Supplementary Lecture H

Parikh's Theorem

Here is a theorem that says a little more about the structure of CFLs. It says that for any CFL A, if we look only at the relative number of occurrences of terminal symbols in strings in A without regard to their order, then A is indistinguishable from a regular set.

Formally, let $\Sigma = \{a_1, \ldots, a_k\}$. The *Parikh map* is the function

$$\psi : \Sigma^* \to \mathbb{N}^k$$

defined by

$$\psi(x) \stackrel{\text{def}}{=} (\#a_1(x), \#a_2(x), \ldots, \#a_k(x)).$$

That is, $\psi(x)$ records the number of occurrences of each symbol in x. The structure Σ^* with binary operation $\cdot$ (concatenation) and constant ϵ forms a monoid,[1] as does the structure $\mathbb{N}^k$ with binary operation $+$ (componentwise addition) and identity $\overline{0} = (0, \ldots, 0)$, and ψ is a monoid homomorphism:

$$\begin{aligned} \psi(xy) &= \psi(x) + \psi(y), \\ \psi(\epsilon) &= \overline{0}. \end{aligned}$$

[1] Recall from Lecture 2 that a *monoid* is an algebraic structure consisting of a set with an associative binary operation and an identity for that operation.

The main difference between the monoids Σ^* and $\mathbb{N}^k$ is that the latter is commutative,[2] whereas the former is not, except in the case $k = 1$. In fact, if $\equiv$ is the smallest monoid congruence[3] on Σ^* such that $a_i a_j \equiv a_j a_i$, $1 \leq i, j \leq k$, then $\mathbb{N}^k$ is isomorphic to the quotient[4] $\Sigma^*/\equiv$. The monoid Σ^* is sometimes called the *free monoid on k generators*, and $\mathbb{N}^k$ is sometimes called the *free commutative monoid on k generators*. The word "free" refers to the fact that the structures do not satisfy any equations besides the logical consequences of the monoid or commutative monoid axioms, respectively.

The *commutative image* of a set $A \subseteq \Sigma^*$ is its image under ψ:

$$\psi(A) \stackrel{\text{def}}{=} \{\psi(x) \mid x \in A\}.$$

If $u_1, \ldots, u_m \in \mathbb{N}^k$, the submonoid of $\mathbb{N}^k$ generated by $u_1, \ldots, u_m$ is denoted $<u_1, \ldots, u_m>$. This is the smallest subset of $\mathbb{N}^k$ containing $u_1, \ldots, u_m$ and the monoid identity $\overline{0}$ and closed under $+$. Equivalently,

$$<u_1, \ldots, u_m> = \{a_1 u_1 + \cdots + a_m u_m \mid a_1, \ldots, a_m \in \mathbb{N}\} \subseteq \mathbb{N}^k.$$

A subset of $\mathbb{N}^k$ is called *linear* if it is a coset of such a finitely generated submonoid; that is, if it is of the form

$$u_0 + <u_1, \ldots, u_m> = \{u_0 + a_1 u_1 + \cdots + a_m u_m \mid a_1, \ldots, a_m \in \mathbb{N}\}.$$

A subset of $\mathbb{N}^k$ is called *semilinear* if it is a union of finitely many linear sets. For example,

$$\begin{aligned} \psi(\{a^n b^n \mid n \geq 0\}) &= \psi(\{x \in \{a,b\}^* \mid \#a(x) = \#b(x)\}) \\ &= \{(n,n) \mid n \geq 0\} \\ &= <(1,1)> \end{aligned}$$

is a semilinear (in fact, linear) subset of $\mathbb{N}^2$, but $\{(n, n^2) \mid n \geq 0\}$ is not.

Theorem H.1 **(Parikh)** *For any context-free language A, $\psi(A)$ is semilinear.*

The converse does not hold: the set $\{a^n b^n c^n \mid n \geq 0\}$ is not context-free but has a semilinear image under ψ. This is also the image of the CFL $\{(ab)^n c^n \mid n \geq 0\}$ and the regular set $(abc)^*$.

