Levin's Proof of Yao's XOR Lemma

Jaikumar Radhakrishnan*

1 Preliminaries

Definition 1.1 (Correlation of functions) *Let $f, g : \{0,1\}^n \to \{+1,-1\}$, and X be a random variable taking values in $\{0,1\}^n$. Then, the correlation of f and g (w.r.t. X) is given by*

$$\underset{X}{corr}(f,g) \stackrel{\text{def}}{=} |\,\mathbf{E}[f(X)\cdot g(X)]|.$$

Remark. Note that $corr_X(f,g) = |\Pr[f(X) = g(X)] - \Pr[f(X) \neq g(X)]|$; when the random variable X is not explicitly given, we will assume it to have uniform distribution over $\{0,1\}^n$.

Definition 1.2 (Hardness) *We say that the function $b : \{0,1,\}^n \to \{+1,-1\}$ is (p,T)-hard if for all circuits C of size T, $corr(b,C) \leq p$.*

In terms of probability, b is (p,T)-hard means that no circuit of size at most T can predict T correctly on more than $\frac{1}{2} + \frac{p}{2}$ fraction of the inputs. Here, $corr_X(b,C)$ should be thought of as the measure of how well C predicts b.

Motivation. Imagine a situation where we have a function $b : \{0,1\}^n \to \{+1,-1\}$, which is mildly unpredictable; say, it is $(1-\delta,T)$-hard for some small but non-negligible δ. This implies that that b is somewhat unpredictable. We would like to use b to produce another function b' that is very unpredictable, say we want its predictability to be smaller than some ϵ. One natural method of reducing the predictability is to XOR several copies of b. That is, we consider the function

$$b^{(t)} : (\{0,1\}^n)^t \to \{+1,-1\},$$

where

$$b^{(t)}(X_1, X_2, \ldots, X_t) \stackrel{\text{def}}{=} \prod_{i=1}^{t} b(X_i).$$

How unpredictable is $b^{(t)}$? If one believes that there is not much one can do to compute $b^{(t)}$ than compute each $b(X_i)$ separately, then it seems that the predictability of $b^{(t)}$ should fall to p^t. *Yao's XOR Lemma* roughly confirms this sentiment. Unfortunately, the proof is not straightforward. The proof given below is due to Levin, and is based on [GNW11, BH89].

Most of the ideas of the proof are contained in the special case of $t = 2$. This case is treated in the following section. For the general case we will apply this lemma repeatedly (see Section 3).

*TIFR, Mumbai. email: jaikumar@tifr.res.in

2 The XOR of two functions

Lemma 2.1 *If $b : \{0,1\}^n \to \{+1,-1\}$ is (p,T)-hard then for all $\epsilon > 0$, $b^{(2)}$ is $(p^2+\epsilon, \epsilon^2 T - O(1))$-hard.*

Proof: *We first present the main ideas of Levin's proof, pointing out the difficulties and making several assumptions (some of them invalid) along the way. Later, we will deal with these difficulties and remove the assumptions.*

Let $C(X,Y)$ be a circuit; let T' be its size. We want to show that if T' is small, then it cannot predict $b^{(2)}(X,Y)$ well. Now,

$$\begin{aligned} corr(C, b^{(2)}) &= \mathop{\mathbf{E}}_{X,Y}[b(X)b(Y)C(X,Y)] \\ &= \mathop{\mathbf{E}}_{X}[b(X)\mathop{\mathbf{E}}_{Y}[b(Y)C(X,Y)]]. \end{aligned} \tag{1}$$

We know from the unpredictability of b that for all circuits D of size at most T,

$$\mathop{\mathbf{E}}_{X}[b(X)D(X)] \leq p. \tag{2}$$

The right hand side of (2) looks remarkably like the right hand side of (1), except that we have the function $g(X) \stackrel{\text{def}}{=} \mathbf{E}_Y[b(Y)C(X,Y)]$ where we want $D(X)$. Before we use this observation, we must somehow take care of the following difficulties.

(a) The function $g(X)$ is not so easy to compute. There are two reasons for this: we need to average over 2^n possible values of Y, and this is too much for a circuit of small size to do; second, we don't know how to compute b (the hypothesis of our lemma, in fact, says that small circuits find b hard to predict!).

