Analysis of Boolean Functions **(CMU 18-859S, Spring 2007)**

Lecture 19: Noise sensitivity of Linear Threshold Functions

March 27, 2007

Lecturer: Ryan O'Donnell *Scribe: Amitabh Basu*

We begin our study of the noise sensitivity of functions with this lecture. We wish to investigate how sensitive the output of a boolean function is to small corruptions in the input bits. This leads to efficient learning algorithms for an important class of functions which we will define in this lecture. But the emphasis in today's lecture will be on achieving tight bounds on the noise sensitivity of this class of functions.

1 Noise sensitivity and Learning

We recapitulate the definition of noise sensitivity. If $f : \{-1,1\}^n \to \{-1,1\}$, then for $0 \le \epsilon \le \frac{1}{2}$

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

where **x** and **y** are chosen by choosing **x** uniformly at random and then forming **y** by flipping each bit of **x** with probability ϵ. We denote this by $\mathbf{y} = \mathbb{N}_\epsilon(\mathbf{x})$

We proved the following fact in Homework 3.

Fact 1.1 *If $\mathbb{NS}_\delta(f) \le \epsilon$, then f is $O(\epsilon)$-concentrated on $\{S \subseteq [n] : |S| \le \frac{1}{\delta}\}$ and therefore the Low Degree Algorithm will learn the function in time poly($n^{\frac{1}{\delta}}, \frac{1}{\epsilon}$).*

Proof: From problem 3 in Homework 2, we know that $\mathbb{NS}_\delta(f) = \frac{1}{2} - \frac{1}{2}\sum_{S\subseteq[n]}(1-2\delta)^{|S|}\hat{f}(S)^2$.

Therefore,

$$\begin{aligned}
\epsilon &\ge \tfrac{1}{2} - \tfrac{1}{2}\textstyle\sum_{S\subseteq[n]}(1-2\delta)^{|S|}\hat{f}(S)^2 \\
&= \tfrac{1}{2}[\textstyle\sum_S \hat{f}(S)^2 - \sum_S(1-2\delta)^{|S|}\hat{f}(S)^2] \qquad \text{(By Parseval's identity)} \\
&= \tfrac{1}{2}[\textstyle\sum_S(1-(1-2\delta)^{|S|})\hat{f}(S)^2] \\
&\ge \tfrac{1}{2}[\textstyle\sum_{|S|\ge\frac{1}{\delta}}(1-(1-2\delta)^{|S|})\hat{f}(S)^2] \\
&\ge \tfrac{1}{2}[(1-(1-2\delta)^{\frac{1}{\delta}})\textstyle\sum_{|S|\ge\frac{1}{\delta}}\hat{f}(S)^2] \\
&\ge \tfrac{1}{2}(1-e^{-2})\textstyle\sum_{|S|\ge\frac{1}{\delta}}\hat{f}(S)^2]
\end{aligned}$$

$\therefore \sum_{|S|\ge\frac{1}{\delta}}\hat{f}(S)^2 \le \frac{2}{1-e^{-2}}\epsilon$.

□

For instance, suppose we prove that $\forall f \in \mathcal{C}$ (where $\mathcal{C}$ is some class of functions), $\mathbb{NS}_\delta(f) \le t\sqrt{\delta}$. Then $\mathcal{C}$ is learnable using the Low Degree Algorithm using time poly($n^{\frac{t^2}{\epsilon^2}}$), where ϵ is the accuracy parameter in the *PAC* learning model. For the parameters to work right, we simply need $t\sqrt{\delta} \le \Omega(\epsilon)$ so that $\frac{1}{\delta} \ge O(\frac{t^2}{\epsilon^2})$.

2 Linear Threshold Functions

In this section we define a very general class of functions called the linear threshold functions. Many of the functions we have studied before fall into this class.

Definition 2.1 *A function $f : \{-1,1\}^n \rightarrow \Re$ is called linear if $\exists$ real numbers $\alpha_i's$ such that $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \ldots \alpha_n x_n$. A function $f : \{-1,1\}^n \rightarrow \{-1,1\}$ is called a linear threshold function (LTF for short) is $\exists \alpha_i's$ such that $f(x) =$ sgn$(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n)$.*

Examples :

- Majority : sgn($\sum_{i=1}^{n} x_i$)
- Dictators : sgn(x_i)
- AND : sgn($\sum_{i=1}^{n} x_i + n - \frac{1}{2}$)
- Decision lists : These are a special case of decision trees where the tree is only one path from the root to a leaf.

Remark 2.2 *We make a technical note here. Firstly, declare sgn(0) = 0. So for LTF's we require that $\alpha_0 + \sum \alpha_i x_i \neq 0 \quad \forall x \in \{-1,1\}^n$. In fact, we will make the stronger assumption that $\sum_{i \in S} \alpha_i x_i \neq 0 \quad \forall x \in \{-1,1\}^n, \phi \neq S \subseteq [n]$.*

A slightly different interpretation of LTFs is from a probability theory point of view. LTFs can be studied as the sum of independent random variables. We now come to the main theorem of today's lecture.

