Lecture 4

Linearity Testing

Definition 1. (BLR Test[1]) Want to test if $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is close to linear. Pick $x, y \in \mathbb{F}_2^n$ u.a.r. Check whether $f(x+y) = f(x) + f(y)$.

Theorem 1. *For all functions f, there exists a linear function g such that $dist(f,g) := Pr_n[f(n) \neq g(n)] \leq \frac{9}{2} Pr[BLR\ rejects\ f]$.*[2]

$\mathbb{F}_2$ **Facts:**

1. $+1 = -1 \pmod 2 \implies x + y = x - y, x + y + y = x$.
2. Fix $x \in \mathbb{F}_2^n$. $y \sim \mathbb{F}_2^n$ u.a.r. $\implies x + y \sim \mathbb{F}_2^n$ u.a.r.
3. Fix $v \neq 0$. $Pr_x[\langle v, x \rangle \neq 0] = \frac{1}{2}$.

Let $g(x) := Maj_{z \in \mathbb{F}_2^n}(f(x+z) - f(z))$. Recall that if f is linear, then $f(x) = f(x+z) - f(z))$. Intuitively, if only a few places in f are not linear, then most majority votes for g will be heavily lopsided.

Claim 2. *g is a linear function if $\delta := Pr[BLR\ rejects\ f]$ is small ($\leq \frac{1}{20}$).*

Claim 3. *$dist(f,g) := Pr_n[f(n) \neq g(n)] \leq 2\delta$.*

Combining Claims 2 and 3 yields proof for Theorem 1 for small δ.

Let $P_x := Pr_y[g(x) = f(x+y) - f(y)]$, i.e. how lopsided the majority vote for g is on the input x. Observe that $P_x \geq \frac{1}{2}$ since it is always the winning majority of two candidates.

Claim 4 ("Surprising" Claim)**.** $\forall x, P_x \geq 1 - 2\delta$.

This is surprising since the BLR test, a "global" estimate of f's linearity, gives a tight bound on the "local" exactness of $g(x)$.

Proof. Let event $A(y,z) := \mathbb{1}\{f(x+y) - f(y) = f(x+z) - f(z)\}$. Consider $Pr_{y,z}[A(y,z)]$ for a fixed x. WLOG, assume $g(x) = 0$. $f(x+y) - f(y)$ and $f(x+z) - f(z)$ are each bits with bias P_x. Then, $Pr[A] = P_x P_x + (1 - P_x)(1 - P_x) = P_x^2 + (1 - P_x)^2$.

[1] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. In *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*, STOC '90, page 73–83, New York, NY, USA, 1990. Association for Computing Machinery

[2] $\frac{9}{2}$ can be eliminated by a different proof technique (Fourier Analysis).

Now, consider a different form of event A: $f(x+y)+f(x+z) = f(y)+f(z)$ (by $\mathbb{F}_2$ Fact #1, we can ignore $\pm$). By using the BLR test, we can conclude that:

$$f(y)+f(z) = f(y+z) \ w.p. \ (1-\delta),$$
$$f(x+y)+f(x+z) = f(x+y+x+z) = f(y+z) \ w.p. \ (1-\delta).$$

Then, by a naive union bound of both sides evaluating to $f(y+z)$, $Pr[A] \geq 1-2\delta$. Substituting the previous result:

$$P_x^2 + (1-P_x)^2 \geq 1-2\delta$$
$$\implies 1 + 2P_x^2 - 2P_x \geq 1-2\delta$$
$$\implies \delta \geq P_x(1-P_x) \geq \frac{1}{2}(1-P_x)$$
$$\implies P_x \geq 1-2\delta$$

where the inequality in the third line holds due to $P_x \geq \frac{1}{2}$. □

Now, we are equipped to prove Claims 2 and 3.

Proof. (Claim 2) Our goal is to prove that $g(x)+g(y) = g(x+y) \ \forall x, y$. To do so, we first construct a "magic square" that relates

$g(x)$	=	$f(x+z)$	-	$f(z)$
+		+		+
$g(y)$	=	$f(y+w)$	-	$f(w)$
=		=		=
$g(x+y)$	=	$f(x+y+z+w)$	-	$f(z+w)$

Table 1: "Magic Square"

functions f and g, as in Table . The key observation is that if all five red equalities hold, then the blue equality must also hold.

Due to the "Surprising" Claim 4, the following three equations:

$$g(x) = f(x+z) - f(z),$$
$$g(y) = f(y+w) - f(w),$$
$$g(x+y) = f(x+y+z+w) - f(z+w)$$

are satisfied with probability at least $1-2\delta$. The other two equations:

$$f(x+z) + f(y+w) = f(x+y+z+w),$$
$$f(z) + f(w) = f(z+w)$$

are satisfied with probability at least $1-\delta$ due to the BLR test.

By taking a naive union bound, the probability that all are satisfied at the same time is at least $1-8\delta$. This gives the boundary condition to have a nonzero probability of satisfying $g(x)+g(y) = g(x+y)$

be $1 - 8\delta > 0 \implies \delta < \frac{1}{8}$, which is always satisfied by a small δ specified in the claim, say $< \frac{1}{20}$. If so, we can find z, w such that all equations are satisfied, which suffice to be witnesses to prove that $g(x) + g(y) = g(x + y)$.[3] Since this scheme is not dependent on the choice of x and y, we can generalize it $\forall x, y$. □

[3] A probabilistic method for existence.

