motivate us to look at the relaxed problem of ℓ_1-regression, namely finding $f : \mathbb{R}^n \to \mathbb{R}$ minimizing $\frac{1}{2}\,\mathbb{E}_{\mathcal{D}}\,|f(\mathbf{x}) - y|$. One key observation is that these problems are actually equivalent — any $f : \mathbb{R}^n \to \mathbb{R}$ can be rounded into a classifier $g : \mathbb{R}^n \to \{\pm 1\}$ without increasing the error. To see why this holds, let $g_t(\mathbf{x}) = \operatorname{sign}(f(\mathbf{x}) - t)$ and draw t at random from the interval $[-1, 1]$. Then,

$$\mathbb{E}_t \operatorname{err}(g_t) = \frac{1}{2}\int_{-1}^{1} \mathbb{E}_{\mathcal{D}}\,\mathbf{1}_{\operatorname{sign}(f(\mathbf{x})-t)\neq y}\,\mathrm{d}t \leq \frac{1}{2}\,\mathbb{E}_{\mathcal{D}}\,|f(\mathbf{x}) - y| = \operatorname{err}(f). \tag{9.2}$$

In particular, there exists a threshold $t \in [-1, 1]$ satisfying $\operatorname{err}(g_t) \leq \operatorname{err}(f)$.

Unfortunately, solving ℓ_1-regression exactly is hard in general. We will therefore further restrict the class of functions to optimize over. From the previous discussion, it makes sense to take this class rich enough to capture good approximations of all halfspaces, while not blowing up the computational cost of the regression problem. When $\mathbf{x}$ is uniform on the hypercube, the theory of Fourier analysis suggest to look at low-degree polynomials.

9.2.1 The polynomial regression algorithm

We are now ready to describe the ℓ_1-polynomial regression algorithm. Given an integer $d \geq 1$ (the degree) and i.i.d. samples $(\mathbf{x}^i, y^i) \sim \mathcal{D}$ for $1 \leq i \leq m$:

1. Pick the multilinear polynomial p of degree at most d that minimizes the ℓ_1-empirical error $\frac{1}{2m}\sum_{i\leq m} |p(\mathbf{x}^i) - y^i|$.
2. Pick the threshold $t \in \mathbb{R}$ that minimizes $\frac{1}{2m}\sum_{i\leq m} |\operatorname{sign}(p(\mathbf{x}^i) - t) - y^i|$.
3. Output the hypothesis $f(\mathbf{x}) = \operatorname{sign}(p(\mathbf{x}) - t)$.

Complexity The optimization problem in Step 1 is equivalent to the following linear program over variables $z_1, \ldots, z_m$ and $\{\alpha_S : |S| \leq d\}$:

$$\begin{aligned}
\min \quad & \sum_{i\leq m} z_i \\
\text{s.t.} \quad & z_i \geq \sum_{|S|\leq d} \alpha_S \prod_{j\in S} \mathbf{x}^i_j - y_i && \text{for } i = 1, \ldots, m \\
& z_i \geq y_i - \sum_{|S|\leq d} \alpha_S \prod_{j\in S} \mathbf{x}^i_j && \text{for } i = 1, \ldots, m
\end{aligned}$$

[1] A corresponding high probability statement can be obtained from Markov's inequality.

The time complexity of solving this linear program is $\mathrm{poly}(m, n^d)$ under mild assumptions on the support of $\mathcal{D}$. Moreover, Step 2 of the algorithm is equivalent to finding a point $t \in \mathbb{R}$ that belongs to a maximum number of intervals, for which there is an easy nearly linear-time algorithm.

9.2.2 Analysis of the algorithm

We now switch to the analysis of the ℓ_1-polynomial regression algorithm. We relate the error of the hypothesis to how well the target optimal classifier can be approximated by degree-d polynomials in ℓ_1-norm. We define the ℓ_1-norm of a function $g : \mathbb{R}^n \to \mathbb{R}$ as $\|g\|_1 = \mathbb{E}_{\mathcal{D}_\mathbf{x}} |g(\mathbf{x})|$.

Theorem 9.2. *There is a universal constant $C > 0$ such that the following holds. For any $g : \mathbb{R}^n \to \{\pm 1\}$, let $\varepsilon = \min \|g - q\|_1$, where the minimum is over all degree-d multilinear polynomials q. Provided that $m \geq n^{Cd}/\varepsilon^2$, the ℓ_1-polynomial regression algorithm achieves expected error at most $\mathrm{err}(g) + \varepsilon$.*

Proof. Let q be the degree-d polynomial that achieves $\|g - q\|_1 = \varepsilon$. First, the argument in (9.2) and Step 1 of the algorithm ensure that

$$\frac{1}{m} \sum_{i \leq m} \mathbf{1}_{f(\mathbf{x}^i) \neq y^i} \leq \frac{1}{2m} \sum_{i \leq m} |p(\mathbf{x}^i) - y^i| \leq \frac{1}{2m} \sum_{i \leq m} |q(\mathbf{x}^i) - y^i|. \tag{9.3}$$

We use standard arguments to argue about uniform concentration of the left-hand side. The VC-dimension of degree-d polynomial threshold functions is at most $n^{O(d)}$, so if $m = n^{O(d)}/\varepsilon^2$,

$$\mathbb{E}_{\mathcal{D}^{\otimes m}} \left[\frac{1}{m} \sum_{i \leq m} \mathbf{1}_{f(\mathbf{x}^i) \neq y^i} \right] \geq \mathbb{E}_{\mathcal{D}^{\otimes m}} \mathrm{err}(f) - \frac{\varepsilon}{2}.$$

On the other hand, the expected value of the right-hand side of (9.3) is

$$\frac{1}{2} \mathbb{E}_{\mathcal{D}} |q(\mathbf{x}) - y| \leq \frac{1}{2} \mathbb{E}_{\mathcal{D}} |q(\mathbf{x}) - g(\mathbf{x})| + \frac{1}{2} \mathbb{E}_{\mathcal{D}} |g(\mathbf{x}) - y| = \frac{\varepsilon}{2} + \mathrm{err}(g).$$

Putting everything together, we get $\mathbb{E}_{\mathcal{D}^{\otimes m}} \mathrm{err}(f) \leq \mathrm{err}(g) + \varepsilon$. □

9.2.3 Proof of Theorem 9.1

To deduce Theorem 9.1 from Theorem 9.2, it remains to bound the approximability of halfspaces by low-degree polynomials. The following is based on an observation of [4] to improve the initial result of [6].

Theorem 9.3. *Assume that $\mathcal{D}_\mathbf{x}$ is the uniform distribution on $\{\pm 1\}^n$. For any $h \in \mathcal{H}$ and $\varepsilon > 0$, there exists a multinear polynomial p of degree $\tilde{O}(1/\varepsilon^2)$ such that $\|h - p\|_1 \leq \varepsilon$.*

We briefly sketch the idea that underlies the proof of Theorem 9.3. We first compare $h(\mathbf{x})$ to its noisy version $\tilde{h}(\mathbf{x}) = \mathbb{E}\left[h(\mathbf{y}) \,|\, \mathbf{x}\right]$, where every entry of $\mathbf{y}$ is the corresponding entry of $\mathbf{x}$ flipped independently with probability ε^2. The high-degree part of $\tilde{h}$ is killed by the noise, so that $\tilde{h}$ is well-approximated by its low-degree part. At this point, it essentially only remains to bound the noisy approximation error $\mathbb{E}\,|h(\mathbf{x}) - \tilde{h}(\mathbf{x})| \leq \mathbb{E}\,|h(\mathbf{x}) - h(\mathbf{y})|$. This last quantity, called noise sensitivity of h, is at most $O(\varepsilon)$ for any halfspace — this follows from a careful charging argument on the set of coordinates that are flipped by the noise.

Proof of Theorem 9.1. Apply Theorem 9.2 with g being the optimal halfspace in (9.1). By Theorem 9.3, one can set the parameter d of the polynomial regression algorithm to $d = \tilde{O}(1/\varepsilon^2)$ to get a hypothesis achieving expected error $\mathrm{opt} + \varepsilon$ in sample and time complexity $n^{\tilde{O}(1/\varepsilon^2)}$. □

9.3 Hardness based on learning parity with noise

In this section, we follow an argument of [5] that gives evidence that the complexity of Theorem 9.1 may not be improvable to $n^{O(1/\varepsilon^{2-\beta})}$ with $\beta > 0$. The idea of [5] is based on an interesting reduction from learning parity with noise, a long-standing open problem in theoretical computer science.

In the learning parity with noise problem, there is an unknown parity function $f(\mathbf{x}) = \prod_{i \in S} \mathbf{x}_i$ with $S \subseteq \{1, \dots, n\}$. The learner receives samples of the form $(\mathbf{x}, y)$ with $\mathbf{x} \sim \{\pm 1\}^n$, $y = f(\mathbf{x})$ with probability 0.9 and $y = -f(\mathbf{x})$ with probability 0.1. The goal of the learner is to recover S.

Despite a lot of effort in the cryptography and complexity theory communities, the best algorithm for learning parity with noise [1] still requires time $2^{O(n/\log n)}$. We show that further improving Theorem 9.1 would result in a subexponential time algorithm for the problem.

Theorem 9.4. *Agnostically learning halfspaces in time $n^{O(1/\varepsilon^{2-\beta})}$ implies learning parity with noise in time $2^{\tilde{O}(n^{1-\beta/2})}$.*

Proof sketch. Fix $i \in \{1, \dots, n\}$. We try to identify whether $i \in S$ by agnostically learning a halfspace on the samples $(\mathbf{x}_{-i}, y)$ (where $\mathbf{x}_{-i} \in \mathbb{R}^{n-1}$ is $\mathbf{x}$ with the i-th coordinate removed). Setting $\varepsilon = \Theta(1/\sqrt{n})$ allows to distinguish between the following two cases:

- If $i \in S$, any function of $\mathbf{x}_{-i}$ has error equal to $\frac{1}{2}$.
- If $i \notin S$, then $h(\mathbf{x}_{-i}) = \mathrm{sign}(\sum_{j \in S} \mathbf{x}_j)$ has error $\frac{1}{2} - \Omega(\frac{1}{\sqrt{n}})$. □

9.4 Generalizations and optimality

Learning other classes The framework we have described in Section 9.2 is not restricted to learning halfspaces. In fact, we can apply the ℓ_1-polynomial regression algorithm and Theorem 9.2 to any class of Boolean-valued functions. Interestingly, this is essentially the best agnostic learning algorithm for *any* such class in the powerful *statistical query* model [2].

Error measurement The results we have stated in this note focus on the error being measured additively. If we ask instead for a hypothesis with error $1.001\mathrm{opt} + \varepsilon$, we can do better — a boosted version of the polynomial regression algorithm achieves agnostic learning in $\mathrm{poly}(n, 1/\varepsilon)$ time [3].

Other distributions The same analysis works for agnostically learning halfspaces when $\mathbf{x}$ is uniform on the sphere, or from a log-concave distribution [5]. Related algorithms also handle the case of adversarially corrupted $\mathbf{x}$'s, but the bounds obtained in [5] for this model are far from optimal.

Bibliography

[1] Blum, A., Kalai, A., Wasserman, H., *Noise-tolerant learning, the parity problem, and the statistical query model.* STOC 2000.

[2] Dachman-Soled, D., Feldman, V., Tan, L.Y., Wan, A., Wimmer, K., *Approximate resilience, monotonicity, and the complexity of agnostic learning.* SODA 2015.

[3] Daniely, A., *A PTAS for Agnostically Learning Halfspaces.* COLT 2015.

[4] Feldman, V., Kothari, P., Vondrak, J., *Tight bounds on ℓ_1-approximation and learning of self-bounding functions.* TCS 2020.

[5] Kalai, A., Klivans, A. , Mansour, Y., Servedio, R., *Agnostically Learning Halfspaces.* FOCS 2005.

[6] Klivans, A., O'Donnell, R., Servedio, R., *Learning intersections and thresholds of halfspaces.* FOCS 2002.

Lucas Pesenti, Bocconi University
email: **lucas.pesenti@phd.unibocconi.it**

Chapter 10

The Correct Exponent for the Gotsman-Linial Conjecture

after D. M. Kane [6]
A summary written by Seung-Yeon Ryoo

Abstract. The asymptotic Gotsman-Linial conjecture claims that a degree-d polynomial threshold function in $n > 1$ variables has average sensitivity at most $O(d\sqrt{n})$. We outline the proof of the partial result due to Kane (2012) that gives an upper bound of $\sqrt{n}(\log n)^{O(d \log d)} 2^{O(d^2 \log d)}$.

10.1 Statement of the conjecture and the main result

Definition 10.1. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a boolean function defined on the hypercube.*

1. *For $i = 1, \cdots, n$, the* i-th influence *of f is defined as*

$$\mathrm{Inf}_i(f) = \frac{1}{4}\mathbb{E}\left[|f(A) - f(A^i)|^2\right],$$

where A is a Bernoulli random vector, i.e., it is chosen uniformly randomly on $\{-1,1\}^n$, and A^i is obtained by negating the i-th entry of A.

2. *The* average sensitivity *or* total influence *of f is defined as*

$$\mathbb{AS}(f) := \sum_{i=1}^{n} \mathrm{Inf}_i(f).$$

3. *For $d \in \mathbb{Z}_{>0}$, the function f is a* degree-d polynomial threshold function *if $f(x) = \operatorname{sgn} p(x)$ for some degree-d real polynomial p.*

Gotsman and Linial [4] conjectured the following regarding the maximum average sensitivity of polynomial threshold functions.

Conjecture 10.2 (Gotsman-Linial conjecture)**.** *Let f be a degree-d polynomial threshold function in $n > 1$ variables. Then*

$$\mathbb{AS}(f) \le 2^{-n+1} \sum_{k=0}^{d-1} \binom{n}{\lfloor (n-k)/2 \rfloor} \left(n - \lfloor \frac{n-k}{2} \rfloor\right).$$

The right-hand side in the Gotsman-Linial conjecture is the average sensitivity of the sign of the polynomial function $p_{n,d}\left(\sum_{i=1}^n x_i\right)$, where $p_{n,d}$ is the monic univariate polynomial of degree-d with non-repeated roots at the d integers closest to 0 of opposite parity from n. One can easily see that this term is of order $\Theta(d\sqrt{n})$ for $n = \Omega(d^2)$.

