Noise Stable Halfspaces are Close to Very Small Juntas

Ilias Diakonikolas[1], Ragesh Jaiswal[2], Rocco A. Servedio[3], Li-Yang Tan[3], and Andrew Wan[4]

[1]University of Edinburgh, `ilias.d@ed.ac.uk`.
[2]Indian Institute of Technology Delhi, `rjaiswal@cse.iitd.ac.in`.
[3]Columbia University, {`rocco, liyang`}`@cs.columbia.edu`.
[4]Harvard University, `atw12@seas.harvard.edu`.

November 6, 2014

Abstract

Bourgain [Bou02] showed that any noise stable Boolean function f can be well-approximated by a junta. In this note we give an exponential sharpening of the parameters of Bourgain's result under the additional assumption that f is a halfspace.

1 Introduction

There is a sequence of results [NS94, Fri98, Bou02] in the theory of Boolean functions which share the following general flavor: if the Fourier spectrum of a Boolean function f is concentrated on low-degree coefficients, then f must be close to a junta (a function that depends only on a small number of its input variables). Bourgain's junta theorem [Bou02] is the most recent and strongest of these results; roughly speaking, it says that if a Boolean function f has low noise sensitivity then f must be close to a junta. See Section 1.1 for definitions and a precise statement of Bourgain's theorem. (Subsequently [KS] generalized Bourgain's result to product distributions, albeit with somewhat weaker parameters. More recently [KO12] gave a sharpening in the parameters of Bourgain's theorem; see Section 1.1.)

The parameters in the statement of Bourgain's theorem are essentially the best possible for general Boolean functions, in the sense that the n-variable Majority function almost (but not quite) satisfies the premise of the theorem – its noise sensitivity is only slightly higher than the bound required by the theorem – but is very far from any junta. It is interesting, though, to consider whether quantitative improvements of the theorem are possible for restricted classes of Boolean functions; this is what we do in this paper, by considering the special case when f is a halfspace. In [DS09] a quantitatively stronger version of an earlier "junta theorem" due to Friedgut [Fri98] was proved for the special case of halfspaces, and it was asked whether a similarly strengthened version of Bourgain's theorem held for halfspaces as well. Intuitively, any halfspace which has noise sensitivity lower than that of Majority should be "quite unlike Majority" and thus could reasonably be expected to depend on few variables; our result makes this intuition precise.

In this note we show that halfspaces do indeed satisfy a junta-type theorem which is similar to Bourgain's but with exponentially better parameters. Our main result shows that if f is a halfspace which (unlike the Majority function) satisfies a noise sensitivity bound similar to the one in Bourgain's original theorem, then f must be close to a junta of exponentially smaller size than

is guaranteed by the original theorem. Our proof does not follow either the approach of Bourgain or of [DS09] but instead is a case analysis based on the value of a structural parameter known as the "critical index" [Ser07, DGJ$^+$10, OS11] of the halfspace.

1.1 Background and Statement of Main Result

We view Boolean functions as mappings $f : \{-1, 1\}^n \to \{-1, 1\}$. All probabilities and expectations over $x \in \{-1, 1\}^n$ are taken with respect to the uniform distribution, unless otherwise specified. We say that $f, g : \{-1, 1\}^n \to \{-1, 1\}$ are ϵ-close to each other (or that g is an ϵ-approximator to f) if $\Pr[f(x) \neq g(x)] \leq \epsilon$.

A function $f : \{-1, 1\}^n \to \{-1, 1\}$ is said to be a "junta on $\mathcal{J} \subseteq [n]$" if f only depends on the coordinates in $\mathcal{J}$. We say that f is a J-junta, $0 \leq J \leq n$, if it is a junta on some set of cardinality at most J.

Definition 1 (Noise sensitivity)**.** *The* noise sensitivity *of a Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ at noise rate ϵ is defined as*

$$\mathbb{NS}_\epsilon(f) = \Pr_{x,y}[f(x) \neq f(y)],$$

where x is uniformly distributed and y is obtained from x by flipping each bit of x independently with probability ϵ.

Theorem 2 ([Bou02], [KO12])**.** *There exists a universal constant $C > 0$ such that the following holds. Fix $f : \{-1, 1\}^n \to \{-1, 1\}$ and ϵ, δ sufficiently small.[1] If $\mathbb{NS}_\epsilon(f) \leq C\delta\sqrt{\epsilon}$, then f is δ-close to a $(\frac{1}{\epsilon\delta})^{O(1/\epsilon)}$-junta.*

Discussion. Bourgain's paper had a somewhat stronger assumption on the noise sensitivity, in particular $\mathbb{NS}_\epsilon(f) \leq (\delta\sqrt{\epsilon})^{1+o(1)}$ for an unspecified function in the $o(1)$. Subsequently Khot and Naor (see Theorem 4.3 of [KN06]) optimized the parameters of Bourgain's proof providing an explicit dependence. The aforementioned tight quantitative statement follows from the recent work of Kindler and O'Donnell [KO12]. It is a slight strengthening of Corollary 3.21 in their paper, whose proof is very similar to the proof of the latter. The essential difference is that one needs to use Theorem 3.2 of [KO12] instead of Theorem 3.19 in the proof [O'D13].

