Lecture 23: Few Applications (BLR Test, LHL)

BLR Linearity Testing I

- We shall consider the following definition of linear functions

Definition (Linear Function)

Let $f\colon \{0,1\}^n \to \{+1,-1\}$ be a boolean function. If $f(x) \cdot f(y) = f(x+y)$, for all $x, y \in \{0,1\}^n$, then the function f is a linear function.

- Note that χ_S is a linear function for all $S \in \{0,1\}^n$. In fact, $\{\chi_0, \dots, \chi_{N-1}\}$ is the set of all linear functions.
- Suppose a function f is provided to us as an oracle. We are interested in testing whether it is close-to some linear function. That it, does there exists S such that f and χ_S agree on a large number of inputs, i.e., $f(x) = \chi_S(x)$ for a large fraction of $x \in \{0,1\}^n$

BLR Linearity Testing II

- Blum–Luby–Rubinfeld provided an algorithm to correctly test this property using only two queries to the f-oracle. This algorithm is known as the BLR linearity testing algorithm
- Here is the pseudo-code of the algorithm

BLR^f: • Pick random $x, y \in \{0,1\}^n$ and query f to obtain $u = f(x)$, $v = f(y)$, and $w = f(x+y)$ • Output true if $u \cdot v == w$

- So, the algorithm is simple. Let us analyze the performance of this algorithm
- We want to claim that "if the algorithm returns true with high probability then the function f agrees with some χ_S with high probability"

- We make the following claim

Lemma

The probability that our algorithm outputs true *is*

$$\frac{1 + \sum_{S \in \{0,1\}^n} \widehat{f}(S)^3}{2}$$

Proof Outline.

- Note that the algorithm says true when $f(x) \cdot f(y) == f(x+y)$. That is, $f(x) \cdot f(y) \cdot f(x+y) = 1$, because the range of f is $\{+1, -1\}$.
- And, similarly, our algorithm says false when $f(x) \cdot f(y) \cdot f(x+y) = -1$.

BLR Linearity Testing IV

- Therefore, we can conclude that

$$\frac{1}{N^2} \sum_{x,y \in \{0,1\}^n} f(x)f(y)f(x+y) = p - (1-p),$$

  where p is the probability that our algorithm says true
- So, to prove the lemma, it suffices to prove that

$$\frac{1}{N^2} \sum_{x,y \in \{0,1\}^n} f(x)f(y)f(x+y) = \sum_{S \in \{0,1\}^n} \widehat{f}(S)^3$$

BLR Linearity Testing V

- Let us prove this

$$
\begin{aligned}
\frac{1}{N^2}\sum_{x,y\in\{0,1\}^n} f(x)f(y)f(x+y) &= \frac{1}{N}\sum_{z\in\{0,1\}^n}\left(\frac{1}{N}\sum_{x\in\{0,1\}^n} f(x)f(z-x)\right)f(z)\\
&= \frac{1}{N}\sum_{z\in\{0,1\}^n}(f*f)(z)\cdot f(z)\\
&= \langle f*f, f\rangle\\
&= \sum_{S\in\{0,1\}^n}\widehat{(f*f)}(S)\cdot\widehat{f}(S)\\
&= \sum_{S\in\{0,1\}^n}\widehat{f}(S)^3
\end{aligned}
$$

BLR Linearity Testing VI

- Okay, back to our main proof now. Suppose p is the probability that our algorithm outputs true. If $p \geqslant 1 - \varepsilon$, then, from the lemma above, we have