The Gotsman-Linial conjecture has been shown to be false in [2]. However, the following weaker version of the Gotsman-Linial conjecture still stands.

Conjecture 10.3 (Asymptotic Gotsman-Linial conjecture)**.** *Let f be a degree-d polynomial threshold function in n variables. Then*

$$\mathbb{AS}(f) = O(d\sqrt{n}).$$

The text under consideration proves the following statement. Previously, the best-known upper bound was $O_{c,d}(n^{5/6+c})$, due to Kane [5].

Theorem 10.4 ([6, Theorem 2])**.** *Let f be a degree-d polynomial threshold function in $n > 1$ variables. Then*

$$\mathbb{AS}(f) \leq \sqrt{n}(\log n)^{O(d \log d)} 2^{O(d^2 \log d)}.$$

In particular, this obtains the correct exponent of n. The author claims without proof in a talk the stronger statement

$$\mathbb{AS}(f) \leq O_d(\sqrt{n})(\log n)^{O(\log d)}.$$

10.2 Sketch of the proof

Due to page limitations we will sketch the proof of the weaker bound

$$\mathbb{AS}(f) \leq \sqrt{n} \exp(O(d \log \log n)^2).$$

Below, A will denote an n-dimensional Bernoulli random vector and X will denote an independent n-dimensional standard Gaussian random vector.

A very high-level description of the proof method is to approximate the distribution of $p(A)$ by that of $p(X)$, and then to use anticoncentration of $p(X)$ to argue that $p(X)$ doesn't change sign that often.

The classical anticoncentration result due to Carbery and Wright [1] that

$$\Pr\left[|p(X)| \leq \varepsilon |p|_2\right] = O(d\varepsilon^{1/d}), \quad \varepsilon > 0$$

is insufficient for the purposes of proving Theorem 10.4 (here $|p|_2 = (\mathbb{E}[|p(X)|^2])^{1/2}$), because of the power in ε. Instead, the author comes up with the following well-behaved alternative.

Lemma 10.5 ([6, Lemma 9])**.** *If p is a degree-d real polynomial, then*

$$\Pr\left[|p(X)| \leq \varepsilon |\nabla p(X)|\right] = O(d^2\varepsilon).$$

"Nice" polynomials of Bernoulli random vectors behave similarly to polynomials of Gaussian random vectors. The definition of "nice" is as follows.

Definition 10.6. *A real polynomial p in n variables is τ-regular, $\tau > 0$, if*

$$\mathrm{Inf}_i(p) \leq \tau \operatorname{Var}(p(A)), \quad i = 1, \cdots, n.$$

For regular and multilinear polynomials we have the following.

Theorem 10.7 ([7, Theorem 2.1])**.** *If p is a degree-d, τ-regular, and multilinear polynomial, then*

$$|\Pr[p(A) > 0] - \Pr[p(X) > 0]| = O(d\tau^{1/8d}).$$

Combining Lemma 10.5 and Theorem 10.7, we have the following.

Proposition 10.8. *If p is a degree-d, τ-regular, multilinear polynomial, then*

$$\Pr[|p(A)| < \varepsilon |\nabla p(A)|] = O(d^2\varepsilon) + O(d\tau^{1/8d}), \quad \varepsilon > 0.$$

We need to reduce to the regular case. (It is straightforward that we can reduce to the multilinear case.) We do this by conditioning on the values of the high-influence coordinates.

Proposition 10.9 ([3, Theorem 1])**.** *Let f be a degree-d polynomial threshold function and $\tau > 0$. After conditioning on the values of $\tau^{-1}O(d \log(\tau^{-1}))^{O(d)}$ coordinates (chosen adaptively), with probability at least $1 - \tau$ the restricted function is τ-close to a τ-regular degree-d polynomial threshold function.*

Here, we say that two boolean functions $f, g : \{-1,1\}^n \to \{-1,1\}$ are τ-close if $\Pr[f(A) \neq f(B)] < \tau$. Proposition 10.9 can be stated more precisely by stating that there is a decision tree of depth $\tau^{-1}O(d\log(\tau^{-1}))^{O(d)}$, the leaves of which are assigned polynomial threshold functions, such that a random path reaches a leaf, with probability $1-\tau$, with assigned polynomial threshold function that is τ-close to a τ-regular degree-d polynomial threshold function.

Let f be a degree-d polynomial threshold function. By Proposition 10.9, we may adaptively condition on $\tau^{-1}O(d\log(\tau^{-1}))^{O(d)}$ coordinates; this process of changing the fixed coordinates gives a contribution of $\tau^{-1}O(d\log(\tau^{-1}))^{O(d)}$ to the average sensitivity. Let us now consider the average sensitivity of the restricted function. It is irregular with probability at most τ, and these cases gives contribution $O(n\tau)$ to the average sensitivity. For the other cases, it is τ-close to a τ-regular polynomial threshold function, so the sensitivity is at most $O(n\tau)$ more than that of a τ-regular polynomial threshold function.

We have just proven that if f is a degree-d polynomial threshold in n variables, then $\mathbb{AS}(f)$ is at most $\tau^{-1}O(d\log(\tau^{-1}))^{O(d)} + O(n\tau)$ more than the maximum average sensitivity of τ-regular degree-d polynomial threshold functions in n variables. We will put $\tau = n^{-1/2}$. It remains to bound the average sensitivity of an $n^{-1/2}$-regular degree-d polynomial threshold function in n variables.

Here is a first attempt. Let p be a $\tau(=n^{-1/2})$-regular degree-d multilinear polynomial in n variables. Since $|p(A)-p(A^i)| = 2|D_ip(A)|$, $\mathbb{AS}(\operatorname{sgn} p)$ is bounded by the expected number of i so that $2|D_ip(A)| \geq |p(A)|$. Thus

$$\begin{aligned}\mathbb{AS}(\operatorname{sgn} p) &\leq \mathbb{E}\left[\min\left(n, \sum_{i=1}^n \frac{4|D_ip(A)|^2}{|p(A)|^2}\right)\right] = \mathbb{E}\left[\min\left(n, \frac{4|\nabla p(A)|^2}{|p(A)|^2}\right)\right] \\ &\leq 1 + \sum_{k=1}^n \Pr\left[|p(A)| \leq 2k^{-1/2}|\nabla p(A)|\right] \\ &\overset{\text{Proposition 10.8}}{\leq} \sum_{k=1}^n \left[O(\frac{d^2}{\sqrt{k}}) + O(d\tau^{1/8d})\right] = O(d^2\sqrt{n}) + O(dn\tau^{1/8d}).\end{aligned}$$

The problem with this argument is that the error from using Proposition 10.8 is very large. The idea to get over this obstacle is to split the coordinates into $b = n^{1/\Theta(d)}$ equally sized blocks. Then

$$\mathbb{AS}(f) = \sum_{\text{blocks } i} \mathbb{E}[\mathbb{AS}(f|_i^A)],$$

where $f|_i^A : \{-1,1\}^i \to \{-1,1\}$ is the function obtained from f by fixing the coordinates of $\{1,\cdots,n\}\setminus i$ to those obtained from A. Likewise, for a real polynomial function $p : \{-1,1\}^n \to \mathbb{R}$ and a block i we define the restricted polynomial $p|_i^A$.

Let C be an appropriate constant.

Definition 10.10. *A degree-d polynomial p is* good *if*

$$\operatorname{Var}(p) > |p|_2^2(C\log n)^{-d},$$

and is bad *otherwise.*

Definition 10.11. *Fix a degree-d real polynomial p and a block decomposition of $\{1,\cdots,n\}$. Given the value of A, a block i is* good *if $p|_i^A$ is good, and is* bad *otherwise.*

If p is a bad polynomial, then $\operatorname{sgn} p(A) = \operatorname{sgn}\mathbb{E}[p(A)]$ with probability $1-O(n^{-2})$, so that $\mathbb{AS}(f) = O(n^{-1})$. We thus only need to consider contributions from good blocks, i.e., blocks such that if f is restricted to it then f is good.

Let p be a good polynomial. Since $\mathbb{E}[|\nabla p(A)|^2] \geq \operatorname{Var}(p)$, we may apply the Paley-Zygmund inequality to obtain that with probability at least $9^{-d}/2$ we have $|\nabla p(A)|^2 > \operatorname{Var}(p)/2$. On the other hand, with probability at least $1-9^{-d}/4$ we have $|p(A)| < (Cd)^{d/2}|p|_2$. Therefore, with probability at least $9^{-d}/4$ we have $|\nabla p(A)|^2 > (Cd\log n)^{-d}|p(A)|^2$.

Therefore, the expected number of good blocks is at most $2^{O(d)}$ times the expected number of blocks i such that $|\nabla_i p(A)|^2 > (Cd\log n)^{-d}|p(A)|^2$. We estimate

$$\begin{aligned}
2^{-O(d)}\mathbb{E}[\#\{\text{good blocks}\}] &\leq \mathbb{E}\left[\min\left(b, \sum_{\text{blocks } i} \frac{O(d\log n)^d|\nabla_i p(A)|^2}{|p(A)|^2}\right)\right] \\
&= \mathbb{E}\left[\min\left(b, \frac{O(d\log n)^d|\nabla p(A)|^2}{|p(A)|^2}\right)\right] \\
&\leq 1 + \sum_{k=1}^{b} \Pr(|p(A)| \leq k^{-1/2}O(d\log n)^{d/2}|\nabla p(A)|) \\
&\overset{\text{Proposition 10.8}}{\leq} \sum_{k=1}^{b}\left(O((d\log n)^{d/2}d^2k^{-1/2}) + O(d\tau^{1/8d})\right) \\
&\leq O(d\log n)^{d/2}(\sqrt{b} + b\tau^{1/8d}).
\end{aligned}$$

Let $\mathrm{MAS}(d,n)$ denote the maximum of $\mathbb{AS}(f)$ over degree-d polynomial threshold functions f in n variables. For a τ-regular polynomial threshold function f, we have that

$$\begin{aligned}
\mathbb{AS}(f) \leq& O(1) + \mathbb{E}[\#\{\text{good blocks}\}]\,\mathrm{MAS}(d, n/b) \\
\leq& O(d\log n)^{d/2}(\sqrt{b} + bn^{-1/16d})\,\mathrm{MAS}(d, n/b).
\end{aligned}$$

Letting $b = n^{1/8d}$ and considering the $\sqrt{n}O(\log n)^{O(d)}$ error involved in the approximation to regular functions, we have that

$$\mathrm{MAS}(d,n) = O(d\log n)^{O(d)}n^{1/16d}\,\mathrm{MAS}(d, n^{1-1/8d}) + \sqrt{n}O(\log n)^{O(d)}.$$

Iterating this recursion gives

$$\mathrm{MAS}(d,n) \leq \sqrt{n}\exp\left(O(d\log\log n)^2\right).$$

The more precise statement of Theorem 10.4 is obtained by keeping more careful track of the "goodness". Very roughly, this means that one defines for each nonzero polynomial p the quantity

$$\alpha(p) := \mathbb{E}\left[\min\left(1, \frac{|\nabla p(A)|^2}{|p(A)|^2}\right)\right]$$

and keeps track of the value of $\alpha(p)$.

Bibliography

[1] Carbery, A., and Wright, J., *Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$.* Math. Res. Lett. 8(3) 233-248 (2001)

[2] Chapman, B., *The Gotsman-Linial conjecture is false.* Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, (2018)

[3] Diakonikolas, I., Servedio, R., Tan, L.-Y., and Wan, A., *A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions.* 25th Conference on Computational Complexity (CCC) (2010)

[4] Gotsman, C. and Linial, N., *Spectral properties of threshold functions* Combinatorica 14(1), 35-50 (1994)

[5] Kane, D. M., *A structure theorem for poorly anticoncentrated polynomials of Gaussians and applications to the study of polynomial threshold functions.* Ann. Probab. 45(3), 1612-1679 (2017)

[6] Kane, D. M., *The correct exponent for the Gotsman-Linial conjecture.* Comput. Complexity 23(2), 151-175 (2014)

[7] Mossel, E., O'Donnell, R., and Oleszkiewicz, K., *Noise stability of functions with low influences: invariance and optimality.* Proceedings of the 46th Symposium on Foundations of Computer Science (FOCS), 21-30 (2005)

Seung-Yeon Ryoo, Princeton University
email: **sryoo@princeton.edu**

Chapter 11

Pseudorandom Generators from the 2nd Fourier Level via Polarizing Random Walks

work of Chattopadhyay et al. [2, 1, 3]
A summary written by Joseph Slote

Abstract. Pseudorandom generators (PRGs) are central objects in complexity theory. We describe a recent framework for their construction via polarizing random walks in the solid hypercube, effective against function classes with $\mathcal{L}_1$-bounded second-level Fourier coefficients. By these constructions, a self-contained conjecture about the second-level Fourier coefficients of $\mathbb{F}_2$ polynomials would imply a PRG against constant depth circuits with parity gates, an important open problem in circuit complexity.

11.1 Introduction

Let $\mathcal{U}^n$ denote the uniform distribution over $\{-1,1\}^n$. Pseudorandom generators, or PRGs, are deterministic functions that induce sparse distributions on $\{-1,1\}^n$ which nonetheless "look like $\mathcal{U}^n$" to a fixed class of statistical tests (Boolean functions).

Definition 11.1. *A* pseudorandom generator with seed length s *is a function* $g : \{-1,1\}^s \to \{-1,1\}^n$. *For* $\mathcal{F}$ *a set of functions from* $\{-1,1\}^n$ *to* $\{-1,1\}$*, we say* g has error ϵ against $\mathcal{F}$ *if for all* $f \in \mathcal{F}$*,*

$$\left|\mathbb{E}[f(g(\mathcal{U}^s))] - \mathbb{E}[f(\mathcal{U}^n)]\right| \leq \epsilon.$$

PRGs have countless applications, from the derandomization of probabilistic algorithms to providing a basis for private key cryptography. In the context of learning theory, PRGs can be used to prove hardness results: roughly, if data is labeled in a way that appears random, there is little chance the labeling function can be learned.[1] In typical applications one wishes to simultaneously minimize seed length and error, goals that are in tension to various degrees depending on the target class of functions $\mathcal{F}$.