Theorem 2.3 (Peres '04) *Let $f : \{-1,1\}^n \rightarrow \{-1,1\}$ be an LTF. Then $\mathbb{NS}_\delta(f) \leq 2\sqrt{\delta}$.*

In light of the discussion at the end of the previous section, we have the following corollary.

Corollary 2.4 *The class of LTFs is learnable in time poly($n^{\frac{1}{\epsilon^2}}$).*

Remark 2.5 *Actually the class of LTFs is learnable in time poly(n/ϵ). This is true even for learning under any distribution and not just the uniform distribution. The idea is to draw a bunch of examples and use a linear program to find a consistent hypothesis.*

We can also learn intersections of LTFs : Let $\mathcal{C} = g_1 \wedge g_2 : g_1, g_2$ are LTFs . The for any $f \in \mathcal{C}$, $\mathbb{NS}_\delta(f) \leq 4\sqrt{\delta}$. Why ? Note that by the union bound the probability that either g_1 or g_2 changes is at most $2\sqrt{\delta} + 2\sqrt{\delta} = 4\sqrt{\delta}$. Hence the probability that $g_1 \wedge g_2$ flips is at most $4\sqrt{\delta}$.

Corollary 2.6 *The class $\mathcal{C}$ defined above is learnable in time poly($n^{\frac{1}{\epsilon^2}}$).*

Arguing along similar lines we have the following corollary.

Corollary 2.7 *Let $\mathcal{C}$ be the collection of functions which are some function of t LTFs. Then $\forall f \in calC, \mathbb{NS}_\delta(f) \leq t \cdot 2\sqrt{\delta}$*

The proof is again simply applying the union bound. Therefore, the above class is learnable in time poly($n^{\frac{t^2}{\epsilon^2}}$).

In the next section we present the proof of Peres's theorem.

3 Proof of Peres's theorem

The proof we present here is a simplification of Peres's original proof as suggested by Parikshit Gopalan. We will prove something a little stronger than the statement of the theorem.

Theorem 3.1 *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF, and let $I_1, \dots, I_m \subseteq [n]$ are disjoint subsets. Then $\mathbf{Pr}_{x,j \in \{1,\dots,m\}}[f(\mathbf{x}) \neq f(\mathbf{x}^{I_j})] \leq \frac{1}{\sqrt{m}}$.*

Let us see why Peres's theorem follows from the above result. Given $0 \leq \delta \leq \frac{1}{2}$, let us first assume that $\delta = \frac{1}{m}$ where m is some integer. Since the theorem holds for all partitions of $[n]$, it holds in expectation if we pick a random m-partition, meaning that for each $i \in [n]$, include it into a random I_j independently. More precisely,

$$\mathbf{E}_{I_1,\dots,I_m}[\mathbf{Pr}_{x,j}(f(\mathbf{x}) \neq f(\mathbf{x}^{I_j}))] \leq \frac{1}{\sqrt{m}} \leq \sqrt{\delta}$$

But $(\mathbf{x}, \mathbf{x}^{I_j})$ has the same distribution as $(\mathbf{x}, \mathbb{N}_{\frac{1}{m}}(\mathbf{x}))$. $\therefore$ $\mathbf{Pr}_{\mathbf{x},\mathbf{y} \in \mathbb{N}_\delta(\mathbf{x})}[f(\mathbf{x}) \neq f(\mathbf{y})] \leq \sqrt{\delta}$.

If on the other hand, $\delta \neq \frac{1}{m}$ for some integer m, let δ' be the next largest integer amount which is $1/integer$. Since the noise sensitivity of a function at δ is an increasing function of δ, we have $\mathbb{NS}_\delta(f) \leq \mathbb{NS}'_\delta \leq \sqrt{\delta'}$. In the worst case, $\sqrt{\delta'} \leq 2\sqrt{\delta}$ (the worst case is when delta is very near $\frac{1}{2}$ and $\delta' = 1$).

Now to prove the theorem.

Proof: Write $f = sgn(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n)$. Given a random x, define random variables σ_j by

$$\sigma_j(x) = sgn(\sum_{i \in I_j} \alpha_i x_i)$$

Because $\sigma_j(x) = -\sigma_j(-x)$, the distribution of σ_j is precisely like that of a random ± 1 bit. Also not that since the I_j's are disjoint, σ_i, σ_j are independent.

Define $g : \{-1,1\}^m \to [-1,1]$ on the σ_i's by $g(\sigma_1, \dots, \sigma_m) = \mathbf{E}_{\mathbf{x}|\sigma}[f(\mathbf{x}]$ where the expectation is taken over $\mathbf{x}$ conditioned on σ, i.e. all $\mathbf{x}$ such the σ_i's are the right sign.

Fact 3.2 $\sum_{j=1}^m \hat{g}(j) \leq \sqrt{m}$

Proof: By the Cauchy-Schwartz inequality, $\sum_{j=1}^m \hat{g}(j) \leq \sqrt{m} \sum_{j=1}^m \hat{g}(j)^2 \leq \sqrt{m}$ □

We now make the following claim, from which the theorem is immediate.