$$\sum_{S \in \{0,1\}^n} \widehat{f}(S)^3 \geqslant 1 - 2\varepsilon$$

- Note that Parseval's identity on f implies that

$$\sum_{S \in \{0,1\}^n} \widehat{f}(S)^2 = \langle f, f \rangle = 1,$$

because the range of f is $\{+1, -1\}$

BLR Linearity Testing VII

- So, we are given two guarantees

$$\sum_{S\in\{0,1\}^n} \widehat{f}(S)^2 = 1$$

$$\sum_{S\in\{0,1\}^n} \widehat{f}(S)^3 \geqslant 1 - 2\varepsilon$$

  We need to prove that $\max_{S\in\{0,1\}^n} \widehat{f}(S)$ is close to 1

- We prove the following result

Lemma

If $\sum_{S\in\{0,1\}^n} \widehat{f}(S)^2 = 1$ *and* $\sum_{S\in\{0,1\}^n} \widehat{f}(S)^3 \geqslant 1 - 2\varepsilon$ *then we have* $\max_{S\in\{0,1\}^n} \widehat{f}(S) \geqslant 1 - 2\varepsilon$.

Proof Outline.

$$\max_{S\in\{0,1\}^n} \widehat{f}(S) = \left(\max_{S\in\{0,1\}^n} \widehat{f}(S)\right)\left(\sum_{S\in\{0,1\}^n} \widehat{f}(S)^2\right)$$
$$\geqslant \sum_{S\in\{0,1\}^n} \widehat{f}(S)^3$$
$$\geqslant 1-2\varepsilon$$

- So, let us recall what we have proven. If the algorithm outputs true with probability $\geqslant (1-\varepsilon)$, then there exists S such that $\widehat{f}(S) \geqslant 1-2\varepsilon$.
- Recall that if q is the probability that f and χ_S agree then we have $\langle f, \chi_S\rangle = q-(1-q)$. So, $q \geqslant 1-\varepsilon$, because $\langle f, \chi_S\rangle = \widehat{f}(S)$.

BLR Linearity Testing IX

- Thus, we conclude that if the algorithm outputs true with probability $p \geqslant (1-\varepsilon)$ then f agrees with some χ_S with probability $q \geqslant p \geqslant (1-\varepsilon)$.

Left-over Hash Lemma I

- We need to introduce the definition of a family of universal hash functions.

Definition (Universal Hash Function Family)

Let $\mathcal{H} = \{h_1, \ldots, h_\alpha\}$ be a set of $\{0,1\}^n \to \{0,1\}^m$ functions such that for any distinct $x, x' \in \{0,1\}^n$ we have

$$\mathbb{P}\left[h(x) = h(x') \colon h \xleftarrow{\$} \mathcal{H}\right] \leqslant \frac{1}{2^m}$$

- Recall that $\mathbb{X}$ has min-entropy k if $\mathbb{P}[\mathbb{X} = x] \leqslant 2^{-k}$ for any x in the sample space

- Left-over Hash Lemma (LHL) states the following. The statistical distance between the distributions $(\mathbb{H}(\mathbb{X}), \mathbb{H})$ and $(\mathbb{U}, \mathbb{H})$ is small, where $\mathbb{U}$ is a uniform distribution over $\{0,1\}^m$ and $\mathbb{H}$ is the uniform distribution over $\mathcal{H}$. Formally, it states the following

Lemma (LHL)

Let $\mathbb{H}$ be a uniform distribution over $\mathcal{H}$, a universal hash function family $\{0,1\}^n \to \{0,1\}^m$, and $\mathbb{X}$ is a random variable over $\{0,1\}^n$. Then, the following holds

$$\mathrm{SD}\left((\mathbb{H}(\mathbb{X}), \mathbb{H}), (\mathbb{U}, \mathbb{H})\right) \leqslant \frac{1}{2}\sqrt{\frac{2^m}{2^{\mathrm{H}_\infty(\mathbb{X})}}}$$

Proof Outline.