Proof. (Claim 3) By the BLR test,

$$Pr_{x,y}[f(x) \neq f(x+y) - f(y)] = Pr_{x,y}[f(x) \neq f(x+y) - f(y)] = \delta.$$

Let $BAD := \{x \mid f(x) \neq g(x)\}$, the set of inputs that f and g disagree on. Then, $Pr_y[f(x) \neq f(x+y) - f(y) \mid x \in BAD] \geq \frac{1}{2}$ since if x is in BAD, $f(x)$ must disagree with at least half of $f(x+y) - f(y)$.[4] Thus, the following inequality can be established:

[4] else, $f(x)$ would have agreed with the majority and thus agreed with $g(x)$.

$$\delta = Pr_{x,y}[f(x) \neq f(x+y) - f(y)]$$
$$\geq Pr_{x,y}[f(x) \neq f(x+y) - f(y) \mid x \in BAD] \cdot Pr_x[x \in BAD]$$
$$\geq \frac{1}{2} Pr_x[x \in BAD]$$

where the first inequality is one partition of the probability space of x and the second inequality is due to the observation above.

It is easy to see from the first and last terms that $dist(f, g) := Pr_x[f(x) \neq g(x)] = Pr_x[x \in BAD] \leq 2\delta$. □

Exponential PCP with Linearity Testing

Recap: The exponential-sized PCP testing is modeled with $QUADEQ$, where a solution $l \in \mathbb{F}_2^n$ satisfies the system of quadratic equations: $\{q_i(l) = 0 \mid i = 1 \ldots m\}$. The prover submits a proof in the form of:

- Hadamard encoding of l that admits a vector x, $\widetilde{L}(x) := \langle l, x \rangle$.
- Hadamard encoding of $H := ll^\top$ that admits a matrix C, $\widetilde{H}(C) := \sum_{i,j} C_{i,j} l_i l_j$.

where $\widetilde{L}$ and $\widetilde{H}$ represent potentially false (nonlinear) proofs.

The process is largely split into four steps:

1. Test whether $\widetilde{L}$ is ϵ-close to linear. If so, gain access to L.
2. Test whether $\widetilde{L}$ is ϵ-close to linear. If so, gain access to H.
3. Test whether H and L agree.
4. Test a random sample of the constraints with H and L.

Steps 1 and 2 are done through BLR and reject with probability $\Omega(\epsilon)$. If $\widetilde{L}$ and $\widetilde{H}$ pass the BLR test, then we can safely assume that there exists a truly linear function L and H that we can access by self-correction with success probability $\geq 1 - 2\epsilon$. For step 3, we test whether $H(xy^\top) = L(x) \cdot L(y)$ for a randomly sampled $x, y \in \mathbb{F}_2^n$. To analyze this step, we introduce another piece of useful $\mathbb{F}_2$ fact.

Another $\mathbb{F}_2$ Fact:

4. $\forall M \neq 0,\ Pr_{x,y}[x^\top My = \sum_{i,j} M_{i,j}x_iy_i \neq 0] \geq \frac{1}{4}$.

Proof. $M \neq 0 \implies \exists M_i \neq \vec{0}$ where M_i is a row of M. Since $\forall M_i \neq \vec{0},\ Pr_y[\langle M_i, y\rangle \neq 0] = \frac{1}{2}$, $My \neq \vec{0}$ w.p. $\geq \frac{1}{2}$. Then, $Pr_x[\langle x, My\rangle \neq 0] \geq Pr_{x,y}[\langle x, My\rangle \neq 0 \mid My \neq \vec{0}] \cdot Pr_y[My \neq \vec{0}] \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. □

Now, we can manipulate the testing equation in step 3 in the following way:

$$H(xy^\top) - L(x) \cdot L(y) = x^\top Hy - x^\top l \cdot l^\top y = x^\top (H - ll^\top)y = 0.$$

If $H - ll^\top$ indeed equals 0, then this test will always succeed. However, if not, by $\mathbb{F}_2$ Fact #4, $Pr_{x,y}[x^\top(H - ll^\top)y \neq 0] \geq \frac{1}{4}$. Thus, we can reject an inconsistent proof w.p. $\geq \frac{1}{4}$.

Finally, for step 4, we take a random linear combination of constraints $\{q_i\}$ and expect it to equal 0. This can just be written as some $\sum_{i,j} A_{i,j}l_il_j + \sum_i b_il_i + c = H(A) + L(b) + c = 0$. If any of $q_i(l) \neq 0$, then w.p. $\frac{1}{2}$, $Pr_l[\sum_i r_iq_i(l) \neq 0] \geq \frac{1}{2}$.

Remarks on Soundness Boosting: For linearity testing, we can actually sample $x, y \in \mathbb{F}_2^n$ multiple times to increase the chance of spotting inconsistent $\widetilde{L}$ and $\widetilde{H}$ early on. Say we sample x, y t times, a total of 6t bits. If f is ϵ-far from linear, $Pr[success] \leq (1-\epsilon)^t$. In general, for t bits, the best possible soundness is $O(\frac{1}{2^t})$.

Soundness of PCP:

Claim 5. *A PCP that reads t bits accepts a wrong proof w.p.* $O(\frac{2t}{2^t})$.

Rather than a proof, a sketch of the justification is as follows. Let $S(\epsilon)$ be the statement that $NP \subseteq PCP(t, poly(n), \epsilon)$. This is equivalent to saying that there exists a CSP with constraints on t bits such that it is NP-Hard to approximate better than a ϵ-factor. Also, for $\epsilon = \frac{2t}{2^t}$, if $S(\epsilon)$ is true, then for all CSPs with constraint on t bits, there exists a $\frac{2t}{2^t}$-factor approximation algorithm.

Bibliography

[BLR90] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. In *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*, STOC '90, page 73–83, New York, NY, USA, 1990. Association for Computing Machinery.