A *halfspace*, or *linear threshold function* (henceforth simply referred to as an LTF), over $\{-1, 1\}^n$ is a Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ of the form $f(x) = \text{sign}(\sum_{i=1}^n w_i x_i - \theta)$, where $w_1, \ldots, w_n, \theta \in \mathbb{R}$. The function $\text{sign}(z)$ takes value 1 if $z \geq 0$ and takes value -1 if $z < 0$; the values $w_1, \ldots, w_n$ are the *weights* of f and θ is the *threshold*. LTFs have been intensively studied for decades in many different fields such as machine learning and computational learning theory, computational complexity, and voting theory and the theory of social choice.

Our main result, given below, is a strengthening of Bourgain's theorem that applies to the special case of halfspaces:

Theorem 3 (Main Result)**.** *There exists a universal constant $C > 0$ such that the following holds. Fix $f : \{-1, 1\}^n \to \{-1, 1\}$ to be any LTF and ϵ, δ sufficiently small. If $\mathbb{NS}_\epsilon(f) \leq C\delta^{(2-\epsilon)/(1-\epsilon)}\sqrt{\epsilon}$, then f is δ-close to an $O\left((1/\epsilon^2) \cdot \log(1/\epsilon) \cdot \log(1/\delta)\right)$-junta.*

---

[1]Here and throughout the paper, "sufficiently small" means "in the interval $(0, c)$" where $c > 0$ is some universal constant that we do not specify.

1.2 Comparison with Previous Work

In comparing Theorem 3 with Bourgain's junta theorem (Theorem 2), it should of course be emphasized that Theorem 3 applies only to LTFs while Theorem 2 applies to any Boolean function. When Theorem 3 does apply it requires a slightly stronger bound on the noise sensitivity in terms of δ, namely as much as $\delta^{(2-\epsilon)/(1-\epsilon)}$ versus essentially δ, but the resulting junta size bound of Theorem 3 is exponentially smaller, both as a function of ϵ and of δ, than the bound of Theorem 2.

The prior work (of which we are aware) that is the most closely related to Theorem 3 is the aforementioned result of [DS09] which gave an LTF analogue of Friedgut's junta theorem. The result of [DS09] is as follows:

Theorem 4 ([DS09]). *Fix $f : \{-1,1\}^n \to \{-1,1\}$ to be any LTF and $\delta > 0$ sufficiently small. Then f is δ-close to an $\mathrm{Inf}(f)^2 \cdot \mathrm{poly}(1/\delta)$-junta.*

Our Theorem 3 directly implies a result that is qualitatively similar to, but somewhat quantitatively weaker than, Theorem 4. To see this, given an LTF f set $\epsilon = C^2\delta^4/\mathrm{Inf}(f)^2$. Then, using the well-known fact that $\mathbb{NS}_\epsilon(f) \leq \epsilon \cdot \mathrm{Inf}(f)$, we get that

$$\mathbb{NS}_\epsilon(f) \leq \epsilon \cdot \mathrm{Inf}(f) = \frac{C^2\delta^4}{\mathrm{Inf}(f)} = C\delta^2\sqrt{\epsilon} < C\delta^{(2-\epsilon)/(1-\epsilon)}\sqrt{\epsilon},$$

so by Theorem 3 we have that f is δ-close to a junta over

$$O\left((1/\epsilon^2) \cdot \log(1/\epsilon) \cdot \log(1/\delta)\right) = O\left(\frac{\mathrm{Inf}(f)^4}{\delta^8} \cdot \log\frac{\mathrm{Inf}(f)}{\delta} \cdot \log\frac{1}{\delta}\right)$$

many variables. (It should be noted that this bound does not give a meaningful result for LTFs unless $\mathrm{Inf}(f) \ll n^{1/4}$, whereas the original result of [DS09] gives a meaningful bound as soon as $\mathrm{Inf}(f) \ll n^{1/2}$, which is the largest possible value for LTFs.)

On the other hand, we observe that Theorem 3 can sometimes give much stronger quantitative bounds for LTFs than Theorem 4. To see this, consider the LTF $f : \{-1,1\}^{n+(\log n)/10} \to \{-1,1\}$,

$$f(x,y) = \mathrm{sign}(10n(x_1 + \cdots + x_{(\log n)/10}) + y_1 + \cdots + y_n - n\log n).$$

(The constant "1/10" is chosen solely for concreteness here; any other constant in $(0,1/2)$ would do as well.) Observing that $f(x,y) = 1$ if and only if both $x_1 = \cdots = x_{(\log n)/10} = 1$ and $\mathrm{Maj}(y_1,\ldots,y_n) = 1$, it is easy to verify that $\mathrm{Inf}(f) = \Theta(n^{0.4})$. Hence taking δ to be (say) 1/1000, Theorem 4 only implies that f is δ-close to a junta over $O(n^{0.8})$ many variables, which is quite a poor bound on junta size. In contrast, Theorem 3 gives a much sharper bound; taking $\epsilon = \Theta(1)$ and recalling that $\mathbb{NS}_\epsilon(f) \leq 2\Pr[f = 1] = O(n^{-1/10})$, we may apply Theorem 3 to obtain that f is δ-close to an $O(1)$-junta.

2 Preliminaries

2.1 Basic Notation

For $n \in \mathbb{Z}_+$, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$. For $a, b, \epsilon \in \mathbb{R}_+$ we write $a \overset{\epsilon}{\approx} b$ to indicate that $|a - b| = O(\epsilon)$. Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian distribution with mean μ and variance σ^2. Let ϕ, Φ denote the probability density function (pdf) and cumulative distribution function (cdf) respectively of a standard Gaussian random variable $X \sim \mathcal{N}(0, 1)$.

