Simple proof of Parikh's theorem *à la* Takahashi

Ryoma Sin'ya

Akita University
ryoma@math.akita-u.ac.jp

Abstract. In this report we describe a simple proof of Parikh's theorem *à la* Takahashi, based on a decomposition of derivation trees. The idea of decomposition is appeared in her master's thesis written in 1970.

1 Preliminaries

For a set S, we denote by $|S|$ the cardinality of S. The set of natural numbers including 0 is denoted by $\mathbb{N}$. Let $G = (V, D, X_0)$ be a context-free grammar over an alphabet A where V $(V \cap A = \emptyset)$ is a finite set of *non-terminals*, $D \subseteq V \times (V \cup A \cup \{\epsilon\})^+$ is a finite set of *derivation rules*, and $X_0 \in V$. The set of (V, A)*-trees*, ranged over by T, is given by the following grammar:

$$T ::= a \ \ (a \in A \cup \{\epsilon\}) \mid X(T_1, \cdots, T_n) \ \ (X \in V, n \geq 1)$$

Namely, (V, A)-trees are trees whose internal nodes are non-terminals, and whose leaves are letters in A or the special symbol $\epsilon \notin A$. For a (V, A)-tree T, we denote by $N(T)$ the set of all non-terminals appeared in T, and denote by $R(T)$ the root of T. The *yield* Y is a function from (V, A)-trees into A^* defined inductively as $Y(a) = a, Y(\epsilon) = \varepsilon$ where ε is the empty string, and $Y(X(T_1, \ldots, T_n)) = Y(T_1) \cdots Y(T_n)$. We call a $(V, A \cup \{[\,]\})$-tree C *context* if exactly one leaf of C is the special symbol $[\,] \notin A$.We denote by $C[T]$ the (V, A)-tree obtained by replacing $[\,]$ in C by T. We define *the set* $\mathcal{T}(G)$ *of derivation trees of* G as

$$\begin{aligned}\mathcal{T}(G) \triangleq \{T \colon (V, A)\text{-tree} \mid R(T) = X_0,\ \text{for each context } C,\\ T = C[X(T_1, \ldots, T_n)] \text{ implies } (X, R(T_1) \cdots R(T_n)) \in D\}\end{aligned}$$

and define $\mathcal{L}(G) \triangleq \{Y(T) \mid T \in \mathcal{T}(G)\}$.

For a non-terminal $X \in V$, we call a $(V, A \cup \{X\})$-tree $\alpha \neq X$ an *adjunct tree* if $R(\alpha) = X$ and exactly one leaf of α is X. For a (V, A)-tree T and an adjunct tree α such that $R(T) = R(\alpha)$, we denote by $\alpha[T]$ the (V, A)-tree obtained by replacing the leaf X in α by T. For a (V, A)-tree T and an adjunct tree α, if the root X of α is appeared in T, *i.e.*, $T = C[X(T_1, \ldots, T_n)]$ for some context C and (V, A)-trees $T_1, \ldots, T_n$, we say that α *is adjoinable to* T, and we say that $T' = C[\alpha[X(T_1, \ldots, T_n)]]$ *is obtained from* T *adjoining* α and write $T \vdash_\alpha T'$. Intuitively, an adjunct tree represents "pump" part, and adjoining corresponds to "pumping" operation for trees. For example, $X(Y(X), a)$ is adjoinable to $Z(X(b))$ and we have $Z(X(b)) \vdash_{X(Y(X),a)} Z(X(Y(X(b)), a))$.

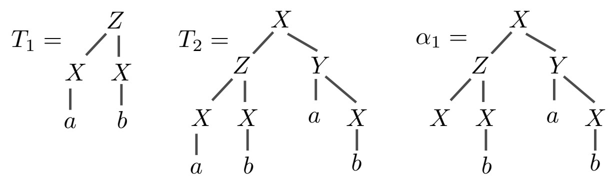

Fig. 1. Example of simple (V, A)-tree T_1, non-simple (V, A)-tree T_2, and simple adjunct tree α_1

We call a (V, A)-tree T *simple* if, for any path in T from the root to a leaf, no non-terminal appears more than once. We call an adjunct tree α *simple* if, for any path in T from a child of the root to a leaf, no non-terminal appears more than once. See Fig. 1 for example. T_1 is simple since all paths $\{(Z, X, a), (Z, X, b)\}$ contain Z and X exactly once. T_2 is not simple since the left-most path (X, Z, X, a) contains X twice. However, the adjunct tree α_1, which is obtained by removing the left-most leave a from T_2 (*i.e.*, $X(a) \vdash_{\alpha_1} T_2$), is simple since all paths from a child of the root to a leaf $\{(Z, X), (Z, X, b), (Y, a), (Y, X, b)\}$ contain no non-terminal more than once.