(b) The second difficulty concerns the type of values g takes. In (2) D is supposed to be circuit returning values in the set $\{+1,-1\}$; $g(x)$ on the other hand, gives the correlation of $b(Y)$ with the circuit $C(x,Y)$, and this is a value in the range $[+1,-1]$. In fact, our hypothesis states that if $C(X,Y)$ has size at most T, then $g(X) \in [-p,p]$. This is actually a good sign: if a correlation greater than p is not possible with circuit of size T that outputs values $\{+1,-1\}$, then, surely, with the output of the circuit scaled down by a factor of p, the correlation should become correspondingly smaller. That is, when we use (2) with $g(X)$ instead of D, then we must have an extra factor of p, because $g(X)$. Thus, we should expect the correlation with $g(X)$ to be about p^2. To exploit this observation, write

$$corr(b^{(2)}, C) = p|\mathop{\mathbf{E}}_{X}[b(X)\frac{\mathbf{E}_Y[b(Y)C(X,Y)]}{p}]|, \tag{3}$$

so that for all $x \in \{0,1\}^n$, the function $\tilde{g}(x) \stackrel{\text{def}}{=} \mathbf{E}_Y[b(Y)C(x,Y)]/p$, takes values in the range $[-1,1]$.

Assumption. Let us ignore the difficulty pointed out in remark (a) above, and go ahead. Suppose, there is a *randomized* circuit $\tilde{D}(X)$ (whose random coin tosses R comes from some distribution) such that

1. For all x, $\mathop{\mathbf{E}}_{R}[\tilde{D}(x)] = \tilde{g}(x)$;

2. The size of $\tilde{D}$ is at most T.

With this assumption, we may write

$$\begin{aligned} corr(b^{(2)}, C) &\leq p|\mathop{\mathbf{E}}_{X}[b(X)\tilde{g}(X)]| \\ &= p|\mathop{\mathbf{E}}_{X,R}[b(X)\tilde{D}(X)] \\ &= p|\mathop{\mathbf{E}}_{R}[\mathop{\mathbf{E}}_{X}[b(X)\tilde{D}(X)]]|. \end{aligned} \tag{4}$$

For each fixed choice of r of R , $\tilde{D}$ is a deterministic circuit of size at most T. Thus, if $T' \leq T$, then $|\mathbf{E}_X[b(X)\tilde{D}_{R=r}(X)]| \leq p$. Hence, by (4),

$$corr(b(2), C) \leq p\mathop{\mathbf{E}}_{R}[p] = p^2,$$

that is, $b^{(2)}$ is (p^2, T)-hard. □

Remark. This, of course, is much better than what the lemma claims. This is because of the assumption we used above. We will not be able to prove the assumption as stated, but instead the following approximate version of it.

Claim 2.1 *Suppose C is a circuit of size at most T. Then, for all δ, there is a randomized circuit $\tilde{D}$ such that*

1. *For all x, $|\mathop{\mathbf{E}}_{R}[\tilde{D}(x)] - \tilde{g}(x)| \leq \dfrac{\delta}{p}$;*

2. *The size of $\tilde{D}$ is at most $\frac{1}{\delta^2}T' + O(\frac{1}{\delta})$.*

This claim will be proved below by supplying an algorithm for approximating $\tilde{g}(x)$. First, let us see how this claim implies our lemma.

Proof of Lemma 2.1. We now return to the proof of Lemma 2.1. To apply Claim 2.1, we will assume that

$$\text{size}(C) \stackrel{\text{def}}{=} T' \leq T. \tag{5}$$

Then, using the estimate provided by the claim with ϵ for δ, we get

$$\begin{aligned} corr(b^{(2)}, C) &\leq p|\mathop{\mathbf{E}}_{X}[b(X)\tilde{g}(X)]| \\ &\leq p|\mathop{\mathbf{E}}_{X,R}[b(X)\tilde{D}(X) + b(X)(\tilde{g}(X) - \tilde{D}(X))]| \\ &\leq p|\mathop{\mathbf{E}}_{X,R}[b(X)\tilde{D}(X)]| + p|\mathop{\mathbf{E}}_{X,R}[b(X)(\tilde{D}(X) - \tilde{g}(X))]| \\ &\leq p^2 + p\mathop{\mathbf{E}}_{X}[b(X)|\mathop{\mathbf{E}}_{R}[\tilde{D}(X) - \tilde{g}(X)]|] \\ &\leq p^2 + \epsilon. \end{aligned} \tag{6}$$