2.2 Probabilistic Facts

We require some basic probability results including the standard additive Hoeffding bound (see e.g., [DP09]):

Theorem 5. *Let $X_1, \ldots, X_n$ be independent random variables such that for each $j \in [n]$, X_j is supported on $[a_j, b_j]$ for some $a_j, b_j \in \mathbb{R}$, $a_j \leq b_j$. Let $X = \sum_{j=1}^n X_j$. Then, for any $t > 0$, $\Pr\left[|X - \mathbb{E}[X]| \geq t\right] \leq 2\exp\left(-2t^2 / \sum_{j=1}^n (b_j - a_j)^2\right)$.*

The Berry-Esséen theorem (see e.g., [Fel68]) gives explicit error bounds for the Central Limit Theorem:

Theorem 6. *(Berry-Esséen) Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbb{E}[X_i] = 0$ for all $i \in [n]$, $\sqrt{\sum_i \mathbb{E}[X_i^2]} = \sigma$, and $\sum_i \mathbb{E}[|X_i|^3] = \rho_3$. Let $S = \sum_i X_i / \sigma$ and let F denote the cumulative distribution function (cdf) of S. Then $\sup_x |F(x) - \Phi(x)| \leq \rho_3 / \sigma^3$.*

Definition 7. *A vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ is said to be τ-regular if $\max_i |w_i| \leq \tau \|w\|_2$.*

An easy consequence of the Berry-Esséen theorem is the following fact, which says that a τ-regular linear form behaves approximately like a Gaussian up to error $O(\tau)$:

Fact 8. *Let $w = (w_1, \ldots, w_n)$ be a τ-regular vector in $\mathbb{R}^n$ with $\|w\|_2 = 1$. Then for any interval $[a, b] \subseteq \mathbb{R}$, we have $\Pr[\sum_{i=1}^n w_i x_i \in (a, b]] \stackrel{\tau}{\approx} \Phi(b) - \Phi(a)$. (In fact, the hidden constant in the $\stackrel{\tau}{\approx}$ is at most 2.)*

We say that two real-valued random variables X, Y are ρ-correlated if $\mathbb{E}[XY] = \rho$. We will need the following generalization of Fact 8 which is a corollary of the two-dimensional Berry-Esséen theorem (see e.g., Theorem 68 in [MORS10]).

Theorem 9. *Let $w = (w_1, \ldots, w_n)$ be a τ-regular vector in $\mathbb{R}^n$ with $\|w\|_2 = 1$. Let (x, y) be a pair of ρ-correlated n-bit binary strings, i.e., a draw of (x, y) is obtained by drawing x uniformly from $\{-1, 1\}^n$ and independently for each i choosing $y_i \in \{-1, 1\}$ to satisfy $\mathbb{E}[x_i y_i] = \rho$. Then for any intervals $I_1 \subseteq \mathbb{R}$ and $I_2 \subseteq \mathbb{R}$ we have $\Pr[(\sum_{i=1}^n w_i x_i, \sum_{i=1}^n w_i y_i) \in (I_1, I_2)] \stackrel{\tau}{\approx} \Pr[(X, Y) \in (I_1, I_2)]$, where (X, Y) is a pair of ρ-correlated standard Gaussians.*

2.3 Fourier Basics over $\{-1, 1\}^n$

We consider functions $f : \{-1, 1\}^n \to \mathbb{R}$ (though we often focus on Boolean-valued functions which map to $\{-1, 1\}$), and we view the inputs x to f as being distributed according to the uniform distribution. The set of such functions forms a 2^n-dimensional inner product space with inner product given by $\langle f, g \rangle = \mathbb{E}[f(x)g(x)]$. The set of functions $(\chi_S)_{S \subseteq [n]}$ defined by $\chi_S(x) = \prod_{i \in S} x_i$ forms a complete orthonormal basis for this space. We will often simply write x_S for $\prod_{i \in S} x_i$. Given a function $f : \{-1, 1\}^n \to \mathbb{R}$ we define its *Fourier coefficients* by $\widehat{f}(S) \stackrel{\text{def}}{=} \mathbb{E}[f(x)x_S]$, and we have that $f(x) = \sum_S \widehat{f}(S) x_S$.

As an easy consequence of orthonormality we have *Plancherel's identity* $\langle f, g \rangle = \sum_S \widehat{f}(S)\widehat{g}(S)$, which has as a special case *Parseval's identity*, $\mathbb{E}[f(x)^2] = \sum_S \widehat{f}(S)^2$. From this it follows that for every $f : \{-1, 1\}^n \to \{-1, 1\}$ we have $\sum_S \widehat{f}(S)^2 = 1$. It is well-known and easy to show that the noise sensitivity of f can be expressed as a function of its Fourier spectrum as follows $\mathbb{NS}_\epsilon(f) = \frac{1}{2} - \frac{1}{2} \cdot \sum_{S \subseteq [n]} (1 - 2\epsilon)^{|S|} \cdot \widehat{f}(S)^2$.

4

3 Proof of Theorem 3

Fix ϵ, δ sufficiently small. Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF satisfying $\mathbb{NS}_\epsilon(f) \leq O(\delta^{\frac{2-\epsilon}{1-\epsilon}} \cdot \sqrt{\epsilon})$. We will show that f is δ-close to an $O\left((1/\epsilon^2) \cdot \log(1/\epsilon) \cdot \log(1/\delta)\right)$-junta.