For a (V, A)-tree T and a set of adjunct trees S, we define

$$\mathrm{Adj}^*(T, S) \triangleq \{T' \mid T = T_0 \vdash_{\alpha_1} T_1 \vdash_{\alpha_2} \cdots \vdash_{\alpha_k} T_k = T', k \in \mathbb{N}, \{\alpha_1, \ldots, \alpha_k\} \subseteq S\}$$
$$\mathrm{Adj}^+(T, S) \triangleq \{T' \mid T = T_0 \vdash_{\alpha_1} T_1 \vdash_{\alpha_2} \cdots \vdash_{\alpha_k} T_k = T', k \in \mathbb{N}, \{\alpha_1, \ldots, \alpha_k\} = S\}$$

Intuitively, $\mathrm{Adj}^*(T, S)$ (resp. $\mathrm{Adj}^+(T, S)$) is the set of all (V, A)-trees obtained from T adjoining each element in S arbitrary number of times (resp. arbitrary *positive* number of times). Clearly, $\mathrm{Adj}^*(T, S) = \bigcup_{U \subseteq S} \mathrm{Adj}^+(T, U)$ and $\mathrm{Adj}^+(T, \emptyset) = \{T\}$. We say that S *is adjoinable to* T if $\mathrm{Adj}^+(T, S)$ is non-empty. Notice that if $\mathrm{Adj}^+(T, S)$ is non-empty then there exists $T' \in \mathrm{Adj}^+(T, S)$ such that T' is obtained from T adjoining each element in S *exactly once*, *i.e.*, $T_0 = T \vdash_{\alpha_1} T_1 \vdash_{\alpha_2} \cdots \vdash_{\alpha_{|S|}} T_{|S|} = T'$ and $S = \{\alpha_1, \ldots, \alpha_{|S|}\}$. Moreover, such $T' \in \mathrm{Adj}^+(T, S)$ contains every root non-terminal of $\alpha \in S$, $\mathrm{Adj}^+(T', S)$ is also non-empty and thus $\mathrm{Adj}^+(T, S)$ should be infinite (if S is non-empty).

Let $A = \{a_1, \cdots, a_d\}$. The *Parikh mapping* $\Phi_A : A^* \to \mathbb{N}^d$ is defined by $\Phi_A(w) \triangleq (|w|_{a_1}, \ldots, |w|_{a_d})$ where $|w|_a$ denotes the number of occurrences of a in w. For a (V, A)-tree T and an adjunct tree α where $X = R(\alpha)$, we can naturally extend the definition of the Parikh mapping as $\Phi_A(T) \triangleq \Phi_A(Y(T))$ and $\Phi_A(\alpha) \triangleq \Phi_A(Y(\alpha[X(\epsilon)]))$. By definition, we have $\Phi_A(\mathcal{L}(G)) = \Phi_A(\mathcal{T}(G))$ for any context-free grammar G. A set $S \subseteq \mathbb{N}^d$ is called *linear* if S is of the form

$$S = \{\boldsymbol{v_0} + x_1\boldsymbol{v_i} + \cdots + x_k\boldsymbol{v_k} \mid x_i \in \mathbb{N} \text{ for each } i\}$$

for some $k \in \mathbb{N}$ and some vectors $\boldsymbol{v_0}, \boldsymbol{v_1}, \ldots, \boldsymbol{v_k} \in \mathbb{N}^d$, and we call a finite union of linear sets *semilinear*.