To bound $|\mathbf{E}_{X,R}[b(X)\tilde{D}(X)]|$ by p in (6) we require to assume that $\tilde{D}$ has size at most T. Since by Claim 2.1, $\tilde{D}$ has size at most $\frac{1}{\epsilon^2}T' + O(\frac{1}{\epsilon})$, this implies that our argument is valid as long as

$$\text{size}(C) \stackrel{\text{def}}{=} T' \leq \epsilon^2 T - O(\epsilon). \tag{7}$$

We have, thus, shown that $b^{(2)}$ is $(p^2 + \epsilon, \epsilon^2 T - O(\epsilon))$-hard. □

2.1 Algorithm for $\tilde{D}$

Given. A function $b : \{0,1\}^n \to \{-1,1\}$ and a circuit $C(X,Y)$. Suppose for all x, $|\mathbf{E}_Y[b(Y)C(x,Y)]| \leq p$.

Task. Design a randomized circuit $\tilde{D}$ that uses random bits R (with some distribution) and meets the requirements of Claim 2.1.

Solution. Given an input x, we want to approximate the expected value of $b(Y)C(x,Y)$ as Y takes value in $\{0,1\}^n$ according to its distribution. Recall that we earlier faced some difficulties in doing this. First, how do we sum over 2^n values using a small circuit? Answer: generate a small sample of values Y and compute the expectation of $b(Y)C(x,Y)$ over this sample instead of the whole set of values. The error will go down exponentially with the size of the sample, so we can approximate $\tilde{g}(X)$ quite efficiently. But, there was another problem. Even if it is possible to generate values for Y at random, we don't know how to computer b efficiently. To circumvent this problem, instead of generating values of Y and computing $b(Y)$ ourselves, we will generate a sample of pairs $\langle Y, b(Y)\rangle$—that is, the computation of b now becomes the headache of the distribution of R.

Remark. The claim that $\tilde{D}$ is a randomized circuit should be taken with a pinch of salt—the random bits it uses, admittedly, come from quite a complicated distribution. Our argument above, however, does not suffer on account of this: we nowhere assumed that the distribution of R was easy to compute.)

Fix the sample size s, and let $\tilde{R}$ be random variable taking values in $(\{0,1\}^n \times \{1,-1\})^s$ such that

$$\Pr[\tilde{R} = \langle\langle y_1, e_1\rangle, \langle y_2, e_2\rangle, \ldots, \langle y_s, e_s\rangle\rangle] = \begin{cases} \prod_{i=1}^s \Pr[Y = y_i] & \text{if } \wedge_{i=1}^t b(y_i) = e_i \\ 0 & \text{otherwise} \end{cases}.$$

Now, the circuit $\tilde{D}$ implements the following algorithm. Let the input be x.

1. Pick $\tilde{R} = \langle\langle y_1, e_1\rangle, \langle y_2, e_2\rangle, \ldots, \langle y_s, e_s\rangle\rangle$.
2. Compute $v = \langle e_1 C(x, y_1), e_2 C(x, y_2), \ldots, e_s C(x, y_s)\rangle$. Thus, v corresponds to values for $b(Y)C(x,Y)$ for a sample of s randomly chosen values for Y. We expect the number of 1's in this list to be between $k_1 = \frac{1-p}{2}t$ and $k_2 = \frac{1-p}{2}t$. Let the actual number of 1's in v be i.
3. If $i \leq k_1$ then output -1. If $i \geq k_2$, then output 1. Otherwise, let $i = \frac{1+q}{2}s$ $(-p < q < p)$; output 1 with probability $\frac{1}{2}(1 + \frac{q}{p})$ and -1 with probability $\frac{1}{2}(1 - \frac{q}{p})$.

Remark. The random bits used by $\tilde{D}$ are $\tilde{R}$ used for generating v and the random bits used for deciding the output when the number of 1's in v is between k_1 and k_2.