A powerful method for creating PRGs was recently introduced by Chattopadhyay et al. in [1] which leads to very good PRGs for function classes possessing bounded Fourier tails. This tail restriction was then dramatically relaxed in [2], building on the work of [3].

Concretely, define the $\mathcal{L}_{1,2}$ Fourier tail of $f : \{-1,1\}^n \to \{-1,1\}$ as

$$\mathcal{L}_{1,2}[f] = \sum_{i<j\in[n]} |\widehat{f}(\{i,j\})| = \sum_{i<j\in[n]} |\mathbb{E}_{x\sim\mathcal{U}^n}[f(x)x_ix_j]|.$$

[1] See the excellent survey of Vadhan [4].

For $i \in [n]$ and $x \in \{-1,1\}^n$ let $y_i(x)$ (resp. $\bar{y}_i(x)$) be the vector x except with the i^{th} coordinate set to 1 (resp. -1). So y_i and $\bar{y}_i$ are functions from $\{-1,1\}^{n-1}$ to $\{-1,1\}^n$. A function family $\mathcal{F} = \cup_{n \geq 0} \mathcal{F}_n$ is *closed under restrictions* if for all $f \in \mathcal{F}$ and for every coordinate i in the domain of f we have $f \circ y_i$ and $f \circ \bar{y}_i \in \mathcal{F}$. Then we have the theorem:

Theorem 11.2 ([2, Thm. 2])**.** *Let $\mathcal{F}$ be a family be closed under restrictions such that for some $t \geq 1, \mathcal{L}_{1,2}[f] \leq t$ for all $f \in \mathcal{F}$. Then for any $\epsilon > 0$ there exists an explicit PRG for $\mathcal{F}$ with error ϵ and seed length* $\text{poly}(t, \log n, 1/\epsilon)$.

This leads to the exciting conjecture,

Conjecture 11.3. *Let* $\text{Poly}_{n,d}$ *denote those Boolean functions computed by n-variate $\mathbb{F}_2$ polynomials p of degree d. Then $\mathcal{L}_{1,2}[\text{Poly}_{n,d}] \in \mathcal{O}(d^2)$.*

Combined with Theorem 11.2, this conjecture would imply PRGs against $AC^0[\oplus]$, a longstanding open problem.

The proof of Theorem 11.2 is in three steps. First, we find that a weaker, real-valued notion of a PRG, called a *fractional PRG*, can be amplified into a standard PRG via random walk techniques. We then show an $\mathcal{L}_{1,2}$ bound on $\mathcal{F}$ is sufficient for a simple kind of random variable (a low-covariance multivariate Gaussian) to be a fractional PRG against $\mathcal{F}$. Finally, we confirm that this multivariate Gaussian can be (approximately) implemented as a function of a uniform distribution over a small space.

We'll focus on the first two steps of this process, which include nice ideas about random walks as well as some very clever random restriction arguments. The third step [2, pg. 22:6] essentially follows from black-box constructions of good linear codes—which are used to map a $\log(n)$-variate zero-covariance Gaussian to an n-variate Gaussian with mild covariance—and well-studied methods for discretely approximating univariate Gaussians.

11.2 PRGs from fractional PRGs

Definition 11.4. *A p-noticeable fractional PRG with error ϵ is a random variable X on $[-1,1]^n$ such that $|\mathbb{E}[f(X)] - f(\mathbf{0})| \leq \epsilon$ and $\mathbb{E}[X_{(i)}^2] \geq p$ for all coordinates $i \in [n]$.*

We use a fractional PRG X to drive a random walk from the origin of the solid hypercube. We'll see that after a short time the walk polarizes towards a corner of the cube and so we may take the PRG output to be the closest vertex $x \in \{\pm 1\}^n$.

Now we construct the walk. For $y \in [-1,1]^n$ define $\delta : [-1,1]^n \to [0,1]$ coordinatewise by $\delta_i(y) = 1 - |y_i|$. For $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$ define the random variables Y_i with $Y_1 = X_1$ and for $i > 1$, $Y_i = Y_{i-1} + \delta(Y_{i-1}) \circ X_i$ where $\circ$ here denotes coordinatewise multiplication. In other words, $2\delta_i(y)$ is the i^{th} sidelength of the largest box B_y that fits inside $[-1,1]^n$ and is centered at y, and Y_i takes a random step within $B_{Y_{i-1}}$ according to an appropriately rescaled sample from X.

It turns out that after a short time, rounding this random walk to $\{-1,1\}^n$ yields a good PRG.

Theorem 11.5 ([1])**.** *Suppose $\mathcal{F}$ is closed under restrictions and X is a symmetric p-noticeable fractional PRG for $\mathcal{F}$ with error ϵ and seed length s. Then for $t \in \mathcal{O}(\log(n/\epsilon)/p)$, $G = \text{sgn}(Y_t)$ is a PRG for $\mathcal{F}$ with error $(1+t)\epsilon$ and seed length ts.*

The proof is by two lemmas: first the *amplification lemma* shows that the random walk $Y_1, Y_2, \ldots$ quickly polarizes to a corner of the hypercube, and second the *rounding lemma* shows that rounding Y_t doesn't increase the PRG error too much.

Lemma 11.6 (Amplification)**.** *The t^{th} walk step Y_t is a $(1-q)$-noticeable fractional PRG with error $t\epsilon$ and $q = 2^{-\Omega(pt)}$.*

Proof. First we estimate the error of Y_t, which amounts the showing the random walk has small drift. The necessary claim is:

Claim. For all $f \in \mathcal{F}$ and $y \in [-1,1]^n$, $\big|f(y) - \mathbb{E}[f(y + \delta(y) \circ X)]\big| \leq \epsilon$.

This is shown using a carefully chosen random restriction. Let μ_f be the distribution over $g \in \mathcal{F}$ defined coordinatewise as follows. On input x, for each $i \in [n]$, replace x_i with $\operatorname{sgn}(y_i)$ with probability $|y_i|$, and otherwise leave it be. Because $\mathcal{F}$ is downward-closed, we have $g \in \mathcal{F}$ for all $g \in \operatorname{Supp}(\mu_f)$, and in particular $\mathbb{E}[g(\mathbf{0}) - g(X)] \le \epsilon$. But we also have for any x, $\mathbb{E}_{g\sim\mu_f}[g(x)] = f(y + \delta(y) \circ x)$ by multilinearity of f. Hence:

$$\begin{aligned} |f(y) - \mathbb{E}[f(y + \delta(y) \circ X)]| &= |\mathbb{E}_{g\sim\mu_f}[g(\mathbf{0})] - \mathbb{E}_{g\sim\mu_f}[g(X)]| \\ &= |\mathbb{E}_{g\sim\mu_f}[g(\mathbf{0}) - g(X)]| \qquad \le \epsilon. \end{aligned}$$

$\diamond$

An iterated triangle inequality then gives $|f(\mathbf{0}) - \mathbb{E}[f(Y_t)]| \le t\epsilon$.

Claim. $\mathbb{E}[Y_t^2] \ge 1 - 4\exp(-tp/4)$.

We'll show this coordinatewise. For any coordinate j, let x_i be the j^{th} coordinate of X_i, and likewise for y_i with Y_i. Then $y_i = y_{i-1} + (1 - |y_{i-1}|)x_i$. Let $z_i = 1 - |y_i|$. Then $z_i \in [0, 1]$ always; we'll find z_i decays to 0 exponentially quickly. We have the recurrence

$$\begin{aligned} z_i &= 1 - |y_{i-1} + z_{i-1}x_i| \\ &\le 1 - \operatorname{sgn}(y_{i-1})\operatorname{sgn}(y_{i-1} + z_{i-1}x_i)|y_{i-1} + z_{i-1}x_i| \\ &= 1 - \operatorname{sgn}(y_{i-1})(y_{i-1} + z_{i-1}x_i) \\ &= z_{i-1}(1 - x_i \operatorname{sgn}(y_{i-1})) \end{aligned}$$

The coordinate y_{i-1} is a symmetric random variable, so $\operatorname{sgn}(y_{i-1})$ is independent from $|y_{i-1}|$ and thus from z_{i-1} too. We may therefore estimate:

$$\begin{aligned} \mathbb{E}\left[\sqrt{z_i}\right] &\le \mathbb{E}\left[\sqrt{z_{i-1}}\right]\mathbb{E}\left[\sqrt{1 - x_i\operatorname{sgn}(y_{i-1})}\right] \\ &\le \mathbb{E}\left[\sqrt{z_{i-1}}\right]\mathbb{E}\left[1 - \frac{1}{2}\left(x_i\operatorname{sgn}(y_{i-1})\right)^2\right] \\ &= \mathbb{E}\left[\sqrt{z_{i-1}}\right]\left(1 - \frac{p}{2}\right) \\ &\le \mathbb{E}\left[\sqrt{z_{i-1}}\right]e^{-p/2} \end{aligned}$$

and so $\mathbb{E}[\sqrt{z_t}] \le \exp(-tp/2)$. Markov's inequality gives $\Pr[z_t \ge \exp(-tp/2)] \le \exp(-tp/4)]$, so we may bound

$$1 - \mathbb{E}[y_i^2] \le 2(\mathbb{E}[z_i]) \le 2(\exp(-tp/2) \cdot 1 + 1 \cdot \exp(-tp/4)) \le 4\exp(-tp/4).$$

Applying this argument to every coordinate completes the claim and the lemma. $\square$

Given a fractional PRG X with error ϵ, a natural way to turn it into a PRG is to take a sample $x \sim X$ and then form the unique distribution W_x over $\{-1, 1\}^n$ with $\mathbb{E}[W_x] = x$ and sample from that. By the multilinearity of f, we'd have

$$|\mathbb{E}[f(\mathcal{U}_n)] - \mathbb{E}_{x\sim X}\mathbb{E}[f(W_x)]| = |\mathbb{E}[f(\mathcal{U}_n)] - \mathbb{E}[f(X)]| \le \epsilon,$$

but this requires at least n bits of randomness. However, as long as X is very noticeable, $\operatorname{sgn}(x)$ is a good approximation for W_x.

Lemma 11.7 (Rounding)**.** *If Y is a $(1-q)$-noticeable fractional PRG for $\mathcal{F}$ with error ϵ, then $\operatorname{sgn}(Y) \in \{-1, 1\}^n$ is a PRG for $\mathcal{F}$ with error $\epsilon + qn$.*

Proof. We compute:

$$\begin{aligned} |f(\mathcal{U}_n) - \mathbb{E}[f(\operatorname{sgn}(Y))]| &\le |f(\mathbf{0}) - \mathbb{E}[f(Y)]| + |\mathbb{E}[f(Y)] - \mathbb{E}[f(\operatorname{sgn}(Y))] \\ &\le \epsilon + \sum_{i=1}^{n} 2\Pr[W_Y \ne \operatorname{sgn}(Y) \text{ on coord } i] \\ &\le \epsilon + n(1 - |Y_{(i)}|) \le \epsilon + nq. \end{aligned}$$

$\square$

11.3 Gaussians are fractional PRGs against $\mathcal{F}$ with $\mathcal{L}_{1,2}$ bounded

Define $\mathrm{trnc} : \mathbb{R} \to [-1, 1]$ by $\mathrm{trnc}(x) = \min(1, \max(-1, x))$.

Theorem 11.8 ([2, Thm. 9])**.** *Let G be an n-variate Gaussian with zero mean, at most σ^2 variance in each coordinate, and all covariances bounded by δ. Suppose $\mathcal{F}$ is a family closed under restrictions with $\mathcal{L}_{1,2}[\mathcal{F}] \leq t$. Then $\mathrm{trnc}(G)$ is a σ^2-noticeable fractional PRG for $\mathcal{F}$ with error $\epsilon \leq 4\delta t + 4n \exp(-\sigma^2/8)$.*

We give a streamlined sketch of the proof; a full version is worked out in [2, §A]. By standard properties of Gaussians, any $Z \sim G$ is distributed exactly as $\sum_{i=1}^{m} pZ_i$ for $Z_1, Z_2 \ldots \overset{\text{iid}}{\sim} G$ and $p = 1/\sqrt{m}$. This is an m-step random walk, and we find it fruitful to examine the hybrids $H_i = \sum_{j=1}^{i} pZ_j$. Each pZ_i has much smaller variance than G and known results about Brownian motion entail that with all-but-negligible probability, for all i, $H_i \in [-1/2, 1/2]^n$—a subcube we'll call the "good zone."

It turns out that when H_i is in the good zone, for a carefully chosen distribution of random restrictions ρ, we have the identity

$$f(H_{i+1}) - f(H_i) = \mathbb{E}_\rho[f_{\lceil\rho}(2pZ_{i+1}) - f_{\lceil\rho}(\mathbf{0})]$$

(Claim 19). Isserlis' Theorem on Gaussian moments then allows us to control the magnitude of the RHS of this expression in terms of $\mathcal{L}_{1,2}[f] = t$, giving

$$|f(H_{i+1}) - f(H_i)| \leq p^2\delta t + \mathcal{O}(p^4 n^4 \delta^4)$$

(Claims 18 & 20).

An iterated triangle inequality takes this bound on the i^{th} increment to a bound on the error for the whole walk, and after giving a crude upper bound on the PRG's performance when H_i leaves the good zone (and dealing with the trnc function in Claim 17), we find

$$|\mathbb{E}[f(\mathrm{trnc}(H_m))] - f(\mathbf{0})| \leq \sum_{k=1}^{\infty} e^{-km/(16n)} m n^k + 4\delta t + \mathcal{O}(n^4\delta^2/m) + 4ne^{-1/8\delta^2}.$$

Taking $m \to \infty$ gives the theorem.