We start by observing that for $\delta^{\frac{1}{1-\epsilon}} < \sqrt{\epsilon}$ the desired statement follows easily; indeed, under the assumption of the theorem f is δ-close to a constant function. This is formalized in the following simple claim which holds for any Boolean function:

Claim 10. *Let* $f : \{-1,1\}^n \to \{-1,1\}$ *be any Boolean function and* $0 < \delta^{\frac{1}{1-\epsilon}} < \sqrt{\epsilon}$. *If* $\mathbb{NS}_\epsilon(f) \leq \delta^{\frac{2-\epsilon}{1-\epsilon}} \cdot \sqrt{\epsilon}$, *then* f *is* δ-close *to a constant function.*

Proof. For any Boolean function we have $\sum_{S \neq \emptyset} (1 - 2\epsilon)^{|S|} \cdot \widehat{f}^2(S) \leq (1 - 2\epsilon) \cdot \sum_{S \neq \emptyset} \widehat{f}^2(S) = (1 - 2\epsilon) \cdot (1 - \widehat{f}^2(\emptyset))$ where the equality follows from Parseval's identity. Therefore, we can write

$$\mathbb{NS}_\epsilon(f) = \frac{1}{2} \cdot \left(1 - \widehat{f}^2(\emptyset) - \sum_{S \neq \emptyset} (1 - 2\epsilon)^{|S|} \cdot \widehat{f}^2(S)\right) \geq \epsilon \cdot \left(1 - \widehat{f}^2(\emptyset)\right)$$

which implies $1 - \widehat{f}^2(\emptyset) \leq \delta^{\frac{2-\epsilon}{1-\epsilon}}/\epsilon^{1/2} \leq \delta$ where the first inequality follows from the assumed upper bound on the noise sensitivity and the second uses the assumption that $\delta^{\frac{1}{1-\epsilon}} < \sqrt{\epsilon}$. It follows that f is δ-close to $\text{sign}(\hat{f}(\emptyset))$ and this completes the proof. ∎

Using the above lemma, for the rest of the proof we can assume that $\delta^{\frac{1}{1-\epsilon}} \geq \sqrt{\epsilon}$.

Fix a weight-based representation of f as $f(x) = \text{sign}(w \cdot x - \theta)$, where we assume, without loss of generality, that $\sum_i w_i^2 = 1$ and $|w_i| \geq |w_{i+1}| > 0$, for all $i \in [n-1]$. For $k \in [n]$, we denote $\sigma_k \stackrel{\text{def}}{=} \sqrt{\sum_{i=k}^n w_i^2}$. The proof proceeds by case analysis based on the value of the ϵ-critical index of the vector w, which we now define.

Definition 11 (critical index, [Ser07]). *We define the* τ-critical index $\ell(\tau)$ *of a vector* $w \in \mathbb{R}^n$ *as the smallest index* $i \in [n]$ *for which* $|w_i| \leq \tau \cdot \sigma_i$. *If this inequality does not hold for any* $i \in [n]$, *we define* $\ell(\tau) = \infty$.

The case analysis is essentially the same as the one used in [Ser07, DGJ+10]. Let $\ell \stackrel{\text{def}}{=} \ell(\epsilon)$ be the ϵ-critical index of f. We fix a parameter

$$L(\epsilon, \delta) \stackrel{\text{def}}{=} \Theta\left(\frac{1}{\epsilon^2} \cdot \log(1/\epsilon) \cdot \log(1/\delta)\right)$$

for an appropriately large value of the constant in the $\Theta(\cdot)$. If $\ell = 1$, then the linear form behaves like a Gaussian and must be either biased or noise sensitive. In Lemma 12, we show that such an f is either δ-close to constant or has noise sensitivity $\Omega(\delta^{\frac{1}{1-\epsilon}}\sqrt{\log(1/\delta)}\sqrt{\epsilon})$. (See Case I below.) If $\ell > L$, then previous results [Ser07] establish that f is δ-close to a junta. (See Case III.) Finally, for $1 < \ell < L$, we consider taking random restrictions to the variables before the critical index. If a $(1 - \delta)$-fraction of these restrictions result in subfunctions which are very biased, then f must be 3δ-close to a junta over the first L variables. Otherwise, a δ-fraction of the restrictions result in regular LTFs which are not very biased, and we can apply the results from Case I to show that the noise sensitivity of f must be too large to satisfy the conditions of Theorem 3. We show this in Lemma 16, Case II. Our requirement on the noise sensitivity in Theorem 3, which is probably stronger than optimal, comes from the analysis of this case.

We now proceed to consider each of these three cases formally.

Case I: [$\ell = 1$, i.e., the vector w is ϵ-regular.] In this case we show that f is δ-close to a constant function. The argument proceeds as follows: If $|\mathbb{E}[f]| < 1 - \delta$, we prove (Lemma 12) that $\mathbb{NS}_\epsilon(f) = \Omega(\delta^{\frac{1}{1-\epsilon}}\sqrt{\log(1/\delta)}\cdot\sqrt{\epsilon})$ contradicting the assumption of the theorem. Hence, $|\mathbb{E}[f]| \geq 1 - \delta$, i.e., f is δ-close to a constant. Our main lemma in this section establishes the intuitive fact that a regular LTF that is not-too-biased towards a constant function has high noise sensitivity.