Thus,

$$\mathbf{E}[\tilde{D}(x) \mid \text{number of 1's in } = i] = \begin{cases} +1 & i \geq \frac{1+p}{2}s \\ -1 & i \leq \frac{1-p}{2}s \\ \frac{2i-s}{sp} & \text{otherwise} \end{cases}.$$

Let $\alpha = \Pr_Y[b(Y)C(x,Y) = 1]$; then $\alpha \in [\frac{1-p}{2}, \frac{1+p}{2}]$, and

$$\tilde{g}(x) \stackrel{\text{def}}{=} \frac{\mathbf{E}_Y[b(Y)C(x,Y)]}{p} = \frac{2\alpha - 1}{p}, \tag{8}$$

and

$$\begin{aligned} \mathop{\mathbf{E}}_R[\tilde{D}(x)] &= -1 \cdot \sum_{i=0}^{k_1} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} \\ &\quad + \sum_{k_1 < i < k_2} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} \frac{2i-s}{sp} \\ &\quad + 1 \cdot \sum_{i=k_2}^{s} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} \end{aligned} \tag{9}$$

From this, it follows that (see below)

$$|\mathop{\mathbf{E}}_R[\tilde{D}(x)] - \tilde{g}(x)| \leq \frac{1}{p}\sqrt{\frac{1-p^2}{2\pi s}}. \tag{10}$$

Now, set $s = 1/\delta^2$. Then the error is now at most δ; furthermore, the algorithm for $\tilde{D}$ can be implemented using a circuit of size $\frac{1}{\delta^2}T' + O(\frac{1}{\delta})$. □

2.2 Deriving (10) from (9)

Boppana and Hirschfeld [BH89] give a nice derivation of (9) from (10). We present their analysis below. If the sum in the middle term of (9) is allowed to range over all values of i (from 0 to s), then we get

$$\sum_{i=0}^{s} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} \frac{2i-s}{sp} = E_I[\frac{2I-s}{sp}] = \frac{2\alpha s - s}{sp} = \frac{2\alpha - 1}{p},$$

where I is the sum of s independent 0-1 random variables, each taking value 1 with probability α. The right hand side above is precisely $\tilde{g}(x)$ (see (8) above), which $\tilde{D}(x)$ is expected to approximate. It is natural, then, to separate out this part from (9), and write

$$\begin{aligned} \mathop{\mathbf{E}}_R[\tilde{D}(x)] &= \sum_{i=0}^{k_1} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} (-1 - \frac{2i-s}{ps}) \\ &\quad + \sum_{i=0}^{s} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} \frac{2i-s}{sp} \\ &\quad + \sum_{i=k_2}^{s} \binom{s}{i} \alpha^i (1-\alpha)^{s-i} (1 - \frac{2i-s}{ps}). \end{aligned}$$

That is, the middle sum gives us precisely $\tilde{g}(x)$, and the first and last sums constitute the *error*. We will now bound this error. In fact, the first and last sums are symmetrical; the first is always positive; the last is always negative. So, it will suffice if we bound the absolute value of one of them. Let us concentrate on the last sum. We write it as

$$-\frac{2}{ps}\sum_{i=k_2}^{s}\binom{s}{i}\alpha^i(1-\alpha)^{s-i}(i-k_2). \tag{11}$$

Now, $\frac{k_1}{s} \leq \alpha \leq \frac{k_2}{s}$; for this range of values of α, (11) is always negative, and maximum in absolute value when $\alpha = \frac{k_2}{s} = \frac{1+p}{2}$. If α is fixed at $\frac{k_2}{s}$, the sum in (11) has a closed form[1].

$$\sum_{i=k_2}^{s}\binom{s}{i}\alpha^i(1-\alpha)^{s-i}(i-k_2) = s\binom{s-1}{k_2-1}\alpha^{k_2}(1-\alpha)^{s-k_2+1}.$$

Thus, the absolute value of (11) is at most

$$\frac{2}{ps}\times s\binom{s-1}{k_2-1}\alpha^{k_2}(1-\alpha)^{s-k_2+1} \leq \frac{1}{p}\sqrt{\frac{1-p^2}{2\pi s}},$$

where, for the inequality, we used $\alpha = \frac{1+p}{2}$, $k_2 = \alpha s$ and the following version of Stirling's formula due to Robbins:

$$\left(\frac{n}{e}\right)^n\sqrt{2\pi n}\times e^{1/(12n+1)} < n! < \left(\frac{n}{e}\right)^n\sqrt{2\pi n}\times e^{1/12n}.$$