Bibliography

[1] Chattopadhyay, E., Hatami, P., Hosseini, K. & Lovett, S. Pseudorandom Generators from Polarizing Random Walks. *Theory Of Computing.* **15**, 1-26 (2019), https://theoryofcomputing.org/articles/v015a010

[2] Chattopadhyay, E., Hatami, P., Lovett, S. & Tal, A. Pseudorandom Generators from the Second Fourier Level and Applications to AC0 with Parity Gates. *10th Innovations In Theoretical Computer Science Conference (ITCS 2019).* **124** pp. 22:1-22:15 (2018), http://drops.dagstuhl.de/opus/volltexte/2018/10115

[3] Raz, R. & Tal, A. Oracle Separation of BQP and PH. *Proceedings Of The 51st Annual ACM SIGACT Symposium On Theory Of Computing.* pp. 13-23 (2019), https://doi.org/10.1145/3313276.3316315

[4] Vadhan, S. Pseudorandomness. *Foundations And Trends® In Theoretical Computer Science.* **7**, 1-336 (2012), http://dx.doi.org/10.1561/0400000010

Joseph Slote, Caltech
email: **jslote@caltech.edu**

Chapter 12

Noise stability of functions with low influences: Invariance and optimality

after E. Mossel, R. O'Donnell and K. Oleszkiewicz [5]
A summary written by Stratos Tsoukanis

Abstract. This text is a summary of the the sections 1 2, and 3 up to theorem 3.18 of the work "Noise stability of functions with low influences: Invariance and optimality" by E. Mossel, R. O'Donnell and K. Oleszkiewicz. In these sections, the authors introduce some very interesting conjectures that can be proved using the invariance theorem 12.14 and they give a formal proof of that theorem.

12.1 Introduction

The paper I will present focuses mainly on Boolean functions of low influence. Boolean functions are functions $f : \{-1, 1\}^n \to \{-1, 1\}$ and they are usually used in theoretical computer science. we say that a Boolean function f has low "influence", if $\mathbb{E}[\mathrm{Var}_i[f]] << \mathrm{Var}[f]$, $\forall i \in [n]$. Functions of low influence are some of the fundamental tools in discrete Fourier analysis. In this paper, the authors show that under some mild conditions the distribution of multilinear polynomials with low influences and bounded degree is essentially invariant for all product spaces. In later chapters, two conjectures, "Majority is Stablest" and "It Ain't Over Till It's Over" are proved as consequences of the invariance principle, but I will focus mostly on the proof of the invariance principle.

12.1.1 Setup and notation

Let $(\Omega_1), \ldots, (\Omega_n)$ be a sequence of probability space, $\Omega = \Omega_1 \times \cdots \times \Omega_n$ and $f \in L^2(\Omega)$.

Definition 12.1. *The "influence" of the i-th component on f is defined by* $\mathrm{Inf}_i(f) = \mathbb{E}_\mu[\mathrm{Var}_{\mu_i}[f]]$

Definition 12.2. *For any $0 \leq \rho \leq 1$ we define the operator T_ρ, where*

$$(T_\rho f)(\Omega_1, \ldots, \Omega_n) = \mathbb{E}[f(\Omega'_1, \ldots, \Omega'_n)],$$

where Ω'_i are independent random variables equal Ω_i with probability ρ and randomly drawn from μ_i with probability $1 - \rho$.

Using the operator T_ρ now we are able to define the noise stability of a function f.

Definition 12.3. *The noise stability of a function f at $\rho \in [0, 1]$ is defined by* $\mathbb{S}\rho(f) = \mathbb{E}_\mu[f \cdot T_\rho f]$

Using the important notions of influence and noice stability we can state some important results now.

12.1.2 Important results involving low influences functions

Theorem 12.4 (Invariance Principle). *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = 1$ and $\mathbb{E}[X_i^3] \leq \beta$. Let Q be a degree d multilinear polynomial of the form $Q(X_1, \ldots, X_n) = \sum_{S \subset [n]} c_S \prod_{i \in S} X_i$, with*

$$\sum_{|S|>0} c_S^2 = 1 \quad \text{and} \quad \sum_{S \ni i} c_S^2 \leq \tau, \ \forall i \in [n].$$

Then

$$\sup_t |P[Q(X_1, \ldots, X_n) \leq t] - P[Q(G_1, \ldots, G_n) \leq t]| \leq O(d\beta^{1/3}\tau^{1/8d}),$$

where $G_1, \ldots, G_n$ are independent standard Gaussians.

If we have information for the upper bound β of $\mathbb{E}[X_i^q]$ for some $q \in (2, 3)$ instead of 3, then using a different approach we have can replace the upper bound of the previous theorem by $O(d\beta^{d/qd+1}\tau^{1/8d}\tau^{(q-2)/(2qd+2)})$.

Using this invariance theorem we can prove two famous conjectures from theoretical computer science and social choice theory.

The "Majority is Stablest" [4] conjecture, states that for $0 \leq \rho \leq 1$ and fixed $\epsilon > 0$, then there exists $t > 0$ such that if $f : \{0,1\}^n \to [-1,1]$ satisfies $\mathbb{E}[f] = 0$ and $\mathrm{Inf}_i(f) \leq \tau$, $\forall i$, then $\mathbb{S}_\rho(f) \leq (2/\pi)\arcsin\rho + \epsilon$.

The "It Ain't Over Till It's Over" [1] conjecture, states that for $0 \leq \rho \leq 1$ and fixed $\epsilon > 0$, there exists $\delta > 0$ and $\tau > 0$ such that if $f : \{0,1\}^n \to [-1,1]$ satisfies $\mathbb{E}[f] = 0$ and $\mathrm{Inf}_i(f) \leq \tau$, $\forall i$, then f has the following property. If V is a random subset of $[n]$ in which each i is included independently with probability ρ, and if the bits $(x_i)_{i \in V}$ are chosen uniformly at random, then

$$P_{V,(x_i)_{i \in V}}[|\mathbb{E}[f|(x_i)_{i \in V}]| > 1 - \delta] \leq \epsilon.$$

12.2 Applications of invariance theorem

12.2.1 Majority is Stablest

"Majority is Stablest" conjecture was first in 2007, but the motivation for getting sharp bounds on the noise stability of low-influence came in 2002 by two different papers. The first important result was stated by Kalai and it referred as "Arrow's Impossibility Theorem" [3] Suppose n voters rank three candidates, A, B, and C, and f is a social choice function which is applied to A vs B, A vs C, B vs C preferences to determine who is globally preferred by each pair. The theorem states that ice states that under some mild restrictions on f , the only functions that never admit nonrational outcomes given rational voters are the dictator function.

The second result was stated by Khot also in 2002 [2] and it was related to theoretical computer science. Constraint satisfaction problems (k-CSPs) are mathematical questions where a set of k objects whose state must satisfy a number of constraints or limitations, for example "Max-2Lin(2)") is the problem of finding a solution to an overconstrained system of linear equations modulo 2 in which each equation has exactly two variables. We say that a k-CSP has (c,s)-hardness" if the problem of finding an algorithm that satisfies an s-fraction of the constraints given that optimal assignment satisfies a c-fraction of the constraints is NP-hard.

Khot showed by using "Unique Games Conjecture" and a sharp inequality about low influence finctions from Bourgain-01.

Theorem 12.5. *If $f : \{-1,1\}^n \to \{-1,1\}$ satisfies $\mathbb{E}[f] = 0$ and $\mathrm{Inf}_i(f) \leq 10^{-d}$ for all $i \in [d]$, then*

$$\sum_{|S|>d} \hat{f}^2(S) \geq d^{-1/2-O(\sqrt{\log\log d/\log d})} = d^{-1/2-0(1)}$$

A corollary of Bourgain theorem is the following.

Corollary 12.6. *If $f : \{-1,1\}^n \to \{-1,1\}$ satisfies $\mathbb{E}[f] = 0$ and $\mathrm{Inf}_i(f) \leq 2^{-O(1/\epsilon)}$ for all $i \in [d]$, then*

$$\S_{1-\epsilon}(f) \leq 1 - \epsilon^{1/2+o(1)}$$

Using this result, Khot showed $(1-\epsilon, 1-\epsilon^{1/2+o(1)})$-hardnes for "Max-2Lin(2)" problem

12.2.2 Consquence of "Majority is Stablest"

A generalization of the "Majority is Stablest" is the following.

Theorem 12.7. *Let $f : \Omega_1 \times \cdots \times \Omega_n \to [0,1]$ be a function on a discrete product probability space and assume that for each i the minimum probability of any atom in Ω_i is at least $a \leq 1/2$. Further assume that $\mathrm{Inf}_i(f) \leq \tau$ for all i. Let $\mu = \mathbb{E}[f]$. Then for any $0 \leq \rho < 1$,*

$$\mathbb{S}_\rho(f) \leq \lim_{n\to\infty} \mathbb{S}_\rho(\mathrm{Thr}_n^{(\mu)}) + O(\log(\frac{\log\log(1/\tau)}{\log(1/\tau)}))$$

Where, $\mathrm{Thr}_n^{(\mu)} : \{-1,1\}^n \to \{0,1\}$ denotes the symmetric threshold function of the form $f(x_1,\ldots,x_n) = 1/2 + (1/2)\mathrm{Sgn}(\sum x_i - r)$ for $r \in \mathbb{R}$ and expectation closest to μ .

A consquence of this theorem is the followng result

Corollary 12.8. *"Max-2Lin(2)" has $(1-\epsilon, 1-O(\epsilon^{1/2}))$-hardness.*

12.2.3 "It Ain't Over Till It's Over"

A generalization of "It Ain't Over Till It's Over" conjecture is the following.

Theorem 12.9. *Let $0 < \rho < 1$, and let $f : \Omega_1 \times \cdots \times \Omega_n \to [0,1]$ be a function on a discrete product probability space; assume that for each i the minimum probability of any atom in Ω_i is at least $a \leq 1/2$. Then there exist $\epsilon(\rho,\mu) > 0$ such that if*

$$\epsilon < \epsilon(\rho,\mu) \text{ and } \mathrm{Inf}_i(f) \leq \epsilon^{O(1/\sqrt{\log(1/\epsilon)})} \ \forall i \text{ and } \mu = \mathbb{E}[f],$$

then,

$$P[V_\rho f > 1-\delta] \leq \epsilon \text{ and } P[V_\rho f > \delta] \leq \epsilon$$

provided $0 < \mu < 1$ and $\delta < \epsilon^{\rho/(1-\rho)+O(1/\sqrt{\log(1/\epsilon)}}$, where the $O(\cdot)$ hides a constant depending only on α, μ and ρ.

12.3 Proof of the main theorem

First lets introduce some definitions that will help us to prove the invariant theorem.

Definition 12.10. *We call a finite collection of orthonormal real random variables, one of which is the constant 1, an orthonormal ensemble.*

Definition 12.11. *Let $\mathcal{X}$ be a sequence of ensembles. For $1 \leq p \leq q < \infty$ and $0 < \eta < 1$, we say that $\mathcal{X}$ is (p,q,η)-hypercontractive if*

$$||(T_\eta Q)(\mathcal{X})||_q \leq ||\mathcal{X})||_q$$

for every multilinear polynomial Q.

Recall that random variable Y is (p,q,η)-hypercontractive for $1 \leq p \leq q < \infty$ and $0 < \eta < 1$ if

$$||\alpha + \eta Y||_q \leq ||\alpha + Y||_p \text{ for all } \alpha \in \mathbb{R}$$

Definition 12.12. *The d-low-degree influence of the i-th ensemble of $Q\mathcal{X}$ is*

$$\mathrm{Inf}_i^{(\leq d)}(Q\mathcal{X}) = \sum_{\substack{\sigma : ||\sigma|| \leq d \\ \sigma_i > 0}} c_\sigma^2$$

Definition 12.13. *Let $\mathcal{X}$ be a sequence of ensembles. For $1 \leq p \leq q < \infty$ and $0 < \eta < 1$, we say that $\mathcal{X}$ is (p,q,η)-hypercontractive if*

$$||(T_n Q)(\mathcal{X})||_q \leq ||Q(\mathcal{X})||_p$$

for every multilinear polynomial Q over $\mathcal{X}$

Now we are ready to state the main theorem. The invariance theorem as it is stated has 4 different variants which have similar hypotheses.

The following hypothesis is a generalization of the other so we will show the proof of the principle just for that.

Hypothesis. *Let $r \geq 3$ be an integer, and let $\mathcal{X}$ and $\mathcal{Y}$ be independent sequences of n ensembles that are $(2, r, \eta)$-hypercontractive. Assume also that for all $1 \leq i \leq n$ and for every sequence $(s_k)_{k=1}^{\infty}$ with $\sum_{k=1}^{\infty} s_K < r$ the sequences $\mathcal{X}$ and $\mathcal{Y}$ satisfy the condition*

$$\mathbb{E}[\prod_{k:s_k>0} X_{i,k}^{s_k}] = \mathbb{E}[\prod_{k:s_k>0} X_{i,k}^{s_k}]$$

Theorem 12.14. *Assume Hypothesis holds. Assume also that*$\mathrm{Var}[Q] \leq 1$, $\deg(Q) \leq d$ *and* $\mathrm{Inf}_i(Q) \leq t$ *for all i. Let $\Psi : \mathbb{R} \to \mathbb{R}$ with $|\Psi(r)| \leq B$ uniformly. Then $|\mathbb{E}[\psi(Q(\mathcal{X}))] - \mathbb{E}[\psi(Q(\mathcal{Y}))]| \leq \epsilon$, where $\epsilon = (2B/r!)d\eta^{-rd}\tau^{r/2-1}$*

Now lets see a summary of the proof. First we will show three propositions that we will need for the proof of main theorem.

Proposition 12.15. *Suppose Q is multilinear polynomial of the form $\sum_\sigma c_\sigma x_\sigma$ then*

$$\sum_i \mathrm{Inf}_i^{\leq d}(Q) \leq d\mathrm{Var}[Q]$$

Proposition 12.16. *Let $\mathcal{X}$ be a sequence of n_1 ensembles and $\mathcal{Y}$ an independent sequence of n_2 ensembles. Assume both are (p, q, η)-hypercontractive. Then the sequence of ensembles $\mathcal{X} \cup \mathcal{Y} = (\mathcal{X}_1, \ldots, \mathcal{X}_{n_1}, \mathcal{Y}_1, \ldots, \mathcal{Y}_{n_2})$, is also (p, q, η)-hypercontractive*

Proposition 12.17. *Suppose $\mathcal{X}$ is a $(2, q, \eta)$-hypercontractive sequence of ensembles and Q is a multilinear polynomial over $\mathcal{X}$ of degree d. Then*

$$||Q(\mathcal{X})||_q \leq \eta^{-d}||Q(\mathcal{X})||_2$$

Sketch of proof: Let $\mathcal{Z}^{(i)}$ be the sequence of ensembles $(\mathcal{X}_1, \ldots, \mathcal{X}_i, \mathcal{Y}_{i+1}, \ldots, \mathcal{Y}_n)$ and $Q^{(i)} = Q(\mathcal{Z}^{(i)})$. Let also

$\tilde{\mathbf{Q}} = \sum_{\sigma:\sigma_i=0} c_\sigma \mathcal{Z}_\sigma^{(i)}$, $\mathbf{R} = \sum_{\sigma:\sigma_i>0} c_\sigma X_{i,\sigma_i} \mathcal{Z}_{\sigma\setminus i}^{(i)}$, $\mathbf{S} = \sum_{\sigma:\sigma_i>0} c_\sigma Y_{i,\sigma_i} \mathcal{Z}_{\sigma\setminus i}^{(i)}$

If we prove that

$$|\mathbb{E}[\psi(\mathbf{Q}^{(i-1)}] - \mathbb{E}[\psi(\mathbf{Q}^{(i)}]| \leq \left(\frac{2B}{r!}\eta^{-rd}\right) \mathrm{Inf}_i(Q)^{r/2}.\forall i \in [n]$$

Then by using 12.15 and the fact that $\mathrm{Var}[Q] \leq 1$ our main theorem is proved.

