## 2.2 Lecture 9 (22/Oct): Fingerprinting with Linear Algebra

### 2.2.1 Verifying Associativity

Let a set $S$ of size $n$ be given, along with a binary operation $\circ : S \times S \to S$. Thus the input is a table of size $n^2$; we call the input $(S, \circ)$. The problem we consider is testing whether the operation is associative, that is, whether for all $a, b, c \in S$,

$$(a \circ b) \circ c = a \circ (b \circ c) \tag{2.1}$$

A triple for which (2.1) fails is said to be a nonassociative triple.

No sub-cubic-time deterministic algorithm is known for this problem. However,

**Theorem 29 (Rajagopalan & Schulman [83])** *There is an $O(n^2)$-time co-RP type algorithm for associativity.*

**Proof:** An obvious idea is to replace the $O(n^3)$-time exhaustive search for a nonassociative triple, by randomly sampling triples and and checking them. The runtime required is inverse to the fraction of nonassociative triples, so this method would improve on exhaustive search if we were guaranteed that a nonassociative operation had a super-constant number of nonassociative triples. However, for every $n \geq 3$ there exist nonassociative operations with only a single nonassociative triple.

So we'll have to do something more interesting.

Let's define a binary operation $(\mathbb{S}, \circ)$ on a much *bigger* set $\mathbb{S}$. Define $\mathbb{S}$ to be the vector space with basis $S$ over the field $\mathbb{Z}/2$, that is to say, an element $x \in \mathbb{S}$ is a formal sum

$$x = \sum_{a \in S} a x_a \quad \text{for } x_a \in \mathbb{Z}/2$$

The product of two such elements $x, y$ is

$$\begin{aligned} x \circ y &= \sum_{a \in S} \sum_{b \in S} (a \circ b) x_a y_b \\ &= \sum_{c \in S} c \bigoplus_{a,b : a \circ b = c} x_a y_b \end{aligned}$$

where of course $\bigoplus$ denotes sum mod 2.

On $(\mathbb{S}, \circ)$ we have an operation that we do not have on $(S, \circ)$, namely, addition:

$$x + y = \sum_{a \in S} a(x_a + y_b)$$

(Those who have seen such constructions before will recognize $(\mathbb{S}, \circ)$ as an "algebra" of $(S, \circ)$ over $\mathbb{Z}/2$.)

The algorithm is now simple: *check the associative identity for three random elements of* $\mathbb{S}$. That is, select $x, y, z$ u.a.r. in $\mathbb{S}$. If $(x \circ y) \circ z) = x \circ (y \circ z)$, report that $(S, \circ)$ is associative, otherwise report that it is not associative. The runtime for this process is clearly $O(n^2)$.

If $(S, \circ)$ is associative then so is $(\mathbb{S}, \circ)$, because then $(x \circ y) \circ z$ and $x \circ (y \circ z)$ have identical expansions as sums. Also, nonnassociativity of $(S, \circ)$ implies nonnassociativity of $(\mathbb{S}, \circ)$ by simply considering "singleton" vectors within the latter.

But this equivalence is not enough. The crux of the argument is the following:

**Lemma 30** *If $(S, \circ)$ is nonnassociative then at least one eighth of the triples $(x, y, z)$ in $\mathbb{S}$ are nonassociative triples.*

**Proof:** The proof relies on a variation on the inclusion-exclusion principle.

For any triple $a, b, c \in S$, let

$$g(a, b, c) = (a \circ b) \circ c - a \circ (b \circ c).$$

Note that $g$ is a mapping $g : S^3 \to \mathbb{S}$. Now extend $g$ to $g : \mathbb{S}^3 \to \mathbb{S}$ by:

$$g(x, y, z) = \sum_{a,b,c} g(a, b, c) x_a y_b z_c$$

If you imagine the $n \times n \times n$ cube indexed by $S^3$, with each position $(a, b, c)$ filled with the entry $g(a, b, c)$, then $g(x, y, z)$ is the sum of the entries in the combinatorial subcube of positions where $x_a = 1, y_b = 1, z_c = 1$. (We say "combinatorial" only to emphasize that unlike a physical cube, here the slices that participate in the subcube are not in any sense adjacent.)

Fix $(a', b', c')$ to be any nonassociative triple of $S$.

Partition $\mathbb{S}^3$ into blocks of eight triples apiece, as follows. Each of these blocks is indexed by a triple $x, y, z$ s.t. $x_{a'} = 0, y_{b'} = 0, z_{c'} = 0$. The eight triples are $(x + \varepsilon_1 a', y + \varepsilon_2 b', z + \varepsilon_3 c')$ where $\varepsilon_i \in \{0, 1\}$.

Now observe that

$$\sum_{\varepsilon_1, \varepsilon_2, \varepsilon_3} g(x + \varepsilon_1 a', y + \varepsilon_2 b', z + \varepsilon_3 c') = g(a', b', c')$$

To see this, note that each of the eight terms on the LHS is, as described above, a sum of the entries in a "subcube" of the "$S^3$ cube". These subcubes are closely related: there is a core subcube whose indicator function is $x \times y \times z$, and all entries of this subcube are summed within all eight terms. Then there are additional width-1 pieces: the entries in the region $a' \times y \times z$ occur in four terms, as do the regions $x \times b' \times z$ and $x \times y \times c'$. The entries in the regions $a' \times b' \times z$, $a' \times y \times c'$ and $x \times b' \times c'$ occur in two terms, and the entry in the region $a' \times b' \times c'$ occurs in one term.

Since the RHS is nonzero, so is at least one of the eight terms on the LHS. □ □

Corollary: in time $O(n^2)$ we can sample $x, y, z$ u.a.r. in $\mathbb{S}$ and determine whether $(x \circ y) \circ z = x \circ (y \circ z)$. If $(S, \circ)$ is associative, then we get $=$; if $(S, \circ)$ is nonassociative, we get $\neq$ with probability $\geq 1/8$.