Converses of Pumping Lemmas

Richard Johnsonbaugh
David P. Miller
Department of Computer Science and Information Systems
DePaul University
Chicago, IL 60604

ABSTRACT

Pumping lemmas appear in courses that study formal languages such as automata theory and the theory of computation. Converses of pumping lemmas, which are generally false, are ignored by most of the books that treat formal languages. This is unfortunate since converses of pumping lemmas arise in a natural way and students typically ask whether converses of particular pumping lemmas are true. We give counterexamples to the converse of a pumping lemma for regular languages and to the converse of Ogden's Lemma, a pumping lemma for context-free languages. We also show that converses to these lemmas are true for languages over a single symbol. We conclude by discussing the counterexample to the converse of Ogden's Lemma with reference to Parikh's necessary condition for a language to be context-free.

INTRODUCTION

A pumping lemma for a class of languages $\mathcal{L}$ states that if a sufficiently long string belongs to some language $L \in \mathcal{L}$, then all strings of a particular form must also belong to L. A frequent application of a pumping lemma is to show that a set L does not belong to a particular class of languages $\mathcal{L}$ by assuming that $L \in \mathcal{L}$ and then deducing a contradiction by showing that the conclusion of the pumping lemma fails. Students frequently raise the following question: Suppose that I try to prove a language L is not in $\mathcal{L}$ by using the pumping lemma, but instead I find that the conclusion of the pumping lemma is true for L. Can I then conclude that $L \in \mathcal{L}$? In other words, is the converse of the pumping lemma true? We will show that converses of two pumping lemmas are false – one for regular languages and Ogden's Lemma for context-free languages.

DEFINITIONS

We use the following definitions and terminology from [HU79]. A finite set Σ is called an *alphabet*. The set Σ^* consists of all strings (including the null string ε) in the symbols Σ. A *language in* Σ is a subset of Σ^*. The *length of a string* x is denoted $|x|$. If x and y are strings, xy is the *concatenation* of x and y. If $L, L_1, L_2 \subseteq \Sigma^*$, we define

$$\begin{aligned} L_1 L_2 &= \{xy \mid x \in L_1, y \in L_2\} \\ L^0 &= \{\varepsilon\} \\ L^i &= LL^{i-1}, i \geq 1 \\ L^* &= \bigcup_{i=0}^{\infty} L^i \\ L^+ &= \bigcup_{i=1}^{\infty} L^i \end{aligned}$$

If $a \in \Sigma$ or $a = \varepsilon$, we denote the set $\{a\}$ by a.

A *context-free grammar* G consists of a finite set N of *variables*, a finite set Σ of *terminals* ($N \cap \Sigma = \emptyset$), a *start symbol* $S \in N$, and a finite subset P of $N \times (N \cup \Sigma)^*$ of *productions*. A context-free grammar is *regular* if every production is of the form (X, wY), where $X \in N$, $w \in \Sigma^*$, and $Y \in N \cup \varepsilon$. If G is a context-free grammar, the set $L(G)$ of strings in Σ^* derivable from S using the productions P is called the *language generated by* G. We say that a subset L of Σ^* is a *context-free* (respectively, *regular*) *language* if there is a context-free (respectively, regular) grammar G such that $L = L(G)$.

A PUMPING LEMMA FOR REGULAR LANGUAGES

We first discuss a pumping lemma for regular languages.

Lemma 1 (Pumping Lemma for Regular Languages [HU79]) *If L is a regular language, there exists a positive integer n such that if $z_1z_2z_3 \in L$ and $|z_2| = n$, there are strings u, v, w, $|v| \geq 1$, such that $z_2 = uvw$ and $z_1uv^iwz_3 \in L$ for every $i \geq 0$.*

We begin constructing a counterexample to the converse of Lemma 1 by defining the set

$$A_1 = \bigcup_{i=1}^{\infty}(10^+)^i(20^+)^i2$$

By considering the string $(10)^n(20)^n2$ in conjunction with Lemma 1, we see that A_1 is not regular. Next let A_2 denote the set of all strings over $\{0,1,2\}$ that contain at least one of 11, 12, or 22 as a substring. Clearly A_2 is a regular language. We show that $A = A_1 \cup A_2$ is a counterexample to the converse of Lemma 1.

We first show that A is not regular. Suppose, by way of contradiction, that A is regular. Then $A_1 = A \cap \overline{A_2}$ is a regular language since regular languages are closed under intersections and complementation [HU79]. ($\overline{\ }$ denotes complementation.) This contradiction shows that A is not regular.