From Taylor theorem we have that

$$|\mathbb{E}[\psi(\tilde{\mathbf{Q}} + \mathbf{R}) - \sum_{k=0}^{r-1} \mathbb{E}[\frac{\psi^{(k)}(\tilde{\mathbf{Q}})\mathbf{R}^k}{k!}]| \leq \frac{B}{r!}\mathbb{E}[|\mathbf{R}|^r] \tag{12.1}$$

$$|\mathbb{E}[\psi(\tilde{\mathbf{Q}} + \mathbf{S}) - \sum_{k=0}^{r-1} \mathbb{E}[\frac{\psi^{(k)}(\tilde{\mathbf{Q}})\mathbf{S}^k}{k!}]| \leq \frac{B}{r!}\mathbb{E}[|\mathbf{S}|^r] \tag{12.2}$$

Using the fact that $\mathcal{Z}_{\sigma\setminus i}^{(i)}$ and $\tilde{\mathbf{Q}}$ are independent we have that

$$\mathbb{E}[\psi^{(k)}(\tilde{\mathbf{Q}}\mathbf{R}^k) = \mathbb{E}[\psi^{(k)}(\tilde{\mathbf{Q}}\mathbf{S}^k) \tag{12.3}$$

So combining relations 12.1,12.2,12.3 we get

$$|\mathbb{E}[\psi(\tilde{\mathbf{Q}}+\mathbf{R})-\mathbb{E}[\psi(\tilde{\mathbf{Q}}+\mathbf{S})|\leq\frac{B}{r!}(\mathbb{E}[|\mathbf{R}|^r]+\mathbb{E}[|\mathbf{S}|^r]).$$

Now, by propositions 12.16 and 12.17 we have that

$$\mathbb{E}[|\mathbf{R}|^r]\leq\eta^{-rd}\mathbb{E}[\mathbf{R}^2]^{r/2}\text{ and }\mathbb{E}[|\mathbf{S}|^r]\leq\eta^{-rd}\mathbb{E}[\mathbf{S}^2]^{r/2}$$

But,

$$\mathbb{E}[\mathbf{S}^2]=\mathbb{E}[\mathbf{R}^2]=\mathrm{Inf}_i(Q)$$

.

So combining the previous relations we have that

$$|\mathbb{E}[\psi(\tilde{\mathbf{Q}}+\mathbf{R})-\psi(\tilde{\mathbf{Q}}+\mathbf{S})]|\leq\left(\frac{2B}{r!}\eta^{-rd}\right)\mathrm{Inf}_i(Q)^{r/2}$$

So the invariance theorem is proved,(Notice that $\mathbf{Q}^{(i-1)}=\tilde{\mathbf{Q}}+\mathbf{R}$ and $\mathbf{Q}^{(i)}=\tilde{\mathbf{Q}}+\mathbf{S}$). □

Bibliography

[1] G. Kalai, E. Friedgut *It Ain't Over Till It's Over*. personal communication, 2006.

[2] S. Khot, *T, On the power of unique 2-prover 1-round games* In: Proc. 34th Ann. ACM Symposium on Theory of Computing (Montreal, 2002), ACM, New York, 2002, pp. 767–775. MR 2121525

[3] G. Kalai, *A Fourier-theoretic perspective on the Condorcet paradox and Arrow's theorem*. In: Advances in Applied Mathematics 29.3 (Oct. 2002), pp.412–426.

[4] S. Khot, G. Kindler, E. Mossel and R. O'Donnell, *Optimal Inapproximability Results for Max-Cut and Other 2-Variable CSPs?* In: 45th Annual IEEE Symposium on Foundations of Computer Science. IEEE.

[5] E. Mossel, R. O'Donnell and K. Oleszkiewicz, *Noise stability of functions with low influences: Invariance and optimality*. In: Annals of Mathematics 171.1 (Mar. 2010), pp. 295–317.

Stratos Tsoukanis, UMD
email: etsoukan@umd.edu

Chapter 13

Majority is Stablest: Discrete and SoS

after A. De, E. Mossel and J. Neeman [3]
A summary written by Dimitris Vardakis

Abstract. This text is a summary of the the sections 1.1, 2, 3 and 4 of the work "Majority is Stablest: Discrete and SoS" by A. De, E. Mossel, and J. Neeman. In these sections, we will cover the "discrete proofs" of Borell's Inequality the and Majority-is-Stablest Theorem, which are done by induction on the dimension.

13.1 Introduction

The "Majority is Stablest" Theorem, proved in [6], is a positive answer to two conjectures, one in hardness of approximation [5] and one in social choice theory [4]. Its proof was based on Gaussian analysis and a form of "Invariance Principle", which was used to connect the Gaussian setting with the discrete nature of the theorem.

Since the "Majority is Stablest" Theorem concerns function on the Hamming cube $\{-1,1\}^n$, it is natural to ask whether there exists a purely "discrete proof". It turns out that it is possible to prove it without relying on any other elaborate machinery by induction on dimension. After all, many results about functions on the Hamming cube can be proven using induction thanks to their discrete nature.

13.1.1 Functions with low-influence variables

Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ play an important role in discrete Fourier analysis. Of particular interest are functions with low "influence". For $X \in \{-1,1\}^n$ let X^{-i} be

$$X^{-i} = (X_1, \ldots, X_{i-1}, -X_i, X_{i+1}, \ldots, X_n).$$

Then, the *i-th influence* of f is

$$\mathrm{Inf}_i f = \mathbb{P}[f(X) \neq f(X^{-i})]$$

where $X \sim \{-1,1\}^n$, that is where X is *uniformly distributed on the Hamming cube.*

Functions of low influence are important tools in hyper-contractive estimates and social choice theory. Two characteristic examples are the *dictator functions* f_j ($j \in [n]$) given by $f_j(x) = x_j$ with influence $\mathrm{Inf}_i f_j$ equal to 1 when $j = i$ and 0 when $j \neq i$, and the *majority function* $\mathrm{Maj}_n(x) = \mathrm{sgn}(\sum_{i=1}^n x_i)$ (for odd n) with $\mathrm{Inf}_i \mathrm{Maj}_n = O(n^{-1/2})$.

The "Majority is Stablest" Theorem states that the expectation of $f(x)f(y)$ cannot be much higher than the corresponding value for the majority function when x and y are "ρ-correlated". Formally, we have the following definitions are result:

Definition 13.1. *The vectors $x, y \in \{-1,1\}^n$ are* ρ-correlated *with $\rho \in [-1,1]$ when the vectors (x_i, y_i) are independent identically distributed random variables with $\mathbb{E}[x_i] = \mathbb{E}[y_i] = 0$ and $\mathbb{E}[x_i y_i] = \rho$. We write $x \sim_\rho y$.*

Definition 13.2. *The* noise stability *of $f : \{-1,1\}^n \to \mathbb{R}$ at $\rho \in (-1,1)$ is defined as*

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

Theorem 13.3 (Majority is Stablest [6])**.** *Let $\rho \in [0,1]$ and $\epsilon > 0$. Then, there exists $\tau > 0$ so that for any $f : \{-1,1\}^n \to [0,1]$ with $\mathbb{E}[f] = 1/2$ and $\max_i \mathrm{Inf}_i f \leq \tau$ it holds*

$$\mathrm{Stab}_\rho f \leq \left(1 - \frac{\arccos \rho}{\pi}\right) + \epsilon.$$

Observe also the (decreasing) limit

$$\lim_{n\to\infty} \mathrm{Stab}_\rho(\mathrm{Maj}_n) = 1 - \frac{\arccos \rho}{\pi},$$

which suggests that no low-influence function can be much more noise-stable than the majority function.

Here, by induction on dimension we will slightly generalise this theorem to functions of any expectation where the right-hand side is replaced by the corresponding quantity for the shifted majority of the same expectation. We will use hypercontractivity to bound certain error terms.

Additionally, we will present an independent proof of Borell's Inequality using similar techniques (but not hypercontractivity). Note that "Majority is Stablest" Theorem implies Borrel's result.

13.2 Tensorisation Theorem

First, we need to introduce the appropriate functions.

Let $\Phi : \mathbb{R} \to (0,1)$ be the cumulative distribution function of a standard normal variable. For every $\rho \in [-1,1]$, we define $J_\rho : (0,1)^2 \to [0,1]$ as

$$J_\rho(x,y) = \mathbb{P}[X \leq \Phi^{-1}(x), Y \leq \Phi^{-1}(y)]$$

where X, Y are jointly normally distributed random variables with covariance matrix

$$\mathrm{Cov}(X,Y) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Note that $J_\rho(\frac{1}{2},\frac{1}{2}) = \frac{1}{2} + \frac{\arcsin(\rho)}{\pi} = \lim_{n\to\infty} \mathrm{Stab}_\rho(\mathrm{Maj}_n)$. Thus, J_ρ is the right function to work with towards generalising the "Majority is Stablest" Theorem.

Ideally, we would like to have an inequality of the form

$$\mathbb{E}[J_\rho(f(X), g(Y))] \leq J_\rho(\mathbb{E}f, \mathbb{E}g)$$

but this won't be possible without the appearance of certain error terms.

Definition 13.4. *For $f : \{-1,1\}^n \to \mathbb{R}$, $S \subset [n]$ and $X \in \{-1,1\}^S$ we define $f_X : \{-1,1\}^{[n]\setminus S} \to \mathbb{R}$ the restriction of f on $\{X\}$.*

Claim.

$$\Delta_n(f) = \sum_{i=1}^{n} \mathbb{E}|f_i - f_{i-1}|^3$$

where $f_i(X_1, \ldots, X_i) = \mathbb{E}[f \mid X_1, \ldots, X_i]$.

The symbols Δ_n will appear in the error terms and measure the "Lipschitzness" of a function. It will be important that they are of 3rd order.

Theorem 13.6. *Let $\epsilon > 0$, $0 < \rho < 1$, and consider two ρ-correlated variables $X, Y \in \{-1,1\}^n$. Then, there exist constants $C, c > 0$ dependent only on ρ such that for any functions $f, g : \{-1,1\}^n \to [\epsilon, 1-\epsilon]$*

$$\mathbb{E}[J_\rho(f(X), g(Y))] \leq J_\rho(\mathbb{E}f, \mathbb{E}g) + C\epsilon^{-c}(\Delta_n(f) + \Delta_n(g)).$$

13.2.1 The base case

We will prove Theorem 13.6 by induction on n. Here we show the base case for $n = 1$:

Claim. *Let $\epsilon > 0$ and $0 < \rho < 1$. There exist $C_\rho, c_\rho > 0$ such that for any random ρ-correlated variables $X, Y \in [\epsilon, 1-\epsilon]$ it holds*

$$\mathbb{E}[J_\rho(X,Y)] \le J_\rho(\mathbb{E}X, \mathbb{E}Y) + C_\rho \epsilon^{-c_\rho}(\mathbb{E}|X - \mathbb{E}X|^3 + \mathbb{E}|Y - \mathbb{E}Y|^3).$$

The proof of this is based on the Taylor expansion of $J_\rho(x,y)$ along with two basic facts about its 2nd and 3rd derivatives:

Claim. *For any $(x,y) \in (0,1)^2$ and $0 < \rho < 1$ the matrix*

$$M_\rho(x,y) = \begin{pmatrix} \frac{\partial^2 J_\rho}{\partial x^2} & \rho \frac{\partial^2 J_\rho}{\partial x \partial y} \\ \rho \frac{\partial^2 J_\rho}{\partial x \partial y} & \frac{\partial^2 J_\rho}{\partial y^2} \end{pmatrix}(x,y)$$

is negative semi-definite.

Proof of Claim 13.7. Using the Taylor expansion of J_ρ around the point $(\mathbb{E}X, \mathbb{E}Y)$, we can write $\mathbb{E}J_\rho$ as

$$\mathbb{E}J_\rho(X,Y) = d_0 + d_1 + d_2 + d_3 + \dots.$$

Set $\tilde{x} = X - \mathbb{E}X$ and $\tilde{y} = Y - \mathbb{E}Y$. Then, $d_0 = J_\rho(\mathbb{E}X, \mathbb{E}Y)$, and

$$d_1 = \frac{\partial J_\rho}{\partial x}(\mathbb{E}X, \mathbb{E}Y)\,\mathbb{E}\tilde{x} + \frac{\partial J_\rho}{\partial y}(\mathbb{E}X, \mathbb{E}Y)\,\mathbb{E}\tilde{y} = 0.$$

The second order term is non-positive:

$$d_2 = \frac{1}{2}\mathbb{E}\left[(\tilde{x},\tilde{y}) \begin{pmatrix} \frac{\partial^2 J_\rho}{\partial x^2} & \frac{\partial^2 J_\rho}{\partial x \partial y} \\ \frac{\partial^2 J_\rho}{\partial x \partial y} & \frac{\partial^2 J_\rho}{\partial y^2} \end{pmatrix}\Big|_{(\mathbb{E}X, \mathbb{E}Y)} \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}\right] = \frac{1}{2}(\tilde{x},\tilde{x}) M_\rho(\mathbb{E}X, \mathbb{E}Y) \begin{pmatrix} \tilde{x} \\ \tilde{x} \end{pmatrix} \le 0.$$

Calculating the 3rd order derivatives of J_ρ we find that

$$d_3 \le C_\rho \epsilon^{-c_\rho}(\mathbb{E}\tilde{x} + \mathbb{E}\tilde{y}).$$

□

13.2.2 The inductive step

Now, we finish with Theorem 13.6.