Lemma 12. *Fix $0 < \epsilon \leq 1/2$. Let $f : \{-1,1\}^n \to \{-1,1\}$ be an ϵ-regular LTF $f(x) = \text{sign}(w\cdot x - \theta)$ that has $|\mathbb{E}[f]| = 1 - p$. Then we have*

$$\mathbb{NS}_\epsilon(f) = \Omega\left(p^{\frac{1}{1-\epsilon}}\sqrt{\log(1/p)}\cdot\sqrt{\epsilon}\right) - O(\epsilon).$$

Case I follows easily from the above lemma. Suppose that $p \leq \delta$. Then the function f is δ-close to a constant. Otherwise, the lemma implies that $\mathbb{NS}_\epsilon(f) = \Omega(\delta^{\frac{1}{1-\epsilon}}\sqrt{\log(1/\delta)}\cdot\sqrt{\epsilon}) - O(\epsilon)$; since $\delta^{\frac{1}{1-\epsilon}} \geq \sqrt{\epsilon}$, this is $\Omega(\delta^{\frac{1}{1-\epsilon}}\sqrt{\log(1/\delta)}\cdot\sqrt{\epsilon})$. This contradicts our assumed upper bound on $\mathbb{NS}_\epsilon(f)$ from the statement of the main theorem.

The proof of Lemma 12 proceeds by first establishing the analogous statement in Gaussian space (Lemma 13 below) and then using invariance to transfer the statement to the Boolean setting.

We start by giving a lower bound on the Gaussian noise sensitivity of any LTF as a function of the noise rate and the threshold of the LTF. The following lemma is classical for $\theta = 0$. We were not able to find an explicit reference for arbitrary θ, so we give a proof for the sake of completeness.

Lemma 13. *Let $0 < \epsilon \leq 1/2$ and $\theta \in \mathbb{R}$. Let X and Y be $\rho \overset{\text{def}}{=} (1 - 2\epsilon)$-correlated standard Gaussians. Then,*

$$\Pr[\text{sign}(X - \theta) \neq \text{sign}(Y - \theta)] \geq (1/\pi)\cdot\arccos(\rho)\cdot e^{-\frac{\theta^2}{1+\rho}}.$$

Proof. Let X and Y be ρ-correlated standard Gaussians. As is well known, (X, Y) can be generated as follows

$$X = Z_1 = (Z_1, Z_2)\cdot(1,0)^T \quad\text{and}\quad Y = \rho\cdot Z_1 + \sqrt{1-\rho^2}\cdot Z_2 = (Z_1, Z_2)\cdot(\rho, \sqrt{1-\rho^2})^T.$$

where Z_1 and Z_2 are *independent* standard Gaussians. For the random variables $X - \theta$ and $Y - \theta$ we can write

$$X - \theta = \left(Z_1 - \theta, Z_2 - \theta\cdot\sqrt{\frac{1-\rho}{1+\rho}}\right)\cdot(1,0)^T \quad\text{and}$$

$$Y - \theta = \left(Z_1 - \theta, Z_2 - \theta\cdot\sqrt{\frac{1-\rho}{1+\rho}}\right)\cdot(\rho, \sqrt{1-\rho^2})^T.$$

Fix $\alpha \overset{\text{def}}{=} \sqrt{\frac{1-\rho}{1+\rho}}$ and consider the 2-dimensional random vector $T = (-Z_2 + \alpha\theta, Z_1 - \theta)$. Note that T is orthogonal to the vector $(Z_1 - \theta, Z_2 - \alpha\theta)$.

We now observe that

$$\Pr[\text{sign}(X - \theta) \neq \text{sign}(Y - \theta)] = \Pr[T \text{ "splits" vectors } (1,0) \text{ and } (\rho, \sqrt{1-\rho^2})]$$

We refer to Figure 1 for the rest of the proof. Let R be the region between the horizontal axis (the line spanned by $(1,0)$) and the line spanned by the vector $(\rho, \sqrt{1-\rho^2})$. The RHS of the above equation is equal to the probability mass of R under a 2-dimensional unit variance Gaussian
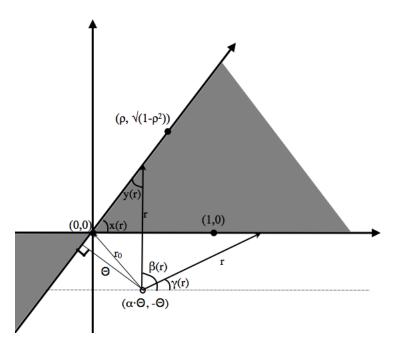
6

Figure 1: Illustration of the integration region for Lemma 13.

centered at $(\alpha\theta, -\theta)$. We estimate the Gaussian integral restricted to the region by considering points at distance $r \geq r_0$ from $(\alpha\theta, -\theta)$. Using polar coordinates to compute the integral, we obtain:

$$
\begin{aligned}
\Pr[T \text{ "splits" vectors } (1,0) \text{ and } (\rho, \sqrt{1-\rho^2})] \;\geq\;& \frac{1}{\pi} \int_{r_0}^{\infty} \int_{\gamma(r)}^{\beta(r)} re^{-r^2/2} d\phi \, dr \\
=\;& \frac{1}{\pi} \int_{r_0}^{\infty} (\beta(r) - \gamma(r)) \, re^{-r^2/2} dr. \qquad (1)
\end{aligned}
$$