We show that A satisfies the conclusion of Lemma 1. Since A_2 is a regular language, it satisfies the conclusion of Lemma 1. Let n' be the constant of Lemma 1 for A_2 and let $n = \max(n', 3)$. Let $z = z_1z_2z_3 \in A$ and $|z_2| = n$. If $z \in A_2$, we may satisfy the conclusion of Lemma 1 since A_2 is itself regular. Suppose that $z \in A_1$. Notice that z_2 contains 0. Let $v = 0$ and define u and w so that $z_2 = uvw$. Certainly, $z_1uv^iwz_3 \in A_1$ for every $i \geq 1$. If the last member of u or the first member of w is 0, $z_1uv^0wz_3 \in A_1$. If the last member of u is not 0 and the first member of w is not 0, uv^0w contains one of the substrings 11, 12, or 22. In this case, $z_1uv^0wz_3 \in A_2$. Therefore A satisfies the conclusion of Lemma 1.

An interesting fact is that the converse of Lemma 1 is true for context-free languages over one symbol. This result is well-known (see [GR62]), but we have not found a proof in the literature.

We show that if $L \subseteq a^*$ and L satisfies the conclusion of Lemma 1, then L is regular.

Let n be as in Lemma 1. We will define a finite number of sets $L_{p,q,r}$, where $0 \leq p < q \leq n$ and r depends on p and q, such that $L_{p,q,r} \subseteq L$. Moreover, we will arrange that for each $m \geq n$, if $a^m \in L$, then there are p, q, and r such that $L_{p,q,r}$ has been defined and $a^m \in L_{p,q,r}$.

Let $a^m \in L$, $m \geq n$. Write $a^m = a^na^{m-n}$. By Lemma 1, there are strings u, v, w, $|v| \geq 1$, such that $a^n = uvw$ and $uv^iwa^{m-n} \in L$ for every $i \geq 0$. If we let $q = |v|$, then $1 \leq q \leq n$ and $a^{m+iq} \in L$ for every $i \geq -1$. Write

$$m = kq + p, \quad 0 \leq p < q$$

Then, for $i \geq k$,

$$a^{p+iq} = a^{m-kq+iq} = a^{m+(i-k)q} \in L$$

Let r be the least integer such that $a^{p+iq} \in L$ for every $i \geq r$. Define

$$L_{p,q,r} = \{a^{p+iq} \mid i \geq r\}$$

Then

$$a^m = a^{p+kq} \in L_{p,q,r} \subseteq L$$

as desired.

It follows that L is the finite union of M, a finite set of strings each of length at most n, and sets of the form $L_{p,q,r}$. Since each of the sets $L_{p,q,r}$ and M is regular, L is a regular language.

Although our counterexample to the converse of Lemma 1 involved three symbols, there are counterexamples that use only two symbols. We leave the construction of such a counterexample to the reader.

In [EPR81], Ehrenfeucht, et al., give a strengthened form of Lemma 1 whose converse is true.

A PUMPING LEMMA FOR CONTEXT-FREE LANGUAGES

Methods similar to those of the previous section can be used to give a counterexample to the converse of Ogden's Lemma [Ogd68], a strong form of a pumping lemma for context-free languages. Our non-context-free language L has the interesting property that almost any single symbol can be "pumped" with the resulting string remaining in L.

Before we discovered our counterexample to the converse of Ogden's Lemma, we did not know whether the converse was true or false. We subsequently found that the problem was solved (see [BH78]), but that the source is not easily obtained. Thus we feel that it will be useful to teachers of courses that treat formal languages to have an easily accessible counterexample.

Lemma 2 (Ogden [Ogd68]) *If L is a context-free language, there is an integer n such that if $z \in L$ and n or more positions in z are designated distinguished, then z may be written $z = uvwxy$ such that*

(a) *w contains at least one distinguished position;*

(b) *either both u and v contain distinguished positions or both x and y do;*

(c) *vwx contains at most n distinguished positions;*

(d) *$uv^iwx^iy \in L$ for every $i \geq 0$.*

We begin by showing that the set

$$L_1 = \bigcup_{i=1}^{\infty} (e^+a^+d^+)^i(e^+b^+d^+)^i(e^+c^+d^+)^i$$

is not a context-free language. If L_1 is a context-free language, there is an integer n as in Ogden's Lemma. Let $z = (ead)^n(ebd)^n(ecd)^n$. Mark the b's as distinguished. Suppose that $z = uvwxy$ satisfies (a)-(d) of Ogden's Lemma. First, consider the case that both u and v contain distinguished positions. In particular, v contains b but v contains no substring of $(ead)^n$. By (d), $z' = uwy \in L_1$. But z' contains exactly n occurrences of ea but less than n b's. Therefore, $z' \notin L_1$. Thus, the first case cannot occur. By (b), both x and y contain distinguished positions. In a similar way, we derive a contradiction in this case also. Therefore, L_1 is not a context-free language.