Claim 3.3 $\hat{g}(j) = \mathbf{Pr}_x[f(\mathbf{x}) \neq f(\mathbf{x}^{(I_j)})]$

Given the claim, $\mathbf{Pr}_{x,j}[f(\mathbf{x}) \neq f(\mathbf{x}^{(I_j)})] \leq (\sum_{j=1}^m \hat{g}(j))/m \leq \frac{1}{\sqrt{m}}$

Now to prove the claim.

$$\begin{aligned} \hat{g}(j) &= \mathbf{E}_\sigma[g(\sigma) \cdot \sigma_j] \\ &= \mathbf{E}_\sigma[\mathbf{E}_{\mathbf{x}|\sigma}[f(\mathbf{x})] \cdot \sigma_j] \\ &= \mathbf{E}_x[f(\mathbf{x}) \cdot \sigma_j(\mathbf{x})] \end{aligned}$$

Now since $\mathbf{x}$ and $\mathbf{x}^{(I_j)}$ have the same distribution, the above expression is also equal to $\mathbf{E}_x[f(\mathbf{x}^{(I_j)}) \cdot \sigma_j(\mathbf{x}^{(I_j)})]$. So $\hat{g}(j)$ is also equal to

$$\mathbf{E}_x[\frac{f(\mathbf{x}) \cdot \sigma_j(\mathbf{x}) + f(\mathbf{x}^{(I_j)}) \cdot \sigma_j(\mathbf{x}^{(I_j)})}{2}]$$

Also note that $\sigma_j(\mathbf{x}) = -\sigma_j(\mathbf{x}^{(I_j)})$. Therefore we have

$$\hat{g}(j) = \mathbf{E}_x[\frac{f(\mathbf{x}) \cdot \sigma_j(\mathbf{x}) - f(\mathbf{x}^{(I_j)}) \cdot \sigma_j(\mathbf{x})}{2}]$$

- If $f(\mathbf{x}) = f(\mathbf{x}^{(I_j)})$ then the expression above is 0.
- If $f(\mathbf{x}) \neq f(\mathbf{x}^{(I_j)})$ then in fact $f(\mathbf{x}) = \sigma_j(\mathbf{x})$. This is because f is changing sign when a subset is changing sign. For example, if f is positive and the subset is negative, then if the subset becomes positive, f cannot become negative.

In other words, the expression inside the expectation is simply the indicator of $[f(\mathbf{x}) \neq f(\mathbf{x}^{(I_j)})]$. And so we get that $\hat{g}(j) = \mathbf{E}[\mathbf{1}[f(\mathbf{x}) \neq f(\mathbf{x}^{(I_j)})]] = \mathbf{Pr}_x[f(\mathbf{x}) \neq f(\mathbf{x}^{(I_j)})]$.
□

If we use the better bound of $\sum_{j=1}^{m} \hat{g}(j) \leq (\sqrt{\frac{2}{\pi}} + o_m(1))\sqrt{m}$, we get the following corollary of Peres's theorem.

Corollary 3.4 *If f is an LTF,* $\mathbb{NS}_\delta(f) \leq (\sqrt{\frac{2}{\pi}} + o_\delta(1))\sqrt{\delta}$

In the next section we show that the majority achieves the bound. and so the theorem is tight.

4 Noise sensitivity of Majority

We use the central limit theorem to prove the following.

Proposition 4.1 $\mathbb{NS}_\delta(Maj_n) \geq \Omega(\sqrt{\delta})$

Proof:[Sketch] $Maj_n(x) = sgn(\sum_i x_i) = sgn(\sum_i \frac{1}{\sqrt{n}} x_i)$. By the Central limit theorem, $\sum_i \frac{1}{\sqrt{n}} x_i$ is distributed very much like a Gaussian with mean 0 and variance 1. The pdf for a standard Gaussian is $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. So

$$\mathbf{Pr}[N(0,1) \in [-\delta/4, \delta/4]] = \int_{-\delta/4}^{\delta/4} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \geq \Omega(\delta)$$

$\therefore \mathbf{Pr}[\sum_i \frac{1}{\sqrt{n}} x_i \in [-\delta/4, \delta/4]] \geq \Omega(\delta)$.

Now suppose the above event happens, and we flip about δn of the $x_i's$. These bits are roughly $\frac{1}{2} - \frac{1}{2}$, so its like adding $-2\sum_{\approx \delta n} \pm \frac{1}{\sqrt{n}}$ which is again approximately like a Gaussian $N(0, 4\delta)$. Therefore, noise will cause a flip for Majority if

1. $|N(0,4\delta)| \geq \frac{\sqrt{\delta}}{4}$
2. The sign of $N(0,4\delta)$ goes the right way.

The first condition asks for a normal to exceed (in magnitude) a constant times the standard deviation. This has probability $\geq \Omega(1)$.

The second condition has probability $\frac{1}{2}$.

Moreover, the two conditions are independent. Therefore, both happen with a probability $\geq \Omega(1)$

So finally, removing the conditioning that $\sum_i \frac{1}{\sqrt{n}} x_i \in [-\delta/4, \delta/4]$, we get a probability $\geq \Omega(\sqrt{\delta})$. $\square$