Left-over Hash Lemma III

- We begin with some simplification

$$\begin{aligned}
\mathrm{SD}\left((\mathbb{H}(\mathbb{X}),\mathbb{H}),(\mathbb{U},\mathbb{H})\right) &= \mathbb{E}\left[\mathrm{SD}\left(h(\mathbb{X}),\mathbb{U}\right) : h \sim \mathbb{H}\right] \\
&\leqslant \mathbb{E}\left[\frac{M}{2}\sqrt{\sum_{S\in\{0,1\}^m} \widehat{h(\mathbb{X})}(S)^2 - \widehat{h(\mathbb{X})}(0)^2} : h \sim \mathbb{H}\right] \\
&\leqslant \frac{M}{2}\sqrt{\mathbb{E}\left[\sum_{S\in\{0,1\}^m} \widehat{h(\mathbb{X})}(S)^2 - \frac{1}{M^2} : h \sim \mathbb{H}\right]}, && \text{Jensen} \\
&\leqslant \frac{M}{2}\sqrt{\mathbb{E}\left[\sum_{S\in\{0,1\}^m} \widehat{h(\mathbb{X})}(S)^2 : h \sim \mathbb{H}\right] - \frac{1}{M^2}} \\
&= \frac{M}{2}\sqrt{\mathbb{E}\left[\left\langle h(\mathbb{X}), h(\mathbb{X})\right\rangle : h \sim \mathbb{H}\right] - \frac{1}{M^2}}, && \text{Parseval} \\
&= \frac{1}{2}\sqrt{M\mathbb{E}\left[\mathrm{col}(h(\mathbb{X})) : h \sim \mathbb{H}\right] - 1}
\end{aligned}$$

- So, we need to estimate $\mathbb{E}\left[\mathrm{col}(h(\mathbb{X}))\colon h \sim \mathbb{H}\right]$. Note that this is equivalent to the probability that we sample two independent samples $x \sim \mathbb{X}$ and $x' \sim \mathbb{X}$ and it turns out that $h(x) = h(x')$, for $h \sim \mathbb{H}$. That is, the following expression

$$\mathbb{E}\left[\mathbf{1}_{\{h(x)=h(x')\}} \colon x \sim \mathbb{X}, x' \sim \mathbb{X}, h \sim \mathbb{H}\right]$$

- Note that if $x = x'$, then we shall definitely have $h(x) = h(x')$ irrespective of the value of h.
- If $x \neq x'$ then the probability that $h(x) = h(x')$ is $\leqslant \frac{1}{M}$, for a random $h \sim \mathbb{H}$
- To use these two observations, we proceed formally as follows. We write

$$\mathbf{1}_{\{h(x)=h(x')\}} = \mathbf{1}_{\{x=x'\}} + \mathbf{1}_{\{(x\neq x') \ \wedge \ (h(x)=h(x'))\}}$$

So, we have

$$\mathbb{E}\left[\mathbf{1}_{\{h(x)=h(x')\}}\right] = \mathbb{E}\left[\mathbf{1}_{\{x=x'\}}\right] + \mathbb{E}\left[\mathbf{1}_{\{(x\neq x') \ \wedge \ (h(x)=h(x'))\}}\right]$$

Left-over Hash Lemma V

- Let p be the collision probability of the random variable $\mathbb{X}$. We know that $p \leqslant \frac{1}{K}$, where k is the min-entropy of $\mathbb{X}$. So, we have $\mathbb{E}\left[\mathbf{1}_{\{x=x'\}}\right] = p$.
- And, by universal hash function family guarantee of $\mathcal{H}$, we have

$$\mathbb{E}\left[\mathbf{1}_{\{(x\neq x')\ \wedge\ (h(x)=h(x'))\}}\right] \leqslant (1-p)\frac{1}{M}$$

- So, we have

$$\mathbb{E}\left[\mathrm{col}(h(\mathbb{X}))\colon h\sim\mathbb{H}\right] \leqslant p + \frac{(1-p)}{M} < \frac{1}{K} + \frac{1}{M}$$

- Now, going back to our original inequality

$$\begin{aligned}\mathrm{SD}\left((\mathbb{H}(\mathbb{X}),\mathbb{H}),(\mathbb{U},\mathbb{H})\right) &\leqslant \frac{1}{2}\sqrt{M\mathbb{E}\left[\mathrm{col}(h(\mathbb{X}))\colon h\sim\mathbb{H}\right]-1} \\ &< \frac{1}{2}\sqrt{\frac{M}{K}}\end{aligned}$$