The angles $\beta(r), \gamma(r)$ are illustrated in Figure 1, and r_0 is the distance of the point $(\alpha\theta, -\theta)$ from the origin, i.e.,

$$
r_0 = \theta\sqrt{1+\alpha^2} = \frac{\sqrt{2}\theta}{\sqrt{1+\rho}} \qquad (2)
$$

where the second equality follows from the definition of α. To compute (1), we need the following claim:

Claim 14. *For all $r > r_0$, it holds that $(\beta - \gamma)(r) = \arccos(\rho)$.*

Proof. Let $x(r)$ and $y(r)$ denote the angles illustrated in Figure 1. First, observe that $\beta(r) = x(r) + y(r)$ and that $x(r) = \arccos(\rho)$. We also have that $\gamma(r) = \arcsin(\theta/r)$ (the vector of length r originates at $(\alpha\theta, -\theta)$ and stops at the origin). Finally, an easy calculation shows that the distance from $(\alpha\theta, -\theta)$ to the line spanned by $(\rho, \sqrt{1-\rho^2})$ is exactly θ, and hence $y(r) = \arcsin(\theta/r)$. ∎

7

Therefore, the RHS of (1) can be written as follows:

$$\frac{1}{\pi} \int_{r_0}^{\infty} (\beta - \gamma)(r) r e^{-r^2/2} dr \quad = \quad (1/\pi) \cdot \arccos(\rho) \cdot \int_{r_0}^{\infty} r e^{-r^2/2} dr \quad \text{(using Claim 14)}$$

$$= \quad (1/\pi) \cdot \arccos(\rho) \left[-e^{-r^2/2} \right]_{r_0}^{\infty}$$

$$= \quad (1/\pi) \cdot \arccos(\rho) \cdot e^{-r_0^2/2}$$

$$= \quad (1/\pi) \cdot \arccos(\rho) \cdot e^{-\frac{\theta^2}{1+\rho}}$$

where the last equality follows from (2). This concludes the proof of Lemma 13. ∎

We are now ready to give the proof of Lemma 12.

Proof of Lemma 12. We first bound from below the Gaussian sensitivity of a halfspace as a function of its bias and the noise rate. Let (X, Y) be a pair of $\rho \stackrel{\text{def}}{=} (1 - 2\epsilon)$-correlated standard Gaussians. Consider the one-dimensional halfspace $h_\theta : \mathbb{R} \to \{-1, 1\}$ defined as $h_\theta(x) = \text{sign}(x - \theta)$ and let $\left| \mathbb{E}_{x \sim \mathcal{N}(0,1)}[h_\theta(x)] \right| = 1 - \widetilde{p}$. We claim that

$$\Pr[h_\theta(X) \neq h_\theta(Y)] = \Omega\left(\widetilde{p}^{\frac{1}{1-\epsilon}} \sqrt{\log(1/\widetilde{p})} \cdot \sqrt{\epsilon} \right). \tag{3}$$

We show (3) as follows: Lemma 13 implies that

$$\Pr[h_\theta(X) \neq h_\theta(Y)] = \Omega\left(\sqrt{\epsilon} \cdot e^{-\frac{\theta^2}{2-2\epsilon}} \right) \tag{4}$$

where we used the elementary inequality $\arccos(1 - 2\epsilon) = \Omega(\sqrt{\epsilon})$. We now relate $\widetilde{p}$ and θ. We claim that

$$\widetilde{p} = \Theta\left(\frac{e^{-\theta^2/2}}{|\theta| + 1} \right).$$

From this it follows that

$$e^{-\theta^2/2} = \Theta\left(\widetilde{p} \sqrt{\log(1/\widetilde{p})} \right)$$

and (4) yields (3). It remains to get the desired bound on $\widetilde{p}$. Assume that $\theta \geq 0$; for $\theta < 0$ the argument is symmetric. First, it is easy to see that

$$\mathbb{E}_{x \sim \mathcal{N}(0,1)}[h_\theta(x)] = -1 + 2\widetilde{\Phi}(\theta)$$

where $\widetilde{\Phi}(\theta) \stackrel{\text{def}}{=} \Pr_{x \sim \mathcal{N}(0,1)}[x \geq \theta]$. Since $\theta \geq 0$, we have $\widetilde{\Phi}(\theta) \leq 1/2$, hence $\widetilde{p} = 2\widetilde{\Phi}(\theta)$. The desired bound on $\widetilde{p}$ now follows from the following elementary fact:

Fact 15. *For all $\theta \geq 0$, it holds $\widetilde{\Phi}(\theta) = \Theta(\frac{e^{-\theta^2/2}}{|\theta|+1})$.*

