CS 4510-X

Parikh's Theorem

Name: Frederic Faulkner

1.1 Deliverables

There are 9 problems on this sheet (five during the proof and four at the end). Turn in 6 of the 9.

1.2 Introduction

Parikh's theorem states that context-free languages are equivalent to regular languages if you ignore the order of the letters in a word. We will prove the theorem together by working through several flawed proofs, and then explore some consequences of the theorem.

Definition 1.1 *Given an alphabet* Σ *with characters* $s_1, ..., s_{|\Sigma|}$*, define the function* $\Psi : \Sigma^* \to \mathbb{N}^{|\Sigma|}$ *by* $\Psi(w) = [n_1, ..., n_{|\Sigma|}]$ *where* n_i *is the number of occurrences of* s_i *in* w*. As an example, let* $\Sigma = \{a, b, c\}$*. Then* $\Psi(ababa) = [3, 2, 0]$*.*

Definition 1.2 *For a language* L*, define* $\Psi(L) = \{\Psi(w) \mid w \in L\}$

Theorem 1.3 (Parikh's Theorem) *For any context-free language* L*, there is a regular language* R *such that* $\Psi(L) = \Psi(R)$*. We say that* L *and* R *are "letter-equivalent".*

<u>Problem 1:</u> Let $L_1 = \{a^n b^n \mid n \in \mathbb{N}\}$ and let $L_2 = \{w \mid w \text{ is a palindrome}\}$. Give regular languages R_1 and R_2 such that $\Psi(L_1) = \Psi(R_1)$ and $\Psi(L_2) = \Psi(R_2)$.

How do we prove Parikh's theorem? The core idea of the proof relies on the concept from the pumping lemma of pumping downwards. Making a shorter string from a longer string allows us to do strong induction on the length of the string. Furthermore, there are a finite number of ways to pump a string downwards, and finite sets play nice with regular languages.

Definition 1.4 *Given a context-free language* L*, let* G *be a grammar that generates* L *and let* p *be the pumping length of* L*. Define* $B = \{w \in L \mid |w| < p\}$ *and* $C = \{xy \in \Sigma^* \mid |xy| \leq p$ *and for some nonterminal* A *in* G*,* $A \overset{*}{\Rightarrow} xAy\}$*.*

Dubious Claim 1.5 *Clearly* BC^* *is a regular language. Perhaps we could show that* $\Psi(L) = \Psi(BC^*)$*.*

<u>Problem 2:</u> Show that $\Psi(L) \subseteq \Psi(BC^*)$, using the pumping lemma and strong induction on the length of the string.

Now we try to prove $\Psi(BC^n) \subseteq \Psi(L)$ for all n by induction. Since $B \subseteq L$, $\Psi(B) \subseteq \Psi(L)$. Now, suppose $\Psi(BC^i) \subseteq \Psi(L)$. If $w \in BC^{i+1}$, we can write $w = w_0 s$, where $w \in BC^i$ and $s \in C$. By the inductive hypothesis, there is a word $w' \in L$ with $\Psi(w') = \Psi(w_0)$. We know that there is some nonterminal A such that $s = xy$ and $A \overset{*}{\Rightarrow} xAy$. Unfortunately, there is no way to complete this step, since the derivation of w doesn't necessarily contain A.[1]

So what do we do? Well, the inductive proof above would have worked out if we had a way of forcing the derivation of w to contain the terminals we need. This gives rise to the following idea: separate the words of L into subsets as determined by the nonterminals used in their derivations. Then, if we can show that each subset is letter-equivalent to a regular language, we will be done. (Why?)

Definition 1.6 *For any U, a subset of the nonterminals of G, define $L_U = \{w \mid w \in L$ and some derivation of w by G contains every nonterminal in $U\}$. (Note that $\cup_U L_U = L$.) Then define $B_U = \{w \in L_U \mid |w| < p\}$ and $C_U = \{xy \in \Sigma^* \mid |xy| \leq p$ and for some nonterminal A in U, $A \overset{*}{\Rightarrow} xAy\}$.*

Dubious Claim 1.7 $\Psi(L_U) = \Psi(B_U C_U^*)$

Problem 3: Following the proof outline at the bottom of page 1, show that $\Psi(B_U C_U^n) \subseteq \Psi(L_U)$ for $n \in \mathbb{N}$.

Unfortunately, now the inductive proof that you gave above for $\Psi(L) \subseteq \Psi(BC^*)$ doesn't work to show $\Psi(L_U) \subseteq \Psi(B_U C_U^*)$.

Problem 4: Why not? (Hint: pumping down doesn't quite work anymore, as we're no longer inducting over strings of L!)

We're almost there! But we need a slightly stronger version of the pumping lemma.

Theorem 1.8 (Stronger Pumping Lemma for Context-Free Languages) *Let L be a context-free language. Then there exists a p such that, if $w \in L$ and $|w| \geq p^k$, then there is some terminal A such that $S \overset{*}{\Rightarrow} uAz \overset{*}{\Rightarrow} uv_1Ay_1z \overset{*}{\Rightarrow} uv_1v_2Ay_2y_1z \overset{*}{\Rightarrow} ... \overset{*}{\Rightarrow} uv_1v_2...v_kAy_k...y_2y_1z \overset{*}{\Rightarrow} uv_1v_2...v_kxy_k...y_2y_1z = w$, with $|v_1v_2...v_kxy_k...y_2y_1| \leq p^k$. Note that for $k = 1$ this is just the ordinary pumping lemma.*

The proof of this stronger form of the pumping lemma follows the same structure as the proof of the normal pumping lemma: if a derivation tree is tall enough, it must contain a long path, and a long enough path must contain some nonterminal repeated at least $k + 1$ times.

Now, let $k = |U|$ and after the following definitions we can finally prove Parikh's theorem.

Definition 1.9 *Let $B'_U = \{w \in L_U \mid |w| < p^k\}$ and $C'_U = \{xy \in \Sigma^* \mid |xy| \leq p^k$ and for some nonterminal A in U, $A \overset{*}{\Rightarrow} xAy\}$.*

Correct Claim 1.10 $\Psi(L_U) = \Psi(B'_U C'^*_U)$

Problem 5: $\Psi(B'_U C'^*_U) \subseteq \Psi(L_U)$ follows almost unchanged from the proof you gave in problem 3. Now, using the Stronger Pumping Lemma for CFL's, prove the reverse direction, i.e. that $\Psi(L_U) \subseteq \Psi(B'_U C'^*_U)$. (The proof closely follows the original proof in problem 2, but now you can pump down in a certain way to stay in L_U.)

[1] Of course, failing to prove a statement doesn't mean that the statement is false. Can you give an example of a grammar G such that $\Psi(BC^*) \not\subseteq \Psi(L(G))$?

1.3 Additional Problems

- Problem 6: Show that L_U is in fact a context-free language.

- Problem 7: Use Parikh's theorem to show that every context-free language over a unary alphabet is regular.

- Problem 8: A set S is linear if for some fixed $a_0, a_1, ..., a_n$, we can write $S = \{a_0+x_1a_1+x_2a_2+ ... + x_na_n \mid x_i \in \mathbb{N}\}$. A set is semi-linear if it can be written as the union of linear sets. Show that if L is context-free, $\Psi(L)$ is semi-linear. (Hint: use the fact that $\Psi(L_U) = \Psi(B'_U C'^*_U))$

- Problem 9: Show that if $\Psi(L)$ is semi-linear, then there is a regular language R such that $\Psi(R) = \Psi(L)$. (Hint: start with the linear case, then use the union property for regular languages.)