Proof of Theorem 13.6. Assuming the theorem holds true for $n-1$, set $X' = (X_1 \dots, X_{n-1})$ and $Y' = (Y_1, \dots, Y_{n-1})$, and consider the functions $f, g : \{-1,1\}^n \to [\epsilon, 1-\epsilon]$. Applying our assumption to f_{X_n} and g_{Y_n}, we get

$$\mathbb{E}_{X',Y'}[J_\rho(f_{X_n}, g_{Y_n})] \le J_\rho(\mathbb{E}[f_{X_n} \mid X_n], \mathbb{E}[g_{Y_n} \mid Y_n]) + C_\rho \epsilon^{-c_\rho}(\Delta_{n-1}(f_{X_n}) + \Delta_{n-1}(g_{Y_n})).$$

Averaging over X_n and Y_n, and using Claim 13.7, we have

$$\begin{aligned}\mathbb{E}J_\rho(f,g) \le J_\rho(\mathbb{E}f, \mathbb{E}g) + C_\rho \epsilon^{-c_\rho}\Big(&\Delta_1(\mathbb{E}[f_{X_n} \mid X_n]) \\ &+ \Delta_1(\mathbb{E}[g_{Y_n} \mid Y_n]) + \mathbb{E}_{X_n}[\Delta_{n-1}(f_{X_n})] + \mathbb{E}_{Y_n}[\Delta_{n-1}(g_{Y_n})]\Big).\end{aligned}$$

The error terms in the parenthesis equal exactly $\Delta_n(f) + \Delta_n(g)$ by Claim 13.5 and the theorem is proved. □

13.3 Borell's Inequality

Theorem 13.9 (Borell's Inequality)**.** *Let $\rho \geq 0$ and consider two Gaussian vectors $G_1, G_2 \in \mathbb{R}^d$ with joint distribution*

$$(G_1, G_2) \sim N\left(0, \begin{pmatrix} I & \rho I \\ \rho I & I \end{pmatrix}\right).$$

For any $f_1, f_2 : \mathbb{R}^d \to [0,1]$ it holds that

$$\mathbb{E} J_\rho(f_1(G_1), f_2(G_2)) \leq J_\rho(\mathbb{E} f_1(G_1), \mathbb{E} f_2(G_2)).$$

Borell's Inequality follows trivially from the "Majority is Stablest" Theorem. Here we have an independent proof based on Theorem 13.6. This will entail the following crude estimate of Δ_n, which will need to be improved for the proof of "Majority is Stablest" Theorem.

Claim. *For $X \in \{-1,1\}^n$ and any $f : \mathbb{R}^n \to \mathbb{R}$ it holds that*

$$\Delta_n(f) \leq \frac{1}{8} \sum_{i=1}^{n} \mathbb{E}|f(X) - f(X^{-i})|^3.$$

Proof. Conditioning on X_n, this follows using induction and Jensen's inequality. □

Proof of Borell's Inequality. First, we write G_1, G_2 as the limits of the averages of independent random variables $X_i, Y_i \in \{-1,1\}$. All the limits exist thanks to central limit theorems.

Set $n = md$ and $X = (X_1, \ldots, X_n)$ so that

$$G_1^n := \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} X_i, \sum_{i=m+1}^{2m} X_i, \ldots, \sum_{i=(d-1)m+1}^{md} X_i \right) \xrightarrow{d} G_1,$$

and similarly for G_2 and Y.

Suppose f_1, f_2 are Lipschitz functions taking values in $[\epsilon, 1-\epsilon]$ and define g_1, g_2 so that

$$g_1(X) = f_1(G_1^n) \quad \text{and} \quad g_2(Y) = f_2(G_2^n).$$

Then, Theorem 13.6 gives

$$\mathbb{E} J_\rho(g_1(X), g_2(Y)) \leq J_\rho(\mathbb{E} g_1, \mathbb{E} g_2) + C_\rho \epsilon^{-c_\rho} (\Delta_n(g_1) + \Delta_n(g_2)). \tag{13.1}$$

Also, by Claim 13.10 and because f_1, f_2 are Lipschitz, $\Delta_n(g_j) = O(m^{-1/2})$; note n goes to infinity while $m \to \infty$. Therefore passing to the limit, (13.1) gives

$$\mathbb{E} J_\rho(f_1(G_1), f_2(G_2)) \leq J_\rho(\mathbb{E} f_1(G_1), \mathbb{E} f_2(G_2)).$$

This proves the theorem for Lipschitz functions with values in $[\epsilon, 1-\epsilon]$. But with this we can approximate any other functions $f_1, f_2 : \mathbb{R}^d \to [0,1]$ and the inequality holds as is. □

13.4 Majority is Stablest

To improve the bound of Claim 13.10 we will need to express $\Delta_n(f)$ in terms of the Fourier coefficients of f. Then, using the hypercontractivity theorem, the error terms will be small whenever the influences of f are small.

Fourier analysis

Consider $\{-1,1\}^n$ equipped with the uniform measure. For any $S \subset [n]$ the *character functions* are given by $\chi_S(x) = \prod_{i \in S} x_i$. Then, every function $f : \{-1,1\}^n \to \mathbb{R}$ can be written in the form

$$f(x) = \sum_{S \subset [n]} \hat{f}(S)\chi_S(x), \qquad \text{where } \hat{f}(S) = \mathbb{E}_{x \sim \{-1,1\}^n}[f(x)\chi_S(x)]$$

are the Fourier coefficients of f.

The i-th influence of f can be written as $\mathrm{Inf}_i f = \sum_{S \ni i} \hat{f}(S)^2$, and also $\mathrm{Var} f = \sum_{S \neq \emptyset} \hat{f}(S)^2$.

The *noise operator* T_ρ is defined by

$$T_\rho f(x) := \mathbb{E}_{y \sim_\rho x} f(y) = \sum_{S \subset [n]} \rho^{|S|} \hat{f}(S)\chi_S(x).$$

It also holds that $\mathrm{Stab}_\rho f = \langle f, T_\rho f \rangle$.

Next, we state some auxiliary yet important properties including Δ_n and f's and $T_\rho f$'s Fourier coefficients.

The first claim, the description of the error term $\Delta_n(f)$, is an easy case of induction:

Claim. *Let $S_i = \{i+1, \dots, n\}$. Then,*

$$\Delta_n(f) = \sum_{i=1}^{n} \mathbb{E}_{X \in \{-1,1\}^{S_i}} |\widehat{f_X}(i)|^3. \tag{13.2}$$

The next two claims follow directly from the definition and properties of the Fourier coefficients and $T_\rho f$. Those along with the Bonami-Beckner Hypercontractivity Theorem will give us a more precise bound from the one used to prove Borell's Inequality.

Claim. *For any disjoint $S, U \subset [n]$, and any $x \in \{-1,1\}^S$ and $i \in U$, we have*

$$\mathbb{E}_{X \in \{-1,1\}^S} |\widehat{f_X}(U)|^2 \leq \mathrm{Inf}_i f. \tag{13.3}$$

Moreover, if $S_i = \{i+1, \dots, n\}$, then

$$\sum_{i=1}^{n} \mathbb{E}_{X \in \{-1,1\}^{S_i}} |\widehat{f_X}(i)|^2 = \mathrm{Var} f. \tag{13.3'}$$

Claim. *Let $\sigma \in (0,1)$. For any disjoint $S, U \subset [n]$*

$$\widehat{(T_\sigma f)_x}(U) = \sigma^{|U|} T_\sigma(\widehat{f_x}(U)) \tag{13.4}$$

as polynomials in $x \in \{-1,1\}^S$.

Theorem 13.14 (Hypercontractivity [1, 2])**.** *Let $f : \{-1,1\}^n \to \mathbb{R}$ and $1 \leq q \leq p$. Then, for any $\rho \leq \sqrt{(q-1)/(p-1)}$*

$$\|T_\rho f\|_p \leq \|f\|_q. \tag{13.5}$$

Now, set $q = 2$ and $p = 1 + \sigma^{-2}$ for $0 < \sigma < 1$. Let $f : \{-1,1\}^n \to [-1,1]$ and consider $\widehat{f_x}(i)$ as a function of $x \in \{-1,1\}^{S_i}$ where $S_i = \{i+1, \dots, n\}$. Applying in order Claim 13.13, Theorem 13.14 and (13.3), we get

$$\mathbb{E}|\widehat{(T_\sigma f)_X}(i))|^p = \sigma^p \mathbb{E}|T_\sigma(\widehat{f_X}(i))|^p \leq (\mathbb{E}|\widehat{f_X}(i)|^2)^{p/2} \leq \mathrm{Inf}_i(f)^{\frac{p-2}{2}} \mathbb{E}|\widehat{f_X}(i)|^2$$

Summing over all $i \in [n]$, (13.3') gives us that

$$\sum_{i=1}^{n} \mathbb{E}_{X \in \{-1,1\}^{S_i}} |\widehat{(T_\sigma f)_X}(i)|^p \leq (\max_i \mathrm{Inf}_i f)^{\frac{p-2}{2}} \mathrm{Var} f.$$

The above along with (13.2) imply the following:

Claim. *If $p = 1 + \sigma^{-2} \leq 3$ it holds*

$$\Delta_n(T_\sigma f) \leq (\max_i \mathrm{Inf}_i f)^{\frac{1-\sigma^2}{2\sigma^2}}. \tag{13.6}$$

And now, we are ready to state and prove the "Majority is Stablest" Theorem:

Theorem 13.16 (Majority is Stablest)**.** *For any $\rho \in (0,1)$ there exists constant $C_\rho > 0$ (dependent only on ρ) such that for any function $f : \{-1,1\}^n \to [0,1]$ with $\max_i \mathrm{Inf}_i f \leq \tau$*

$$\mathrm{Stab}_\rho f \leq J_\rho(\mathbb{E}f, \mathbb{E}f) + C_\rho \frac{\log\log(1/\tau)}{\log(1/\tau)}.$$

Proof. First, we deal with functions $f : \{-1,1\}^n \to [\epsilon, 1-\epsilon]$. Towards this, consider X, Y to two ρ-correlated uniformly random variables on $\{-1,1\}^n$. If we apply Theorem 13.6 to $T_\sigma f$ for appropriate $\sigma \geq \sqrt{\rho}$ and Claim 13.15 (in this order), then for some exponent $\tilde{\sigma}$ we have

$$\mathbb{E}J_\rho(T_\sigma f(X), T_\sigma f(Y)) \leq J_\rho(\mathbb{E}[T_\sigma f], \mathbb{E}[T_\sigma f]) + C\epsilon^{-c}\Delta_n(T_\sigma f) \leq J_\rho(\mathbb{E}f, \mathbb{E}f) + C\epsilon^{-c}\tau^{\tilde{\sigma}},$$

where c and C depend only on ρ. Since $xy \leq J_\rho(x,y)$, we get

$$\mathrm{Stab}_{\rho\sigma^2} f = \mathbb{E}[T_\sigma f(X) T_\sigma f(Y)] \leq J_\rho(\mathbb{E}f, \mathbb{E}f) + C\epsilon^{-c}\tau^{\tilde{\sigma}}$$

and with appropriate relabelling

$$\mathrm{Stab}_\rho f \leq J_{\rho\sigma^{-2}}(\mathbb{E}f, \mathbb{E}f) + C\epsilon^{-c}\tau^{\tilde{\sigma}}. \tag{13.7}$$

In order to pass to functions $f : \{-1,1\}^n \to [0,1]$ let f^ϵ be the truncation of f to $[\epsilon, 1-\epsilon]$; it holds that $|f - f^\epsilon| \leq \epsilon$ and $\mathbb{E}|f - \mathbb{E}f^\epsilon| \leq \epsilon$. Lipschitz properties of $J_\rho(x,y)$ and elementary computations imply, through (13.7), that

$$\mathrm{Stab}_\rho f \leq \mathrm{Stab}_\rho f^\epsilon + 2\epsilon \leq J_{\rho\sigma^{-2}}(\mathbb{E}f, \mathbb{E}f) + 4\epsilon + C\epsilon^{-c}\tau^{\tilde{\sigma}}.$$

Next, notice that $J_{\rho\sigma^{-2}}(x,y) \leq J_\rho(x,y) + O_\rho(\tilde{\sigma})$, and by carefully choosing ϵ we have

$$\mathrm{Stab}_\rho f \leq J_\rho(\mathbb{E}f, \mathbb{E}f) + C(\tilde{\sigma} + \tau^{\tilde{\sigma}}).$$

Finally, when τ is small, it is possible to pick σ —and thus $\tilde{\sigma}$— so that

$$\tilde{\sigma} + \tau^{\tilde{\sigma}} \leq \frac{\log\log(1/\tau)}{\log(1/\tau)}$$

and the theorem is proved. □

Bibliography

[1] W. Beckner, *Inequalities in Fourier Analysis* In: The Annals of Mathematics 102.1 (July 1975), p 159.

[2] A. Bonami, *Étude des coefficients de Fourier des fonctions de $L^p(G)$*. In: Annales de l'institut Fourier 20.2 (1970), pp. 335–402.

[3] A. De, E. Mossel and J. Neeman, *Majority is stablest: Discreet and SoS* In: Proceedings of the 45th annual ACM symposium on Symposium on theory of computing - STOC-13 ACM Press, 2013.

[4] G. Kalai, *A Fourier-theoretic perspective on the Condorcet paradox and Arrow's theorem*. In: Advances in Applied Mathematics 29.3 (Oct. 2002), pp.412–426.

[5] S. Khot, G. Kindler, E. Mossel and R. O'Donnell, *Optimal Inapproximability Results for Max-Cut and Other 2-Variable CSPs?* In: 45th Annual IEEE Symposium on Foundations of Computer Science. IEEE.

[6] E. Mossel, R. O'Donnell and K. Oleszkiewicz, *Noise stability of functions with low influences: Invariance and optimality*. In: Annals of Mathematics 171.1 (Mar. 2010), pp. 295–341.