Next, we let L_2 be the set of strings over $\{a, b, c, d, e\}$ that do not begin with e or do not end with d or contain at least one of the substrings ed, da, db, dc, ae, be, or ce. We note that $L_1 \cap L_2 = \emptyset$. Clearly, L_2 is a regular language. We show that $L = L_1 \cup L_2$ is a counterexample to the converse of Ogden's Lemma.

First, we note that L is not context-free. For suppose, by way of contradiction, that L is context-free. Then $L_1 = L \cap \overline{L_2}$ is context-free since context-free languages are closed under intersections with regular languages [HU79]. This contradiction shows that L is not context-free.

It remains to show that L satisfies the conclusion of Ogden's Lemma. Since L_2 is context-free, it satisfies the conclusion of Ogden's Lemma. Let n' be the constant of Ogden's Lemma for L_2 and let $n = \max(n', 3)$. We show that L satisfies the conclusion of Ogden's Lemma for n.

Let $z \in L$ and designate n or more positions distinguished. Clearly, if $z \in L_2$, we may satisfy the conclusion of Ogden's Lemma. Suppose that $z \in L_1$. Let v be the second distinguished position from the right and let $x = y = \varepsilon$. Define u and w so that $z = uvwxy$. Part (a) of the Lemma is satisfied since w contains the rightmost distinguished position. Since vwx contains exactly two distinguished positions, part (c) is satisfied. Since u and v contain distinguished positions, part (b) is satisfied. Finally, for all $i \geq 1$, $uv^iwx^iy \in L_1$. If v is not adjacent to an identical character, $uv^0wx^0y \in L_2$, and $uv^0wx^0y \in L_1$ otherwise. Thus, part (d) is also satisfied.

By arguing as in the previous section, one can show that if $L \subseteq a^*$ and L satisfies the conclusion of Ogden's Lemma, then L is a regular language.

PARIKH'S THEOREM

Parikh [Par66] gave a necessary condition for a language L to be context-free in terms of the distribution of the symbols in the strings of L. We conclude by discussing the converse of Parikh's result.

A set S of n-tuples of nonnegative numbers is *linear* if for some k, there exist n-tuples $v_0, \ldots, v_k$ such that

$$S = \{v_0 + \sum_{i=1}^{k} m_i v_i \mid m_i \text{ are nonnegative integers}\}$$

A set of n-tuples is *semi-linear* if it is the finite union of linear sets. Let $\Sigma = \{a_1, \ldots, a_n\}$. If $z \in \Sigma^*$, let $f_i(z)$ denote the number of occurrences of the symbol a_i in z. The *Parikh map* q is defined by

$$q(z) = (f_1(z), \ldots, f_n(z)), z \in \Sigma^*$$

If L is a subset of Σ^*, we define

$$q(L) = \{q(z) \mid z \in L\}$$

We may now state Parikh's result.

Theorem 3 (Parikh [Par66]) *If L is a context-free language, then $q(L)$ is semi-linear.*

Wise [Wis76] gave an example of a non-context-free language L for which $q(L)$ is semi-linear, thus showing that the converse of Parikh's Theorem is false. The non-context-free language L of the previous section gives a particularly dramatic counterexample to the converse of Parikh's Theorem: all distributions of symbols appear. More precisely, $q(L)$ is the linear set consisting of *all* 5-tuples of nonnegative integers.

CONCLUSIONS

Converses of pumping lemmas should be considered in courses and books that deal with formal languages since the issue of whether the converse of a pumping lemma for a class of languages $\mathcal{L}$ is true will arise as one tries to use the pumping lemma to show that various languages do not belong to $\mathcal{L}$.

References

[BH78] L. Boasson and S. Horvath. On languages satisfying Ogden's Lemma. *R.A.I.R.O. Theoretical Comp. Sci.*, 12:201–202, 1978.

[EPR81] A. Ehrenfeucht, R. Parikh, and G. Rozenberg. Pumping lemmas for regular sets. *Siam J. Comput.*, 10:536–541, 1981.

[GR62] S. Ginsburg and H. G. Rice. Two families of languages related to ALGOL. *J. ACM*, 9:350–371, 1962.

[HU79] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 1979.

[Ogd68] W. Ogden. A helpful result for proving inherent ambiguity. *Math. Systems Theory*, 2:191–194, 1968.

[Par66] R. J. Parikh. On context-free languages. *J. ACM*, 13:570–581, 1966.

[Wis76] D. S. Wise. A strong pumping lemma for context-free languages. *Theoretical Comp. Sci.*, 3:359–369, 1976.