---

[2] A monoid is *commutative* if $xy = yx$ for all x and y.

[3] A *monoid congruence* is an equivalence relation $\equiv$ on the monoid that respects the monoid structure in the sense that if $x \equiv x'$ and $y \equiv y'$, then $xy \equiv x'y'$.

[4] The *quotient* of a monoid M by a congruence $\equiv$ is a monoid whose elements are the congruence classes $[x] \stackrel{\text{def}}{=} \{y \mid y \equiv x\}$, binary operation $[x] \cdot [y] \stackrel{\text{def}}{=} [xy]$, and constant $[1]$, where 1 is the identity of M. See Supplementary Lectures C and D for more information on these concepts.

For every semilinear set $S \subseteq \mathbb{N}^k$, it is not hard to construct a regular set $R \subseteq \Sigma^*$ such that $\psi(R) = S$. For example, $\{(n,n) \mid n \geq 0\} = \psi((ab)^*)$. For this reason, Parikh's theorem is sometimes stated as follows:

Every context-free language is letter-equivalent to a regular set,

where letter-equivalence means the sets have the same commutative image.

In order to prove Parikh's theorem, we need some definitions. Let $G = (N, \Sigma, P, S)$ be an arbitrary CFG in Chomsky normal form. Let $s, t, \ldots$ denote parse trees of G with a nonterminal at the root, nonterminals labeling the internal nodes, and terminals or nonterminals labeling the leaves. Define

$\mathbf{root}(s) \stackrel{\text{def}}{=}$ the nonterminal at the root of s;

$\mathbf{yield}(s) \stackrel{\text{def}}{=}$ the string of terminals and nonterminals at the leaves of s, reading left to right;

$\mathbf{depth}(s) \stackrel{\text{def}}{=}$ the length of the longest path in s from a leaf up to the root (the *length* of a path is the number of edges, or the number of nodes less one);

$N(s) \stackrel{\text{def}}{=}$ the set of nonterminals appearing in s.

Define a *pump* to be a parse tree s such that

(i) s contains at least two nodes; and

(ii) $\mathbf{yield}(s) = x \cdot \mathbf{root}(s) \cdot y$ for some $x, y \in \Sigma^*$; that is, all leaves are labeled with terminal symbols except one, and the nonterminal labeling that leaf is the same as the one labeling the root.

These objects arose in the proof of the pumping lemma for CFLs in Lecture 22: wherever a nonterminal A appears in a parse tree s, the tree can be split apart at that point and a pump u with $\mathbf{root}(u) = A$ inserted to get a larger parse tree t.

For parse trees s, t, define $s \lhd t$ if t can be obtained from s by splitting s at a node labeled with some nonterminal A and inserting a pump with root labeled A. The relation $\lhd$ is not a partial order (it is not reflexive or transitive), but it is *well founded* in the sense that there exists no infinite descending chain $s_0 \rhd s_1 \rhd s_2 \rhd \cdots$, because if $s \lhd t$, then s has strictly fewer nodes than t.

Define a pump t to be a *basic pump* if it is $\lhd$-minimal among all pumps; that is, if it does not properly contain another pump that could be cut out. In other words, a pump t is a *basic pump* if the only s such that $s \lhd t$ is

the trivial one-node parse tree labeled with the nonterminal $\mathbf{root}(t)$. Basic pumps cannot be too big:

Lemma H.2 *If s is a basic pump, then $\mathbf{depth}(s) \leq 2n$, where n is the number of nonterminals in N.*

Proof. Let π denote the path in s from the unique leaf with label $\mathbf{root}(s)$ up to the root. The path π can be no longer than n, because if it were, it would have a repeated nonterminal and would therefore contain a pump that could be removed, contradicting the $\lhd$-minimality of s. For any other leaf, the path from that leaf up to the first node on the path π can be no longer than $n+1$ for the same reason. Thus the total length of any path from a leaf to the root can be no longer than $2n$. □

It follows from Lemma H.2 and the fact that there are only finitely many productions in G that the number of basic pumps is finite, say p.