Dimitris Vardakis, MSU
email: **vardakis@msu.edu**

Chapter 14

Low Degree Learning and the Metric Entropy of Polynomials

after A. Eskenazis, P. Ivanisvili, and L. Streck [2]
A summary written by Thomas Winckelman

Abstract. We investigate how many data points are needed to recover a function on $H_n = \{\pm 1\}^n$. Since exact recovery is achieved by $|H_n| = 2^n$ data points, the paradigm is to achieve non-exponential dependence on n at the expense of assumptions on the function and "probably only epsilon." The term "metric entropy" refers to a general argument for deriving lower bounds based on covering/packing numbers. We survey results for approximating real-valued functions on H_n, and we outline the metric entropy argument.

14.1 Real-Valued Functions on the Hamming Cube

The set $H_n := \{\pm 1\}^n$ is a group with operation $(x_1, \dots, x_n)(y_1, \dots, y_n) = (x_1 y_1, \dots, x_n y_n)$ and identity $(1, \dots, 1)$. The characters of H_n are, precisely, the **Walsh functions**, defined $w_S(x) := \prod_{j \in S} x_j$ for each $S \subseteq [n]$. Therefore, given $g : H_n \to \mathbb{R}$, we use the Fourier notation $\widehat{g}(S) := \langle g, w_S \rangle_{L^2(H_n)}$ where $L^2(H_n)$ is always with respect to the uniform probability measure on H_n. We call the real values $\{\widehat{g}(S) : S \subseteq [n]\}$ the **Walsh coefficients** of g. For further discussion, see, for instance, section E of [4].

Obs. Every function $H_n \to \mathbb{R}$ extends *uniquely* to a **multi-linear polynomial on $\mathbb{R}^n$ with real coefficients**, that is, a function of the form $p(x) = \sum_{S \subseteq [n]} a_S \prod_{j \in S} x_j$. Indeed, take a_S to the Walsh coefficients.

Def. Given a function $f : H_n \to \mathbb{R}$, a collection $\mathcal{S}$ of subsets of $[n]$, and $\eta \geq 0$, we define the statement that f **is η-concentrated on $\mathcal{S}$** to mean

$$\sum_{S \notin \mathcal{S}} \left|\widehat{f}(S)\right|^2 \quad \leq \quad \eta.$$

Remark: In the extreme case, a function f which depends on only k variables is 0-concentrated on the collection of sets $\{S \subseteq \sigma\}$, where $\sigma \subseteq [n]$ is the set of indices of the variables on which f depends.

Convention. We let $\mathcal{F}_{n,d}(t)$ denote the functions $f : H_n \to \mathbb{R}$ which are t-concentrated on the collection of sets $\{|S| \leq d\}$. We call $\mathcal{F}_{n,d} := \mathcal{F}_{n,d}(0)$, and we refer to functions $f \in \mathcal{F}_{n,d}$ as having **degree at most** d.

14.2 Select Concepts and Results from Learning Theory

Since functions $f : H_n \to \mathbb{R}$ correspond to polynomials, results on polynomial interpolation are applicable. In general, however, the multi-variate case is difficult, which is why we focus on the more specific setting of functions $f : H_n \to \mathbb{R}$. To begin, we present some baseline observations for context.

Thm (Exact Deterministic Learning I; [2]). A function $f \in \mathcal{F}_{n,d}$ is completely determined by the points $x \in H_n$ with at most d negative coordinates, of which there are $Q := Q_{n,d} := \sum_{k \leq d} \binom{n}{k} \in \big[(n/d)^d, (en/d)^d\big]$. In fact, for a fixed enumeration $\{x^{(1)}, \dots, x^{(Q)}\}$ of these points, there is an honest-to-goodness formula for the function which maps each tuple $(a_1, ..., a_Q) \in \mathbb{R}^Q$ to the unique $f \in \mathcal{F}_{n,d}$ that satisfies $f(x^{(k)}) = a_k$ for all $k \in [Q]$.

Def (Query Complexity). Given $\mathcal{F} \subseteq L^2(H_n)$ and $\varepsilon \geq 0$, we write $Q(\mathcal{F}, \varepsilon)$ to denote the smallest positive integer Q such that there exists a function $H : (H_n \times \mathbb{R})^Q \to L_2(H_n)$ along with a point $x^{(1)} \in H_n$ and functions $\varphi_1, \dots, \varphi_{Q-1}$ each mapping $\varphi_q : (H_n \times \mathbb{R})^q \to H_n$ such that, for any given $f \in \mathcal{F}$, if $x_{q+1} = \varphi_q[(x^{(1)}, f(x^{(1)})), \dots, (x^{(q)}, f(x^{(q)}))]$ for every $q < Q$, then

$$\Big\| f \; - \; H\big[(x^{(1)}, f(x^{(1)})), \dots, (x^{(Q)}, f(x^{(Q)}))\big] \Big\|^2_{L^2(H_n)} \quad \leq \quad \varepsilon.$$

Remark: The intention is that $x^{(q+1)}$ is allowed to depend on $x^{(1)}, \dots, x^{(q)}$ and on $f(x^{(1)}), \dots, f(x^{(q)})$, but not on anything more. The function φ_q merely serves to formalize this.

Remark: Such an integer Q always exists, and is at most 2^n.

Remark: The function $\varepsilon \mapsto Q(\mathcal{F}, \varepsilon^2)$ is analogous the *inverse* function of what is called *m-th minimal adaptive intrinsic error* in optimal recovery. However, in optimal recovery, more diverse measurements are typically allowed, not merely point evaluations.

Thm (Exact Deterministic Learning II; [2]). While the above theorem says that $Q(\mathcal{F}_{n,d}, 0) \leq \sum_{k \leq d} \binom{n}{k}$, it is in fact true that $Q(\mathcal{F}_{n,d}, 0) = \sum_{k \leq d} \binom{n}{k}$.

Remark: The statement "$Q(\mathcal{F}_{n,d}, 0) \leq \sum_{k \leq d} \binom{n}{k}$," by itself, does not tell the complete story. For instance, it actually did not matter in original theorem the order in which the data points are given.

Def (Randomized Query Complexity). Given $\mathcal{F} \subseteq L^2(H_n)$ and $\varepsilon \geq 0$ and $\delta \in [0, 1]$, we write $Q_r(\mathcal{F}, \varepsilon, \delta)$ to denote the infimum of the set of integers $Q > 0$ for which there exists a (Borel) function $H : (H_n \times \mathbb{R})^Q \to L_2(H_n)$ such that, for any $f \in \mathcal{F}$, if $X_1, \dots, X_Q$ are drawn uniformly IID from H_n,

$$P\Big(\big\| f \; - \; H\big[(X_1, f(X_1)), \dots, (X_Q, f(X_Q))\big] \big\|^2_{L^2(H_n)} \leq \varepsilon \Big) \quad \geq \quad 1 - \delta.$$

Remark: As a sanity check, the assumption that H is Borel ensures that the thing of which we're taking the probability is, indeed, a random variable.

Remark: This is very analogous to the notion of *sample complexity* in statistical learning theory. However, in statistical learning theory, more diverse probability distributions are typically allowed, not merely uniform.

Obs. Given $\mathcal{B} \subseteq \mathcal{F}$ and $0 \leq \varepsilon \leq E$ and $0 \leq \delta \leq \Delta \leq 1$, we have $Q_r(\mathcal{B}, E, \Delta) \leq Q_r(\mathcal{F}, \varepsilon, \delta)$ and $Q(\mathcal{B}, E) \leq Q(\mathcal{F}, \varepsilon)$.

14.3 Upper Bounds

An intuitive means of attempting to reconstruct a function $g : H_n \to \mathbb{C}$ based only on a finite data set $\mathcal{D} = \{(X_1, Y_1), \dots, (X_Q, Y_Q)\} \subseteq (H_n \times \mathbb{R})^Q$ is to estimate each of g's Walsh coefficients by taking the empirical counterparts

$$\alpha_S(\mathcal{D}) \quad := \quad \frac{1}{Q} \sum_{j=1}^{Q} Y_j w_S(X_j).$$

In fact, the below two theorems are each proven *constructively*, by exhibiting a quite practical algorithm based on this notion of coefficient estimation.

Obs (Uniform Distribution is Special). If X_j has the uniform distribution and $g(X_j)$ is a version of $\mathbb{E}(Y_j \mid X_j)$ for every j, then $\alpha_S(\mathcal{D})$ is an unbiased estimator of $\widehat{g}(S)$. This unbiasedness is used in order to apply concentration inequalities, thus explaining the role of the uniform distribution.

Remark: Since $\mathbb{E}(Y_j \mid X_j) = g(X_j)$ is the key property, not $Y_j = g(X_j)$, the algorithms implicit in the following proofs are even somewhat robust to noisy measurements.

Thm (Random Learning; [1]). Call $U := \{f \in L^2(H_n) : \|f\|_{L^2(H_n)} \leq 1\}$. There is $C > 0$ such that, for all $\varepsilon, \delta > 0$ and $d \in [n]$, we have

$$Q_r\big(U \cap \mathcal{F}_{n,d}, \varepsilon, \delta\big) \quad \leq \quad \min\left\{\frac{\exp\big(Cd^{1.5}\sqrt{\ln(d)}\big)}{\varepsilon^{d+1}},\ \frac{4dn^d}{\varepsilon}\right\}\ln(n/\delta).$$

Thm (Robust Random Learning; [2]). Let $R, \eta, t \geq 0$ and $m \in [n]$. Let $\mathcal{L}$ be a non-empty collection of subsets of $[n]$. Let $\mathcal{F}$ be a collection of functions $H_n \to \mathbb{R}$ such that each $f \in \mathcal{F}$ is t-concentrated on $\mathcal{L}$, satisfies $\|f\|_{H^2(H_n)} \leq R$, and is η-concentrated on some (unknown) collection $\mathcal{S}$ of subsets of $[n]$ for which $|\mathcal{S}| \leq m$. Then, for every $\varepsilon, \delta > 0$, we have

$$Q_r\big(\mathcal{F}, \min\{R^2, \eta + t + \varepsilon\}, \delta\big) \quad \leq \quad \left\lceil 18R^2\frac{m}{\varepsilon}\ln(2|\mathcal{L}|/\delta)\right\rceil.$$

Remark: Even in the extreme case $\mathcal{L} = 2^{[n]}$, our requirement on Q is still only $Q \gtrsim R^2\frac{mn}{\varepsilon}\ln(2/\delta)$ where m is the smallest cardinality such that every $f \in \mathcal{F}$ is η-concentrated on asomecollection of that cardinality. Thus, the dependence on n is only *linear*, though scales like the number m of "non-negligible" Walsh coefficients of functions in the class $\mathcal{F}$.

Remark: The parameters $\mathcal{L}$ and δ are chosen by the user, however, crucially, $\mathcal{S}$ is not. The idea is that $\mathcal{L}$ is overly conservative in the sense that, even while g is t-concentrated on $\mathcal{L}$, there might still be a much smaller (unknown) set $\mathcal{S}$ on which g is η-concentrated. Logically, "$\exists\mathcal{L} : \forall f : \exists\mathcal{S}$."

Cor (Boolean Case; [2]). Using deep structural results regarding the existence of collections on which degree d and/or Boolean functions are concentrated, we can extract many further estimates, For instance,

$$Q_r\big(\mathcal{B}_{n,d}, \varepsilon, \delta\big) \quad \leq \quad 36\,\frac{d\,2^{d^2}}{\varepsilon}\,\ln(n/\delta)$$

where $\mathcal{B}_{n,d}$ denotes the ± 1-valued functions on H_n of degree at most d.

14.4 Lower Bounds

A reasonable notion of *query complexity* can be defined in numerous different ways, and variations of this definition appear throughout applied mathematics, as already remarked. What unifies these diverse concepts is the notion of a "best possible worst case scenario." Indeed, in our definitions, idealized algorithms are allowed, as are the most pathaological f which $\mathcal{F}$ has to offer. Subjectively, this can make upper bounds on query complexity more "impressive" than lower bounds, especially if the upper bound is realized through a concrete algorithm, as has been the case for all the estimates presented thus far, though the opaque notation "$Q_r(\mathcal{F}, \varepsilon, \delta)$" fails to reflect this.

Lower bounds are most convincing when the cause for failure is some *non-pathological* f. This can be achieved by lower-bounding $Q_r(\mathcal{B}, \varepsilon, \delta)$ of a class $\mathcal{B}$ which does not contain any pathologies. Indeed, we will derive lower bounds on query complexity for the relatively "nice" classes of Boolean functions. Since errors are measured as metric distances, *packing numbers* emerge as a useful tool for estimating complexity. This principle is difficult to articulate in general, yet widespread (for a recent example in the context of shallow ReLU networks, see [3]). Here is a specific instance of the principle.

Prop (Metric Entropy Bound Argument; [2]). Let $\mathcal{B}$ be *any* collection of functions $H_n \to \{\pm 1\}$. For all $\varepsilon, \delta > 0$, the packing numbers satisfy

$$\begin{aligned} Q(\mathcal{B}, \varepsilon) &\geq \log_2\big(M\big(\mathcal{B}, \|\cdot\|_{L^2(H_n)}, 2\sqrt{\varepsilon}\big)\big), \qquad \text{AND} \\ Q_r(\mathcal{B}, \varepsilon, \delta) &\geq \log_2\big(M\big(\mathcal{B}, \|\cdot\|_{L^2(H_n)}, 2\sqrt{\varepsilon}\big)\big) + \log_2(1-\delta). \end{aligned}$$

Remark: As is usually the case with such terse notation, in fact, much more can be said. For instance, the latter bound is still true even if the variables $X_1, \ldots, X_Q$ are allowed to have *any* joint distribution.

Remark: Sharp bounds on the packing number often boil down to combinatorial bottlenecks, the study of which is, basically, "discrepancy theory."

Obs. The set $\mathcal{W}_{n,d} := \{w_S : |S| \leq d\}$ is *discrete*, by orthogonality, with $\|f - g\|_{L^2(H_n)} = \sqrt{2}$ for any distinct $f, g \in \mathcal{W}_{n,d}$. Therefore, whenever $\varepsilon < \sqrt{2}$, we have $M\big(\mathcal{W}_{n,d}, \|\cdot\|_{L^2(H_n)}, \varepsilon\big) = |\mathcal{W}_{n,d}| = \sum_{k \leq d} \binom{n}{k} \geq (n/d)^d$.