We now turn to the Boolean setting to finish the proof of Lemma 12. Let $f = \text{sign}(w \cdot x - \theta)$ be a Boolean ϵ-regular LTF (where without loss of generality $\|w\|_2 = 1$) that has $|\mathbb{E}[f]| = 1 - p$. We use (3) and invariance to prove the lemma. In particular, we have the following sequence of inequalities:

$$\mathbb{NS}_\epsilon(f) \quad = \quad \Pr[\text{sign}(w \cdot x - \theta) \neq \text{sign}(w \cdot y - \theta)]$$

$$\stackrel{2\epsilon}{\approx} \quad \Pr[\text{sign}(X - \theta) \neq \text{sign}(Y - \theta)] \tag{5}$$

$$= \quad \Omega(\widetilde{p}^{\frac{1}{1-\epsilon}} \sqrt{\log(1/\widetilde{p})} \cdot \sqrt{\epsilon}) - O(\epsilon) \tag{6}$$

$$= \quad \Omega(p^{\frac{1}{1-\epsilon}} \sqrt{\log(1/p)} \cdot \sqrt{\epsilon}) - O(\epsilon) \tag{7}$$

where (5) follows from Theorem 9 and (6) is an application of (3). To see (7), note that, by Fact 8 (a corollary of the Berry-Esséen theorem) we get that $p \overset{\epsilon}{\approx} \widetilde{p}$, and hence

$$|p^{1/(1-\epsilon)}\sqrt{\log(1/p)} - \widetilde{p}^{1/(1-\epsilon)}\sqrt{\log(1/\widetilde{p})}| = O(\epsilon).$$

This completes the proof of the lemma. ∎

Case II: $[1 < \ell \leq L.]$ In this case, we show that f is δ-close to an ℓ-junta.

Consider the partition of the set $[n]$ into a set of head variables $H = [\ell]$ and a set of tail variables $T = [n] \setminus H$. Let us write $H(x_H)$ to denote $w_H \cdot x_H$ and $T(x_T)$ to denote $w_T \cdot x_T$, the linear forms corresponding to the head and the tail.

The argument proceeds as follows: If a non-trivial fraction of restrictions to the head variables lead to a not-too-biased LTF, we show that the original LTF has high noise sensitivity contradicting the assumption of the theorem. On the other hand, if most restrictions to the head lead to a substantially biased LTF, we argue that the original LTF is close to a junta over the head coordinates.

Let $\rho \in \{-1,1\}^{|H|}$ denote an assignment to the head coordinates and f_ρ be the corresponding restriction of f. Note that for any restriction ρ of the head variables the resulting f_ρ is an ϵ-regular LTF (with a threshold of $H(\rho) - \theta$). Formally, we consider two sub-cases depending on the distribution of $|\mathbb{E}[f_\rho]|$ for a random choice of ρ.

Case IIa: [This case corresponds to $\Pr_\rho\left[|\mathbb{E}[f_\rho]| \leq 1 - \delta\right] > \delta.$] That is, at least a δ fraction of restrictions to the head variables result in a "not-too-biased" LTF. Since each of these restricted sub-functions has high noise-sensitivity, we can show that the overall noise-sensitivity is also somewhat high. This intuitive claim is quantified in the following lemma.

Lemma 16. *Let ϵ, δ be sufficiently small values that satisfy $\delta^2 \geq \sqrt{\epsilon}$. Let the ϵ-critical index ℓ of f satisfy $1 < \ell \leq L$. If $\Pr_\rho\left[|\mathbb{E}[f_\rho]| \leq 1 - \delta\right] > \delta$, then $\mathbb{NS}_\epsilon(f) = \Omega(\delta^{\frac{2-\epsilon}{1-\epsilon}}\sqrt{\log(1/\delta)} \cdot \sqrt{\epsilon})$.*

Therefore, in Case IIa we reach a contradiction. To prove the above lemma, we need the following claim, whch implies that if a noticeable fraction of restrictions to a Boolean function have high noise sensitivity, then so does the original function.

Claim 17. *Let $f : \{-1,1\}^n \to \{-1,1\}$, $R \subseteq [n]$ and $\rho \in \{-1,1\}^{|R|}$ be a random restriction to the variables in R. For any $\epsilon > 0$, if $\Pr_\rho[\mathbb{NS}_\epsilon(f_\rho) > \tau] > \delta$, then $\mathbb{NS}_\epsilon(f) \geq \tau\delta$.*

Proof. The following elementary fact will be useful for the proof:

Fact 18. *Let $f : \{-1,1\}^n \to \{-1,1\}$, $R \subseteq [n]$ and $\rho \in \{-1,1\}^{|R|}$. For any $S \subseteq ([n] \setminus R)$,*

$$\mathbb{E}_\rho[\widehat{f_\rho}(S)^2] = \sum_{T \subseteq R} \widehat{f}(S \cup T)^2.$$

By linearity of expectation and Fact 18 we get that

$$\mathbb{E}_\rho[\mathbb{NS}_\epsilon(f_\rho)] = \frac{1}{2} \cdot \sum_{S \subseteq ([n] \setminus R)} (1 - (1-2\epsilon)^{|S|}) \cdot \sum_{T \subseteq R} \hat{f}(S \cup T)^2 \tag{8}$$

On the other hand, we have:

$$
\begin{aligned}
\mathbb{NS}_\epsilon(f) &= \frac{1}{2} \cdot \sum_{S \subseteq ([n] \setminus R)} \sum_{T \subseteq R} \left(1 - (1 - 2\epsilon)^{|S|+|T|}\right) \cdot \hat{f}(S \cup T)^2 \\
&\geq \frac{1}{2} \cdot \sum_{S \subseteq ([n] \setminus R)} \sum_{T \subseteq R} \left(1 - (1 - 2\epsilon)^{|S|}\right) \cdot \hat{f}(S \cup T)^2 \\
&= \frac{1}{2} \cdot \sum_{S \subseteq ([n] \setminus R)} (1 - (1 - 2\epsilon)^{|S|}) \cdot \sum_{T \subseteq R} \hat{f}(S \cup T)^2 \qquad (9)
\end{aligned}
$$

Combining equations 8 and 9, we obtain

$$
\mathbb{NS}_\epsilon(f) \geq \mathbb{E}_\rho[\mathbb{NS}_\epsilon(f_\rho)] \geq \delta\tau.
$$

∎

Using the above claim we can prove Lemma 16.