[1]This closed form has a combinatorial interpretation. Let

$$f(p) = \sum_{i=k}^{s}\binom{s}{i}p^i(1-p)^{s-i}.$$

That is, $f(p)$ is the probability that in s independent trials of a 0-1 random variable that takes the value 1 with probability p, we see at least k ones. Then,

$$\frac{df}{dp} = \sum_{i=k}^{s}\binom{s}{i}p^i(1-p)^{s-i-1}(i-ps). \tag{12}$$

On the other hand, the derivative of $f(p)$ can be calculated directly. Consider s independent $0-1$ variables where the ith variable takes the value 1 with probability p_i. Let $g(p_1,\ldots,p_s)$ be the probability that the sum of these variables is at least k. Note that $f(p) = g(p,p,\ldots,p)$; thus,

$$\frac{df}{dp} = \sum_{i=1}^{s}\left.\frac{\partial g}{\partial p_i}\right|_{\forall j:p_j=p}.$$

Now, $\left.\frac{\partial g}{\partial p_i}\right|_{\forall j:p_j=p}$ is exactly the probability that the ith variable has an influence on g (probability computed over the choices of the other variables). It follows that

$$\left.\frac{\partial g}{\partial p_i}\right|_{\forall j:p_j=p} = \binom{s-1}{k-1}p^{k-1}(1-p)^{s-k},$$

and

$$\frac{df}{dp} = s\binom{s-1}{k-1}p^{k-1}(1-p)^{s-k}.$$

We have, thus, obtained a closed form expression for the sum in (12).

3 XOR of more than two functions

As stated above, the general case follows by repeated application of the special case. But, for that, we need to slightly strengthen what we proved for the XOR of two functions.

Lemma 3.1 (Isolation Lemma) *Suppose $b_1 : \{0,1\}^{n_1} \to \{+1,-1\}$ and $b_2 : \{0,1\}^{n_2} \to \{+1,-1\}$ are boolean functions, such that b_1 is (p_1, T_1)-hard and b_2 is (p_2, T_2) hard, then for all $\epsilon > 0$, $b_1 \cdot b_2$ (which is a function from $\{0,1\}^{n_1+n_2}$ to $\{+1,-1\}$) is $(p_1p_2 + \epsilon, T)$-hard, where $T = \min\{\epsilon^2 T_1 - O(1), T_2\}$.*

Proof: The proof is similar to the proof of Lemma 2.1 of the previous section. In that proof, we used the hardness of the function b twice: in assumptions (5) and (7). When we repeat the proof with b_1 and b_2 instead of b, we observe that the T in (5) now gets replaced by T_1, whereas the T in (7) becomes T_2. Thus, the argument is valid as long as the circuit predicting $b_1(x_1)b_2(x_2)$ has size at most $\min\{\epsilon^2 T_1 - O(\epsilon), T_2\}$. □

Lemma 3.2 (Yao's XOR Lemma) *If for $i = 1, 2, \ldots, \ell$, $b_i(x_i)$ is (p, T)-hard, then for all $\epsilon > 0$, the function $\prod_{i=1}^{\ell} b_i(x_i)$ is $(p^\ell + \epsilon, \epsilon^2(1-p)^2 T)$-hard.*

Proof: Let us first consider the case $\ell = 3$. By the Isolation Lemma, $b(x_2)b(x_3)$ is $(p^2 + \delta, \delta^2 T - O(1))$-hard, for all $\delta > 0$. Then, by applying the Isolation Lemma again to $b(x_1)$ and $b(x_2)b(x_3)$, we get that $b(x_1)b(x_2)b(x_3)$ is $(p^3 + p\delta + \delta, \delta^2 T - O(1))$-isolated.

In general, we can show that $\prod_{i=1}^{\ell} b_i(x_i)$, is $(p^\ell + (p^{\ell-2} + p^{\ell-3}\delta + \cdots + 1)\delta, \delta^2 T - O(1))$-hard. This implies, that $\prod_{i=1}^{\ell} b_i(x_i)$ is $(p^\ell + \delta/(1-p), \delta^2 T - O(1))$-hard. To get the claim above, set $\delta = \epsilon(1-p)$. □

References

[BH89] Ravi B. Boppana and Rafael Hirschfeld. *Pseudorandom generators and complexity classes.* In Silvio Micali, ed., *Randomness and Computation*, volume 5 of *Advances in Computing Research*, pages 1–26. JAI Press, Greenwich, Connecticut, 1989. 1, 5

[GNW11] Oded Goldreich, Noam Nisan, and Avi Wigderson. *On Yao's XOR-Lemma*. In Oded Goldreich, ed., *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, volume 6650 of *LNCS*, pages 273–301. Springer, 2011. eccc: 1995/TR95-050. 1