Cor. Our bound $Q_r(U \cap \mathcal{F}_{n,d}, \varepsilon, \delta) \leq \exp\big(Cd^{1.5}\sqrt{\ln(d)}\big)\ln(n/\delta)/\varepsilon^{d+1}$ is "asymptotically sharp in n" in the sense that, fixing $\varepsilon, \delta, d > 0$, if $\varepsilon < 1/\sqrt{2}$,

$$\limsup_{n\to\infty} \frac{\exp\big(Cd^{1.5}\sqrt{\ln(d)}\big)\ln(n/\delta)/\varepsilon^{d+1}}{Q_r(U \cap \mathcal{F}_{n,d}, \varepsilon, \delta)} \quad \leq \quad \frac{\exp\big(Cd^{1.5}\sqrt{\ln(d)}\big)}{d\,\varepsilon^{d+1}} \quad < \quad \infty.$$

Indeed, $U \cap \mathcal{F}_{n,d} \supseteq \mathcal{W}_{n,d}$, so that $Q_r(U \cap \mathcal{F}_{n,d}, \varepsilon, \delta) \geq d\log_2(n/d) + \log_2(1-\delta)$, but $\log_2(x/\delta)/\log_2(x/d) \to 1$ as $x \to \infty$, by L' Hôspital's rule.

Bibliography

[1] Alexandros Eskenazis, Paata Ivanisvili (2022). Learning Low-Degree Functions from a Logarithmic Number of Random Queries. To appear in *Proceedings of STOC, 2022*. Preprint available at https://arxiv.org/abs/2109.10162

[2] Alexandros Eskenazis, Paata Ivanisvili, Lauritz Streck (2022). Low-degree Learning and the Metric Entropy of Polynomials. Preprint available at https://arxiv.org/abs/2203.09659

[3] Jonathan Siegel, Jinchao Xu (2021). Sharp Bounds on the Approximation Rates, Metric Entropy, and n-Widths of Shallow Neural Networks. Preprint available at https://arxiv.org/abs/2101.12365

[4] My senior thesis https://www.overleaf.com/read/vdxpkcvzsgqk

Thomas Winckelman, Texas A&M
email: **winckelman@tamu.edu**

Chapter 15

On the Gaussian noise sensitivity and Gaussian surface area of polynomial threshold functions

after D. Kane [1]
A summary written by Qiang Wu

Abstract. In this note, we summarize the main results and its proof ideas in [1], where Kane proved sharp results of Gaussian noise sensitivity for general degree-d polynomial threshold functions. Furthermore, this result was used to establish some sharp results about the Gaussian surface area.

15.1 Introduction

Polynomial threshold functions are a class of binary-valued functions associated with some polynomial. It plays an important role in several different subjects, such as in learning theory, theoretical computer science etc. To understand the noise sensitivity of those functions is a fundamental problem. To begin with, let us present the precise definitions of the related concepts first.

Definition 15.1 (Polynomial threshold functions)**.** *A given function $f : \mathbb{R}^N \to \{-1, +1\}$ is called a polynomial threshold function if*

$$f(x) = sign(p(x))$$

for some polynomial $p(x)$. If the degree of associated $p(x)$ is at most d, we call $f(x)$ is a degree-d polynomial threshold function.

If one restricts the domain from $\mathbb{R}^N$ to $\{-1, +1\}^N$, PTFs become particular standard boolean functions. For general boolean functions, an important question is about its noise sensitivity. Specifically, the question is asking how stable of the function's value under a small perturbation of the argument x. In this article, we will instead focus on the notion of Gaussian noise sensitivity to deal with continuous inputs. The special properties of Gaussian also enable us to exploit the symmetry and thus can obtain some sharp results. Here is a formal definition of the Gaussian noise sensitivity.

Definition 15.2 (Gaussian noise sensitivity)**.** *Given noise rate $\epsilon > 0$, for a PTF f, the Gaussian noise sensitivity is*

$$GNS_\epsilon(f) := \mathbb{P}(f(X) \neq f(Z)), \tag{15.1}$$

where $Z := (1 - \epsilon) \cdot X + \sqrt{2\epsilon - \epsilon^2}Y$, and X, Y are independent N-dimensional Gaussian vectors.

One of the main results in the paper [1] is a sharp asymptotic bound on the Gaussian noise sensitivity of PTFs. It turns out that the noise sensitivity has some intimate connections with the Gaussian surface area. Heuristically, noise sensitivity characterizes the probability of X near the boundary, where a small perturbation will push the function over the boundary. This in particular is related the the area of the boundary surface.

In general, the Gaussian surface area of a set $A \subseteq \mathbb{R}^N$ is defined as

$$\Gamma(A) := \liminf_{\delta \to 0} \frac{1}{\delta} \cdot \mathbb{P}(X \in A_\delta \setminus A) \quad \text{for } X \text{ Gaussian vector}, \tag{15.2}$$

where the set $A_\delta := \{x \in \mathbb{R}^N : d(x, A) \leq \delta\}$ is the augmented set of A within distance δ. In particular, for PTF f, the Gaussian surface area is

$$\Gamma(f) := \Gamma(f^{-1}(1)).$$

15.2 Main results

The first main result is about the Gaussian noise sensitivity.

Theorem 15.3 ([1]-Theorem 1.1)**.** *If f is a degree-d PTF, then*

$$GNS_\epsilon(f) \leq \frac{d \arcsin(\sqrt{2\epsilon - \epsilon^2})}{\pi} \sim \frac{d\sqrt{2\epsilon}}{\pi} = O(d\sqrt{\epsilon}). \tag{15.3}$$

Furthermore the above bound is asymptotically tight as $\epsilon \to 0$ for threshold functions of any square free product of homogeneous linear functions.

Due to the intimate connection of Gaussian surface area and Gaussian noise sensitivity, similar tight bound for surface area can also be obtained.

Theorem 15.4 ([1]-Theorem 1.2)**.** *If f is a degree-d PTF, then*

$$\Gamma(f) \leq \frac{d}{\sqrt{2\pi}}. \tag{15.4}$$

Furthermore the above bound is optimal for threshold functions of any square free product of homogeneous linear functions.

Before discussing the proof details, let us briefly mention one notable conjecture by Gotsman and Linial [2] about the average noise sensitivity of PTFs.

15.3 Connections to Gotsman-Linial conjecture

Consider PTFs on the hypercube $\{-1, +1\}^N$, we introduced the noise sensitivity of PTFs, the average noise sensitivity can be heuristically defined as the expected number of bits flipped that will change the function's value. Specifically, Gotman-Linial conjecture says that the average noise sensitivity of degree-d PTFs is maximized by product of linear threshold functions cutting the middle d layers of the hypercube. This original statement was recently refuted in [3]. However, a weaker form about the asymptotic bound is expected to be true and still open. For more details about this conjecture, see [3, 4] and references therein.

15.4 Proof of Theorem 15.3

The idea of the proof heavily depends on the Gaussianity assumption, in particular, the key step using symmetrization to simplify the problem is based on the rotational invariance property of Gaussian distribution. We start with reformulating the original problem.

Step 1: Rewrite

$$\mathrm{GNS}_\epsilon(f) = \mathbb{P}(f(X) \neq f(Z)) = \mathbb{P}(f(X) \neq f(\cos(\theta)X + \sin(\theta)Y)),$$

where $\theta = \arcsin(\sqrt{2\epsilon - \epsilon^2})$. For notation convenience, let $X_\theta := \cos(\theta)X + \sin(\theta)Y$.

Step 2: Instead of directly computing the probability over all X, Y, one can utilize the rotational invariance property of Gaussian. This implies for any fixed $\phi \in [0, 2\pi]$,

$$\mathrm{GNS}_\epsilon(f) = \mathbb{P}(f(X_\phi) \neq f(X_{\phi+\theta})).$$

One way to look at this property is selecting the independent Gaussians as X_ϕ and $X_{\phi+\pi/2}$, then it's easy to verify $X_{\phi+\theta} = \cos(\theta)X_\phi + \sin(\theta)X_{\phi+\pi/2}$. Therefore

$$\mathrm{GNS}_\epsilon(f) = \frac{1}{2\pi}\int_0^{2\pi} \mathbb{P}(f(X_\phi) \neq f(X_{\phi+\theta}))d\phi.$$

Step 3: One observation is that in order to evaluate the above integral, one needs to understand the event $f(X_\phi) \neq f(X_{\phi+\theta})$. Recall f is binary-valued, then the problem reduces to counting the number of sign changes of f in $[\phi, \phi + \theta]$. therefore

$$\mathrm{GNS}_\epsilon(f) = \frac{1}{2\pi}\mathbb{E}_{X,Y}\int_0^{2\pi} 1_{f(X_\phi)\neq f(X_{\phi+\theta})}d\phi \leq \frac{\theta}{2\pi}\mathbb{E}_{X,Y}[W],$$

where W is a random variable representing the number of sign changes of $f(X_\phi)$ for $\phi \in [0, 2\pi]$.

Step 4: Since f is a degree-d PTF, that is, there exists a degree-d polynomial $p(x)$ such that $f(x) = sign(p(x))$. Notice that the number of zeros of $p(x)$ relates to the number of sign changes. For $p(\cos(\phi)X + \sin(\phi)Y)$, this can be treated as the intersections of $p(ax + by) = 0$ and $a^2 + b^2 = 1$, by Bezout's Theorem [1], there can be at most $2d$ solutions. Thus

$$\mathrm{GNS}_\epsilon(f) \leq \frac{d\theta}{\pi}.$$

A final remark is in some cases, the number of sign changes can be less than $2d$, then the bound will not be sharp for those functions. But for functions as product of d homogeneous linear functions, the bound is asymptotic sharp.

15.5 Proof of the Gaussian surface area

The following key lemma basically relates the Gaussian noise sensitivity and Gaussian surface area.

Lemma 15.5 ([1]-Lemma 3.1)**.** *For PTF f, and X, Y are independent Gaussians, then*

$$\lim_{\epsilon\to 0}\frac{\mathbb{P}(f(X) = -1, f(X + \epsilon Y) = 1)}{\epsilon} = \frac{\Gamma(f)}{\sqrt{2\pi}}$$

The left hand side is some sort of noise sensitivity but not exactly. The right hand side is exactly the Gaussian surface area. It needs to further formalize the relation of LHS and the Gaussian noise sensitivity. Before that, we roughly sketch the idea of the proof for Lemma 15.5.

Suppose X is in distance t from the boundary, it is expected that the probability is roughly $\Gamma(f)dt$. To compute the LHS probability, one needs the size of ϵY's projection onto the normal direction to the boundary is larger than t, otherwise $X + \epsilon Y$ can not be pushed over the boundary. Thus the LHS probability is

$$\int_{\epsilon y > t > 0} \phi(y)\Gamma(f)dtdy = \int_0^\infty \epsilon\Gamma(f)y\phi(y)dy = \frac{\Gamma(f)\epsilon}{\sqrt{2\pi}},$$

where $\phi(y)$ is the Gaussian density function.

[1] Bezout's theorem is a classical result in algebraic geometry, which asserts that for two nonzero polynomials P, Q with no common factor, the number of intersections of $P(x, y) = 0$ and $Q(x, y) = 0$ can be at most $\mathrm{degree}(P) \cdot \mathrm{degree}(Q)$.

15.6 Proof of Theorem 15.4

The proof can be roughly divided into following steps.

Step 1: First it is clear to notice that for a given PTF f, $\Gamma(f) = \Gamma(-f)$. Since in any case, it measures the boundary between $f^{-1}(-1)$ and $f^{-1}(1)$.

Step 2: To bound the LHS probability $\mathbb{P}(f(X) \neq f(X + \epsilon Y))$, we are seeking for a connection to the standard Gaussian noise sensitivity. Recall

$$\mathrm{GNS}_\epsilon(f) = \mathbb{P}(f(X) \neq f(Z)) \text{ for } Z = \cos(\theta)X + \sin(\theta)Y.$$

Taking $\theta = \arctan(\epsilon)$, then $X + \epsilon Y = rZ$ with $r = \sqrt{1+\epsilon^2}$. Finally

$$\mathbb{P}(f(X) \neq f(X + \epsilon Y)) \leq \mathbb{P}(f(X) \neq f(Z)) + \mathbb{P}(f(Z) \neq f(rZ))$$

The first term can be bounded using Gaussian noise sensitivity, and the second term also can be controlled using the Gaussian symmetry. We skip the details here. Eventually, the bound is

$$\mathbb{P}(f(X) \neq f(X + \epsilon Y)) \leq \frac{d\epsilon}{\pi} + \frac{d\epsilon^2}{4}\sqrt{\frac{n}{\pi}}.$$

Step 3: Finally, collecting all the previous results,

$$\begin{aligned}
\Gamma(f) &= \frac{1}{2}\left(\Gamma(f) + \Gamma(-f)\right) \\
&= \sqrt{\frac{\pi}{2}} \cdot \lim_{\epsilon \to 0} \frac{\mathbb{P}(f(X) = -1, f(X + \epsilon Y) = 1) + \mathbb{P}(f(X) = 1, f(X + \epsilon Y) = -1)}{\epsilon} \\
&= \sqrt{\frac{\pi}{2}} \cdot \lim_{\epsilon \to 0} \frac{\mathbb{P}(f(X) \neq (X + \epsilon Y))}{\epsilon} \leq \frac{d}{\sqrt{2\pi}}
\end{aligned}$$

where the first equality utilized the symmetry property in Step 1. The second step is based on the identity in Lemma 15.5, finally the conclusion in Step 2 gives the desired bound.

Bibliography

[1] Kane, Daniel M., *The Gaussian surface area and noise sensitivity of degree-D polynomial threshold functions.* Comput. Complexity, 20, 389-412 (2011).

[2] Gotsman, C. and Linial, N., *Spectral properties of threshold functions.* Combinatorica, 14, 35-50 (1994)

[3] Chapman B., *The Gotsman-Linial conjecture is false.* Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM, 2018, pp. 692-699.

[4] Kane, Daniel M., *The correct exponent for the Gotsman-Linial conjecture.* Comput. Complexity, 23, 151-175 (2014)

Qiang Wu, University of Illinois at Urbana-Champaign
email: **qiangwu2@illinois.edu**