2 Proof *à la* Takahashi

Definition (decomposition). A decomposition $\Delta(T)$ of a (V, A)-tree T is defined inductively as follows. If $T = a \in A \cup \{\epsilon\}$, define $\Delta(T) \triangleq (a, \emptyset)$. If $T = X(T_1, \ldots, T_n)$, let $(T'_1, S_1) = \Delta(T_1), \ldots, (T'_n, S_n) = \Delta(T_n)$ and define

$$\Delta(T) \triangleq \begin{cases} (X(T'_1, \ldots, T'_n), S_1 \cup \cdots \cup S_n) & X \notin N(T'_1) \cup \cdots \cup N(T'_n) \\ (T', \{\alpha\} \cup S_1 \cup \cdots \cup S_n) & X \in N(T'_1) \cup \cdots \cup N(T'_n) \end{cases}$$

where T' is the left-most X-rooted proper subtree of $X(T'_1, \ldots, T'_n)$, *i.e.*, the left-most X-rooted subtree of T'_i (where $X \in N(T'_i)$ and $X \notin N(T'_j)$ for each $1 \leq j < i$), and α is the adjunct tree obtained by replacing T' by X in $X(T'_1, \ldots, T'_n)$.

See Fig. 1 for example. The non-simple tree T_2 is decomposed as $\Delta(T_2) = (X(a), \{\alpha_1\})$; it is clear that $X(a)$ is the left-most X-rooted proper subtree of T_2 and $X(a) \vdash_{\alpha_1} T_2$.

Let $G = (V, D, X_0)$ be a context-free grammar over A.

Lemma. For any $T \in \mathcal{T}(G)$ and $(T', S) = \Delta(T)$, (1) T' is simple and $T' \in \mathcal{T}(G)$, (2) S is a set of simple adjunct trees, and (3) $T \in \mathrm{Adj}^+(T', S) \subseteq \mathcal{T}(G)$.

Proof. Straightforward induction on T.

We define $\mathcal{S}(G) \triangleq \{T' \mid (T', S) = \Delta(T) \text{ for some } T \in \mathcal{T}(G) \text{ and } S\}$ and define $\mathcal{A}(G) \triangleq \{\alpha \in S \mid (T', S) = \Delta(T) \text{ for some } T \in \mathcal{T}(G) \text{ and } S\}$. Because there are only finitely many simple (V, A)-trees (resp. simple adjunct trees), $\mathcal{S}(G)$ and $\mathcal{A}(G)$ are both finite by Claim (1)–(2) of Lemma.

Proposition (Takahashi [1]). $\mathcal{T}(G) = \bigcup_{T \in \mathcal{S}(G)} \mathrm{Adj}^*(T, \mathcal{A}(G))$.

Proof. Left-to-right inclusion $\subseteq$ is clear by Lemma. Right-to-left inclusion $\supseteq$ is shown by induction. The base case $T' \in \mathcal{S}(G) \subseteq \mathcal{T}(G)$ is trivial. Assume $T' \in \mathcal{T}(G)$. Then for any $\alpha \in \mathcal{A}(G)$ such that α is adjoinable to T', since α is extracted from some valid derivation tree in $\mathcal{T}(G)$, $T' \vdash_\alpha T''$ is also in $\mathcal{T}(G)$.

Theorem (Parikh [2]). $\Phi_A(\mathcal{L}(G))$ is semilinear.

Proof.

$$\Phi_A(\mathcal{L}(G)) = \Phi_A(\mathcal{T}(G)) = \bigcup_{T \in \mathcal{S}(G)} \bigcup_{S \subseteq \mathcal{A}(G)} \Phi_A(\mathrm{Adj}^+(T, S))$$

holds by Proposition. If S is not adjoinable to T then $\Phi_A(\mathrm{Adj}^+(T, S)) = \emptyset$. Otherwise, $\Phi_A(\mathrm{Adj}^+(T, S)) = \{\Phi_A(T) + \sum_{i=1}^{|S|} x_i \Phi_A(\alpha_i) \mid S = \{\alpha_1, \ldots, \alpha_{|S|}\}, x_i \in \mathbb{N} \setminus \{0\}\}$ holds since $T' \vdash_\alpha T''$ implies $\Phi_A(T'') = \Phi_A(T') + \Phi_A(\alpha)$. In both cases, $\Phi_A(\mathrm{Adj}^+(T, S))$ is semilinear, hence those finite union $\Phi_A(\mathcal{L}(G))$ is semilinear.

References

1. Takahashi, M.: A characterization of the derivation trees of a context-free grammar and an intercalation theorem. Master's thesis, University of Pennsylvania (1970)
2. Parikh, R.J.: Language generating devices. Quart. Prog. Rep. (60) (1961) 199–212