Proof of Lemma 16. By Claim 17 and the assumption that $\Pr_\rho\left[|\mathbb{E}[f_\rho]| \leq 1 - \delta\right] > \delta$, it suffices to show that f_ρ is noise sensitive whenever $|\mathbb{E}[f_\rho]| \leq 1 - \delta$, i.e., that

$$
\mathbb{NS}_\epsilon(f_\rho) = \Omega(\delta^{\frac{1}{1-\epsilon}} \sqrt{\log(1/\delta)} \cdot \sqrt{\epsilon}).
$$

This follows from the fact that f_ρ is an ϵ-regular LTF. Applying Lemma 12 with $p = \delta \geq \epsilon^{\frac{1-\epsilon}{2}}$ completes the proof.

∎

Case IIb: [The complementary case corresponds to $\Pr_\rho\left[|\mathbb{E}[f_\rho]| \leq 1 - \delta\right] \leq \delta$.] That is, with probability at least $1 - \delta$ over a random restriction of the head, the bias of the corresponding restriction is "large." In this case, a simple argument yields the following:

Lemma 19. *Let $f : \{-1, 1\}^n \to \{-1, 1\}$, $H \subseteq [n]$, and $0 < \delta \leq 1$. Suppose $\Pr_{\rho \sim H}\left[|\mathbb{E}[f_\rho]| \leq 1 - \delta\right] \leq \delta$. Then f is 3δ-close to a junta over H.*

Proof. Let $B \subseteq \{-1, 1\}^{|H|}$ denote the set of bad restrictions, where we say that a restriction $\rho \in \{-1, 1\}^{|H|}$ is bad if $|\mathbb{E}[f_\rho]| \leq 1 - \delta$. Define $g : \{-1, 1\}^n \to \{-1, 1\}$ to be:

$$
g(x) = \begin{cases} 1 & \text{if } x_H \in B \\ f(x) & \text{otherwise,} \end{cases}
$$

and note that g is δ-close to f since $|B| \leq \delta \cdot 2^{|H|}$ by assumption. We also have that g satisfies $|\widehat{g_\rho}(\emptyset)| = |\mathbb{E}[g_\rho]| > 1 - \delta$ for all $\rho \in \{-1, 1\}^{|H|}$. Now consider $h(x) = \sum_{S \subseteq H} \hat{g}(S) x_S$ and note that

$$
\|h - g\|_2^2 = \sum_{T \not\subseteq H} \hat{g}(T)^2 = \mathbb{E}_{\rho \sim H}[\text{Var}(g_\rho)] = 1 - \mathbb{E}_{\rho \sim H}[\widehat{g_\rho}(\emptyset)^2] < 1 - (1 - 2\delta) = 2\delta.
$$

Since f is δ-close to g and g is 2δ-close to $\text{sign}(h)$ (a junta over H), this completes the proof. ∎

This completes Case II.

Case III: [$\ell > L$]. In this case, we merely observe that f is δ-close to an L-junta. This follows immediately from the arguments in [Ser07, DGJ+10]. In particular,

Lemma 20 (Case II(a) of Theorem 1 of [Ser07])**.** *Fix $\epsilon, \delta > 0$. Let f be an LTF with ϵ-critical index $\ell > L$. Then f is δ-close to an L-junta.*

The proof of Theorem 3 is now complete.

References

[Bou02]   J. Bourgain. On the distributions of the fourier spectrum of boolean functions. *Israel J. Math.*, 131:269–276, 2002.

[DGJ$^+$10]   I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM J. on Comput.*, 39(8):3441–3462, 2010.

[DP09]   D. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms.* Cambridge University Press, Cambridge, 2009.

[DS09]   I. Diakonikolas and R. Servedio. Improved approximation of linear threshold functions. In *Proc. 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 161–172, 2009.

[Fel68]   W. Feller. *An introduction to probability theory and its applications.* John Wiley & Sons, 1968.

[Fri98]   E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):474–483, 1998.

[KN06]   S. Khot and A. Naor. Nonembeddability theorems via Fourier analysis. *Mathematische Annalen*, 334(4):821–852, 2006.

[KO12]   G. Kindler and R. O'Donnell. Gaussian noise sensitivity and fourier tails. In *CCC*, pages 137–147, 2012.

[KS]   G. Kindler and S. Safra. Noise resistant boolean functions are juntas. Manuscript, 2004.

[MORS10]   K. Matulef, R. O'Donnell, R. Rubinfeld, and R. Servedio. Testing halfspaces. *SIAM J. on Comput.*, 39(5):2004–2047, 2010.

[NS94]   N. Nisan and M. Szegedy. On the degree of Boolean functions as real polynomials. *Comput. Complexity*, 4:301–313, 1994.

[O'D13]   R. O'Donnell. Personal communication, 2013.

[OS11]   R. O'Donnell and R. Servedio. The Chow Parameters Problem. *SIAM J. on Comput.*, 40(1):165–199, 2011.

[Ser07]   R. Servedio. Every linear threshold function has a low-weight approximator. *Comput. Complexity*, 16(2):180–209, 2007.