Lemma H.3 *Every parse tree t with $\mathbf{yield}(t) \in \Sigma^*$ is either $\lhd$-minimal or contains a basic pump.*

Proof. If t is not $\lhd$-minimal, then by definition it contains a pump s. Let s be $\lhd$-minimal among all pumps contained in t. Then s is a basic pump, because if it were not, then it would contain a smaller pump u, and u would be a smaller pump contained in t, contradicting the minimality of s. □

Define $s \leq t$ if t can be obtained from s by some finite sequence of insertions of basic pumps u such that $N(u) \subseteq N(s)$. In other words, starting from s, we are allowed to choose any occurrence of a nonterminal A in s and insert a basic pump u with $\mathbf{root}(u) = A$ at that point, provided u contains no new nonterminals that are not already contained in s. If t can be obtained from s by a finite number of repetitions of this process, then $s \leq t$.

If $\alpha \in (N \cup \Sigma)^*$, define $\psi(\alpha) = \psi(x)$, where x is the string obtained from α by deleting all nonterminals. Let $\psi(t)$ abbreviate $\psi(\mathbf{yield}(t))$.

Lemma H.4 *The set $\{\psi(t) \mid s \leq t\}$ is linear.*

Proof.

$$\{\psi(t) \mid s \leq t\} = \psi(s) + <\{\psi(u) \mid u \text{ is a basic pump with } N(u) \subseteq N(s)\}>.$$

□

Lemma H.5 *If s is $\leq$-minimal, then $\mathbf{depth}(s) \leq (p+1)(n+1)$, where p is the number of distinct basic pumps and n is the size of N.*

Proof. If s had a path longer than $\mathbf{depth}(s) \leq (p+1)(n+1)$, then that path could be broken up into $p+1$ segments, each of length at least $n+1$, and each segment would have a repeated nonterminal. Then there would be

$p + 1$ disjoint pumps. (Two pumps are considered *disjoint* if they have no nodes in common, or if the root of one is a leaf of the other.) Each of these $p + 1$ pumps either is basic or contains a basic pump by Lemma H.3; thus there would be $p + 1$ disjoint basic pumps. But there are only p distinct basic pumps in all, so by the pigeonhole principle there must be two disjoint occurrences of the same basic pump. But this contradicts the $\leq$-minimality of s, since one of these basic pumps could be deleted without changing the set of nonterminals contained in the tree. □

Proof of Theorem H.1. Let

$$M = \{s \mid s \text{ is } \leq\text{-minimal}, \mathbf{root}(s) = S, \mathbf{yield}(s) \in \Sigma^*\}.$$

We show that

$$\psi(L(G)) = \bigcup_{s \in M} \{\psi(t) \mid s \leq t\}.$$

This set is semilinear by Lemma H.5, which implies that M is finite, and by Lemma H.4. Any t such that $s \leq t$ for some $s \in M$ has $\mathbf{root}(t) = S$ and $\mathbf{yield}(t) \in \Sigma^*$; thus $\mathbf{yield}(t) \in L(G)$ and $\psi(t) \in \psi(L(G))$. Conversely, any string $x \in L(G)$ has a parse tree t with $\mathbf{root}(t) = S$ and $\mathbf{yield}(t) = x$, and there must exist a $\leq$-minimal $s \leq t$. Then $s \in M$ and

$$\psi(x) \in \{\psi(t) \mid s \leq t\}.$$ □

Historical Notes

Parikh's theorem was first proved by Rohit Parikh [98]. Alternative proofs have been given by Goldstine [52], Harrison [55], and Kuich [75].